

ВСТУП

В сучасний час енергія, а саме енергозабезпечення є невід'ємним чинником не тільки для забезпечення благополучного та комфортного життя населення, а також для зросту соціально-економічного стану у всіх країнах світу. Можна з упевненістю сказати у сучасному розумінні, що існує прямо пропорційна залежність між якістю життя та споживанням енергії. На даний момент людство використовує велику кількість енергії для виготовлення харчових продуктів, для приємного проведення вільного часу, та в багатьох інших видів діяльності, які асоціюються з сучасним способом життя.

Не беручи до уваги те, що енергозберігаючі технології у країнах, які розвинуті на високому рівні та з високою якістю життя та те, що значну кількість виробництва було переведено в країни Азії, а також те, що країни, які на даний момент розвиваються, мають низьку енергоефективність, саме споживання в цих країнах є низьким. У співвідношенні між високорозвиненими країнами та країнами, що розвиваються, розвинуті країни споживають у середньому у 5-6 разів більше енергії. Очевидним є те, що збільшення використання енергії супроводжується підвищенням рівня життя.

Зараз сучасний світ вступає до нового етапу розвитку, коли після безперервного використання великих кількостей енергії, благополуччя життя населення почало відриватись від зростання споживання енергії. У багатьох економічно та технологічно високорозвинених країнах світу кількість споживання енергії стабілізуються, а в деяких вже почали знижуватись. Прикладом може бути Данія, яка найбільш скоротила цю кількість з 91% до 69% і на даний момент являється передовою країною світу, яка використовує відновлювані джерела енергії, головним чином за допомогою вітру.

У сучасному суспільстві існують головні запити до систем енергозабезпечення, такі як: доступність енергії у достатній кількості та за прийнятними цінами, висока

безпека при споживанні енергії та висока екологічність. Світова енергетична рада об'єднала напрями енергетичної політики у три групи, завдяки яким чиновники різних країн світу прагне знайти загальний баланс:

1) Забезпечення енергобезпеки. Для цього потрібно забезпечити надійний імпорту, підтримку нових видів та джерел енергії, подолання енергетичної бідності, а також – відмовитись від джерел енергії, які вважаються небезпечними.

2) Доступність. Для цього потрібно зробити енергії загальнодоступною за прийнятними цінами для споживачів різних категорій.

3) Екологія та стабільний розвиток. Це може означати те, що потрібно зменшити кількість шкідливих викидів та стимулювати ефективне споживання енергії.

Дана дипломна робота виконується по плану наукових робіт Інституту демографії та соціальних досліджень ім. М.В. Птухи НАН України, що підтверджує актуальність даної теми. Також актуальність розробки даного програмного забезпечення для даної теми дипломної роботи полягає у тому, що немає систем, які виконують аналіз відразу відносно існуючих даних. Тому були прийнято рішення створити відповідну систему.

1 ПОСТАНОВКА ЗАДАЧІ

Метою даної дипломної роботи є створення математичного та програмного забезпечення системи для оцінювання взаємозв'язку енергоспоживання та якості життя населення.

Для вирішення поставленої задачі необхідно розв'язати такі завдання:

- а) Провести огляд та якісний аналіз існуючих рішень;
- б) Провести огляд та якісний аналіз математичних методів розв'язання задачі, які будуть використовуватись для оцінювання взаємозв'язку якості життя та енергозабезпечення;
- в) Підготувати вхідні дані для аналізу;
- г) Розробити програмне забезпечення для реалізації обраного методу;
- д) Провести тестування розробленого продукту, а також аналіз для отриманих результатів.

2 ОГЛЯД ІСНУЮЧИХ РІШЕНЬ ДЛЯ ПОСТАВЛЕНОЇ ЗАДАЧІ

2.1 Огляд існуючих рішень

2.1.1 Дослідження програми розвитку Організації Об'єднаних Націй

Далеко не секрет, що індекс людського розвитку (ІЛР) сильно залежить від енергоспоживання і чим краще країна буде забезпечена різними видами енергії, які мають позитивний вплив на розвиток держави в цілому та не забруднюють навколишнє середовище, тим вище буде індекс людського розвитку країни.

Даний факт підтверджує доповідь програми розвитку Організації Об'єднаних Націй (ПРООН) про людський розвиток за 2015 рік.

Звіт про людський розвиток, як правило, концентрується не на економічному стані країни, а саме на людському розвитку. Даний звіт починається з чіткого питання: як саме можна сприяти розвитку людства? У звіті підкреслюється дивовижний прогрес розвитку людства за останні чверть століття. Сьогодні люди живуть набагато довше, більше дітей мають можливість отримати освіту, більше людей мають доступ до очищеної води та основних санітарних умов. Також можна впевнено сказати, що доходи на душу населення у світі зросли, а відповідно знизилась бідність населення, що призвело до покращення життя багатьох людей. Сучасний час – це час цифрових технологій. Цифрова революція змогла з'єднати людей з різних куточків світу.

Звіт про людський розвиток за 2015 рік є продуктом бюро звітів про людський розвиток (HDRO) ПРООН [1].

Відповідно до тематики поставленої задачі в даному звіті є співвідношення споживання електроенергії та індексом людського розвитку, а саме існує сильний позитивний зв'язок між споживанням енергії та індексом людського розвитку для країн, що розвиваються (рис. 2.1) [1].

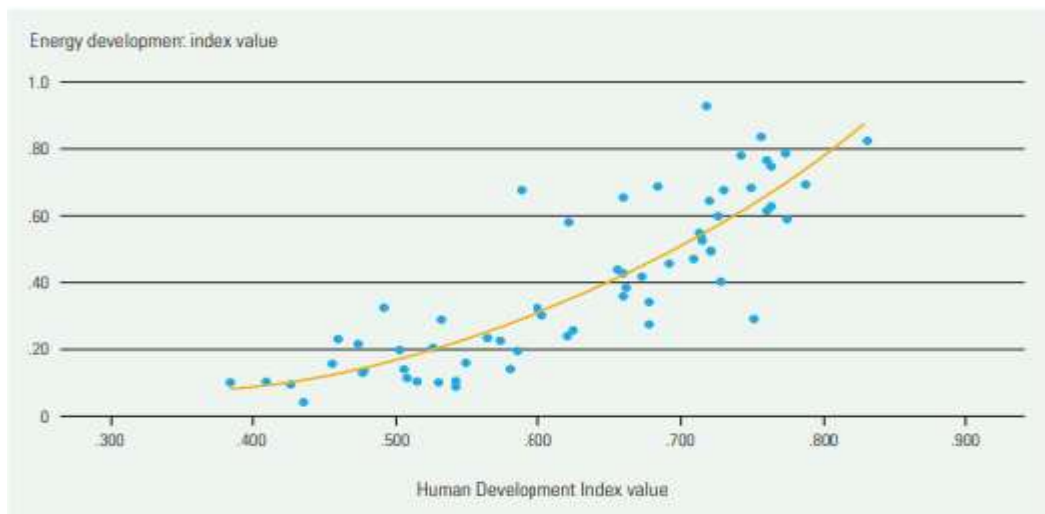


Рисунок 2.1 – Взаємозв'язок споживання енергії та індексу людського розвитку[1]

Дивлячись на рисунок 2.1 не важко побачити, що разом із збільшенням енергоспоживання, а саме, як вказано на графіку, індексу енергетичного розвитку збільшується індекс розвитку людства. Завдяки цьому можна впевнено говорити те, що дана залежність існує. Це також підтверджується великою кількістю емпіричних доказів щодо доступу до сучасних енергетичних послуг, покращення здоров'я, зменшення бідності та підвищення рівня життя.

Те, як було побудовано даний графік та які методи було використано для того, щоб його побудувати, інформації в даній роботі не наведено.

2.1.2 Дослідження токійського технологічного інституту

В даному дослідженні будується певна функція для якості життя за допомогою факторного аналізу, де факторний аналіз являється статистичним методом, який аналізує вплив окремих факторів на результат. Після проведених нормалізацій процесів рівнянь враховуються показник якості життя щодо споживання електроенергії на душу населення та споживання електроенергії на душу населення.

На рисунку 2.2 та рисунку 2.3 показано споживання електроенергії та енергії на душу населення відносно показника якості життя за 2013 рік.

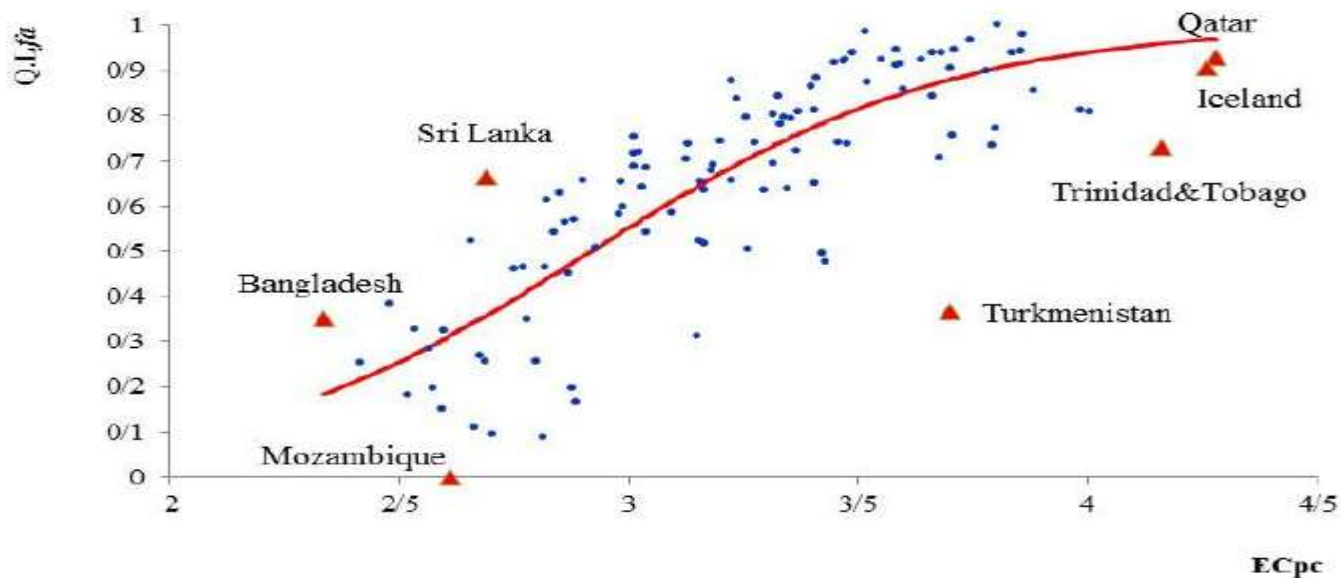


Рисунок 2.2 – Показник якості життя у зв'язку з споживанням енергії на душу населення[2]

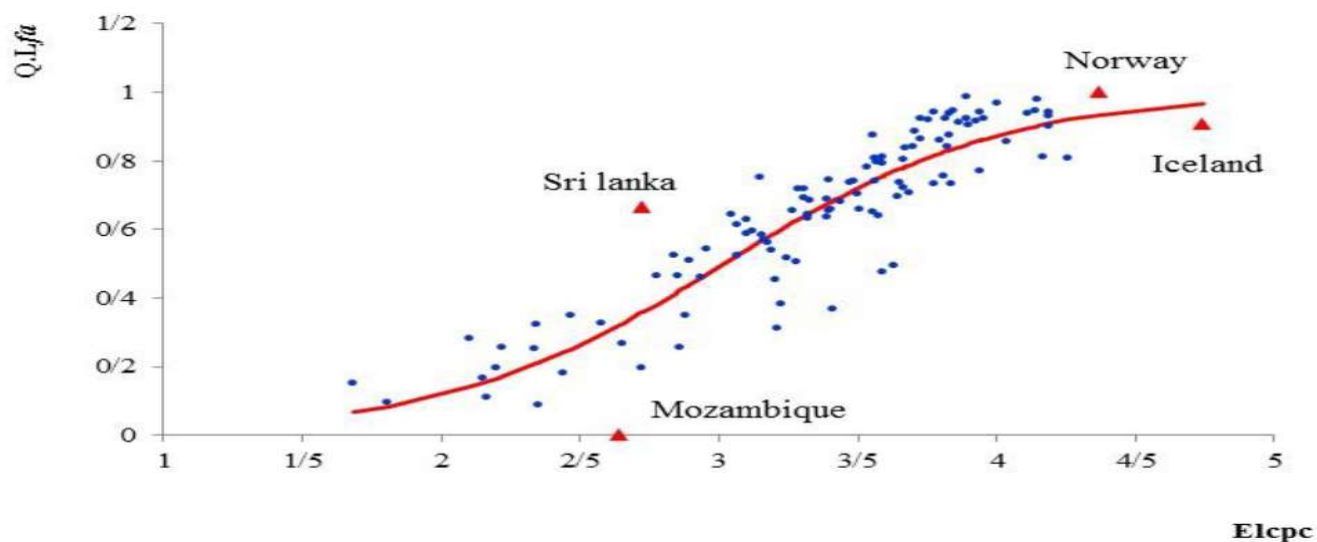


Рисунок 2.3 - Показник якості життя у зв'язку з споживанням електроенергії на душу населення[2]

Відносно вищевказаних рисунків було проведено аналіз. Країни, що розташовані в правому нижньому куті під сигмоїдальною кривою показують

неефективне споживання енергії та електроенергії на душу населення та низьку якість життя населення. З іншого боку, в лівому верхньому куті над кривою знаходяться країни, які демонструють відносно високу якість життя, але показують не високе використання енергії та електроенергії.

Також можна спостерігати те, що в таких країнах, як Ісландія, Катар та Тринідад і Тобаго споживання енергії на душу населення порівняно вище, чим у країнах, які мають схожий показник якості життя. Такими країнами є Шрі-Ланка та Бангладеш. Дані країни мають незвичну поведінку, оскільки вони мають відносно високу якість життя при невисокому споживанні енергії на душу населення. Такі країни, як Мозамбік та Туркменістан є неефективними країнами, оскільки дані країни споживають велику кількість енергії на душу населення, хоча рівень життя населення являється низьким.

Із споживанням електроенергії на душу населення Ісландія та Норвегія знаходяться в самій віддаленій точці кривої. З розташування Мозамбіку зрозуміло, що він знаходиться в неефективній області, оскільки високе споживання енергії не призвело до підвищення якості життя населення. Шрі-Ланка проявляє ефективну поведінку відносно споживання енергії на душу населення та значення якості населення, оскільки при споживанні малої кількості електроенергії отримується високе значення якості життя населення.

На рисунку 2.2 та рисунку 2.3 видно, що у відповідності показників якості життя до споживання енергії та електроенергії на душу населення сигмоїдальна крива починається повільно, потім рухається до фази швидкого розвитку і після цього починається плато. Дану функцію можна класифікувати на 3 фази, які відповідають трьом типам країн у світі (рис. 2.4), а саме:

- 1) слаборозвинуті країни;
- 2) країни, що розвиваються;
- 3) розвинуті країни.

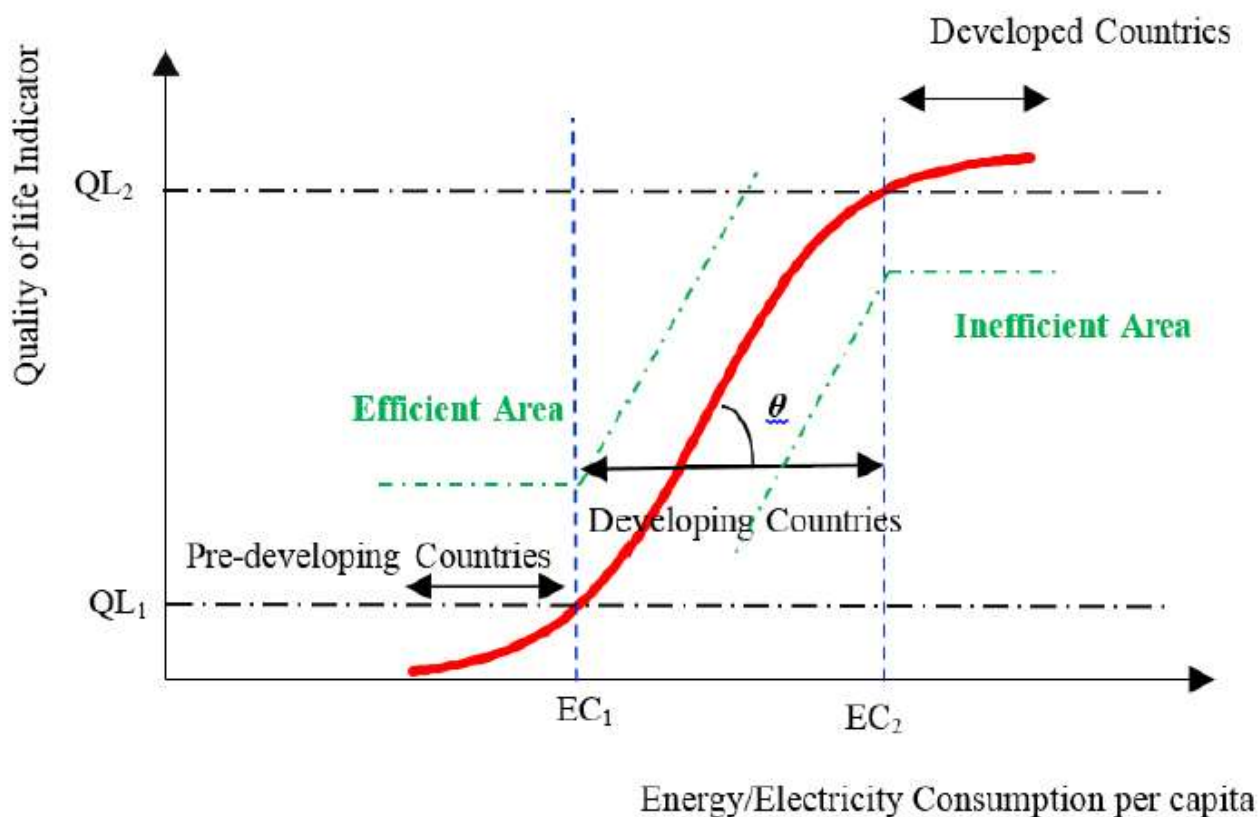


Рисунок 2.4 – Три області відносно відношення якості життя та споживання енергії/електроенергії на душу населення[2]

Повільний ріст кривої можна спостерігати на початку графіку. Це зумовлено тим, що в слаборозвинутих країнах існує недостача засобів інфраструктури і недостатньо знань, що заважає і викликає повільний ріст показника якості життя.

Швидкий ріст показника якості життя можна спостерігати в середній фазі сигмоїдальної функції. Саме ця область належить країнам, що розвиваються. Це означає те, що в цих країнах є доступ до технологій з високою ефективністю, що значно змінює якість життя в цих країнах. Також можна помітити те, що в даних країнах впровадження нових технологій значно збільшує споживання енергії/електроенергії в країнах, що розвиваються.

Неважко побачити і те, що розвинуті країни знаходяться на плато сигмоїдальної функції, на якій показники якості життя населення досягнули свого максимального рівня.

Для того, щоб класифікувати спостереження на три групи, яке засноване на споживанні енергії/електроенергії на одну людину і якості життя, використовується ієрархічний метод, який заснований на основі алгоритму кластерного аналізу, який називається K-Means.

У 2013 році споживання енергії на душу населення в слаборозвинутих країнах становить 509 кілограмів нафтового еквівалента(kg_{oe}), коли для розвинутих країн це значення становить щонайменше 5470 кілограмів нафтового еквівалента. Для країн, що розвиваються це значення коливається від 509 до 5470 кілограмів нафтового еквівалента на душу населення. Подібне тлумачення можна провести для споживання електроенергії на душу населення [2]. Класифікація країн за енергетикою та споживанням енергії на душу населення разом із значенням якості життя в відповідних країнах наведено у таблиці 2.1.

Таблиця 2.1 – Класифікація країн на основі (якості життя, споживання енергії (kg_{oe})/електроенергії(kWh) на душу населення)[2]

Модель	Слаборозвинуті країни	Країни, що розвиваються	Розвинуті країни
Енергія	$0 < \dots \leq 0.32, \leq 483$	$(0.32 < \dots < 0.87, 483 < \dots < 5204)$	$0.87 \leq \dots \leq 1, \geq 5204$
Електрика	$0 < \dots \leq 0.27, \leq 227$	$(0.27 < \dots < 0.87, 227 < \dots < 8158)$	$0.87 \leq \dots \leq 1, \geq 8158$

2.1.3 Порівняння існуючих рішень

В таблиці 2.2 демонструються переваги та недоліки приведених відповідно до 2.1.1 та 2.1.2 вже існуючих рішень для поставленої задачі.

Таблиця 2.2 – Порівняння існуючих рішень в області поставленої задачі

Рішення Категорія порівняння	Дослідження ПРООН	Дослідження токійського технологічного інституту
Тип надання інформації	Графічно	Графічно та за допомогою чисельних обчислень, де результати занесені в відповідні таблиці
Детальність надання інформації	Текстове описання тематики та надання графічного представлення результату без детального опису	Текстове описання тематики, надання достатньої кількості результатів за допомогою графічних представлень для поставленої задачі та чисельні обчислення, які занесені в відповідні таблиці, що підтверджують правильність графічних представлень.
Описання отриманих результатів	Не надано	Надано в детальному описі

2.2 Порівняльний аналіз підходів

2.2.1 Засоби для виконання поставленої задачі та їх порівняння

Поставлену задачу можна вирішувати відразу в декількох системах, а саме в яких можна проводити відповідні аналізи для поставленої задачі. У таблиці 2.3 наведено приклади найвідоміших таких систем, які виконують потрібні дії.

Таблиця 2.3 – Аналіз існуючих засобів

Категорія Система	Ціна	Студентська підписка	Зручність	Спосіб подання результатів
MATLAB (для студентів)	55\$ / місяць	-	Велика кількість методів для вирішення поставленої задачі	Графіки, чисельні дані
SPSS	99\$ / місяць	-	Можливість виконувати аналіз даних	Графіки, чисельні дані
PyCharm	19.90\$ / місяць	+ (безкоштовна ліцензія з можливістю оновлення кожного року)	Велика кількість бібліотек для аналізу даних та можливостей роботи з даними	Графіки, чисельні дані, виведення результатів в браузері

Продовження таблиці 2.3

Категорія Система	Ціна	Студентська підписка	Зручність	Спосіб подання результатів
Microsoft Excel	139.99\$	+(Для Office 365, потрібна пошта університету)	Потрібно використовувати спеціальні налаштування, немає великої кількості можливостей	Графіки

2.2.2 Методи для виконання поставленої задачі

Поставлена задача, а саме створення математичного та програмного забезпечення для оцінювання взаємозв'язку енергоспоживання та якості життя населення розв'язується за допомогою наступних методів, за допомогою яких являється можливим проведення аналізу даних.

Такими методами являються кластерний аналіз, факторний аналіз, регресійний аналіз та кореляційний аналіз.

Кластерний аналіз – це техніка машинного навчання, яка має на меті групування даних. Алгоритм кластеризації використовується для того, щоб класифікувати дані в певні групи. Таким шляхом можна визначити дані, які мають схожі властивості, оскільки вони будуть знаходитись в одній групі, тоді як дані, які мають інші властивості, будуть знаходитись в інших групах [3].

Для кластеризації часто потрібно визначити відстань між об'єктами. Найбільш популярним типом відстані являється Евклідова відстань, яка має наступний загальний вигляд:

$$\sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (2.1)$$

Факторний аналіз – це спосіб скоротити дані до меншого набору даних та класифікувати признаки. Також факторний аналіз використовується для створення набору змінних для схожих елементів в наборі даних, тобто використовується для вивчення взаємозв'язку між значеннями елементів. Факторний аналіз також використовується для того, щоб визначати признаки і, в ідеалі, будується на інтервальних ознаках та на метричних шкалах [4].

Модель факторної системи можна представити у вигляді наступної формули:

$$y = f(x_1, x_2, \dots, x_n), \quad (2.2)$$

де y – результативний показник, а x – факторні признаки.

Регресійний аналіз – це метод аналізу даних, який оснований на визначенні відокремлених впливів факторів на результативну ознаку, також за допомогою використання певних критеріїв на кількісну оцінку впливів факторів.

Для проведення даного аналізу будується рівняння регресії та визначається те, як впливає кожна незалежна змінна на варіацію досліджуваної (прогнозованої) залежної змінної величини [5].

Представлення рівняння регресії має наступний вигляд:

$$Y_x = f(x_1, x_2, \dots, x_n), \quad (2.3)$$

де Y_x – залежна змінна величина, а x – незалежні змінні величини (фактори).

Кореляційний аналіз – це метод, за допомогою якого досліджується взаємозалежність ознак у генеральній сукупності, які являються випадковими величинами з нормальним характером розподілу. Також використання кореляційного аналізу дуже пов'язане з регресійним аналізом даних, тому часто говорять – «Кореляційно-Регресійний аналіз» [5].

Для того, щоб виконати кореляційний аналіз, як це описано в [5] для початку необхідно виконати оцінювання типовості та однорідності даних спостереження. Це можливо зробити при використанні таких критеріїв як середньоквадратичне відхилення (σ) та коефіцієнт варіації (V).

Середньоквадратичне відхилення обчислюється наступним чином:

$$\sigma = \sqrt{\frac{\sum(x-\bar{x})^2}{n}}, \quad (2.4)$$

а коефіцієнт варіації обчислюється наступним чином:

$$V = \frac{\sigma}{\bar{x}} \quad (2.5)$$

Далі вимірюється тіснота зв'язку за допомогою лінійного коефіцієнту кореляції (Пірсона) r за наступною формулою:

$$r = \frac{\overline{xy} - \bar{x} * \bar{y}}{\sigma_x \sigma_y} \quad (2.6)$$

Залежність тісноти зв'язку від величини коефіцієнту кореляції наведена у таблиці 2.4, яка також має назву «Таблиця Чеддока» [5].

Таблиця 2.4 – Величина коефіцієнта кореляції та тіснота зв'язку (таблиця Чеддока)

1,00	Зв'язок функціональний
0,90 — 0,99	Дуже сильний
0,70 — 0,89	Сильний
0,50 — 0,69	Значний
0,30 — 0,49	Помірний
0,10 — 0,29	Слабкий
0,10 — 0,29	Слабкий

2.2.3 Порівняння методів розв'язання задачі

В таблиці 2.5 приведено порівняння методів (див. 2.2.2) за певними категоріями для розв'язання поставленої задачі.

Таблиця 2.5 – Порівняння методів для розв'язання задачі

Метод Категорія	Кластерний аналіз	Факторний аналіз	Регресійний аналіз	Кореляційний аналіз
Мета	Групування даних	Виявлення факторів та оцінка їх впливу	Визначення впливу факторів на результат	Виявлення існування залежності однієї змінної від інших
Способи	Два основні способи: ієрархічні процедури та ітераційні процедури	Ланцюгові підстановки, різниця абсолютних величин, балансовий, індексний та інтегральний методи	Парний (простий) регресійний аналіз та регресійний аналіз, що основується на множинний регресії, або багатофакторний	Лінійний, ранговий, парний та множинний
Області використання	Маркетинг, менеджмент, медицина, соціологія	В різних науках: психологія, економіка тощо	Використовується для прогнозування, дослідження та моделювання в різних сферах	Економіка, соціологія, психологія, медицина, біометрія та в інших сферах

Продовження таблиці 2.5

Метод Категорія	Кластерний аналіз	Факторний аналіз	Регресійний аналіз	Кореляційний аналіз
Основні формули	Евклідова відстань за формулою (2.1)	Модель факторної системи за формулою (2.2)	Рівняння регресії за формулою (2.3)	Середньоквадратичне відхилення за формулою (2.4), коефіцієнт варіації за формулою (2.5) та лінійний коефіцієнт кореляції (Пірсона) за формулою (2.6)

2.3 Висновки до розділу

У даному розділі було чітко окреслено проблемну область поставленої задачі у розділі 1. Було розглянуто вже існуючі рішення (див. 2.1) відносно тематики завдання, такі як звіт програми розвитку Організації Об'єднаних Націй та дослідження токійського технологічного інституту. Було надано відповідні графіки та обчислення як результати досліджень, які були проведені в даних роботах та їх описання. Після детального огляду та вивчення даних досліджень було проведено їх порівняння за певними критеріями, що відображені в таблиці 2.2.

Далі було проведено порівняння засобів, яке було занесено до таблиці 2.3, за допомогою яких являється можливим проведення розв'язку поставленої задачі. Не важко зрозуміти, що найкращим засобом для створення програмного забезпечення, за допомогою якого можна розв'язати поставлену задачу, являється PyCharm та мова Python, оскільки ліцензоване програмне забезпечення являється безкоштовним для студентів та має багато переваг відносно інших засобів, за допомогою яких теж можна розв'язати задачу.

Після огляду можливих програмних засобів було проведено детальний огляд та вивчення методів, які надають можливість проводити аналіз даних відносно певних властивостей, які має відповідний метод. Порівняння даних методів за певними критеріями було занесено у таблицю 2.5. Серед усіх запропонованих та розглянутих методів був обраний кластерний аналіз. Оскільки даний аналіз найбільше підходить до поставленої задачі і дозволяє визначити кластери (групи) елементів зі схожими властивостями, що дозволяє розв'язати поставлену задачу. Також використання методу кластерного аналізу було запропоновано Інститутом демографії та соціальних досліджень ім. М.В. Птухи НАН України у плані наукових робіт.

3 МАТЕМАТИЧНЕ ЗАБЕЗПЕЧЕННЯ

3.1 Загальний опис алгоритму аналізу даних

Процес оцінювання та аналізу завжди починається з того, що потрібно отримати певні вхідні дані, після чого з цими даними будуть виконуватись певні маніпуляції. В даному випадку вхідними даними є файл Excel з певними даними, які потрібно проаналізувати.

Отриманий файл Excel містить такі стовпці, як назва країни, індекс людського розвитку, кількість використаної енергії та континент, в якому розташована країна. Кожна країна, яка знаходиться у файлі, має відповідний індекс людського розвитку та загальне споживання первинної енергії у розрахунку на душу населення.

Після зчитування даних, з ними будуть проводитись маніпуляції для того, щоб в кінцевому рахунку провести кластерний аналіз даних та отримати бажаний результат. Результатом будуть відповідні графіки та діаграми, які будуть демонструвати взаємозв'язок якості життя населення та споживання електроенергії.

3.2 Побудова матриці кореляцій Scatter-Matrix

Діаграма матриці розсіювання (scatter matrix) насамперед являється інструментом візуалізації даних, який дозволяє проводити порівнювання декількох наборів даних між собою. Іншими словами, матриця розсіювання – це матриця, яка компактно відображає всі числові змінні, які ми маємо в наборі даних [6] і дозволяє вивчати відношення між багатьма змінними. В даному випадку діаграма буде мати лише дві основні частини (індекс людського розвитку та споживання енергії).

На рисунку 3.1 відображається матриця розсіювання для вхідних даних, що використовуються у роботі [6].

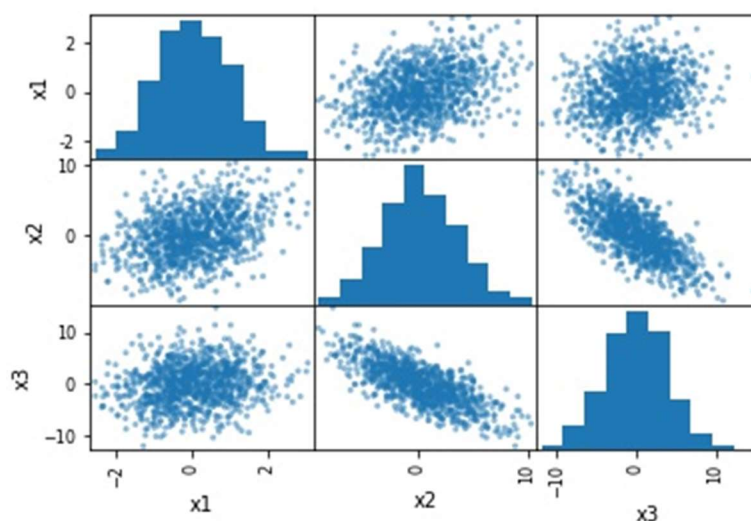


Рисунок 3.1 – Матриця розсіяння відповідно до роботи [6]

Відносно даної матриці можна зробити певне її описання:

- діагональ показує розподілення даних відповідного прикладу;
- в усіх інших комірках матриці знаходяться діаграми розсіяння, які також називають графіками кореляцій, для кожної комбінації змінних. Тобто в другій та в третій клітинці показано кореляцію між x_1 та x_2 , x_1 та x_3 відповідно. Аналогічно і для інших клітинок.

Для того, щоб виконати обчислення матриці розсіяння, використовується наступна формула [7]:

$$S = \sum_{k=1}^n (x_k - m)(x_k - m)^T, \quad (3.1)$$

де m – це середній вектор, який обчислюється за наступною формулою:

$$S = \frac{1}{n} \sum_{k=1}^n x_k, \quad (3.2)$$

3.3 Алгоритм зменшення розмірності (t-SNE)

t-SNE (t-distributed Stochastic Neighbor Embedding) – розподілене стохастичне вкладення сусідів [8].

Даний метод є одним з методів, що використовується для зменшення розмірності, який також дуже добре підходить для того, щоб виконувати візуалізацію багатомірних наборів даних. Також це не є математична техніка, а більш ймовірнісна.

В оригінальному документі [9] робота t-SNE описується таким чином, що даний алгоритм виконує мінімізацію розбіжності між двома розподілами:

- розподілом, яке виконує вимірювання попарних подібностей вхідних даних;
- розподілом, яке вимірює попарні подібності відповідних точок.

Тобто це означає те, що сам алгоритм оглядає оригінальні дані, які подаються до нього, після чого він визначає, найкращий спосіб для того, щоб подати ці дані, використовуючи меншу кількість обчислень.

Якщо дві точки розташовані близько одна відносно одної, то потрібно зробити так, щоб ці точки у відображенні були також близько одна до одної. В даному випадку, між точками даних та точками відображень існує бієкція, тобто кожна точка має одне своє представлення у відображенні.

У роботі [9] показано, що даний алгоритм починається з того, що перетворення багатовимірного евклідового відношення між точками даних в умовні ймовірності. За допомогою умовної ймовірності $p_{j|i}$ обчислюється збіжність точок, тобто вірогідність того, що точка x_i вибере точку x_j в якості свого сусіда, якщо сусіди будуть обрані пропорційно їх ймовірності при нормальному розподілі (розподілі Гауса), де центром буде точка x_i .

Для обчислення умовної ймовірності використовується наступна формула:

$$p_{j|i} = \frac{\exp(-|x_i - x_j|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-|x_i - x_k|^2 / 2\sigma_i^2)}, \quad (3.3)$$

де $|x_i - x_j|$ – евклідова відстань між двома точками, а σ_i – задана дисперсія.

Для маломірних аналогів y_i та y_j можна вирахувати аналогічну умовну ймовірність $q_{j|i}$ за наступною формулою:

$$q_{j|i} = \frac{\exp(-|y_i - y_j|^2)}{\sum_{k \neq i} \exp(-|y_i - y_k|^2)}, \quad (3.4)$$

де $|y_i - y_j|$ – відстань між точками відображення.

Після цього алгоритм намагається мінімізувати невідповідності між $p_{j|i}$ та $q_{j|i}$. Для вимірювання мінімізації суми різниць умовної ймовірності даний алгоритм мінімізує суму розходження Кульбака-Лейблера спільних точок за допомогою методу градієнтного спуску. Дана процедура обчислюється за формулами (3.5) та (3.6):

$$C = \sum_i KL(P_i || Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}, \quad (3.5)$$

де P_i – умовне розподілення ймовірностей для точок x_i , а Q_i – умовне розподілення ймовірностей по іншим точкам y_i .

Мінімізація формули 3.5 виконується за допомогою методу градієнтного спуску:

$$\frac{\partial C}{\partial y_i} = 2 \sum_j (p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})(y_i - y_j) \quad (3.6)$$

В даному алгоритмі значення $p_{i|j}$ є визначеними як симетричні умовні ймовірності (за формулою 3.7), а значення $q_{i|j}$ отримуються шляхом розподілення Стюдента з одним степенем свободи (за формулою 3.8).

$$p_{i|j} = \frac{p_{j|i} + p_{i|j}}{2n}, \quad (3.7)$$

де $p_{i|j}$ та $p_{j|i}$ – обчислюються за формулою 3.3.

$$q_{i|j} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|y_i - y_j\|^2)^{-1}}, \quad (3.8)$$

Отримуються дві матриці подібності для даних $p_{i|j}$ та $q_{i|j}$ відповідно, де перша являється постійною, а інша матриця подібності для відображення $q_{i|j}$ залежить від точок відображення. Потрібно досягти того, щоб отримані матриці подібності $p_{i|j}$ та $q_{i|j}$ були подібними між собою.

Також в даній роботі [9] наведено приклад візуалізації даних, який наведено на рисунку 3.2:

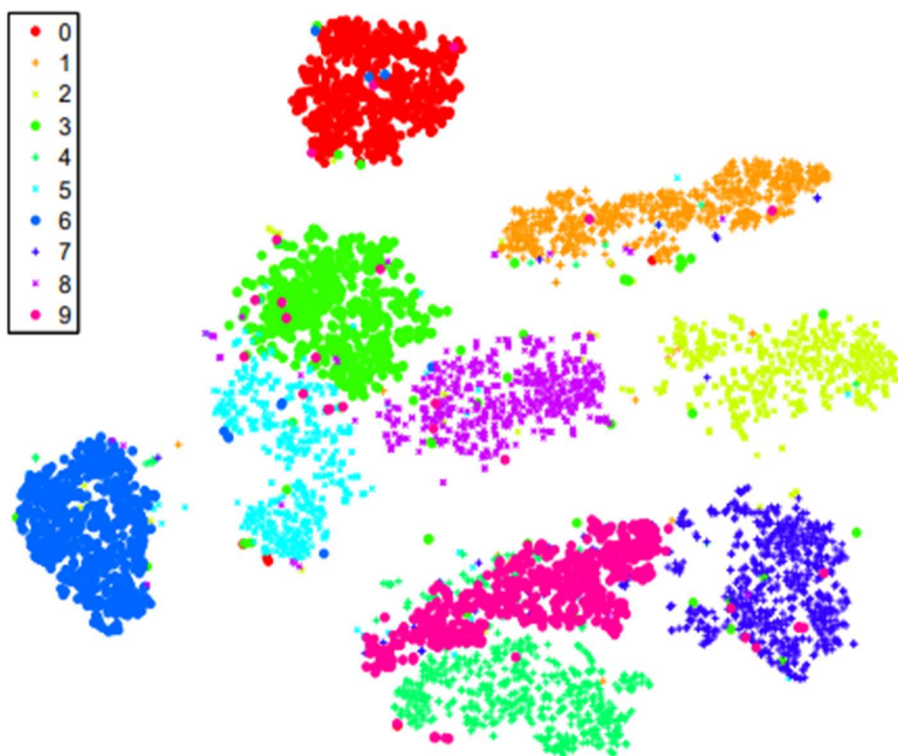


Рисунок 3.2 – Візуалізація методом t-SNE [9]

3.4 Метод кластеризації K-Means

Метод кластеризації K-Means - це найбільш відомий метод кластеризації, який відноситься до ієрархічних методів кластеризації даних. Також даний метод ще називається швидким кластерним аналізом. Також кажуть, що даний алгоритм кластеризації являється неконтрольованим алгоритмом машинного навчання.

Алгоритм K-Means проводить класифікацію даних без їх попередньої підготовки, але потрібно мати гіпотезу про те, яка найбільш ймовірна кількість кластерів. Іншими словами: даний алгоритм вимагає, щоб число кластерів було вказаним.

Даний алгоритм ділить набір з N вибірок X на K непересічних кластерів C , кожний з яких описується як середнє μ_j зразків в кластері [10].

Алгоритм K-середніх призначений для того, щоб обирати центри, яку зазвичай прийнято називати центроїдами, які мінімізують інерцію або критерій суми квадратів всередині кластеру. Дана дія проводиться за наступною формулою [10]:

$$\sum_{i=0}^n \min_{\mu_j \in C} (\|x_i - \mu_j\|^2) \quad (3.8)$$

Для цього алгоритму його основна ідея має наступний вигляд: для заданого фіксованого числа кластерів спостереження зіставляються так, щоб центроїди кожного кластеру відрізнялись між собою якнайбільше.

В роботі [11] наведено описання даного алгоритму:

а) Перше розподілення об'єктів по кластерам:

Обирається деяке число k і це число на першому кроці буде вважатись центром кластеру. Вибір початкових центроїдів може бути виконане наступним чином:

- обрати деяке k - спостережень для того, щоб максимізувати початкові відстані;
- обрати k -спостережень випадковим чином;

- обрання перших k -спостережень.

б) Ітеративний процес:

В даному пункті обчислюються центри кластерів, після чого виконується перерозподіл кластерів. Ця дія буде виконуватись до тих пір, поки не буде виконана одна з наступних умов:

- центри кластерів являються стабільними;
- число ітерацій має дорівнювати максимальному числу ітерацій.

Даний метод кластеризації даних являється простим та швидким у використанні, що набагато спрощує роботу, але також він може бути неточним та працювати доволі повільно, коли вхідні дані є дуже великими.

На рис. 3.3 зображено приклад результату виконання методу K-Means у роботі [12]:

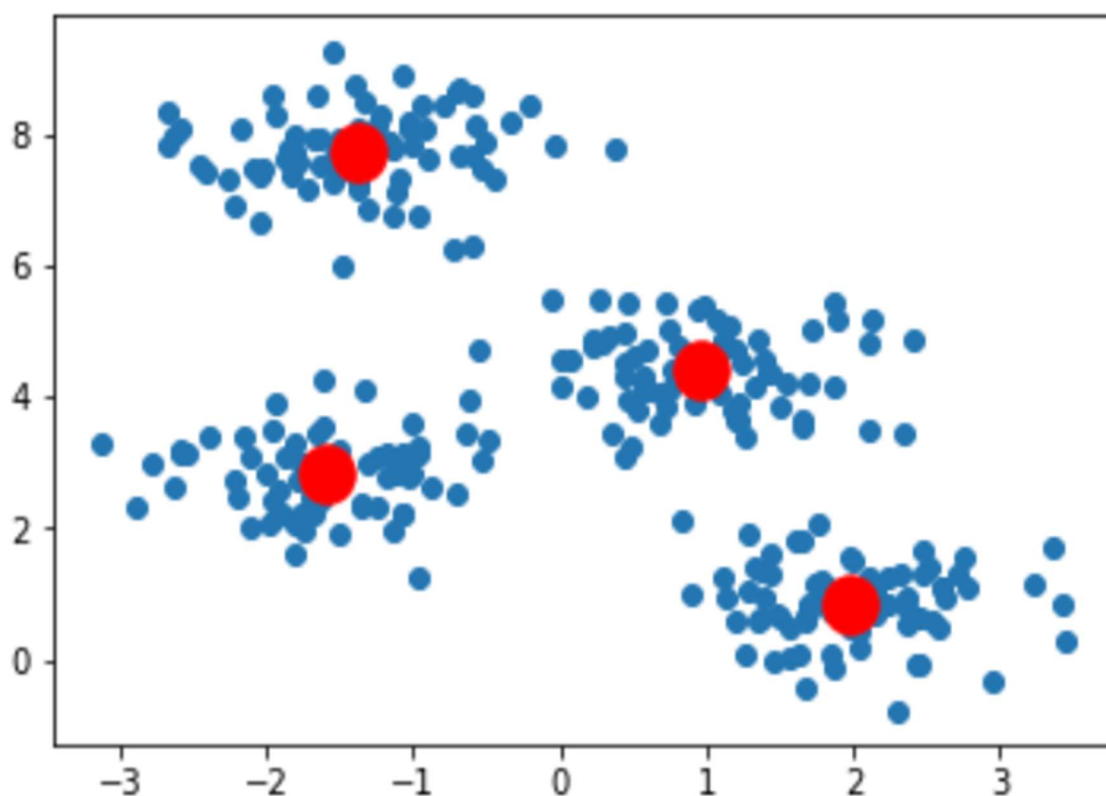


Рисунок 3.3 – Результат виконання методу K-Means[12]

Даний метод являється також дуже зручним у використанні при громіздких вхідних даних, коли оптимальна кількість кластерів більше не являється очевидною. В цьому випадку можна використовувати метод ліктя (The Elbow Method), який призначений для оцінювання оптимальної кількості кластерів k для поставленої задачі та являється для цього графічним інструментом. Даний графічний інструмент виконує побудову графіку залежності між кількістю кластерів та сумою квадратів в кластері (WCSS). Після цього обирається кількість кластерів, після якої зміни в WCSS починають вирівнюватись [12], тобто починають змінюватись не дуже значно. Це і називається «методом ліктя».

Величина WCSS визначається як сума квадратів відстаней між членом кластеру та його центроїдом [12] і визначається за наступною формулою:

$$WSS = \sum_{i=1}^m (x_i - c_i)^2 \quad (3.9)$$

На рисунку 3.4 зображено приклад результату роботи методу ліктя, який використовується у роботі [12], який показує, що оптимальною кількістю кластерів для використаних у даному прикладі вхідних даних – буде 4 кластери:

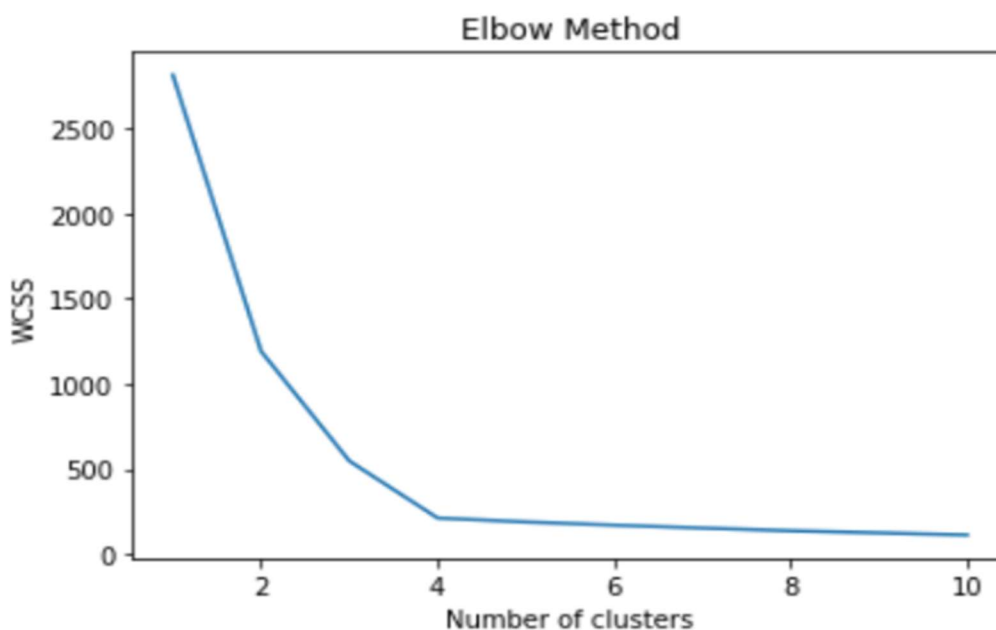


Рисунок 3.4 – Результат виконання методу ліктя [12]

3.5 Агломеративний метод кластеризації

Агломераційна кластеризація є одним з найвідоміших методів ієрархічної кластеризації, який використовує підхід, що називається підходом знизу вгору.

Як сказано у роботі [13]: основна ідея даного підходу полягає в тому, що кожний елемент спочатку являється окремим кластером і потім послідовно виконується об'єднання всіх кластерів, поки всі кластери не будуть об'єднані в один великий кластер, який містить в собі всі елементи. Тому даний тип ієрархічної кластеризації називається знизу вгору, оскільки інший тип – зверху вниз має на меті метод розділення кластеру. Ця дія виконується шляхом рекурсивного розділення кластерів, поки не буде досягнуто окремих елементів.

Об'єднання кластерів виконується відповідно до певних критеріїв. Критерії, по яким виконується визначення показнику, який використовується для об'єднання, наведено у роботі [10] та у роботі [14] наведена візуалізація цих зв'язків:

– Ward linkage. Використовується для мінімізації сум квадратів відмінностей у всіх кластерах. Цей підхід мінімізує значення дисперсії та є аналогічним до цільової функції методу кластеризації k-Means (див. 3.4), але розв'язується за допомогою агломераційного ієрархічного підходу.

На рисунку 3.5 зображено даний зв'язок:

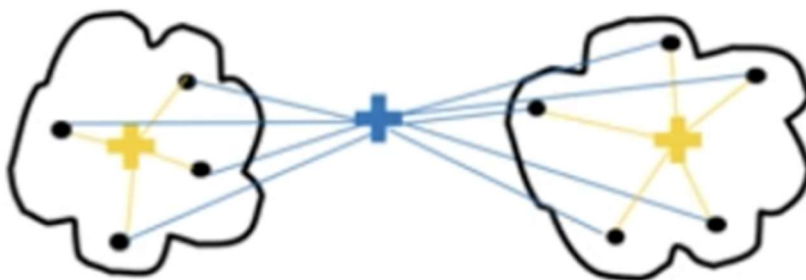


Рисунок 3.5 – візуалізація Ward Linkage [14]

Дистанція в цьому випадку дорівнює сумі квадратів різниць у всіх даних кластерах.

– Maximum or complete linkage (Максимальний або повний зв'язок). Виконується мінімізація відстані між парами кластерів і відстань при цьому є максимальною. Відстань між двома кластерами в цьому випадку обчислюється за наступною формулою:

$$L(r, s) = \max(D(x_{ri}, x_{sj})), \quad (3.10)$$

де r, s – позначення для кластерів, а x_{ri}, x_{sj} – найвіддаленіші точки в даних кластерах.

На рисунку 3.6 зображено візуальний приклад для повного зв'язку:

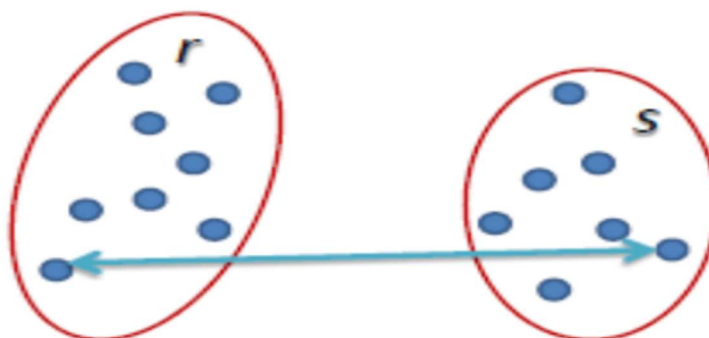


Рисунок 3.6 – Візуалізація Maximum or complete linkage [14]

– Average linkage (Середній зв'язок). Виконується мінімізація середньої відстані між усіма можливими парами. В даному випадку відстань між наведеними двома кластерами рахується за наступною формулою:

$$L(r, s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D(x_{ri}, x_{sj}), \quad (3.11)$$

де r, s – позначення для кластерів, x_{ri}, x_{sj} – найвіддаленіші точки в даних кластерах, n_r, n_s – кількість елементів у відповідному кластері.

На рисунку 3.7 зображено візуальний приклад для середнього зв'язку:

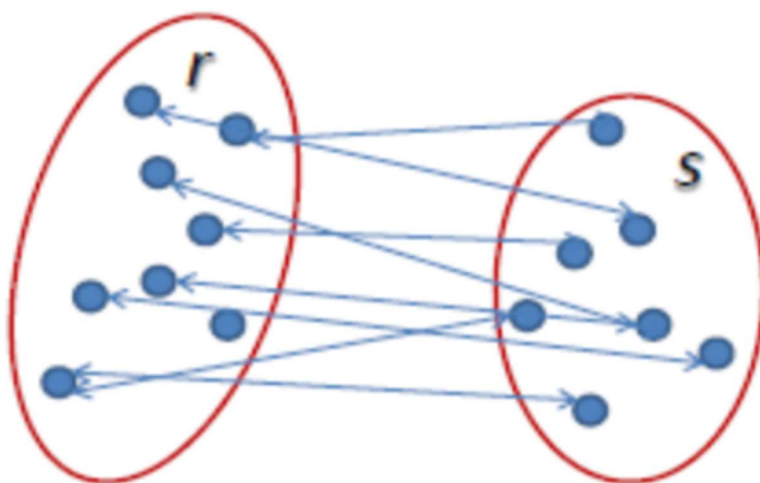


Рисунок 3.7 - Візуалізація Average linkage [14]

– Single linkage (Одиночний зв'язок). Виконується мінімізація відстані між тими парами скупчень, які є найближчими. Даний зв'язок є повною протилежністю повного зв'язку. В цьому випадку обираються точки з кожного кластеру, відстань між якими є найменшою, і це буде відстанню між двома кластерами. Дана відстань рахується за наступною формулою:

$$L(r, s) = \min(D(x_{ri}, x_{sj})), \quad (3.12)$$

де r, s – позначення для кластерів, а x_{ri}, x_{sj} – найвіддаленіші точки в даних кластерах.

На рисунку 3.8 зображено візуальний приклад для одиничного зв'язку:

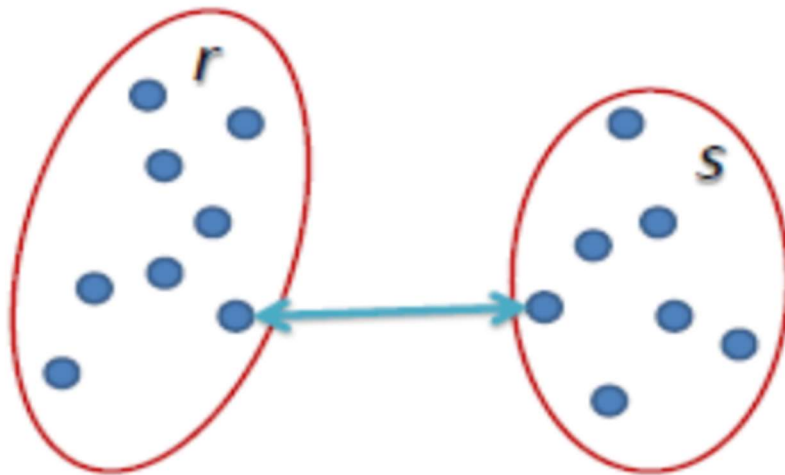


Рисунок 3.8 – Візуалізація Single linkage [14]

На рис. 3.9 зображено приклад результату виконання агломеративного алгоритму кластеризації для вхідних даних, що наведено у роботі [14]:

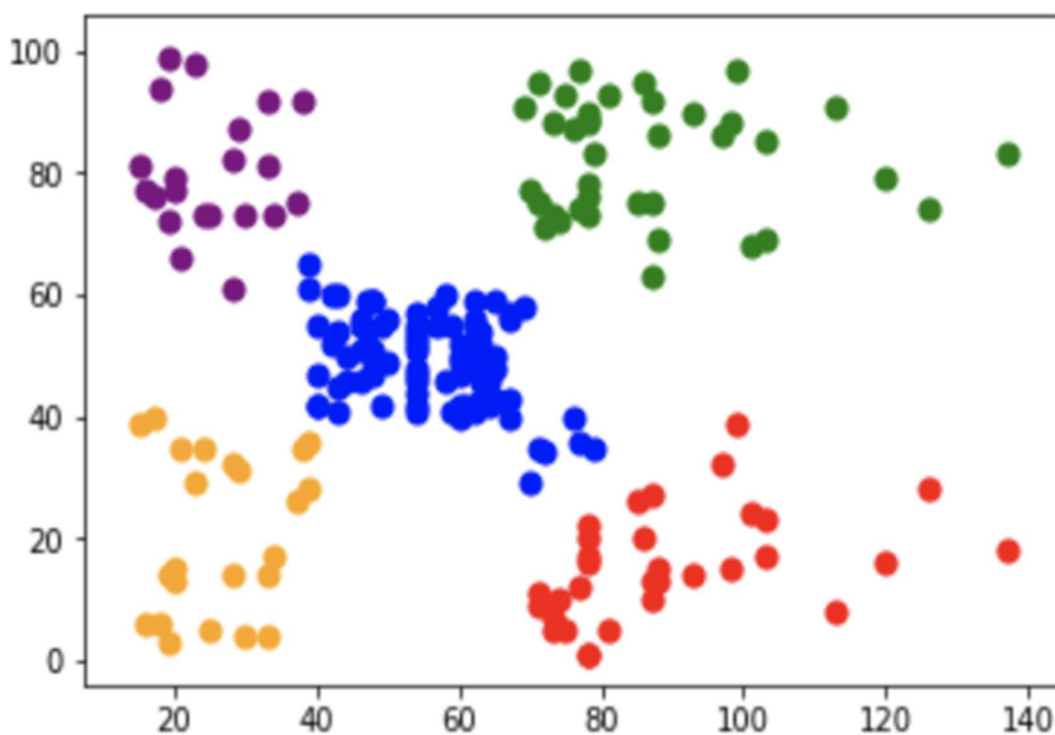


Рисунок 3.9 – Результат виконання агломеративного методу [14]

Оскільки агломеративний метод кластеризації являється ієрархічним методом кластеризації даних, то для даного типу кластеризації результатом може бути Дендрограма.

Як наведено у офіційній документації [15]: Дендрограма – це діаграма, яка демонструє ієрархічне відношення між елементами. Дендрограма містить в собі множину U-подібних ліній, які поєднують між собою елементи ієрархічного дерева. Висота кожної такої лінії вказує на відстань між точками даних, що є зв'язаними.

На рис. 3.10 зображено приклад дендрограми, що наведено у роботі [15]:

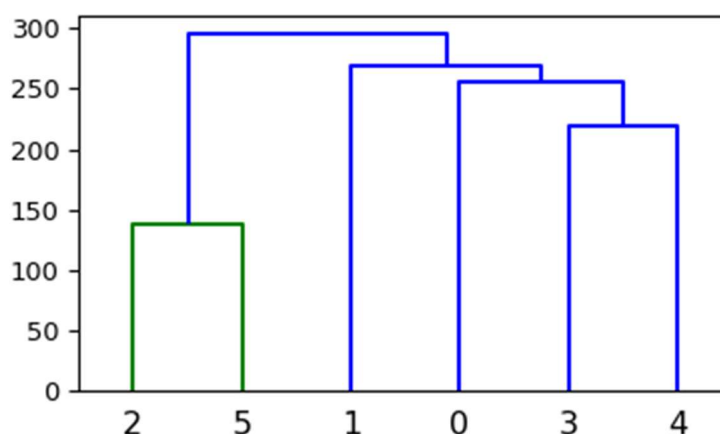


Рисунок 3.10 – Приклад дендрограми [15]

На даному рисунку видно, що між кластерами (2, 5) та (1, 0, 3, 4) існує велика відстань між ними, і так далі.

3.6 Спектральний метод кластеризації

Спектральний метод кластеризації являється популярним методом кластеризації, оскільки має просту реалізацію та може бути вирішено за допомогою стандартних методів лінійної алгебри. Даний метод може бути ефективнішим

методом, ніж стандартні методи кластеризації, такі як k-Means. Також даний метод являється зручним для пошуку кластерів, коли вхідні дані, з якими потрібно провести кластерний аналіз – не відповідають вимогам інших відомих алгоритмів.

Типова реалізація даного алгоритму наведена у роботі [16]. Вона складається з трьох основних етапів:

а) Побудова графу подібності. На даному етапі будується граф подібності у формі матриці A , що є суміжною.

б) Проектування даних на простір з більш низькою розмірністю. Цей крок робиться саме для того, щоб була можливість врахувати те, що члени кластеру можуть знаходитись далеко один від одного. За допомогою даної процедури сам простір зменшується і з ним зменшується відстань між точками, що дає можливість для їх групування за допомогою стандартного методу кластеризації. Це робиться шляхом обчислення матриці Кірхгофа. Щоб це виконати, потрібно спочатку підрахувати ступінь вузла графу. Це робиться за допомогою наступної формули:

$$d_i = \sum_{j=1}^n w_{ij}, \quad (3.13)$$

де w_{ij} – це ребро між вузлами i та j в матриці суміжності.

Матриця ступенів визначається наступним чином:

$$D_{ij} = \begin{cases} d_i, & i = j \\ 0, & i \neq j \end{cases} \quad (3.14)$$

Виходячи з даних формул, матриця Кірхгофа буде визначатись так:

$$L = D - A \quad (3.15)$$

Після цього виконується нормалізація даної матриці для математичної ефективності. Для того, щоб зробити розміри меншими, спочатку виконують

обчислення власних значень та відповідних власних векторів. Власні вектори додаються до матриці таким чином, щоб власні вектори були стовпцями матриці.

в) Проведення кластеризації даних. В основному кластеризація виконується традиційним методом кластеризації даних, але як правило – це метод k-Means. Кластеризація проходить таким чином, що кожному вузлу присвоюється рядок з нормалізованої матриці Кірхгофа, після чого дані починають групуватись з використанням будь-якого традиційного методу кластеризації.

Результат даного алгоритму кластеризації має схожий вигляд, як у агломеративного методу кластеризації даних (рис. 3.9) та у методі кластеризації даних k-Means (рис. 3.3).

3.7 Метрики

3.7.1 ARI (Adjusted rand index)

Детальне пояснення даної міри наведено у роботі [10]. Даний індекс обчислює міру збіжності між двома кластеризаціями. Необроблений індекс обчислюється за наступною формулою:

$$RI = \frac{a+b}{C_2^{n_{samples}}}, \quad (3.16)$$

де a – число пар, які знаходяться в одному кластері, b – кількість пар елементів, що знаходяться в різних кластерах, $C_2^{n_{samples}}$ – загальна кількість можливих елементів в наборі даних.

Для того, щоб даний індекс давав значення, яке буде наближено до нуля, то необхідно його нормалізувати. Нормалізація виконується за допомогою наступної формули:

$$ARI = \frac{RI - E[RI]}{\max(RI) - E[RI]} \quad (3.17)$$

3.7.2 AMI (Adjusted mutual info)

Дана міра дуже схожа з мірою ARI. Вона також являється симетричною і не залежить від перестановок.

Детальне пояснення даної міри наведено у роботі [10]. Нехай є два елементи U та V , які взяті з різних класів U_i та V_j відповідно. Для даної міри використовуються наступні формули:

$$MI(U, V) = \sum_{i=1}^{|U|} \sum_{j=1}^{|V|} P(i, j) \log\left(\frac{P(i, j)}{P(i)P(j)}\right), \quad (3.18)$$

де $P(i, j) = \frac{|U_i \cap V_j|}{N}$ – ймовірність того, що елемент, який був обраний випадково, потрапляє в відповідні класи.

Нормалізація попередньої формули виконується по аналогії до формули 3.17:

$$AMI = \frac{MI - E[MI]}{\max(H(U), H(V)) - E[MI]} \quad (3.19)$$

3.7.3 Homogeneity, Completeness, V-measure

Дані міри детально описуються в офіційному джерелі [10].

Homogeneity – дана метрика являється метрикою однорідності. Результат буде задовольняти однорідність, якщо кластери містять лише точки даних, які входять лише до одного кластеру. За допомогою наступної формули вимірюється, наскільки кожний кластер складається з об'єктів одного класу:

$$h = 1 - \frac{H(C | K)}{H(C)} \quad (3.20)$$

Completeness – всі члени одного кластеру відносяться до одного кластеру. За допомогою наступної формули вимірюється, наскільки об'єкти одного класу відносяться до одного кластеру:

$$c = 1 - \frac{H(K | C)}{H(K)} \quad (3.21)$$

В формулі 3.20 та в формулі 3.21 K – результат кластеризації, C – розбиття вибірки на класи.

Дані міри не являються нормалізованими, тому вони залежать від кількості кластерів.

V-measure – називається їх гармонічне середнє, яке використовується для врахування обох величин h та c . Дана величина являється симетричною і показує, наскільки обидві кластеризації схожі між собою. За наступною формулою виконується підрахунок даної величини:

$$v = 2 \frac{hc}{h+c} \quad (3.22)$$

3.7.4 Silhouette

Описання даної метрики наведено у роботі [10]. Даний коефіцієнт дозволяє проводити оцінювання якості кластеризації, використовуючи для цього лише саму вибірку та результат кластеризації. Силует обчислюється за наступною формулою:

$$s = \frac{b-a}{\max(a,b)}, \quad (3.23)$$

де a – це середня відстань від даного об'єкту до об'єктів того ж кластеру, а b – середня відстань від даного об'єкту до об'єктів найближчого кластеру.

3.8 Висновки до розділу

У цьому розділі було розглянуто та проаналізовано математичні підходи і алгоритми для розв'язання поставленої задачі. Також було визначено загальний опис алгоритму для розв'язання задачі.

Обране математичне забезпечення будується на різних методах, які призначенні для кластеризації даних. До цих методів належить метод кластеризації k-Means, агломеративний кластерний аналіз та спектральний кластерний аналіз.

Також було проведено детальний розгляд та аналіз мір, за допомогою яких можна переглянути ефективність певних методів кластеризації даних для вхідних даних, що використовуються в цій роботі. Цими мірами є ARI, AMI, Homogeneity, Completeness, V-measure та Silhouette.

Для візуалізації вхідних даних, які використовуються в даному дипломному проекті, було обрано метод побудови матриці кореляцій Scatter-Matrix та алгоритм зменшення розмірності (t-SNE).

4 ОПИС ПРОГРАМНИХ ЗАСОБІВ

4.1 Архітектура розробленої системи

На рисунку 4.1 зображено архітектуру розроблюваної системи, яка була створена за допомогою онлайн сервісу для створення діаграм Draw IO.

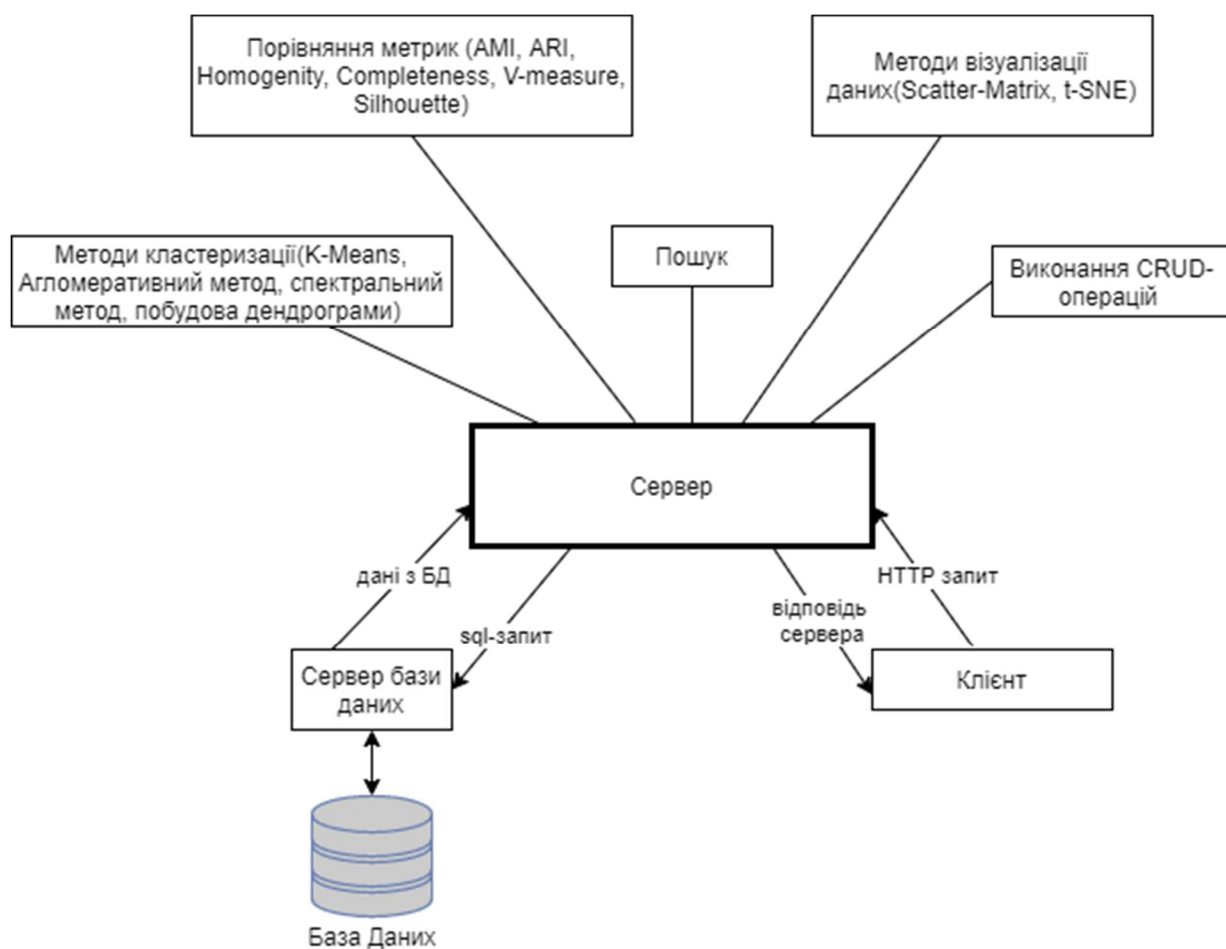


Рисунок 4.1 – Структура системи

В даній архітектурі присутні такі складові, як «Клієнт», «Сервер» та «Сервер бази даних». Робота даної системи є наступною:

Коли «Клієнт» хоче отримати результат виконання певної задачі, то відправляє HTTP-запит на «Сервер». Після того, як «Сервер» отримав даний запит, він починає оброблювати його. Оскільки для того, щоб виконати певну задачу

– потрібні вхідні дані. Для того, щоб отримати вхідні дані, сервер надсилає SQL-запит до серверу баз даних, після чого отримує відповідні дані з цієї бази даних. Після того, як «Сервер» отримав дані, починається виконання певної задачі. Коли задача виконана, то «Клієнт» отримує відповідь серверу у вигляді відповідного до задачі результату.

В даній системі за базу даних було обрано PostgreSQL, тому що вона добре підходить для зберігання та подальшого використання даних за допомогою відповідних засобів мови Python для даного дипломного проекту.

Для створення серверу було використано фреймворк Flask. Клієнтську частину було створено за допомогою мови HTML та CSS фреймворку Bootstrap 4. Екранні форми було створено за допомогою мови Python.

4.2 Аналіз вхідних даних

4.2.1 Загальний опис вхідних даних

В даному дипломному проектуванні вхідними даними являються дані з таблиці, яка містить інформацію відповідно до таких полів:

- назва країни;
- індекс людського розвитку (ІЛР);
- загальне споживання енергії у розрахунку на душу населення;
- континент.

Аналіз проводиться для 76 країн, тобто дана таблиця має 4 стовпці та 76 рядків з даними.

Дану таблицю було створено в базі даних PostgreSQL за допомогою засобів мови Python, а саме з використанням бібліотеки SQLAlchemy, яка створена саме для

того, щоб працювати з реляційними базами даних. Далі, з допомогою можливостей даної бази даних, дані було імпортовано в уже створену таблицю.

4.2.2 Опис ER-діаграми

База даних для даного дипломного проекту складається лише з однієї сутності, яка має назву «rate». На рисунку 4.2 зображено вигляд сутності в базі даних, яке створено за допомогою інструменту моделювання PowerDesigner.

rate	
country_name	<pi> Variable characters (20)
index_value	Float (10)
usage_value	Float (10)
continent_name	Variable characters (20)
Identifier_1	<pi>

Рисунок 4.2 – Вигляд сутності

У таблиці 4.1 наведено опис даної сутності.

Таблиця 4.1 – Опис сутності

Атрибути сутності	Тип атрибуту	Опис атрибуту
country_name	VARCHAR(20)	Містить в собі назву країни та є ключем. Має бути унікальним
index_value	FLOAT(10)	Містить в собі числові дані для індексу людського розвитку (ІЛР)

Продовження таблиці 4.1

Атрибути сутності	Тип атрибуту	Опис атрибуту
usage_value	FLOAT(10)	Містить числові дані про загальне споживання первинної енергії на душу населення
continent_name	VARCHAR(20)	Містить в собі назву континенту

4.3 Елементарні події

В створеному програмному проекті користувач може виконувати певні дії, які мають відповідні сценарії. До таких дій відноситься виконання CRUD-операцій, виконання пошуку даних відповідно до обраного критерію.

У таблицях 4.2- наведено різні сценарії відповідно до того, що хоче виконати користувач.

Таблиця 4.2 – Сценарій для кейсу «Перегляд даних»

Актори	Система, Користувач
Мета	Переглянути існуючі дані
Передумови	Користувач хоче переглянути дані
Успішний сценарій: Для того, щоб переглянути дані, користувачу потрібно перейти на відповідну сторінку. Після переходу користувач може побачити всі існуючі дані.	
Результат	Перегляд існуючих даних

Таблиця 4.3 – Сценарій для кейсу «Створення даних»

Актори	Система, Користувач
Мета	Створити нові дані
Передумови	Користувач хоче створити нові дані
<p>Успішний сценарій:</p> <p>Для того, щоб виконати редагування вже існуючих даних, користувачу потрібно спочатку перейти на відповідну сторінку, де виводяться всі дані сутності. Після переходу та натискання відповідної клавіші, користувачу потрібно ввести нові дані, після дані будуть оновлені.</p>	
Результат	Дані створено

Таблиця 4.4 – Сценарій для кейсу «Редагування даних»

Актори	Система, Користувач
Мета	Відредагувати існуючі дані
Передумови	Користувач хоче відредагувати дані
<p>Успішний сценарій:</p> <p>Для того, щоб виконати редагування вже існуючих даних, користувачу потрібно спочатку перейти на відповідну сторінку, де виводяться всі дані сутності. Після переходу та вибору відповідного екземпляру сутності, користувач повинен відредагувати існуючі дані або замінити їх на нові, після чого дані будуть оновлені.</p>	
Результат	Дані відредаговано

Таблиця 4.5 – Сценарій для кейсу «Видалення даних»

Актори	Система, Користувач
Мета	Видалити існуючі дані
Передумови	Користувач хоче видалити дані

Продовження таблиці 4.5

<p>Успішний сценарій:</p> <p>Для того, щоб виконати редагування вже існуючих даних, користувачу потрібно спочатку перейти на відповідну сторінку, де виводяться всі дані сутності. Після переходу, користувачу потрібно обрати екземпляр сутності, який користувач хоче видалити та натиснути відповідну клавішу, після чого дані будуть оновлені.</p>	
Результат	Дані видалено

Таблиця 4.6 – Сценарій для кейсу «Пошук»

Актори	Система, Користувач
Мета	Виконати пошук відповідно до обраного категорії
Передумови	Користувач хоче виконати пошук даних
<p>Успішний сценарій:</p> <p>Для того, щоб виконати пошук даних, користувачу потрібно спочатку перейти на відповідну сторінку. Після переходу, користувачу потрібно обрати категорію, відносно якої буде проводитись пошук і ввести дані для пошуку, після чого користувачу виводиться результат пошуку.</p>	
Результат	Бажані дані

4.4 Опис програмного забезпечення

Для створення програмного забезпечення було використано ряд бібліотек, які потрібні для того, щоб розв'язати поставлену задачу. У таблиці 4.7 наведено використані бібліотеки та їх описання.

Таблиця 4.7 – Опис бібліотек

Назви бібліотек	Імпортовані класи	Описання бібліотеки
numpy	-	Бібліотека, яка використовується для підтримки масивів, різних математичних функцій тощо.
pandas	-	Бібліотека, яка призначена для того, щоб виконувати аналіз даних та оброблювати дані відповідно до певних методів.
flask	Flask, render_template, request, redirect	Бібліотека, за допомогою якої стає можливим використання даного фреймворку, який використовується для того, щоб створювати веб-застосунки. За допомогою даної бібліотеки стає
flask_sqlalchemy	SQLAlchemy	Спеціальна бібліотека, що використовується для роботи з базами даних.
matplotlib	Pyplot	Спеціальна бібліотека, яка використовується для побудови графіків та діаграм.
scipy	linkage, dendrogram, fcluster, pdist	Це спеціальна бібліотека, яку використовують для різних наукових та інженерних підрахунків.
sklearn	metrics, preprocessing, KMeans, SpectralClustering, AffinityPropagation, AgglomerativeClustering	Бібліотека, яка використовується для машинного навчання.

Продовження таблиці 4.7

Назви бібліотек	Імпортовані класи	Описання бібліотеки
flask_wtf	FlaskForm	Бібліотека, яка використовується для полегшення роботи з формами.
wtforms	StringField, SubmitField, HiddenField, FloatField, SelectField, validators	Дана бібліотека призначена для перевірки та візуалізації створених форм.

З використанням даних бібліотек проводилась розробка та було створено програмне забезпечення.

За допомогою бібліотеки «Jinja», яка являється стандартом при створенні веб-застосунків з використанням «Flask», мови HTML та CSS, було створено сторінки з використанням стилів, на яких користувач має можливість бачити вже існуючі дані, виконувати дії з даними та бачити результат виконання поставленої задачі.

За допомогою мови Python та з використанням бібліотек flask_wtf та wtforms було розроблено відповідні форми, за допомогою яких користувач має можливість виконувати додавання нових даних, редагування та видалення існуючих даних. Додавання нового елементу повністю створює новий кортеж даних в базі даних. Редагування та видалення користувач може використовувати лише для обраного кортежу даних. Також користувач має можливість виконувати пошук по вже існуючим даним за певним критерієм. В даному випадку користувач може обрати або пошук по назві країни, або пошук по назві континенту. Для кожної форми є відповідна сторінка. Також для кожної форми було створено відповідну перевірку на правильність введених даних.

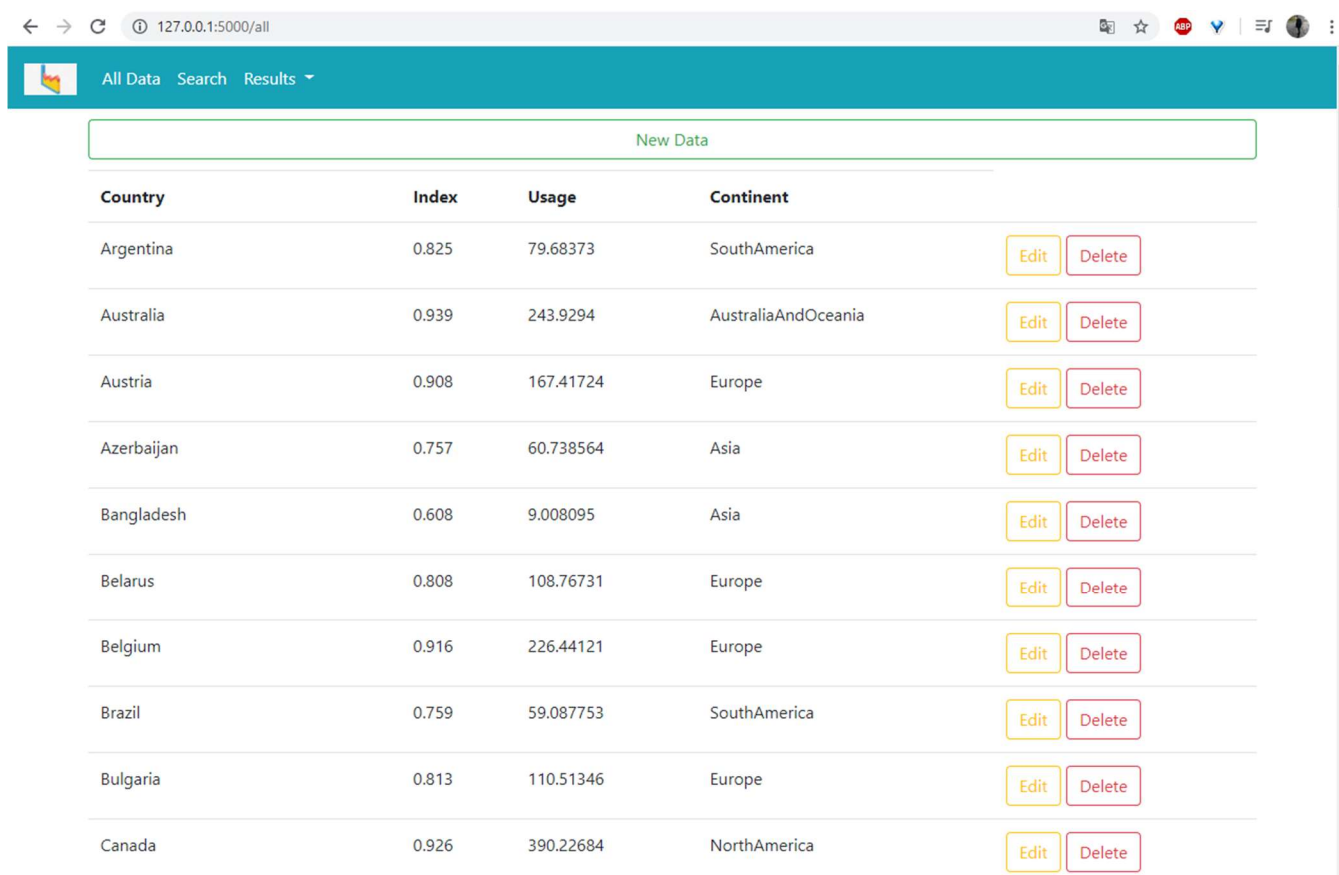
За допомогою мови Python та бібліотек, що наведено у таблиці 4.7, було створено програмне забезпечення, яке виконує аналіз даних за допомогою кластерного аналізу та виводить результати обчислень на сторінки, які відповідають методу певному методу кластеризації. Результати обчислень виводяться на відповідні до кожного методу сторінки.

В даному розробленому програмному забезпеченні результатами являються графіки та діаграми. За допомогою бібліотеки «matplotlib» було побудовано відповідні графічні об'єкти, після чого вони були збережені в локальну пам'ять комп'ютера, після чого виконується їх завантаження до веб-застосунку на відповідних сторінках.

4.5 Результати роботи програми та їх опис

4.5.1 Вивід існуючих даних та дії з ними

На рисунку 4.3 зображено виведення всіх існуючих даних на створену веб-сторінку:



The screenshot shows a web browser window with the address bar displaying '127.0.0.1:5000/all'. The page has a teal header with a search bar and a 'New Data' button. Below the header is a table with the following data:

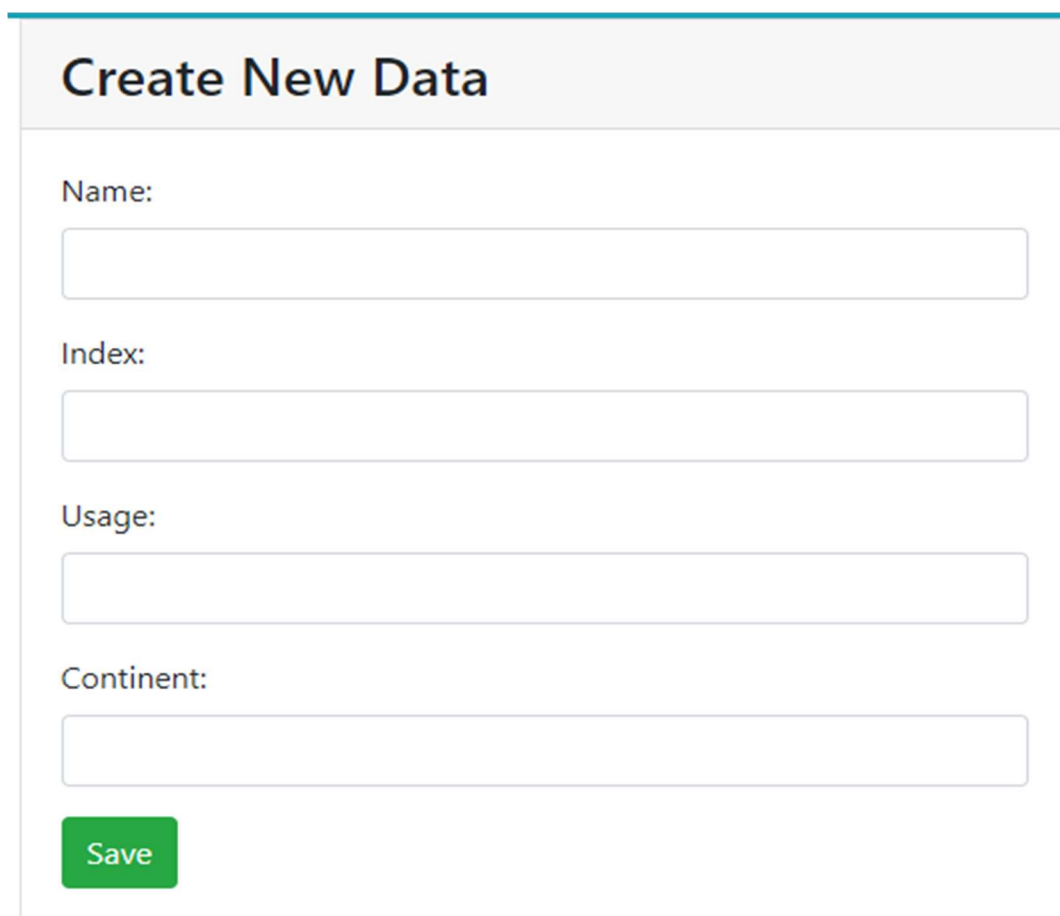
Country	Index	Usage	Continent		
Argentina	0.825	79.68373	SouthAmerica	Edit	Delete
Australia	0.939	243.9294	AustraliaAndOceania	Edit	Delete
Austria	0.908	167.41724	Europe	Edit	Delete
Azerbaijan	0.757	60.738564	Asia	Edit	Delete
Bangladesh	0.608	9.008095	Asia	Edit	Delete
Belarus	0.808	108.76731	Europe	Edit	Delete
Belgium	0.916	226.44121	Europe	Edit	Delete
Brazil	0.759	59.087753	SouthAmerica	Edit	Delete
Bulgaria	0.813	110.51346	Europe	Edit	Delete
Canada	0.926	390.22684	NorthAmerica	Edit	Delete

Рисунок 4.3 – Виведення всіх існуючих даних

Також на попередньому рисунку 4.3 видно загальний вигляд розробленого стилю сторінки, відповідні посилання на інші сторінки та клавiші для виконання операцій створення, редагування та видалення даних:

- All Data – сторінка виведення всіх існуючих даних;
- Search – сторінка для виконання пошуку;
- Result – випаданий список з посиланнями на результати виконання відповідного методу;
- New Data – клавiша, яка відповідає за виклик форми для додавання нових даних;
- Edit – клавiша для виклику форми редагування існуючих даних;
- Delete – клавiша для видалення даних.

На рисунку 4.4 зображена вигляд створеної форми для додавання нових даних:



The image shows a web form titled "Create New Data". It contains four text input fields, each preceded by a label: "Name:", "Index:", "Usage:", and "Continent:". Below these fields is a green button with the text "Save". The form is enclosed in a light gray border.

Рисунок 4.4 – Створення даних

Як видно на рисунку 4.4, користувач отримує чотири поля для введення даних, які будуть занесені до бази даних. Цими полями являються:

- а) назва країни;
- б) ІЛР;
- в) Кількість використаної енергії у розрахунку на душу населення;
- г) назва континенту.

Для назви країни та континенту застосовується перевірка на введення коректної інформації, а саме користувачу потрібно ввести назву у відповідне поле і ця назва має бути більше трьох символів та менше 20 символів. Якщо дані вимоги будуть порушені, то користувач отримає повідомлення про помилку.

Для ІЛР та кількості використаної енергії також застосовується перевірка на коректність введених даних. В даному випадку вимагається, щоб ці дані були більшими за нуль та не пустими.

На рисунку 4.5 показано повідомлення про помилки при введенні користувачем у всі поля значення, які не відповідають вимогам.

The screenshot shows a web form titled "Create New Data". It contains four input fields, each with a validation error message displayed in a red box below it:

- Name of country:** The input field contains "s". The error message is "Name should be from 3 to 20 symbols".
- Index:** The input field contains "-1".
- Usage:** The input field contains "-1". The error message is "should be > 0".
- Continent:** The input field contains "s". The error message is "Name should be from 3 to 20 symbols".

At the bottom of the form is a green "Save" button.

Рисунок 4.5 – Показ повідомлень про помилки

Якщо всі дані введено відповідно до встановлених норм та являються унікальними, то новий кортеж даних з'явиться в створеній базі даних. Якщо такі дані вже існують, то користувачу буде виведено відповідне повідомлення.

Редагування у даному випадку відбувається відповідно до назви країни, яку потрібно відредагувати. Для даної форми застосовані ті ж самі перевірки на правильність введення, що й для форми, яка відповідає за створення нових даних.

На рисунку 4.6 зображено форму, яка використовується для редагування даних, яке виконується для країни з назвою «Argentina»:



Edit For Country: Argentina

Index:
0.825

Usage:
79.68373

Save

Рисунок 4.6 – Форма для редагування даних відповідної країни

На рисунку видно, що при використанні форми для редагування у поля заносяться дані, які вже існують для даної країни. Змінювати в даному випадку можливо лише ІЛР та значення використаної енергії.

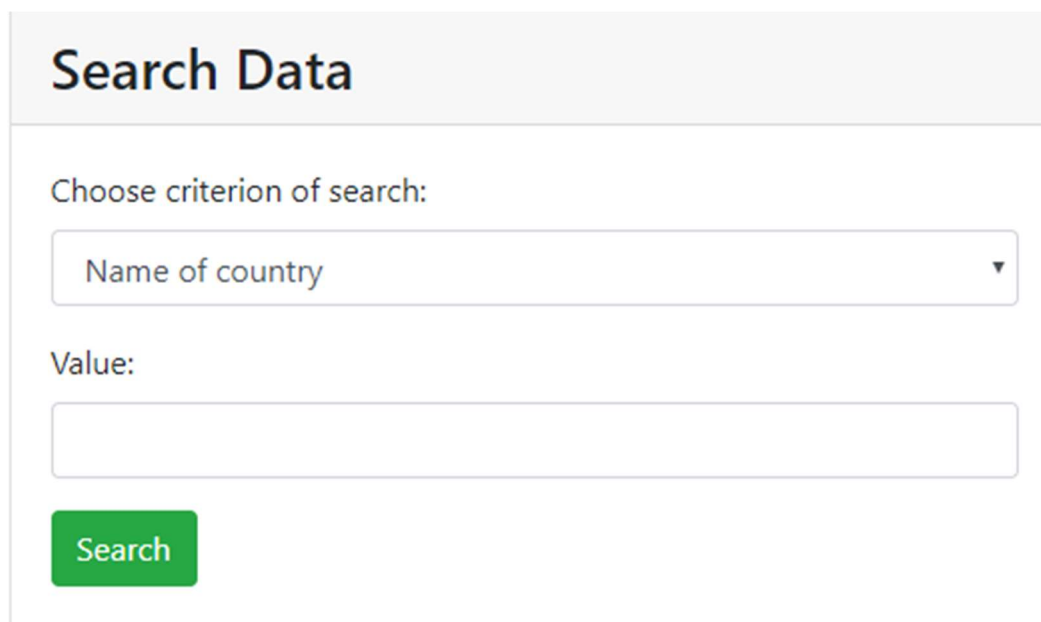
Після того, як користувач введе коректні дані, які задовольняють поставленим вимогам, відповідні дані будуть відредаговані, а зміни додані в базу даних. Після цього на сторінці виведення всіх існуючих даних дані для відредагованої країни будуть змінені.

Видалення елементів в розробленому програмному виконується відповідно до назви країни, яку користувач хоче видалити, тобто відповідно до ключа бази даних. Після натискання на клавішу видалення, таблиця буде оновлена, а обрані дані – видалені.

4.5.2 Пошук даних

Пошук даних в даному програмному забезпеченні виконується по всім існуючим даним відносно двох критеріїв, а саме назви країни або назви континенту. Користувач може сам обрати, по якому критерію виконувати пошук.

На рисунку 4.7 зображено форму, яка призначена для виконання пошуку відповідно до обраного критерію:



Search Data

Choose criterion of search:

Name of country ▼

Value:

Search

Рисунок 4.7 – Форма для виконання пошуку

Перше поле являється випадаючим списком, з якого користувач може обрати критерій пошуку. В другому полі користувач повинен написати назву країни або континенту, дані яких потрібно переглянути.

Після вибору критерію та введення даних, користувачу виведеться результат пошуку.

На рисунку 4.8 зображено виведення результату пошуку по назві країни для країни «Argentina»:

Country	Index	Usage	Continent
Argentina	0.825	79.68373	SouthAmerica

Рисунок 4.8 – Результат пошуку відносно країни

На рисунку 4.9 зображено виведення результату пошуку по назві країни для континенту «SouthAmerica»:

Country	Index	Usage	Continent
Brazil	0.759	59.087753	SouthAmerica
Chile	0.843	92.255424	SouthAmerica
Colombia	0.749	39.738113	SouthAmerica
Mexico	0.774	59.848297	SouthAmerica
Peru	0.75	34.732155	SouthAmerica
Venezuela	0.761	83.57831	SouthAmerica
Argentina	0.825	79.68373	SouthAmerica

Рисунок 4.9 – Результат пошуку відносно континенту

4.5.3 Вивід результатів обчислень

Для виведення результатів обчислень використовується 7 різних сторін, які містять графіка та діаграми відповідно до методу, за допомогою якого виконуються

На попередньому рисунку видно, що по головній діагоналі матриці зображено розподілення між числовими елементами вхідних даних. З матриці розсіювання (рис. 4.11) видно, що кореляції в даному випадку немає, тобто елементи не схожі між собою, що показано в інших клітинках матриці розсіювання. Також в цих комірках показано звичайна візуалізація вхідних даних для даного програмного забезпечення.

Другим елементом випадаючого списку з посиланнями на сторінки результатів являється Dendrogram. Дана сторінка відповідає за виведення дендрограми.

На рисунку 4.12 зображено дендрограму відносно вхідних даних для поставленої задачі:

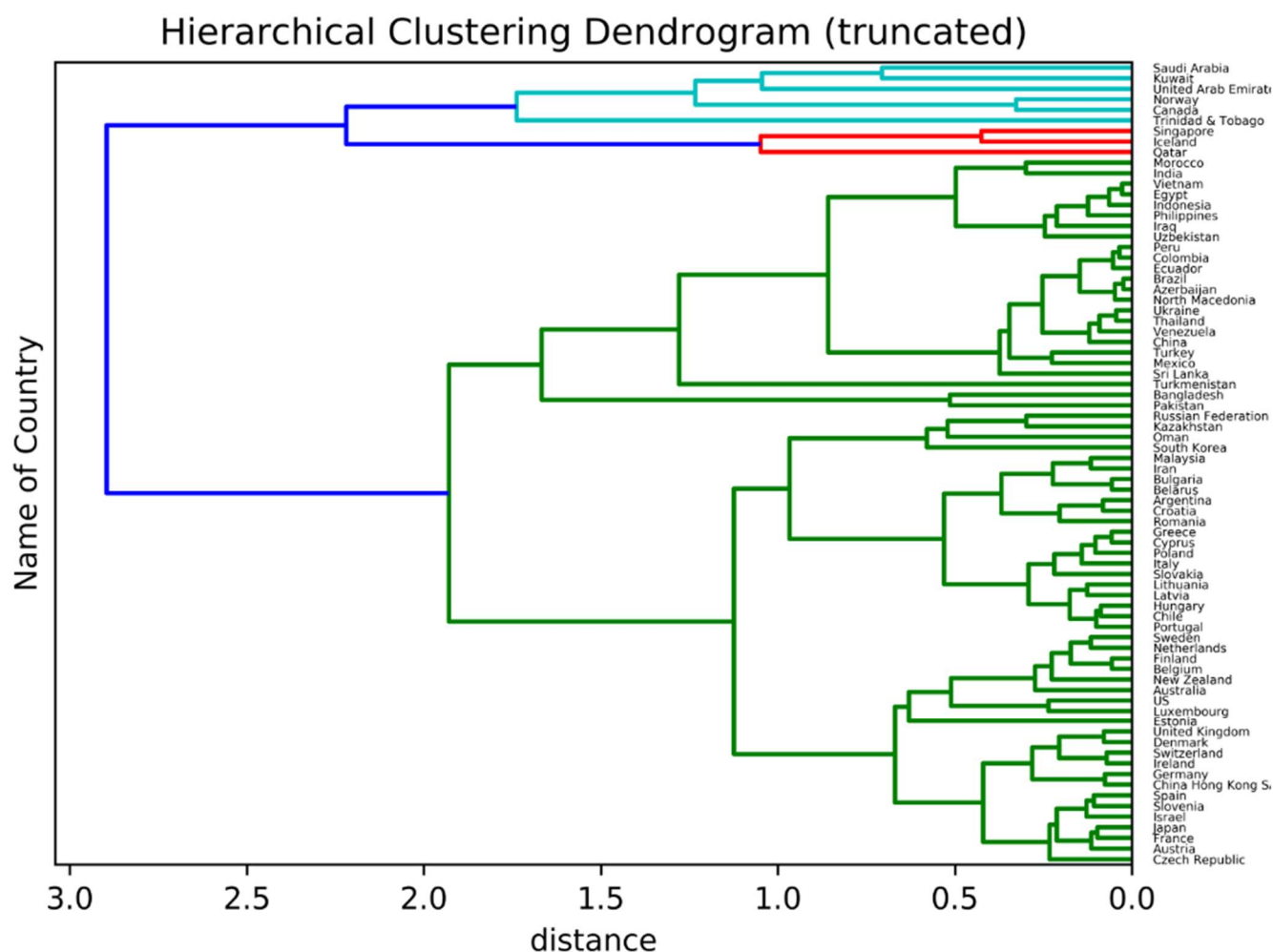


Рисунок 4.12 – Дендрограма

Дана дендрограма має метод, який має назву «Average». Даний метод побудови дендрограми являється найбільш популярним методом для побудови ієрархічних

методів кластеризації. З назви можна зрозуміти, що даний метод побудови виконується таким чином, що відстань між двома групами елементів визначається за допомогою середньої відстані між кожним елементом, що знаходиться в групі.

Також на даній сторінці виводяться кластери при діленні дендрограми по відстані 0.5. Після ділення створюється таблиця, до якої заноситься номер кластеру (всього вийшов 21 кластер) та список країни, які відносяться до цього кластеру. Створена таблиця виводиться на даній сторінці під дендрограмою.

На рисунку 4.13 зображено виведення даних відносно дендрограми.

Number of Cluster	Countries	
1	['Austria' 'China Hong Kong SAR' 'Czech Republic' 'Denmark' 'France' 'Germany' 'Ireland' 'Israel' 'Japan' 'Slovenia' 'Spain' 'Switzerland' 'United Kingdom']	Show
2	['Luxembourg' 'US']	Show
3	['Belgium' 'Finland' 'Netherlands' 'New Zealand' 'Sweden' 'Australia']	Show
4	['Estonia']	Show
5	['Chile' 'Cyprus' 'Greece' 'Hungary' 'Italy' 'Latvia' 'Lithuania' 'Poland' 'Portugal' 'Slovakia']	Show
6	['Belarus' 'Bulgaria' 'Croatia' 'Iran' 'Malaysia' 'Romania' 'Argentina']	Show
7	['Kazakhstan' 'Russian Federation']	Show
8	['Oman']	Show
9	['South Korea']	Show
10	['Pakistan']	Show
11	['Bangladesh']	Show

Рисунок 4.13 – Виведення даних відносно дендрограми

Також відповідно до кожного кластеру існує клавiша, при натисканні на яку у користувача з’явиться можливість переглянути дані кожної країни, що знаходиться в кластері.

На наступній сторінці випадаючого списку виводяться результати роботи алгоритму зменшення розмірності t-SNE (див. 3.3).

На рисунку 4.14 та на рисунку 4.15 зображено початкова візуалізація вхідних даних та результат роботи даного алгоритму візуалізації даних відповідно.

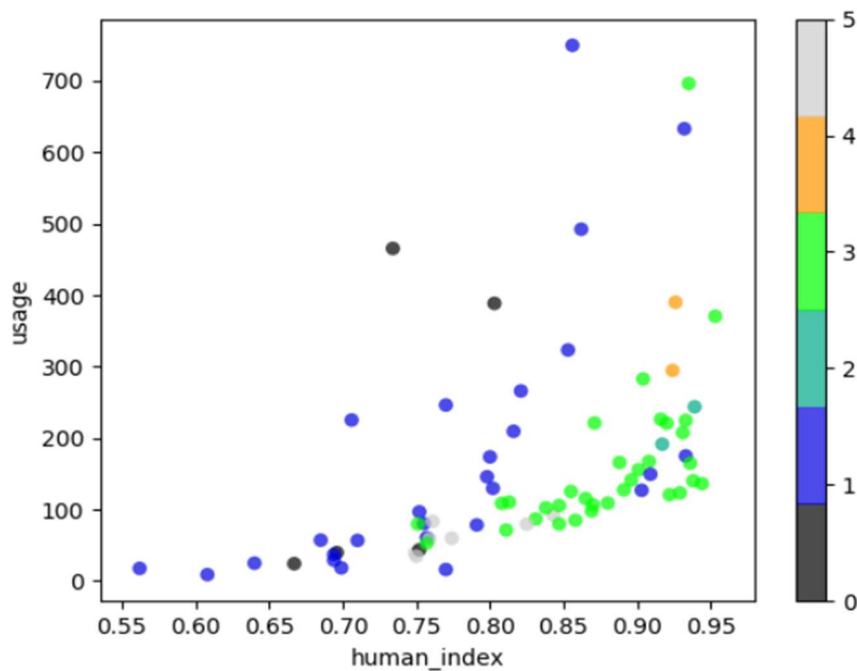


Рисунок 4.14 – Візуалізація вхідних даних

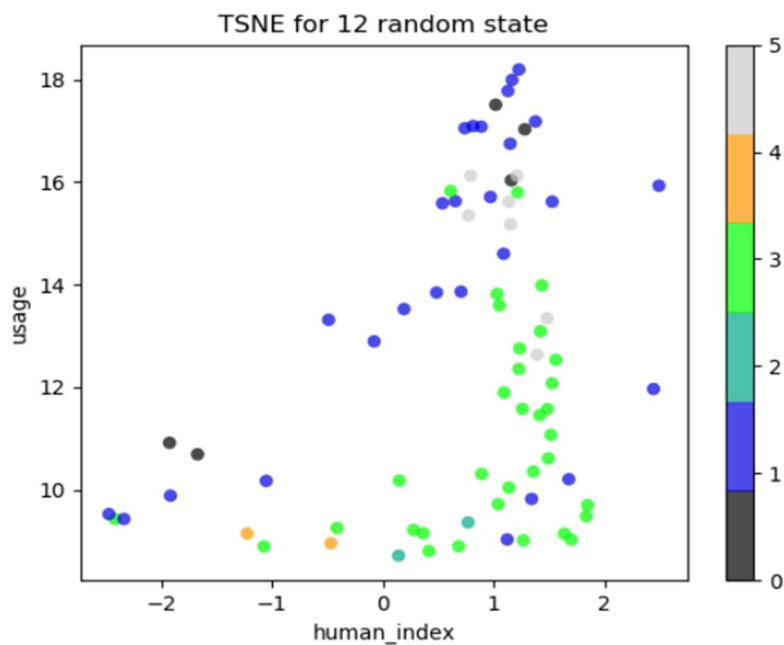


Рисунок 4.15 – Результат роботи t-SNE

На рисунку 4.15 можна побачити, що алгоритм певним чином згрупував схожі елементи між собою і вони стали знаходитись поруч один від одного. Також було обрано 12-те виконання даного алгоритму, оскільки в цьому випадку найкраще видно те, що дані групуються між собою.

Наступна (четверта) сторінка, яка призначена для виведення результатів роботи, являється результатом роботи методу кластеризації K-Means (див. 3.4). На даній сторінці користувач може побачити результат роботи методу ліктя для даного набору вхідних даних та сам метод кластеризації K-Means.

На рисунку 4.16 показано результат роботи методу ліктя:

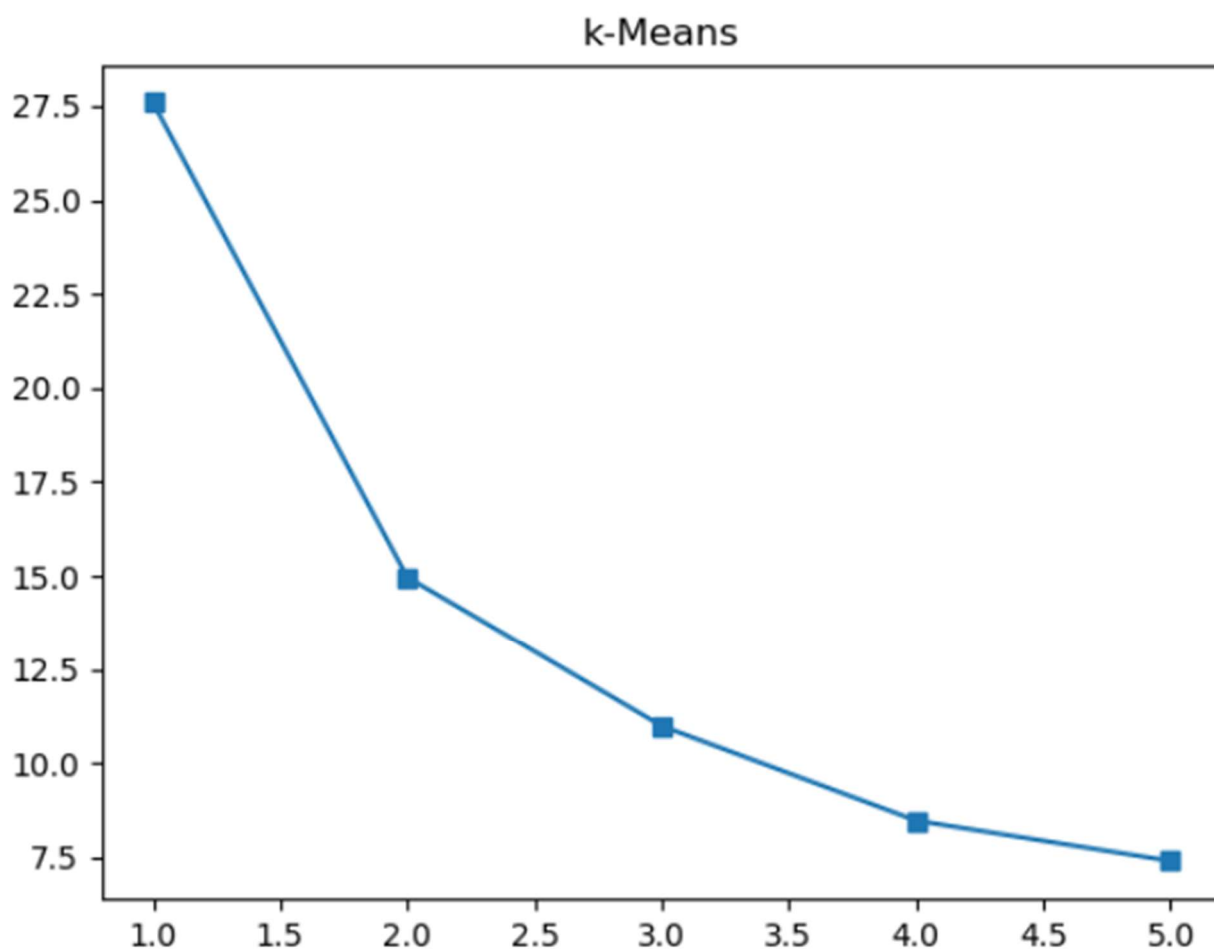


Рисунок 4.16 – Метод ліктя

На попередньому рисунку видно, що для даної задачі та для її вхідних даних оптимальна кількість кластерів дорівнює п'яти.

Після візуалізації методу ліктя представлено результат виконання методу кластеризації K-Means. На рисунку 4.17 зображено візуалізацію роботи методу даного методу.

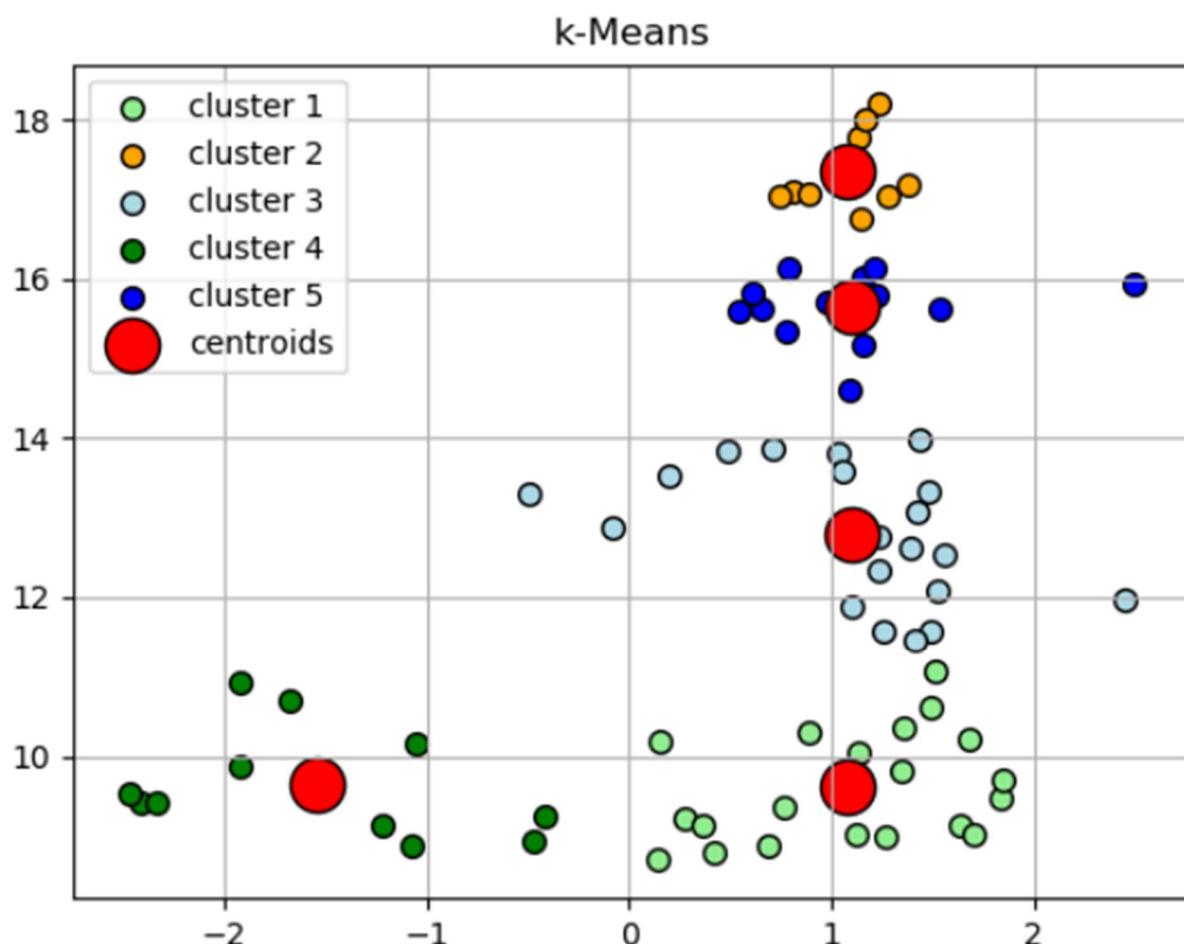


Рисунок 4.17 – Результат виконання методу K-Means

На попередньому рисунку видно візуалізацію роботи даного методу. Також можна побачити, що даний метод розбив вхідні дані на п'ять кластерів (зображені різнокольоровими крапками). Великі червоні круги вказують на центроїди відповідних кластерів.

П'ятий результат роботи створеного програмного забезпечення являється сторінка, на якій виводиться результат роботи спектрального методу кластеризації (див. 3.6). В даному випадку використовується властивість, яка має назву найближчих сусідів, який використовується для спектральної кластеризації даних.

На рисунку 4.18 зображено результат роботи даного методу в порівнянні з початковою візуалізацією даних.

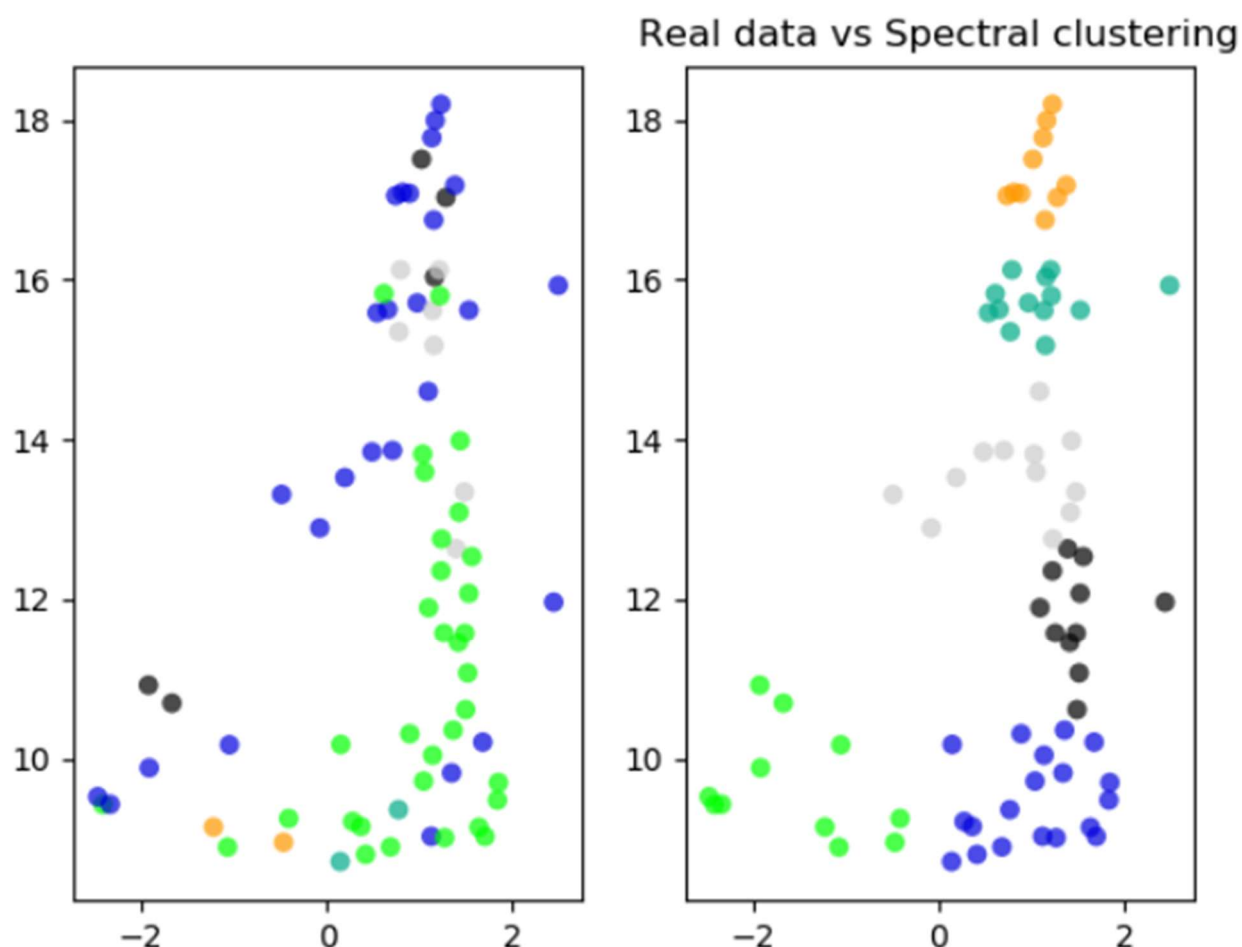


Рисунок 4.18 – Результат роботи спектральної кластеризації

На попередньому рисунку видно, що початкові дані добре групуються в кластери та самі кластери чітко виражені. Всього отримано п'ять кластерів, які виділені різними кольорами.

Передостання сторінка призначена для виведення результату роботи агломеративного кластерного аналізу. Дана сторінка має аналогічний вигляд, що й попередня сторінка для виведення результату роботи алгоритму спектральної кластеризації. Також на даній сторінці представлено порівняння початкової візуалізації даних з результатом роботи даного алгоритму. Кластери даного методу також виділяються різними кольорами.

На рисунку 4.19 зображено результат роботи алгоритму агломеративної кластеризації.

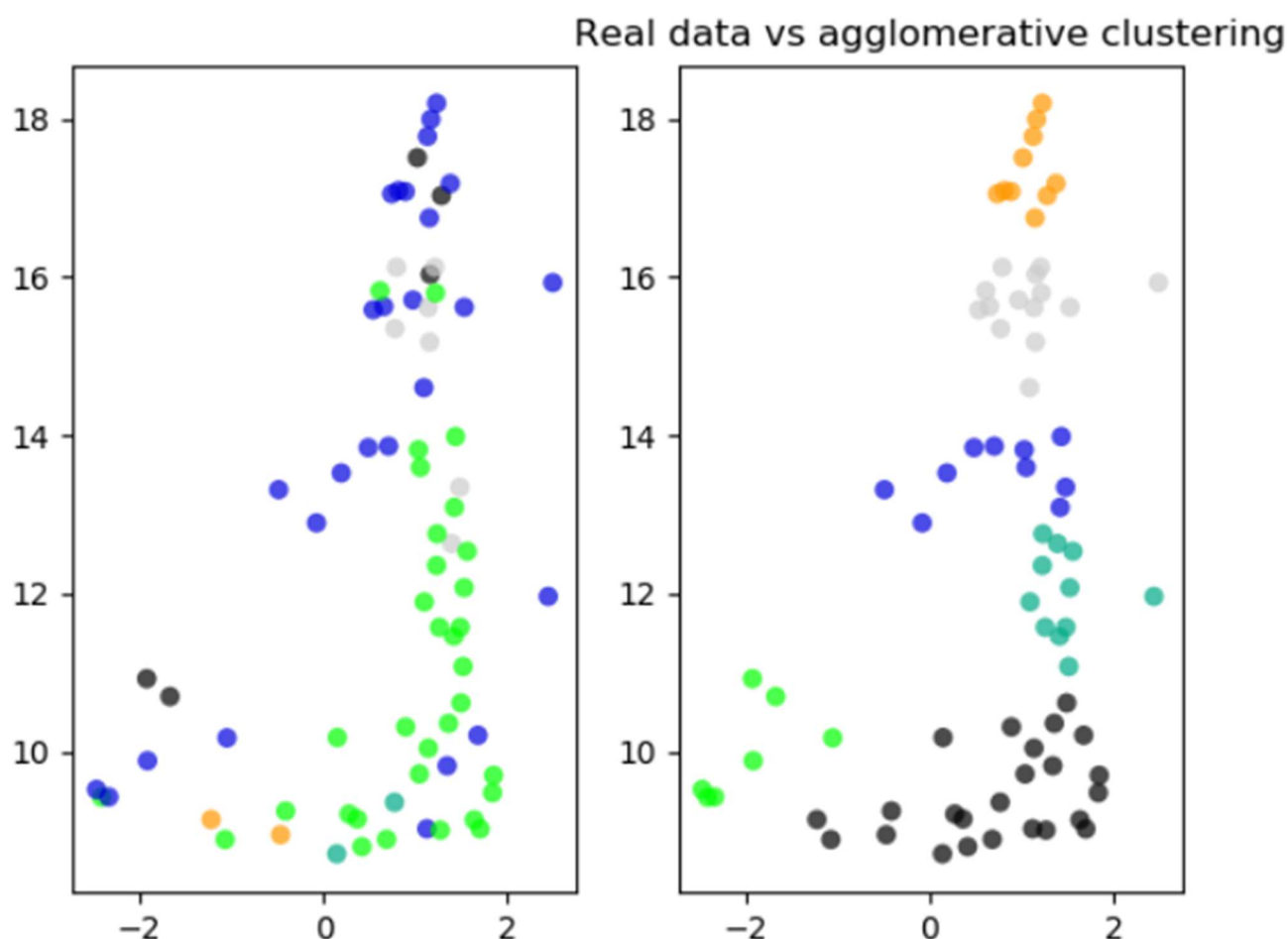


Рисунок 4.19 - Результат роботи агломеративної кластеризації

На останній сторінці зображено таблицю, в якій виконується порівняння ефективності деяких методів кластеризації відносно існуючих метрик (див. 3.7). Порівняння методів відбувається відносно вхідних даних для даної задачі.

За допомогою побудованої таблиці можна визначити, який з наведених методів являється ефективнішим та в якій метриці. Ефективнішим являється той метод, який має нижче число у відповідній метриці.

На рисунку 4.20 зображено таблицю, яка була отримана після порівняння ефективності метрик якості кластеризації.

	ARI	AMI	Homogeneity	Completeness	V-measure	Silhouette
K-means	0.130834	0.176303	0.316403	0.238915	0.272253	0.048383
Affinity	0.155040	0.192060	0.331586	0.251738	0.286197	0.060586
Spectral	0.123196	0.184244	0.326477	0.243972	0.279258	0.056052
Agglomerative	0.152988	0.200434	0.336815	0.260192	0.293586	0.004454

Рисунок 4.20 – Порівняння методів

З даної таблиці можна визначити, який алгоритм являється кращим відносно розглядуваної метрики.

- Для метрики ARI якіснішим являється агломеративний метод кластеризації;
- Для метрики AMI – метод кластеризації K-Means;
- Для Homogeneity – алгоритм кластеризації K-Means;
- Для Completeness – також алгоритм K-Means;
- Для V-measure – K-Means;
- Для Silhouette – агломеративна кластеризація.

4.6 Висновки до розділу

У цьому розділі дипломної роботи було описано архітектуру розробленої системи, проаналізовано вхідні дані, які отримуються з бази даних PostgreSQL та використані бібліотеки для виконання поставленої задачі. Також було розглянуто елементарні події, які можуть виконуватись користувачем під час роботи з системою (див. 4.3).

Також було створене відповідне програмне забезпечення для поставленої задачі з виконанням методів, що були описані у розділі 3. У даній системі користувач має

можливість виконувати маніпуляції з існуючими даними, результат яких буде зберігатися до бази даних. Також користувач може виконувати пошук по критерію, який можна обрати самостійно. Можна переглядати всі результати, які отримуються після виконання алгоритмів кластерного аналізу. Було побудовано низку діаграм та графіків, які демонструють роботу відповідних алгоритмів.

ВИСНОВКИ

Для того, щоб виконати поставлену задачу, було зроблено наступні кроки:

а) На початку було проведено огляд та якісний аналіз існуючих рішень для поставленої задачі та виконано порівняння даних рішень між собою відносно певних критеріїв. Також було проаналізовано методи, які підходять для поставленої задачі та виконано їх порівняння. Було визначено, що метод кластерного аналізу підходить більше для поставленої задачі, оскільки він має низку переваг відносно інших, а також даний метод кластеризації даних був запропонованим до виконання у рамках наукових робіт Інституту демографії та соціальних досліджень ім. М.В. Птухи НАН України;

б) Далі було проведено детальний огляд алгоритмів, за допомогою який виконується методу кластерного аналізу. Після огляду було проведено їх детальне вивчення та описання математичної складової цих методів з використанням відповідних формул. Також було переглянути приклади виконання даних методів та які існують можливості для реалізації даних алгоритмів;

в) Було розроблено веб-застосунок для поставленої задачі з простим та зрозумілим у використанні інтерфейсом. За допомогою даного програмного забезпечення у користувача є можливість переглядати всі існуючі дані, які занесено до бази даних PostgreSQL, та робити з ними такі дії, як створення нового екземпляру даних, редагування та видалення існуючих даних. Користувач також має можливість виконувати пошук, який ведеться по всім існуючим даним відносно критерію, який користувач обирає самостійно. Для цих операцій було створено відповідні форми для полегшення використання даної системи. Також було створено сторінки для виведення результатів виконання обраних методів кластерного аналізу та для візуалізації існуючих даних.

Проаналізувавши отримані результати кластерного аналізу, а саме дендрограму та відповідні кластери, можна сказати, що країни, які мають вищий рівень якості життя, тобто індекс людського розвитку, використовують більшу кількість енергії. Також можна побачити те, що країни, індекс людського розвитку яких є меншим, використовують відповідно меншу кількість енергії. З цього можна зробити висновок, що якість життя населення має сильну залежність від кількості використаної енергії. Хоча також можна помітити той фактор, що деякі країни хоча мають високий рівень людського розвитку, але використовують меншу кількість енергії. Це зумовлено тим, що ці країни відносяться до вже розвинутих країн, що вказує на те, що вони використовують альтернативні способи електроенергії.

Подальшим вдосконаленням даного програмного продукту може бути розширення бази даних, тобто додавання нових даних, нових сутностей тощо, після чого можливо збільшити кількість використовуваних методів, які відносяться до методів аналізу даних. Також, як вдосконалення програмного продукту, може слугувати деталізація результатів виконання обраних алгоритмів.