МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ
ІМЕНІ ВАДИМА ГЕТЬМАНА

Навчально-науковий інститут
«Інститут інформаційних технологій в економіці»

Кафедра математичного моделювання та статистики

Освітньо-професійна програма	**«Економічна кібернетика»**
Галузь знань	**05 «Соціальні та поведінкові науки»**
Спеціальність	**051 «Економіка»**

Форма навчання: очна (денна)
очна (денна)/ дистанційна

КВАЛІФІКАЦІЙНА БАКАЛАВРСЬКА РОБОТА

на тему **«Аналіз та оцінювання впливу макроекономічних змін на індекс розвитку людини в Україні та світі»**
(*назва теми*)

здобувача Гринька Іллі Євгеновича _____
(*ПІБ*) *(підпис)*

Науковий керівник: кандидат економічних наук, доцент
Осипова Ольга Ігорівна

_____
(підпис)

Робота допущена до захисту перед екзаменаційною комісією з атестації здобувачів вищої освіти (ЕК)

В.о. завідувача кафедри: кандидат фізико-математичних наук,
професор Великоіваненко Г.І.

_____
(підпис)

Київ 2024

ЗМІСТ

ВСТУП

Актуальність теми. Одним із ключових питань в умовах трансформаційних перетворень соціально-економічної системи України є не лише збереження, а й нарощування людського потенціалу країни. Це питання набуває особливої гостроти у зв'язку з сучасними викликами, які стоять перед українським суспільством, зокрема, у контексті війни, економічних потрясінь та соціальних змін.

Водночас нагальним щодо теоретичного і практичного аспектів є всебічне розуміння наявних тенденцій, залежностей і факторів, спрямованих на формування стану, рівня та перспективи забезпечення високоякісної динаміки людського розвитку українського населення за непростих поточних воєнних умов. Це включає подальше дослідження та розробку наукових підходів та ефективних інструментів для забезпечення цього процесу, що є важливим не лише з точки зору стабільності, але й для довгострокового прогресу країни.

Новий характер економічного зростання, в якому визначальну роль відіграють нематеріальні фактори, зокрема позитивні людські якості, робить розвиток людського потенціалу особливо актуальним. Позитивні людські якості, такі як інноваційність, креативність, адаптивність та соціальна згуртованість, є невід'ємною складовою людського розвитку. Їхній розвиток є критично важливим для забезпечення конкурентоспроможності країни на світовій арені.

Консолідовані державні стратегії, спрямовані на забезпечення розвитку людського потенціалу, мають враховувати регіональні розбіжності у рівні людського розвитку та його фундаментальних аспектах. Це вимагає детального опрацювання та дослідження, оскільки регіональні відмінності можуть значно впливати на ефективність реалізації державної політики та програм. Розуміння цих відмінностей дозволить розробити більш цільові та ефективні заходи, спрямовані на розвиток людського потенціалу в різних частинах країни, що в кінцевому результаті сприятиме загальному прогресу та стабільності.

Аналіз останніх досліджень і публікацій. Проблематиці вивчення людського розвитку присвячені наукові дослідження вчених України і світу. Питання застосування статистичних методів і моделей для вивчення соціально-економічних явищ та процесів знайшли відображення в наукових розробках О. Власюка, В. Гейця, Е. Лібанової, О. Макарової. Серед дослідження теоретичних і практичних аспектів рівня життя населення в Україні можна виділити досідження Д. Богиня, О. Гладуна, В. Мандибури, С. І. Пирожкова. Щодо вітчизняних авторів, роботи яких доповнюють і розвивають концептуальні ідеї людського потенціалу, людського капіталу й людського розвитку, слід назвати таких вчених: Н. Борецька, О. Грішнова, Г. Дмитренко, І. Каленюк, О. Кизима, Н. Ковтун, І. Кочума, І. Куценко, Л. Лісогор, О. Мельниченко, Б. Міланович, В. Никифоренко, В. Новіков, В. Онікієнко, О. Палій, А. Ревенко, В. Стешенко, В. Тропіна, С. Тютюнникова, В. Мельник, А. Чухно, В. Шишкін.

Мета та завдання виконання кваліфікаційної бакалаврської роботи. Дослідження впливу макроекономічних змін на індекс розвитку людини в Україні та світі з використанням економетричних методів та кластерного аналізу та запропонування рекомендації щодо вдосконалення політики національного та міжнародного розвитку з урахуванням отриманих результатів.

Задля досягнення поставленої мети визначено такі завдання:

У розділі 1:

- опис історичних та економічних засад використання індексу людського розвитку;
- огляд основних складових індексу розвитку людини та їх значення дослідити методами економічно-математичного аналізу регресійні чинники, дати їм характеристику;
- огляд математичних методів для аналізу та оцінювання рівня людського розвитку.

У розділі 2:

- опис інформаційної бази для регресійного аналізу;

- побудова економетричної моделі: вибір типу регресійної моделі та розрахунок параметрів впливу макроекономічних факторів на індекс розвитку людини;
- перевірка адекватності та статистичної значущості моделі;
- аналіз отриманих результатів та їх інтерпретація: висновки щодо впливу макроекономічних факторів на індекс розвитку людини.

У розділі 3:

- опис інформаційної бази для кластерного аналізу;
- проведення кластерного аналізу;
- дослідження різниці між отриманими в результаті побудови моделі кластерами;
- аналіз місця України серед країн Європи за соціально-економічними показниками та розробка рекомендацій для покращення її позицій.

Об’єктом дослідження є індекс розвитку людини в Україні та світі.

Предметом дипломної роботи є аналіз впливу макроекономічних змін на показники індексу розвитку людини в Україні та світі.

Методи дослідження. Методологічну основу дослідження становлять наукові розробки вітчизняних та зарубіжних вчених-економістів. Зокрема, при написанні кваліфікаційної бакалаврської роботи були використані сучасні методи інтелектуального аналізу даних. Для здійснення комплексного дослідження було необхідно застосування таких методів:

- теоретичний та економіко-логічний аналіз, синтез, систематизація і класифікація матеріалів та друкованих джерел з досліджуваної проблеми;
- порівняльний (дозволив визначити розбіжності і визначити методологічні аналогії та відмінності у поглядах дослідників);
- розвідувальний аналіз даних (використовувався для підготовчого аналізу наборів даних задля визначення спільних тенденцій та закономірностей, характеру та особливостей даних аналізу; для пошук найважливіших ознак та узагальнення їх основних характеристик, законів розподілу величин, виявлення відхилень та аномалій);

- динамічний та статистичний аналіз, порівняння, узагальнення;
- регресійний (дозволив виявити виокремлений і загальний вплив чинників на індекс розвитку людини, зокрема макроекономічних, а також розробити відповідну модель завдяки використанню відповідних критеріїв для перевірки гіпотези);
- кластерний аналіз (цей статистичний метод було застосовано для групування спостережень у кластери);
- тощо.

Питання тенденцій розвитку людини потребує подальшого дослідження. Розроблені теоретичні та методичні положення можуть бути використані бути використані в навчальному процесі для економічних дисциплін. Одержані результати та рекомендації формують наукове підґрунтя для подальших досліджень.

Інформаційна база дослідження. Кваліфікаційна бакалаврська робота виконується на матеріалах сайтів ООН, Світового банку, МВФ, Transparency International, Worldometers [1-13].

Структура роботи. Робота складається зі вступу, трьох розділів, висновків, переліку використаних джерел та додатків.

Зміст розділів такий:

Розділ 1 – Теоретичні засади оцінювання рівня людського розвитку;

Розділ 2 – Побудова економетричної моделі оцінювання впливу макроекономічних факторів на індекс розвитку людини;

Розділ 3 – Кластерний аналіз як інструмент систематизації країн за рівнем людського розвитку та іншими макроекономічними показниками.

РОЗДІЛ 1
ТЕОРЕТИЧНІ ЗАСАДИ ОЦІНЮВАННЯ РІВНЯ ЛЮДСЬКОГО РОЗВИТКУ

1.1 Опис об’єкту дослідження

Індекс людського розвитку вперше був розрахований у 1990 році в прагненні вийти за рамки ВВП як мірила добробуту. А вже сьогодні Організація Об’єднаних Націй прогнозує зниження якості життя людства. Через COVID-19, війни в Україні та наслідки зміни клімату спричинили регрес у глобальному розвитку [14].

Війна перетворює людей назавжди. Боротьба за розвиток та освіченість нації – один з-поміж головних пріоритетів в умовах війни. Українська хоробра і непереможна нація мусить вибудовувати сильну державу.

Індекс людського розвитку (ІЛР) – один з найбільш складних комплексних показників рівня людського потенціалу та рівня якості життя – є поєднанням трьох аспектів (індикаторів, факторів): очікуваної тривалості життя при народженні, середньої кількості років освіти та очікуваної кількості років навчання, об’єднаних в єдиний індекс освіти, та економічних благ, виражених обсягом виробництва, або ВВП за купівельною спроможністю [15].

ІЛР був створений для того, щоб підкреслити, що люди та їхні можливості мають бути головним критерієм оцінки розвитку країни, а не лише економічне зростання. Також індекс можна використовувати для того, щоб поставити під сумнів вибір національної політики, запитуючи, як дві країни з однаковим рівнем ВНД на душу населення можуть мати різні показники людського розвитку. Ці контрасти можуть стимулювати дебати про пріоритети державної політики.

ІЛР спрощує і відображає лише частину того, що включає в себе людський розвиток. Він не відображає нерівність, бідність, людську безпеку, розширення прав і можливостей тощо.

Загалом у світі індекс людського розвитку залежить від стану національної економіки, політичної та соціальної галузі, зовнішніх умов та варіюється від низького до дуже високого рівня.

Держави із соціально-ринковим господарством мають значний рівень розвитку людського потенціалу, відтак дослідження їхнього доробку становить неабияку науково-практичну цінність для нашої держави, котра відчутно поступається показниками людського розвитку перед розвиненими країнами. Незважаючи на те, що доходи громадян не виступають кінцевою ціллю та центральним індикатором людського прогресу, саме вони слугують його фундаментом.

В рамках публікації дослідження «Україна 2030: Доктрина збалансованого розвитку» експерти проаналізували перспективи соціально-економічного розвитку України на основі положень Декларації G20 зі сталого розвитку. Фахівці виділили 19 індикаторів, які дозволяють оцінити рівень суспільного розвитку. Автори зазначають, що головним рушієм і джерелом збалансованого розвитку є щаслива людина, її творчий (креативний) потенціал, відтак першочергової уваги набуває інструментарій Індексу щастя та Індексу людського розвитку [16].

У цьому плані макроекономічний рівень – це процес створення державою умов для можливостей вибору та руху людей від нижчого до вищого рівня шляхом визначення стратегічних пріоритетів і використання суспільних (громадських) фінансових ресурсів, в результаті чого відбувається формування й подальші якісні зміни у процесі функціонування людського капіталу та досягаються конкурентні переваги держави [17].

Власне, за критерієм нагальності потреб та ієрархічності взаємовідносин поміж ними виділяють моделі Ф. Герцберга, А. Маслоу, К. Альдерфера. Попри наявність певних відмінностей, всі ці моделі вирізняють потреби найнижчого порядку та найвищого. Відповідно до такого підходу найвищі потреби індивіда не виходять на перший порядок допоки не задовольняться найбільш нагальні.

Модель Маслоу, зокрема, структурована у вигляді піраміди, де на нижньому рівні знаходяться фізіологічні потреби, такі як їжа, вода та сон, які необхідно

задовольнити перш за все. Наступним рівнем є потреби безпеки, що включають захист від фізичних та емоційних загроз. Лише після задоволення цих базових потреб виникають соціальні потреби у відносинах та приналежності, далі йдуть потреби у повазі та самоповазі, а на вершині піраміди розташовуються потреби у самореалізації.

Модель Альдерфера, відома як ERG теорія, спрощує цю ієрархію, об'єднуючи потреби в три категорії: існування (Existence), зв'язок (Relatedness) і зростання (Growth). Теорія ERG допускає, що потреби можуть задовольнятися одночасно на різних рівнях, і якщо задоволення вищого рівня є недосяжним, людина може повернутися до потреб нижчого рівня для отримання більшої задоволеності.

Ф. Герцберг запропонував двофакторну теорію, що розділяє фактори задоволення на гігієнічні (контекстуальні) та мотиваційні (внутрішні). Гігієнічні фактори, такі як умови праці, зарплата і безпека, є базовими і їх відсутність може викликати незадоволення, але їх наявність не обов'язково стимулює вищий рівень задоволення або мотивації. Мотиваційні фактори, навпаки, пов'язані з самореалізацією та визнанням і можуть істотно підвищувати рівень задоволеності роботою та мотивацію, коли базові гігієнічні потреби вже задоволені.

1.2. Огляд методів і моделей, що застосовувались

1.2.1 Розвідувальний аналіз даних (exploratory data analysis, EDA, РАД)

Після збору даних, аналізу та їхньої інтерпретації переходимо до наступного логічного етапу – розвідувального аналізу даних. Цей підхід допомагає дізнатися, яку інформацію можуть дати нам дані, поза формальним завданням моделювання або перевірки гіпотез. Метод EDA допомагає проаналізувати набори даних для узагальнення їх статистичних характеристик; це спосіб узагальнення даних шляхом

виокремлення їхніх основних характеристик та візуалізації за допомогою відповідного представлення [18].

Типи розвідувального аналізу даних [19]:

• одновимірний неграфічний аналіз – найпростіша форма аналізу даних, який використовується для вивчення однієї змінної. Цей підхід дозволяє описати дані та виявити закономірності, що існують в цій змінній. Однак він не розглядає причинно-наслідкові зв’язки між змінними;

• одновимірний графічний аналіз – важливі інструменти для вивчення даних. До типів одновимірного графічного зображення відносять:

- стовпчасті діаграми, які відображають всі значення змінної та їхній розподіл;

- гістограми, стовпчасті діаграми, в яких кожен стовпчик представляє частоту (кількість) або частку (відношення кількості до загальної кількості) випадків для діапазону значень;

- коробчасті діаграми, які графічно відображають статистичні характеристики, як-от, зведення мінімуму, першого квартилю, медіани, третього квартилю та максимуму;

• багатовимірний неграфічний аналіз – аналіз даних, який включає більше однієї змінної. Цей підхід EDA дозволяє вивчати взаємозв’язки між двома або більше змінними за допомогою перехресних таблиць або статистики;

• багатовимірний графічний аналіз – графічний аналіз, який використовує графіку для відображення взаємозв’язків двома або більше наборами даних. Наприклад, найпоширенішими є згруповані гістограми або стовпчикові діаграми, які демонструють взаємозв’язок між рівнями різних змінних.

Часто під час математичного моделювання дослідники прагнуть встановити причинно-наслідковий вплив однієї змінної на іншу. Регресійний аналіз – це статистичний інструмент для дослідження взаємозв’язків між змінними. Зібравши дані про основні змінні, що становлять інтерес, цей метод дозволяє використати регресію для оцінки кількісного впливу факторних змінних на залежні.

–На меті виявити закономірності, що пояснюють розвиток нашої держави та нашого суспільств, виявити показники, що мають найбільший вплив на рівень життя та розвиток громадян. Наступний крок – створення регресійної моделі для перевірки гіпотези, а потім слідує оцінка статистичної значущості оцінюваних взаємозв'язків, тобто визначення рівня впевненості в тому, що реальний взаємозв'язок близький до оцінюваного взаємозв'язку. Підсумком стануть статистичні висновки.

Початковий набір даних складається з 28-ти спостережень – років.

Джерела інформації – дані ООН, Світового банку, МВФ, Transparency International, Worldometers [1-13].

Дослідження буде полягати у побудові множинної регресійної моделі для виявлення того, наскільки індекс людського розвитку визначається іншими показниками. Регресійний аналіз дозволяє зробити обґрунтовані висновки стосовно обраного напрямку, який спирається на конкретні математичні обчислення.

1.2.2 Регресійний аналіз

Регресія – це статистичний метод, що використовується в економіці, фінансах, інвестуванні та інших науках. Регресійний зв'язок між змінними Y та X є таким, коли одна з них, наприклад X, вибирається як незалежна змінна, її називають пояснювальною змінною (регресором), а друга Y – залежна (пояснювана, регресант). В цьому випадку пояснююча змінна X (регресор) є причиною зміни залежної змінної Y. Під поняттям регресії розуміють функціональну залежність між умовним математичним сподіванням випадкової величини Y від значень пояснювальної змінної X [20, 360 с.].

Моделі лінійної регресії здобули широке застосування в економічних розвідках, однак вони є найбільш спрощеним засобом у моделюванні реальних економічних процесів.

Лінійна регресія, яку також називають простою регресією або методом найменших квадратів (МНК) встановлює лінійний зв'язок між двома змінними на основі лінії найкращої відповідності. Таким чином, лінійна регресія графічно зображується за допомогою прямої лінії з нахилом, що визначає, як зміна однієї змінної впливає на зміну іншої.

Метою ж множинного регресійного аналізу є використання незалежних змінних, значення яких відомі, для прогнозування значення однієї залежної величини. Кожна предикторна величина зважується, причому ваги позначають її відносний внесок у загальний прогноз. Рівняння множинної регресії задане через формулу (див. формулу 1.1):

$$Y = a + b_1X_1 + b_2X_2 + \ldots + b_nX_n, \quad (1.1)$$

де Y – залежна змінна;

a – вільний член;

$b_1 \ldots b_n$ – вагові коефіцієнти;

$X_1 \ldots X_n$ – незалежні змінні.

Множинна лінійна регресія базується на таких припущеннях [20, 21]:

1. Лінійний зв'язок між залежною та незалежними змінними:

Перше припущення множинної лінійної регресії полягає в тому, що існує лінійний зв'язок між залежною змінною та кожною з незалежних змінних. Найкращим способом перевірки лінійного зв'язку є створення діаграм розсіювання, та подальша візуальна перевірка діаграм розсіювання на лінійність. Якщо зв'язок, відображений на діаграмі розсіювання, не є лінійним, то необхідно провести нелінійну регресію або трансформувати дані;

2. Незалежні змінні значно не корелюють між собою:

Дані не повинні демонструвати мультиколінеарність, яка виникає, коли незалежні змінні (пояснювальні змінні) сильно корелюють між собою. Коли незалежні змінні демонструють мультиколінеарність, виникають проблеми з визначенням конкретної змінної, яка робить внесок у дисперсію залежної змінної. Для перевірки на мультиколінеарність буде використаний vif-критерій;

3. Дисперсія залишків є постійною:

Множинна лінійна регресія припускає, що величина помилки в залишках є однаковою в кожній точці лінійної моделі. Такий сценарій відомий як гомоскедастичність. Аналізуючи дані, аналітик повинен побудувати графік стандартизованих залишків у порівнянні з прогнозованими значеннями, щоб визначити, чи справедливо розподілені точки по всіх значеннях незалежних змінних. Для перевірки припущення дані можна нанести на діаграму розсіювання або за допомогою статистичного програмного забезпечення створити діаграму розсіювання, яка включає всю модель;

4. Незалежність спостережень:

Модель передбачає, що спостереження повинні бути незалежними одне від одного. Тобто, модель припускає, що значення залишків є незалежними;

5. Багатовимірна нормальність:

Багатовимірна нормальність має місце, коли залишки розподілені нормально. Щоб перевірити це припущення, треба подивитись, як розподілені значення залишків. Це також може бути перевірено за допомогою двох основних методів, тобто гістограми з накладеною нормальною кривою або методу нормального розподілу ймовірностей.

Розрахунки та візуалізація виконувалися за допомогою мови програмування Python з використанням бібліотек pandas, numpy, sklearn, matplotlib, Prophet, statsmodels та seaborn.

1.2.3 Кластерний аналіз

Кластерний аналіз – це техніка статистичної класифікації, за якої набір об’єктів зі схожими характеристиками групується в кластери. Він охоплює низку різних алгоритмів і методів, які використовуються для групування об’єктів подібного типу у відповідні категорії. Мета кластерного аналізу полягає в тому, щоб організувати спостережувані дані в значущі структури, щоб отримати подальше розуміння з них.

Існує кілька різних методів, які можна використовувати для проведення кластерного аналізу. Ці методи можна класифікувати наступним чином [22-23]:

- Ієрархічні методи:

- Агломераційні методи, за яких суб’єкти починають із власного окремого кластера. Потім два «найближчих» (найбільш схожих) кластери об’єднуються, і це робиться неодноразово, доки всі об’єкти не опиняться в одному кластері. Зрештою, з усіх кластерних рішень обирається оптимальна кількість кластерів;

- Методи розділення, за яких усі суб’єкти починаються з одного кластеру, а наведена вище стратегія застосовується у зворотному порядку, поки кожен суб’єкт не опиниться в окремому кластері;

- Неієрархічні методи (часто відомі як методи кластеризації *k*-середніх).

Під час кластерного аналізу важливо враховувати тип даних, які використовуються. Дані, які використовуються в кластерному аналізі, можуть бути інтервальними, порядковими або категоріальними. Однак наявність різних типів змінних може ускладнити аналіз. Це пояснюється тим, що під час кластерного аналізу вам потрібно мати певний спосіб вимірювання відстані між спостереженнями, а тип вимірювання залежатиме від типу даних, які ми маємо. Для інтервальних даних найпоширенішою мірою відстані є евклідова відстань. Вона дозволяє обчислити відстань між двома точками у просторі.

Евклідова відстань між двома об’єктами задається через формулу (див. формулу 1.2):

$$d_{ij} = \sqrt{\left(x_{i1} - x_{j1}\right)^2 + \left(x_{i2} - x_{j2}\right)^2 + \ldots + \left(x_{ip} - x_{jp}\right)^2}, \quad (1.2)$$

де d – відстань;

p *– кількість змінних;*

x_i *– дані першого об’єкту;*

x_j *– дані другого об’єкту.*

При використанні такої міри, як евклідова відстань (1.1), виникає проблема масштабу вимірювання змінних, що розглядаються, оскільки зміна масштабу, очевидно, вплине на відстань між суб’єктами: наприклад, різниця в 10 см може бути також різницею в 100 мм. Крім того, якщо одна змінна має значно ширший діапазон, ніж інші, вона може домінувати в обчисленні відстані. До прикладу, якщо виміри тіла були зроблені для кількох різних людей, діапазон (в мм) зросту буде набагато ширшим, ніж діапазон, скажімо, обхвату зап’ястя. Щоб уникнути цього, кожна змінна може бути стандартизована.

Однак це саме по собі створює проблему, оскільки має тенденцію до зменшення відстані між кластерами. Це відбувається тому, що якщо певна змінна добре розділяє спостереження, то, за визначенням, вона матиме велику дисперсію. Якщо ж цю змінну стандартизувати, то розбіжність між кластерами стане меншою. Якщо виникають сумніви, можна провести кластерний аналіз двічі: один раз без стандартизації та інший раз зі стандартизацією, щоб оцінити вплив на отримані кластери.

Основне питання в ієрархічній кластеризації полягає в тому, як обчислювати відстань між кластерами та оновлювати матрицю близькості. Існує багато різних підходів для відповіді на це питання. Кожен підхід має свої переваги та недоліки. Вибір буде залежати від того, чи є в наборі даних шум, чи є форма кластерів круглою чи ні, а також від щільності точок даних. Наведемо основні методи [23]:

- Метод найближчого сусіда (метод одного зв’язку): один із способів виміряти відстань між кластерами – це знайти мінімальну відстань між точками в цих кластерах. Тобто, можна знайти точку в першому кластері, найближчу до точки в іншому кластері, і обчислити відстань між цими точками.

- Метод найдальшого сусіда (метод повного зв’язку): у цьому випадку відстань між двома кластерами визначається як максимальна відстань між членами - тобто відстань між двома об’єктами, які знаходяться найбільш далеко один від одного. Цей метод має тенденцію до створення компактних кластерів однакового розміру, але, як і метод найближчого сусіда, не враховує структуру кластерів. Він також є досить чутливим до викидів.

- Метод середнього (між групами) зв’язку (іноді його називають UPGMA): відстань між двома кластерами розраховується як середня відстань між усіма парами респондентів у двох кластерах.

- Метод центроїдного зв’язку: цей метод визначає відстань між кластерами як відстань між їхніми центрами/центроїдами. Після обчислення центроїда для кожного кластера, відстань між центроїдами обчислюється за допомогою функції відстані.

- Метод Варда: у цьому методі об’єднуються всі можливі пари кластерів і обчислюється сума квадратів відстаней всередині кожного кластера. Потім ця сума підсумовується по всіх кластерах. Обирається комбінація, яка дає найменшу суму квадратів. Іншими словами, метод Варда намагається мінімізувати суму квадратів відстаней точок від центрів кластерів.

Порівняно з описаними вище відстанями, метод Варда менш чутливий до шуму та викидів. Тому методу Варда надають перевагу при кластеризації більше, ніж іншим.

Цей метод має тенденцію створювати кластери приблизно однакового розміру, що не завжди бажано. Він також досить чутливий до викидів. Незважаючи на це, він є одним з найпопулярніших методів, поряд з методом середнього зв’язку. Саме цей метод буде використаний при проведенні кластерного аналізу.

Як ми зазначали, після проведення кластерного аналізу необхідно обрати «найкраще» кластерне рішення.

При проведенні ієрархічного кластерного аналізу процес може бути представлений на діаграмі, відомій як дендрограма. Ця діаграма ілюструє, які кластери були об’єднані на кожному етапі аналізу, а також відстань між кластерами на момент об’єднання.

Якщо відстань між кластерами різко змінюється від одного етапу до іншого, це свідчить про те, що на одному етапі були об’єднані кластери, які знаходяться відносно близько один до одного, тоді як на наступному етапі об’єднані кластери були відносно далеко один від одного.

Метод *k*-середніх [23]:

1. Дослідник визначає кількість кластерів, що необхідно утворити. Найчастіше, за дендрограмою, отриманою за допомогою ієрархічного кластерного аналізу.

2. Випадковим чином обирається *k* спостережень, які на цьому кроці вважаються центрами кластерів.

3. Кожне спостереження «приписується» до одного з n кластерів — того, відстань до якого найкоротша.

4. Розраховується новий центр кожного кластера як елемент, ознаки якого розраховуються як середнє арифметичне ознак об’єктів, що входять у цей кластер.

5. Відбувається така кількість ітерацій (повторюються кроки 3-4), поки кластерні центри стануть стійкими (тобто при кожній ітерації в кожному кластері опинятимуться одні й ті самі об’єкти), дисперсія всередині кластера буде мінімізована, а між кластерами – максимізована.

Перед тим, як розподіляти спостереження на кластери відповідно до обраних вихідних ознак, варто визначити, чи відповідають дані вимогам однорідності (відсутність пропусків, відхилень та аномалій), чи відповідають вони нормальному закону розподілу тощо. Тому перед застосуванням обраного методу також проводять розвідувальний аналіз даних, який дозволяє виявити можливі проблеми, такі як наявність пропущених значень, аномальних спостережень або

інших невідповідностей, що можуть суттєво вплинути на результати кластеризації. Розвідувальний аналіз включає в себе різноманітні методи та техніки для дослідження структури даних, виявлення закономірностей і тенденцій, оцінки розподілу змінних та їх взаємозв'язків. Завдяки виконанню цього етапу можна забезпечити кращу якість даних, коректність вибору моделі та точність отриманих кластерів.

Застосовуючи методи кластерного аналізу, ми маємо на меті виокремити кластери країн за рівнем людського розвитку. Цей метод передбачає проведення дослідження розподілу вибірки вхідних даних за однорідними групами так, щоб об'єкти в межах групи були подібними між собою згідно з деяким критерієм, а об'єкти з різних груп істотно розрізнялися між собою.

Для побудови економіко-математичної моделі обрано такі показники:

• індекс розвитку людини;

• ВВП на душу населення;

• рівень інфляції;

• рівень безробіття;

• індекс сприйняття корупції;

• індекс інновацій; індекс права на власність;

• відсоток міського населення; очікувана тривалість життя.

Ми вважаємо, що оптимальним є розбиття на 4 кластери. Для дослідження країн за рівнем людського розвитку був обраний метод *k*-середніх. Принцип роботи методу ґрунтується на побудові наперед заданої кількості кластерів *k*. Згідно з методом *k*-середніх країни розбивалися на чотири групи (*k*=4), розрахунки та візуалізація виконувалися за допомогою мови програмування Python з використанням бібліотек pandas, numpy, sklearn, matplotlib, Prophet, statsmodels та seaborn.

Усі дані перед початком кластерного аналізу були нормалізовані. Будуть виокремлені такі кластери: країни з дуже високим, високим, середнім та низьким рівнем людського розвитку. Найбільше нас буде цікавити кластер, до якого увійде Україна.

РОЗДІЛ 2
ПОБУДОВА ЕКОНОМЕТРИЧНОЇ МОДЕЛІ ОЦІНЮВАННЯ ВПЛИВУ МАКРОЕКОНОМІЧНИХ ФАКТОРІВ НА ІНДЕКС РОЗВИТКУ ЛЮДИНИ

2.1 Інформація про набір даних, інструменти аналізу та заповнення пропущених даних

Основне завдання нашої моделі – проаналізувати фактори, які так чи інакше впливають на індекс людського розвитку. Аби здійснити аналіз індикаторів, було зібрано статистичні дані за 1995-2022 рр. Набір даних складається із 28 спостережень – років.

Дослідження виконувалися за допомогою мови програмування Python. Ця багатофункціональна мова має кілька переваг для своїх користувачів. Вона допомагає аналітикам даних розібратися в складних наборах даних і зробити їх більш зрозумілими. Ще однією перевагою використання Python є його висока читабельність. Код на Python легше використовувати для співпраці з іншими аналітиками, для спілкування з іншими технічними зацікавленими сторонами, а також його легше підтримувати, коли настає час адаптувати його до нових джерел даних і потреб [24].

Набір даних (Додаток А.1) оформлений у файлі зі значеннями, розділеними комою (.csv) і називається Ukraine_regression. У цьому наборі даних містяться значення українських економічних показників протягом 28 років.

Усього маємо 8 числових ознак, а саме:

- Індекс людського розвитку (англ. Human development index (HDI));
- ВВП на душу населення у сучасних доларах (англ. GDP per capita, current dollars);

- Кінцеві споживчі витрати домогосподарств та неприбуткових установ, що допомагають домогосподарствам % ВВП (англ. Households and NPISHs final consumption expenditure, percent of GDP);
- Інфляція (англ. Inflation rate);
- Рівень безробіття (англ. Uneployment rate);
- Індекс сприйняття корупції (англ. Corruption perceptions);
- Відсоток міського населення (англ. Percent urban population);
- Очікувана тривалість життя (англ. Life expectancy).

У ознаці «Індекс сприйняття корупції» (англ. Corruption perceptions index) відсутні дані за перші три роки обраного періоду – за 1995, 1996, та 1997 роки. На жаль, дані щодо сприйняття корупції в Україні у вказаний період відсутні на сайті Transparency International. Організація не проводила досліджень щодо цього питання у зазначені роки.

Реальні дані майже завжди містять пропущені значення. Причинами такого явища можуть бути помилки при введенні даних або проблеми зі збиранням даних. Незалежно від причин, обробка пропущених даних є важливим кроком, оскільки статистичні результати, отримані на основі набору даних з невипадковими пропущеними значеннями, можуть бути зміщеними.

Існує три типи відсутніх даних [25]:

• відсутні абсолютно випадково (англ. Missing Completely At Random, MCAR): найвищий рівень випадковості. Це означає, що відсутні значення в будь-якій ознаці не залежать від значень інших ознак – бажаний сценарій у випадку відсутності даних;

• випадкова відсутність (англ. Missing At Random, MAR): відсутні значення в будь-якій ознаці залежать від значень інших ознак;

• відсутні не випадково (англ. Missing Not At Random, MNAR): є більш серйозною проблемою, і в цьому випадку може бути доцільно додатково перевірити процес збору даних і спробувати зрозуміти, чому інформація відсутня.

Щоб впоратись із проблемою відсутніх значень, можна [25]:

• ігнорування відсутніх значень: відсутність даних менше 10% для окремого випадку або спостереження, як правило, можна ігнорувати, за винятком випадків, коли відсутні дані є MAR або MNAR. Кількість повних випадків, тобто спостережень без пропущених даних, має бути достатньою для обраного методу аналізу, якщо неповні випадки не враховуються;

• вилучити пропущені значення/вилучити змінну: якщо дані є MCAR або MAR і кількість пропущених значень за ознакою є дуже високою, то цю ознаку слід виключити з аналізу. Якщо пропущені дані для певної ознаки або вибірки перевищують 5%, то, ймовірно, слід виключити цю ознаку або вибірку. Якщо у спостереженнях відсутні значення цільових змінних, доцільно видалити залежну змінну (змінні), щоб уникнути штучного посилення взаємозв'язку з незалежними змінними;

• видалення спостережень: за допомогою цього методу видаляються спостереження, які мають пропущені значення для однієї або декількох ознак. Якщо спостережень з відсутніми значеннями небагато, їх краще відкинути. Хоча це простий підхід, він може призвести до значного зменшення розміру вибірки. Крім того, дані не завжди можуть бути відсутніми повністю випадково. Це може призвести до зміщеної оцінки параметрів;

• обчислення («Imputation»): процес заміни відсутніх даних за допомогою деяких статистичних методів. Обчислення є корисним у тому сенсі, що воно зберігає всі випадки, замінюючи відсутні дані оціночним значенням, яке базується на іншій доступній інформації. Але ці методи слід використовувати обережно, оскільки більшість з них вносять значну похибку і зменшують дисперсію в наборі даних;

• обчислення за середнім значенням/модою/медіаною: якщо відсутні значення в стовпчику або ознаці є числовими, їх можна обчислити за середнім значенням усіх повних випадків змінної. Середнє значення можна замінити медіаною, якщо є підозра, що ознака має викиди. Для категоріальної ознаки відсутні значення можна замінити модою стовпчика. Основним недоліком цього методу є те, що він зменшує дисперсію обчислених змінних, що означає меншу

точність та надійність побудованої моделі. Цей метод також зменшує кореляцію між обчисленими змінними та іншими змінними, оскільки обчислена величина є лише оцінкою і не буде пов’язана з іншими величинами за своєю суттю;

• регресійні методи: змінні з пропущеними значеннями розглядаються як залежні змінні, а змінні з заповненими випадками – як предиктори або незалежні змінні. Незалежні змінні використовуються для підбору лінійного рівняння для спостережуваних значень залежної змінної. Це рівняння потім використовується для прогнозування значень для відсутніх точок даних.

Недоліком цього методу є те, що визначені незалежні змінні матимуть високу кореляцію із залежною змінною внаслідок відбору. Це призведе до того, що відсутні значення будуть підібрані занадто добре і зменшить невизначеність щодо цих значень. Крім того, це припускає, що зв’язок є лінійним, що може бути не так у реальності [25].

Перед проведенням попереднього аналізу в цій роботі було застосовано прогнозування за допомогою бібліотеки Prophet для мови програмування Python.

Prophet – це метод прогнозування даних часових рядів на основі адитивної моделі, в якій нелінійні тренди підганяються під річну, тижневу та денну сезонність. Найкраще працює з часовими рядами, які мають сильні сезонні ефекти та кілька сезонів історичних даних.

Prophet стійкий до пропущених даних і зсувів у тренді, і, як правило, коректно справляються з викидами. Prophet – це програмне забезпечення з відкритим вихідним кодом, випущене командою Core Data Science компанії Facebook [26].

Наведемо деякі з можливостей бібліотеки Prophet [26]:

- прогнозування майбутніх значень часових рядів;
- виявлення сезонності, трендів та інших закономірностей у даних;
- візуалізація часових рядів та прогнозів;
- обчислення метрик точності прогнозів;
- використання попередніх значень для покращення прогнозів.

Погляньмо на результат прогнозування Індексу сприйняття корупції, де вищий бал вказує на меншу корупцію, за 1995, 1996 та 1997 роки (див рис. 2.1).

Year	Corruption perceptions
1995	18.626717
1996	20.148450
1997	15.787242

Рисунок 2.1 – Прогноз для Індексу сприйняття корупції за 1995-1997 роки

Джерело: розроблено автором за даними [1-13]

З рис. 2.1 видно, прогнозовані результати свідчать про значний рівень корупції в Україні за досліджуваний період. З отриманими результатами можемо приступати до подальшого дослідження.

2.2 Розвідувальний аналіз інформаційної бази для побудови регресійної моделі

Почнімо з візуалізацій для кращого розуміння вмісту набору даних (див. рис. 2.2):

Рисунок 2.2 – Візуалізація зміни Індексу розвитку людини в Україні

Джерело: розроблено автором за даними [1-13]

З рис. 2.2 можна побачити, що Індекс розвитку людини вийшов на певне плато у 2012 році і кожна наступна криза – АТО, Covid-19, Повномасштабне вторгнення 24-го лютого 2022-го року – спричиняла різке його падіння.

Для оцінювання впливу визначення індексу людського розвитку іншими показниками була побудована множинна регресійна модель. Спершу на базі наведених даних (Додаток А.1) проведемо їхній попередній аналіз для запобігання великих значень кореляції між незалежними змінними та високого рівня автокореляції.

Кореляція в широкому сенсі – це міра зв’язку між змінними. У корельованих даних зміна величини однієї змінної пов’язана зі зміною величини іншої змінної або в тому ж (позитивна кореляція), або в протилежному (негативна кореляція) напрямку.

Найчастіше термін кореляція використовується в контексті лінійного зв’язку між двома неперервними змінними і виражається як коефіцієнт кореляції Пірсона (Pearson product-moment correlation). Коефіцієнт кореляції Пірсона зазвичай використовується для нормально розподілених даних. Для ненормально розподілених безперервних даних, для порядкових даних або для даних з суттєвими викидами можна використовувати рангову кореляцію Спірмена як міру монотонного зв’язку.

Обидва коефіцієнти кореляції масштабуються таким чином, що вони знаходяться в діапазоні від -1 до +1, де 0 вказує на відсутність лінійного або монотонного зв’язку, а зв’язок стає сильнішим і в кінцевому підсумку наближається до прямої лінії (кореляція Пірсона) або постійно зростаючої або спадаючої кривої (кореляція Спірмена) в міру того, як коефіцієнт наближається до абсолютного значення 1.

Тести гіпотез і довірчі інтервали можуть бути використані для перевірки статистичної значущості результатів і для оцінки сили зв’язку в сукупності, з якої були відібрані дані [27].

Побудуймо та проаналізуймо матрицю парних коефіцієнтів кореляції (див. рис. 2.3):

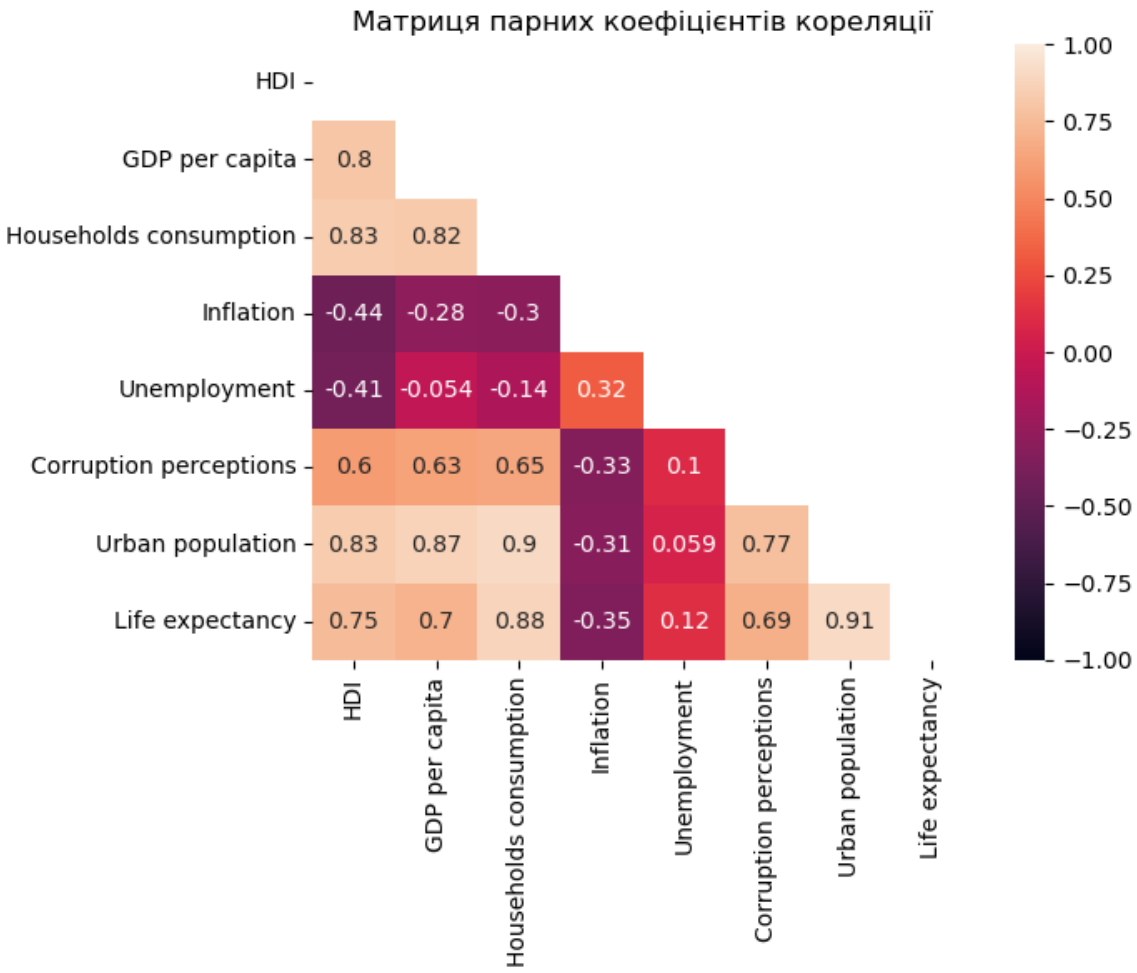

Рисунок 2.3 – Матриця парних коефіцієнтів кореляції

Джерело: розроблено автором за даними [1-13]

У результаті аналізу даних, наведених на рис. 2.3, можна зробити висновок про те, що кілька ознак мають тісний зв’язок між собою (коефіцієнт кореляції більше за 0.5). Прикладом таких пар можуть послугувати:

- ВВП на душу населення та Кінцеві споживчі витрати домогосподарств
- ВВП на душу населення та очікувана тривалість життя
- Індекс сприйняття корупції та Відсоток міського населення
- та інші

Це може свідчити про наявність мультиколінеарності у датасеті. Мультиколінеарність означає високий ступінь лінійної взаємозалежності між пояснювальними змінними в моделі множинної регресії і призводить до некоректних результатів регресійного аналізу. Інструментами діагностики мультиколінеарності є фактор інфляції дисперсії (variance inflation factor, VIF), індекс умов та число умов, а також частка декомпозиції дисперсії (variance decomposition proportion, VDP). Мультиколінеарність може бути виражена коефіцієнтом детермінації множинної регресійної моделі, в якій одна пояснювальна змінна є змінною відгуку моделі, а інші – її пояснювальними змінними. Загалом, мультиколінеарність може призвести до ширших довірчих інтервалів, які дають менш надійні ймовірності з точки зору впливу незалежних змінних у моделі. Хоча мультиколінеарність не впливає на оцінки регресії, вона робить їх нечіткими, неточними та ненадійними. Таким чином, може бути важко визначити, як незалежні змінні впливають на залежну змінну окремо. Це призводить до завищення стандартних похибок деяких або всіх коефіцієнтів регресії. Мультиколінеарність присутня, коли VIF перевищує 5 [28, 29]. Розрахуємо VIF-критерій для пояснювальних змінних нашої моделі (див. рис. 2.4).

Variable	VIF
GDP per capita	5.896305
Households consumption	10.652110
Inflation	1.575362
Unemployment	1.940329
Corruption perceptions	2.733268
Urban population	18.955260
Life expectancy	11.302849

Рисунок 2.4 – VIF-критерії для кожної пояснювальної змінної

Джерело: розроблено автором за даними [1-13]

З рис 2.4 видно, що мультиколінеарність присутня для змінних ВВП на душу населення, споживання домогосподарств, відсоток міського населення, очікувана тривалість життя.

Один з найпоширеніших способів усунення проблеми мультиколінеарності полягає в тому, щоб спочатку визначити колінеарні незалежні предиктори, а потім вилучити один або декілька з них. Як правило, в статистиці для визначення ступеня мультиколінеарності розраховують VIF-критерій. Альтернативним методом усунення мультиколінеарності є збір більшої кількості даних [29, 30].

У цій роботі був застосований метод, згідно з яким один за одним видаляються предиктори з найбільшим значенням критерію. Через видалення одного такого предиктора з дослідження, значення критеріїв інших пояснювальних змінних зменшаться, тому одразу видаляти всі предиктори зі значенням критерію більше 5, є недоцільним.

Послідовне видалення пояснювальних змінних з найбільшими значеннями vif-критерію (див. рис. 2.5-2.6).

Variable	VIF
GDP per capita	3.425642
Households consumption	10.325076
Inflation	1.555711
Unemployment	1.935036
Corruption perceptions	2.137255
Life expectancy	7.868960

Рисунок 2.5 – Таблиця значень VIF-критеріїв після видалення ознаки «Відсоток міського населення»

Джерело: розроблено автором за даними [1-13]

Variable	VIF
GDP per capita	2.271566
Inflation	1.388878
Unemployment	1.271676
Corruption perceptions	2.134454
Life expectancy	2.704048

Рисунок 2.6 – Таблиця значень VIF-критеріїв після видалення ознаки «Кінцеві споживчі витрати домогосподарств та неприбуткових установ, що допомагають домогосподарствам»

Джерело: розроблено автором за даними [1-13]

З рис. 2.5-2.6 можемо зробити висновки, що після проведення аналізу видалення двох ознак, а саме: «Відсоток міського населення» та «Кінцеві споживчі витрати домогосподарств та неприбуткових установ, що допомагають домогосподарствам,» є достатнім для позбавлення від мультиколінеарності.

Далі потрібно приступити до розбиття набору даних на кілька частин для побудови регресійної моделі. По-перше, було відділено залежну та незалежні змінні та записано у різні масиви. По-друге, було проведено розбиття датасету на тренувальну(навчальну) та тестову вибірки. Оцінка навичок моделі на навчальній вибірці призвела б до зміщеної оцінки. Тому модель оцінюється на збереженій вибірці, щоб отримати незміщену оцінку навичок моделі. Це зазвичай називається підходом до оцінки алгоритму з розділенням тренувань і тестів (train-test approach) [31].

«Припустимо, що ми хочемо оцінити похибку тесту, пов’язану з пристосуванням певного статистичного методу навчання до набору спостережень. Підхід валідаційної множини (...) є дуже простою стратегією для цього завдання. Він передбачає випадковий розподіл наявного набору спостережень на дві частини, навчальну та валідаційну. Модель припасовується на навчальній множині, і

припасована модель використовується для прогнозування спостережень у валідаційній множині. Отримана в результаті валідаційної вибірки частота помилок – зазвичай оцінюється за допомогою MSE у випадку кількісної відповіді – дає оцінку частоти помилок тесту»[32].

2.3 Побудова моделі множинної регресії

Побудова відбувалась за допомогою бібліотеки scikit-learn для мови програмування Python, а саме з використанням функції LinearRegression().

Scikit-learn демонструє широкий спектр алгоритмів машинного навчання, як керованих, так і некерованих, використовуючи послідовний, орієнтований на завдання інтерфейс, що дозволяє легко порівнювати методи для конкретного застосування. Оскільки він спирається на наукову екосистему Python, його можна легко інтегрувати в додатки, що виходять за рамки традиційного аналізу статистичних даних. Важливо, що алгоритми, реалізовані мовою високого рівня, можна використовувати як будівельні блоки для підходів, специфічних для конкретних випадків використання [33].

Функція LinearRegression() підбирає лінійну модель з коефіцієнтами w = (w1, ..., wp) для мінімізації залишкової суми квадратів між спостережуваними цілями в наборі даних і цілями, передбаченими лінійною апроксимацією [34].

Після побудови моделі переглянемо оцінки її точності (див. рис. 2.7).

```
R^2 train: 0.926
R^2 test: 0.785
Root Mean Squared Error train: 0.009
Root Mean Squared Error test: 0.014
```

Рисунок 2.7 – Оцінки точності побудованої моделі

Джерело: розроблено автором за даними [1-13]

З рис. 2.7 можемо зробити такі висновки:

- R^2 (коефіцієнт детермінації) на тренувальній вибірці: 0.926, що означає, що модель пояснює 92.6% варіації в даних на тренувальній вибірці. Це дуже високий показник, який свідчить про те, що модель добре підходить для навчальних даних;
- R^2 на тестовій вибірці: 0.785, тобто на тестовій вибірці модель пояснює 78.5% варіації в даних. Це все ще досить хороший показник, але помітно нижчий, ніж на тренувальній вибірці. Це може свідчити про певну «перенавченість» (overfitting) моделі;
- Root Mean Squared Error (RMSE) на тренувальній вибірці: 0.009 – низьке значення RMSE на тренувальній вибірці означає, що середня помилка передбачення на тренувальних даних дуже мала. Це ще раз підтверджує, що модель добре навчається на тренувальних даних;
- RMSE на тестовій вибірці: 0.014, хоч значення RMSE на тестовій вибірці трохи вище, ніж на тренувальній, воно все ще залишається досить низьким, що вказує на те, що модель добре передбачає нові дані. Проте різниця між RMSE на тренувальній і тестовій вибірках підтверджує наявність певної «перенавченості».

Переглянемо коефіцієнти побудованої моделі (див. рис. 2.8).

GDP per capita	5.396559e-06
Inflation	8.537243e-07
Unemployment	-3.994517e-03
Corruption perceptions	1.467431e-03
Life expectancy	8.690700e-03

Рисунок 2.8 – Коефіцієнти моделі множинної регресії

Джерело: розроблено автором за даними [1-13]

Як це видно з рис. 2.8, коефіцієнти моделі множинної регресії показують, як зміна кожної незалежної змінної впливає на залежну змінну, при фіксованих значеннях інших змінних. Кожен коефіцієнт вказує на величину та напрямок зміни залежної змінної при зміні відповідної незалежної змінної на одиницю, за умови, що всі інші незалежні змінні залишаються незмінними. Розглянемо детальніше кожен з коефіцієнтів у контексті цієї моделі:

- GDP per capita (ВВП на душу населення): 0.000005 – коефіцієнт показує, що зі збільшенням ВВП на душу населення на 1 долар, Індекс людського розвитку зростає на 0.000005 одиниць, за інших рівних умов. Оскільки коефіцієнт дуже малий, вплив кожного додаткового долара ВВП на душу населення є незначним;
- Inflation (Інфляція): 0.00000085 – цей коефіцієнт вказує, що зі збільшенням інфляції на 1%, Індекс людського розвитку збільшується на 0.00000085 одиниць, за інших рівних умов. Вплив інфляції також дуже незначний;
- Unemployment (Рівень безробіття): -0.00399 – коефіцієнт означає, що зі збільшенням рівня безробіття на 1%, Індекс людського розвитку зменшується на 0.00399 одиниць, за інших рівних умов. Це від'ємне значення свідчить про негативний вплив безробіття на залежну змінну;
- Corruption perceptions (Індекс сприйняття корупції): 0.00146 – цей коефіцієнт показує, що зі збільшенням індексу сприйняття корупції на 1 пункт, Індекс людського розвитку зростає на 0.00146 одиниць, за інших рівних умов. Індекс сприйняття корупції вимірюється від 0 до 100, тому навіть незначні зміни можуть бути суттєвими;
- Life expectancy (Очікувана тривалість життя): 0.0087 – коефіцієнт показує, що зі збільшенням очікуваної тривалості життя на 1 рік, Індекс людського розвитку зростає на 0.0087 одиниць, за інших рівних умов. Це означає, що очікувана тривалість життя має позитивний і більш суттєвий вплив на Індекс людського розвитку порівняно з іншими факторами.

Розрахуємо критерій Стьюдента та його p-значення для кожної ознаки.

Т-критерій Стьюдента – метод перевірки гіпотез про середнє значення невеликої вибірки, взятої з нормально розподіленої генеральної сукупності, коли стандартне відхилення генеральної сукупності невідоме.

Зазвичай спочатку формулюється нульова гіпотеза, яка стверджує, що не існує ефективної різниці між спостережуваним вибірковим середнім і гіпотетичним або заявленим генеральним середнім – тобто, що будь-яка виміряна різниця пояснюється лише випадковістю. Загалом, t-тест може бути або двостороннім, який просто констатує, що середні не є еквівалентними, або одностороннім, який визначає, чи є спостережуване середнє більше або менше за гіпотетичне середнє.

Кожного разу, коли застосовується лінійна регресія, необхідно визначити, чи існує статистично значущий зв’язок між змінною-предиктором та залежною змінною. Якщо p-значення, яке відповідає t, менше певного порогового значення (наприклад, α = 0.05), нульова гіпотеза відкидається і можна зробити висновок про те, що існує статистично значущий зв’язок між предикторною змінною і змінною відгуку [35-36].

Погляньмо на результати розрахунків t-статистики та відповідного p-значення (див. рис. 2.9).

```
GDP per capita: t-статистика = 24.583671135401065, p-значення = 0.00011960636292520332
Inflation: t-статистика = 6.4566452498484885, p-значення = 0.02109707747915825
Unemployment: t-статистика = 2.2969423178880293, p-значення = 0.14800182248767527
Corruption perceptions: t-статистика = 25.734231696124322, p-значення = 9.410480727152082e-05
Life expectancy: t-статистика = 29.38366165902005, p-значення = 4.589327639682535e-05
```

Рисунок 2.9 – Результати розрахунку t-статистики та відповідного p-значення

Джерело: розроблено автором за даними [1-13]

Як це видно з рис. 2.9, для ВВП на душу населення: t-статистика: 24.584, p-значення: 0.0001.

Низьке p-значення (набагато менше 0.05) свідчить про те, що можна з великою впевненістю відкинути нульову гіпотезу і стверджувати, що ВВП на душу населення має значний вплив на залежну змінну.

Рівень інфляції: t-статистика: 6.457, p-значення: 0.021, тобто інфляція також є значущим предиктором у моделі.

Рівень безробіття: t-статистика: 2.297, p-значення: 0.148, тож для рівня безробіття t-статистика є значно меншою, а p-значення перевищує 0.05. Це означає, що немає достатніх доказів для відхилення нульової гіпотези, тобто рівень безробіття не є статистично значущим предиктором у цій моделі.

Індекс сприйняття корупції: t-статистика: 25.734, p-значення: 0.00009, тобто Індекс сприйняття корупції є дуже значущим предиктором, про що свідчать як висока t-статистика, так і дуже низьке p-значення.

Очікувана тривалість життя: t-статистика: 29.384, p-значення: 0.0004, тож Очікувана тривалість життя також є дуже значущим предиктором.

Тож з рис. 2.9 можемо зробити певні висновки.

Отримані результати вказують на важливість різноманітних аспектів економіки, соціуму та політики для розуміння та передбачення динаміки Індексу людського розвитку у отриманій моделі множинної регресії. Високі значення t-статистики та низькі p-значення для таких показників, як ВВП на душу населення, рівень інфляції, індекс сприйняття корупції та очікувана тривалість життя, демонструють суттєвий вплив цих факторів на залежну змінну. Це свідчить про те, що економічний, соціальний та політичний контекст сильно впливають на рівень та тенденції розвитку суспільства.

Ці висновки будуть корисними для формулювання та впровадження ефективних стратегій економічного управління та соціальної політики, спрямованих на забезпечення сталого розвитку та підвищення якості життя населення України. Крім того, вони можуть слугувати основою для подальших досліджень у галузі економіки, соціології та політики з метою глибшого розуміння взаємозв’язків та механізмів функціонування сучасного суспільства.

РОЗДІЛ 3
КЛАСТЕРНИЙ АНАЛІЗ ЯК ІНСТРУМЕНТ СИСТЕМАТИЗАЦІЙ КРАЇН ЗА РІВНЕМ ЛЮДСЬКОГО РОЗВИТКУ ТА ІНШИМИ МАКРОЕКОНОМІЧНИМИ ПОКАЗНИКАМИ

3.1 Інформація про набір даних та інструменти аналізу

Датасет (Додаток Б.1) оформлений у файлі зі значеннями, розділеними комою (.csv) і називається Countries_cluster.

У цьому наборі даних міститься значення економічних і соціальних показників 32-х європейської країни за 2022 рік. Всього є 9 числових ознак, а саме: Індекс людського розвитку (англ. Human development index (HDI)), ВВП на душу населення у сучасних доларах (англ. GDP per capita, current dollars), інфляція (англ. Inflation rate), рівень безробіття (англ. Uneployment rate), індекс сприйняття корупції (англ. Corruption perceptions), Індекс іновацій (англ. Innovation Index), Індекс прав власності (англ. Property rights Index), Відсоток міського населення (англ. Percent urban population) та очікувана тривалість життя (англ. Life expectancy). Країни, що присутні в датасеті: Австрія, Бельгія, Боснія і Герцеговина, Болгарія, Хорватія, Чехія, Данія, Естонія, Фінляндія, Франція, Німеччина, Греція, Угорщина, Ісландія, Ірландія, Італія, Латвія, Литва, Нідерланди, Північна Македонія, Норвегія, Польща, Португалія, Румунія, Сербія, Словаччина, Словенія, Іспанія, Швеція, Швейцарія, Україна, Велика Британія.

Кластеризація – це некерований розподіл шаблонів (спостережень, елементів даних або векторів ознак) за групами (кластерами). Завдання кластеризації розглядалося в багатьох контекстах, різними дослідниками в багатьох дисциплінах; це відображає його широку привабливість і корисність як одного з етапів дослідницького аналізу даних [37].

Для розбиття набору даних на кластери, був використаний метод k-середніх.

Основна ідея кластеризації за методом k-середніх полягає в побудові кластерів таким чином, щоб загальна варіація всередині кластера була мінімальною. Для цього існує декілька алгоритмів K-середніх. Стандартним алгоритмом є алгоритм Хартігана-Вонга, який визначає загальну варіацію всередині кластера як суму евклідових відстаней між значеннями ознаки спостереження і та відповідним центроїдом

Алгоритм K-середніх є одним з найпопулярніших методів кластеризації ітеративного пошуку і є найпоширенішим некерованим алгоритмом машинного навчання для розбиття заданого набору даних на k груп (тобто k кластерів), де k – це кількість груп, попередньо визначена аналітиком. Він класифікує об’єкти в декілька груп (тобто кластерів) таким чином, що об’єкти в межах одного кластера є максимально схожими (тобто висока внутрішньокласова схожість), тоді як об’єкти з різних кластерів є максимально несхожими (тобто низька міжкласова схожість). При кластеризації за методом k-середніх кожен кластер представлений своїм центром (тобто центроїдом), який відповідає середньому значенню балів, віднесених до кластеру [38, 39].

3.2 Розвідувальний аналіз інформаційної бази для побудови моделі кластеризації

Для більшого розуміння змісту датасету, був реалізований ряд візуалізацій. Перша з них – теплова мапа кореляцій (рис 3.1):

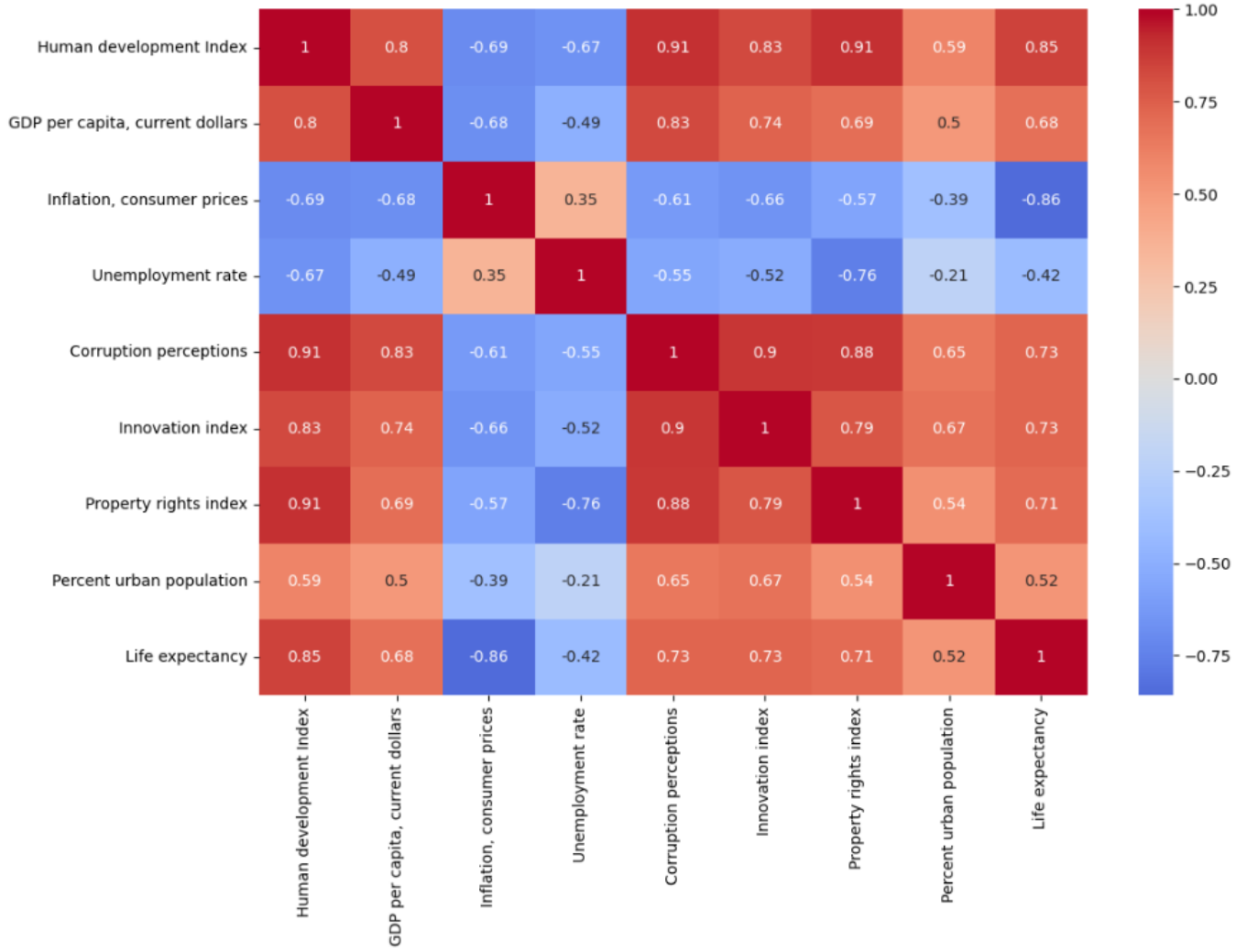

Рисунок 3.1 – Теплова мапа кореляцій

Джерело: розроблено автором за даними [1-13]

З рис 3.1, можна зробити висновки, що на тепловій карті можна побачити наступні кореляції:

- сильна позитивна кореляція між ІЛР та сприйняттям корупції;
- сильна позитивна кореляція між ІЛР та індексом інновацій;
- сильна позитивна кореляція між ІЛР та індексом прав власності;
- сильна позитивна кореляція між ІЛР та тривалістю життя;
- сильна негативна кореляція між ІЛР та інфляцією;
- сильна негативна кореляція між ІЛР та рівнем безробіття.

Ці кореляції дозволяють зрозуміти те, як різні фактори впливають одне на одне. Наприклад, високий ІЛР, як правило, пов’язаний з низьким рівнем корупції, високим рівнем інновацій, чітко визначеними та захищеними правами власності, більш тривалим життям, низьким рівнем інфляції та низьким рівнем безробіття.

Важливо зазначити, що кореляція не дорівнює причинно-наслідковому зв’язку. Тільки тому, що дві змінні корелюють, не означає, що одна з них спричинює іншу. Наприклад, можливо, що високий ІЛР та низький рівень корупції спричинені третьою змінною, такою як сильна демократія.

Однак кореляції можуть бути корисними для виявлення потенційних причинно-наслідкових зв’язків, які можна далі дослідити.

Для вивчення характерних для Європи взаємозв’язків між двома окремими показниками, були побудовані точкова діаграма взаємозв’язку ІЛР та ВВП на душу населення (див. рис. 3.2).

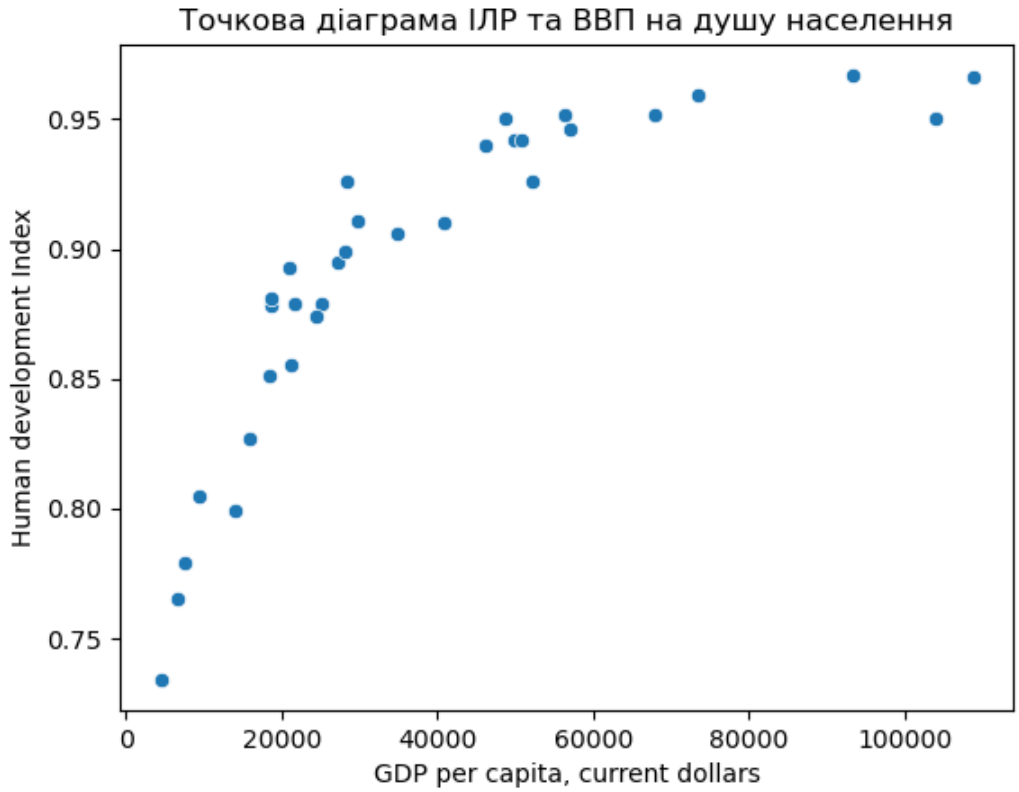

Рисунок 3.2 – Точкова діаграма взаємозв’язку ІЛР та ВВП на душу населення

Джерело: розроблено автором за даними [1-13]

З рис. 3.2, можна зробити висновок, що між ІЛР та ВВП на душу населення в Європі існує позитивний зв'язок. Це означає, що країни з більш високим рівнем доходу, як правило, мають кращі показники за такими параметрами, як очікувана тривалість життя, рівень грамотності та доступ до освіти. На рівні ВВП на душу населення близько 30000 тис. дол. помітна «точка перелому», коли зростання ВВП на душу населення слабше пов'язано зі зростанням Індексу людського розвитку, ніж до цієї точки. Важливо зазначити, що зв'язок не є абсолютно лінійним. Це означає, що зростання ВВП на душу населення не завжди призводить до пропорційного зростання ІЛР. Крім того, не можна однозначно стверджувати, що ВВП на душу населення є причиною високого ІЛР.

Далі для вивчення характерних для Європи взаємозв'язків між двома окремими показниками, пропонуємо побудувати точкову діаграму взаємозв'язку ІЛР та Рівня інфляції (див. рис. 3.3).

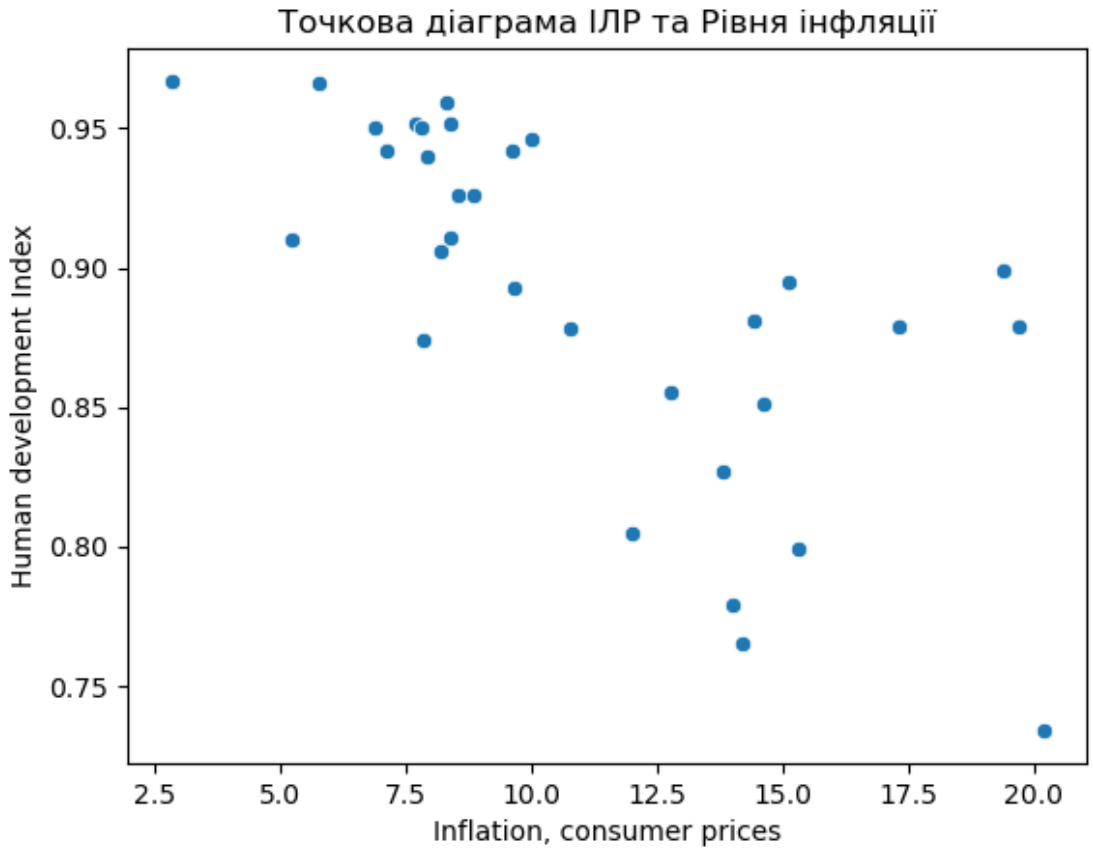

Рисунок 3.3 – Точкова діаграма взаємозв'язку ІЛР та Рівня інфляції

Джерело: розроблено автором за даними [1-13]

З рис. 3.3, можна зробити висновок, що між ІЛР та рівнем інфляції в Європі існує негативний зв'язок. Це означає, що країни з високим рівнем інфляції, як правило, мають гірші показники ІЛР. Спостерігається певна дифузія точок, що свідчить про те, що зв'язок між ІЛР та рівнем інфляції скоріше не є лінійним.

Далі побудуймо та проаналізуймо точкову діаграму взаємозв'язку ВВП на душу населення та Індексу сприйняття корупції (див. рис. 3.4).

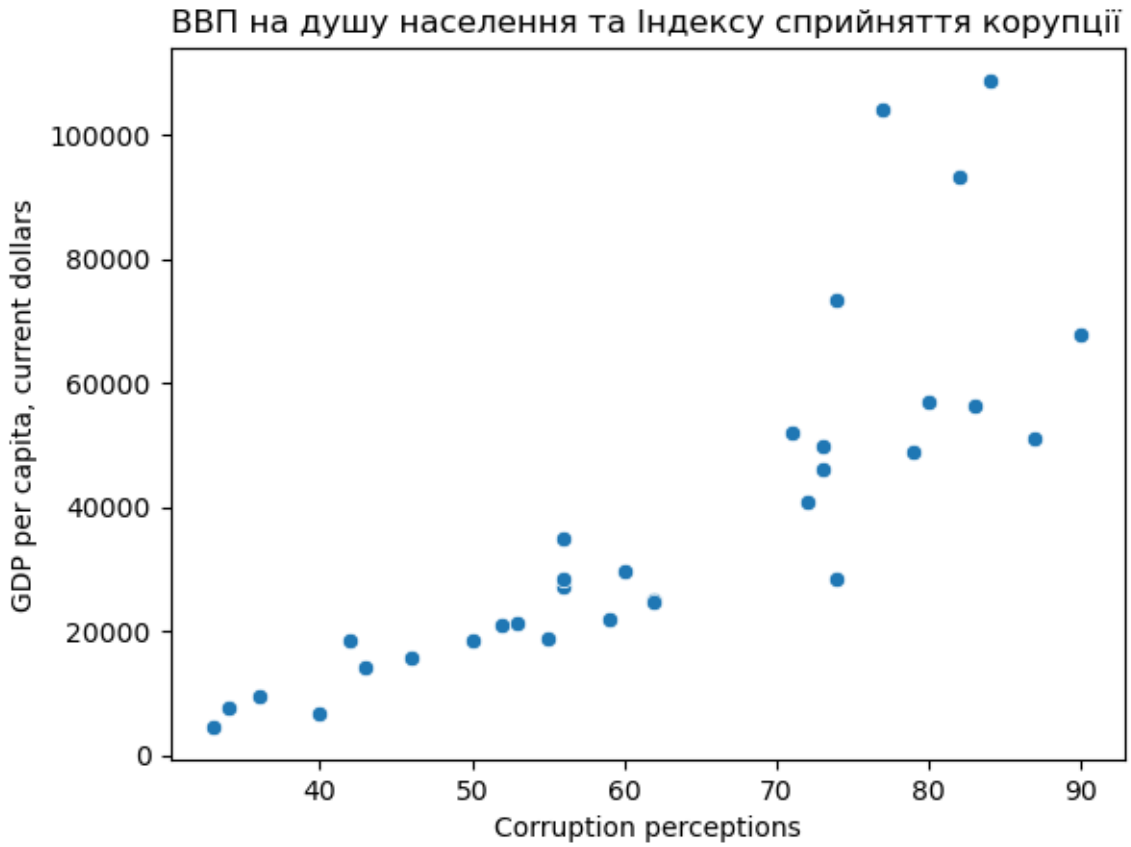

Рисунок 3.4 – Точкова діаграма взаємозв'язку ВВП на душу населення та Індексу сприйняття корупції

Джерело: розроблено автором за даними [1-13]

З рис. 3.4 можна зробити висновок, що між ВВП на душу населення та Індексом сприйняття корупції в Європі існує позитивний зв'язок. Це означає, що країни з більш високим рівнем доходу, як правило, сприймаються як менш корумповані. На діаграмі відсутня чітко виражена лінія тренду, що може бути пов'язано з нелінійним характером зв'язку між ВВП на душу населення та Індексом сприйняття корупції. З діаграми можна зробити висновок, що між ВВП на душу

населення та Індексом сприйняття корупції в Європі існує позитивний зв’язок, що означає, що країни з більш високим рівнем доходу, як правило, сприймаються як менш корумповані.

Хоча на початку діаграми можна прослідкувати певний лінійний зв’язок між показниками, після значення Індексу сприйняття корупції у 60, дисперсія спостережень значно збільшується і будь-яку залежність прослідкувати важко.

Далі побудуймо точкову діаграму взаємозв’язку Рівня інфляції та Рівня безробіття (див. рис. 3.5).

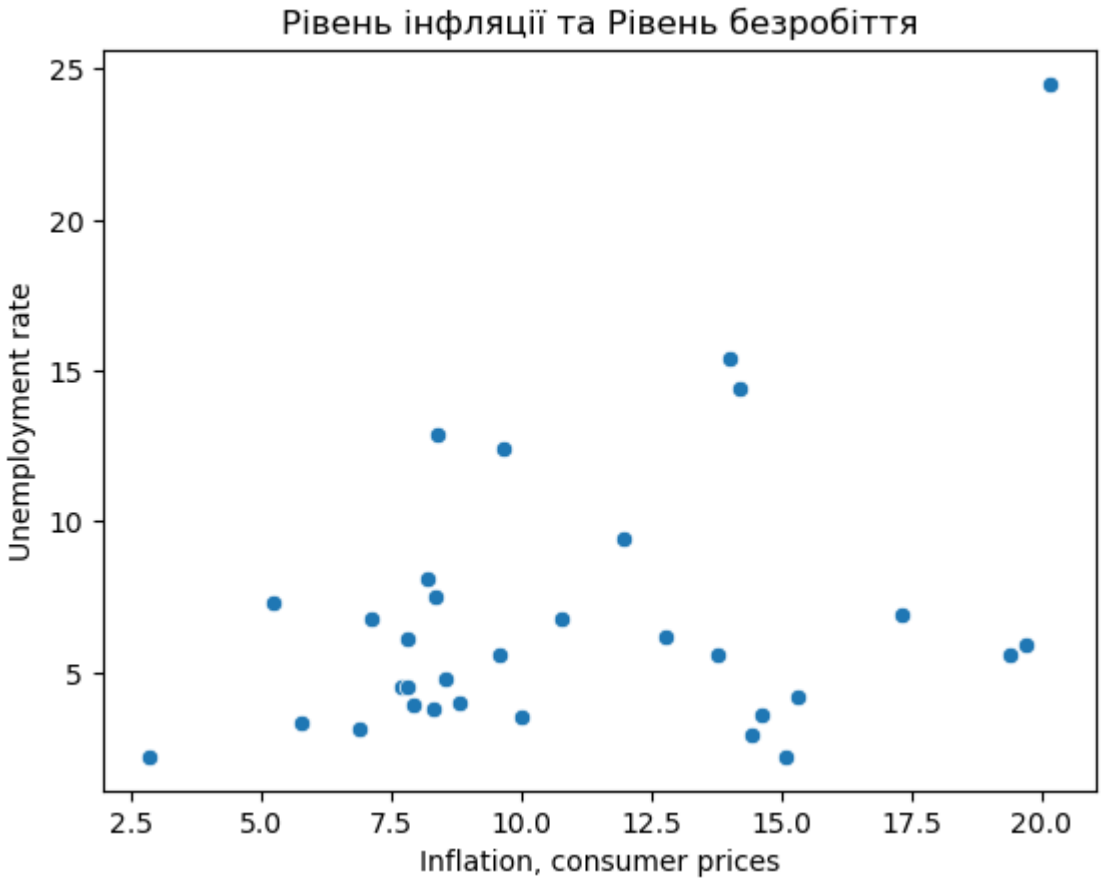

Рисунок 3.5 – Точкова діаграма взаємозв’язку Рівня інфляції та Рівня безробіття

Джерело: розроблено автором за даними [1-13]

З рис. 3.5 можна зробити висновок, що в країнах Європи майже немає залежності між рівнем інфляції та рівнем безробіття. Більшість точок на діаграмі сконцентровані в лівому нижніх квадрантах, що підтверджує відсутність таких тенденцій.

3.3 Побудова моделі кластеризації

Для побудови якісної моделі кластеризації за методом *k*-середніх потрібно спочатку визначитись із бажаною кількістю кластерів. Для цього потрібно скористатись ієрархічною кластеризацією та її візуалізацією у вигляді дендрограми. Щоб цього досягнути, були використані дві функції з бібліотеки sciру – linkage та dendrogram відповідно. Функція linkage підтримує розрахунок відстаней між спостереженнями за більшістю популярних методів(метод найближчого сусіда, метод найдальшого сусіда, метод середнього (між групами) зв'язку, метод Варда). Для цього дослідження був обраний метод Варда. Вертикальні лінії (гілки). На початку вони представляють окремі точки даних. По мірі просування вгору по дендрограмі ці лінії починають об'єднуватися в кластери.

Інтерпретування дендрограми ієрархічної кластеризації [40]:

• горизонтальні лінії (зв'язки) – горизонтальні лінії, що з'єднують вертикальні лінії, відображають злиття кластерів. Висота цих ліній відповідає відстані між кластерами, що об'єднуються. Чим вища лінія, тим більша відстань між кластерами, що об'єднуються, і тим більша різниця між ними;

• кластери – листки дендрограми – це окремі точки даних. По мірі просування вгору по дендрограмі, кластери утворюються шляхом об'єднання точок даних або менших кластерів. Гілки внизу дендрограми представляють найдрібніший рівень кластеризації (окремі точки даних), тоді як гілки на вищих рівнях представляють більші кластери, утворені об'єднанням менших кластерів;

• обрізання дендрограми – для визначення кількості кластерів, ми можемо провести на дендрограмі горизонтальну лінію (яку часто називають «зріз») на певній висоті. Кількість кластерів визначається кількістю перетинів горизонтальної лінії з вертикальними лініями. Кожна точка перетину відповідає кластеру.

Погляньмо на побудовану дендрограму ієрархічної кластеризації за методом Варда (див. рис. 3.6).

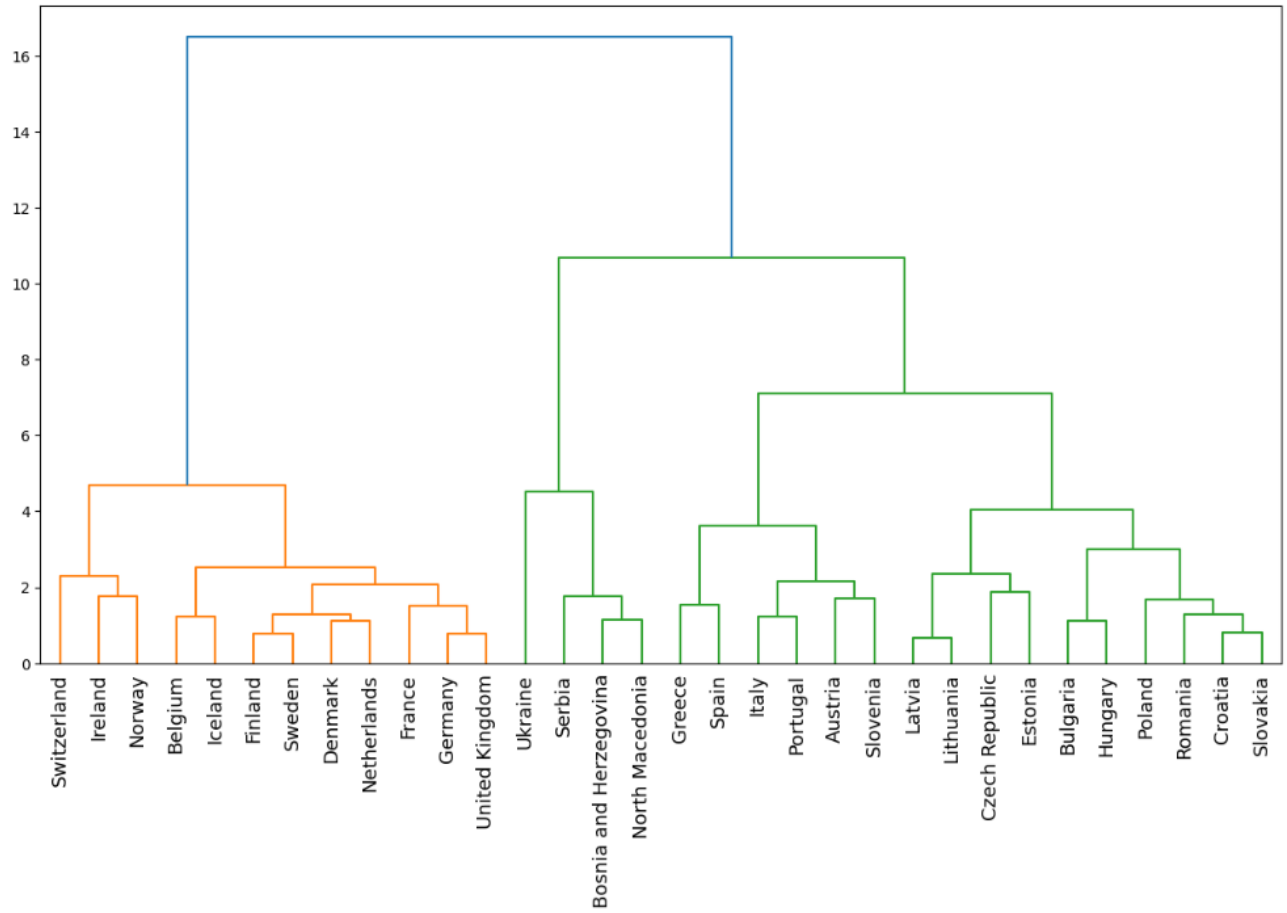

Рисунок 3.6 – Дендрограма ієрархічної кластеризації, побудованої за методом Варда

Джерело: розроблено автором за даними [1-13]

Після проведення візуального аналізу рис. 3.6 було прийнято рішення про побудову моделі кластеризації за допомогою методу k-середніх на основі 4 кластерів. Навіть до розподілу можна помітити, що країни вже згруповані певним чином: з лівого боку дендрограми можна побачити групу з найбагатших країн Західної Європи, посередині є група найбідніших країн Європи, з Україною у своєму складі. Також видно кластер бідніших країн Західної Європи та кластер країн Східної Європи. Поглянемо на лінію відтинання, зображену на дендрограмі (див. рис. 3.7):

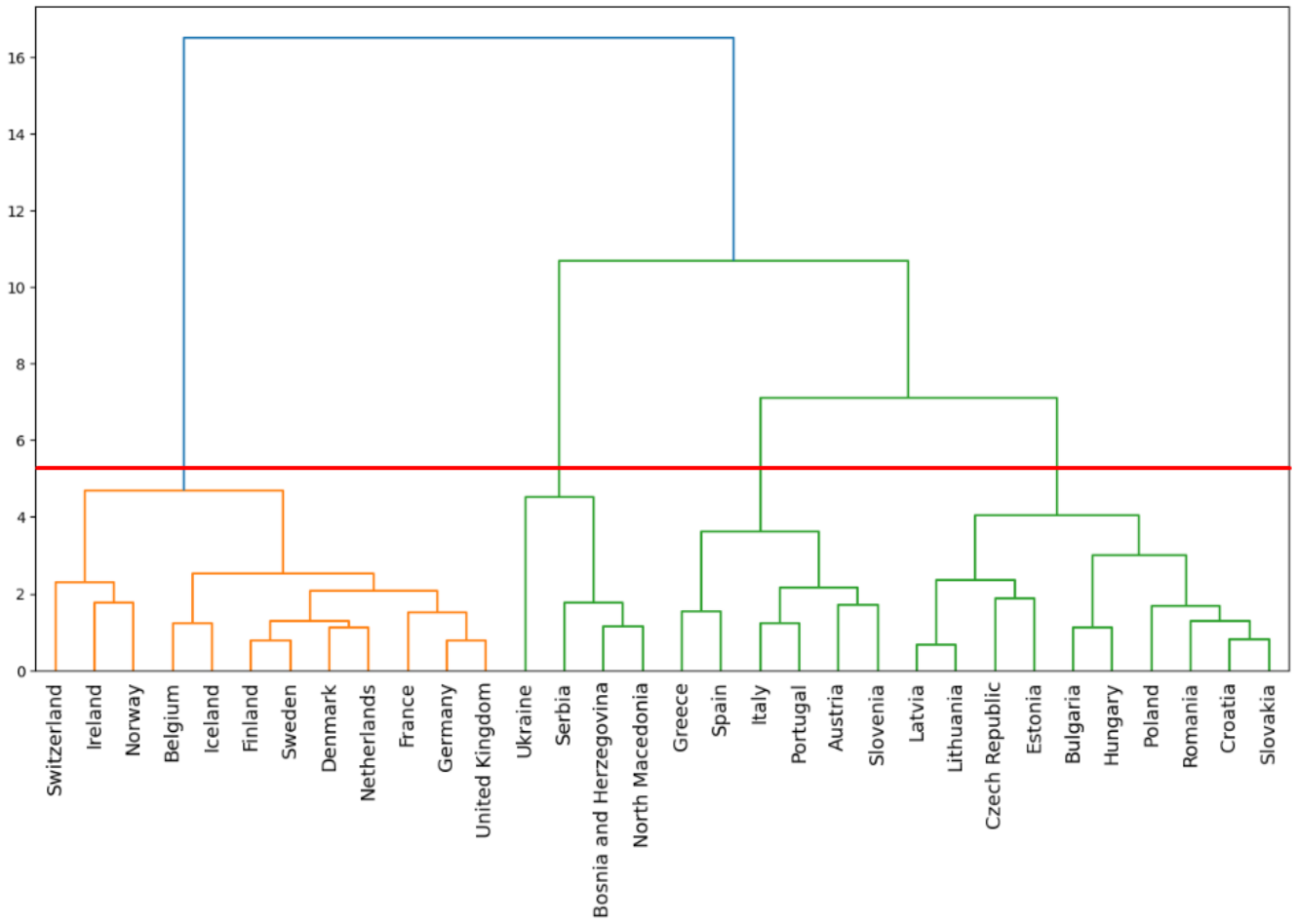

Рисунок 3.7 – Лінія відтинання на дендрограмі

Джерело: розроблено автором за даними [1-13]

Як це видно на рис. 3.7, червона лінія на дендрограмі демонструє місце зрізу. Кожна горизонтальна лінія на дендрограмі позначає кластер, висота якого вказує на рівень близькості кластерів. Точка зрізу дендрограми – це рівень подібності, на якому дендрограма розрізається для отримання розбиття сутностей. Кожна точка зрізу дає розбиття на кластери, які дають рекомендації щодо реструктуризації модуля. Вибір відповідних точок зрізу завжди був культовою проблемою, що включає в себе.

У мові програмування Python застосування методу k-середніх може бути реалізовано за допомогою функції KMeans з бібліотеки scikit-learn. Погляньмо на результат застосування моделі кластеризації (див. рис. 3.8).

	Human development Index	GDP per capita, current dollars	Inflation, consumer prices	Unemployment rate	Corruption perceptions - Transparency International	Innovation index	Property rights index	Percent urban population	Life expectancy	Cluster
United Kingdom	0.940	46125.255750	7.922049	3.9	73	59.7	96	84.398	81.77	0
Norway	0.966	108729.186900	5.764123	3.3	84	48.8	100	83.664	82.94	0
Netherlands	0.946	57025.012460	10.001208	3.5	80	58.0	96	92.886	82.78	0
Ireland	0.950	103983.291300	7.807375	4.5	77	48.5	93	64.183	82.81	0
Iceland	0.959	73466.778670	8.308755	3.8	74	49.5	97	93.992	83.52	0
Germany	0.950	48717.991140	6.872574	3.1	79	57.2	96	77.648	81.88	0
France	0.910	40886.253270	5.222367	7.3	72	55.0	94	81.509	83.13	0
Finland	0.942	50871.930450	7.123508	6.8	87	56.9	100	85.681	82.48	0
Sweden	0.952	56424.284700	8.369291	7.5	83	61.6	97	88.492	83.33	0
Denmark	0.952	67790.053990	7.696567	4.5	90	55.9	99	88.367	81.40	0
Switzerland	0.967	93259.905720	2.835028	2.2	82	64.6	95	74.092	84.25	0
Belgium	0.942	49926.825430	9.597512	5.6	73	46.9	93	98.153	82.17	0
Czech Republic	0.895	27226.615640	15.100165	2.2	56	42.8	89	74.377	79.85	1
Slovakia	0.855	21256.808430	12.774146	6.2	53	34.3	83	53.909	78.00	1
Hungary	0.851	18390.185000	14.608144	3.6	42	39.8	76	72.552	77.31	1
Croatia	0.878	18570.404000	10.780581	6.8	50	35.6	81	58.219	79.02	1
Bulgaria	0.799	13974.449250	15.325259	4.2	43	39.5	77	76.363	75.49	1
Latvia	0.879	21779.504260	17.310283	6.9	59	36.5	89	68.540	75.73	1
Lithuania	0.879	25064.808910	19.705046	5.9	62	37.4	89	68.465	76.41	1
Poland	0.881	18688.004490	14.429451	2.9	55	37.5	72	60.134	79.27	1
Estonia	0.899	28247.095990	19.398263	5.6	74	50.2	92	69.609	79.18	1
Romania	0.827	15786.801740	13.795489	5.6	46	34.1	81	54.489	76.50	1
Spain	0.911	29674.544290	8.390576	12.9	60	44.6	88	81.304	83.99	2
Slovenia	0.926	28439.334100	8.833699	4.0	56	40.6	90	55.751	81.85	2
Austria	0.926	52084.681200	8.546870	4.8	71	50.2	98	59.256	82.05	2
Portugal	0.874	24515.265850	7.832691	6.1	62	42.1	90	67.381	82.65	2
Greece	0.893	20867.269090	9.645260	12.4	52	34.5	76	80.357	82.80	2
Italy	0.906	34776.423230	8.201290	8.1	56	46.1	82	71.657	84.01	2
North Macedonia	0.765	6591.471314	14.204717	14.4	40	28.8	57	59.118	76.26	3
Ukraine	0.734	4533.975586	20.183637	24.5	33	31.0	40	69.919	72.50	3
Bosnia and Herzegovina	0.779	7568.798480	14.000000	15.4	34	28.5	49	49.841	77.93	3
Serbia	0.805	9537.682867	11.981512	9.4	36	32.3	59	56.873	76.47	3

Рисунок 3.8 – Результат застосування моделі кластеризації

Джерело: розроблено автором за даними [1-13]

З рис. 3.8 можемо спостерігати, які країни складають певний кластер. Видно, що розміри кластерів розподілені досить нерівномірно. Погляньмо на кількість країн (спостережень), що потрапили до кожного кластера (див. рис. 3.9).

```
Cluster
0          12
1          10
2           6
3           4
dtype: int64
```

Рисунок 3.9 – Кількість спостережень у кожному кластері

Джерело: розроблено автором за даними [1-13]

З рис 3.9 можна зробити такі висновки щодо розподілу кластерного аналізу, де кількість країн у кожному кластері представлена таким чином:

Кластери мають різну кількість країн, що вказує на нерівномірний розподіл даних. Найбільший кластер (кластер 1) містить 12 країн, тоді як найменший (кластер 4) — лише 4 країни, включаючи Україну. Це може свідчити про те, що деякі групи країн мають більш схожі характеристики, які об’єднують їх у більші кластери. Кластер з 12 країнами (кластер 1) може мати більш гомогенні характеристики, які об’єднують ці країни разом. У той же час, кластер з 4 країнами (кластер 4) має більш специфічні або відмінні риси, які не так часто зустрічаються серед інших країн. Наявність кластерів різного розміру може вказувати на те, що дані, які використовувалися для кластерного аналізу, мають значну диверсифікацію. Деякі країни мають характеристики, які сильно відрізняються від інших, що призводить до формування невеликих кластерів. Кластер 4, який має лише 4 країни, може вказувати на аномальні або унікальні характеристики цих країн.

Для повного розуміння характеристик отриманих кластерів, погляньмо на графік середніх значень для кожного кластеру (див. рис. 3.10).

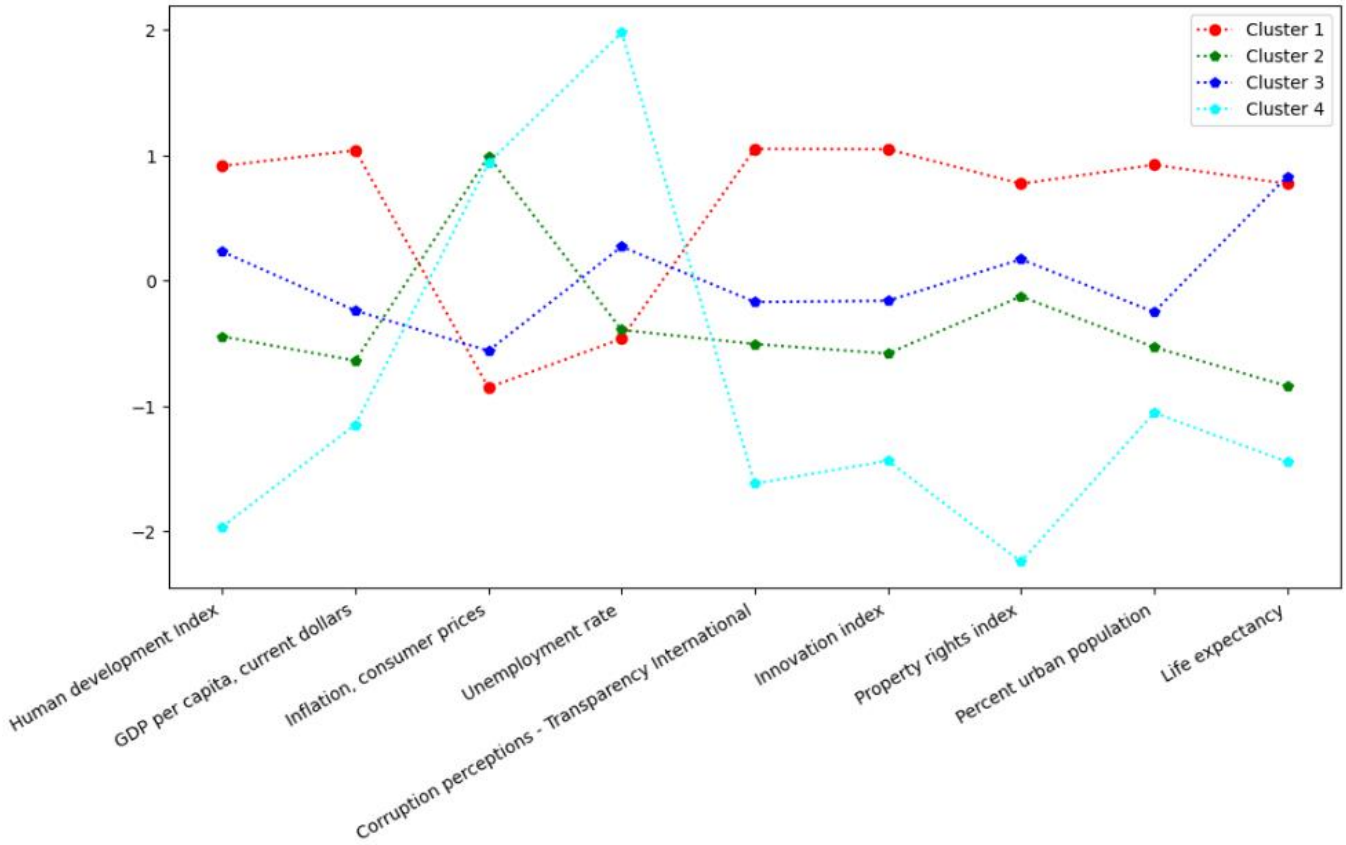

Рисунок 3.10 – Графік середніх значень для кожного кластеру

Джерело: розроблено автором за даними [1-13]

Як це видно на рис. 3.10, середні значення для кожного кластеру по різних індексах. Розгляньмо ключові моменти:

• Кластер 1 (червоний) містить 12 країн, включаючи такі, як Великобританія, Норвегія, Нідерланди, Ірландія, Ісландія, Німеччина, Франція, Фінляндія, Швеція, Данія, Швейцарія та Бельгія. Має високі значення для таких показників, як Індекс людського розвитку та ВВП на душу населення. Показники інфляції, рівня безробіття та корупції знаходяться на середньому рівні. Високий індекс інновацій та індекс прав власності;

• Кластер 2 (зелений) включає 10 країн: Чехія, Словаччина, Угорщина, Хорватія, Болгарія, Латвія, Литва, Польща, Естонія та Румунія. Має нижчі значення для ВВП на душу населення та індексу інновацій у порівнянні з кластером 1. Помірні значення індексу прав власності та корупції;

• Кластер 3 (блакитний) включає 6 країн: Іспанія, Словенія, Австрія, Португалія, Греція та Італія. Помірні значення індексу людського розвитку та ВВП на душу населення. Вищі значення рівня безробіття та інфляції;

• Кластер 4 (голубий) включає 4 країни: Північна Македонія, Україна, Боснія і Герцеговина та Сербія. Найнижчі значення для більшості індексів, включаючи Індекс людського розвитку та ВВП на душу населення. Високий рівень безробіття та інфляції. Найнижчі значення індексу прав власності та індексу інновацій.

Загалом, графік і дані показують, що країни з кластеру 1 мають найвищі показники розвитку та економічної стабільності, тоді як країни з кластеру 4 мають найнижчі значення для більшості показників.

Україна потрапила до кластеру 4 (голубий), який включає також Північну Македонію, Боснію і Герцеговину та Сербію. Характеристики цього кластеру такі:

- найнижчі значення для більшості індексів, включаючи Індекс людського розвитку (ІЛР) та ВВП на душу населення;
- високий рівень безробіття та інфляції;
- низькі значення індексу прав власності та індексу інновацій.

Позиція України відносно інших кластерів:

Відносно кластеру 1:

Цей кластер включає країни з високими значеннями ІЛР та ВВП на душу населення, високими індексами інновацій та прав власності, і середніми рівнями інфляції, безробіття та корупції. Україна значно відстає від країн цього кластеру за всіма показниками, особливо в економічному розвитку, інноваційній діяльності та рівні життя.

Відносно кластеру 2:

Країни цього кластеру мають нижчі значення для ВВП на душу населення та індексу інновацій у порівнянні з кластером 1, але кращі, ніж у кластері 4. Вони також мають помірні значення індексу прав власності та корупції. Україна також відстає від цих країн, хоча, різниця не така велика, як з країнами кластеру 1. Однак

все одно ці країни демонструють вищий рівень розвитку та економічної стабільності.

Відносно кластеру 3:

Цей кластер має помірні значення ІЛР та ВВП на душу населення, але вищі рівні безробіття та інфляції. Україна має гірші показники розвитку в порівнянні з цими країнами, хоча рівень безробіття та інфляції може бути подібним або навіть вищим.

Україна значно відстає за рівнем економічного розвитку, людського розвитку, інноваційної діяльності та прав власності від більшості європейських країн, особливо тих, що знаходяться в кластері 1. Також Україна разом з іншими країнами кластеру 4 має найнижчі показники рівня життя та економічної стабільності, що відображається у високих рівнях безробіття та інфляції. Україні необхідно зосередитись на покращенні ключових показників, таких як ВВП на душу населення, індекс людського розвитку, індекс прав власності та інновацій, щоб наблизитись до країн з вищими показниками розвитку.

ВИСНОВКИ

Соціально-економічний розвиток держави безпосередньо залежить від рівня людського розвитку. Головною метою розвитку є створення умов для того, щоб люди могли прожити довге і здорове життя, здобути освіту, мати гідні фінансові можливості.

У першій частині роботи представлена основна теоретична інформація, що стосується розглядуваних макроекономічних показників, їх використання у аналізі економічного стану країни, методології проведення регресійного та кластерного методів аналізу, їх вимог до даних та умови їх успішного застосування.

Регресійні моделі використовуються для опису взаємозв’язків між змінними шляхом підбору лінії до спостережуваних даних. Регресія дозволяє оцінити, як змінюється залежна змінна при зміні незалежної змінної.

Кластеризація – це процес об’єднання групи абстрактних об’єктів у класи схожих об’єктів:

- кластер об’єктів даних можна розглядати як одну групу;
- при проведенні кластерного аналізу ми спочатку розбиваємо набір даних на групи на основі схожості даних, а потім присвоюємо групам мітки;
- основна перевага кластеризації перед класифікацією полягає в тому, що вона адаптується до змін і допомагає виділити корисні ознаки, які відрізняють різні групи.

У практичній частині було застосовано регресійну модель для дослідження впливу восьми макроекономічних показників на індекс людського розвитку з ціллю виявлення економічних закономірностей, що спричиняють зростання або спад якості життя громадян України. З різних джерел були зібрані значення цих показників у діапазоні з 1995-го по 2022-й роки. Аналіз відбувався за допомогою мови програмування Python у середовищі Jupyter Notebook із застосуванням бібліотек, призначених для аналізу та візуалізації даних, таких як, Prophet, Statsmodels, numpy, pandas , matplotlib, seaborn, sklearn.

Збільшення ВВП на душу населення на 1 долар приводить до зростання Індексу людського розвитку на 0.000005 одиниць. Хоча цей вплив дуже малий, низьке p-значення (0.0001) і висока t-статистика (24.584) свідчать про статистично значущий вплив цієї змінної на залежну змінну.

Збільшення інфляції на 1% призводить до збільшення Індексу людського розвитку на 0.00000085 одиниць. Хоча вплив також дуже незначний, інфляція є статистично значущим фактором у моделі (p-значення 0.021 і t-статистика 6.457).

Збільшення рівня безробіття на 1% зменшує Індекс людського розвитку на 0.00399 одиниць. Однак, з огляду на високе p-значення (0.148) і низьку t-статистику (2.297), цей фактор не є статистично значущим предиктором у моделі.

Збільшення індексу сприйняття корупції на 1 пункт приводить до зростання Індексу людського розвитку на 0.00146 одиниць. Це свідчить про значущий позитивний вплив, підкріплений дуже низьким p-значенням (0.00009) і високою t-статистикою (25.734).

Збільшення очікуваної тривалості життя на 1 рік приводить до зростання Індексу людського розвитку на 0.0087 одиниць. Це вказує на позитивний і суттєвий вплив, підтверджений дуже низьким p-значенням (0.0004) і найвищою t-статистикою (29.384).

Отже, ВВП на душу населення, індекс сприйняття корупції та очікувана тривалість життя є дуже значущими предикторами, їхній вплив на залежну змінну є статистично значущим. Інфляція також є значущим фактором, хоча її вплив менш суттєвий порівняно з іншими змінними. Рівень безробіття не є статистично значущим предиктором у цій моделі.

У кластерному ж аналізі метою дослідження було розділення країн Європи на кластери за схожою низкою показників: Індекс людського розвитку (англ. Human development index (HDI)), ВВП на душу населення у сучасних доларах (англ. GDP per capita, current dollars), інфляція (англ. Inflation rate), рівень безробіття (англ. Uneployment rate), індекс сприйняття корупції (англ. Corruption perceptions), Індекс іновацій (англ. Innovation Index), Індекс прав власності (англ. Property rights Index), Відсоток міського населення (англ. Percent urban population) та очікувана

тривалість життя (англ. Life expectancy). Усього, зі стовпчиком з назвами країн, показників десять. Країни, що присутні в датасеті: Австрія, Бельгія, Боснія і Герцеговина, Болгарія, Хорватія, Чехія, Данія, Естонія, Фінляндія, Франція, Німеччина, Греція, Угорщина, Ісландія, Ірландія, Італія, Латвія, Литва, Нідерланди, Північна Македонія, Норвегія, Польща, Португалія, Румунія, Сербія, Словаччина, Словенія, Іспанія, Швеція, Швейцарія, Україна, Велика Британія.

Для дослідження було обрано 32 країни. Для визначення кількості кластерів був застосований ієрархічний агломеративний метод, для розділення на кластери – метод k-середніх.

У результаті розподілу було отримано чотири кластери, один з яких містить 12 найбагатших країн Західної Європи з найвищими показниками якості життя.

Другий кластер включає 10 країн, має нижчі значення для ВВП на душу населення та індексу інновацій у порівнянні з кластером 1 та помірні значення індексу прав власності та корупції. Абсолютна більшість країн цього кластеру – в минулому країни Варшавського договору, тобто, країни Східної Європи із соціалістичним минулим.

Третій кластер, що складається з 11 бідніших країни Західної Європи, країни Середземномор’я має дуже високі значення рівня безробіття, решта показників трохи краща за третій кластер. Країни ж цього третього кластеру, у якому знаходиться 8 об’єктів мають високий рівень інфляції.

Україна потрапила до кластеру 4 (голубий), який включає також Північну Македонію, Боснію і Герцеговину та Сербію. Цей кластер характеризується найнижчими значеннями для більшості індексів, включаючи Індекс людського розвитку (ІЛР) та ВВП на душу населення, низькі значення індексу прав власності та індексу інновацій та високий рівень безробіття та інфляції. Україна значно відстає за рівнем економічного розвитку, людського розвитку, інноваційної діяльності та прав власності від більшості європейських країн. Нашій державі та громадянам необхідно зосередитись на покращенні ключових показників, таких як ВВП на душу населення, індекс людського розвитку, індекс прав власності та інновацій, щоб наблизитись до країн з вищими показниками розвитку.

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Human Development Index. *Human Development Reports*. URL: https://hdr.undp.org/data-center/human-development-index#/indicies/HDI (дата звернення: 25.05.2024).

2. Indicators. *World Bank Open Data*. URL: https://data.worldbank.org/indicator (дата звернення: 25.05.2024).

3. GDP per capita (current US$). *World Bank Open Data*. URL: https://data.worldbank.org/indicator/NY.GDP.PCAP.CD (дата звернення: 25.05.2024).

4. Households and NPISHs final consumption expenditure (% of GDP). *World Bank Open Data*. URL: https://data.worldbank.org/indicator/NE.CON.PRVT.ZS (дата звернення: 25.05.2024).

5. Inflation, consumer prices (annual %). *World Bank Open Data*. URL: https://data.worldbank.org/indicator/FP.CPI.TOTL.ZG (дата звернення: 25.05.2024).

6. Urban population (% of total population). *World Bank Open Data*. URL: https://data.worldbank.org/indicator/SP.URB.TOTL.IN.ZS (дата звернення: 25.05.2024).

7. Life expectancy at birth, total (years). *World Bank Open Data*. URL: https://data.worldbank.org/indicator/SP.DYN.LE00.IN (дата звернення: 25.05.2024).

8. Unemployment rate. *International Monetary Fund*. URL: https://www.imf.org/external/datamapper/LUR@WEO/OEMDC/ADVEC/WEOWORLD (дата звернення: 25.05.2024).

9. Corruption Perceptions Index. *Transparency.org*. URL: https://www.transparency.org/en/cpi/2022 (дата звернення: 25.05.2024).

10. Life Expectancy by Country and in the World. *Worldometer*. URL: https://www.worldometers.info/demographics/life-expectancy/ (дата звернення: 25.05.2024).

11. Ligata R., Stojanović G. Indeks potrošačkih cijena u Bosni i Hercegovini Sarajevo. Consumer Price Index in Bosnia and Herzegovina 2022. *Thematic bulletin*.

2023. 35 p. URL: https://bhas.gov.ba/data/Publikacije/Bilteni/2023/PRI_00_2022_TB_1_BS.pdf (дата звернення: 25.05.2024).

12. Innovation index by country, around the world. *TheGlobalEconomy*. URL: https://www.theglobaleconomy.com/rankings/GII_Index/ (дата звернення: 25.05.2024).

13. List of available indicators. *TheGlobalEconomy*. URL: https://www.theglobaleconomy.com/indicators_list.php (дата звернення: 25.05.2024).

14. Hegarty B. S. UN sees life expectancy, education and income fall. *BBC News*. URL: https://www.bbc.com/news/world-62824357 (дата звернення: 25.05.2024).

15. Dasic B., Devic Z., Denic N., Zlatkovic D., Ilic I. D., Cao Y., Jermsittiparsert K., Van Le H. Human development index in a context of human development: Review on the western Balkans countries. *Brain and Behavior*. 2020. № 10 (9). 10 p. DOI: https://doi.org/10.1002/brb3.1755

16. Україна 2030: Доктрина збалансованого розвитку. Львів: Кальварія, 2017. 164 с.

17. Вараксіна О., Карлінська О., Петренко В. Сучасні виклики менеджменту людського капіталу. *Економіка та суспільство*. 2022. № 43. DOI: https://doi.org/10.32782/2524-0072/2022-43-53

18. Sahoo K., Samal A. K., Pramanik J., Pani S. K. Exploratory Data Analysis using Python. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*. 2019. № 8 (12). URL: https://www.researchgate.net/profile/Dr-Subhendu-Pani/publication/337146539_IJITEE/links/5dc70b124585151435fb427e/IJITEE.pdf (дата звернення: 25.05.2024).

19. What is exploratory data analysis (EDA)? *IBM*. URL: https://www.ibm.com/topics/exploratory-data-analysis (дата звернення: 25.05.2024).

20. Економіко-математичне моделювання : навч. посіб. / В. В. Вітлінський, С. І. Наконечний, О. Д. Шарапов, П. І. Верченко та ін. ; за заг. ред. В. В. Вітлінського. Київ : КНЕУ, 2008. 536 с.

21. Taylor S. Multiple Linear Regression. *Corporate Finance Institute.* URL: https://corporatefinanceinstitute.com/resources/data-science/multiple-linear-regression/ (дата звернення: 25.05.2024).

22. Cluster Analysis (5th ed.). / Everitt B. S., Landau S., Leese M., Stahl D. London : Wiley, 2011. 348 p. URL: https://cicerocq.wordpress.com/wp-content/uploads/2019/05/cluster-analysis_5ed_everitt.pdf (дата звернення: 25.05.2024).

23. Cornish R. Statistics: Cluster Analysis. *Mathematics Learning Support Centre*. 2007. P. 1-5. URL: https://www.lboro.ac.uk/departments/mlsc/ (дата звернення: 25.05.2024).

24. Why Do Data Analysts Use Python? *UCD Professional Academy*. URL: https://www.ucd.ie/professionalacademy/resources/why-do-data-analysts-use-python/ (дата звернення: 25.05.2024).

25. How to Deal with Missing Values in Your Dataset. *KDnuggets.* URL: https://www.kdnuggets.com/2020/06/missing-values-dataset.html (дата звернення: 25.05.2024).

26. Forecasting at scale. *Prophet*. URL: https://facebook.github.io/prophet/ (дата звернення: 25.05.2024).

27. Schober P., Boer C., Schwarte L. Correlation Coefficients: Appropriate Use and Interpretation. *Anesth Analg*. 2018. 126(5). P. 1763-1768. URL: https://pubmed.ncbi.nlm.nih.gov/29481436/ (дата звернення: 25.05.2024).

28. Kim J. H. Multicollinearity and misleading statistical results. *Korean J Anesthesiol.* 2019. 72(6). P. 558–569. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6900425/ (дата звернення: 25.05.2024).

29. Hayes A. Multicollinearity: Meaning, Examples, and FAQs. *Investopedia.* URL: https://www.investopedia.com/terms/m/multicollinearity.asp (дата звернення: 25.05.2024).

30. Reducing Data-based Multicollinearity. *The Pennsylvania State University*. URL: https://online.stat.psu.edu/stat462/node/181/ (дата звернення: 25.05.2024).

31. Brownlee J. What is the Difference Between Test and Validation Datasets? *Machine Learning Mastery*. URL: https://machinelearningmastery.com/difference-test-validation-datasets (дата звернення: 25.05.2024).

32. An Introduction to Statistical Learning: with Applications in R (2nd ed.). / James G., Witten D., Hastie T., Tibshirani R. Springer. 2013. 176 p. URL: https://www.statlearning.com/ (дата звернення: 25.05.2024).

33. Brucher M. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research.* 2011. 12. P. 2826-2830. URL: https://jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf (дата звернення: 25.05.2024).

34. LinearRegression. *scikit-learn*. URL: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html (дата звернення: 25.05.2024).

35. Student’s t-test. Definition, Formula, & Example. *Encyclopedia Britannica*. URL: https://www.britannica.com/science/Students-t-test (дата звернення: 25.05.2024).

36. Understanding the t-Test in Linear Regression. *Statology.* URL: https://www.statology.org/t-test-linear-regression/ (дата звернення: 25.05.2024).

37. Jain A. K., Murty M. N., Flynn P. J. Data clustering: a review. *ACM Computing Surveys.* 1999. № 3, vol. 31. URL: https://dl.acm.org/doi/10.1145/331499.331504 (дата звернення: 25.05.2024).

38. Abdulhafedh A. Incorporating K-means, Hierarchical Clustering and PCA in Customer Segmentation. *Open Access Peer Reviewed Journals*. URL: https://pubs.sciepub.com/jcd/3/1/3/index.html (дата звернення: 25.05.2024).

39. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning Data Mining, Inference, and Prediction. *Springer Series in Statistics*. URL: https://hastie.su.domains/Papers/ESLII.pdf (дата звернення: 25.05.2024).

40. Trotta F. Unsupervised Learning in Python: A Gentle Introduction to Clustering Techniques for Discovering Patterns. Codemotion. URL: https://www.codemotion.com/magazine/ai-ml/machine-learning/clustering-python-patterns/ (дата звернення: 25.05.2024).

ДОДАТКИ

Додаток А.

	Year	HDI	GDP per capita	Households consumption	Inflation	Unemployment	Corruption perceptions	Urban population	Life expectancy
0	1995	0.683	935.958740	55.1	376.70	14.8	NaN	66.951	66.741951
1	1996	0.683	872.713501	58.1	80.30	10.0	NaN	66.990	67.022439
2	1997	0.688	991.242798	54.2	15.90	9.8	NaN	67.029	67.576341
3	1998	0.694	835.236633	56.8	10.60	11.3	28.0	67.067	68.374878
4	1999	0.696	635.757935	57.2	22.70	11.9	26.0	67.106	67.980976
5	2000	0.698	658.344604	57.4	28.20	11.5	15.0	67.145	67.675610
6	2001	0.713	807.801941	57.9	12.00	10.8	21.0	67.183	67.837073
7	2002	0.721	911.906860	57.8	0.80	9.6	24.0	67.283	68.275610
8	2003	0.730	1087.788086	57.3	5.20	9.1	23.0	67.427	68.210732
9	2004	0.739	1416.603760	54.6	9.00	8.6	22.0	67.597	68.185366
10	2005	0.743	1894.460083	59.1	13.50	7.2	26.0	67.790	67.956829
11	2006	0.752	2391.323975	60.5	9.10	6.8	28.0	67.969	68.077561
12	2007	0.759	3197.934326	60.6	12.80	6.4	27.0	68.147	68.222195
13	2008	0.763	4066.531738	63.2	25.20	6.4	25.0	68.325	68.251463
14	2009	0.761	2639.378174	65.1	15.90	8.8	22.0	68.502	69.190000
15	2010	0.763	3078.414795	64.7	9.40	8.1	24.0	68.596	70.265366
16	2011	0.772	3704.842285	67.8	8.00	7.9	23.0	68.689	70.809268
17	2012	0.774	4004.789795	69.3	0.60	7.5	26.0	68.782	70.944146
18	2013	0.774	4187.739746	72.8	-0.20	7.2	25.0	68.875	71.159512
19	2014	0.774	3104.653809	71.4	12.10	9.3	26.0	68.968	71.186585
20	2015	0.764	2124.662598	67.7	48.70	9.1	27.0	69.061	71.189512
21	2016	0.767	2187.727539	66.5	13.90	9.5	29.0	69.154	71.476341
22	2017	0.771	2638.325439	67.1	14.40	9.7	30.0	69.246	71.780976
23	2018	0.771	3096.562500	69.3	11.00	9.0	32.0	69.352	71.582683
24	2019	0.774	3661.457764	74.3	7.90	8.5	30.0	69.473	71.827317
25	2020	0.762	3751.737305	73.2	2.70	9.2	33.0	69.608	71.185122
26	2021	0.765	4827.845703	69.1	9.30	9.8	32.0	69.757	69.647805
27	2022	0.734	4533.975586	65.9	20.18	24.5	33.0	69.919	72.500000

Рисунок А.1 – Таблиця початкових даних вибірки для регресійної моделі

Джерело: розроблено автором за даними [1-13]

Додаток Б.

	Human development Index	GDP per capita, current dollars	Inflation, consumer prices	Unemployment rate	Corruption perceptions	Innovation index	Property rights index	Percent urban population	Life expectancy
Austria	0.926	52084.681200	8.546870	4.8	71	50.2	98	59.256	82.05
Belgium	0.942	49926.825430	9.597512	5.6	73	46.9	93	98.153	82.17
Bosnia and Herzegovina	0.779	7568.798480	14.000000	15.4	34	28.5	49	49.841	77.93
Bulgaria	0.799	13974.449250	15.325259	4.2	43	39.5	77	76.363	75.49
Croatia	0.878	18570.404000	10.780581	6.8	50	35.6	81	58.219	79.02
Czech Republic	0.895	27226.615640	15.100165	2.2	56	42.8	89	74.377	79.85
Denmark	0.952	67790.053990	7.696567	4.5	90	55.9	99	88.367	81.40
Estonia	0.899	28247.095990	19.398263	5.6	74	50.2	92	69.609	79.18
Finland	0.942	50871.930450	7.123508	6.8	87	56.9	100	85.681	82.48
France	0.910	40886.253270	5.222367	7.3	72	55.0	94	81.509	83.13
Germany	0.950	48717.991140	6.872574	3.1	79	57.2	96	77.648	81.88
Greece	0.893	20867.269090	9.645260	12.4	52	34.5	76	80.357	82.80
Hungary	0.851	18390.185000	14.608144	3.6	42	39.8	76	72.552	77.31
Iceland	0.959	73466.778670	8.308755	3.8	74	49.5	97	93.992	83.52
Ireland	0.950	103983.291300	7.807375	4.5	77	48.5	93	64.183	82.81
Italy	0.906	34776.423230	8.201290	8.1	56	46.1	82	71.657	84.01
Latvia	0.879	21779.504260	17.310283	6.9	59	36.5	89	68.540	75.73
Lithuania	0.879	25064.808910	19.705046	5.9	62	37.4	89	68.465	76.41
Netherlands	0.946	57025.012460	10.001208	3.5	80	58.0	96	92.886	82.78
North Macedonia	0.765	6591.471314	14.204717	14.4	40	28.8	57	59.118	76.26
Norway	0.966	108729.186900	5.764123	3.3	84	48.8	100	83.664	82.94
Poland	0.881	18688.004490	14.429451	2.9	55	37.5	72	60.134	79.27
Portugal	0.874	24515.265850	7.832691	6.1	62	42.1	90	67.381	82.65
Romania	0.827	15786.801740	13.795489	5.6	46	34.1	81	54.489	76.50
Serbia	0.805	9537.682867	11.981512	9.4	36	32.3	59	56.873	76.47
Slovakia	0.855	21256.808430	12.774146	6.2	53	34.3	83	53.909	78.00
Slovenia	0.926	28439.334100	8.833699	4.0	56	40.6	90	55.751	81.85
Spain	0.911	29674.544290	8.390576	12.9	60	44.6	88	81.304	83.99
Sweden	0.952	56424.284700	8.369291	7.5	83	61.6	97	88.492	83.33
Switzerland	0.967	93259.905720	2.835028	2.2	82	64.6	95	74.092	84.25
Ukraine	0.734	4533.975586	20.183637	24.5	33	31.0	40	69.919	72.50
United Kingdom	0.940	46125.255750	7.922049	3.9	73	59.7	96	84.398	81.77

Рисунок Б.1 – Таблиця початкових даних вибірки для моделі кластеризації

Джерело: розроблено автором за даними [1-13]