МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Сумський державний університет
Навчально-науковий інститут бізнесу, економіки та менеджменту
Кафедра економічної кібернетики

«До захисту допущено»

Завідувач кафедри

_____ Віталія КОЙБІЧУК
(підпис) (Ім'я та ПРІЗВИЩЕ)

_____ 2024 р.

КВІЛІФІКАЦІЙНА РОБОТА

на здобуття освітнього ступеня магістр
(бакалавр / магістр)

зі спеціальності 051 «Економіка»,
(код та назва)

освітньо-професійної програми Економічна кібернетика
(освітньо-професійної / освітньо-наукової) (назва програми)

на тему: Прогнозування розвитку цифрової економіки країни на основі методів машинного навчання

Здобувачки групи ЕК.мз-31с Солярової Катерини Геннадіївни
(шифр групи) (прізвище, ім'я, по-батькові)

Кваліфікаційна робота містить результати власних досліджень.

Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело.

Керівник доцентка, д. е. н., Ганна Яровенко _____
(посада, науковий ступінь, вчене звання, Ім'я та ПРІЗВИЩЕ) (підпис)

Суми – 2024

Міністерство освіти і науки України
Сумський державний університет
Навчально-науковий інститут бізнесу, економіки та менеджменту
Кафедра економічної кібернетики

ЗАТВЕРДЖУЮ
Завідувач кафедри
к.е.н., доцент
_____Віталія КОЙБІЧУК
“___” _____ 2024 р.

ЗАВДАННЯ НА КВАЛІФІКАЦІЙНУ БАКАЛАВРСЬКУ РОБОТУ

(спеціальність 051 Економіка «Економічна кібернетика», «Бізнес аналітика»)

студенту 2 курсу, групи ЕК.мз-31с

Соляровій Катерині Геннадіївні
(прізвище, ім’я, по батькові студента)

1. Тема роботи Прогнозування розвитку цифрової економіки країни на основі методів машинного навчання

затверджена наказом по університету від «1» жовтня 2024 року № 1002-VI

2. Термін подання студентом закінченої роботи «5» грудня 2024 року

3. Метою дослідження є побудова моделей прогнозування розвитку цифрової економіки країни на основі методів машинного навчання.

4. Об’єкт дослідження — економічні процеси, що відображають вплив цифрових технологій на розвиток економіки країни.

5. Предмет дослідження — методи машинного навчання та аналітичні підходи для прогнозування розвитку цифрової економіки країни.

6. Кваліфікаційна робота виконується на матеріалах інформаційно-аналітичної платформи United Nations.

7. Орієнтовний план кваліфікаційної роботи, терміни подання розділів керівникові та зміст завдань для виконання поставленої мети

Розділ 1 Аналіз інформаційних трендів кібератак «15» жовтня 2024 рокуУ розділі 1необхідно визначити сутність та значення цифрової економіки країни, проаналізувати стан цифрової економіки України та світу, визначити проблему прогнозування розвитку цифрової економіки.

Розділ 2 Статистичні тести інформаційних трендів кібератак «3» листопада 2024 року

У розділі 2 необхідно провести статистичний аналіз вхідного масиву дослідження, провести регресійний аналіз, охарактеризувати XGBoost, Дерево рішень та Випадковий ліс, як методи прогнозування розвитку цифрової економіки.

Розділ 3 Побудова моделей та прогнозів «15» листопада 2024 року
У розділі 3 необхідно оцінити якість прогнозних моделей, реалізувати прогнози на основі побудованих моделей, перевірити якість отриманих прогнозів.

8. Консультації з роботи:

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
1			
2			
3			

9. Дата видачі завдання: «15» жовтня 2024 року

Керівник кваліфікаційної роботи _____ Г. М. Яровенко
(підпис) (ініціали, прізвище)

Завдання до виконання одержав _____ К. Г. Солярова
(підпис) (ініціали, прізвище)

SUMMARY

master's qualification thesis on the topic

“FORECASTING the DEVELOPMENT of the COUNTRY'S DIGITAL ECONOMY BASED on MACHINE LEARNING METHODS”

student of Solarova Kateryna Gennadiyevna

The relevance of the topic considered within the scope of the study is determined by the fact that forecasting the development of the digital economy based on machine learning methods allows to effectively assess the impact of modern digital technologies on economic processes. This is important for the formation of sound strategies for the development of national economies, in particular in the context of global changes, such as the introduction of e-government and the improvement of infrastructure.

The purpose of the study is to build models for forecasting the development of the country's digital economy based on machine learning methods.

The object of the study is economic processes that reflect the impact of digital technologies on the development of the country's economy.

The subject of the research is machine learning methods and analytical approaches for forecasting the development of the country's digital economy based on key indicators of digital transformation.

The objectives of the research are:

1) to determine the essence and significance of the country's digital economy;

2) to analyze the state of the digital economy of Ukraine and the world;

3) to determine the problem of forecasting the development of the digital economy;

4) to conduct a statistical analysis of the input array of the study;

5) to conduct a regression analysis;

6) to characterize XGBoost, Decision Tree and Random Forest as methods of forecasting the development of the digital economy;

7) to evaluate the quality of predictive models;

8) to implement forecasts based on built models;

9) to check the quality of received forecasts.

The following research methods were used to achieve the set goal and objectives of the study: data processing and integration, generalization of the main results, in-depth study of individual aspects, argumentation of the obtained conclusions, comparative analysis and arrangement of data, thanks to which the main conclusions were formulated. Statistical analysis methods were used for calculations.

The information base of the qualification work is the information and analytical platform of the United Nations.

The main scientific result of the qualifying master's thesis is as follows: models for forecasting the development of the digital economy were developed and tested using machine learning methods. The models were checked for the presence of significant economic factors affecting the development of the digital economy, which made it possible to create a reliable forecast for future periods.

The obtained results can be used by government bodies to develop effective strategies in the field of digital transformation, which will contribute to the sustainable development of the digital economy.

The work was carried out within the framework of the NDR № 0124U000544 Cybersecurity and digital transformations of the country's wartime economy: the fight against cybercrimes, corruption and the shadow sector.

Keywords: XGBoost, Decision Tree, Random Forest, digital economy, machine learning, forecasting, innovative technologies.

АНОТАЦІЯ

кваліфікаційної роботи магістра на тему

«ПРОГНОЗУВАННЯ РОЗВИТКУ ЦИФРОВОЇ ЕКОНОМІКИ КРАЇНИ НА ОСНОВІ МЕТОДІВ МАШИННОГО НАВЧАННЯ»

студентки Солярової Катерини Геннадіївни

(прізвище, ім'я, по батькові)

Актуальність теми, розглянутої в межах дослідження, обумовлюється тим, що прогнозування розвитку цифрової економіки на основі методів машинного навчання дозволяє ефективно оцінювати вплив сучасних цифрових технологій на економічні процеси. Це важливо для формування обґрунтованих стратегій розвитку національних економік, зокрема в умовах глобальних змін, таких як впровадження електронного урядування та покращення інфраструктури.

Метою дослідження є побудова моделей прогнозування розвитку цифрової економіки країни на основі методів машинного навчання.

Об'єктом дослідження є економічні процеси, що відображають вплив цифрових технологій на розвиток економіки країни.

Предметом дослідження є методи машинного навчання та аналітичні підходи для прогнозування розвитку цифрової економіки країни на основі ключових індикаторів цифрової трансформації.

Задачами дослідження є:

1) визначити сутність та значення цифрової економіки країни;

2) проаналізувати стан цифрової економіки України та світу;

3) визначити проблему прогнозування розвитку цифрової економіки;

4) провести статистичний аналіз вхідного масиву дослідження;

5) провести регресійний аналіз;

6) охарактеризувати XGBoost, Дерево рішень та Випадковий ліс, як методи прогнозування розвитку цифрової економіки;

7) оцінити якість прогнозних моделей;

8) реалізувати прогнози на основі побудованих моделей;

9) перевірити якість отриманих прогнозів.

Для досягнення поставленої мети та задач дослідження були використані такі методи дослідження: обробка та інтеграція даних, узагальнення основних результатів, поглиблене вивчення окремих аспектів, аргументація отриманих висновків, порівняльний аналіз та впорядкування даних, завдяки яким були сформульовані основні висновки. Для проведення розрахунків застосовувалися методи статистичного аналізу.

Інформаційною базою кваліфікаційної роботи є інформаційно-аналітична платформа United Nations.

Основний науковий результат кваліфікаційної магістерської роботи полягає у такому: були розроблені та протестовані моделі прогнозування розвитку цифрової економіки, використовуючи методи машинного навчання. Моделі були перевірені на наявність суттєвих економічних факторів, що впливають на розвиток цифрової економіки, що дозволило створити надійний прогноз для майбутніх періодів.

Одержані результати можуть бути використані державними органами та для розробки ефективних стратегій у сфері цифрової трансформації, що сприятимуть сталому розвитку цифрової економіки.

Роботу було виконано в рамках НДР № 0124U000544 Кібербезпекові та цифрові трансформації економіки країни воєнного часу: боротьба із кіберзлочинами, корупцією та тіньовим сектором.

Ключові слова: XGBoost, Дерево рішень, Випадковий ліс, цифрова економіка, машинне навчання, прогнозування, інноваційні технології.

Зміст кваліфікаційної магістерської роботи викладено на 86 сторінках. Список використаних джерел із 56 найменувань, розміщений на 7 сторінках. Робота містить 78 рисунків, а також 1 додаток, розміщених на 18 сторінках.

Рік виконання кваліфікаційної роботи – 2024 рік.

Рік захисту роботи – 2024 рік.

ЗМІСТ

ВСТУП

Цифрова економіка є одним із найважливіших напрямів сучасного економічного розвитку, що суттєво впливає на ефективність бізнесу, державного управління та рівень життя суспільства. В умовах глобальної конкуренції та швидких технологічних змін виникає потреба у точному прогнозуванні її розвитку для прийняття обґрунтованих стратегічних рішень.

Методи машинного навчання відкривають нові можливості для аналізу економічних процесів, завдяки здатності працювати з великими масивами даних, враховувати складні залежності та адаптуватися до динамічних умов. У цій роботі розглядається підхід до прогнозування розвитку цифрової економіки країни, що базується на застосуванні методів машинного навчання, таких як XGBoost, Дерево рішень та Випадковий ліс із використанням панельних даних.

Актуальність теми, розглянутої в межах дослідження, обумовлюється тим, що прогнозування розвитку цифрової економіки на основі методів машинного навчання дозволяє ефективно оцінювати вплив сучасних цифрових технологій на економічні процеси. Це важливо для формування обґрунтованих стратегій розвитку національних економік, зокрема в умовах глобальних змін, таких як впровадження електронного урядування та покращення інфраструктури.

Метою дослідження є побудові моделей прогнозування розвитку цифрової економіки країни на основі методів машинного навчання.

Об’єктом дослідження є економічні процеси, що відображають вплив цифрових технологій на розвиток економіки країни.

Предметом дослідження є методи машинного навчання та аналітичні підходи для прогнозування розвитку цифрової економіки країни на основі ключових індикаторів цифрової трансформації.

Для досягнення мети дослідження необхідно виконати ряд ключових завдань, зокрема:

— визначити сутність та значення цифрової економіки країни;

— проаналізувати стан цифрової економіки України та світу;

— визначити проблему прогнозування розвитку цифрової економіки;

— провести статистичний аналіз вхідного масиву дослідження;

— провести регресійний аналіз;

— охарактеризувати XGBoost, Дерево рішень та Випадковий ліс, як методи прогнозування розвитку цифрової економіки;

— оцінити якість прогнозних моделей;

— реалізувати прогнози на основі побудованих моделей;

— перевірити якість отриманих прогнозів.

Для дослідження було використано: набір даних (1386 спостереження та 7 змінних), на основі якого було проведено аналіз, побудовано модель та зроблено прогнози. Для виконання розрахунків та візуалізації отриманих результатів використовувалася документація з мови програмування Python.

Роботу було виконано в рамках НДР № 0124U000544 Кібербезпекові та цифрові трансформації економіки країни воєнного часу: боротьба із кіберзлочинами, корупцією та тіньовим сектором.

РОЗДІЛ 1 СТАН ТА ПЕРСПЕКТИВИ РОЗВИТКУ ЦИФРОВОЇ ЕКОНОМІКИ КРАЇН

1.1 Сутність та значення цифрової економіки країни

Цифрова економіка стала основною рушійною силою для сучасного розвитку країни, забезпечуючи нові можливості для економічного зростання, підвищення конкурентоспроможності та покращення якості життя населення [1]. Це поняття охоплює усі економічні процеси, які використовують цифрові технології, зокрема, інформаційно-комунікаційні технології, для створення вартості продукції та послуг. Цифровізація забезпечує нові можливості для соціальної інтеграції, дозволяючи менш захищеним верствам населення брати участь в економічних процесах через доступ до онлайн-освіти, віддаленої роботи або електронної комерції [2]. Це особливо важливо для регіонів, де традиційна інфраструктура слабко розвинена. Разом із тим, цифрова нерівність залишається викликом, що потребує подолання через розширення доступу до швидкісного інтернету та навчання цифровим навичкам.

Сучасні тенденції розвитку цифрової економіки охоплюють широкий спектр напрямів, які сприяють впровадженню новітніх технологій, трансформують ринки праці та впливають на соціальні зміни. Цифрова економіка є багатовимірним явищем, яке охоплює не лише економічні аспекти, але й суттєво впливає на соціальні, культурні та екологічні процеси. Її розвиток змінює структуру зайнятості, способи споживання, бізнес-моделі, а також підходи до освіти та професійної підготовки. У 2024 році визначальними є наступні напрями, що відображають ці зміни [3].

Одним із перших факторів, що займає ключове місце серед тенденцій є штучний інтелект (AI) та автоматизація. Їх застосування у виробництві, сфері охорони здоров’я, фінансовому секторі та державному управлінні суттєво трансформує бізнес-моделі, створюючи нові можливості для впровадження

інновацій. Зокрема, автоматизація процесів, включаючи використання роботизованих систем, забезпечує підвищення продуктивності та ефективності. Такі зміни викликають занепокоєння щодо потенційного скорочення робочих місць у традиційних секторах, що потребує відповідної адаптації працівників до нових економічних умов [4].

Значення даних і аналітики стало невід'ємною характеристикою сучасної цифрової економіки. Дані стали одним із ключових ресурсів, що визначає конкурентоспроможність організацій та є потужним інструментом для оптимізації процесів, прийняття рішень та інновацій. Сучасні технології аналізу великих даних та хмарні обчислення дозволяють здійснювати ефективну обробку великих обсягів інформації, що надходить з численних джерел, наприклад, соціальні медіа, онлайн-транзакції, дані мобільних додатків, GPS-сигнали [5]. Це допомагає підприємствам отримувати важливу інформацію, яка дозволяє точніше прогнозувати ринкові зміни та визначати потенційні можливості для оптимізації продуктів і послуг. Застосування автоматизованих системи аналізу даних, технологій машинного навчання та штучного інтелекту, дозволяє компаніям досягати високої точності в ухваленні рішень і підвищувати ефективність бізнес-процесів [6].

Цифрова економіка також відіграє важливу роль у досягненні глобальних соціальних та екологічних цілей. В умовах стрімкої автоматизації та цифровізації економічних процесів виникає потреба у підвищенні цифрової грамотності населення. Низький рівень цифрових навичок може стати значною перешкодою для впровадження інноваційних технологій, зокрема у віддалених регіонах, що створює цифровий розрив [7]. Це, у свою чергу, впливає на соціальну нерівність і потребує впровадження державних програм для навчання та перекваліфікації працівників.

Екологічний аспект цифровізації також має неабиякий вплив. Хоча цифрові технології сприяють оптимізації ресурсів і зменшенню викидів у багатьох галузях, їхній розвиток супроводжується значними викликами,

такими як утилізація електронних відходів і високий рівень енергоспоживання дата-центрів [8]. Інтеграція принципів сталого розвитку в цифрову економіку є необхідною умовою для її подальшого прогресу.

Розвиток платформної економіки визначає трансформацію ринків і створює нові можливості для бізнесу [9]. Такі платформи, як Alibaba, Amazon та Uber значно змінюють традиційні моделі бізнесу, дозволяючи споживачам і підприємцям швидко отримувати доступ до товарів та послуг через цифрові канали. Ці платформи виступають у якості посередників між постачальниками і споживачами, створюючи нові можливості для малих підприємств, які раніше не мали доступу до глобальних ринків. Також вони сприяють більшому залученню стартапів, які можуть масштабувати свої операції завдяки цим платформам.

Інтернет речей також є важливим компонентом сучасної цифрової економіки, який визначає зміни в різних секторах економіки. Розширення інтернету речей охоплює поєднання фізичних пристроїв з мережею Інтернет для забезпечення автоматизованого збору та обміну даними, що має важливе значення для підвищення ефективності процесів в таких сферах, як виробництво, енергетика, охорона здоров'я та транспорт [10]. Впровадження цього компоненту сприяє оптимізації операцій, зменшенню витрат і покращенню управлінських рішень. Однак цей процес також створює нові виклики для забезпечення кібербезпеки, оскільки зростає кількість підключених пристроїв і обсяг переданих даних, що вимагає посиленої уваги до захисту та конфіденційності інформації.

Зі зростанням цифровізації у нашому житті, кібербезпека стала однією з найважливіших задач, що стоїть перед урядом, бізнесом та суспільством. Адже зі збільшення фінансових транзакцій, надання державних послуг зростає і кількість кіберзагроз, які можуть призвести до серйозних фінансових втрат та порушення конфіденційності даних. Незахищеність інформаційних систем, що використовуються як державними установами, так і приватними

компаніями, є одним із основних факторів, які спонукають до необхідності значних інвестицій у захист даних та цифрових інфраструктур [11]. Тому глобальні й національні організації спрямовують значні ресурси на розробку та впровадження сучасних технологій кіберзахисту.

Цифровізація державних послуг також є ключовим елементом сучасної цифрової трансформації, спрямованим на спрощення доступу громадян до адміністративних послуг та підвищення ефективності управлінських процесів. Чимало країн активно розвивають і впроваджують цифрові платформи, що дозволяють громадянам отримувати необхідні адміністративні послуги в онлайн-форматі. Наприклад, у Великій Британії платформа Gov.uk [12] дає змогу громадянам звертатися до різних державних служб, подавати заяви та отримувати інформацію про послуги в електронному вигляді. Схожу платформу має й Естонія, яка є одним із лідерів у цифровізації державних послуг, забезпечуючи своїх громадян можливістю голосувати, отримувати медичні послуги, а також реєструвати бізнес через e-Estonia [13].

В Україні прикладом цифровізації є платформа «Дія» [14], яка спрямована на покращення доступу громадян до адміністративних послуг. Вона надає можливість безпосередньо взаємодіяти з різними державними установами в електронному вигляді, включаючи подачу заяв, отримання необхідних документів і сертифікатів, а також доступ до широкого спектра державних послуг через інтернет. Ця платформа значно знижує бюрократичні бар'єри, оптимізує процеси взаємодії між громадянами та державою, роблячи ці процеси швидкими, зручними та прозорими.

Посилення міжнародної співпраці у сфері цифрової економіки стало ще одним важливим трендом. В епоху глобалізації цифрові платформи, фінансові системи та технології забезпечують інтеграцію національних економік у світову цифрову інфраструктуру. Україна активно розвиває партнерства з міжнародними організаціями, такими як Європейський Союз, для сприяння

впровадженню єдиних цифрових стандартів, кібербезпеки та інтеграції в глобальні технологічні ланцюги.

1.2 Аналіз стану цифрової економіки України та світу

Цифрова економіка є ключовим чинником сучасного економічного розвитку, що впливає на всі сфери суспільного життя: від економічної діяльності та управління державою до освіти та соціальних відносин. У глобальному контексті «цифровізація» перетворюється на основу конкурентоспроможності країн, оскільки інноваційні технології створюють нові можливості для бізнесу, оптимізації процесів і доступу до інформації. Аналіз стану цифрової економіки України та світу дозволяє оцінити, які є перспективи розвитку, виявити сильні сторони та визначити ключові виклики.

Для глибшого розуміння ролі цифрової економіки у сучасному світі необхідно звернути увагу на ключові тенденції та особливості її розвитку [15]. Порівняльний аналіз досвіду провідних країн та України дозволяє виділити найкращі практики, а також визначити бар’єри, які заважають повноцінній інтеграції України до глобального цифрового простору. Особливий акцент слід зробити на розвитку інфраструктури, інновацій, кібербезпеки та підвищенні цифрової грамотності населення як основних драйверів цифровізації. Саме ці напрями можуть стати основою для визначення стратегічних пріоритетів у процесі подальшої трансформації економіки.

Одним із ключових досягнень України у сфері цифрової економіки є IT-галузь, яка забезпечує вагому частину експорту послуг. Значний поштовх розвитку IT-галузі в Україні забезпечує спеціальний правовий режим Дія.City, запроваджений Міністерством цифрової трансформації. Ця ініціатива пропонує низку фінансових переваг, таких як знижені податкові ставки: 5%

ПДФО замість стандартних 18%, 1,5% військовий збір та 9% податок на виведений капітал замість традиційного податку на прибуток у розмірі 18% [16]. Такі умови значно зменшують фінансове навантаження на компанії, стимулюють розвиток стартапів та сприяють залученню висококваліфікованих фахівців. За інформацією Міністерства цифрової трансформації, режим Дія.City є одним із найкращих у світі інструментів для підтримки продуктових IT-компаній, навіть у складних умовах військового часу [17].

Згідно з даними Української асоціації IT [18], у 2022 році обсяг експорту українських IT-послуг становив рекордні $7,3 млрд, що стало піковим показником за всю історію галузі. Але в 2023 році цей показник скоротився на $0,6 млрд (8,4 %) і склав $6,7 млрд. Таким чином, у 2023 році галузь зазнала першого значного падіння після багаторічного стабільного зростання [19]. Це скорочення може бути , як наслідок глобальних економічних труднощів, впливу війни на бізнес-середовище та зменшення попиту на міжнародному ринку.

Світові економічні лідери активно інвестують у цифрову трансформацію своїх економік, що підтверджується збільшенням частки цифрової економіки в глобальному ВВП, яка, за даними Світового економічного форуму [20], перевищила 20% у 2023 році.

США досягли значного прогресу в розвитку фінансових технологій, створенні нових платформ та застосуванні великих даних, що сприяє зростанню таких компаній, як PayPal і Stripe. Це також відкриває нові можливості для стартапів у фінансовому секторі, дозволяючи їм впроваджувати інноваційні рішення та залучати інвестиції.

У той же час, Китай активно розвиває інфраструктуру цифрових технологій, зокрема мережі 5G, штучний інтелект (ШІ) і електронну комерцію. Китайські компанії, такі як Alibaba та Tencent, стали лідерами на світовому

ринку електронної комерції, що підтверджується даними про домінування китайських платформ в обсягах онлайн-продажів [21].

У Європейському Союзі країни активно працюють над створенням єдиного цифрового ринку, що включає як юридичні ініціативи, так і інфраструктурні рішення. Одним із важливих кроків стало введення Регламенту GDPR, який забезпечує захист персональних даних та регулює діяльність цифрових платформ. Згідно з даними Єврокомісії, цифрова економіка ЄС зросла на 12% у 2023 році, де основними драйверами стали розвиток цифрових послуг, автоматизація виробництва та блокчейн-технології [22]. Ці ініціативи сприяють поглибленій інтеграції цифрових технологій у різні галузі економіки ЄС, що робить європейський ринок більш ефективним та конкурентоспроможним.

Цифрова економіка України демонструє значний потенціал для розвитку, незважаючи на складнощі сьогодення, пов’язані з економічною нестабільністю та військовими діями. Впровадження сучасних технологій, таких як штучний інтелект, блокчейн і цифрові послуги, сприяє модернізації основних секторів економіки, забезпечуючи їх ефективність та конкурентоспроможність [23]. Державні та приватні ініціативи відіграють ключову роль у залученні України до глобальних цифрових процесів, створюючи передумови для зміцнення її конкурентних переваг та розв'язання нагальних проблем [24].

Однією з ключових перспектив є розвиток цифрової інфраструктури, включаючи доступ до швидкісного Інтернету в усіх регіонах країни. За даними ITU, станом на 2023 рік рівень проникнення інтернету в Україні становить близько 68% [25]. Для порівняння, у Польщі цей показник становить понад 80%, у Румунії – близько 78%. Така різниця свідчить про необхідність пріоритетного інвестування у розвиток мереж 4G та 5G в Україні, що дозволить суттєво підвищити доступність цифрових послуг [26], [27].

Ще одним важливим напрямом є стимулювання інновацій у сферах штучного інтелекту, фінансових технологій та кібербезпеки. За прогнозами Організації економічного співробітництва та розвитку за 2023 рік, застосування технологій штучного інтелекту може збільшити продуктивність праці в країнах з перехідною економікою, таких як Україна. Збільшення прогнозується на 20 – 30 % у наступні 10 років. Важливим етапом для цього є залучення іноземних інвестицій і створення сприятливих умов для розвитку стартапів, що займаються сучасними технологіями [28]. Перспективи цифровізації України тісно пов’язані з інтеграцією в європейський цифровий простір. В межах Угоди про асоціацію з Європейським Союзом [29] Україна адаптує своє законодавство до європейських стандартів, що створює умови для залучення іноземних інвестицій і доступу до передових технологій. Зокрема, синхронізація з європейськими ініціативами, такими як Єдиний цифровий ринок, сприяє покращенню регулювання електронної комерції, кібербезпеки та управління даними [30].

Підвищення цифрової грамотності населення є не менш важливим аспектом розвитку цифрової економіки України. За даними Міжнародного союзу електрозв'язку, лише близько 40 % українців володіють достатніми цифровими навичками для ефективного використання інтернет-ресурсів і сучасних технологічних платформ. Цей показник є значно нижчим порівняно з провідними країнами Європи, де рівень цифрової компетентності перевищує 70 % [26]. Але уряд активно впроваджує ініціативи, спрямовані на підвищення цифрових навичок населення. Однією з основних стратегій є Концепція розвитку цифрових компетентностей до 2025 року, затверджена Кабінетом Міністрів України [31]. Її метою є навчання цифровій грамотності 6 мільйонів українців протягом трьох років, що сприятиме загальному зростанню рівня цифрової обізнаності, а також їх відповідності вимогам ринку праці.

Для цього Міністерство цифрової трансформації запустило платформу «Дія.Освіта», де доступні курси у форматі освітніх серіалів, спрямовані на

розвиток цифрових навичок у різних сферах. На порталі також діє національний тест «Цифрограм», що дозволяє перевірити базові цифрові компетенції, важливі для повсякденного життя та професійної діяльності [32].

Для подальшого розвитку цифрової економіки України необхідно зосередитись на кількох ключових аспектах: забезпеченні доступу до технологій, зміцненні цифрової інфраструктури, підтримці інновацій та залученні міжнародних партнерів і інвестицій. Розвиток цифрової економіки сприятиме створенню нових робочих місць у сфері високих технологій, особливо у фінансових технологіях, кібербезпеці та штучному інтелекті. Це дозволить Україні не лише стимулювати економічне зростання, але й інтегруватися у глобальний цифровий простір. Ініціативи, такі як «Дія», а також міжнародні угоди з країнами ЄС і іншими партнерами, мають стати рушійними силами для розвитку цифрової економіки.

1.3 Постановка проблеми прогнозування розвитку цифрової економіки

Прогнозування розвитку цифрової економіки є важливим завданням сучасної економічної науки, що дозволяє розробляти ефективні стратегії цифровізації на національному та глобальному рівнях. Однак, цей процес супроводжується численними проблемами, які обумовлюють необхідність використання новітніх методів аналізу і прогнозування, що базуються на великих масивах даних і машинному навчанні [33].

Цифрова економіка характеризується високою динамічністю та залежністю від інноваційних технологій. Однією з основних проблем є відсутність єдиної концептуальної моделі для аналізу та прогнозування її розвитку. Сучасні підходи до прогнозування часто ігнорують багатовимірність економічних процесів і обмежуються традиційними

методами, які можуть бути недостатньо точними для аналізу складних взаємозв’язків між показниками.

Ще однією важливою проблемою є обмеження в структурі даних. У цьому дослідженні використовуються панельні дані по країнам за період з 2003 по 2022 роки, причому дані з 2008 року збиралися з інтервалом у два роки. Така специфіка формату створює додаткові труднощі для аналізу, адже перерви в часових рядах можуть спричинити втрату інформації про динамічні тренди. Це вимагає застосування методів, які здатні ефективно працювати з неповними наборами даних і коректно враховувати тимчасові розриви.

Одним із критичних етапів підготовки до прогнозування є забезпечення якості даних. На основі наданого масиву були проведені такі кроки:

1. Перевірка відсутніх значень із використанням візуалізації. Це дозволило ідентифікувати потенційні прогалини.
2. Описова статистика для аналізу основних характеристик показників.
3. Кореляційний аналіз, який дозволив виявити взаємозв’язки між числовими змінними та оцінити ризик мультиколінеарності.
4. Стандартизація даних для забезпечення порівнянності змінних з різними масштабами.

На основі попереднього аналізу була розроблена концептуальна модель, що містить кілька ключових етапів:

1. Побудова головних компонент (PCA) для зменшення розмірності даних. Метод дозволив виділити три основні компоненти, що пояснюють найбільшу частку варіації у даних [34].
2. Кластеризація на основі методу локтя для визначення оптимальної кількості кластерів. Це дозволило сегментувати країни за рівнем цифрового розвитку.
3. Побудова регресійних моделей (OLS Regression) для кожного кластеру, що забезпечило адаптивний підхід до прогнозування.

4. Використання методів машинного навчання, таких як XGBoost, Дерево рішень та Випадковий ліс, для прогнозування показників цифрової економіки з урахуванням специфіки кожного кластеру.

Особливе значення має порівняння метрик моделей, таких як MSE та R^2, для кожного кластеру, що дозволяє визначити найефективніший підхід до прогнозування.

Побудова концептуальної моделі для прогнозування розвитку цифрової економіки базується на використанні комплексного підходу, який поєднує традиційні статистичні методи, машинне навчання та аналіз багатовимірних даних. Подальший розвиток цієї моделі потребує вдосконалення методів роботи з даними та розширення доступності інформації для глибшого аналізу.

РОЗДІЛ 2. ПОПЕРЕДНЯ ОБРОБКА ДАНИХ ТА ВИБІР МЕТОДУ ПРОГНОЗУВАННЯ

2.1 Проведення статистичного аналізу вхідного масиву дослідження

Статистичний аналіз є ключовим етапом підготовки даних для побудови прогнозних моделей, особливо у дослідженнях, що базуються на панельних даних. Мета цього етапу полягає в оцінці структури та властивостей вхідного масиву, виявлення проблем, таких як пропущені значення чи мультиколінеарність, а також забезпечення готовності даних до моделювання, які можуть вплинути на якість моделювання.

У межах цього дослідження використовується набір панельних даних, що охоплює період з 2003 до 2022 року [35]. Ряд показників відображають цифрову економіку та її вплив на економічний розвиток країн. Дані за 2008 – 2022 роки представлені з інтервалом у два роки, що дозволяє відстежувати динаміку показників у середньостроковій перспективі. До ключових змінних належать:

- GDP (ВВП) — основний економічний показник, що характеризує рівень розвитку країни.
- E-Government Index — індекс розвитку електронного уряду.
- E-Participation Index — індекс електронної участі громадян.
- Online Service Index — індекс розвитку онлайн-сервісів.
- Human Capital Index — індекс розвитку людського капіталу.
- Telecommunication Infrastructure Index — індекс телекомунікаційної інфраструктури.
- Individuals using the Internet — частка населення, що використовує інтернет.

Перше завдання при роботі з вхідними даними — це перевірка наявності пропущених значень. Для кращого представлення візуалізуємо результати, що дозволить наочно підтвердити відсутність пропусків у даних (рисунок 2.1).

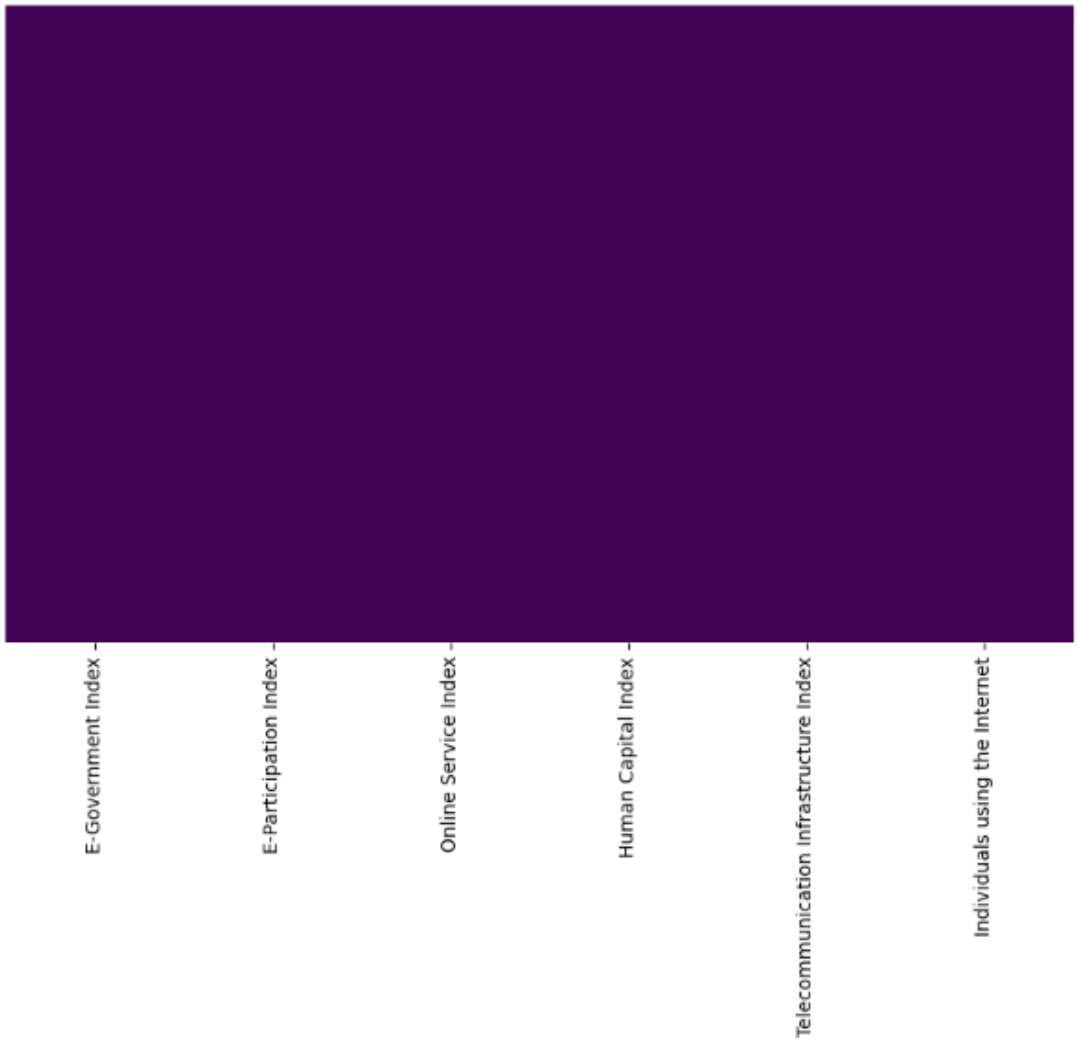


Рисунок 2.1 — Візуалізація відсутніх значень

Перевірка показала, що всі стовпці в наборі даних не містять пропущених значень, оскільки для кожного з них отримано нулі. Це свідчить про високу якість даних, оскільки всі значення є повними та валідними. Відсутність пропусків дозволяє проводити аналіз без потреби застосовувати методи обробки пропусків, що підвищує надійність результатів.

Наступним етапом було проведення описової статистики, яка включала розрахунок середнього значення, медіани, стандартного відхилення та

екстремальних значень для кожного показника (рисунок 2.2). Ці характеристики дозволяють отримати базове уявлення про структуру даних, оцінити їх варіативність і наявність аномальних значень.

	GDP	E-Government Index	E-Participation Index	Online Service Index	Human Capital Index	Telecommunication Infrastructure Index	Individuals using the Internet
count	1386.000000	1386.000000	1386.000000	1386.000000	1386.000000	1386.000000	1386.000000
mean	20280.967327	0.529796	0.381217	0.479331	0.748909	0.369185	42.811920
std	21053.282496	0.215610	0.298779	0.265237	0.199223	0.278849	32.084228
min	497.312058	0.013870	0.000415	0.004360	0.096000	0.001540	0.031011
25%	4815.494214	0.356507	0.105300	0.263978	0.648432	0.111493	10.524500
50%	12781.220454	0.524105	0.312595	0.475445	0.803050	0.321250	39.862300
75%	30560.767623	0.711575	0.623600	0.692003	0.900000	0.613165	72.772300
max	146457.020544	0.975800	1.000000	1.000000	1.330160	0.997900	100.000000

Рисунок 2.2 — Описова статистика

Загалом було проаналізовано 1386 країн, що забезпечує репрезентативність і різноманітність. Середнє значення, наприклад, індексу електронного урядування дорівнює 0.529, що свідчить про середній рівень розвитку в більшості країн. Стандартне відхилення демонструє ступінь варіації: чим воно більше, тим більша різниця між країнами. Мінімальні та максимальні значення, як-от ВВП від 497 до 146457, відображають значний діапазон рівнів економічного розвитку. Квартилі допомагають глибше зрозуміти розподіл даних: наприклад, 25% країн мають індекс електронного урядування нижче 0.356, тоді як 75% — нижче 0.711. Це дозволяє ідентифікувати групи країн із різним рівнем розвитку для подальшого аналізу.

Далі було обчислено кореляційну матрицю для всіх числових показників (рисунок 2.3). Це дозволило оцінити взаємозв'язок між змінними. Аналіз коефіцієнтів кореляції дозволяє виявити сильні (близькі до 1 або -1) та слабкі (близькі до 0) зв'язки між показниками, що є ключовим для розуміння можливих залежностей між ними [36]. Для наочності результатів було виконано візуалізацію кореляційної матриці у вигляді теплової карти (heatmap). Ця візуалізація дозволила швидко виявити пари змінних з високим або низьким рівнем кореляції, що є важливим для подальшого аналізу мультиколінеарності.

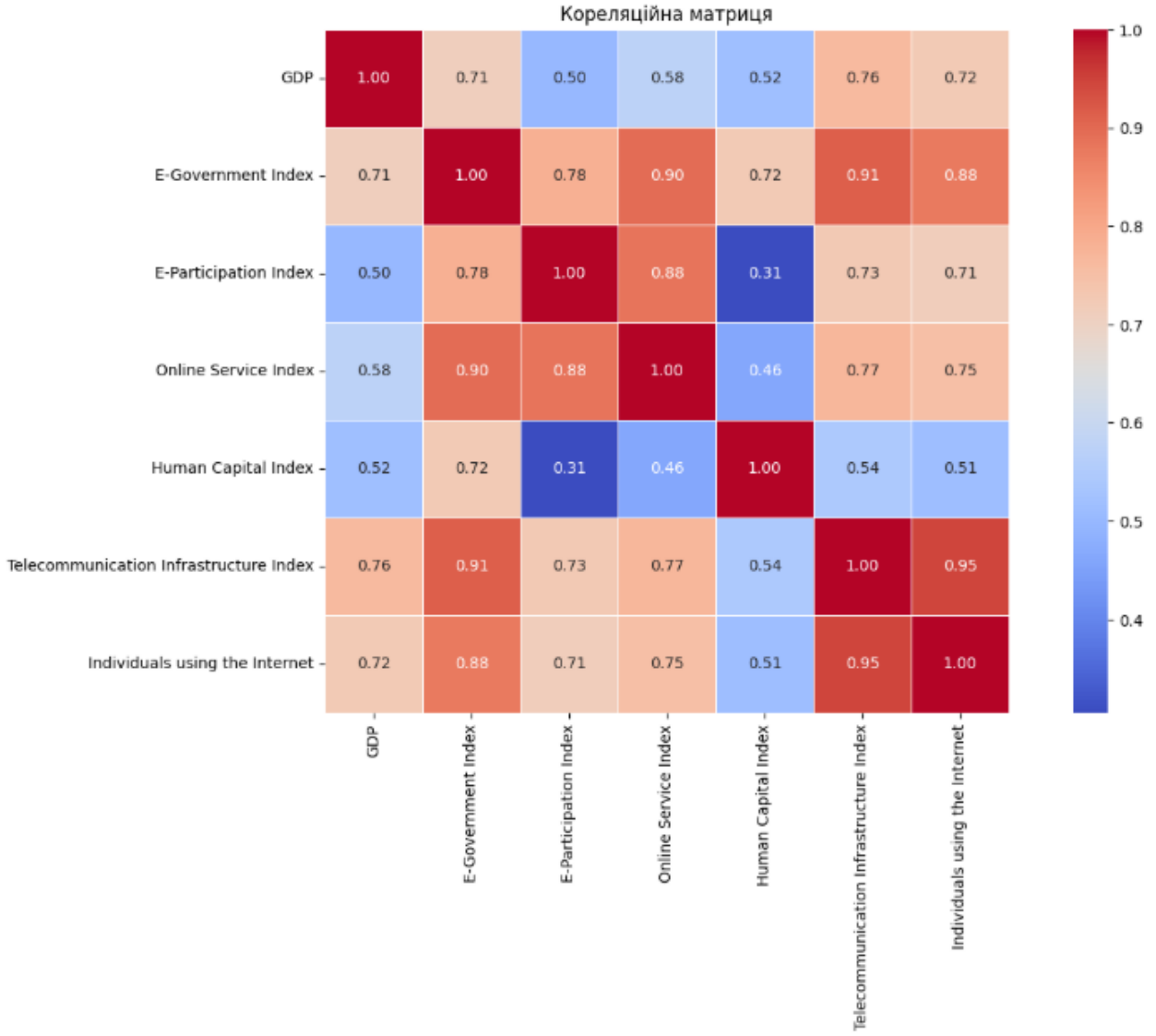


Рисунок 2.3 — Обчислення та візуалізація кореляційної матриці

Кореляційна матриця показала сильні позитивні взаємозв'язки між ВВП та індексами електронного урядування, онлайн-сервісів, телекомунікаційної інфраструктури і кількістю користувачів Інтернету. Це свідчить, що країни з вищим ВВП мають кращий рівень цифрової інфраструктури та електронного врядування. Висока кореляція між індексами електронного урядування і телекомунікаційною інфраструктурою вказує на їх паралельний розвиток.

Менш виражена кореляція між ВВП та індексом людського капіталу може свідчити про те, що економічне зростання не завжди супроводжується

пропорційним розвитком людського потенціалу. Загалом, результати підтверджують, що цифровий розвиток і електронне урядування є багатофакторними процесами, що залежать від економічного рівня, інфраструктури та людського капіталу.

Наступним кроком було проведено перевірку на мультиколінеарність. Для цього обчислювався коефіцієнт дисперсії інфляції (VIF) для кожної змінної (рисунок 2.4). Значення VIF, які перевищують критичне значення (зазвичай 10), свідчать про наявність мультиколінеарності, що може негативно впливати на регресійні моделі.

index	feature	VIF
0	const	17.03648
1	E-Government Index	72.2103
2	E-Participation Index	5.24522
3	Online Service Index	19.01204
4	Human Capital Index	8.01595
5	Telecommunication Infrastructure Index	21.43529
6	Individuals using the Internet	9.76013

Рисунок 2.4 — Перевірка на мультиколінеарність

Аналіз таблиці VIF показав, що деякі змінні мають значення VIF, які значно перевищують критичне значення 10, що вказує на наявність сильної мультиколінеарності. Це означає, що змінні, такі як E-Government Index, Online Service Index, та Telecommunication Infrastructure Index, сильно взаємопов’язані.

Оскільки змінні мають різні одиниці виміру та масштаби, їх стандартизація була необхідною для забезпечення коректності порівняння та аналізу (рисунок 2.5). Стандартизовані дані дозволяють уникнути впливу масштабів змінних на результати моделей, що є особливо важливим при використанні методів машинного навчання, таких як PCA або XGBoost.

	GDP	E-Government Index	E-Participation Index	Online Service Index	Human Capital Index	Telecommunication Infrastructure Index	Individuals using the Internet
0	-0.91754	-1.91283	-1.16086	-1.49495	-2.41479	-1.31870	-1.33210
1	-0.72607	-1.01612	-1.21879	-1.49495	0.25654	-1.14722	-1.30454
2	-0.44844	-0.74173	-1.10328	-0.35853	-0.29580	-1.19701	-1.26639
3	0.87455	-1.65004	-1.16086	-1.00083	-0.28844	-0.21760	-0.91248
4	-0.78121	-1.56624	-1.04535	-1.01731	-1.95283	-1.29904	-1.32328
...	...	...	...	...	...	...	...
1381	-0.53401	0.91264	0.77806	0.99822	0.14507	1.03432	1.28109
1382	-0.81144	-0.14381	-0.13499	-0.21321	-0.74320	0.37136	0.84424
1383	-0.29601	0.69087	0.51188	0.63766	-0.29429	1.17710	1.11553
1384	-0.78002	-0.12804	-0.02082	-0.14306	-0.37413	0.07790	-0.36098
1385	-0.78972	-0.26955	-0.51534	-0.35766	-0.51523	0.05423	-0.31960

Рисунок 2.5 — Стандартизація даних

Всі змінні приведені до однакового масштабу. Стандартизовані дані готові для побудови моделей машинного навчання.

Для перевірки нормальності розподілу даних було застосовано тест Шапіро-Уілка [37], який є одним із найбільш розповсюджених і потужних методів для оцінки того, чи відповідають дані нормальному розподілу (рисунок 2.6). Цей тест перевіряє нульову гіпотезу про те, що дані мають нормальний розподіл. Якщо p-значення тесту менше 0.05, це вказує на відхилення даних від нормальності, і ми можемо відкинути нульову гіпотезу.

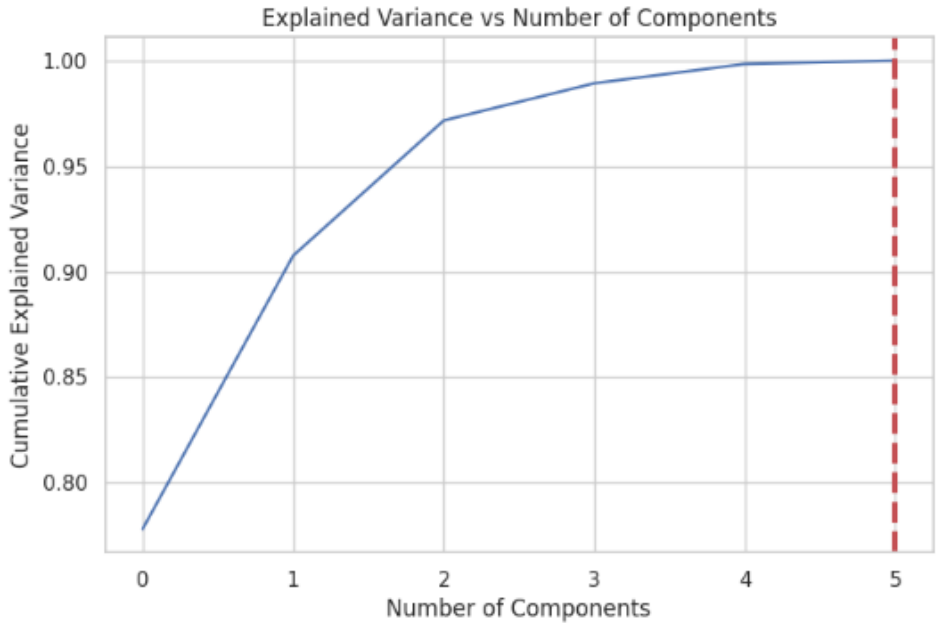
```
Shapiro-Wilk test for GDP: stat=0.8123465441363971, p_value=2.2136822189242006e-37
Shapiro-Wilk test for E-Government Index: stat=0.9773774908154986, p_value=6.088767351619679e-14
Shapiro-Wilk test for E-Participation Index: stat=0.915419602697971, p_value=3.953870957311857e-27
Shapiro-Wilk test for Online Service Index: stat=0.9698553632485623, p_value=2.0593188689293138e-16
Shapiro-Wilk test for Human Capital Index: stat=0.9126054793897814, p_value=1.6213931972204583e-27
Shapiro-Wilk test for Telecommunication Infrastructure Index: stat=0.9273760911011418, p_value=2.298697283781847e-25
Shapiro-Wilk test for Individuals using the Internet: stat=0.9106958144284503, p_value=8.966491941780137e-28
```

Рисунок 2.6 — Тест Шапіро-Вілка

У всіх випадках p-значення значно менше за 0.05, що дозволяє відкинути гіпотезу про нормальний розподіл даних для кожного з цих показників. Це означає, що для подальшого аналізу варто розглядати застосування методів, які не вимагають нормальності.

Для зменшення розмірності даних була застосована методологія аналізу головних компонент (PCA) із п’ятьма компонентами (рисунок 2.7). Це

дозволяє виділити найбільш значущі фактори, що пояснюють варіативність даних.



	Cumulative Variance Ratio	Explained Variance Ratio
0	0.77815	0.77815
1	0.90763	0.12948
2	0.97163	0.06399
3	0.98924	0.01762
4	0.99840	0.00915
5	1.00000	0.00160

Рисунок 2.7 — Створення PCA з 5 компонентами

Отриманий графік демонструє ефективність використання методу головних компонент (PCA) для зменшення розмірності даних. Як бачимо, вже перші три головні компоненти пояснюють понад 97% загальної дисперсії даних. Це свідчить про те, що основну інформацію про дані можна уявити у вигляді лише трьох нових змінних. Додавання четвертого та п'ятого компонентів призводить до незначного збільшення пояснюваної дисперсії, що вказує на те, що вони несуть менш значущу інформацію.

Отже, наступним кроком створимо об'єкт PCA з трьома компонентами (рисунок 2.8).

	principal component 1	principal component 2	principal component 3
0	-3.82221	-1.28199	0.56428
1	-2.55004	1.08457	-0.20968
2	-2.07069	0.31531	-0.59757
3	-2.20021	0.37755	0.45612
4	-3.26659	-1.02898	0.15936

Рисунок 2.8 — Створюємо об'єкт PCA з 3-ма компонентами

Результати показують, що основні компоненти ефективно зменшують розмірність даних, зберігаючи основну варіативність, що дозволяє спростити подальший аналіз, та зберегти суттєві тренди для кластеризації та побудови моделей. Principal Component 1 показує, наскільки кожне спостереження пов'язане з основною варіацією даних. Значення Principal Component 2 пояснює другу за важливістю варіацію в даних. Наприклад, спостереження 1 має значення 1.08457, що свідчить про значний внесок у цю компоненту. Значення Principal Component 3 пояснює залишкову варіацію, важливу для розуміння структури даних, але менш значну порівняно з PC1 та PC2. Наприклад, спостереження 2 має значення -0.59757.

Після виконання аналізу головних компонент (PCA) з трьома компонентами, до даних додаємо стовпці “Year” і “Country Name”. Це дозволяє зберегти початкову інформацію про часовий період та географічну приналежність кожного спостереження в контексті панельних даних. Далі згрупуємо дані за країнами (рисунок 2.9). Цей підхід допомагає врахувати особливості кожної країни при проведенні кластерного аналізу та регресійного моделювання, що підвищує точність і змістовність отриманих результатів.

Country Name	principal component 1	principal component 2	principal component 3	Year	GDP
Afghanistan	-2.94210	-1.45785	0.33586	2012.00000	1787.74306
Albania	-0.19072	0.27733	-0.16372	2012.00000	10494.28974
Algeria	-1.62632	0.35798	0.42195	2012.00000	13506.48781
Andorra	0.30277	0.33138	1.18751	2012.00000	49951.67813
Angola	-2.20387	-0.81103	-0.03571	2012.00000	6367.58923
...	...	...	...	...	...
Uzbekistan	-0.33429	0.32491	-0.54983	2012.00000	5769.11085
Vanuatu	-2.17262	-0.24502	0.43070	2012.00000	2661.77556
Viet Nam	-0.34976	0.00218	-0.04440	2012.00000	7162.78779
Zambia	-2.01765	-0.18145	-0.21533	2012.00000	2989.01832
Zimbabwe	-1.83133	0.23348	-0.06326	2012.00000	2286.42385

126 rows × 5 columns

Рисунок 2.9 — Групування за країною

Перша головна компонента, яка пояснює найбільшу частку дисперсії, виділяє країни з високим ВВП і розвиненими інфраструктурними показниками, такими як Andorra. Друга компонента акцентує на специфічних відмінностях, зокрема рівнях розвитку онлайн-сервісів і телекомунікацій. Третя головна компонента враховує відносно менш помітні, але важливі аспекти, як-от особливості людського капіталу.

Перед тим, як виконувати кластеризацію, необхідно визначити оптимальну кількість кластерів. Один із найпоширеніших способів — це метод ліктя (рисунок 2.10) [38]. Він полягає у побудові графіка, на якому показується залежність індексу якості кластеризації від кількості кластерів. Оптимальна кількість кластерів визначається на основі точки "ліктя", коли зменшення інерції при додаванні нових кластерів значно сповільнюється. Цей метод дозволяє вибрати таку кількість кластерів, яка мінімізує складність моделі, одночасно зберігаючи її високу якість.

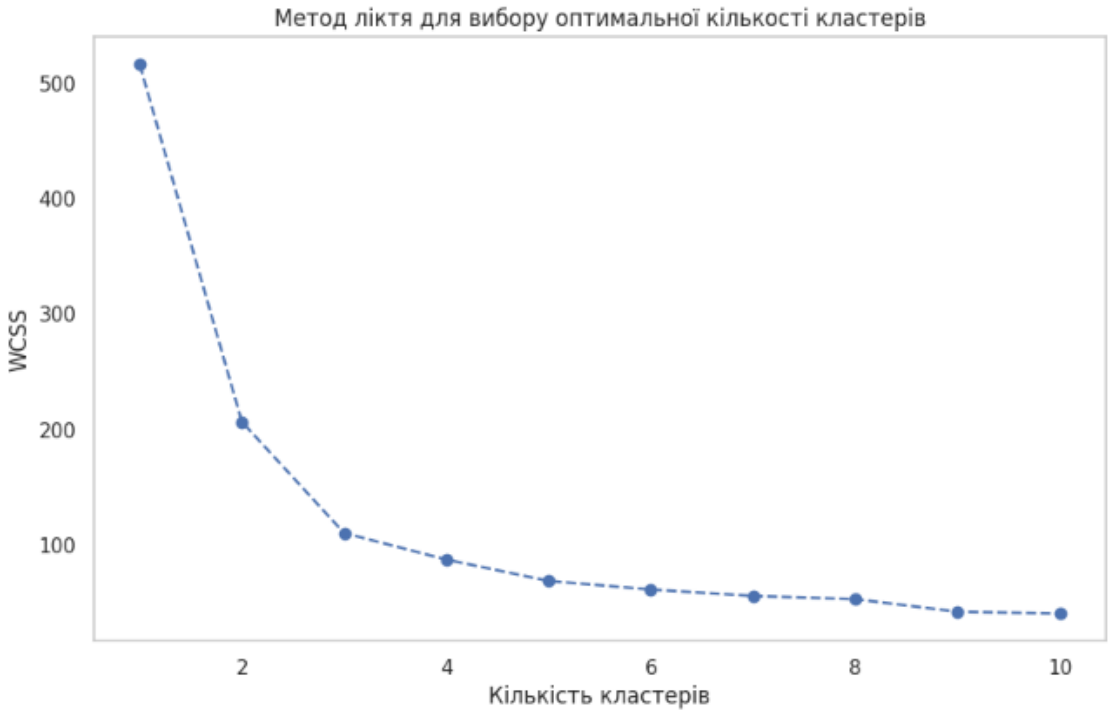


Рисунок 2.10 — Метод ліктя для вибору оптимальної кількості кластерів

На основі отриманого графіка можемо визначити, що оптимальною кількістю кластерів є три. Це підтверджується значним зламом у графіку при переході від двох до трьох кластерів, після чого подальше зменшення стає менш вираженим. Кількість кластерів три дозволяє отримати хорошу кластеризацію, де кожен кластер має внутрішню однорідність і чіткі відмінності між групами.

Для побудови двовимірної візуалізації кластерів застосовано метод головних компонентів (PCA), який зменшує кількість вимірів даних до двох, водночас зберігаючи фактичні характеристики їх варіації (рисунок 2.11). Це дозволяє наочно побачити структуру кластерів, де кожен об'єкт відображається точкою, а колір чи маркер позначає кластер. Така візуалізація допомагає оцінити ефективність кластеризації та виявити природні групи в даних.

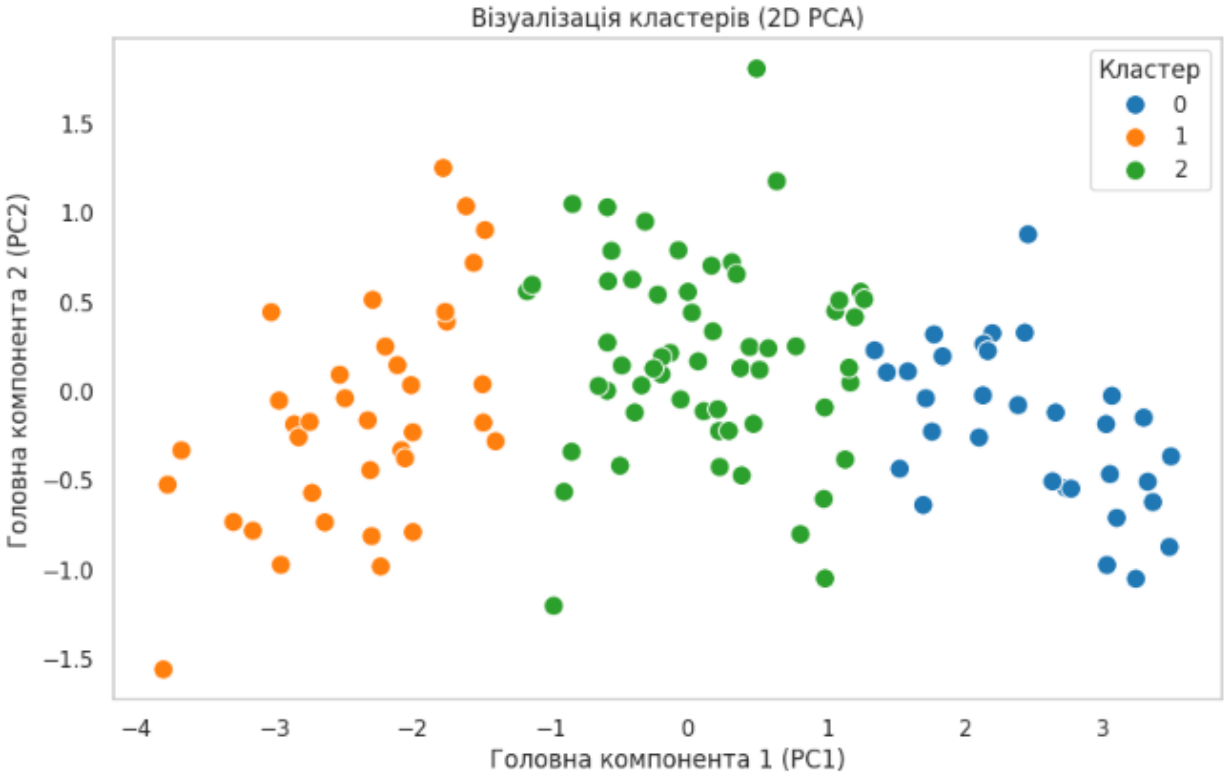


Рисунок 2.11 — Візуалізація кластерів (2D PCA)

Візуалізація підтверджує ефективність кластеризації. Графік з трьома чітко відокремленими групами даних демонструє, що обраний параметр кількості кластерів є оптимальним, оскільки на ньому немає значних перетинів. Кожен кластер утворює окрему область, що свідчить про однорідність характеристик об'єктів усередині кожної групи. Це також узгоджується з результатами методу ліктя, де точка перегину на кривій WCSS співпадає з числом кластерів, рівним трьом.

Наступним етапом інтегруємо результати кластеризації з основним набором даних (рисунок 2.12). Основною метою є додавання до DataFrame стовпця "Cluster", який відображатиме належність кожної країни до одного з кластерів, визначених за допомогою алгоритму k-means. Це важливий крок, який забезпечує зв'язок між результатами кластеризації та оригінальними даними, дозволяючи подальший аналіз з урахуванням кластерів.

	principal component 1	principal component 2	principal component 3	Year	Country Name	GDP	Cluster
0	-3.82221	-1.28199	0.56428	2003	Afghanistan	970.71623	1
1	-2.55004	1.08457	-0.20968	2003	Albania	5000.30908	2
2	-2.07069	0.31531	-0.59757	2003	Algeria	10843.16846	1
3	-2.20021	0.37755	0.45612	2003	Andorra	38686.47947	2
4	-3.26659	-1.02898	0.15936	2003	Angola	3839.85414	1
...	...	...	...	...	...	...	...
1381	2.18124	-0.42730	0.40441	2022	Uzbekistan	9042.34392	2
1382	0.08954	-0.54719	1.05046	2022	Vanuatu	3203.61662	1
1383	1.68333	-0.59945	0.78864	2022	Viet Nam	14051.24877	2
1384	-0.35878	-0.27080	0.02682	2022	Zambia	3864.89437	1
1385	-0.74315	-0.11911	0.37162	2022	Zimbabwe	3660.83550	1

1386 rows × 7 columns

Рисунок 2.12 — Присвоєння значення кластера країнам у новому стовпці “Cluster” в основному DataFrame

Присвоєння значення кластера країнам у новому стовпці "Cluster" дозволило чітко визначити кластерну належність кожної країни на основі результатів PCA та кластеризації. Наприклад, Афганістан у 2003 році належить до кластера 1, а Албанія та Андорра до кластера 2. Усього кластери поділили країни на групи з подібними характеристиками, такими як GDP, що надає основу для подальшого аналізу їхніх соціально-економічних особливостей. Додавання цього стовпця полегшує інтерпретацію та аналіз результатів, дозволяючи порівнювати країни між собою у межах кластерів або досліджувати зміни кластерної належності з часом. Це сприяє більш глибокому розумінню структури даних та характеристик кожного кластеру.

Отже, проведений статистичного аналізу вхідного масиву даних підтвердив відсутність пропущених значень, що спрощує подальшу обробку та підвищує достовірність результатів. Дослідження кореляцій показало сильний позитивний зв’язок між економічним розвитком і цифровими показниками, такими як індекси електронного врядування та телекомунікацій. Наявність мультиколінеарності потребувала застосування методів зменшення розмірності, зокрема PCA, що ефективно зменшило кількість змінних,

зберігаючи суттєву варіативність даних. Поділ на три кластери дозволив виділити групи країн із подібними соціально-економічними характеристиками, дозволяючи здійснювати подальший аналіз їхніх соціально-економічних особливостей.

2.2 Реалізація регресійного аналізу

Метою регресійного аналізу є оцінка впливу різних економічних та соціальних індикаторів на розвиток цифрової економіки. Регресійний аналіз є потужним статистичним методом, що дозволяє встановити взаємозв'язки між залежними та незалежними змінними [39].

При побудові регресійних моделей, особливо для даних панельного типу, важливим кроком є створення дамі-змінних для категоріальних змінних. Дамі-змінні дозволяють представити категоріальні змінні в числовій формі, що необхідно для аналізу в статистичних моделях, таких як регресія. Це особливо актуально, коли категоріальні змінні, такі як країни, регіони, або типи економіки, мають суттєвий вплив на результати моделей.

Створюються дамі-змінні за наступним принципом, для кожної категорії в змінній створюється окремий стовпець, де значення 1 вказує на приналежність до цієї категорії, а 0 — на відсутність приналежності. Наприклад, якщо є змінна «країна», то для кожної країни створюється окрема дамі-змінна. Для цього, якщо в спостереженні є значення для країни X, то в стовпці «країна_X» буде стояти 1, а в інших стовпцях — 0. Якщо змінна має n категорій, створюється n-1 дамі-змінних, оскільки одна з категорій повинна бути базовою для порівняння. Це дозволяє уникнути проблеми мультиколінеарності, яка може виникнути, якщо в модель будуть включені всі категорії.

Тепер створюємо дамі-змінні для cluster_0, cluster_1, cluster_2 (рисунок 2.13 – рисунок 2.15).

	principal component 1	principal component 2	principal component 3	Year	Country Name	GDP	Cluster	Country_Australia	Country_Austria	Country_Belgium	...
7	2.38402	0.48311	-0.70715	2003	Australia	30121.81842	0	1	0	0	...
8	0.64422	1.35642	0.31985	2003	Austria	32146.01529	0	0	1	0	...
13	0.90683	1.19708	0.02433	2003	Belgium	30934.58265	0	0	0	1	...
24	2.73412	0.12668	-0.56065	2003	Canada	32351.54570	0	0	0	0	...
27	1.30808	-0.31725	-1.90761	2003	Chile	10820.67162	0	0	0	0	...
...	...	...	...	...	...	...	...	...	...	...	...
1363	3.56902	-0.12845	0.39192	2022	Spain	48685.49631	0	0	0	0	...
1366	3.93028	0.13902	0.36842	2022	Sweden	68088.19327	0	0	0	0	...
1367	3.47540	0.04333	0.72689	2022	Switzerland	90746.45328	0	0	0	0	...
1377	3.80482	-0.37295	0.49607	2022	United Arab Emirates	78915.25476	0	0	0	0	...
1378	4.04985	-0.34064	0.09634	2022	United Kingdom of Great Britain and Northern I...	56761.51730	0	0	0	0	...

352 rows × 39 columns

Рисунок 2.13 — Створення дамі-змінних для cluster_0

	principal component 1	principal component 2	principal component 3	Year	Country Name	GDP	Cluster	Country_Afghanistan	Country_Algeria	Country_Angola	...
0	-3.82221	-1.28199	0.56428	2003	Afghanistan	970.71623	1	1	0	0	...
2	-2.07069	0.31531	-0.59757	2003	Algeria	10843.16846	1	0	1	0	...
4	-3.26659	-1.02898	0.15936	2003	Angola	3839.85414	1	0	0	1	...
11	-3.52597	-0.68237	0.35582	2003	Bangladesh	1803.27026	1	0	0	0	...
15	-3.01459	-0.86170	0.00860	2003	Benin	1902.64231	1	0	0	0	...
...	...	...	...	...	...	...	...	...	...	...	...
1375	-1.00925	-0.87478	-0.59937	2022	Uganda	2920.66477	1	0	0	0	...
1379	-1.05377	-0.87384	0.19434	2022	United Republic of Tanzania	3751.16675	1	0	0	0	...
1382	0.08954	-0.54719	1.05046	2022	Vanuatu	3203.61662	1	0	0	0	...
1384	-0.35878	-0.27080	0.02682	2022	Zambia	3864.89437	1	0	0	0	...
1385	-0.74315	-0.11911	0.37162	2022	Zimbabwe	3660.83550	1	0	0	0	...

396 rows × 43 columns

Рисунок 2.14 — Створення дамі-змінних для cluster_1

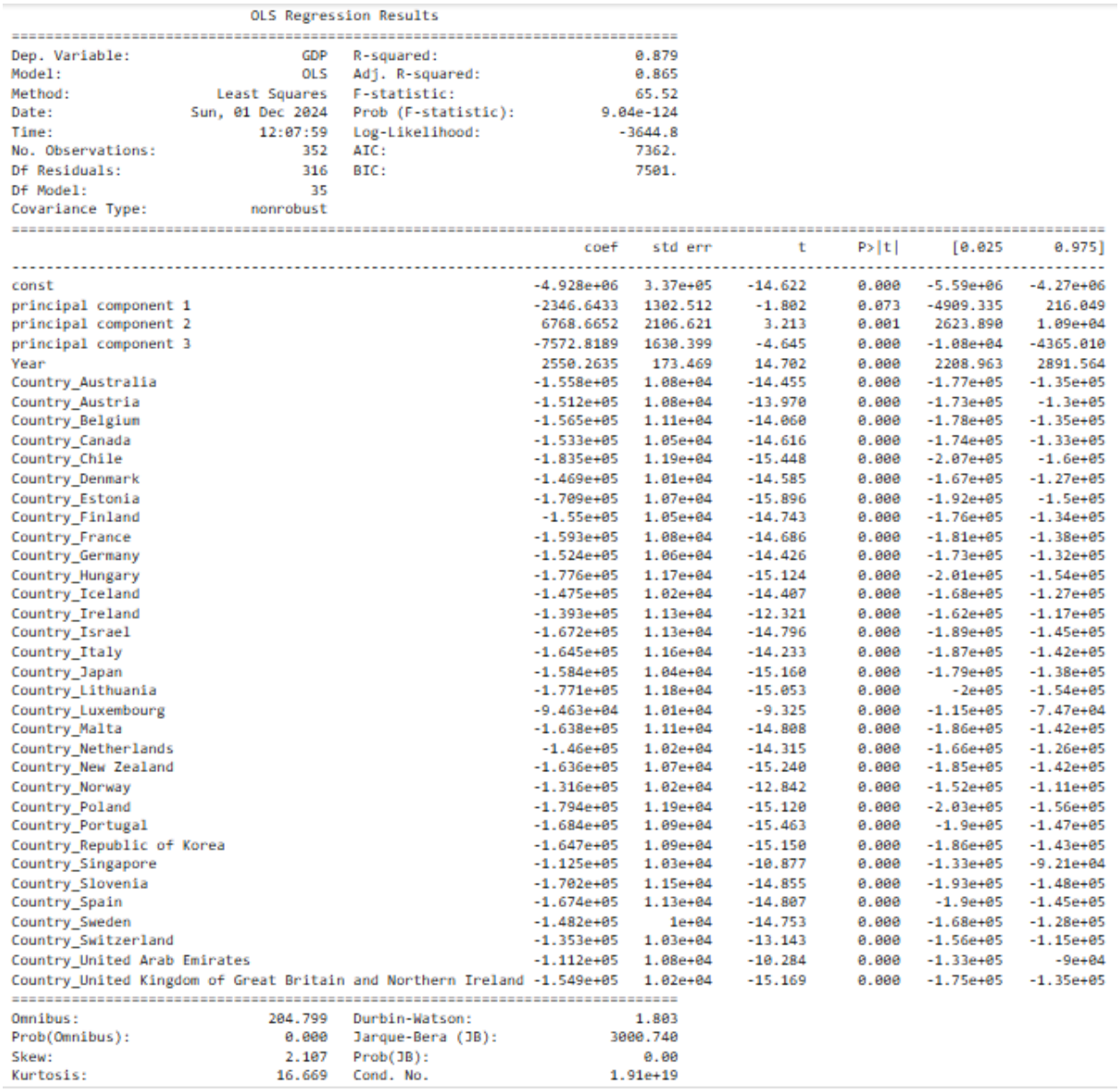
	principal component 1	principal component 2	principal component 3	Year	Country Name	GDP	Cluster	Country_Albania	Country_Andorra	Country_Argentina	...
1	-2.55004	1.08457	-0.20968	2003	Albania	5000.30908	2	1	0	0	...
3	-2.20021	0.37755	0.45612	2003	Andorra	38686.47947	2	0	1	0	...
5	0.16341	0.31226	-1.65109	2003	Argentina	10975.12824	2	0	0	1	...
6	-2.01374	1.51299	-0.45413	2003	Armenia	3958.62000	2	0	0	0	...
9	-2.16317	1.40194	-0.33546	2003	Azerbaijan	4640.13437	2	0	0	0	...
...	...	...	...	...	...	...	...	...	...	...	...
1373	2.88587	-0.37063	-0.17044	2022	Turkiye	38355.15397	2	0	0	0	...
1376	2.73103	-0.06331	0.32468	2022	Ukraine	16080.22852	2	0	0	0	...
1380	2.98760	0.15107	0.62731	2022	Uruguay	32746.31528	2	0	0	0	...
1381	2.18124	-0.42730	0.40441	2022	Uzbekistan	9042.34392	2	0	0	0	...
1383	1.68333	-0.59945	0.78864	2022	Viet Nam	14051.24877	2	0	0	0	...

638 rows × 65 columns

Рисунок 2.15 — Створення дамі-змінних для cluster_2

Побудовані таблиці включають основні компоненти (Principal Components 1, 2, 3), Year, Country Name, GDP, а також дамі-змінні для кожної країни. Кожен рядок представляє собою спостереження для конкретної країни в конкретному році, що дозволяє включити категоріальні змінні для кожного кластеру в подальший аналіз.

Для реалізації регресійного аналізу використано метод найменших квадратів (OLS), який є стандартним для оцінки лінійних взаємозв'язків між змінними. Метод OLS дозволяє з мінімальною помилкою оцінити коефіцієнти регресії для кожної незалежної змінної, а також оцінити їх значущість для моделі [40]. Оцінка моделі OLS включає розрахунок статистики, таких як коефіцієнти регресії, t-статистика, p-значення, а також аналіз залишків для виявлення потенційних проблем з моделлю, таких як гетероскедастичність або автокореляція [41].



OLS Regression Results

Dep. Variable:	GDP	R-squared:	0.879
Model:	OLS	Adj. R-squared:	0.865
Method:	Least Squares	F-statistic:	65.52
Date:	Sun, 01 Dec 2024	Prob (F-statistic):	9.04e-124
Time:	12:07:59	Log-Likelihood:	-3644.8
No. Observations:	352	AIC:	7362.
Df Residuals:	316	BIC:	7501.
Df Model:	35		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-4.928e+06	3.37e+05	-14.622	0.000	-5.59e+06	-4.27e+06
principal component 1	-2346.6433	1302.512	-1.802	0.073	-4909.335	216.049
principal component 2	6768.6652	2106.621	3.213	0.001	2623.890	1.09e+04
principal component 3	-7572.8189	1630.399	-4.645	0.000	-1.08e+04	-4365.010
Year	2550.2635	173.469	14.702	0.000	2208.963	2891.564
Country_Australia	-1.558e+05	1.08e+04	-14.455	0.000	-1.77e+05	-1.35e+05
Country_Austria	-1.512e+05	1.08e+04	-13.970	0.000	-1.73e+05	-1.3e+05
Country_Belgium	-1.565e+05	1.11e+04	-14.060	0.000	-1.78e+05	-1.35e+05
Country_Canada	-1.533e+05	1.05e+04	-14.616	0.000	-1.74e+05	-1.33e+05
Country_Chile	-1.835e+05	1.19e+04	-15.448	0.000	-2.07e+05	-1.6e+05
Country_Denmark	-1.469e+05	1.01e+04	-14.585	0.000	-1.67e+05	-1.27e+05
Country_Estonia	-1.709e+05	1.07e+04	-15.896	0.000	-1.92e+05	-1.5e+05
Country_Finland	-1.55e+05	1.05e+04	-14.743	0.000	-1.76e+05	-1.34e+05
Country_France	-1.593e+05	1.08e+04	-14.686	0.000	-1.81e+05	-1.38e+05
Country_Germany	-1.524e+05	1.06e+04	-14.426	0.000	-1.73e+05	-1.32e+05
Country_Hungary	-1.776e+05	1.17e+04	-15.124	0.000	-2.01e+05	-1.54e+05
Country_Iceland	-1.475e+05	1.02e+04	-14.407	0.000	-1.68e+05	-1.27e+05
Country_Ireland	-1.393e+05	1.13e+04	-12.321	0.000	-1.62e+05	-1.17e+05
Country_Israel	-1.672e+05	1.13e+04	-14.796	0.000	-1.89e+05	-1.45e+05
Country_Italy	-1.645e+05	1.16e+04	-14.233	0.000	-1.87e+05	-1.42e+05
Country_Japan	-1.584e+05	1.04e+04	-15.160	0.000	-1.79e+05	-1.38e+05
Country_Lithuania	-1.771e+05	1.18e+04	-15.053	0.000	-2e+05	-1.54e+05
Country_Luxembourg	-9.463e+04	1.01e+04	-9.325	0.000	-1.15e+05	-7.47e+04
Country_Malta	-1.638e+05	1.11e+04	-14.808	0.000	-1.86e+05	-1.42e+05
Country_Netherlands	-1.46e+05	1.02e+04	-14.315	0.000	-1.66e+05	-1.26e+05
Country_New Zealand	-1.636e+05	1.07e+04	-15.240	0.000	-1.85e+05	-1.42e+05
Country_Norway	-1.316e+05	1.02e+04	-12.842	0.000	-1.52e+05	-1.11e+05
Country_Poland	-1.794e+05	1.19e+04	-15.120	0.000	-2.03e+05	-1.56e+05
Country_Portugal	-1.684e+05	1.09e+04	-15.463	0.000	-1.9e+05	-1.47e+05
Country_Republic of Korea	-1.647e+05	1.09e+04	-15.150	0.000	-1.86e+05	-1.43e+05
Country_Singapore	-1.125e+05	1.03e+04	-10.877	0.000	-1.33e+05	-9.21e+04
Country_Slovenia	-1.702e+05	1.15e+04	-14.855	0.000	-1.93e+05	-1.48e+05
Country_Spain	-1.674e+05	1.13e+04	-14.807	0.000	-1.9e+05	-1.45e+05
Country_Sweden	-1.482e+05	1e+04	-14.753	0.000	-1.68e+05	-1.28e+05
Country_Switzerland	-1.353e+05	1.03e+04	-13.143	0.000	-1.56e+05	-1.15e+05
Country_United Arab Emirates	-1.112e+05	1.08e+04	-10.284	0.000	-1.33e+05	-9e+04
Country_United Kingdom of Great Britain and Northern Ireland	-1.549e+05	1.02e+04	-15.169	0.000	-1.75e+05	-1.35e+05

Omnibus:	204.799	Durbin-Watson:	1.803
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3000.740
Skew:	2.107	Prob(JB):	0.00
Kurtosis:	16.669	Cond. No.	1.91e+19

Рисунок 2.16 — Регресійний аналіз для cluster_0

Результати регресійного аналізу для cluster_0 демонструють високу якість моделі, що підтверджується значенням коефіцієнта детермінації R^2 = 0.879, яке свідчить про те, що модель пояснює 87.9% варіації GDP. З головних компонент, значущими є друга (PC2, t = 3.213, p = 0.001) та третя (PC3, t = -4.645, p < 0.001), що вказує на структурний вплив специфічних економічних факторів. Однак, перша головна компонента (PC1) виявилась незначущою (t = -1.802, p = 0.073). Критерій F-статистика (F = 65.52, Prob < 0.001) підтверджує загальну значущість моделі. Із додаткових тестів бачимо, що тест Durbin-

Watson показує відсутність автокореляції, що є позитивним фактором для стабільності моделі, хоча залишки не відповідають нормальному розподілу.

OLS Regression Results

Dep. Variable:	GDP	R-squared:	0.935
Model:	OLS	Adj. R-squared:	0.928
Method:	Least Squares	F-statistic:	130.8
Date:	Sun, 01 Dec 2024	Prob (F-statistic):	1.20e-186
Time:	12:07:59	Log-Likelihood:	-3271.7
No. Observations:	396	AIC:	6623.
Df Residuals:	356	BIC:	6783.
Df Model:	39		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.217e+05	3.27e+04	-3.720	0.000	-1.86e+05	-5.74e+04
principal component 1	386.2564	113.161	3.413	0.001	163.707	608.805
principal component 2	-125.0498	132.783	-0.942	0.347	-386.188	136.088
principal component 3	707.0627	156.272	4.525	0.000	399.730	1014.396
Year	64.4169	16.643	3.871	0.000	31.687	97.147
Country_Afghanistan	-5411.1824	948.950	-5.702	0.000	-7277.435	-3544.930
Country_Algeria	5965.5341	945.780	6.308	0.000	4105.516	7825.552
Country_Angola	-772.8689	944.833	-0.818	0.414	-2631.024	1085.286
Country_Bangladesh	-3185.9970	920.865	-3.460	0.001	-4997.016	-1374.978
Country_Benin	-4574.5785	940.427	-4.864	0.000	-6424.070	-2725.087
Country_Burundi	-6202.3723	1006.108	-6.165	0.000	-8181.034	-4223.711
Country_Cameroon	-3816.0706	969.404	-3.937	0.000	-5722.548	-1909.593
Country_Central African Republic	-6185.3335	1001.581	-6.176	0.000	-8155.092	-4215.575
Country_Chad	-5511.1125	989.427	-5.570	0.000	-7456.969	-3565.256
Country_Comoros	-4399.2983	1006.825	-4.369	0.000	-6379.372	-2419.225
Country_Congo	-2040.5972	1004.387	-2.032	0.043	-4015.875	-65.319
Country_Ethiopia	-5645.6122	952.035	-5.930	0.000	-7517.931	-3773.293
Country_Gabon	7299.8986	945.303	7.722	0.000	5440.818	9158.980
Country_Gambia	-5363.0044	924.703	-5.800	0.000	-7181.572	-3544.437
Country_Ghana	-2920.7410	909.472	-3.211	0.001	-4709.355	-1132.127
Country_Iraq	3760.0883	962.902	3.905	0.000	1866.396	5653.780
Country_Kenya	-3742.7248	952.669	-3.929	0.000	-5616.292	-1869.158
Country_Liberia	-5834.5203	973.667	-5.992	0.000	-7749.382	-3919.659
Country_Madagascar	-5371.6558	1000.482	-5.369	0.000	-7339.254	-3404.057
Country_Malawi	-5542.6948	1002.234	-5.530	0.000	-7513.739	-3571.651
Country_Mali	-5536.0021	933.181	-5.932	0.000	-7371.242	-3700.762
Country_Mozambique	-5943.9684	956.652	-6.213	0.000	-7825.368	-4062.568
Country_Nicaragua	-2548.8594	941.212	-2.708	0.007	-4399.893	-697.825
Country_Nigeria	-2667.6729	928.174	-2.874	0.004	-4493.066	-842.279
Country_Pakistan	-2954.2187	915.215	-3.228	0.001	-4754.127	-1154.310
Country_Senegal	-4523.5712	884.373	-5.115	0.000	-6262.824	-2784.319
Country_Somalia	-6135.1108	957.994	-6.404	0.000	-8019.151	-4251.071
Country_Sudan	-3375.9570	951.589	-3.548	0.000	-5247.400	-1504.514
Country_Tajikistan	-4388.8688	1029.998	-4.261	0.000	-6414.515	-2363.223
Country_Togo	-5217.9450	966.604	-5.398	0.000	-7118.917	-3316.973
Country_Turkmenistan	4299.0675	1045.847	4.111	0.000	2242.253	6355.882
Country_Uganda	-4837.9854	986.744	-4.903	0.000	-6778.565	-2897.406
Country_United Republic of Tanzania	-4750.0579	962.553	-4.935	0.000	-6643.062	-2857.054
Country_Vanuatu	-4749.7629	941.417	-5.045	0.000	-6601.200	-2898.326
Country_Zambia	-4017.6419	980.543	-4.097	0.000	-5946.027	-2089.257
Country_Zimbabwe	-4847.8388	983.524	-4.929	0.000	-6782.086	-2913.591

Omnibus:	104.464	Durbin-Watson:	1.910
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4283.502
Skew:	0.083	Prob(JB):	0.00
Kurtosis:	19.111	Cond. No.	3.59e+19

Рисунок 2.17 — Регресійний аналіз для cluster_1

Результати регресійного аналізу для cluster_1 показують, що модель має високий коефіцієнт детермінації $R^2 = 0.935$, що свідчить її високу точність. Модель є статистично значущою (F-статистика = 130.8, $p < 0.001$). PC1 (коеф. = 386.26, $p = 0.001$) та PC3 (коеф. = 707.06, $p = 0.000$) вказують на її високу

точність, а PC2 не має значущого впливу на GDP (p = 0.347). Durbin-Watson тест (DW = 1.91) свідчить про відсутність серйозної автокореляції. Високе значення тесту Jarque-Bera (JB = 4283.5) вказує на наявність проблем із нормальністю залишків, проте інші статистичні показники підтверджують загальну ефективність моделі.

OLS Regression Results

Dep. Variable:	GDP	R-squared:	0.942
Model:	OLS	Adj. R-squared:	0.936
Method:	Least Squares	F-statistic:	154.6
Date:	Sun, 01 Dec 2024	Prob (F-statistic):	4.64e-317
Time:	12:07:59	Log-Likelihood:	-6086.2
No. Observations:	638	AIC:	1.230e+04
Df Residuals:	576	BIC:	1.257e+04
Df Model:	61		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.228e+06	1.78e+05	-6.884	0.000	-1.58e+06	-8.78e+05
principal component 1	1191.9777	345.019	3.455	0.001	514.330	1869.626
principal component 2	2272.2037	434.594	5.228	0.000	1418.621	3125.786
principal component 3	-1037.3511	518.015	-2.003	0.046	-2054.779	-19.923
Year	629.2218	90.182	6.977	0.000	452.097	806.347
Country_Albania	-2.78e+04	3356.268	-8.284	0.000	-3.44e+04	-2.12e+04
Country_Andorra	1.235e+04	2767.896	4.461	0.000	6909.960	1.78e+04
Country_Argentina	-2.164e+04	3206.456	-6.749	0.000	-2.79e+04	-1.53e+04
Country_Armenia	-2.877e+04	3403.582	-8.454	0.000	-3.55e+04	-2.21e+04
Country_Azerbaijan	-2.538e+04	3259.180	-7.788	0.000	-3.18e+04	-1.9e+04
Country_Bahamas	-8465.7286	3070.131	-2.757	0.006	-1.45e+04	-2435.711
Country_Barbados	-2.391e+04	2978.447	-8.027	0.000	-2.98e+04	-1.81e+04
Country_Belize	-2.706e+04	3498.552	-7.735	0.000	-3.39e+04	-2.02e+04
Country_Bolivia	-3.181e+04	3541.620	-8.981	0.000	-3.88e+04	-2.49e+04
Country_Bosnia and Herzegovina	-2.696e+04	3282.537	-8.214	0.000	-3.34e+04	-2.05e+04
Country_Botswana	-2.36e+04	3479.954	-6.782	0.000	-3.04e+04	-1.68e+04
Country_Brazil	-2.513e+04	3170.549	-7.926	0.000	-3.14e+04	-1.89e+04
Country_Brunei Darussalam	3.622e+04	3012.254	12.023	0.000	3.03e+04	4.21e+04
Country_Bulgaria	-2.188e+04	3143.711	-6.960	0.000	-2.81e+04	-1.57e+04
Country_China	-2.694e+04	3268.937	-8.240	0.000	-3.34e+04	-2.05e+04
Country_Colombia	-2.655e+04	3299.555	-8.048	0.000	-3.3e+04	-2.01e+04
Country_Costa Rica	-2.336e+04	3156.004	-7.402	0.000	-2.96e+04	-1.72e+04
Country_Cyprus	-4957.6502	2906.712	-1.706	0.089	-1.07e+04	751.397
Country_Czech Republic	-8172.8250	2933.094	-2.786	0.006	-1.39e+04	-2411.962
Country_Dominica	-2.642e+04	3103.540	-8.514	0.000	-3.25e+04	-2.03e+04
Country_Ecuador	-2.82e+04	3429.429	-8.224	0.000	-3.49e+04	-2.15e+04
Country_Egypt	-2.547e+04	3382.584	-7.530	0.000	-3.21e+04	-1.88e+04
Country_El Salvador	-2.969e+04	3463.737	-8.570	0.000	-3.65e+04	-2.29e+04
Country_Georgia	-2.853e+04	3418.068	-8.346	0.000	-3.52e+04	-2.18e+04
Country_Greece	-1.238e+04	3090.680	-4.006	0.000	-1.85e+04	-6312.104
Country_Grenada	-2.702e+04	3328.955	-8.118	0.000	-3.36e+04	-2.05e+04
Country_India	-3.026e+04	3627.614	-8.341	0.000	-3.74e+04	-2.31e+04
Country_Indonesia	-2.855e+04	3577.250	-7.982	0.000	-3.56e+04	-2.15e+04
Country_Jamaica	-2.902e+04	3347.347	-8.669	0.000	-3.56e+04	-2.24e+04
Country_Jordan	-2.915e+04	3357.028	-8.682	0.000	-3.57e+04	-2.26e+04
Country_Kazakhstan	-1.859e+04	3292.196	-5.647	0.000	-2.51e+04	-1.21e+04
Country_Latvia	-1.736e+04	2932.164	-5.919	0.000	-2.31e+04	-1.16e+04
Country_Malaysia	-1.597e+04	2883.392	-5.538	0.000	-2.16e+04	-1.03e+04
Country_Maldives	-2.312e+04	3333.568	-6.935	0.000	-2.97e+04	-1.66e+04
Country_Mexico	-2.157e+04	3309.384	-6.519	0.000	-2.81e+04	-1.51e+04
Country_Mongolia	-2.936e+04	3552.215	-8.265	0.000	-3.63e+04	-2.24e+04

Country_Morocco	-2.749e+04	3147.321	-8.736	0.000	-3.37e+04	-2.13e+04
Country_North Macedonia	-2.451e+04	3114.838	-7.870	0.000	-3.06e+04	-1.84e+04
Country_Oman	5711.8664	3075.556	1.857	0.064	-328.806	1.18e+04
Country_Panama	-1.739e+04	3285.899	-5.292	0.000	-2.38e+04	-1.09e+04
Country_Paraguay	-2.631e+04	3453.813	-7.617	0.000	-3.31e+04	-1.95e+04
Country_Peru	-2.863e+04	3393.733	-8.435	0.000	-3.53e+04	-2.2e+04
Country_Philippines	-3.247e+04	3513.636	-9.240	0.000	-3.94e+04	-2.56e+04
Country_Romania	-1.834e+04	3199.350	-5.733	0.000	-2.46e+04	-1.21e+04
Country_San Marino	1.854e+04	2976.676	6.228	0.000	1.27e+04	2.44e+04
Country_Serbia	-2.417e+04	3136.670	-7.704	0.000	-3.03e+04	-1.8e+04
Country_Seychelles	-1.548e+04	3151.399	-4.913	0.000	-2.17e+04	-9294.048
Country_Slovakia	-1.342e+04	2873.942	-4.669	0.000	-1.91e+04	-7773.903
Country_South Africa	-2.596e+04	3362.763	-7.720	0.000	-3.26e+04	-1.94e+04
Country_Sri Lanka	-2.835e+04	3712.186	-7.637	0.000	-3.56e+04	-2.11e+04
Country_Thailand	-2.402e+04	3346.684	-7.178	0.000	-3.06e+04	-1.74e+04
Country_Tonga	-3.366e+04	3661.418	-9.193	0.000	-4.08e+04	-2.65e+04
Country_Tunisia	-2.68e+04	3289.882	-8.147	0.000	-3.33e+04	-2.03e+04
Country_Turkiye	-1.756e+04	3227.839	-5.440	0.000	-2.39e+04	-1.12e+04
Country_Ukraine	-2.958e+04	3420.959	-8.647	0.000	-3.63e+04	-2.29e+04
Country_Uruguay	-2.083e+04	3090.245	-6.740	0.000	-2.69e+04	-1.48e+04
Country_Uzbekistan	-3.286e+04	3549.204	-9.260	0.000	-3.98e+04	-2.59e+04
Country_Viet Nam	-3.019e+04	3307.553	-9.129	0.000	-3.67e+04	-2.37e+04

Omnibus:	82.456	Durbin-Watson:	1.981
Prob(Omnibus):	0.000	Jarque-Bera (JB):	703.897
Skew:	0.168	Prob(JB):	1.42e-153
Kurtosis:	8.135	Cond. No.	4.47e+19

Рисунок 2.18 — Регресійний аналіз для cluster_2

Модель для cluster_2 є найефективнішою, з коефіцієнтом детермінації R^2 = 0.942, це означає, що 94.2% варіації GDP можна пояснити за допомогою змінних, включених у модель. Значення F-статистики (F = 154.6) та ймовірність її статистичної значущості свідчать про те, що модель значуща і добре описує залежність між змінними. Всі основні компоненти мають статистичну значущість, крім PC3 (коеф. = -1037.35, p = 0.046), що свідчить про слабший вплив у порівнянні з іншими компонентами. Високі значення F-статистики підтверджують, що модель добре описує залежність між змінними.

Для оцінки ефективності побудованих моделей використовувалися основні метрики якості, такі як R^2 (коефіцієнт детермінації) та MSE (середньоквадратична помилка). Величина R^2 дозволяє оцінити, яку частку варіації залежної змінної пояснюють незалежні змінні, а MSE визначає середнє квадратичне відхилення між прогнозованими та реальними значеннями [42].

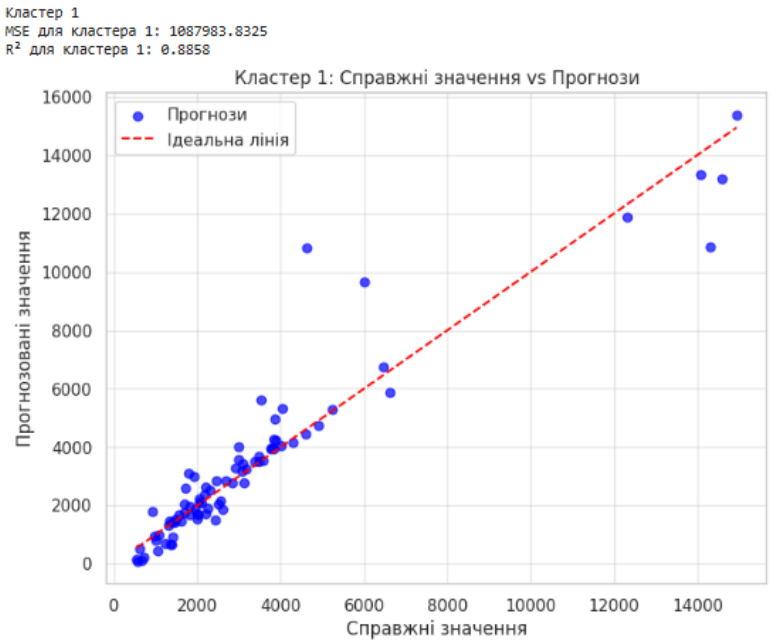


Рисунок 2.19 — Побудова регресії для Кластер 1

Модель, побудована для кластера 1, демонструє високу якість моделі, що вказує на добру відповідність прогнозів фактичним значенням, хоча значення MSE свідчить про можливість покращення точності в окремих спостереженнях. Графік прогнозу показує, що більшість точок розташовані поблизу ідеальної лінії, що свідчить про загальну адекватність моделі. Однак, спостерігається певний розкид точок навколо ідеальної лінії, що пояснюється значенням MSE. Це означає, що для деяких спостережень похибка прогнозу може бути досить значною.

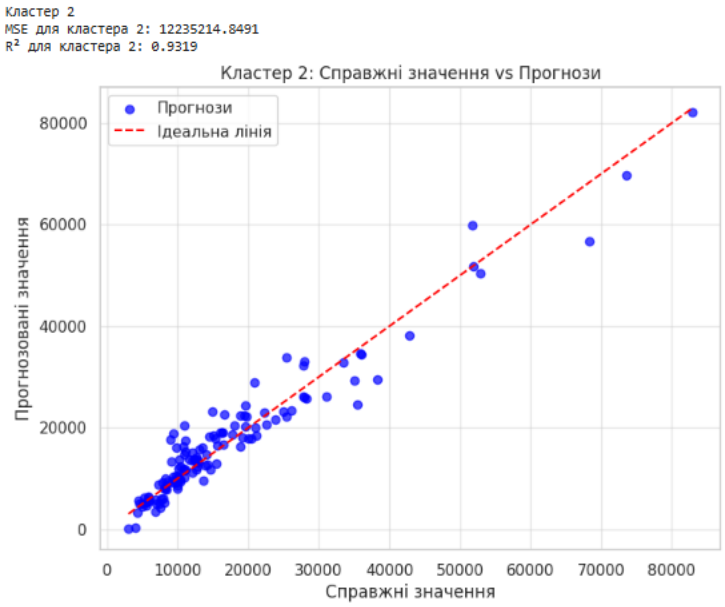


Рисунок 2.20 — Побудова регресії для Кластер 2

Результати регресійного аналізу для кластера 2 характеризується найвищою якістю моделі серед усіх, що свідчить про чудову відповідність даних. Незважаючи на це, середньоквадратична похибка (MSE) вказує на наявність певних відхилень прогнозів від фактичних значень. Графік прогнозу демонструє лінійну залежність між фактичними значеннями та прогнозами, однак спостерігається певний розкид точок навколо лінії, що пояснюється значенням MSE. Загалом, модель демонструє високу точність прогнозування для кластера 2.

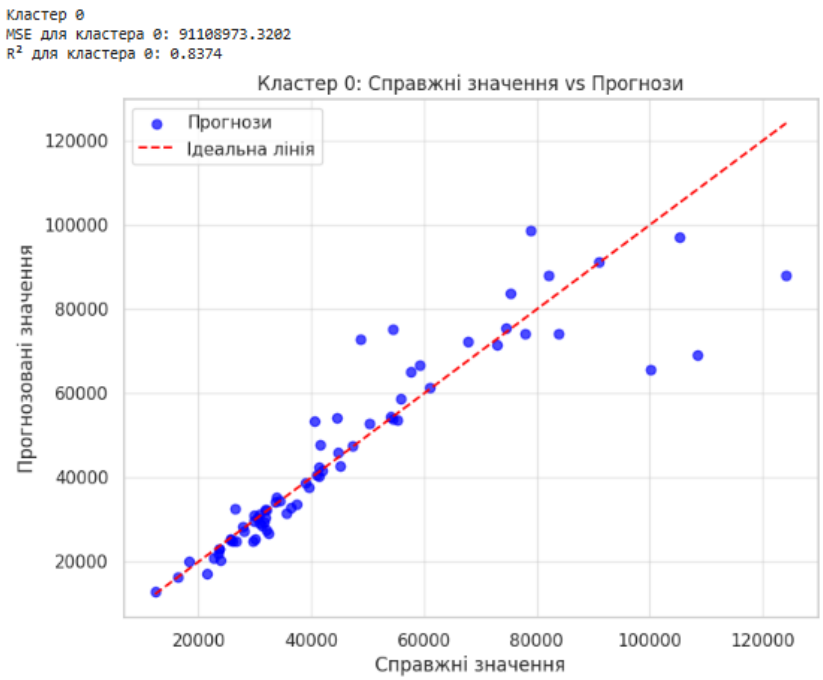


Рисунок 2.21 — Побудова регресії для Кластер 0

Кластер 0 має найнижчий серед аналізованих кластерів коефіцієнт детермінації, але свідчить про задовільну якість моделі. Значення MSE вказує на вищі відхилення прогнозів, що потребує особливої уваги при використанні цієї моделі для практичних прогнозів. Графік прогнозу показує лінійну залежність між фактичними значеннями та прогнозами, більшість точок близько до ідеальної лінії регресії, хоча спостерігається певний розкид точок навколо цієї лінії, що пояснюється значенням MSE.

Отже, всі побудовані моделі виявилися статистично значущими та пояснюють значну частину варіації залежної змінної. Однак, для підвищення точності прогнозування для кластерів з більшими відхиленнями (кластер 0) можуть знадобитися додаткові дослідження та використання більш складних моделей.

2.3 XGBoost, Дерево рішень та Випадковий ліс, як методи прогнозування розвитку цифрової економіки

Розвиток цифрової економіки є складним процесом, який залежить від великої кількості змінних, таких як розвиток інформаційних технологій, індекси електронного урядування, телекомунікаційна інфраструктура, індекс людського капіталу, а також економічні показники. Для аналізу та прогнозування цього процесу використовуються сучасні методи машинного навчання, серед яких XGBoost, Дерево рішень та Випадковий ліс, які є найпоширенішими за рахунок ефективності, адаптивності та здатності обробляти великі обсяги даних.

Прогнозування розвитку цифрової економіки вимагає застосування передових методів машинного навчання, здатних ефективно аналізувати великі обсяги даних та виробляти складні взаємозв'язки між абсолютно змінними. У цьому контексті алгоритми, такі як XGBoost, дерево рішень та випадковий ліс, визначають особливу важливість [43]. Кожен із цих методів має свої специфічні характеристики, переваги та обмеження, що дозволяють ефективно використовувати їх для детального аналізу та прогнозування економічних процесів у процесі цифрової трансформації.

XGBoost є одним із найпотужніших методів машинного навчання, заснованих на градієнтному бустингу. Цей алгоритм особливо добре працює зі структурованими даними і часто використовується для регресії та класифікації [44] і демонструє високу точність прогнозів .

Модель XGBoost має наступні особливості:

- Градієнтний бустинг — метод побудови послідовних дерев рішень, де кожне наступне дерево коригує помилки попередніх;
- Алгоритм оптимізовано для високої швидкодії, що дозволяє працювати з великими наборами даних.
- Завдяки вбудованим механізмам регуляризації, XGBoost зменшує ризик перенавчання, що є критично важливим для задач прогнозування у цифровій економіці [45].

- XGBoost дозволяє оцінювати важливість змінних, що допомагає визначити ключові фактори впливу на розвиток цифрової економіки, такі як рівень інтернет-проникнення, розвиток телекомунікаційної інфраструктури або людський капітал.

XGBoost використовує принцип більшого вдосконалення моделі за допомогою створення компонентів дерев, кожна з яких намагається компенсувати помилки попередніх. Такий підхід дозволяє значно зменшити залишкову похибку, що робить модель більш точною. Для роботи з великими наборами даних XGBoost підтримує паралельну обробку, що суттєво прискорює навчання [46]. Крім того, він може автоматично втратити пропущені значення, визначаючи найбільш оптимальний шлях розгалуження для кожного з таких випадків [47]. Ця особливість важлива в умовах економічного аналізу, де є неповнота даних.

XGBoost ідеально підходить для прогнозування розвитку цифрової економіки, не дозволяє уникнути нелінійних взаємозв'язків між такими показниками, як індекс телекомунікаційної інфраструктури, електронного урядування, а також рівень ВВП. Його здатність адаптуватися до складної структури даних дає можливість будувати прогнози з високою точністю, що є критичним для прийняття стратегічних рішень урядами та компаніями в цифровій сфері.

Метод Дерево рішень є базовим алгоритмом для побудови моделей прогнозування, який часто використовується через простоту та інтуїтивну зрозумілість. Алгоритм поділяє дані на основі логічних умов і створює ієрархічну структуру, де кожне «гілкування» відповідає певному набору правил [48].

Модель Дерево рішень має наступні особливості:

- Дерево рішень легко візуалізувати, що дозволяє зрозуміти ключові відносини з даними;
- Алгоритм швидко навчається навіть на великих наборах даних;

- Обробка як числових, так і категоріальних змінних робить модель Дерево рішень універсальним інструментом для аналізу різнорідних даних [49].

Дерева рішень часто застосовуються як базовий інструмент для визначення взаємозв'язків між факторами та результуючою змінною. Алгоритм працює за принципом послідовного розподілу даних на підгрупи, ґрунтуючись на певному критерію, що дозволяє ідентифікувати ключові залежності. У контексті прогнозування цифрової економіки цей метод може використовуватися для виявлення впливу різних показників, таких як індекс електронного урядування, рівень телекомунікаційної інфраструктури чи людського капіталу, на економічні результати [50].

Дерева рішень забезпечують прозорість і зрозумілість процесу аналізу, що робить їх корисними для попередньої оцінки та виявлення найбільш значущих факторів. Наприклад, за допомогою цього методу можна визначити, які аспекти цифрової інфраструктури мають найбільший вплив на ВВП, що дозволяє сфокусувати зусилля на оптимізації цих сфер [51]. Однак, як правило, метод схильний до перенавчання, особливо якщо дерево є занадто складним і враховує незначні коливання у вихідних даних.

Для вирішення цієї проблеми можна застосовувати методи обрізання дерев або впроваджувати цей алгоритм у більш складні об'єднані моделі, такі як Випадковий ліс або XGBoost. Завдяки цьому дерева рішень перетворюються на базовий компонент для побудови більш точних і надійних прогнозних моделей, які ефективно використовуються в аналізі та прогнозуванні розвитку цифрової економіки.

Випадковий ліс — це комбінований метод, який базується на об'єднанні великої кількості дерев рішень, створених на основі випадкових підмножин даних та змінних. Цей підхід дозволяє знизити ризик перенавчання і забезпечує високу точність прогнозування [52].

Модель Випадковий ліс має наступні особливості:

- За рахунок усереднення прогнозів окремих дерев алгоритм досягає більшої стабільності та точності;
- Випадковий ліс добре працює з даними, що містять шум, що робить його корисним для аналізу великих масивів економічних даних;
- Допомагає визначити, які фактори найбільше впливають на розвиток цифрової економіки;
- Вимагає більше обчислювальних ресурсів, ніж Дерева рішень, і може бути повільнішим при роботі з великими наборами даних [53], [54].

Кожен з описаних методів має свої переваги та можливі сфери застосування. Наприклад, XGBoost ідеально підходить для високоточних прогнозів і складних завдань, які потребують аналізу взаємодії між численними факторами. Дерево рішень може використовуватись для початкового аналізу даних і побудови простих, легко інтерпретованих моделей. А Випадковий ліс забезпечує баланс між простотою та високою точністю, що робить його універсальним інструментом для аналізу цифрової економіки.

Поєднання цих методів дає змогу провести глибокий аналіз і створити прогнози, які враховують як загальні тенденції, так і особливості окремих факторів. У сфері цифрової економіки ці алгоритми можуть бути застосовані для прогнозування впливу технологічних інновацій, моделювання розвитку електронної комерції, оцінки впровадження цифрових технологій і аналізу політик, спрямованих на вдосконалення цифрової інфраструктури.

РОЗДІЛ 3 АНАЛІЗ ПРОГНОЗІВ РОЗВИТКУ ЦИФРОВОЇ ЕКОНОМІКИ ТА ОЦІНКА ЇХ ЯКОСТІ

3.1 Оцінка якості прогнозних моделей

У цьому розділі здійснюється оцінка якості прогнозних моделей за допомогою ключових метрик — середньоквадратичної похибки (MSE) та коефіцієнта детермінації (R^2) для Кластер 1, Кластер 2 та Кластер 0. Аналіз проводиться таких моделей: XGBoost, дерева рішень і випадкового лісу.

Показник середньоквадратичної похибки (MSE) відображає середню величину похибки між прогнозованими та фактичними значеннями. Менші значення MSE свідчать про високу точність моделі.

Показник коефіцієнт детермінації (R^2) визначає частку варіації залежної змінної, яка пояснюється моделлю. Значення R^2, що наближається до 1, вказує на високу пояснювальну здатність моделі [55].

Розрахуємо значення MSE та R^2 для кожного кластеру обраних моделей (рисунок 3.1 – 3.9).

```
Кластер 1
MSE для кластера 1: 757479.4562
R² для кластера 1: 0.9205
```

Рисунок 3.1 — Значень MSE та R^2 (XGBoost) для Кластер 1

```
Кластер 2
MSE для кластера 2: 12506671.0764
R² для кластера 2: 0.9304
```

Рисунок 3.2 — Значень MSE та R^2 (XGBoost) для Кластер 2

```
Кластер 0
MSE для кластера 0: 85180848.6047
R² для кластера 0: 0.8480
```

Рисунок 3.3 — Значень MSE та R^2 (XGBoost) для Кластер 0

Для моделі XGBoost для Кластер 1 (рисунок 3.1) розраховане значення MSE є найменшим серед кластерів, що свідчить про високу точність прогнозів, значення R^2 показує, що модель пояснює 92.05% варіації залежної змінної, що свідчить про високу відповідність моделі реальним даним. Кластер 2 (рисунок 3.2) має розраховане значення MSE більше за кластер 1, але значно меншим, ніж для кластера 0. Це свідчить про достатньо високий рівень точності прогнозів для кластера 2. Значення R^2 показує, що модель пояснює 93.04% варіації залежної змінної, підтверджуючи стабільну якість моделі. Для Кластер 0 (рисунок 3.3) розраховане значення MSE є найбільшим серед усіх кластерів, що вказує на значне відхилення прогнозів від фактичних значень, тим самим знижує точність прогнозів у цьому кластері. Значення R^2 є показує, що модель пояснює 84.8% варіації залежної змінної, це свідчить про адекватну, але менш точну модель порівняно з іншими кластерами.

```
Кластер 1
MSE для кластера 1: 1462042.5229
R² для кластера 1: 0.8466
```

Рисунок 3.4 — Значень MSE та R^2 (Дерево рішень) для Кластер 1

```
Кластер 2
MSE для кластера 2: 36149939.8800
R² для кластера 2: 0.7987
```

Рисунок 3.5 — Значень MSE та R^2 (Дерево рішень) для Кластер 2

```
Кластер 0
MSE для кластера 0: 187638537.0812
R² для кластера 0: 0.6652
```

Рисунок 3.6 — Значень MSE та R^2 (Дерево рішень) для Кластер 0

Для моделі Дерево рішень для Кластер 1 (рисунок 3.4) розраховане значення MSE свідчить про високу точність моделі для цього кластера, оскільки прогнозовані значення близькі до фактичних. Значення R^2 показує, що модель пояснює 84.66% варіації залежної змінної, це вказує на те, що

модель якісна. Для Кластер 2 (рисунок 3.5) розраховане значення MSE порівняно з кластером 1 вказує на суттєвіші відхилення прогнозів від фактичних значень, це говорить про більшу варіативність даних. Значення R^2 показує, що модель пояснює 79.87% варіації залежної змінної, це свідчить про її адекватність. Для Кластер 0 (рисунок 3.6) розраховане значення MSE є найбільшим серед усіх кластерів, що вказує на значні розбіжності між фактичними та прогнозованими значеннями. Значення R^2 показує, що модель пояснює 66.52% варіації залежної змінної, що вказує на посередню якість прогнозування. Для покращення цієї моделі може знадобитися врахування додаткових змінних.

```
Кластер 1
MSE для кластера 1: 991266.7417
R² для кластера 1: 0.8960
```

Рисунок 3.7 — Значень MSE та R^2 (Випадковий ліс) для Кластер 1

```
Кластер 2
MSE для кластера 2: 20966984.2695
R² для кластера 2: 0.8832
```

Рисунок 3.8 — Значень MSE та R^2 (Випадковий ліс) для Кластер 2

```
Кластер 0
MSE для кластера 0: 175936964.8707
R² для кластера 0: 0.6861
```

Рисунок 3.9 — Значень MSE та R^2 (Випадковий ліс) для Кластер 0

Для моделі Випадковий ліс для Кластер 1 (рисунок 3.7) розраховане значення MSE дуже близьке до фактичного, що свідчить про високу точність моделі. Значення R^2 показує на рівні 89.60% високу якість моделі в кластері 1. Для Кластер 2 (рисунок 3.8) розраховане значення MSE вище у порівнянні з кластером 1, що вказує на більше відхилення прогнозованих значень від фактичних. Значення R^2 на рівні 88.32% демонструє хорошу якість моделі, модель пояснює більшу частину варіації залежної змінної, хоча є можливість

подальшого удосконалення. Для Кластер 0 (рисунок 3.9) розраховане значення MSE є найвищим, що вказує на значну похибку прогнозування. Значення R^2 вказує, що модель пояснює 68.61% варіації залежної змінної. Це прийнятний, але відносно низький показник, що свідчить про необхідність покращення моделі для цього кластера.

3.2 Реалізація прогнозів на основі побудованих моделей

Реалізація прогнозів здійснюється для кожного кластера за допомогою моделей XGBoost, дерево рішень та випадковий ліс. В результаті отримаємо прогнозні значення залежної змінної (GDP), базуючись на значеннях незалежних показників, що характеризують економічну ситуацію в кожному кластері. Для наочного порівняння побудовано графіки «Справжні значення vs Прогнози» для кожного кластера і кожної моделі (рисунок 3.10 – 3.18). Вони дозволять оцінити ступінь відповідності прогнозованих значень фактичним даним.

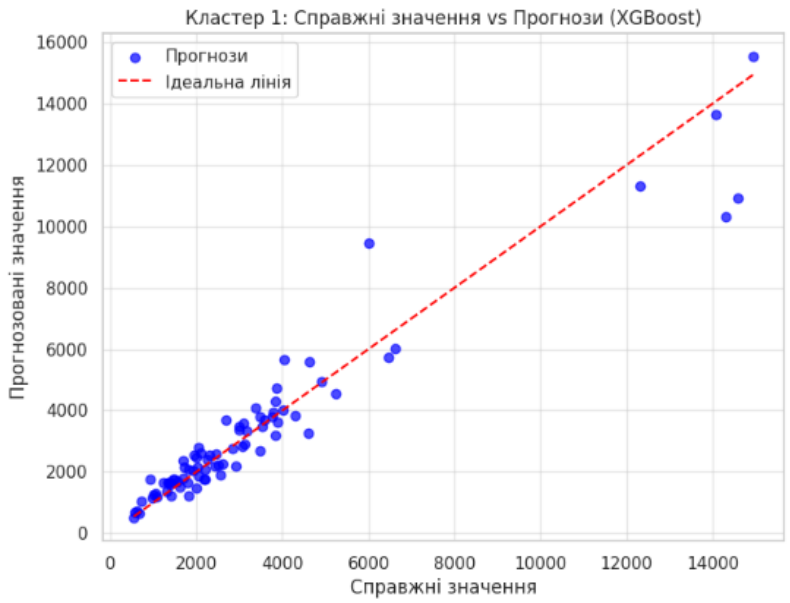


Рисунок 3.10 — Графік прогнозування (XGBoost) для Кластер 1

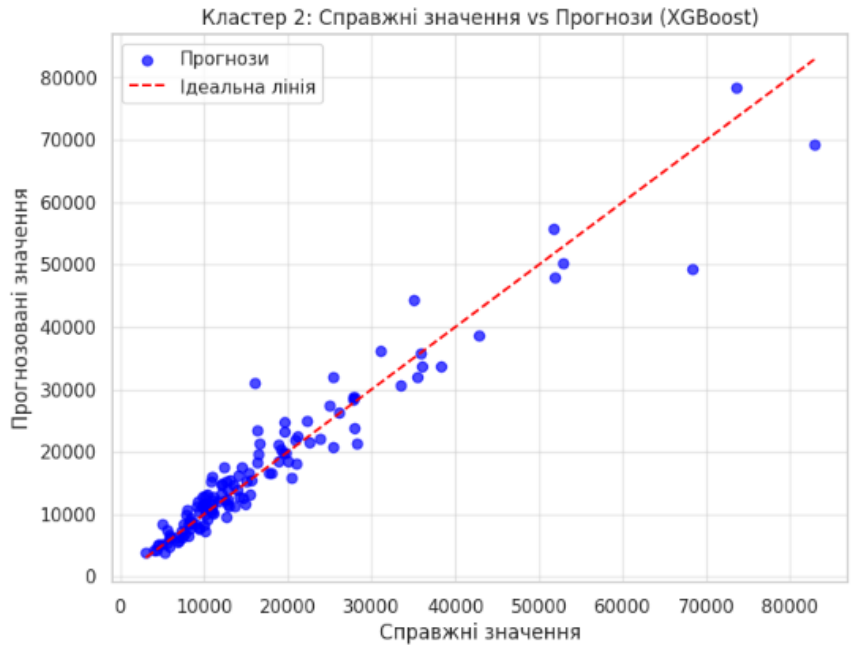


Рисунок 3.11 — Графік прогнозування (XGBoost) для Кластер 2

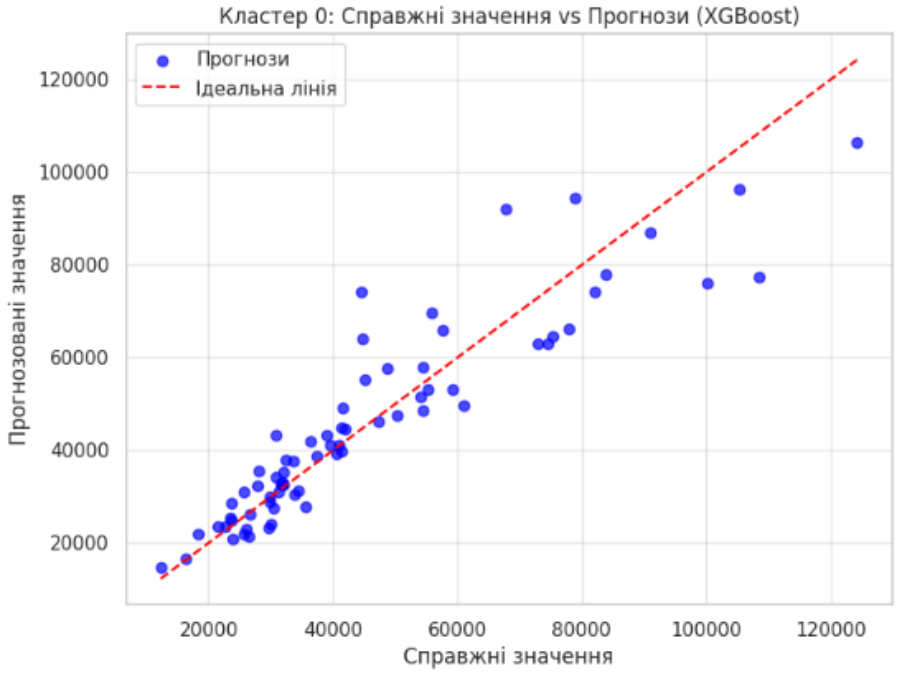


Рисунок 3.12 — Графік прогнозування (XGBoost) для Кластер 0

Модель XGBoost демонструє високу відповідність між прогнозованими та фактичними значеннями, що підтверджується розташуванням більшості точок поблизу ідеальної лінії регресії. Це свідчить про здатність моделі

ефективно пояснювати основні тенденції в даних. Однак, певне розсіювання точок навколо цієї лінії вказує на наявність похибок у прогнозуванні для окремих спостережень, що може бути наслідком варіативності даних. Загалом, модель XGBoost демонструє високу якість прогнозування, забезпечуючи надійну основу для аналізу та прийняття рішень у межах цих кластерів.

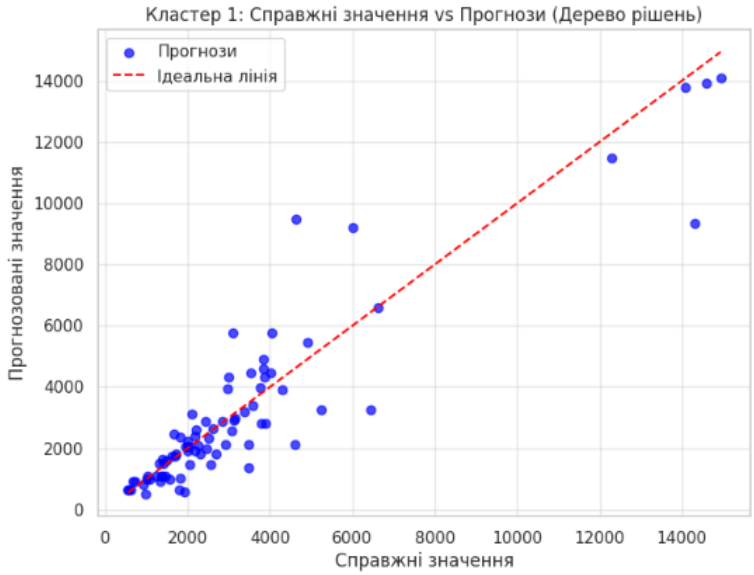


Рисунок 3.13 — Графік прогнозування (Дерево рішень) для Кластер 1

Рисунок 3.14 — Графік прогнозування (Дерево рішень) для Кластер 2

Рисунок 3.15 — Графік прогнозування (Дерево рішень) для Кластер 0

Модель Дерева рішень демонструє задовільні результати прогнозування для Кластеру 1, забезпечуючи високу кореляцію між прогнозованими та фактичними значеннями, що говорить про здатність моделі адекватно відображати структуру даних. Проте наявність розсіювання точок у Кластері 2 та значні відхилення у Кластері 3 вказують на похибки, зумовлені варіативністю даних або складністю їхньої внутрішньої структури. Це підкреслює необхідність подальшої оптимізації моделі або використання більш складних методів прогнозування для підвищення її точності.

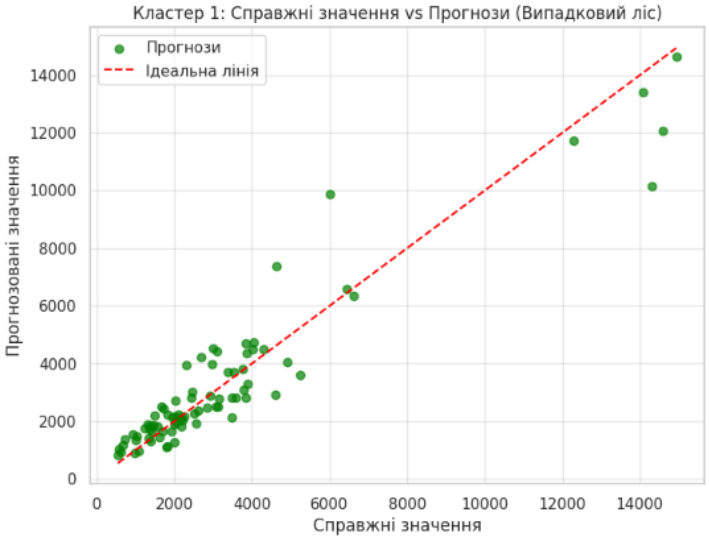


Рисунок 3.16 — Графік прогнозування (Випадковий ліс) для Кластер 1

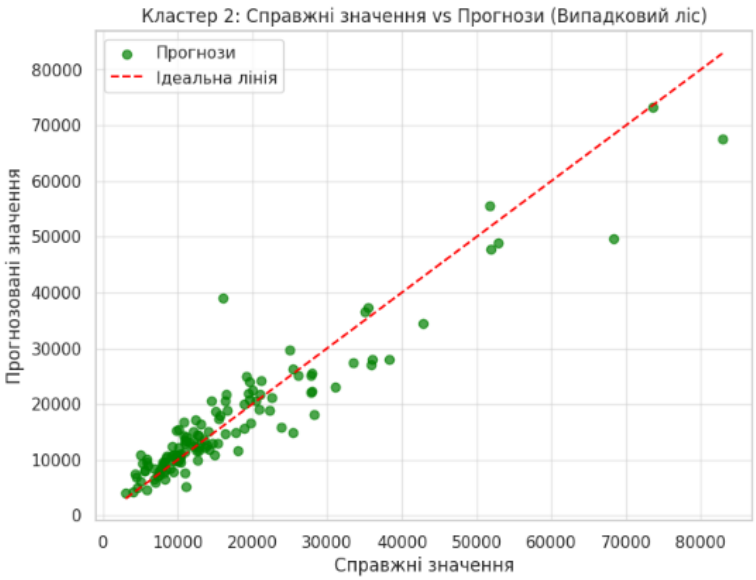


Рисунок 3.17 — Графік прогнозування (Випадковий ліс) для Кластер 2

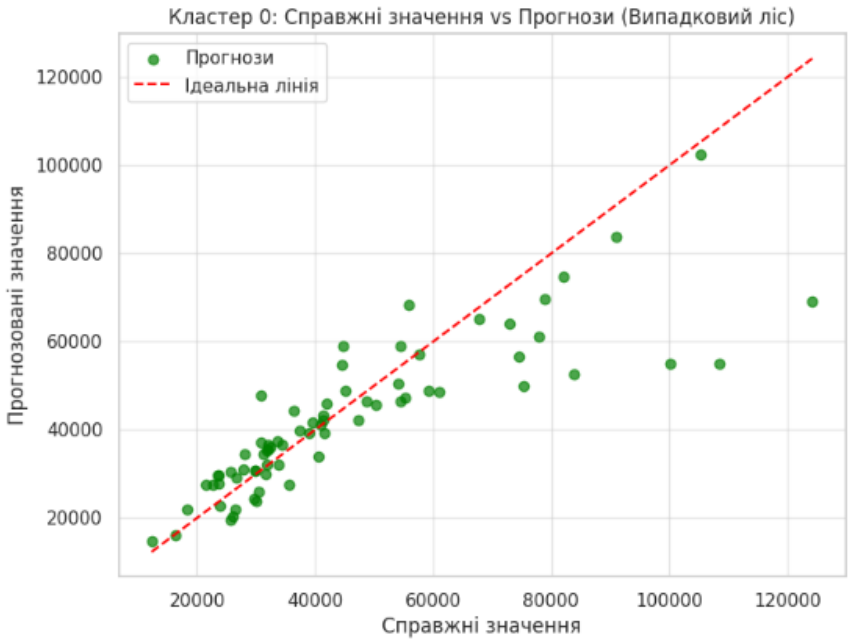


Рисунок 3.18 — Графік прогнозування (Випадковий ліс) для Кластер 0

Модель Випадковий ліс демонструє високу точність у прогнозуванні для всіх кластерів, забезпечуючи добру відповідність між фактичними та прогнозованими значеннями. Графіки показують, що більшість точок розташовані поблизу ідеальної лінії регресії, що вказує на загальну ефективність моделі. Незначний розкид точок навколо лінії у кластерах 1 і 2

свідчить про високий рівень точності прогнозів. У кластері 0 спостерігається дещо більший розкид, однак модель все ще зберігає свою адекватність.

3.3 Перевірка якості отриманих прогнозів

Тепер проведемо детальну оцінку якості отриманих прогнозів для кожного з кластерів, використовуючи такі дві основні метрики, як середньоквадратичну похибку (MSE) та коефіцієнт детермінації (R^2). MSE дозволяє визначити середню величину відхилення прогнозованих значень від фактичних, що дає змогу оцінити точність моделі на рівні кількісних показників. У свою чергу, R^2 відображає частку варіації залежної змінної, яку пояснює модель, і слугує показником відповідності моделі реальним даним [56]. Аналіз метрик MSE та R^2 на рівні кластерів дозволяє не лише оцінити загальну якість моделі, але й визначити особливості її роботи з окремими групами даних.

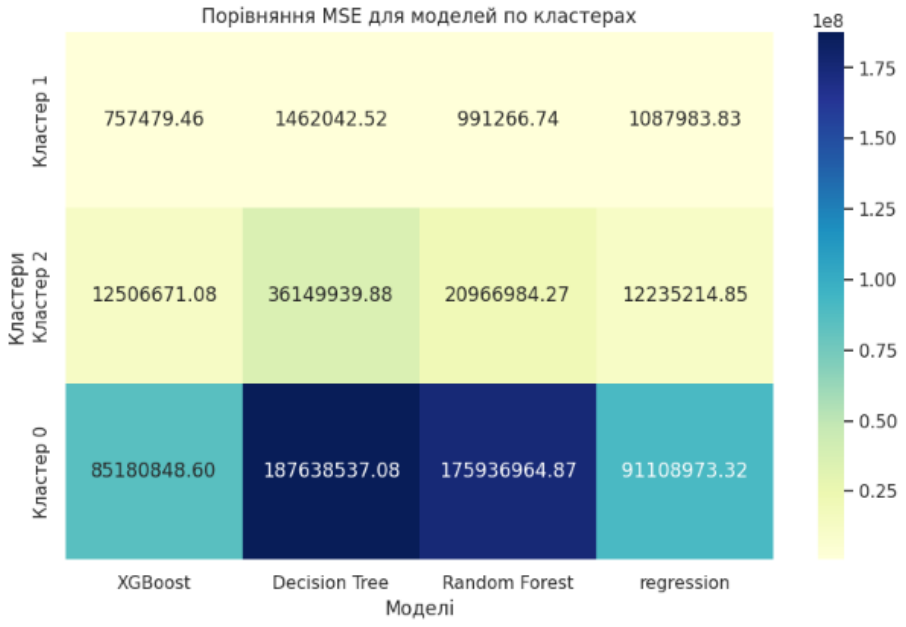


Рисунок 3.19 — Порівняння MSE для моделей по кластерах

Для кластеру 1 найкращі результати демонструє модель regression з найменшим значенням MSE. Моделі XGBoost та Random Forest також показують досить хороші результати, тоді як модель Decision Tree має найбільше значення MSE серед усіх моделей для цього кластера.

Для кластера 2 найкращі результати демонструє модель XGBoost. Модель regression також показує відносно хороші результати, тоді як моделі Decision Tree та Random Forest мають значно більші значення MSE.

Аналогічно до кластера 2, для кластера 0 найкращі результати демонструє модель XGBoost. Модель regression також показує відносно хороші результати, тоді як моделі Decision Tree та Random Forest мають значно більші значення MSE.

Отже, XGBoost стабільно забезпечує найкращу продуктивність у всіх кластерах, тоді як моделі дерева рішень і випадкового лісу відстають у плані точності, особливо в кластерах із більшою складністю чи дисперсією. Модель регресії добре працює в деяких кластерах, але XGBoost залишається найнадійнішим вибором для більшості випадків.

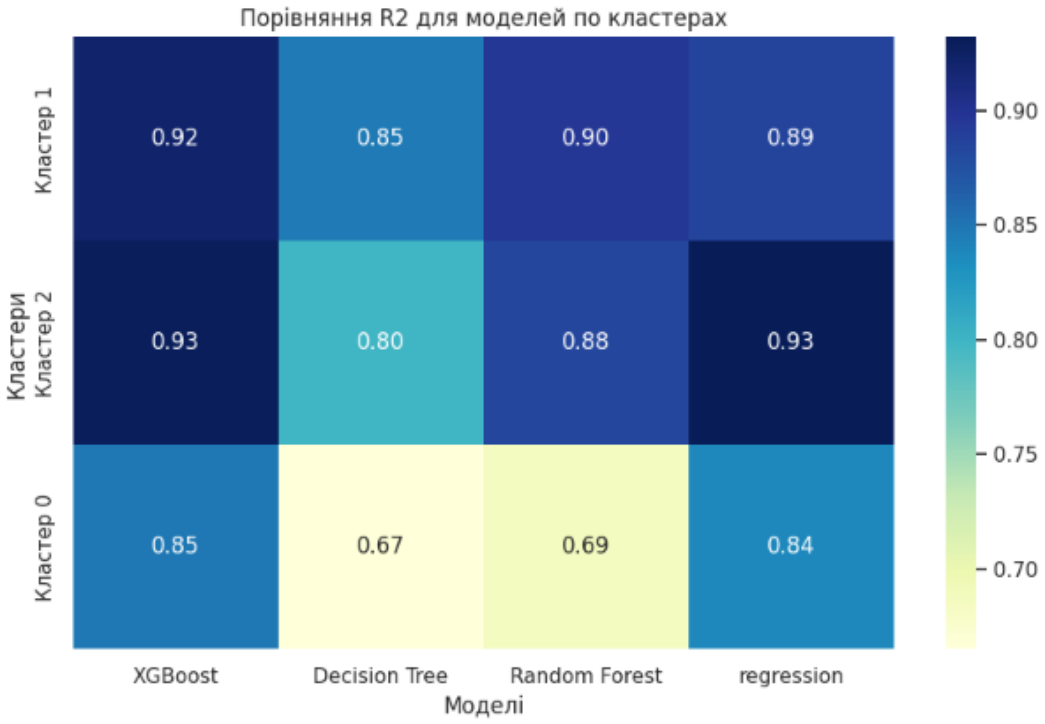


Рисунок 3.20 — Порівняння R^2 для моделей по кластерах

Найкращі результати для кластера 1 демонструє модель XGBoost з найвищим значенням R^2. Модель Random Forest також показує досить хороші результати, тоді як моделі Decision Tree та regression мають дещо нижчі значення R^2. Для кластера 2 найкращі результати демонструють моделі XGBoost та regression з майже однаковими високими значеннями R^2. Моделі Decision Tree та Random Forest мають дещо нижчі значення R^2. Аналогічно до кластера 2, для кластера 0 найкращі результати демонструє модель XGBoost. Модель regression також показує відносно хороші результати, тоді як моделі Decision Tree та Random Forest мають значно нижчі значення R^2.

Отже, виявлена висока ефективність XGBoost для Кластера 1 показує, що ця модель краще справляється з аналізом складних взаємозв'язків у даних. Однак для Кластера 2 і Кластера 0 також можна отримати непогані результати за допомогою регресії, що можна свідчити про збільшену структуру залежностей у цих підгрупах.

ВИСНОВКИ

У ході виконання дослідження на тему «Прогнозування розвитку цифрової економіки країни на основі методів машинного навчання» було виявлено та застосовано сучасні методи машинного навчання для аналізу та прогнозування ключових аспектів цифрової економіки. Проведений аналіз дозволив використовувати алгоритми ефективності XGBoost, Decision Tree та Random Forest у моделюванні розвитку цифрової економіки, зокрема у прогнозуванні показників, що характеризують її стан, таких як ВВП, індекс електронного урядування, рівень участі громадян в онлайн-сервісах, індекс людського капіталу та інші.

Статистичний аналіз вхідних даних показав відсутність пропущених значень, що спрощує подальшу обробку та справжню надійність результатів. Сильні кореляції між економічними та цифровими показниками підтверджують взаємозв'язок між цими факторами. Використання PCA дозволило зменшити вплив мультиколінеарності та виділити основні компоненти, що пояснюють варіативність даних. Кластерний аналіз дозволив розподілити країни на три групи з подібними характеристиками для подальших досліджень.

Результати регресійного аналізу для різних кластерів показали високу якість моделей, зокрема для кластерів 1 і 2, де коефіцієнти детермінації R^2 склали 0,935 і 0,942 відповідно, що вказує на високу точність прогнозування. Моделі для кластерів 1 і 2 показали статистичну значимість, де головні компоненти (PC1, PC3) мали сильний вплив на ВВП, хоча в деяких частинах залишки не відповідали нормальному розподілу. Для кластера 0 модель також показала хорошу якість, але з меншим R^2 (0,879) та вищими відхиленнями в прогнозах, але вона потребує вдосконалення для зменшення помилок в окремих випадках.

Проведені регресійні аналізи для кожного з трьох кластерів продемонстрували високу якість побудованих моделей, що підтверджується високими значеннями коефіцієнта детермінації (R^2) та статистичною значущістю моделей. Це свідчить про те, що розроблені моделі ефективно пояснюють варіацію ВВП та можуть бути використані для прогнозування.

Середньоквадратична помилка та коефіцієнт детермінації підтвердили, що модель для кластеру 2 є найефективнішою, з високою точністю прогнозування, хоча відхилення у вигляді високого значення MSE показують на потенціал для подальшого вдосконалення точності. Для кластеру 1 і 0 має певний розкид точок навколо ідеальної лінії на графіку прогнозів, що також вимагає уточнення для зменшення похибки в прогнозах.

Розраховані значення MSE та R^2 для кожного кластеру обраних моделей показали наступне. Результати моделювання за допомогою XGBoost для різних кластерів свідчать про високу якість моделей. Найкращі результати отримано для кластера 2 з найменшим значенням MSE та найвищим значенням R^2. Модель для кластера 1 також демонструє добру якість, хоча і з дещо більшими помилками прогнозування. Найнижчу точність має модель для кластера 0, що проявляється у значному значенні MSE та нижчому значенні R^2.

Аналіз моделей Дерева рішень показав, що модель для кластера 1 має найменше значення MSE та найвище значення R^2, що свідчить про найкращу якість прогнозування серед усіх кластерів. Модель для кластера 2 також демонструє задовільну якість, хоча і з більшими відхиленнями. Найнижчу точність має модель для кластера 0, що проявляється у значному значенні MSE та нижчому значенні R^2.

Результати моделювання за допомогою Випадкового лісу показали високу якість моделей для всіх кластерів. Незначний розкид точок на графіках прогнозування для кластерів 1 і 2 свідчить про високу точність моделей. Для

кластера 0 спостерігається дещо більший розкид, однак загалом модель також демонструє добру якість.

Проведена детальна оцінка якості отриманих прогнозів для кожного з кластерів показала, що для моделі XGBoost значення MSE демонструє найкращі результати для всіх трьох кластерів. Вона найкраще справляється з прогнозуванням і найменше схильна до помилок. Модель лінійної регресії також показує непогані результати, особливо для кластера 1, що може свідчити про лінійну залежність між змінними в цьому кластері. Моделі дерева рішень та випадкового лісу мають більшу похибку, що може бути пов'язано з їхньою меншою гнучкістю або перенавчанням.

Для моделі XGBoost значення R^2 демонструє найкращі результати для всіх кластерів. Це свідчить про її високу здатність пояснювати дисперсію залежної змінної та робити точні прогнози. Модель лінійної регресії також показує досить хороші результати для всіх кластерів, особливо для кластерів 1 та 2. Це може вказувати на лінійну залежність між змінними в цих кластерах. Моделі Decision Tree та Random Forest демонструють гірші результати порівняно з моделями XGBoost та Regression. Це може бути пов'язано з складністю даних або невідповідністю структури даних цим моделям.

На основі цього аналізу можна зробити висновок, що модель XGBoost є найбільш універсальною та ефективною для всіх розглянутих кластерів.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Ратинський В. В., Співак С. М., Синькевич Н. І. РОЗВИТОК ЕКОНОМІЧНОГО АНАЛІЗУ В УКРАЇНІ В КОНТЕКСТІ ЦИФРОВОЇ ЕКОНОМІКИ. Ефективна економіка. 2023. № 7. URL: https://doi.org/10.32702/2307-2105.2023.7.16 (дата звернення: 11.09.2024).

2. Т. М. Куценко та ін. ТЕХНОЛОГІЯ БЛОКЧЕЙНУ В СИСТЕМІ ПРИЙНЯТТЯ УПРАВЛІНСЬКИХ РІШЕНЬ ПРИ ЦИФРОВІЗАЦІЇ ЗОВНІШНЬОЕКОНОМІЧНОЇ ДІЯЛЬНОСТІ ПІДПРИЄМСТВА. Інвестиції: практика та досвід. 2024. № 11. С. 166–173. URL: https://doi.org/10.32702/2306-6814.2024.11.166 (дата звернення: 11.09.2024).

3. Білик М. Ю., Яковенко Я. Ю., Олійник Є. В. РОЗВИТОК ЦИФРОВОЇ ЕКОНОМІКИ: СУЧАСНІ НАПРЯМИ ТА ПЕРСПЕКТИВИ. Ефективна економіка. 2023. № 1. URL: https://doi.org/10.32702/2307-2105.2023.1.15 (дата звернення: 13.09.2024).

4. The Future of Jobs Report 2023. World Economic Forum. URL: https://www.weforum.org/publications/the-future-of-jobs-report-2023/ (date of access: 13.09.2024).

5. Uddin A. S. M. A. The Era of AI: Upholding Ethical Leadership. Open Journal of Leadership. 2023. Vol. 12, no. 04. P. 400–417. URL: https://doi.org/10.4236/ojl.2023.124019 (date of access: 13.09.2024).

6. Кіншаков Е. В. Моделювання та прогнозування великих наборів даних засобами машинного навчання. 2020. URL: https://essuir.sumdu.edu.ua/handle/123456789/81370 (дата звернення: 15.09.2024).

7. Калініченко О., Вікарчук О., Ніколаєнко С. ФІНАНСОВА ГРАМОТНІСТЬ – ЗАПОРУКА УСПІШНОГО НАСЕЛЕННЯ. Economics.

Management. Innovations. 2019. № 1(24). URL: https://doi.org/10.35433/issn2410-3748-2019-1(24)-3 (дата звернення: 16.09.2024).

8. Ушкальов В., Мартіянова М. ПОВЕДІНКОВІ АСПЕКТИ ЦИФРОВІЗАЦІЇ БІЗНЕСУ. Наукові інновації та передові технології. 2023. № 14(28). URL: https://doi.org/10.52058/2786-5274-2023-14(28)-805-815 (дата звернення: 16.09.2024).

9. Khlivniuk, T. P. ПЛАТФОРМНА ЕКОНОМІКА ЯК ЧИННИК МОДЕРНІЗАЦІЇ СОЦІАЛЬНОЇ ДЕРЖАВИ. Епістемологічні дослідження в філософії, соціальних і політичних науках. 2021. Т. 4, № 1. С. 123–131. URL: https://doi.org/10.15421/342114 (дата звернення: 16.09.2024).

10. Петросян А. Р., Петросян Р. В., Колос К. Р. Розробка платформи віддаленого управління інфраструктурою Інтернет речей. Технічна інженерія. 2021. № 1(87). С. 73–80. URL: https://doi.org/10.26642/ten-2021-1(87)-73-80 (дата звернення: 16.09.2024).

11. В.Д. Голь, А. Ю. Раківська, Д. Ю.Раківський. ЗАСОБИ КІБЕРЗАХИСТУ НА РІВНІ МЕРЕЖНОЇ ІНФРАСТРУКТУРИ. Системи управління, навігації та зв’язку. Збірник наукових праць. – Полтава: ПНТУ, 2022. Т. 3 № 69. С. 116-120. URL: https://doi.org/10.26906/SUNZ.2022.3.116 (дата звернення: 17.09.2024).

12. Welcome to GOV.UK. URL: https://www.gov.uk/ (date of access: 17.09.2024).

13. e-Estonia. URL: https://e-estonia.com/ (date of access: 17.09.2024).

14. Дія. Державні послуги онлайн. URL: https://diia.gov.ua/ (дата звернення: 17.09.2024).

15. Дашко І. М., Михайліченко Л. В. ТЕНДЕНЦІЇ РОЗВИТКУ ЦИФРОВОЇ ЕКОНОМІКИ В УКРАЇНІ ТА КРАЇНАХ ЄС. Efektyvna ekonomika. 2024. № 7. URL: https://doi.org/10.32702/2307-2105.2024.7.30 (дата звернення: 17.09.2024).

16. Diia.City. Diia.City. URL: https://city.diia.gov.ua/ (date of access: 17.09.2024).

17. Міністерство цифрової трансформації України. URL: https://thedigital.gov.ua/news/mikhaylo-fedorov-diya-city-stvoryue-naykrashchi-u-sviti-umovi-dlya-rozvitku-produktovikh-it-kompaniy (дата звернення: 20.09.2024).

18. Асоціація "IT Ukraine". URL: https://itukraine.org.ua/report/pidsumki-diyalnosti-it-ukraine/ (дата звернення: 20.09.2024).

19. Прасад А., Тарасовський Ю. Експорт IT-послуг з України у вересні продовжив падати – Forbes.ua. Forbes.ua | Бізнес, мільярдери, новини, фінанси, інвестиції, компанії. URL: https://forbes.ua/news/eksport-it-poslug-z-ukraini-u-veresni-prodovzhiv-padati-31102024-24525 (дата звернення: 20.09.2024).

20. The World Economic Forum. URL: https://www.weforum.org/ (date of access: 21.09.2024).

21. China Internet Watch. URL: https://www.chinainternetwatch.com/ (date of access: 21.09.2024).

22. Eurostat. Language selection. European Commission. URL: https://ec.europa.eu/eurostat/data/database (date of access: 21.09.2024).

23. Журнал «НТІ» – Наука Технології Інновації. URL: https://nti.ukrintei.ua/ (дата звернення: 21.09.2024).

24. Новини IT і бізнесу в Україні. URL: https://ain.ua/2023/12/15/minczyfry-predstavylo-strategiyu-rozvytku-innovaczij-ukrayiny/ (дата звернення: 21.09.2024).

25. ITU. Committed to connecting the world. URL: https://www.itu.int/en/Pages/default.aspx (date of access: 23.09.2024).

26. Національний інститут стратегічних досліджень. URL: https://niss.gov.ua/news/komentari-ekspertiv/tsyfrova-transformatsiya-ekonomiky-ukrayiny-v-umovakh-viyny-sichen-2024 (дата звернення: 23.09.2024).

27. Havrylenko N., Tarasenko I. CURRENT TRENDS OF DIGITALIZATION OF THE ECONOMY: PROBLEMS AND PROSPECTS OF DEVELOPMENT. 2017. № 3 (47). URL: https://doi.org/10.25313/2520-2294-2021-3-7046 (date of access: 25.09.2024).

28. Акулюшина М., Ісламова А., Біюк В. ПЕРСПЕКТИВИ РОЗВИТКУ ЦИФРОВОЇ ЕКОНОМІКИ В УКРАЇНІ. Економіка та суспільство. 2024. № 61. URL: https://doi.org/10.32782/2524-0072/2024-61-11 (дата звернення: 25.09.2024).

29. Офіційний веб-портал парламенту України. Угода про асоціацію між Україною, з однієї сторони, та Європейським Союзом, Європейським співтовариством з атомної енергії і їхніми державами-членами, з іншої сторони. URL: https://zakon.rada.gov.ua/laws/show/984_011#Text (дата звернення: 25.09.2024).

30. Котелевець Д. О. ТЕНДЕНЦІЇ РОЗВИТКУ ЦИФРОВОЇ ЕКОНОМІКИ В УКРАЇНІ. Проблеми сучасних трансформацій. Серія: економіка та управління. 2022. № 5. URL: https://doi.org/10.54929/2786-5738-2022-5-03-01 (дата звернення: 27.09.2024).

31. Кабінет Міністрів України. https://www.kmu.gov.ua/. URL: https://www.kmu.gov.ua/ (дата звернення: 27.09.2024).

32. Дія.Освіта. URL: https://osvita.diia.gov.ua/ (дата звернення: 27.09.2024).

33. Михайлов Н. О. МЕТОДИ ВИСОКОЕФЕКТИВНОГО ПЛАНУВАННЯ ПРОЄКТІВ: ТРАДИЦІЙНІ ПІДХОДИ ТА МАШИННЕ НАВЧАННЯ. Таврійський науковий вісник. Серія: Технічні науки. 2024. № 4. С. 186–192. URL: https://doi.org/10.32782/tnv-tech.2024.4.18 (дата звернення: 27.09.2024).

34. Шадура О. В. МОДИФІКАЦІЯ ГЕНЕТИЧНИХ АЛГОРИТМІВ НА ОСНОВІ МЕТОДУ НЕЦЕНТРОВАНИХ ГОЛОВНИХ КОМПОНЕНТ ТА СТАНДАРТНІ ТЕСТИ. World Science. 2019. Т. 1, № 4(44). С. 4–10. URL:

https://doi.org/10.31435/rsglobal_ws/30042019/6464 (дата звернення: 30.09.2024).

35. United Nations. URL: https://publicadministration.un.org/egovkb/en-us/Data-Center (date of access: 11.09.2024).

36. Корякіна А. А. Матриця прийняття рішень. 2019. URL: https://er.knutd.edu.ua/handle/123456789/13897 (дата звернення: 03.10.2024).

37. Royston J. P. An Extension of Shapiro and Wilk's W Test for Normality to Large Samples. Applied Statistics. 1982. Vol. 31, no. 2. P. 115. URL: https://doi.org/10.2307/2347973 (date of access: 03.10.2024).

38. D. Marutho, S. Hendra Handaka, E. Wijaya and Muljono, "The Determination of Cluster Number at k-Mean Using Elbow Method and Purity Evaluation on Headline News", 2018 International Seminar on Application for Technology of Information and Communication , Semarang, Indonesia, 2018, p. 533-538, https://doi.org/10.1109/ISEMANTIC.2018.8549751 (date of access: 03.10.2024).

39. Jia C, Shang M, Cao J, Liu Y (2023) Емпіричний аналіз впливу цифрової економіки на зелену трансформацію виробництва: докази Китаю. PLoS ONE 18(8). https://doi.org/10.1371/journal.pone.028996840 (дата звернення: 05.10.2024).

40. Gow, I.D., & Ding, T. Empirical Research in Accounting: Tools and Methods (1st ed.). Chapman and Hall/CRC. 2024. https://doi.org/10.1201/9781003456230 (date of access: 07.10.2024).

41. Ярещенко Н. В. Оцінка параметрів моделі прогнозування розрахункових характеристик. Сучасні технології та методи розрахунків у будівництві. 2024. № 20. С. 205–211. URL: https://doi.org/10.36910/6775-2410-6208-2023-10(20)-22 (дата звернення: 07.10.2024).

42. Smith H., Draper N. R. Applied Regression Analysis. Wiley & Sons, Incorporated, John, 2014. P. 736 (date of access: 07.10.2024).

43. Stock-Price Forecasting Based on XGBoost and LSTM / P. Hoang Vuong et al. Computer Systems Science and Engineering. 2022. Vol. 40, №. 1. P. 237–246. URL: https://doi.org/10.32604/csse.2022.017685 (date of access: 09.10.2024).

44. Chen, T. and Guestrin, C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, San Francisco, 13-17 August 2016, 785-794. 2016. URL: https://doi.org/10.1145/2939672.2939785 (date of access: 11.10.2024).

45. OUKHOUYA, H., KADIRI, H., EL HIMDI, K., & GUERBAZ, R. Forecasting International Stock Market Trends: XGBoost, LSTM, LSTM-XGBoost, and Backtesting XGBoost Models. Statistics, Optimization & Information Computing, 12(1), 200-209. 2023. URL: https://doi.org/10.19139/soic-2310-5070-1822 (date of access: 14.10.2024).

46. What is the XGBoost Algorithm in ML. Explained With Steps. theiotacademy. URL: https://www.theiotacademy.co/blog/xgboost-algorithm/ (date of access: 15.10.2024).

47. Z. E. Aydin and Z. K. Ozturk, PERFORMANCE ANALYSIS of XGBOOST CLASSIFIER WITH MISSING DATA, 1st Int. Conf. Comput. Mach. Intell., № March, 2021. URL: https://www.researchgate.net/publication/350135431 (date of access: 15.10.2024).

48. Нестеренко Б.М. Брич О.І. Шахматов І.О. ДЕРЕВО РІШЕНЬ – ЕФЕКТИВНИЙ ІНСТРУМЕНТ ПРИЙНЯТТЯ РІШЕНЬ В УМОВАХ НЕВИЗНАЧЕНОСТІ. 2008. URL: http://er.nau.edu.ua/handle/NAU/20785 (дата звернення: 17.10.2024).

49. Breiman L., Last M., Rice J. Random Forests: Finding Quasars. Statistical Challenges in Astronomy. New York. P. 243–254. URL: https://doi.org/10.1007/0-387-21529-8_16 (date of access: 18.2024).

50. Оттен Н. В. Decision Trees In ML Complete Guide [How To Tutorial, Examples, 5 Types & Alternatives]. Spot Intelligence. 2024. URL:

https://spotintelligence.com/2024/05/22/decision-trees-in-ml/ (date of access: 19.10.2024).

51. 8 Key Advantages and Disadvantages of Decision Trees. Inside Learning Machines. Inside Learning Machines. URL: https://insidelearningmachines.com/advantages_and_disadvantages_of_decision_trees/ (date of access: 19.10.2024).

52. IBM. What Is Random Forest? United States. URL: https://www.ibm.com/topics/random-forest (date of access: 25.10.2024).

53. Random Forest and Boosting. Data Analysis for Business, Economics, and Policy. P. 438–456. 2021. URL: https://doi.org/10.1017/9781108591102.016 (date of access: 25.10.2024).

54. Quinlan, J.R. Induction of Decision Trees. Machine Learning, 1, 81-106. 1986. URL:http://dx.doi.org/10.1007/BF00116251 (date of access: 25.10.2024).

55. Zach Bobbitt. MSE vs. RMSE: WHICH METRIC SHOULD YOU USE?. Statology. URL: https://www.statology.org/mse-vs-rmse/ (date of access: 08.11.2024).

56. Regression Model Accuracy (MAE, MSE, RMSE, R-squared) Check in R. DataTechNotes. URL: https://www.datatechnotes.com/2019/02/regression-model-accuracy-mae-mse-rmse.html (date of access: 14.11.2024).

ДОДАТКИ

Додаток А

```
# Встановлення бібліотек
!pip install pandas matplotlib seaborn statsmodels openpyxl
```

```
Requirement already satisfied: pandas in /usr/local/lib/python3.10/dist-packages (2.2.2)
Requirement already satisfied: matplotlib in /usr/local/lib/python3.10/dist-packages (3.8.0)
Requirement already satisfied: seaborn in /usr/local/lib/python3.10/dist-packages (0.13.2)
Requirement already satisfied: statsmodels in /usr/local/lib/python3.10/dist-packages (0.14.4)
Requirement already satisfied: openpyxl in /usr/local/lib/python3.10/dist-packages (3.1.5)
Requirement already satisfied: numpy>=1.22.4 in /usr/local/lib/python3.10/dist-packages (from pandas) (1.26.4)
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.10/dist-packages (from pandas) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-packages (from pandas) (2024.2)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.10/dist-packages (from pandas) (2024.2)
Requirement already satisfied: contourpy>=1.0.1 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (1.3.1)
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (4.55.0)
Requirement already satisfied: kiwisolver>=1.0.1 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (1.4.7)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (24.2)
Requirement already satisfied: pillow>=6.2.0 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (11.0.0)
Requirement already satisfied: pyparsing>=2.3.1 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (3.2.0)
Requirement already satisfied: scipy!=1.9.2,>=1.8 in /usr/local/lib/python3.10/dist-packages (from statsmodels) (1.13.1)
Requirement already satisfied: patsy>=0.5.6 in /usr/local/lib/python3.10/dist-packages (from statsmodels) (1.0.1)
Requirement already satisfied: et-xmlfile in /usr/local/lib/python3.10/dist-packages (from openpyxl) (2.0.0)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-packages (from python-dateutil>=2.8.2->pandas) (1.16.0)
```

Рисунок А.1 — Встановлення бібліотеки

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from statsmodels.stats.outliers_influence import variance_inflation_factor
from statsmodels.tools.tools import add_constant
from sklearn.preprocessing import StandardScaler
from scipy import stats
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA
import statsmodels.api as sm
import xgboost as xgb
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.tree import DecisionTreeRegressor
from sklearn import tree
from sklearn.ensemble import RandomForestRegressor
from IPython.display import display

# Завантаження даних з Excel-файлу
df = pd.read_excel('Data.xlsx')
```

Рисунок А.2 — Встановлення бібліотек

```
df = df.drop(["Year", "Country Name"], axis=1)
```

```
df
```

Рисунок А.3 — Виведення вхідних даних

```
numeric_columns = df.select_dtypes(include=['float64', 'int64']).columns

# Проходимо по кожному стовпцю та будуємо графік
for column in numeric_columns:
    plt.figure(figsize=(8, 4))
    df[column].plot(kind='line', title=column)
    plt.xlabel('Index')  # Підпис осі X
    plt.ylabel(column)  # Підпис осі Y
    plt.gca().spines[['top', 'right']].set_visible(False)  # Відключення верхньої та правої рамки
    plt.grid(axis='y', linestyle='--', alpha=0.7)  # Додавання сітки
    plt.show()
```

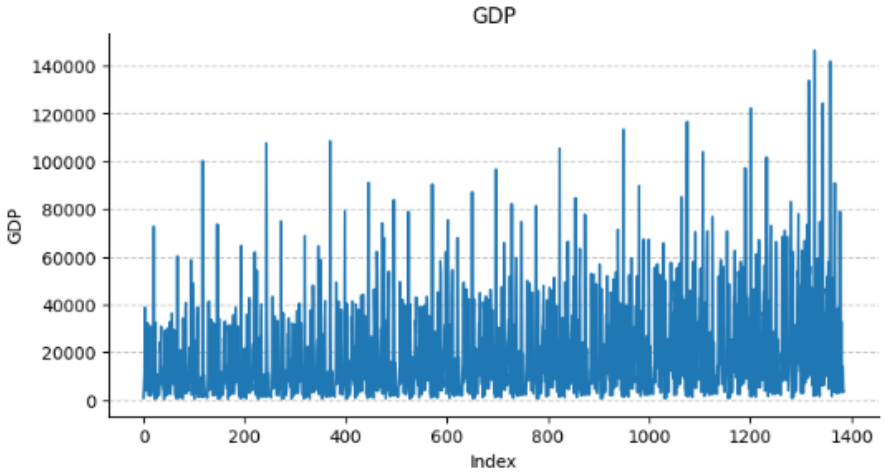


Рисунок А.4 — Графіки для кожної змінної

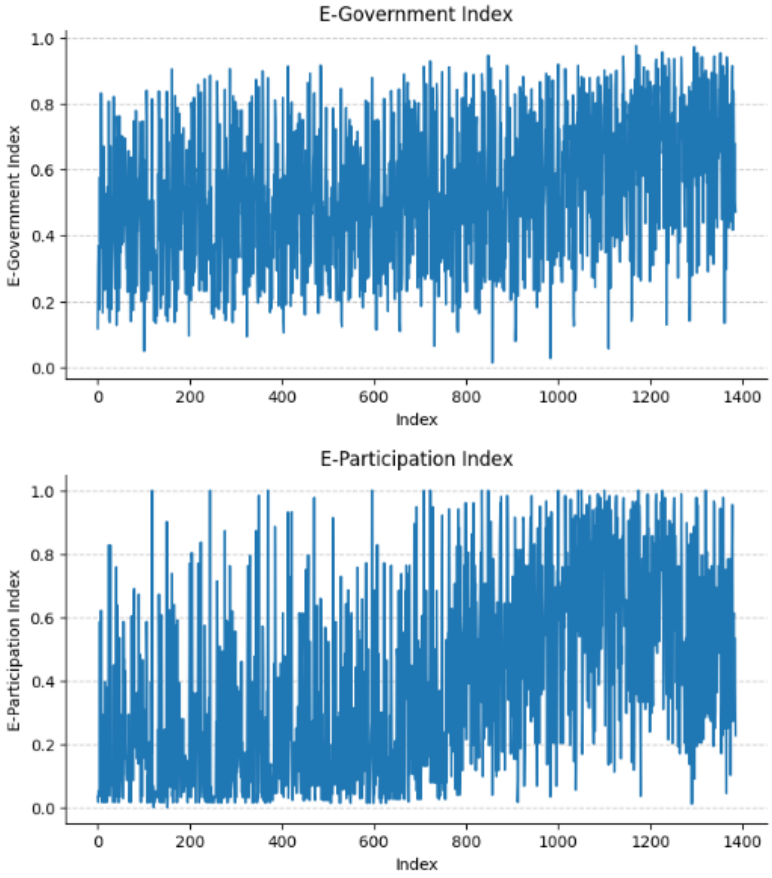


Рисунок А.5 — Продовження рисунку Б.4

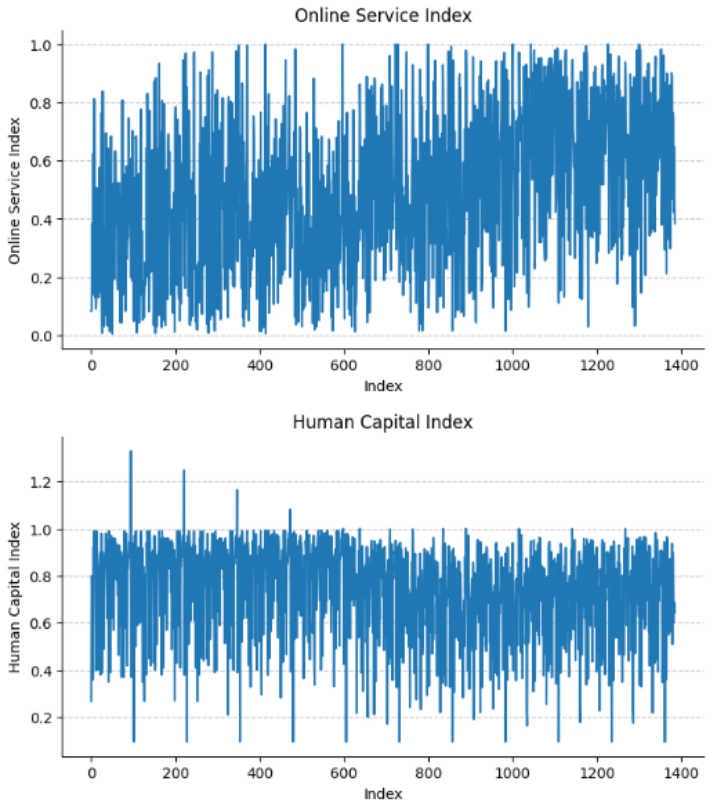


Рисунок А.6 — Продовження рисунку Б.4

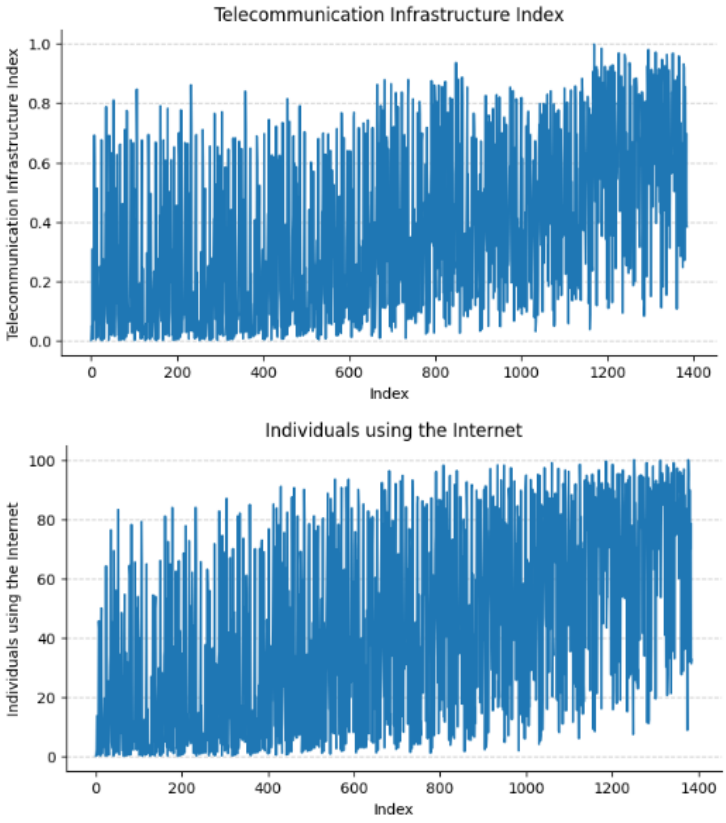


Рисунок А.7 — Продовження рисунку Б.4

```python
# Фільтруємо числові стовпці
numeric_columns = df.select_dtypes(include=['float64', 'int64']).columns

# Створюємо гістограми для кожного показника
fig, axes = plt.subplots(nrows=len(numeric_columns), ncols=1, figsize=(8, len(numeric_columns) * 4))
fig.tight_layout(pad=5.0)

for i, col in enumerate(numeric_columns):
    ax = axes[i] if len(numeric_columns) > 1 else axes  # Якщо лише один графік, не використовуємо список
    df[col].plot(kind='hist', bins=20, alpha=0.7, color='blue', ax=ax, edgecolor='black')
    ax.set_title(f'Histogram of {col}', fontsize=14)
    ax.set_xlabel(col, fontsize=12)
    ax.set_ylabel('Frequency', fontsize=12)
    ax.grid(axis='y', linestyle='--', alpha=0.7)
    ax.spines[['top', 'right']].set_visible(False)

plt.show()
```

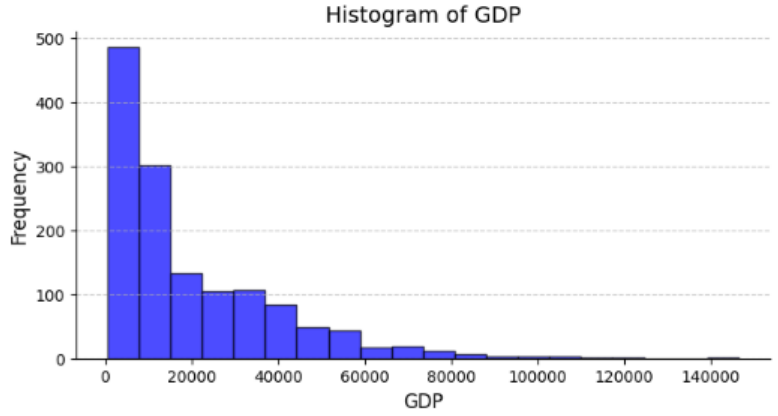


Рисунок А.8 — Гістограми для кожної змінної

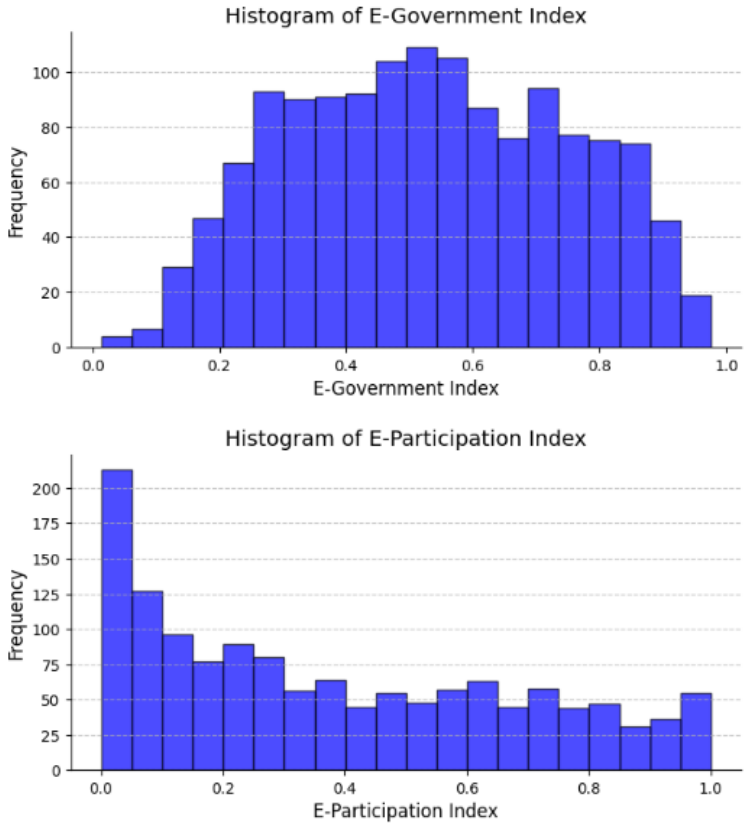


Рисунок А.9 — Продовження рисунку Б.8

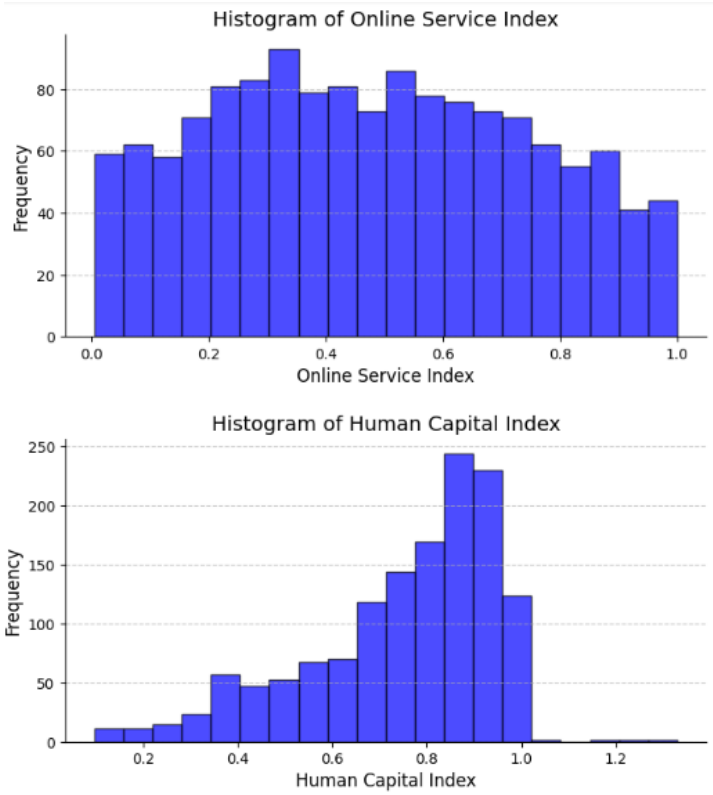


Рисунок А.10 — Продовження рисунку Б.8

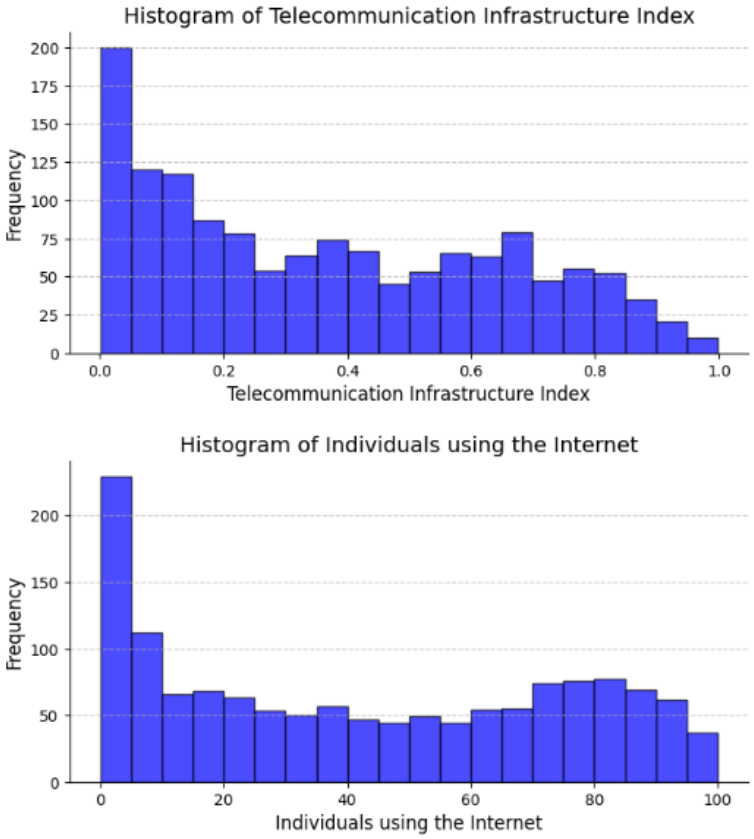


Рисунок А.11 — Продовження рисунку Б.8

```
#перевірка наявності/відсутності нульових значень
df.isnull().sum()
```

	0
GDP	0
E-Government Index	0
E-Participation Index	0
Online Service Index	0
Human Capital Index	0
Telecommunication Infrastructure Index	0
Individuals using the Internet	0

dtype: int64

Рисунок А.12 — Перевірка наявності/відсутності нульових значень

```
# Візуалізація відсутніх значень
plt.figure(figsize=(10, 6))

# Вибір незалежних змінних (усі, крім 'y')
independent_vars = df.drop(columns=['GDP'])
sns.heatmap(independent_vars.isnull(), yticklabels=False, cmap='viridis', cbar=False) #df.isnull()
```

Рисунок А.13 — Візуалізація відсутніх значень

```
#описова статистика
df.describe()
```

Рисунок А.14 — Описова статистика

```
# Обчислення кореляційної матриці для даних
correlation_matrix = df.select_dtypes(include=['float64', 'int64']).corr()

# Візуалізація кореляційної матриці
plt.figure(figsize=(15, 8))
sns.heatmap(correlation_matrix, annot=True, fmt=".2f", cmap='coolwarm', cbar=True, square=True, linewidths=0.5)
plt.title('Кореляційна матриця')
plt.show()
```

Рисунок А.15 — Обчислення кореляційної матриці

```python
# Перевірка мультиколінеарності
# Вибір усіх стовпців, що містять дані для перевірки мультиколінеарності
columns_to_check_vif = [col for col in df.columns if any(keyword in col for keyword in
                                                         ['E-Government Index', 'E-Participation Index', 'Online Service Index', 'Human Capital Index',
                                                          'Telecommunication Infrastructure Index', 'Individuals using the Internet'])]

# Додавання константи для коректного обчислення VIF
X = df[columns_to_check_vif]
X = add_constant(X)

# Обчислення VIF для кожного стовпця
vif_data = pd.DataFrame()
vif_data["feature"] = X.columns
vif_data["VIF"] = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]

# Налаштування відображення чисел з 5 знаками після коми
pd.set_option('display.float_format', '{:.5e}'.format)


# Округлення VIF до 5 знаків після коми
vif_data["VIF"] = vif_data["VIF"].round(5)

# Виведення результату
vif_data
```

Рисунок А.16 — Перевірка мультиколінеарності

```python
# Стандартизація даних


scaler = StandardScaler()
df_scaled = pd.DataFrame(scaler.fit_transform(df))

# Відновлення індексів та стовпців
df_scaled.columns = df.columns
df_scaled.index = df.index

# Округлення до 5 знаків після коми
pd.set_option('display.float_format', '{:.5f}'.format)

# Виведення результату
df_scaled
```

Рисунок А.17 — Стандартизація даних

```python
# Статистичний аналіз
# Перевірка на нормальність для всіх числових стовпців
for column in df.columns:
    if df[column].dtype != 'object':  # Перевірка, чи є стовпець числовим
        stat, p_value = stats.shapiro(df[column].dropna())  # Видалення NaN значень перед тестом
        print(f"Shapiro-Wilk test for {column}: stat={stat}, p_value={p_value}")
```

Рисунок А.18 — Тест Шапіро-Вілка

```
# Побудова коробкових діаграм
# Створення фігури з оптимальними розмірами
plt.figure(figsize=(15, 10))

# Побудова коробкових діаграм
sns.boxplot(data=df_scaled, orient="v")

# Налаштування підписів по осі X
plt.xticks(rotation=90)

# Додавання заголовка
plt.title("Коробкові діаграми для показників")

# Відображення графіка
plt.show()
```

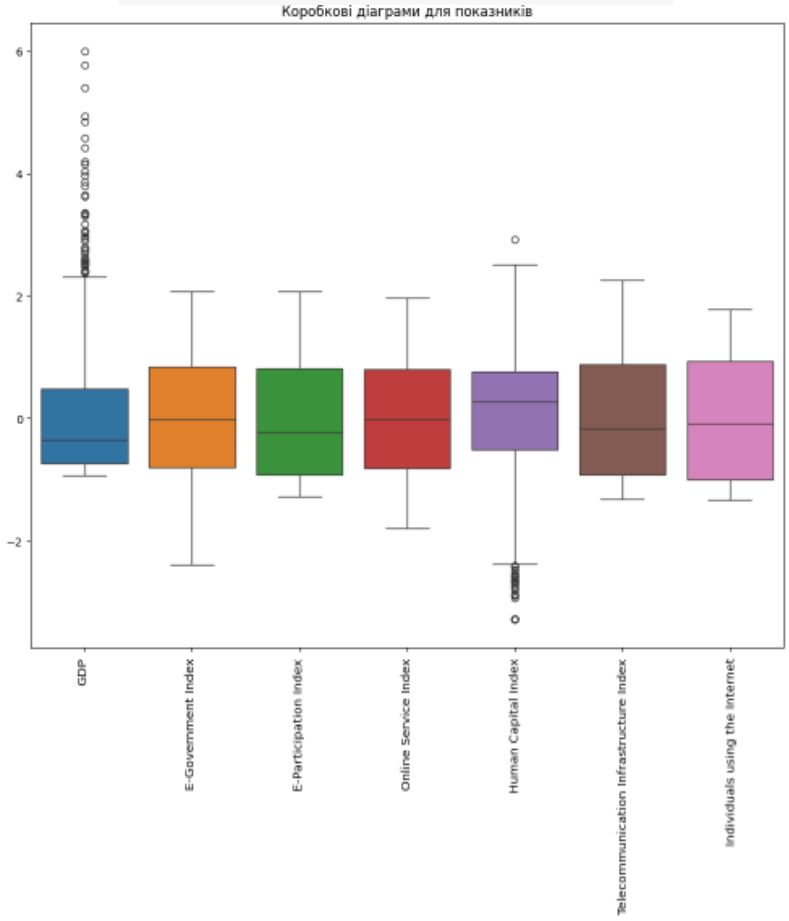


Рисунок А.19 — Побудова коробкових діаграм

```
# Створення PCA з 5 компонентами
# Вибираємо стовпці з роками
#x = df.filter(like='200').dropna()  # Вибираємо дані за всі роки та видаляємо рядки з NaN

# Стандартизація даних
#scaler = StandardScaler()
#x_scaled = scaler.fit_transform(x)

# PCA з трьома компонентами
x=df_scaled.drop(['GDP'], axis=1)
pca_test = PCA(n_components=6)
pca_test.fit(x)

# Графік накопиченої дисперсії
sns.set(style='whitegrid')
plt.figure(figsize=(8, 5))
plt.plot(np.cumsum(pca_test.explained_variance_ratio_))
plt.xlabel('Number of Components')
plt.ylabel('Cumulative Explained Variance')
plt.title('Explained Variance vs Number of Components')
plt.axvline(linewidth=3, color='r', linestyle='--', x=5, ymin=0, ymax=1)
plt.grid(True)
plt.show()

# Розрахунок поясненої та накопиченої дисперсії
evr = pca_test.explained_variance_ratio_
cvr = np.cumsum(evr)

# Створення датафрейму з результатами
pca_df = pd.DataFrame({
    'Cumulative Variance Ratio': cvr,
    'Explained Variance Ratio': evr
})

# Відображення результатів
display(pca_df.head(10))
```

Рисунок А.20 — Створення PCA з 5 компонентами

```
# Створюємо об'єкт PCA з 3-ма компонентами
pca = PCA(n_components=3)

# Застосовуємо PCA до даних x
principalComponents = pca.fit_transform(x)

# Створюємо DataFrame з результатами перших двох головних компонент
principalDf = pd.DataFrame(data=principalComponents, columns=['principal component 1', 'principal component 2', 'principal component 3'])

# Виводимо перші рядки DataFrame для перевірки
principalDf.head()
```

Рисунок А.21 — Створення PCA з 3 компонентами

```
#додавання стовпців 'Year' та 'Country Name'
df2 = pd.read_excel('Data.xlsx')
df2 = df2[['Year', 'Country Name', 'GDP']]
result = pd.concat([principalDf, df2], axis=1)
```

```
result
```

	principal component 1	principal component 2	principal component 3	Year	Country Name	GDP
0	-3.82221	-1.28199	0.56428	2003	Afghanistan	970.71623
1	-2.55004	1.08457	-0.20968	2003	Albania	5000.30908
2	-2.07069	0.31531	-0.59757	2003	Algeria	10843.16846
3	-2.20021	0.37755	0.45612	2003	Andorra	38686.47947
4	-3.26659	-1.02898	0.15936	2003	Angola	3839.85414
...	...	...	...	...	...	...
1381	2.18124	-0.42730	0.40441	2022	Uzbekistan	9042.34392
1382	0.08954	-0.54719	1.05046	2022	Vanuatu	3203.61662
1383	1.68333	-0.59945	0.78864	2022	Viet Nam	14051.24877
1384	-0.35878	-0.27080	0.02682	2022	Zambia	3864.89437
1385	-0.74315	-0.11911	0.37162	2022	Zimbabwe	3660.83550

1386 rows × 6 columns

Рисунок А.22 — Додавання стовпців

```
# Групування за країною
result_agg = result.groupby('Country Name').mean(numeric_only=True)
result_agg
```

Country Name	principal component 1	principal component 2	principal component 3	Year	GDP
Afghanistan	-2.94210	-1.45785	0.33586	2012.00000	1787.74306
Albania	-0.19072	0.27733	-0.16372	2012.00000	10494.28974
Algeria	-1.62632	0.35798	0.42195	2012.00000	13506.48781
Andorra	0.30277	0.33138	1.18751	2012.00000	49951.67813
Angola	-2.20387	-0.81103	-0.03571	2012.00000	6367.58923
...	...	...	...	...	...
Uzbekistan	-0.33429	0.32491	-0.54983	2012.00000	5769.11085
Vanuatu	-2.17262	-0.24502	0.43070	2012.00000	2661.77556
Viet Nam	-0.34976	0.00218	-0.04440	2012.00000	7162.78779
Zambia	-2.01765	-0.18145	-0.21533	2012.00000	2989.01832
Zimbabwe	-1.83133	0.23348	-0.06326	2012.00000	2286.42385

126 rows × 5 columns

Рисунок А.23 — Групування за країною

```
scaled_data = result_agg[['principal component 1', 'principal component 2', 'principal component 3']]
#  Метод локтя для вибору оптимальної кількості кластерів
wcss = []  # Within-Cluster Sum of Squares
max_clusters = 10

for k in range(1, max_clusters + 1):
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(scaled_data)
    wcss.append(kmeans.inertia_)

#  Візуалізація методу локтя
plt.figure(figsize=(10, 6))
plt.plot(range(1, max_clusters + 1), wcss, marker='o', linestyle='--')
plt.title('Метод ліктя для вибору оптимальної кількості кластерів')
plt.xlabel('Кількість кластерів')
plt.ylabel('WCSS')
plt.grid()
plt.show()

#  Вибір оптимальної кількості кластерів (наприклад, з графіка — припустимо 3)
optimal_k = 3

#  Кластеризація з оптимальною кількістю кластерів
kmeans = KMeans(n_clusters=optimal_k, random_state=42)
clusters = kmeans.fit_predict(scaled_data)

# Додавання результатів кластеризації до DataFrame
result_agg['Cluster'] = clusters


#  Виведення перших рядків із результатами кластеризації
print(df.head())

#  Зменшення до 2D для візуалізації
pca = PCA(n_components=2)
reduced_data = pca.fit_transform(scaled_data)

# Створення DataFrame з PCA-результатами та кластерами
visualization_df = pd.DataFrame(reduced_data, columns=['PC1', 'PC2'])
visualization_df['Cluster'] = clusters
```

```
#  Візуалізація кластерів
plt.figure(figsize=(10, 6))
sns.scatterplot(data=visualization_df, x='PC1', y='PC2', hue='Cluster', palette='tab10', s=100)
plt.title('Візуалізація кластерів (2D PCA)')
plt.xlabel('Головна компонента 1 (PC1)')
plt.ylabel('Головна компонента 2 (PC2)')
plt.legend(title='Кластер')
plt.grid()
plt.show()
```

Рисунок А.24 — Метод ліктя

```
#присвоюємо значення кластера країнам у новому стовпці 'Cluster' в основному DataFrame
result['Cluster'] = result['Country Name'].map(result_agg['Cluster'])

result
```

Рисунок А.25 — Значення кластерів для країн

```
cluster_0 = result[result['Cluster'] == 0]
cluster_1 = result[result['Cluster'] == 1]
cluster_2 = result[result['Cluster'] == 2]
```

```
# Створення дамі-змінних для cluster_0
dummy_variables = pd.get_dummies(cluster_0['Country Name'], prefix='Country', dtype=int)

# Додавання дамі-змінних до початкового датафрейму
cluster_0 = pd.concat([cluster_0, dummy_variables], axis=1)
```

Рисунок А.26 — Дамі-змінні для cluster_0

```
y = cluster_0['GDP']
x = cluster_0.drop(['GDP', 'Country Name','Cluster'], axis=1)
```

```
x = sm.add_constant(x)

# Побудова моделі OLS для cluster_0
model = sm.OLS(y, x).fit()

# Виведення результатів для кожного року
#print(f"OLS Regression Results for {year}:")
print(model.summary())
print("\n")
```

Рисунок А.27 — Побудова моделі OLS для cluster_0

```
# Створення дамі-змінних cluster_1
dummy_variables = pd.get_dummies(cluster_1['Country Name'], prefix='Country', dtype=int)

# Додавання дамі-змінних до початкового датафрейму
cluster_1 = pd.concat([cluster_1, dummy_variables], axis=1)
```

```
cluster_1
```

Рисунок А.28 — Дамі-змінні для cluster_1

```
y = cluster_1['GDP']
x = cluster_1.drop(['GDP', 'Country Name','Cluster'], axis=1)
```

```
x = sm.add_constant(x)

# Побудова моделі OLS для cluster_1
model = sm.OLS(y, x).fit()

# Виведення результатів для кожного року
#print(f"OLS Regression Results for {year}:")
print(model.summary())
print("\n")
```

Рисунок А.29 — Побудова моделі OLS для cluster_1

```
# Створення дамі-змінних cluster_2
dummy_variables = pd.get_dummies(cluster_2['Country Name'], prefix='Country', dtype=int)

# Додавання дамі-змінних до початкового датафрейму
cluster_2 = pd.concat([cluster_2, dummy_variables], axis=1)
```

```
cluster_2
```

Рисунок А.30 — Дамі-змінні для cluster_2

```
y = cluster_2['GDP']
x = cluster_2.drop(['GDP', 'Country Name','Cluster'], axis=1)
```

```
x = sm.add_constant(x)

# Побудова моделі OLS для cluster_2
model = sm.OLS(y, x).fit()

# Виведення результатів для кожного року
#print(f"OLS Regression Results for {year}:")
print(model.summary())
print("\n")
```

Рисунок А.31 — Побудова моделі OLS для cluster_2

```
!pip install xgboost
```

```
Requirement already satisfied: xgboost in /usr/local/lib/python3.10/dist-packages (2.1.3)
Requirement already satisfied: numpy in /usr/local/lib/python3.10/dist-packages (from xgboost) (1.26.4)
Requirement already satisfied: nvidia-nccl-cu12 in /usr/local/lib/python3.10/dist-packages (from xgboost) (2.23.4)
Requirement already satisfied: scipy in /usr/local/lib/python3.10/dist-packages (from xgboost) (1.13.1)
```

Рисунок А.32 — Встановлення бібліотеки

```python
#регресія
import matplotlib.pyplot as plt

# Список унікальних кластерів
clusters = df_long['Cluster'].unique()

# Візуалізація для кожного кластеру
for cluster in clusters:
    print(f"\nКластер {cluster}")

    # Вибір даних для конкретного кластера
    cluster_data = df_long[df_long['Cluster'] == cluster]

    # Вибір предикторів (X) і цільової змінної (y)
    X = cluster_data.drop(['GDP', 'Cluster'], axis=1)
    y = cluster_data['GDP']

    # Перетворення категорійних змінних у дамі-змінні
    categorical_features = ['Country Name', 'Year']
    preprocessor = ColumnTransformer(
        transformers=[('cat', OneHotEncoder(drop='first'), categorical_features)],
        remainder='passthrough'
    )
    X_transformed = preprocessor.fit_transform(X)

    # Розділення даних на тренувальний і тестовий набори
    X_train, X_test, y_train, y_test = train_test_split(X_transformed, y, test_size=0.2, random_state=42)

    # Ініціалізація і навчання моделі
    model = LinearRegression()
    model.fit(X_train, y_train)

    # Прогнозування
    y_pred = model.predict(X_test)

    # Обчислення метрик
    mse = mean_squared_error(y_test, y_pred)
    r2 = r2_score(y_test, y_pred)

    # Вивід метрик для поточного кластера
    print(f"MSE для кластера {cluster}: {mse:.4f}")
    print(f"R² для кластера {cluster}: {r2:.4f}")
```

```python
    # Візуалізація
    plt.figure(figsize=(8, 6))
    plt.scatter(y_test, y_pred, alpha=0.7, label='Прогнози', color='blue')
    plt.plot([y_test.min(), y_test.max()], [y_test.min(), y_test.max()],
             color='red', linestyle='--', label='Ідеальна лінія')
    plt.title(f'Кластер {cluster}: Справжні значення vs Прогнози')
    plt.xlabel('Справжні значення')
    plt.ylabel('Прогнозовані значення')
    plt.legend()
    plt.grid(alpha=0.5)
    plt.show()
```

Рисунок А.33 — Регресія для кластерів

```python
# Список унікальних кластерів
clusters = df_long['Cluster'].unique()

# Візуалізація для кожного кластеру
for cluster in clusters:
    print(f"\nКластер {cluster}")

    # Вибір даних для конкретного кластера
    cluster_data = df_long[df_long['Cluster'] == cluster]

    # Вибір предикторів (X) і цільової змінної (y)
    X = cluster_data.drop(['GDP', 'Cluster'], axis=1)
    y = cluster_data['GDP']

    # Перетворення категорійних змінних у дамі-змінні
    categorical_features = ['Country Name', 'Year']
    preprocessor = ColumnTransformer(
        transformers=[('cat', OneHotEncoder(drop='first'), categorical_features)],
        remainder='passthrough'
    )
    X_transformed = preprocessor.fit_transform(X)

    # Розділення даних на тренувальний і тестовий набори
    X_train, X_test, y_train, y_test = train_test_split(X_transformed, y, test_size=0.2, random_state=42)

    # Ініціалізація і навчання моделі XGBoost
    model = xgb.XGBRegressor(objective='reg:squarederror', n_estimators=100, random_state=42)
    model.fit(X_train, y_train)

    # Прогнозування
    y_pred = model.predict(X_test)

    # Обчислення метрик
    mse = mean_squared_error(y_test, y_pred)
    r2 = r2_score(y_test, y_pred)

    # Вивід метрик для поточного кластера
    print(f"MSE для кластера {cluster}: {mse:.4f}")
    print(f"R² для кластера {cluster}: {r2:.4f}")

    # Візуалізація результатів
    plt.figure(figsize=(8, 6))
    plt.scatter(y_test, y_pred, alpha=0.7, label='Прогнози', color='blue')
    plt.plot([y_test.min(), y_test.max()], [y_test.min(), y_test.max()],
             color='red', linestyle='--', label='Ідеальна лінія')
    plt.title(f'Кластер {cluster}: Справжні значення vs Прогнози (XGBoost)')
    plt.xlabel('Справжні значення')
    plt.ylabel('Прогнозовані значення')
    plt.legend()
    plt.grid(alpha=0.5)
    plt.show()
```

Рисунок А.34 — Прогнозування (XGBoost)

```
# Список унікальних кластерів
clusters = df_long['Cluster'].unique()

# Візуалізація для кожного кластера
for cluster in clusters:
    print(f"\nКластер {cluster}")

    # Вибір даних для конкретного кластера
    cluster_data = df_long[df_long['Cluster'] == cluster]

    # Вибір предикторів (X) і цільової змінної (y)
    X = cluster_data.drop(['GDP', 'Cluster'], axis=1)
    y = cluster_data['GDP']

    # Перетворення категорійних змінних у дамі-змінні
    categorical_features = ['Country Name', 'Year']
    preprocessor = ColumnTransformer(
        transformers=[('cat', OneHotEncoder(drop='first'), categorical_features)],
        remainder='passthrough'
    )
    X_transformed = preprocessor.fit_transform(X)

    # Розділення даних на тренувальний і тестовий набори
    X_train, X_test, y_train, y_test = train_test_split(X_transformed, y, test_size=0.2, random_state=42)

    # Ініціалізація та навчання моделі дерева рішень
    tree_model = DecisionTreeRegressor(random_state=42)
    tree_model.fit(X_train, y_train)

    # Прогнозування
    y_pred = tree_model.predict(X_test)

    # Обчислення метрик
    mse = mean_squared_error(y_test, y_pred)
    r2 = r2_score(y_test, y_pred)

    # Виведення метрик для поточного кластера
    print(f"MSE для кластера {cluster}: {mse:.4f}")
    print(f"R² для кластера {cluster}: {r2:.4f}")

    # Візуалізація результатів
    plt.figure(figsize=(8, 6))
    plt.scatter(y_test, y_pred, alpha=0.7, label='Прогнози', color='blue')
    plt.plot([y_test.min(), y_test.max()], [y_test.min(), y_test.max()],
             color='red', linestyle='--', label='Ідеальна лінія')
    plt.title(f'Кластер {cluster}: Справжні значення vs Прогнози (Дерево рішень)')
    plt.xlabel('Справжні значення')
    plt.ylabel('Прогнозовані значення')
    plt.legend()
    plt.grid(alpha=0.5)
    plt.show()
```

Рисунок А.35 — Прогнозування (Дерево рішень)

```
# Список унікальних кластерів
clusters = df_long['Cluster'].unique()

# Візуалізація для кожного кластера
for cluster in clusters:
    print(f"\nКластер {cluster}")

    # Вибір даних для конкретного кластера
    cluster_data = df_long[df_long['Cluster'] == cluster]

    # Вибір предикторів (X) і цільової змінної (y)
    X = cluster_data.drop(['GDP', 'Cluster'], axis=1)
    y = cluster_data['GDP']

    # Перетворення категорійних змінних у дамі-змінні
    categorical_features = ['Country Name', 'Year']
    preprocessor = ColumnTransformer(
        transformers=[('cat', OneHotEncoder(drop='first'), categorical_features)],
        remainder='passthrough'
    )
    X_transformed = preprocessor.fit_transform(X)

    # Розділення даних на тренувальний і тестовий набори
    X_train, X_test, y_train, y_test = train_test_split(X_transformed, y, test_size=0.2, random_state=42)

    # Ініціалізація та навчання моделі випадкового лісу
    rf_model = RandomForestRegressor(n_estimators=100, random_state=42)
    rf_model.fit(X_train, y_train)

    # Прогнозування
    y_pred = rf_model.predict(X_test)

    # Обчислення метрик
    mse = mean_squared_error(y_test, y_pred)
    r2 = r2_score(y_test, y_pred)

    # Виведення метрик для поточного кластера
    print(f"MSE для кластера {cluster}: {mse:.4f}")
    print(f"R² для кластера {cluster}: {r2:.4f}")

    # Візуалізація результатів
    plt.figure(figsize=(8, 6))
    plt.scatter(y_test, y_pred, alpha=0.7, label='Прогнози', color='green')
    plt.plot([y_test.min(), y_test.max()], [y_test.min(), y_test.max()],
             color='red', linestyle='--', label='Ідеальна лінія')
    plt.title(f'Кластер {cluster}: Справжні значення vs Прогнози (Випадковий ліс)')
    plt.xlabel('Справжні значення')
    plt.ylabel('Прогнозовані значення')
    plt.legend()
    plt.grid(alpha=0.5)
    plt.show()
```

Рисунок А.36 — Прогнозування (Випадковий ліс)

```python
# Дані для графіків
clusters = ['Кластер 1', 'Кластер 2', 'Кластер 0']
models = ['XGBoost', 'Decision Tree', 'Random Forest', 'regression']

# MSE для кожної моделі і кластеру

mse_values = {
    ('Кластер 1', 'XGBoost'): 757479.4562,
    ('Кластер 1', 'Decision Tree'): 1462042.5229,
    ('Кластер 1', 'Random Forest'): 991266.7417,
    ('Кластер 1', 'regression'): 1087983.8325,
    ('Кластер 2', 'XGBoost'): 12506671.0764,
    ('Кластер 2', 'Decision Tree'): 36149939.8800,
    ('Кластер 2', 'Random Forest'): 20966984.2695,
    ('Кластер 2', 'regression'): 12235214.8491,
    ('Кластер 0', 'XGBoost'): 85180848.6047,
    ('Кластер 0', 'Decision Tree'): 187638537.0812,
    ('Кластер 0', 'Random Forest'): 175936964.8707,
    ('Кластер 0', 'regression'): 91108973.3202
}

# R2 для кожної моделі і кластеру
r2_values = {
    ('Кластер 1', 'XGBoost'): 0.9205,
    ('Кластер 1', 'Decision Tree'): 0.8466,
    ('Кластер 1', 'Random Forest'): 0.8960,
    ('Кластер 1', 'regression'): 0.8858,
    ('Кластер 2', 'XGBoost'): 0.9304,
    ('Кластер 2', 'Decision Tree'): 0.7987,
    ('Кластер 2', 'Random Forest'): 0.8832,
    ('Кластер 2', 'regression'): 0.9319,
    ('Кластер 0', 'XGBoost'): 0.8480,
    ('Кластер 0', 'Decision Tree'): 0.6652,
    ('Кластер 0', 'Random Forest'): 0.6861,
    ('Кластер 0', 'regression'): 0.8374
}

# Побудова графіків

#  Графік стовпців для MSE
fig, ax = plt.subplots(figsize=(10, 6))
mse_matrix = np.array([[mse_values[(cluster, model)] for model in models] for cluster in clusters])

# Створення графіку
sns.heatmap(mse_matrix, annot=True, xticklabels=models, yticklabels=clusters, cmap="YlGnBu", fmt=".2f", ax=ax)
ax.set_title('Порівняння MSE для моделей по кластерах')
ax.set_xlabel('Моделі')
ax.set_ylabel('Кластери')
plt.show()
```

```python
#  Графік стовпців для R2
fig, ax = plt.subplots(figsize=(10, 6))
r2_matrix = np.array([[r2_values[(cluster, model)] for model in models] for cluster in clusters])

# Створення графіку
sns.heatmap(r2_matrix, annot=True, xticklabels=models, yticklabels=clusters, cmap="YlGnBu", fmt=".2f", ax=ax)
ax.set_title('Порівняння R2 для моделей по кластерах')
ax.set_xlabel('Моделі')
ax.set_ylabel('Кластери')
plt.show()
```

Рисунок А.37 — Порівняння MSE та R2 для моделей по кластерах