

**НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»**

Навчально-науковий інститут прикладного системного аналізу

Кафедра математичних методів системного аналізу

До захисту допущено:

Завідувач кафедри

_____ Оксана ТИМОЩУК

«__» _____ 2026 р.

Дипломна робота

на здобуття ступеня бакалавра

за освітньо-професійною програмою «Системний аналіз і управління»

спеціальності 124 «Системний аналіз»

**на тему: «Математичне моделювання та прогнозування
показників економічного стану населення в умовах глобальних
демографічних трансформацій»**

Виконав (-ла):

студент (-ка) IV курсу, групи КА-23,

Коваленко Олександра Павлівна _____

Керівник:

доцент кафедри ШІ, доктор філософії, доцент,

Гуськова Віра Геннадіївна _____

Консультант з економічного розділу:

доцент кафедри ЕК, к. е. н., доцент,

Рощина Надія Василівна _____

Консультант з нормконтролю:

асистент кафедри ММСА, доктор філософії,

Канцедал Георгій Олегович _____

Рецензент:

асистент кафедри ШІ, доктор філософії,

Гаврилович Марія Павлівна _____

Засвідчую, що у цій дипломній роботі
немає запозичень з праць інших авторів
без відповідних посилань.

Студентка _____

Київ – 2026 року

Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Навчально-науковий інститут прикладного системного аналізу
Кафедра математичних методів системного аналізу

Рівень вищої освіти – перший (бакалаврський)

Спеціальність – 124 «Системний аналіз»

Освітньо-професійна програма «Системний аналіз і управління»

ЗАТВЕРДЖУЮ

Завідувач кафедри

_____ Оксана ТИМОЩУК

«____» _____ 2026 р.

ЗАВДАННЯ
на дипломну роботу студенту
Коваленко Олександрі Павлівни

1. Тема роботи «Математичне моделювання та прогнозування показників економічного стану населення в умовах глобальних демографічних трансформацій», керівник роботи Гуськова Віра Геннадіївна, доцент кафедри штучного інтелекту, д.ф., затверджені наказом по університету від «____» _____ 2026 р. № 1944-с
2. Термін подання студентом роботи 10.06.2026_____
3. Вихідні дані до роботи: статистичні дані World Bank Open Data та OECD Data за 2000–2024 роки. для п'яти європейських країн (Україна, Польща, Німеччина, Італія, Іспанія): індекс Джині, рівень безробіття, частка населення 65+, коефіцієнт фертильності, інфляція, чиста міграція, ВВП на душу населення.

4. Зміст роботи: 1 – Аналіз предметної галузі, огляд існуючих програмних рішень та методів моделювання соціально-економічної нерівності; 2 – Теоретичні основи методів математичного моделювання та прогнозування; 3 – Практична реалізація системи моделювання: регресійний аналіз, Random Forest, ARIMA, кластеризація K-Means, сценарне моделювання та DSS; 4 – Функціонально-вартісний аналіз програмного продукту.

5. Перелік ілюстративного матеріалу (із зазначенням плакатів, презентацій тощо): схема аналітичного pipeline DSS-системи; графіки та таблиці порівняння регресійних моделей та Random Forest; графіки прогнозування часових рядів ARIMA для України; кластерна карта країн за показниками нерівності; матриця сценарного аналізу; графіки бальних залежностей параметрів та таблиці до функціонально-вартісного аналізу; презентація.

6. Консультанти розділів роботи

	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
Економічний	Доцент, к. е. н. Рощина Надія Василівна		

7. Дата видачі завдання _____

Календарний план

№ з/п	Назва етапів виконання дипломної роботи	Термін виконання етапів роботи	Примітка
1	Затвердження теми дипломної роботи	07.02.2026	Виконано
2	Огляд та аналіз наукової літератури, огляд аналогів	27.02.2026	Виконано
3	Збір та попередня обробка даних (World Bank API)	7.03.2026	Виконано
4	Побудова та верифікація математичних моделей (регресія, Random Forest, ARIMA, K-Means)	29.03.2026	Виконано
5	Розробка DSS-блоку та сценарного модуля	13.04.2026	Виконано
6	Написання основної частини дипломної роботи	20.04.2026	Виконано

7	Проходження нормоконтролю	04.06.2026	Виконано
8	Підготовка презентації	05.06.2026	Виконано
9	Попередній захист дипломної роботи	06.06.2026	Виконано

Студент

Олександра КОВАЛЕНКО

Керівник

Віра ГУСЬКОВА

РЕФЕРАТ

Дипломна робота містить 100 с., 11 табл., 16 рис., 1 дод., 15 джерела.

МАТЕМАТИЧНЕ МОДЕЛЮВАННЯ, ПРОГНОЗУВАННЯ, СОЦІАЛЬНО-ЕКОНОМІЧНА НЕРІВНІСТЬ, ІНДЕКС ДЖІНІ, ДЕМОГРАФІЧНІ ТРАНСФОРМАЦІЇ, РЕГРЕСІЙНИЙ АНАЛІЗ, RANDOM FOREST, ARIMA, КЛАСТЕРИЗАЦІЯ K-MEANS, СЦЕНАРНЕ МОДЕЛЮВАННЯ, СИСТЕМА ПІДТРИМКИ ПРИЙНЯТТЯ РІШЕНЬ, WORLD BANK API, PYTHON.

Тема: Математичне моделювання та прогнозування показників економічного стану населення в умовах глобальних демографічних трансформацій.

Об'єкт дослідження — соціально-економічні та демографічні процеси в п'яти європейських країнах (Україна, Польща, Німеччина, Італія, Іспанія) в умовах сучасних демографічних трансформацій за 2000–2024 роки.

Предмет дослідження — математичні моделі та методи прогнозування індексу Джині на основі демографічних (частка населення 65+, коефіцієнт фертильності, чиста міграція) та макроекономічних (ВВП на душу населення, рівень безробіття, інфляція) чинників, реалізовані засобами регресійного аналізу, ансамблевого машинного навчання, моделей часових рядів і кластеризації.

Мета роботи — розробити систему математичного моделювання та прогнозування індексу Джині для п'яти європейських країн, що поєднує множинну лінійну регресію, алгоритм Random Forest, модель ARIMA, кластеризацію K-Means та сценарний аналіз в єдиному аналітичному контурі системи підтримки прийняття рішень (DSS).

Результат роботи — аналітична DSS-система, реалізована мовою Python, що автоматизує збір даних через World Bank API, формує багатовимірну матрицю показників (125×9), порівнює якість чотирьох прогнозних моделей

за метриками MAE, RMSE, MAPE та R^2 , будує середньостроковий прогноз індексу Джині для України методом ARIMA, виконує кластеризацію країн і генерує три сценарії розвитку (базовий, оптимістичний, песимістичний). Результати збережено у форматі CSV для подальшого використання в DSS.

Програмний продукт розроблено із використанням мови програмування Python та бібліотек pandas, numpy, scikit-learn, statsmodels, matplotlib, seaborn, wldata.

ABSTRACT

The diploma work contains 100 pages, 11 tables, 16 figures, 1 appendices, 15 references.

MATHEMATICAL MODELLING, FORECASTING, SOCIO-ECONOMIC INEQUALITY, GINI INDEX, DEMOGRAPHIC TRANSFORMATIONS, REGRESSION ANALYSIS, RANDOM FOREST, ARIMA, K-MEANS CLUSTERING, SCENARIO MODELLING, DECISION SUPPORT SYSTEM, WORLD BANK API, PYTHON.

Topic: Mathematical Modelling and Forecasting of Population Economic Indicators under Global Demographic Transformations.

Object of the study – socio-economic and demographic processes in five European countries (Ukraine, Poland, Germany, Italy, Spain) under contemporary demographic transformations over the period 2000–2024.

Subject of the study – mathematical models and methods for forecasting the Gini index based on demographic factors (share of population aged 65+, fertility rate, net migration) and macroeconomic indicators (GDP per capita, unemployment rate, inflation), implemented through regression analysis, ensemble machine learning, time series models and clustering.

Purpose of the work – to develop a mathematical modelling and forecasting system for the Gini index across five European countries, integrating multiple linear regression, Random Forest, ARIMA, K-Means clustering and scenario analysis into a unified analytical pipeline of a decision support system (DSS).

Result of the work – an analytical DSS prototype implemented in Python that automates data collection via the World Bank API, constructs a multidimensional indicator matrix (125×9), compares the performance of four forecasting models using MAE, RMSE, MAPE and R^2 metrics, builds a medium-term ARIMA forecast of Ukraine's Gini index, performs country clustering, and generates three

development scenarios (baseline, optimistic, pessimistic). Outputs are saved in CSV format for further use in the DSS.

The software product was developed using the Python programming language and the pandas, numpy, scikit-learn, statsmodels, matplotlib, seaborn and wldata libraries.

ЗМІСТ

ПЕРЕЛІК ПРИЙНЯТИХ СКОРОЧЕНЬ	11
ВСТУП	12
РОЗДІЛ 1 ТЕОРЕТИЧНІ ОСНОВИ АНАЛІЗУ ТА ПРОГНОЗУВАННЯ ЕКОНОМІЧНОГО СТАНУ НАСЕЛЕННЯ	14
1.1 Актуальність роботи	14
1.2 Аналіз існуючої літератури.....	15
1.3 Аналіз існуючих рішень	17
1.3.1 Аналіз існуючих програмних рішень.....	18
1.3.2 Аналіз мов програмування та бібліотек.....	20
1.4 Постановка задачі.....	21
1.5 Висновки до розділу 1	23
РОЗДІЛ 2 МАТЕМАТИЧНА ПОСТАНОВКА ТА ВИКОРИСТАННЯ МОДЕЛЕЙ ПРОГНОЗУВАННЯ ТА ПІДГОТОВКИ ДАНИХ.....	25
2.1 Соціально-економічний стан населення та вплив демографічних трансформацій	25
2.2 Методи математичного моделювання та прогнозування соціально- економічних процесів	27
2.2.1 Модель множинної лінійної регресії.....	27
2.2.2 Модель випадкового лісу (Random Forest Regressor).....	28
2.2.3 Методи прогнозування часових рядів.....	29
2.2.4 Кластеризація та сценарне моделювання	30
2.3 Архітектура DSS та критерії оцінки моделей	31
2.4 Джерела даних та сучасні інформаційні технології аналізу.....	33
2.5 Висновки до розділу 2	35
РОЗДІЛ 3 АНАЛІЗ РЕЗУЛЬТАТІВ ДОСЛІДЖЕНЬ ЗА ТЕМОЮ ДИПЛОМНОЇ РОБОТИ	36
3.1 Формування та попередня обробка набору даних	36

	10
3.2 Аналіз соціально-економічних та демографічних показників	39
3.3 Побудова моделей прогнозування та оцінка їх ефективності	44
3.3.1 Регресійні моделі та Random Forest	44
3.3.2 Прогнозування часових рядів методом ARIMA	48
3.4 Кластеризація та сценарне моделювання соціально-економічних процесів	50
3.5 Висновки до розділу 3	53
РОЗДІЛ 4 ФУНКЦІОНАЛЬНО-ВАРТІСНИЙ АНАЛІЗ СИСТЕМИ	
МАТЕМАТИЧНОГО МОДЕЛЮВАННЯ ТА ПРОГНОЗУВАННЯ	
ПОКАЗНИКІВ ЕКОНОМІЧНОГО СТАНУ НАСЕЛЕННЯ	55
4.1 Постановка задачі проектування	55
4.2 Обґрунтування функцій програмного продукту	56
4.3 Обґрунтування системи параметрів програмного продукту	59
4.4 Аналіз експертного оцінювання параметрів	62
4.5 Аналіз рівня якості варіантів реалізації функцій	65
4.6 Економічний аналіз варіантів розробки програмного продукту	67
4.7 Вибір кращого варіанту за техніко-економічним рівнем	70
4.8 Висновки до розділу 4	71
ВИСНОВКИ	73
ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ	76
ДОДАТОК А Лістинг програми	78
ДОДАТОК Б Презентація	101

ПЕРЕЛІК ПРИЙНЯТИХ СКОРОЧЕНЬ

ARIMA — AutoRegressive Integrated Moving Average
SARIMA — Seasonal AutoRegressive Integrated Moving Average
XGBoost — eXtreme Gradient Boosting
AIC — Akaike Information Criterion
BIC — Bayesian Information Criterion
MSE — Mean Squared Error
RMSE — Root Mean Squared Error
MAE — Mean Absolute Error
MAPE — Mean Absolute Percentage Error
DSS — Decision Support System (система підтримки прийняття рішень)
API — Application Programming Interface
VAR — Vector AutoRegression
ADF — Augmented Dickey-Fuller (test)
ВВП — валовий внутрішній продукт
ООН — Організація Об'єднаних Націй
МНК — метод найменших квадратів
CSV — Comma-Separated Values

ВСТУП

У сучасних умовах глобальних демографічних трансформацій дослідження економічного стану населення набуває особливої наукової та прикладної значущості. Старіння населення, міграційні процеси, зміна рівня народжуваності, трансформація структури ринку праці та нерівномірність розподілу доходів формують складну систему взаємопов'язаних соціально-економічних чинників, які безпосередньо впливають на розвиток держав, стійкість бюджетної політики, функціонування системи соціального забезпечення та якість життя населення.

Актуальність теми дипломної роботи зумовлена необхідністю кількісного оцінювання впливу демографічних і макроекономічних факторів на показники економічного стану населення, зокрема на рівень соціально-економічної нерівності. В умовах високої динамічності досліджуваних процесів традиційні описові методи аналізу є недостатніми, оскільки вони не забезпечують повного розуміння причинно-наслідкових зв'язків і не дають змоги формувати обґрунтовані прогнози. Саме тому виникає потреба у використанні сучасних методів математичного моделювання, економетричного аналізу, машинного навчання та аналізу часових рядів.

Особливого значення ця проблема набуває для країн Європи, зокрема України, Польщі, Німеччини, Італії та Іспанії, які мають різні демографічні структури, рівні економічного розвитку та соціальні моделі. Порівняльний аналіз таких країн дає змогу виявити спільні закономірності та специфічні особливості впливу демографічних змін на економічний стан населення.

Метою дипломної роботи є розроблення системи математичного моделювання та прогнозування показників економічного стану населення під впливом демографічних і макроекономічних чинників.

Об'єктом дослідження є соціально-економічні та демографічні процеси в європейських країнах в умовах сучасних глобальних трансформацій.

Предметом дослідження є математичні моделі, методи прогнозування та алгоритми аналізу взаємозв'язків між демографічними факторами і показниками економічного стану населення.

Практична значущість роботи полягає у можливості використання отриманих результатів для підтримки прийняття управлінських рішень у сфері соціально-економічної політики, оцінювання ризиків нерівності доходів та формування сценаріїв подальшого розвитку досліджуваних країн.

РОЗДІЛ 1 ТЕОРЕТИЧНІ ОСНОВИ АНАЛІЗУ ТА ПРОГНОЗУВАННЯ ЕКОНОМІЧНОГО СТАНУ НАСЕЛЕННЯ

В умовах сучасних глобальних демографічних трансформацій аналіз економічного стану населення набирає все більше актуальності. Питання застосування математичних методів та алгоритмів машинного навчання до аналізу соціально-економічних процесів активно висвітлюються у науково-технічному журналі «Системні дослідження та інформаційні технології» [1] та журналі «Наукові вісті КПП» [2], що підтверджує актуальність обраного напрямку дослідження. Старіння населення, міграційні процеси, розподіл доходів, зміни структур ринку праці та зростання макроекономічних процесів створюють багатофакторні системи, для дослідження яких необхідно використовувати методи системного аналізу, економетричного моделювання та машинного навчання [3-5]. У таких умовах традиційні методи описової статистики є недостатніми для прийняття обґрунтованих управлінських рішень, з'являється необхідність у переході до кількісного моделювання та середньострокового прогнозування.

1.1 Актуальність роботи

Актуальність теми визначається декількома взаємопов'язаними чинниками. По-перше, в умовах воєнного стану Україна зіштовхнулася з великим відтоком населення. За результатами оцінки ООН, станом на 2024 рік кількість зареєстрованих українських біженців у Європі перевищила 6 мільйонів осіб, що вплинуло на пропозицію робочої сили та розподіл доходів. По-друге, загальноєвропейська тенденція зниження коефіцієнта фертильності нижче критичного порогу 2.1 дитини на жінку означає

поступове скорочення частки економічно активного населення та зростання демографічного навантаження. По-третє, збільшення частки осіб віком 65+ у країнах із розвиненою економікою (Німеччина, Італія) посилює нерівномірність розподілу доходів, оскільки пенсійні виплати формують окремий клас споживачів з обмеженим доходом.

Розглянуті обставини формують запит на розробку системи математичного моделювання та прогнозування, що зможе кількісно оцінити вплив демографічних і макроекономічних факторів соціально-економічної нерівності, зокрема індексу Джині, та генерувати аналітичні сценарії для підтримки прийняття рішень. Саме тому тема дипломної роботи є досить своєчасною та науково значущою [1].

1.2 Аналіз існуючої літератури

Теоретичною базою для дослідження є велика кількість існуючих наукових публікацій, що охоплюють необхідні напрями: економіку нерівності та демографії, методи математичного й економетричного моделювання, архітектуру систем підтримки прийняття рішень.

Сучасні дослідження у суміжних напрямках системного аналізу та інформаційних технологій публікуються у науково-технічному журналі «Системні дослідження та інформаційні технології» [1] та журналі «Наукові вісті КПІ» [2] Національного технічного університету України «КПІ ім. Ігоря Сікорського». Зазначені видання висвітлюють питання застосування математичних методів та алгоритмів машинного навчання до аналізу соціально-економічних процесів, що зайвий раз підкреслює актуальність обраного напрямку.

Концептуальна база аналізу нерівності сформована на основі праць класиків економічної теорії. Фундаментальний внесок у розуміння

вимірювання нерівності зробив А. Сен, що у роботі «On Economic Inequality» [6] продемонстрував обмеженість агрегованих показників та обґрунтував необхідність використання кривої Лоренца у динамічному контексті. Питання зв'язку між демографічними зрушеннями та соціально-економічними системами розглянуто в роботі Д. Медоуз «Thinking in Systems» [5], що заклала основу для системного підходу до аналізу складних взаємозалежних процесів.

У економічному моделюванні ж ключовими є роботи Дж. Вулдріджа «Introductory Econometrics: A Modern Approach» [7] та Д. Гуджараті «Basic Econometrics» [8]. У цих роботах детально розглянуто математичний апарат множинної лінійної регресії, методи діагностики мультиколінеарності та гетероскедастичності, а також принципи побудови панельних моделей для даних, що є методологічною основою даного дослідження.

Ансамблеві методи машинного навчання, зокрема алгоритм Random Forest, ґрунтуються на роботі Л. Бреймана «Random Forests» [9], де запропоновано теоретичне обґрунтування зменшення дисперсії через усереднення дерев рішень та введено поняття важливості ознак (feature importance). Методологія прогнозування часових рядів базується на класичному підручнику Бокса, Дженкінса та Рейнсела «Time Series Analysis: Forecasting and Control» [10], де викладено повний цикл ідентифікації, оцінювання та верифікації моделей класу ARIMA. Теоретичні основи кластеризації та статистичного навчання без учителя систематизовано у праці Гасті, Тібширані та Фрідмана «The Elements of Statistical Learning» [11].

Методологія сценарного моделювання базується на роботі М. Годе «Creating Futures: Scenario Planning as a Strategic Management Tool» [12], де систематизовано підходи до побудови альтернативних сценаріїв розвитку складних систем. Питання побудови систем підтримки прийняття рішень для соціально-економічних завдань висвітлено у роботі Турбана, Шарди та Делена «Decision Support and Business Intelligence Systems» [13], де визначено

архітектурні принципи багаторівневих DSS та описано методологію інтеграції аналітичних блоків.

Аналіз розглянутих літературних джерел свідчить про те, що, незважаючи на розвиненість окремих методів, комплексне поєднання регресійного аналізу, ансамблевого машинного навчання, моделей часових рядів і кластеризації в єдиній системі для кількісного оцінювання впливу демографічних трансформацій на соціально-економічну нерівність у контексті порівняльного аналізу європейських країн залишається недостатньо вивченим.

1.3 Аналіз існуючих рішень

Перед розробкою власної системи проведено аналіз існуючих підходів, програмних засобів і мов програмування, які використовуються для вирішення задач математичного моделювання та прогнозування соціально-економічних показників.

Серед існуючих підходів до аналізу соціально-економічної нерівності можна виділити декілька основних напрямів. Перший напрям — це використання регресійних моделей на панельних даних (fixed effects, random effects), що дає можливість контролювати незмінні характеристики країн і є стандартом у роботах Міжнародного валютного фонду та Світового банку [3]. Другий напрям — застосування методів машинного навчання, зокрема градієнтного бустингу (XGBoost, LightGBM) та нейронних мереж, для прогнозування макроекономічних показників, що описано в ряді публікацій ОЕСР [4]. Третій — побудова векторних авторегресійних моделей (VAR) для аналізу взаємовпливу між демографічними та економічними часовими рядами. Хоча все ж більшість існуючих рішень мають вузьку спрямованість

і не поєднують зазначені підходи в єдиній аналітичній платформі з можливістю сценарного моделювання.

1.3.1 Аналіз існуючих програмних рішень

Серед спеціалізованого програмного забезпечення варто виділити рішення.

1. EViews — пакет для економетричного аналізу і моделювання часових рядів, широко використовується у банківських і урядових структурах. Має потужні інструменти для побудови моделей ARIMA та VAR, але є закритим і дорогим, не підтримує сучасних алгоритмів машинного навчання та погано масштабується на великі набори даних.
2. STATA — статистичний пакет, популярний у соціальних науках, з розвиненими засобами регресійного аналізу панельних даних. Проте вартість ліцензії та відсутність нативної інтеграції з відкритими API (такими як World Bank API) ускладнюють автоматизований збір і оновлення даних.
3. Tableau / Power BI — інструменти бізнес-аналітики з сильними засобами візуалізації, але з наявними обмеженими можливостями для побудови математичних прогнозних моделей та кластерного аналізу.
4. Відкриті бібліотеки Python (такі як scikit-learn, statsmodels, pandas) — забезпечують повний аналітичний цикл від збору даних до побудови моделей і візуалізації результатів. Цей підхід обрано у роботі як найбільш гнучкий і здатний до інтеграції різнорідних методів.

Для узагальнення результатів аналізу розглянутих програмних засобів необхідне їх порівняння за ключовими критеріями в таблиці 1.1.

Таблиця 1.1 – Порівняння інструментів аналізу та прогнозування соціально-економічних показників

Інструмент	Переваги	Недоліки	Підтримка ML	Інтеграція з API	Вартість
EViews	Сильний економетричний аналіз, підтримка ARIMA та VAR	Закритий код, обмежені можливості ML	Ні	Обмежена	Висока
STATA	Розвинений регресійний аналіз, підтримка панельних даних	Платна ліцензія, складна інтеграція з відкритими джерелами даних	Частково	Обмежено	Висока
Tableau / Power BI	Зручна візуалізація, інтерактивні дашборди	Слабкі можливості прогнозування та кластеризації	Частково	Так	Середня / висока

Продовження таблиці 1.1

Python (pandas, scikit- learn, statsmodel s)	Повний аналітичний цикл, відкритий код, масштабовані сть, велика кількість бібліотек	Потребує програмува ння та налаштуван ня середовища	Так	Так	Безкош товно
---	--	--	-----	-----	-----------------

Як видно з таблиці 1.1, використання відкритих бібліотек Python є найбільш доцільним для реалізації поставлених завдань, оскільки вони забезпечують повний цикл обробки даних, побудови прогнозних моделей та візуалізації результатів без додаткових витрат на ліцензування.

1.3.2 Аналіз мов програмування та бібліотек

Вибір мови програмування зупинився на мові Python, яка є стандартом у сфері Data Science та аналізу соціально-економічних даних. Python має широку екосистему бібліотек загального призначення та кращі засоби для розробки аналітичних pipeline. Додатковою перевагою Python є його активна підтримка з боку open-source спільноти, що забезпечує регулярне оновлення бібліотек та швидке впровадження нових методів аналізу даних. Це особливо важливо для задач економічного моделювання, де часто виникає потреба в адаптації алгоритмів до нових типів даних або змін у структурі показників.

Порівняно з MATLAB, Python є відкритим програмним забезпеченням, що забезпечує актуальність і підтримку пакетів. Використання бібліотеки

wbdata дозволило автоматизувати збір статистичних даних безпосередньо з World Bank API, забезпечивши відтворюваність дослідження, що також підвищує актуальність інформації, що використовується в моделюванні.

Важливою складовою реалізації дослідження є використання спеціалізованих бібліотек Python. Наприклад, бібліотека pandas забезпечує зручну роботу з табличними даними та їх трансформацію, NumPy використовується для ефективних чисельних обчислень, matplotlib та seaborn — для візуалізації результатів аналізу. Для побудови моделей машинного навчання застосовується бібліотека scikit-learn, яка надає широкий набір алгоритмів класифікації, регресії та кластеризації. У свою чергу, бібліотека statsmodels використовується для реалізації класичних економетричних підходів, зокрема регресійного аналізу та моделей часових рядів, таких як ARIMA.

Отже, проведений аналіз підтвердив, що жодне з існуючих рішень не забезпечує повного функціонального покриття задач даного дослідження від автоматизованого збору і очищення міжнародних статистичних даних до одночасного застосування регресійних моделей, ансамблевого машинного навчання, ARIMA-прогнозування, кластеризації та сценарного аналізу в єдиному аналітичному контурі. Це обґрунтовує актуальність розробки власного рішення на основі Python.

1.4 Постановка задачі

Спираючись на проведений аналіз літератури, існуючих рішень та актуальність теми, можемо сформулювати актуальну постановку задачі дипломної роботи.

Метою роботи є розробка системи математичного моделювання та прогнозування показників економічного стану населення (зокрема, індексу

Джині) в умовах демографічних трансформацій на прикладі п'яти європейських країн — України, Польщі, Німеччини, Італії та Іспанії, за період 2000–2024 роки.

Для реалізації задачі необхідно виконати поставлені задачі.

1. Сформувати репрезентативну багатовимірну матрицю соціально-економічних та демографічних показників на основі даних World Bank Open Data та OECD Data шляхом автоматизованого збору через World Bank API.
2. Провести попередню обробку даних, де виявити та заповнити пропущені значення методами *ffill/bfill*, видалити спостереження без цільової змінної, здійснити стандартизацію ознак.
3. Виконати аналіз даних: розрахувати дескриптивні статистики, побудувати кореляційну матрицю Пірсона та візуалізувати динаміку ключових показників.
4. Побудувати та порівняти дві прогностні моделі — множинну лінійну регресію та Random Forest Regressor, за метриками MAE, RMSE, MAPE та R^2 .
5. Перевірити стаціонарність часового ряду індексу Джині для України (ADF-тест) та побудувати середньостроковий прогноз методом ARIMA.
6. Виконати кластеризацію країн алгоритмом K-Means за показниками нерівності, безробіття та демографічного старіння.
7. Розробити три сценарії розвитку (базовий, оптимістичний, песимістичний) та реалізувати їх у межах аналітичного блоку системи підтримки прийняття рішень (DSS).

Об'єктом дослідження є соціально-економічні та демографічні процеси в п'яти обраних європейських країнах в умовах сучасних демографічних трансформацій. Предметом дослідження є математичні моделі та методи прогнозування індексу Джині на основі демографічних і макроекономічних чинників.

Вхідними даними є 7 статистичних індикаторів (індекс Джині, рівень безробіття, частка населення 65+, коефіцієнт фертильності, інфляція, чиста міграція, ВВП на душу населення) для 5 країн за 25 років, що формує

матрицю розміром 125×9 . Вихідними результатами є порівняльна таблиця якості моделей, точковий прогноз індексу Джині для України на 2021–2025 роки, матриця сценарного аналізу та кластерна карта країн, збережені у форматі CSV для подальшого використання в DSS.

1.5 Висновки до розділу 1

У розділі було проведено комплексний розгляд з обґрунтуванням вибору методів та засобів реалізації дипломної роботи. Визначено актуальність дослідження, що зумовлена демографічними шоками в Україні та загальноєвропейськими тенденціями старіння населення і зниження народжуваності, які безпосередньо впливають на розподіл доходів і соціальну нерівність.

Аналіз літератури засвідчив наявність розвиненої теоретичної бази у вимірюванні нерівності (Сен, Гіні–Лоренц), економетричного моделювання (Вулдрідж, Гуджараті), машинного навчання (Брейман, Гасті) та прогнозування часових рядів (Бокс–Дженкінс). Водночас встановлено, що комплексне поєднання зазначених методів в єдиній аналітичній системі для задачі оцінювання демографічного впливу на нерівність у порівняльному контексті потребує подальшого розвитку.

Аналіз існуючих програмних рішень (EViews, STATA, Tableau) показав їх недостатність для вирішення поставлених задач: комерційні пакети обмежені у гнучкості та інтеграції з відкритими API, тоді як інструменти бізнес-аналітики не забезпечують необхідного математичного апарату. Обґрунтовано доцільність використання мови Python та відкритих бібліотек (scikit-learn, statsmodels, pandas, wbdata), що забезпечують повний аналітичний цикл — від автоматизованого збору даних до побудови прогнозних моделей і сценарного аналізу.

Визначена постановка задачі визначає конкретні цілі, об'єкт і предмет дослідження, склад вхідних даних та очікувані результати, що створює чіткий методологічний орієнтир для практичної реалізації, описаної у другому розділі.

РОЗДІЛ 2 МАТЕМАТИЧНА ПОСТАНОВКА ТА ВИКОРИСТАННЯ МОДЕЛЕЙ ПРОГНОЗУВАННЯ ТА ПІДГОТОВКИ ДАНИХ

2.1 Соціально-економічний стан населення та вплив демографічних трансформацій

Соціально-економічний стан населення базується на характеристиці рівнів доходів, зайнятості, безробіття, бідності, споживчих витрат та доступу до соціальних послуг [3]. Для дослідження таких складних систем необхідне чітке математичне визначення метрик, які будуть описувати стан населення та макроструктурні зрушення, аналіз яких відіграє важливу роль для формування ефективної державної соціальної політики та прогнозування економічного розвитку країни [4]. Як ключовий інтегральний індикатор розподілу доходів розглядається індекс Джині (Gini Index).

Індекс Джині базується на концепції кривої Лоренца, яка описує кумулятивний розподіл доходів між населення. Сен зазначає, що коефіцієнт Джині є зручним скалярним показником нерівності, але він маскує внутрішній розподіл, тому криву Лоренца слід розглядати у повноті [6]. Нехай x — кумулятивна частка населення (впорядкована за зростанням доходів), де $x \in [0, 1]$, а $L(x)$ — кумулятивна частка сукупного доходу, яку отримує цей відсоток населення. За умови повної рівності крива Лоренца стає лінійною функцією $L(x) = x$.

Коефіцієнт Джині (G) визначається як відношення площі між лінією абсолютної рівності та фактичною кривою Лоренца до площі під лінією абсолютної рівності [6]:

$$G = \frac{\int_0^1 (x - L(x)) dx}{\int_0^1 x dx} = 2 \int_0^1 (x - L(x)) dx = 1 - 2 \int_0^1 L(x) dx$$

Для дискретного набору даних, де населення розбите на n підгруп, коефіцієнт Джині розраховується за формулою:

$$G = \frac{\sum_{i=1}^n \sum_{j=1}^n |y_i - y_j|}{2n^2 \bar{y}}$$

де y_i — дохід i -ї особи (або підгрупи), а $\bar{y}$ — середнє значення доходу по всій вибірці.

Динаміка індексу Джині є результуючим вектором впливу багатьох факторів, серед яких значну роль відіграють саме демографічні трансформації:

- 1) старіння населення з часткою з віком 65+, що чинить великий тиск на пенсійні системи, змінюючи співвідношення між працездатними та непрацездатними громадянами;
- 2) коефіцієнт фертильності (Fertility Rate), який визначає довгостроковий потенціал відтворення людського капіталу;
- 3) чиста міграція (Net Migration), що зумовлює відтік та притік робочої сили, яка безпосередньо впливає на пропозицію на ринку праці та рівень оплати праці.

Сукупність демографічних та економічних показників формує простір ознак, що потребує використання сучасних методів математичного моделювання та прогнозування [5]. Аналогічні підходи до аналізу складних систем та моделювання соціально-економічних процесів розглядаються у сучасних дослідженнях НН ІПСА КПІ [1].

2.2 Методи математичного моделювання та прогнозування соціально-економічних процесів

Для дослідження причинно-наслідкових зав'язків та визначення впливу факторів на рівень соціально-економічних нерівностей застосовуються методи економетричного аналізу та машинного навчання [7].

2.2.1 Модель множинної лінійної регресії

Модель описує залежність між залежною цільовою змінною Y та вектором незалежних ознак (регресорів) $X = (X_1, X_2, \dots, X_k)$ [8]. Матрична форма моделі має вигляд:

$$Y = X\beta + \varepsilon$$

де Y — вектор-стовпець спостережень змінної розмірністю $(n \times 1)$; X — матриця плану розмірністю $(n \times (k + 1))$, перший стовпець якої складається з одиниць; β — вектор-стовпець невідомих параметрів (коефіцієнтів регресії) розмірністю $((k + 1) \times 1)$; ε — вектор випадкових похибок (залишків) розмірністю $(n \times 1)$.

Оцінка вектора параметрів $\hat{\beta}$ здійснюється методом найменших квадратів, шляхом мінімізації суми квадратів залишків:

$$S(\beta) = \varepsilon^T \varepsilon = (Y - X\beta)^T (Y - X\beta) \rightarrow \min$$

Диференціюючи за вектором β та прирівнюючи до нуля, ми отримуємо систему рівнянь, розв'язок яких має вигляд:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Згідно з теоремою Гаусса-Маркова, оцінки МНК є найкращими лінійними незміщеними оцінками (BLUE), якщо похибки мають нульове математичне сподівання, є гомоскедастичними та некорельованими [8].

2.2.2 Модель випадкового лісу (Random Forest Regressor)

Random Forest є ансамблевим методом машинного навчання, що базується на побудові великої кількості базових моделей — дерев рішень та усередненні їх прогнозів. Як доведено Брейманом, при збільшенні кількості дерев N узагальнена помилка лісу збігається до граничного значення, яке визначається силою окремих класифікаторів та їхньою кореляцією між собою [9]. Математична основа алгоритму будується на двох концепціях: бегінгу (Bagging) та методі випадкових підпросторів.

Кожне дерево рішень розщеплює простір ознак на гіперпрямокутники. Для регресійного дерева критерієм розщеплення у вузлі є мінімізація середньоквадратичної похибки (MSE):

$$Q = \frac{1}{|R_m|} \sum_{i \in R_m} (y_i - \hat{y}_{R_m})^2$$

де R_m — множина об'єктів у поточному вузлі m , а $\hat{y}_{R_m}$ — середнє значення цільової змінної для об'єктів у цьому вузлі.

Random Forest усереднює прогнози побудованих N дерев, що значно знижує дисперсію моделі без збільшення її зсуву:

$$\hat{y} = \frac{1}{N} \sum_{b=1}^N T_b(x)$$

де $T_b(x)$ — вихідні дані b -го дерева для вектора ознак x .

Однією з переваг Random Forest є можливість оцінювання важливості ознак, що надає можливість визначити найбільш впливові демографічні та економічні фактори [9].

2.2.3 Методи прогнозування часових рядів

Прогнозування динаміки індексу Джині реалізується в межах методології Бокса-Дженкінса [10] за допомогою лінійних стохастичних моделей часових рядів.

Часовий ряд Y_t називається слабо стаціонарним, якщо його математичне сподівання та дисперсія є константними у часі, а автоковаріаційна функція залежить лише від часового зсуву τ :

$$E[Y_t] = \mu, \quad D[Y_t] = \sigma^2, \quad \gamma(\tau) = \text{cov}(Y_t, Y_{t+\tau})$$

Для перевірки стаціонарності використовується розширений тест Дікі–Фуллера (ADF-test) [10]:

$$\Delta Y_t = \alpha + \beta t + \gamma Y_{t-1} + \sum_{i=1}^p \delta_i \Delta Y_{t-i} + \varepsilon_t$$

Якщо часовий ряд є інтегрованим порядку d , використовується клас загальних моделей ARIMA.

1. Авторегресійна модель (AR) порядку p , що описує залежність поточного значення ряду від його попередніх значень:

$$Y_t = c + \sum_{i=1}^p \varphi_i Y_{t-i} + \varepsilon_t$$

де p — порядок авторегресії, φ_i — параметри моделі, ε_t — білий шум.

2. Модель ковзного середнього (MA) описує залежність від попередніх випадкових шоків:

$$Y_t = \mu + \sum_{i=1}^q \theta_i \varepsilon_{t-i} + \varepsilon_t$$

де: q — порядок ковзного середнього, θ_i — параметри моделі.

3. Модель ARMA(p, q), поєднує авторегресійну частину та ковзне середнє:

$$Y_t = c + \sum_{i=1}^q \theta_i \varepsilon_{t-i} + \varepsilon_t + \sum_{i=1}^p \varphi_i Y_{t-i}$$

Вибір порядків p та q здійснюється на основі спільного аналізу ACF та PACF, а також інформаційних критеріїв якості моделі (AIC, BIC), де оптимальною вважається модель з мінімальними значеннями цих критеріїв.

4. Модель ARIMA(p, d, q), для нестационарних рядів і включає операцію диференціювання:

$$\Phi_p(L)(1-L)^d Y_t = \Theta_q(L) \varepsilon_t$$

де L — оператор лагу, $\Phi_p(L)$ — поліном авторегресії, $\Theta_q(L)$ — поліном ковзного середнього.

5. Модель SARIMA(p, d, q)(P, D, Q) s для рядів із сезонністю:

$$\Phi_p(L)\Phi_p(L^s)(1-L)^d(1-L^s)^D Y_t = \Theta_q(L)\Theta_q(L^s)\varepsilon_t$$

де s — довжина сезонного циклу, (P, D, Q) — сезонні порядки AR, диференціювання та MA, (p, d, q) — несезонні параметри.

Таким чином, методологія Бокса–Дженкінса дозволяє послідовно здійснювати ідентифікацію, оцінювання та перевірку адекватності моделей AR, MA, ARMA, ARIMA та SARIMA, що забезпечує побудову якісних прогнозів часових рядів економічних показників.

2.2.4 Кластеризація та сценарне моделювання

Для сегментації країн або соціальних груп за економічними та демографічними характеристиками використовується алгоритм К-середніх

(K-Means) [11]. Функція мінімізації всерединікластерної дисперсії має вигляд:

$$J = \sum_{k=1}^K \sum_{x_i \in S_k} ||x_i - \mu_k||^2 \rightarrow \min$$

де μ_k — центроїд кластера S_k :

$$\mu_k = \frac{1}{|S_k|} \sum_{x_i \in S_k} x_i$$

Для оцінювання можливих траєкторій розвитку соціально-економічної системи застосовується сценарне моделювання [12]. Основними сценаріями є:

- 1) базовий;
- 2) оптимістичний;
- 3) песимістичний.

2.3 Архітектура DSS та критерії оцінки моделей

Система підтримки прийняття рішень (Decision Support System, DSS) є багаторівневою інформаційно-аналітичною системою, яка використовується для збору, обробки, аналізу та прогнозування соціально-економічних показників. Турбан, Шарда та Делен визначають DSS як інтерактивну комп'ютерну систему, що допомагає особам, які приймають рішення, використовувати дані та моделі для вирішення неструктурованих і напівструктурованих проблем [13]. Архітектура DSS складається з таких рівнів (рис. 2.1):

- 1) Data Layer — інтеграція статистичних та демографічних даних;
- 2) Processing Layer — очищення, нормалізація та агрегація даних;
- 3) Modeling Layer — побудова моделей прогнозування;

4) Decision Layer — формування аналітичних рекомендацій.

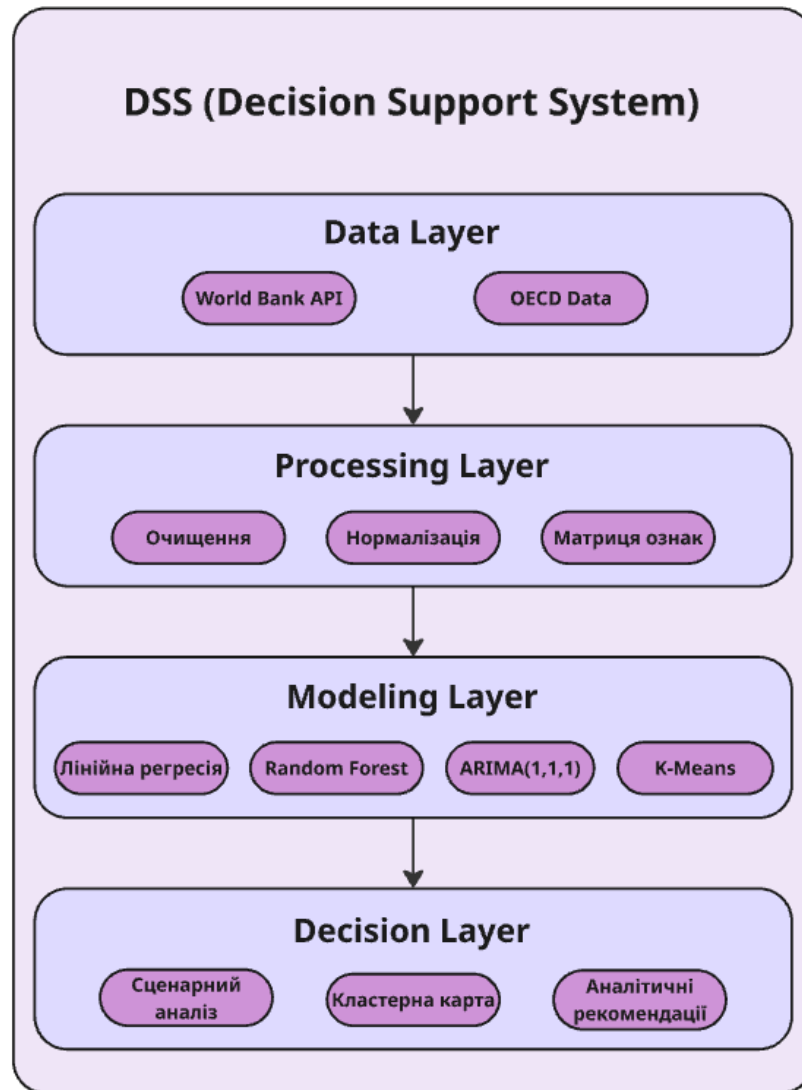


Рисунок 2.1 – Візуалізація архітектурних рівней DSS-системи

Для оцінювання якості побудованих моделей та точності прогнозування використовуються наступні статистичні метрики та критерії адекватності:

1) середня абсолютна похибка (MAE) оцінює середній модуль відхилення прогнозних значень від фактичних:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

2) середньоквадратична похибка (MSE) штрафує модель за великі аномальні помилки через піднесення різниць до квадрата:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

3) корінь із середньоквадратичної похибки (RMSE) повертає похибку до розмірності досліджуваної змінної:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

4) середня абсолютна відсоткова похибка (MAPE) визначає відносну точність моделі у відсотках, що є зручним для порівняння часових рядів із різними масштабами:

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100\%$$

5) коефіцієнт детермінації (R^2) показує частку дисперсії залежної змінної, яка пояснюється побудованою моделлю:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

2.4 Джерела даних та сучасні інформаційні технології аналізу

Інформаційною базою дослідження є міжнародні статистичні ресурси та відкриті API, зокрема World Bank Open Data та OECD Data [14]. Зазначені джерела містять офіційні статистичні показники соціально-економічного розвитку країн світу, що регулярно оновлюються та використовуються у наукових дослідженнях, аналітичних звітах і державному плануванні.

Використання мови програмування Python та бібліотек Data Science дозволяє реалізувати аналітичний процес використовуючи:

- 1) `wbdata` — для роботи безпосередньо з World Bank API [14], що забезпечує автоматизоване завантаження статистичних показників за обраними країнами та часовими періодами;
- 2) `pandas`, `numpy` — базові бібліотеки для зберігання, очищення, трансформації та статистичної обробки даних;
- 3) `scikit-learn` — бібліотека машинного навчання, яка використовується для побудови моделей прогнозування, оцінювання якості моделей та виконання процедур попередньої обробки даних;
- 4) `statsmodels` — пакет для економетричного аналізу, що містить інструменти перевірки статистичних гіпотез, аналізу часових рядів та для реалізації моделей AR, ARMA, ARIMA та SARIMA;
- 5) `matplotlib`, `seaborn` — бібліотеки для побудови графіків і візуалізації результатів аналізу, вони використовуються для дослідження динаміки показників, аналізу залежностей між змінними та представлення результатів прогнозування.

Використання зазначених інформаційних технологій забезпечує відтворюваність результатів дослідження, автоматизацію аналітичних процедур та можливість оперативного оновлення даних при появі нових статистичних спостережень. Це особливо важливо в умовах дослідження демографічних та економічних процесів, які характеризуються постійною зміною та потребують регулярного моніторингу.

2.5 Висновки до розділу 2

У другому розділі сформовано комплексний математичний та методологічний базис для системного аналізу соціально-економічних та демографічних процесів. Формалізовано концепцію індексу Джині через інтегральне обчислення кривої Лоренца та визначено характер впливу демографічних показників на існуючу соціальну нерівність. Описано математичний апарат багатofакторного моделювання, що включає метод найменших квадратів для лінійної регресії та мінімізацію MSE-критерію для Random Forest. Обґрунтовано застосування стохастичного апарату теорії часових рядів (ADF-тест, ARIMA) для задач короткострокового та середньострокового прогнозування соціальних індикаторів, визначено систему критеріїв оцінки (MAE, RMSE, MAPE, R^2) для вибору оптимальних рішень у межах розробленої архітектури аналізу даних.

РОЗДІЛ 3 АНАЛІЗ РЕЗУЛЬТАТІВ ДОСЛІДЖЕНЬ ЗА ТЕМОЮ ДИПЛОМНОЇ РОБОТИ

У третьому розділі проводиться практичне дослідження взаємозв'язку між демографічними трансформаціями та показниками економічного стану населення, як приклад розглядаються статистичні показники європейських країн. Реалізовано процес збору, очищення та попередньої обробки даних, проведено аналіз основних соціально-економічних показників, побудовано моделі прогнозування та виконано оцінювання їх ефективності. Для дослідження використано методи регресійного аналізу, ансамблевого машинного навчання, моделі часових рядів та кластеризацію.

3.1 Формування та попередня обробка набору даних

Практична реалізація досліджень починається з автоматизованого збору та консолідації інформації. Як джерела даних взято міжнародні статистичні ресурси World Bank Open Data та OECD Data [3; 4], статистичні індикатори використовувалися через World Bank API бібліотеку *wbdata*. Для аналізу було обрано репрезентативну вибірку з 5 європейських країн: Україна, Польща, Німеччина, Італія та Іспанія. Період дослідження охоплює 2000–2024 роки, що дозволяє відстежити довгострокові структурні зрушення. Виведення перших сформованих 20 рядків розглянуто на фото (рис. 3.1).

Розглянемо перші 20 рядків датасету:

	Country	Year	gini_index	unemployment_rate	population_65_plus	\
1	Germany	2024	NaN	3.400	23.195587	
2	Germany	2023	NaN	3.071	22.790044	
3	Germany	2022	33.7	3.137	22.458906	
4	Germany	2021	33.6	3.594	22.134537	
5	Germany	2020	32.5	3.881	21.844728	
6	Germany	2019	32.2	3.163	21.628393	
7	Germany	2018	31.9	3.384	21.449761	
8	Germany	2017	31.3	3.781	21.293641	
9	Germany	2016	31.3	4.104	21.134996	
10	Germany	2015	31.4	4.612	21.053354	
11	Germany	2014	30.8	4.979	20.958669	
12	Germany	2013	31.4	5.316	20.812976	
13	Germany	2012	30.9	5.372	20.637649	
14	Germany	2011	30.4	5.967	20.521498	
15	Germany	2010	30.2	7.043	20.565445	
16	Germany	2009	30.4	7.880	20.521596	
17	Germany	2008	30.6	7.508	20.287315	
18	Germany	2007	31.2	8.732	19.991194	
19	Germany	2006	31.0	10.277	19.495622	
20	Germany	2005	31.6	11.193	18.858685	

	fertility_rate	inflation	net_migration	gdp_per_capita
1	1.360	2.256498	36954.0	56103.732318
2	1.390	5.946437	609553.0	54776.766824
3	1.455	6.872574	981552.0	50506.517964
4	1.580	3.066667	310264.0	52349.245999
5	1.530	0.144878	203468.0	47394.873450
6	1.540	1.445660	319082.0	47656.199740
7	1.570	1.732169	403908.0	48916.168661
8	1.570	1.509495	452988.0	45553.934150
9	1.600	0.491747	517480.0	42948.938193
10	1.500	0.514426	1175283.0	41929.754911
11	1.470	0.906794	592338.0	48959.599120
12	1.420	1.504723	462384.0	47206.835351
13	1.410	2.008489	199143.0	44718.015921
14	1.390	2.075173	295282.0	47630.976188
15	1.390	1.103810	130116.0	42396.966598
16	1.360	0.312739	46533.0	42471.886863
17	1.380	2.628380	-23341.0	46379.164795
18	1.370	2.298344	-24772.0	42350.921721
19	1.330	1.577426	-489056.0	36980.334995
20	1.340	1.546911	81351.0	35084.436355

Рисунок 3.1 – Вивод даних

У межах системного аналізу сформовано багатовимірну матрицю ознак, що включає індекс Джині як цільову змінну та 6 регресорів: рівень безробіття, частку населення віком 65 років і старше, сумарний коефіцієнт фертильності, рівень інфляції, обсяг чистої міграції та ВВП на душу населення. Початковий розмір сформованого масиву даних становить 125 рядків на 9 стовпців спостережень.

Проблема збору даних з міжнародних джерел полягає в наявності великої кількості пропусків, що зумовлено періодичністю статистичних звітів країн. Первинний аналіз виявив таку кількість припущених значень у розрізі індикаторів:

- 1) *gini_index*: 14 пропущених значень;
- 2) *unemployment_rate*: 3 пропущені значення;
- 3) *population_65_plus*: 0 пропусків;
- 4) *fertility_rate*: 0 пропусків;
- 5) *inflation*: 0 пропусків;
- 6) *net_migration*: 0 пропусків;
- 7) *gdp_per_capita*: 0 пропусків.

Для відновлення цілісності часових рядів та демографічних факторів, було застосовано послідовний алгоритм заповнення: спершу метод прямого копіювання останнього відомого значення вперед у часі (*ffill*), потім — заповнення пропусків, що залишилися, копіюванням назад (*bfill*) у межах кожної країни відповідно. Оскільки індекс Джині є цільовою міткою моделі, то об'єкти, для яких він був повністю відсутній, було вилучено (*dropna*), щоб уникнути штучного зсуву при навчанні. Після очищення матриця набула повної статистичної повноти з 0 пропусків по всіх змінних.

3.2 Аналіз соціально-економічних та демографічних показників

Для первинного виявлення структурних особливостей досліджуваних країн було розраховано показники описової статистики на основі масиву даних, який до цього було очищено. Таблиця 3.1 демонструє ці статистики.

Таблиця 3.1 – Основні дескриптивні статистики набору даних

Показник (метрика)	Кількість (count)	Середнє (mean)	Ск. відхи- л. (std)	Мінім ум (min)	Медіа на (50%)	Максим ум (max)
gini_index	111	31.952	3.285	24	33	38
unemploy- ment_rate	122	9.81	4.944	2.743	9.099	26.094
population_ 65_plus	125	18.08	3.058	12.244	18.217	24.625
fertility_rat- e	125	1.326	0.122	0.897	1.33	1.6
inflation	125	4.45	6.38	-0.874	2.459	48.699
net_migrati- on	125	114,739	590.77	-5,699, 445	81,351	1,175,28 3
gdp_per_ca- pita	125	23,874	15.407	653.29	26,451	56,104

Аналіз розрахованих дескриптивних статистик дозволяє зробити важливі системні висновки щодо досліджуваної панелі країн.

1. Середнє значення індексу Джині становить 31.95%. Мінімальне зафіксоване значення дорівнює 24.0%, що характерно для періодів відносно низького розшарування, а максимальне сягає 38.0%, що свідчить про

наявність суттєвої соціально-економічної поляризації в окремих регіонах або часових проміжках.

2. Демографічне старіння демонструє великий демографічний зсув — у середньому частка літнього населення становить аж 18.08%. Максимальне значення сягає 24.63% (процеси глибокого старіння в Італії та Німеччині), тоді як мінімальне становить лише 12.24%.

3. Середній коефіцієнт фертильності у вибірці становить 1.33, що свідчить про наявність демографічної кризи народжуваності (3.1). Демографічний мінімум зафіксований на рівні 0.897 дитини на одну жінку (це історичні мінімуми народжуваності), а максимум становить лише 1.6.

4. По економічному розвитку та міграції спостерігається велика полярність між країнами за рівнем багатства: ВВП на душу населення коливається від мінімальних \$653.29 до значних \$56,104. Обсяг чистої міграції показує як катастрофічний відтік населення (мінімум -5.69 млн осіб, що безпосередньо відображає міграційні шоки в Україні), так і потужний приплив робочої сили до країн-реципієнтів (максимум +1.17 млн осіб у лідерах економічного розвитку, наприклад, Німеччині).

Аналіз графіків динаміки (рис. 3.2 — рис. 3.4) підтверджує, що процеси демографічного старіння носять безперервний характер, що вимагає регресійного оцінювання їхнього безпосереднього внеску у зміну макроекономічного стану населення.

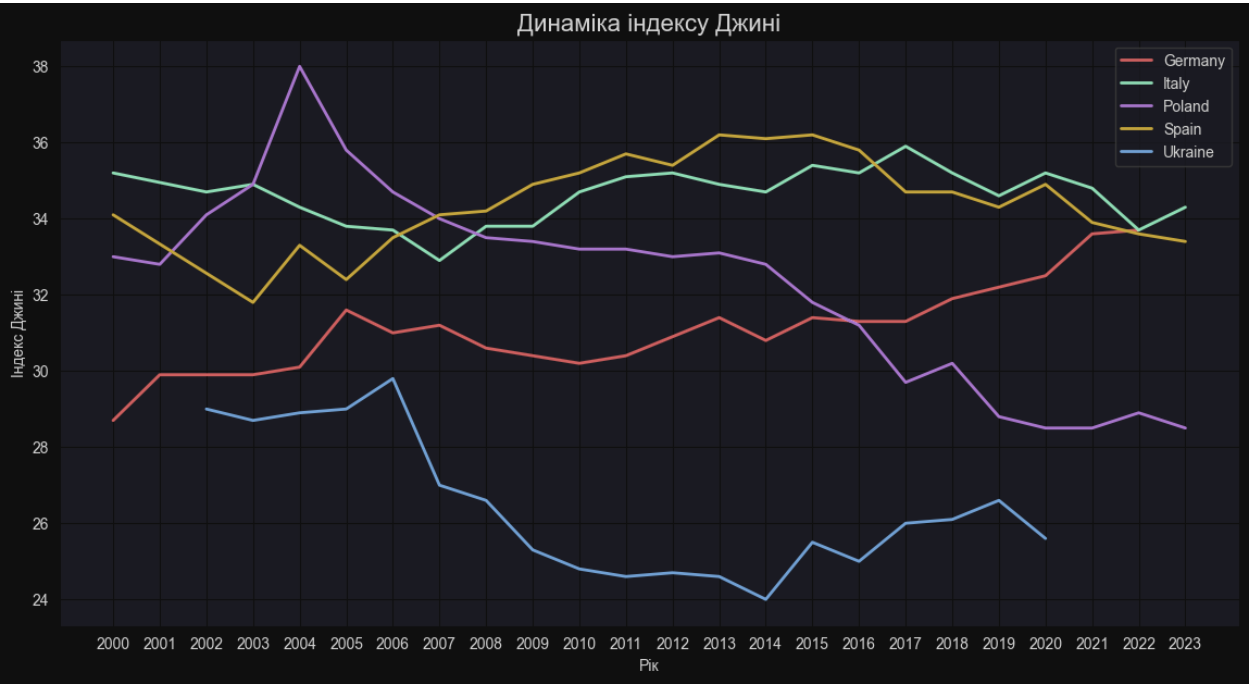


Рисунок 3.2 – Графік індексу Джині

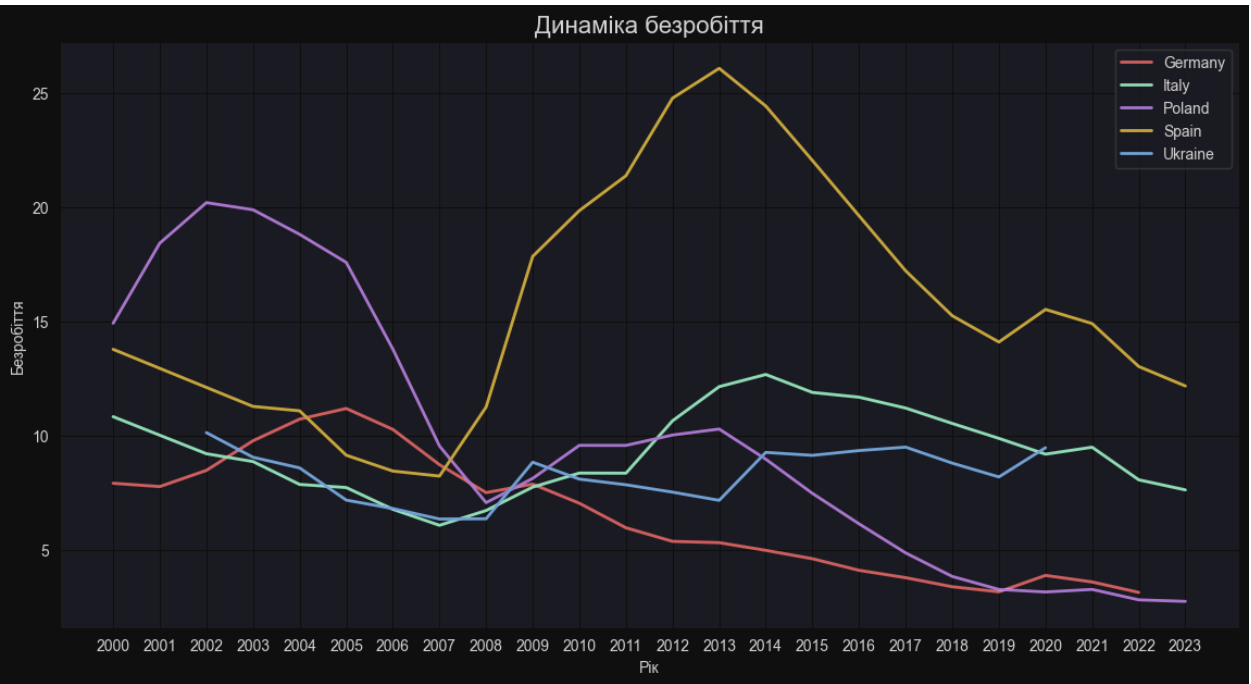


Рисунок 3.3 – Графік безробіття

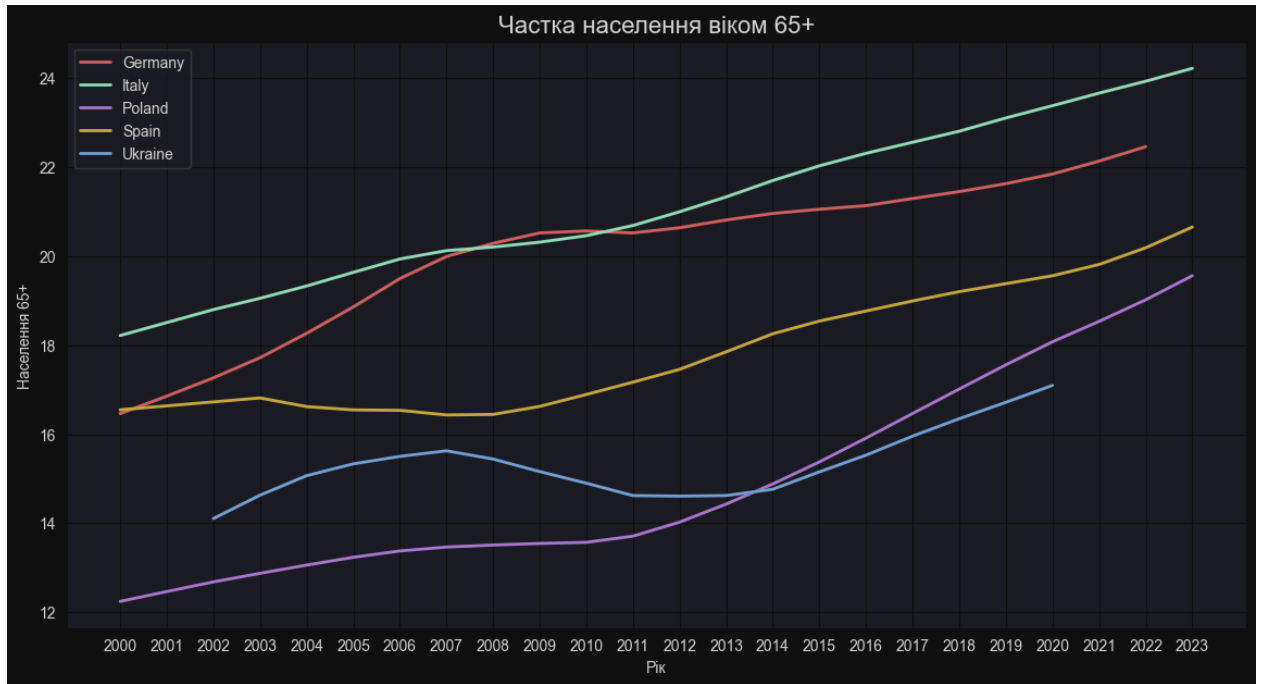


Рисунок 3.4 – Графік старіння населення

Для оцінювання щільності лінійної залежності між соціально-економічними та демографічними маркерами побудовано кореляційну матрицю Пірсона (рис. 3.5).

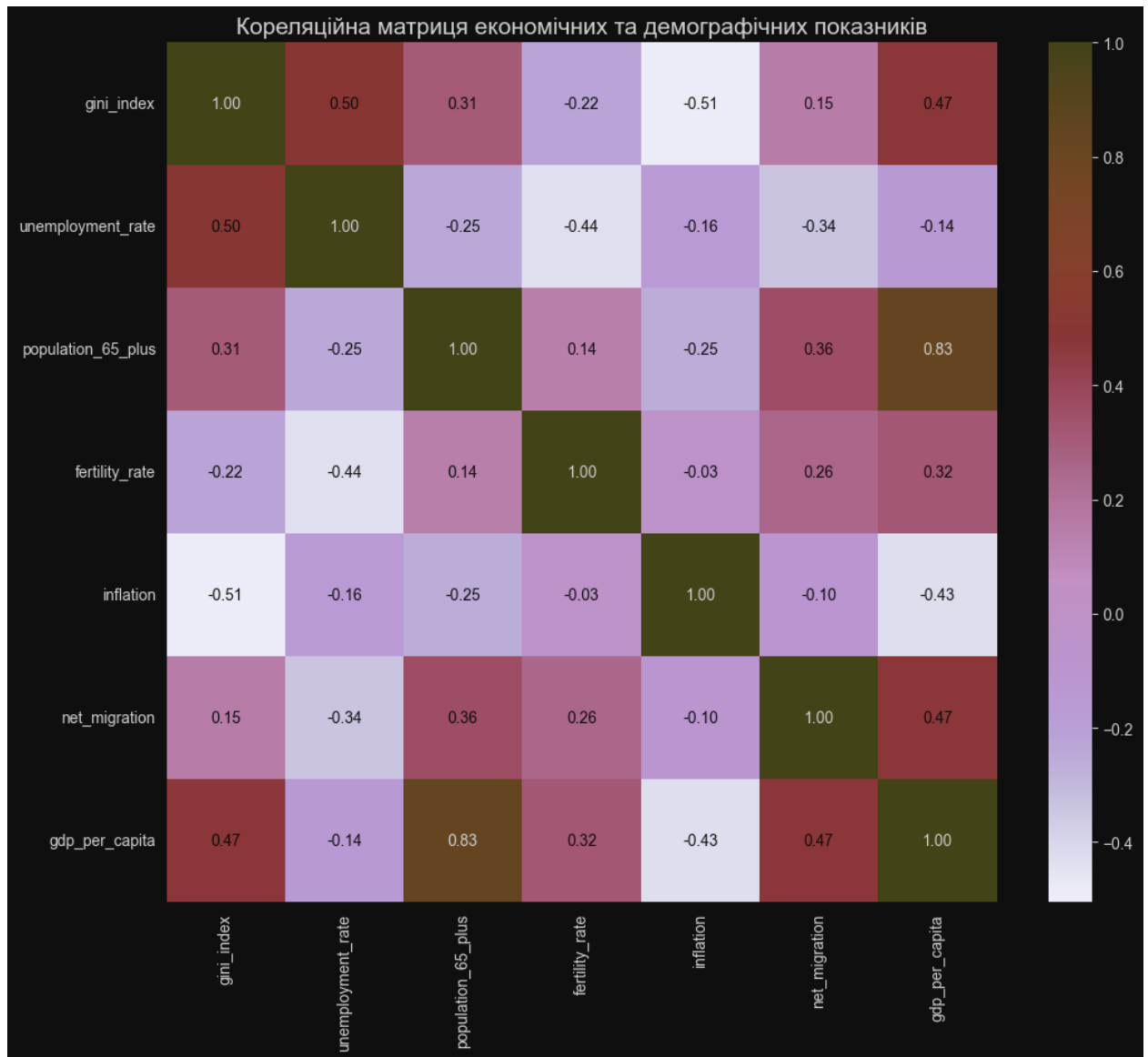


Рисунок 3.5 – Кореляційна матриця економічних та демографічних показників

Отже, як можемо помітити, індекс Джині має помірний позитивний зв'язок із рівнем безробіття ($r = 0.5$). Це математично підтверджує, що зростання безробіття є одним із ключових факторів, який безпосередньо призводить до збільшення соціального розшарування та бідності.

Між ВВП на душу населення (*gdp_per_capita*) та часткою літнього населення (*population_65_plus*), яка становить $r = 0.83$. Це відбиває загальноєвропейський тренд: багатші та розвиненіші країни (Німеччина,

Італія) мають значно вищу тривалість життя і, як наслідок, вищий ступінь демографічного старіння.

Рівень фертильності (*fertility_rate*) має помірний зворотний зв'язок із часткою населення 65+ ($r = 0,14$). При цьому суттєвіший зворотний зв'язок спостерігається між фертильністю та рівнем безробіття ($r = -0,44$): в країнах із вищим безробіттям народжуваність, як правило, нижча, що свідчить про вплив економічної нестабільності на репродуктивні рішення домогосподарств.

Чиста міграція (*net_migration*) виявляє помірну позитивну залежність із ВВП на душу населення ($r = 0.35$), що свідчить про те, що потоки мігрантів спрямовані переважно до економічно стабільніших та багатших країн-реципієнтів.

3.3 Побудова моделей прогнозування та оцінка їх ефективності

3.3.1 Регресійні моделі та Random Forest

Оскільки досліджувані макроекономічні та демографічні фактори мають різні одиниці виміру, перед моделюванням було виконано їх стандартизацію за допомогою *StandardScaler*. Масив даних розділено на навчальну та тестову вибірки класичному співвідношенні 80% на 20%.

Для ідентифікації ступеня впливу факторів на індекс Джині побудовано та порівняно дві моделі: параметричну (множинна лінійна регресія) та непараметричну ансамблеву (Random Forest Regressor). Обчислені на тестовій вибірці метрики якості зведено в порівняльну таблицю 3.2.

Таблиця 3.2 – Порівняльний аналіз якості регресійних моделей

Модель (Model)	MAE	RMSE	MAPE	R^2 (Коеф. детермінації)
Linear Regression	1.7367	2.0474	0.0549	0.4565
Random Forest Regressor	1.1022	1.4808	0.0356	0.7157

Аналіз економетричних метрик свідчить про помітну перевагу ансамблевої моделі машинного навчання. Модель Random Forest краще описує досліджуваний процес: її коефіцієнт детермінації становить $R^2 = 0.7157$, що означає, що вона пояснює 71.57% дисперсії індексу Джині на основі вхідних факторів. У той ж час лінійна регресія зафіксувала значно слабший результат ($R^2 = 0.4565$). Крім того, відносна похибка прогнозу MAPE є мінімальною і становить всього 3.56% (проти 5.49% у лінійної моделі).

Перевага Random Forest пояснюється здатністю ансамблевих моделей враховувати нелінійні залежності між демографічними та економічними показниками. На відміну від лінійної регресії, випадковий ліс не вимагає припущення про лінійний характер взаємозв'язків між змінними, що є особливо важливим для складних соціально-економічних систем.

Для детального аналізу структури обох моделей та характеру впливу регресорів було побудовано графіки вагових коефіцієнтів лінійної регресії (Рис. 3.6) та оцінки важливості ознак для випадкового лісу (Рис. 3.7).

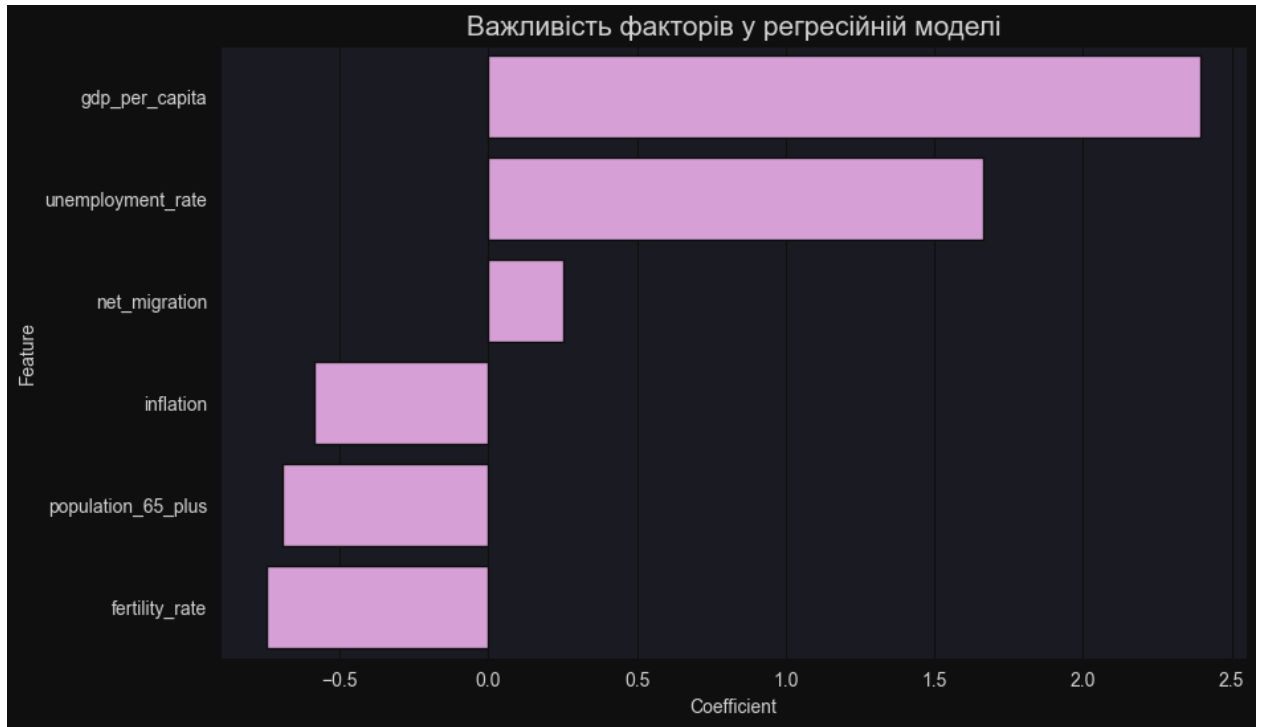


Рисунок 3.6 – Важливість факторів у лінійній регресійній моделі

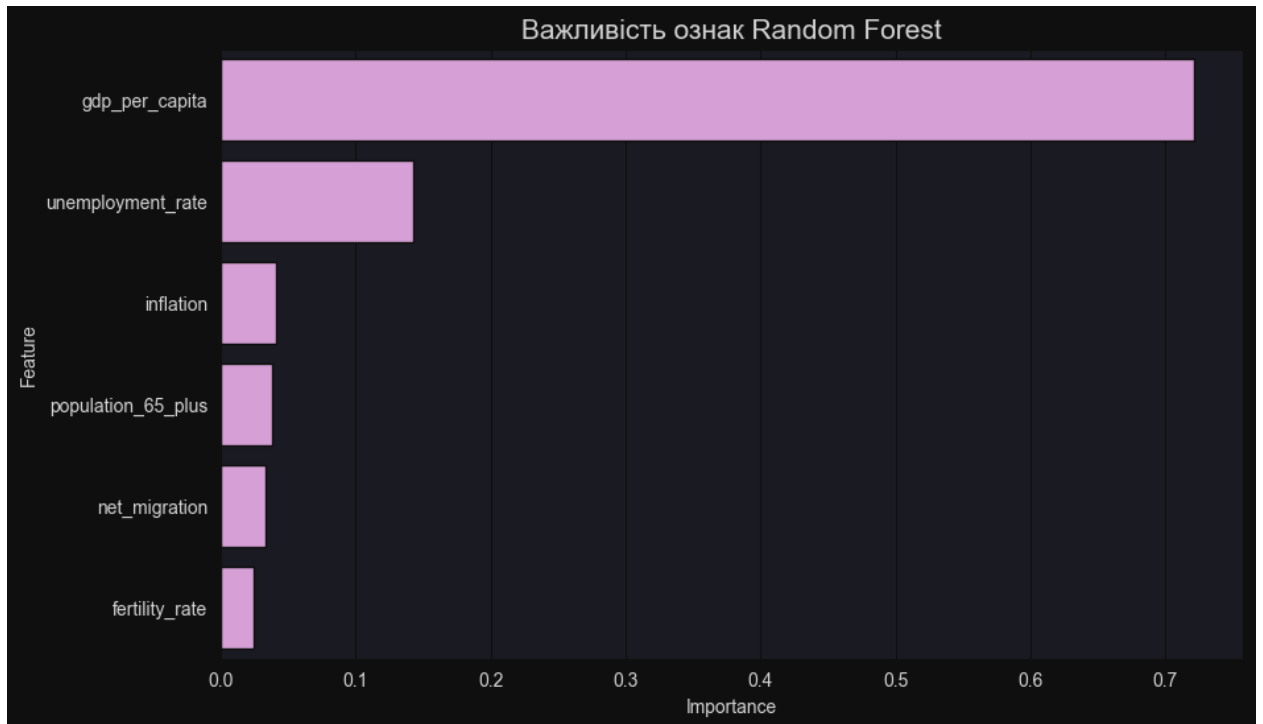


Рисунок 3.7 – Важливість ознак у моделі Random Forest Regressor

Аналіз результатів параметричного моделювання дозволяє визначити безпосередній напрямок впливу кожного фактора. Модель лінійної регресії

зафіксувала позитивний зв'язок індексу Джині з показниками ВВП на душу населення (*gdp_per_capita*, коефіцієнт 2.3949) та рівнем безробіття (*unemployment_rate*, коефіцієнт 1.6630). Це свідчить про те, що лінійний тренд інтерпретує зростання безробіття та збільшення масштабів ВВП як фактори, що поглиблюють соціальне розшарування. Також маємо сумарний коефіцієнт фертильності (*fertility_rate*, коефіцієнт -0.7442) та частка літнього населення з віком 65+ (*population_65_plus*, коефіцієнт -0.6914) мають негативні коефіцієнти, що вказує на їхній зворотній лінійний вплив у межах досліджуваної панелі країн.

Алгоритм Random Forest оцінює не напрямок, а чисту інформаційну вагу (внесок) кожного чинника під час розщеплення вузлів дерев рішення. Аналіз архітектури навченого «випадкового лісу», представлений на рис. 3.7, демонструє суттєву нерівномірність розподілу значущості.

1. ВВП на душу населення (*gdp_per_capita*) визначається як абсолютний домінуючий драйвер системи з показником значущості 0.7217. Модель ухвалила майже три чверті рішень на основі саме цього індикатора, що свідчить про ключову роль рівня економічного розвитку країни у формуванні соціально-економічної нерівності
2. Рівень безробіття (*unemployment_rate*) посідає друге місце за ступенем впливу з оцінкою значущості 0.1430. Це підтверджує значну роль кон'юнктури ринку праці у формуванні первинних доходів домогосподарств.
3. Усі інші ознаки мають сумарну вагу трохи більше 13%: рівень інфляції (*inflation*) — 0.0406, частка літнього населення (*population_65_plus*) — 0.0380, обсяг чистої міграції (*net_migration*) — 0.0328, а сумарний коефіцієнт фертильності (*fertility_rate*) має найменший прямий вплив — 0.0239.

Низька пряма значущість демографічних факторів у моделі Random Forest пояснюється їхньою опосередкованою дією: як показав кореляційний аналіз, процеси старіння нації мають критично високий зв'язок із багатством

країн ($r = 0.83$), тому їхній реальний деструктивний або конструктивний внесок акумулюється всередині головного макроекономічного чинника — ВВП.

3.3.2 Прогнозування часових рядів методом ARIMA

Для побудови середньострокового прогнозу індексу Джині безпосередньо для України проведено дослідження властивостей її одновимірному часового ряду. Спершу виконано розширений тест Дікі-Фуллера (ADF test [10]) на наявність одиничних коренів. Результати тесту виявилися наступними:

- 1) *ADF Statistic*: -4.9858
- 2) *p – value*: 0.0000

Оскільки обчислене значення $p - value = 0.0000$, що значно менше за критичний рівень значущості $\alpha = 0.05$, нульова гіпотеза про нестационарність ряду відхиляється. Згенерований вивід математичного апарату чітко демонструє, що ряд є стаціонарним. Оскільки модель ARIMA [10] допускає інтегрування ряду через операцію диференціювання, для подальшого прогнозування було використано модель ARIMA(1,1,1), яка враховує авторегресійну компоненту, першу різницю та ковзне середнє. На основі відомих даних за період 2000–2020 років для України було згенеровано точковий прогноз динаміки нерівності до 2025 року. Результати моделювання представлені на рис. 3.8.



Рисунок 3.8 – Графік прогнозування індексу Джині для України за допомогою моделі ARIMA

Аналіз побудованого часового ряду та прогнозової траєкторії (Рис. 3.8) вказує на те, що за умови збереження поточних макроекономічних тенденцій без урахування екстремальних шоків, індекс Джині в Україні демонструватиме тенденцію до поступової стабілізації. Модель ARIMA згладжує короткострокові коливання та фіксує вирівнювання показника, що є математичним наслідком властивості стаціонарності ряду. Проте точковий прогноз описує лише інерційний хід системи, що зумовлює необхідність побудови багатоваріантних сценарних меж у межах розробленого аналітичного блоку системи підтримки прийняття рішень (DSS).

3.4 Кластеризація та сценарне моделювання соціально-економічних процесів

З метою типізації країн та виявлення прихованих закономірностей на ринку праці, було застосовано алгоритм некерованого навчання $K - Means$. Було обрано три ознаки кластеризації: *gini_index*, *unemployment_rate* та *population_65_plus*. Результати геометричного розбиття простору на три кластери представлено на діаграмі розсіювання (рис. 3.9).

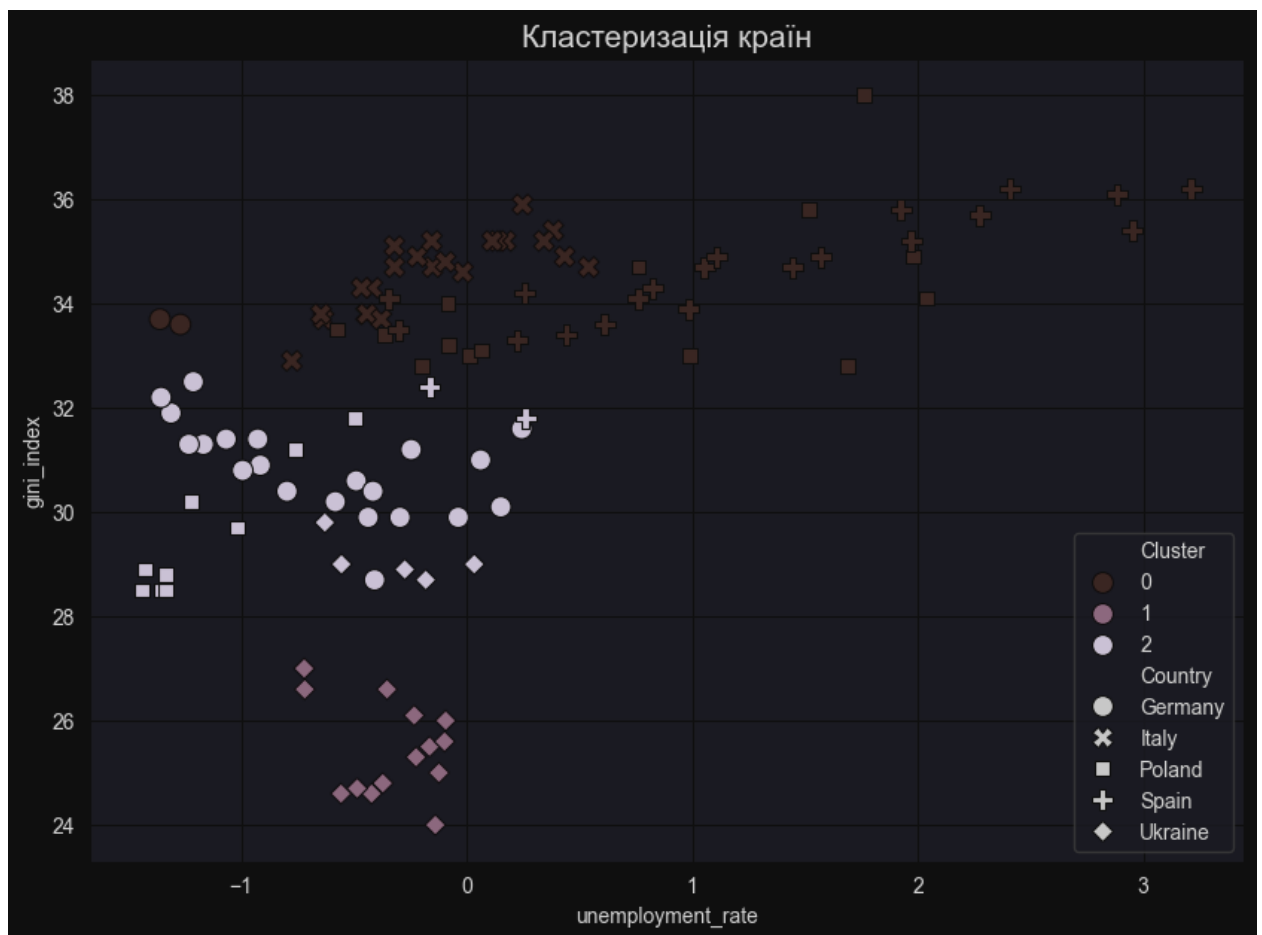


Рисунок 3.9 – Кластеризація країн

Аналіз сформованих кластерів дозволяє провести чітку макроструктурну сегментацію.

1. Кластер 0 об'єднує країни та періоди з високим рівнем соціально-економічного розшарування (індекс Джині зміщений вгору, 33–36%). Тут спостерігаються переважно спостереження Італії та Іспанії, де вищий рівень безробіття супроводжується значною поляризацією доходів суспільства.
2. Кластер 1 складає унікальний ізольований сегмент, утворений спостереженнями України. Ця група характеризується найнижчими значеннями індексу Джині у вибірці (24–27%) та помірним рівнем безробіття, що відображає специфіку вітчизняної економічної моделі минулих періодів (низькі офіційні доходи за високої вирівняності соціальних трансфертів).
3. Кластер 2 агрегує стабільні європейські економіки (зокрема Німеччину та Польщу), які утримують баланс між низьким або помірним безробіттям та контрольованим рівнем нерівності (Джині у межах 28–32%) на тлі інтенсивного демографічного старіння.

Для детального оцінювання ризиків та реакції соціально-економічної системи України на демографічні трансформації було реалізовано сценарне моделювання (Scenario Analysis). Використовуючи результати регресійного аналізу та прогнозні межі ARIMA, розраховано три альтернативні траєкторії розвитку до 2030 року (Рис. 3.10).



Рисунок 3.10 – Графік сценарного моделювання (Scenario Analysis) індексу Джині для України

Сценарне моделювання дає змогу оцінити чутливість системи до зовнішніх шоків та структурних змін на ринку праці:

- 1) базовий сценарій відображає продовження інерційного тренду, сам індекс Джині стабілізується у діапазоні 25.5%–26.2%, що передбачає відсутність різких демографічних зсувів, помірне відтворення робочої сили та збереження поточних обсягів соціальних трансфертів;
- 2) у песимістичному сценарії індекс Джині зазнає стрімкого зростання (приблизно на 15% від базового рівня) і досягне позначки 29.5%–30.1%, що означатиме перехід України до моделі з вищою соціальною полярністю;
- 3) оптимістичний сценарій передбачає успішну реалізацію стратегій демографічної стабілізації та активне реформування ринку праці, де індекс Джині планомірно знизиться приблизно на 10% — до рівня 23.1%–23.8%, що мінімізує бідність та наблизить показники України до стандартів найбільш соціально стабільних країн ЄС.

Згенеровані прогнозні траєкторії та аналітичні матриці сценарного простору були експортовані з програмного комплексу у файли

forecast_results.csv, model_comparison.csv та scenario_analysis.csv. Отримані файли та графічний матеріал є математичним та емпіричним базисом для формування керуючих впливів у блоці DSS, а також розробки прикладних рекомендацій щодо стабілізації економічного стану населення.

3.5 Висновки до розділу 3

У третьому розділі було реалізовано практичне дослідження соціально-економічних та демографічних процесів на основі статистичних даних європейських країн за 2000–2024 роки. Сформовано багатовимірний набір даних із використанням ресурсів World Bank Open Data та OECD Data, проведено очищення, нормалізацію та попередню обробку інформації.

У результаті аналізу описових статистик та кореляційної матриці встановлено наявність взаємозв'язків між демографічними трансформаціями та показниками економічної нерівності. Зокрема, виявлено позитивний зв'язок між рівнем безробіття та індексом Джині, а також суттєву залежність між старінням населення та рівнем економічного розвитку країн.

Під час моделювання побудовано та порівняно дві моделі прогнозування: множинну лінійну регресію та Random Forest Regressor. За результатами оцінювання встановлено, що ансамблева модель Random Forest забезпечує вищу точність прогнозування та краще описує нелінійні взаємозв'язки між факторами.

Також проведено прогнозування індексу Джині для України за допомогою моделі ARIMA та реалізовано сценарне моделювання для базового, оптимістичного та песимістичного сценаріїв розвитку. Додатково виконано кластеризацію країн методом K-Means, що дозволило виділити групи держав із подібними соціально-економічними характеристиками.

Отримані результати підтверджують доцільність використання методів машинного навчання, економетричного аналізу та систем підтримки прийняття рішень для аналізу впливу демографічних факторів на економічний стан населення.

РОЗДІЛ 4 ФУНКЦІОНАЛЬНО-ВАРТІСНИЙ АНАЛІЗ СИСТЕМИ МАТЕМАТИЧНОГО МОДЕЛЮВАННЯ ТА ПРОГНОЗУВАННЯ ПОКАЗНИКІВ ЕКОНОМІЧНОГО СТАНУ НАСЕЛЕННЯ

У цьому розділі проводиться техніко-економічне обґрунтування розробки програмного продукту — системи математичного моделювання та прогнозування індексу Джині в умовах демографічних трансформацій методом функціонально-вартісного аналізу (ФВА).

Функціонально-вартісний аналіз — методологія оцінки реальної вартості продукту незалежно від організаційної структури. Мета ФВА — знайти ефективніші технічні рішення та досягти балансу між споживчою цінністю і витратами на розробку. Алгоритм ФВА включає: визначення функцій ПП, обґрунтування параметрів і варіантів реалізації, розрахунок коефіцієнтів якості та вартості кожного варіанту, вибір оптимального.

4.1 Постановка задачі проєктування

Метою розробки аналітична система підтримки прийняття рішень (DSS), яка реалізує математичне моделювання залежності індексу Джині від демографічних і макроекономічних факторів для п'яти європейських країн (Україна, Польща, Німеччина, Італія, Іспанія) за 2000–2024 роки з урахуванням специфіки інерційного моделювання часових рядів. Система повинна забезпечувати виконання таких основних функцій:

- 1) функціонування на стандартних персональних комп'ютерах без спеціалізованого апаратного забезпечення;
- 2) автоматизований збір статистичних даних через World Bank API (бібліотека wbdata);

- 3) реалізація щонайменше двох прогнозних моделей з порівнянням метрик якості (MAE, RMSE, MAPE, R^2);
- 4) підтримка сценарного моделювання (базовий, оптимістичний, песимістичний сценарії);
- 5) мінімальні витрати на впровадження.

Об'єктом проєктування виступає процес аналізу та прогнозування економічного стану населення. Предметом проєктування є програмні засоби автоматизації аналізу статистичних даних та прогнозування соціально-економічних процесів.

4.2 Обґрунтування функцій програмного продукту

Головна функція F_0 — створення аналітичної DSS-системи прогнозування соціально-економічної нерівності. Виділено три основні підфункції з кількома варіантами реалізації кожної:

- 1) F_1 — вибір мови програмування (Python, R);
- 2) F_2 — вибір підходів, методів і бібліотек для виконання завдання (готові бібліотеки scikit-learn, statsmodels, wldata, власні алгоритми);
- 3) F_3 — вибір середовища розробки (Jupyter Notebook / Google Colab, PyCharm IDE, VS Code).

За розглянутими варіантами побудовано морфологічну карту системи на рисунку 4.1.

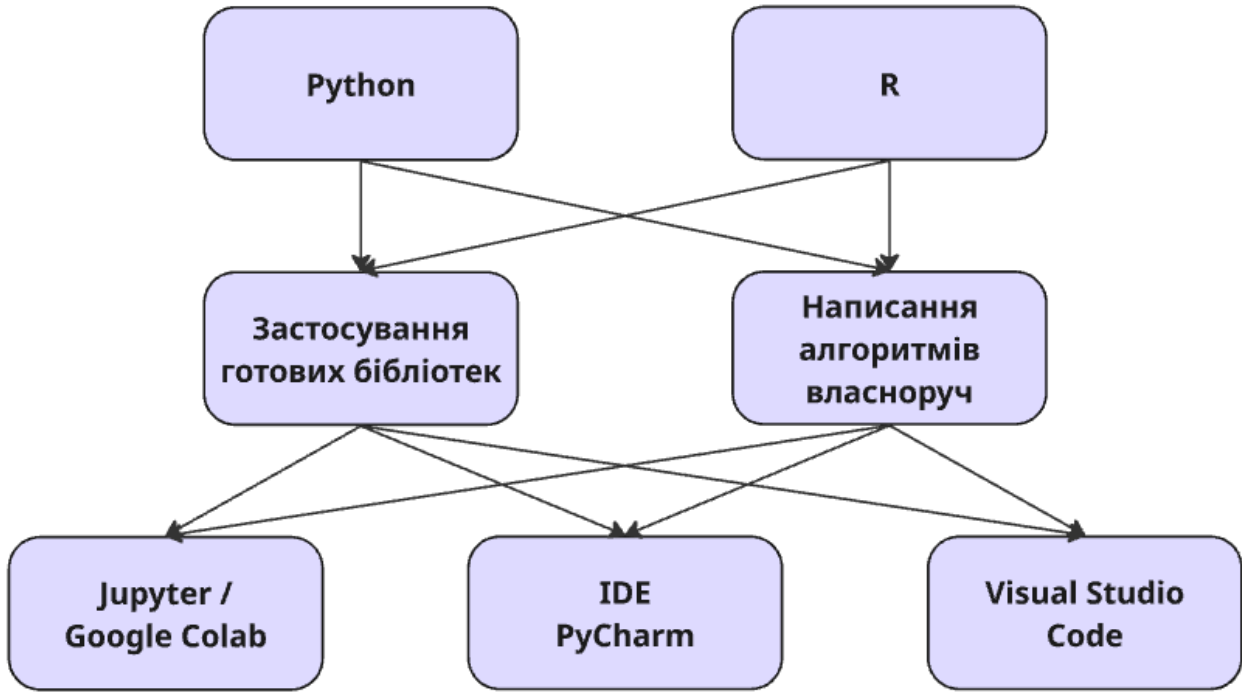


Рисунок 4.1 – Морфологічна карта

Морфологічна карта демонструє множину можливих комбінацій реалізації функцій програмного продукту. На основі морфологічної карти побудовано позитивно-негативну матрицю з усіма варіантами (табл. 4.1).

Таблиця 4.1 – Позитивно-негативна матриця варіантів реалізації функцій

Функція	Варіант	Переваги	Недоліки
F ₁	A (Python)	Широка екосистема бібліотек для Data Science (pandas, scikit-learn, statsmodels, wldata), читабельний синтаксис, інтеграція з World Bank API	Може бути повільнішим за компільовані мови при обробці надвеликих масивів

Продовження таблиці 4.1

F ₁	Б (R)	Сильні інструменти для статистичного аналізу та роботи з часовими рядами	Обмежені можливості для складних моделей ML, вузька підтримка World Bank API
F ₂	А (бібліотеки)	Оптимізовані та протестовані реалізації (Random Forest, ARIMA, K-Means), значно скорочує час розробки та знижує ризик помилок	Обмежена гнучкість, залежність від оновлень сторонніх пакетів
F ₂	Б (власні)	Повний контроль над логікою алгоритмів	Значно більший час розробки, виска ймовірність помилок, складне тестування
F ₃	А (Colab)	Інтерактивне середовище, зручне для ітеративного аналізу даних і візуалізації, безкоштовні хмарні обчислювальні ресурси	Не дуже зручний для великих проєктів зі складною структурою коду
F ₃	Б (PyCharm)	Повноцінне IDE з рефакторингом, відлагоджувачем, інтеграцією Git	Ресурсоємний, менш зручний для інтерактивної роботи з даними

Продовження таблиці 4.1

F_3	B (VS Code)	Легке та швидке середовище, підтримує Jupyter-розширення та Git	Вимагає налаштування середовища, менш потужне за PyCharm для великих проєктів
-------	-------------	---	---

Згідно з аналізом позитивно-негативної матриці виключаємо варіанти, що не відповідають вимогам задачі.

1. Функція F_1 , варіант А (Python) — єдиний придатний, оскільки забезпечує доступ до World Bank API та всіх необхідних бібліотек.
2. Функція F_2 , варіант А (готові бібліотеки) — обрано через необхідність реалізувати чотири різних методи (регресія, Random Forest, ARIMA, K-Means) в обмежені терміни.
3. Функція F_3 , варіант В (VS Code) виключається через меншу зрілість інструментів для науково-аналітичних проєктів.

Залишаються варіанти А та Б. Таким чином, два варіанти реалізації ПП:

- 1) $F_{1a} - F_{2a} - F_{3a}$ (Python + бібліотеки + Jupyter / Google Colab);
- 2) $F_{1a} - F_{2a} - F_{3b}$ (Python + бібліотеки + PyCharm IDE).

4.3 Обґрунтування системи параметрів програмного продукту

Взявши за основу аналіз технічних вимог і специфіки завдань моделювання визначено чотири ключові параметри для оцінки варіантів програмного продукту:

- 1) X_1 — обсяг програмного коду (рядків);
- 2) X_2 — час для навчання та верифікації всіх моделей (хвилини);

- 3) X_3 — загальний час розробки та налагодження програмного продукту (дні);
 4) X_4 — точність прогнозової моделі (R^2 , %).

Із даних практичного досвіду розробки аналітичних систем визначено значення параметрів. Для X_1 , X_2 , X_3 менше значення відповідає кращій якості; для X_4 — більше значення є кращим. Результати наведено в табл. 4.2.

Таблиця 4.2 – Основні параметри програмного продукту

Позн.	Назва параметра	Одиниці виміру	Гірше	Середнє	Краще	Вплив на якість
X_1	Обсяг програмного коду	рядків	1 200	900	600	Менше краще
X_2	Час навчання та верифікації моделей	хвилини	60	35	15	Менше краще
X_3	Загальний час розробки	дні	70	50	35	Менше краще
X_4	Точність прогнозової моделі (R^2)	%	45	65	75	Більше краще

За даними табл. 4.2 побудовано графічні залежності бальної оцінки кожного параметра від його абсолютного значення (рис. 4.2 — 4.5). У цих графіках краще значення параметра оцінюється в 10 балів, гірше — в 1 бал.

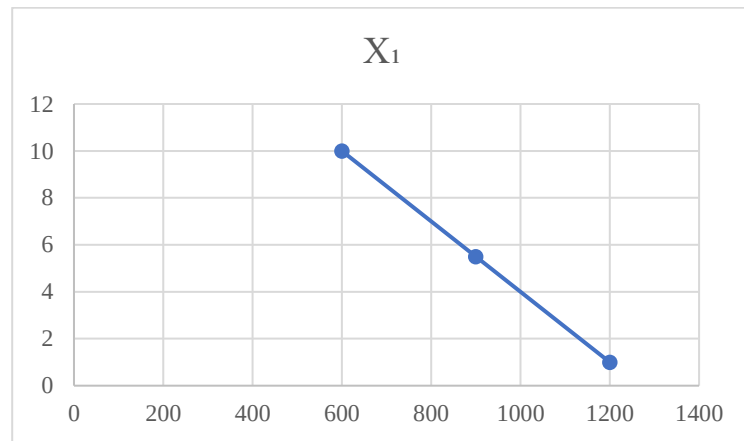


Рисунок 4.2 – X_1 , обсяг коду

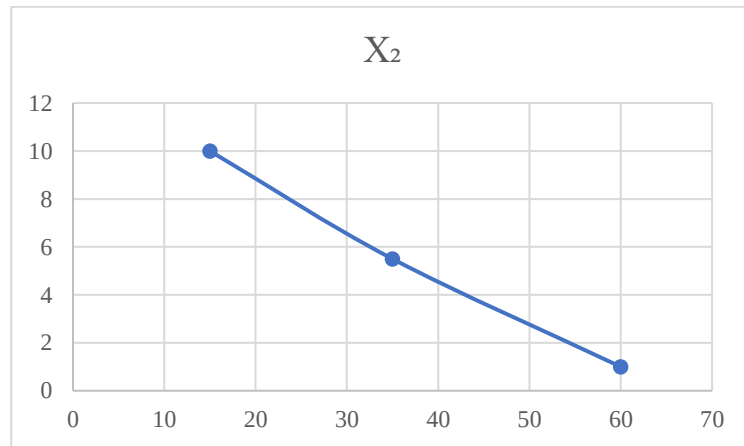


Рисунок 4.3 – X_2 , час для навчання та верифікації

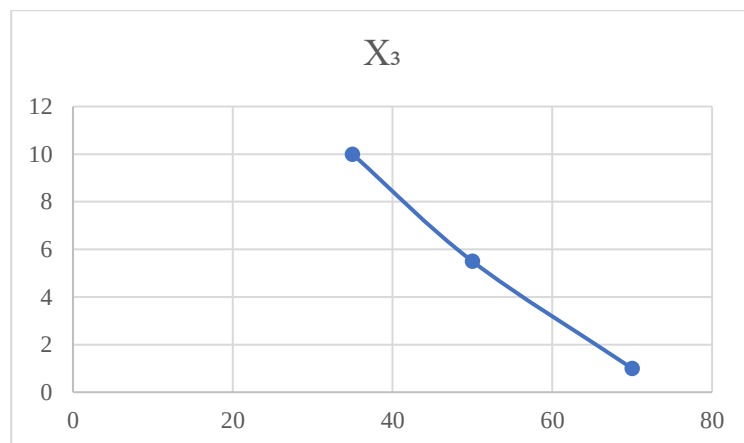


Рисунок 4.4 – X_3 , загальний час розробки та налагодження

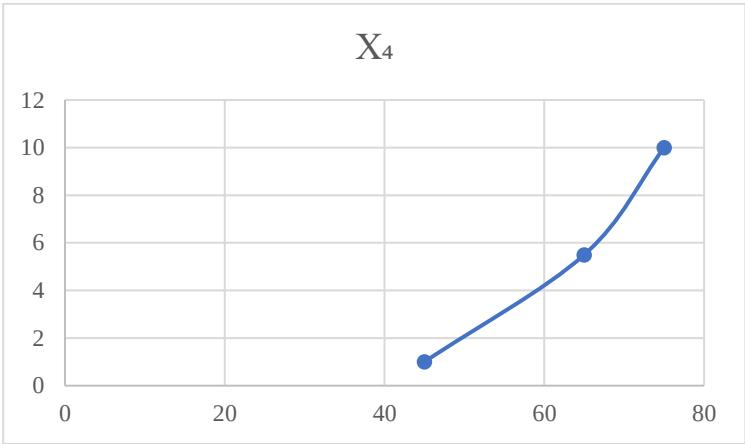


Рисунок 4.5 – X_4 , точність

4.4 Аналіз експертного оцінювання параметрів

Вагомість кожного параметра визначається методом попарного порівняння. До роботи залучено 7 експертів — фахівців у галузі системного аналізу, економетрики та розробки аналітичного програмного забезпечення. Після детального обговорення кожен експерт присвоїв параметрам ранги. Результати ранжування наведено в табл. 4.3.

Таблиця 4.3 – Результати ранжування параметрів експертами

Позн.	Назва парам.	1	2	3	4	5	6	7	R_i	Δ_i	Δ_i^2
X_1	Обсяг коду	2	2	1	2	2	1	2	12	– 5,5	30,25
X_2	Час навчання	3	4	3	4	3	4	3	24	6,5	42,25
X_3	Час розробки	1	1	2	1	1	2	1	9	– 8,5	72,25

Продовження таблиці 4.3

X ₄	Точність моделі	4	3	4	3	4	3	4	25	7,5	56,25
Разом		10	10	10	10	10	10	10	70	0	201

Для того, щоб оцінити достовірність експертних рейтингів, визначаємо такі параметри:

1) сума рангів кожного з параметрів і загальна сума рангів:

$$R_i = \sum_{j=1}^N r_{ij} R_{ij} = \frac{Nn(n+1)}{2} = 70$$

де N — кількість експертів, n — число параметрів;

2) середня сума рангів:

$$T = \frac{1}{n} R_{ij} = \frac{70}{4} = 17,5$$

3) відхилення суми рангів кожного параметра від середньої суми рангів:

$$\Delta_i = R_i - T = 70 - 17,5 = 52,5$$

Сума, що визначає відхилення кожного параметру повинна дорівнювати 0;

4) загальна сума квадратів відхилення:

$$S = \sum_{j=1}^N \Delta_i^2 = 30,25 + 42,25 + 72,25 + 56,25 = 201$$

5) розрахуємо коефіцієнт конкордації:

$$W = \frac{12S}{N^2(n^3 - n)} = \frac{12 * 201}{49 * (64 - 4)} = 0,821 > W_k = 0,67$$

Оскільки розрахований коефіцієнт конкордації перевищує нормативний показник, то експертні оцінки вважаються суттєво узгодженими та придатними для подальшого використання. Результати попарного порівняння параметрів наведено в таблиці 4.4.

Таблиця 4.4 – Попарне порівняння параметрів

Параметри	1	2	3	4	5	6	7	Кінцева оцінка	Числове значення X_{ij}
X_1 і X_2	<	<	>	<	<	<	<	$X_1 < X_2$	0,5
X_1 і X_3	>	>	<	>	>	<	>	$X_1 > X_3$	1,5
X_1 і X_4	<	<	<	<	<	<	<	$X_1 < X_4$	0,5
X_2 і X_3	>	>	>	>	>	>	>	$X_2 > X_3$	1,5
X_2 і X_4	<	<	<	<	<	<	<	$X_2 < X_4$	0,5
X_3 і X_4	<	<	<	<	<	<	<	$X_3 < X_4$	0,5

Ступінь переваги i -го параметра над j -тим визначається числовим значенням a_{ij} по формулі:

$$a_{ij} = \begin{cases} 1,5 & \text{при } X_i > X_j \\ 1,0 & \text{при } X_i = X_j \\ 0,5 & \text{при } X_i < X_j \end{cases}$$

З отриманих числових оцінок переваги складемо матрицю $A = \|a_{ij}\|$. Для кожного параметра зробимо розрахунок вагомості K_{bi} за наступними формулами:

$$K_{bi} = \frac{b_i}{\sum_{i=1}^n b_i}$$

$$b_i = \sum_{j=1}^N a_{ij}$$

Відносні оцінки перераховуються по декілька разів, поки наступні значення не будуть незначно відрізнятися від попередніх (менше 2%). На наступних етапах розрахунку відносні оцінки визначаються за вказаними формулами:

$$K_{bi} = \frac{b'_i}{\sum_{i=1}^n b'_i}$$

$$b_i = \sum_{j=1}^N a_{ij} b_j$$

Розрахунок коефіцієнтів вагомості параметрів проводиться ітераційним шляхом, зображено в таблиці 4.5.

Таблиця 4.5 – Розрахунок вагомості параметрів

X_i	X_1	X_2	X_3	X_4	b_i (1 ітер.)	K_{bi} (1 ітер.)	b_i' (2 ітер.)	K_{bi}' (2 ітер.)
X_1	1	0,5	1,5	0,5	3,5	0,219	12,25	0,208
X_2	1,5	1	1,5	0,5	4,5	0,281	16,25	0,275
X_3	0,5	0,5	1	0,5	2,5	0,156	9,25	0,157
X_4	1,5	1,5	1,5	1	5,5	0,344	21,25	0,36
Всього					16	1	59	1

Оскільки відхилення між значеннями K_{bi} на першій та другій ітераціях не перевищує 2%, розрахунок вважається завершеним. Підсумкові вагомості параметрів: $K(X_1) = 0,208$, $K(X_2) = 0,275$, $K(X_3) = 0,157$, $K(X_4) = 0,360$. Найбільш пріоритетною є точність прогнозової моделі (X_4).

4.5 Аналіз рівня якості варіантів реалізації функцій

Узагальнений показник якості для кожного варіанта обчислюється за формулою:

$$K_k(j) = \sum_{j=1}^N K_{bi,j} B_{i,j}$$

де n — кількість параметрів, K_{bi} - коефіцієнт вагомості i -го параметра, B_i - бальна оцінка i -го параметра j -го варіанту.

Бальні оцінки визначено згідно з графіками на рисунках 4.2 — 4.5. Розрахунок показників технічного рівня наведено в таблиці 4.6.

Таблиця 4.6 – Розрахунок показників рівня якості варіантів розробки

Варіант	Параметр	Абсолютне значення	Бальна оцінка B_{ij}	K_{bi}	$K_{bi} \cdot B_{ij}$
1	X_1	750 рядків	7,5	0,208	1,56
	X_4	71,6 %	8,2	0,36	2,95
	X_2	18 хв	9,1	0,275	2,5
	X_3	40 днів	7,3	0,157	1,15
2	X_1	850 рядків	5,8	0,208	1,21
	X_4	71,6 %	8,2	0,36	2,95
	X_2	20 хв	8,2	0,275	2,26
	X_3	44 дні	6	0,157	0,94

За результатами таблиці 4.6 визначаємо рівень якості кожного з варіантів за формулою:

$$K_k = K_{Ty}[F_{1k}] + K_{Ty}[F_{2k}] + \dots + K_{Ty}[F_{zk}]$$

Визначимо рівень якості кожного варіанту:

$$1) K_{K1} = 1,56 + 2,95 + 2,5 + 1,15 = 8,16$$

$$2) K_{K2} = 0,208 + 0,36 + 0,275 + 0,157 = 7,36$$

4.6 Економічний аналіз варіантів розробки програмного продукту

Для розрахунку собівартості розробки спершу визначимо трудомісткість. Програмний комплекс складається з двох основних завдань.

Завдання 1 — аналітичний модуль автоматизованого збору та інтелектуального моделювання даних (ступінь новизни — Б, група складності алгоритму — 1, вид інформації — нормативно-довідкова (НДІ)). Базова трудомісткість $T_{p1} = 64\$$ людино-днів. Поправковий коефіцієнт $K_{\Pi} = 1,021$. Складність контролю вхідної та вихідної інформації — стандартна (5,2 / 5,4), тому $K_{СК} = 1$. Мова програмування Python — рівень високий, $K_{\text{м}} = 1$. Використовуються стандартні бібліотеки: $K_{\text{ст}} = 0,6$. $K_{\text{ст.м}} = 1$: Складність контролю вхідної та вихідної інформації — стандартна (5,2 / 5,4), тому $K_{СК} = 1$. Мова програмування Python — рівень високий, $K_{\text{м}} = 1$. Використовуються стандартні бібліотеки: $K_{\text{ст}} = 0,6$. $K_{\text{ст.м}} = 1$:

Завдання 2 — розробка DSS-блоку та інтерактивного сценарного модуля (ступінь новизни — Б, група складності — 2, вид інформації — банк даних (БД)). Базова трудомісткість $T_{p2} = 27\$$ люд.-дн. Поправковий коефіцієнт $K_{\Pi} = 0,90$. $K_{СК} = 1$, $K_{\text{м}} = 1$, $K_{\text{ст}} = 0,7$.

Загальна трудомісткість обчислюється за формулою:

$$T_0 = T_p \cdot K_{\Pi} \cdot K_{СК} \cdot K_{\text{м}} \cdot K_{\text{ст}} \cdot K_{\text{ст.м}}$$

де T_p — трудомісткість розробки, K_{Π} — поправочний коефіцієнт, $K_{СК}$ — коефіцієнт на складність вхідної інформації, $K_{\text{м}}$ — коефіцієнт рівня мови програмування, $K_{\text{ст}}$ — коефіцієнт використання стандартних модулів і прикладних програм, $K_{\text{ст.м}}$ — коефіцієнт стандартного математичного забезпечення.

Обчислимо трудомісткість для кожного варіанту:

- 1) $T_1 = 64 \cdot 1,021 \cdot 1 \cdot 1 \cdot 0,6 \cdot 1 = 39,2$ людино-днів;
- 2) $T_2 = 27 \cdot 0,9 \cdot 1 \cdot 1 \cdot 0,7 \cdot 1,0 = 17$ людино-днів.

Загальна нормативна трудомісткість розробки з урахуванням коефіцієнтів якості варіантів становить:

- 1) $T_I = (39,2 + 17 + 8,16 + 17) \cdot 8 = 648,5$ людино-годин;
- 2) $T_{II} = (39,2 + 17 + 7,36 + 17) \cdot 8 = 642,1$ людино-годин.

У розробці проекту беруть участь один програміст-аналітик (посадовий оклад 28 000 грн) та науковий керівник із коефіцієнтом зайнятості 0,3 (оклад 32 000 грн). Середня погодинна ставка розробників вираховується за формулою:

$$C_{\text{ч}} = \frac{M}{T_m \cdot t} = \frac{28000 + 32000 \cdot 0,3}{2 \cdot 21,1 \cdot 8} = 108,4 \text{ грн/год}$$

де M — місячний оклад працівників, T_m — кількість робочих днів на тиждень, t — кількість робочих годин в день.

Заробітна плата обчислюється за формулою:

$$C_{\text{зп}} = C_{\text{ч}} \cdot T_i \cdot K_{\text{д}}$$

де $C_{\text{ч}}$ — величина погодинної оплати, T_i трудомісткість відповідного завдання, $K_{\text{д}}$ — норматив, який враховує додаткову заробітну плату.

Витрати на заробітну плату розробників з урахуванням додаткової заробітної плати ($K_{\text{д}} = 1,3$) становлять:

- 1) Варіант 1: $C_{\text{зп}} = 108,4 \cdot 648,5 \cdot 1,3 = 91408$ грн;
- 2) Варіант 2: $C_{\text{зп}} = 108,4 \cdot 642,1 \cdot 1,3 = 90506$ грн.

Відрахування на єдиний соціальний внесок ($\text{ЄСВ} = 22\%$) становлять:

- 1) Варіант 1: $C_{\text{від}} = 91408 \cdot 0,22 = 20110$ грн;
- 2) Варіант 2: $C_{\text{від}} = 90506 \cdot 0,22 = 19911$ грн.

Визначимо витрати на експлуатацію однієї ЕОМ вартістю 42 000 грн за норми амортизації $H_{\text{а}} = 25\%$ та витрат на ремонт $K_{\text{р}} = 3\%$. Ефективний фонд часу роботи ПК за рік становить $T_{\text{еф}} = 1408$ год. Витрати на електроенергію обчислені при потужності 0,25 кВт та тарифі 9,43 грн/кВт·год. Річні експлуатаційні витрати ПК становлять $C_{\text{ЕКС}} = 158922$ грн. Собівартість однієї машино-години роботи дорівнює:

$$C_{M-Г} = \frac{C_{EКС}}{T_{еф}} = \frac{158922}{1408} = 112,9 \text{ грн/год}$$

Витрати на оплату машинного часу роботи ПК за варіантами становлять:

- 1) Варіант 1: $C_M = 112,9 \cdot 648,5 = 73234$ грн;
- 2) Варіант 2: $C_M = 112,9 \cdot 642,1 = 72513$ грн.

Накладні витрати становлять 67% від заробітної плати:

- 1) Варіант 1: $C_H = 91408 \cdot 0,67 = 61243$ грн;
- 2) Варіант 2: $C_H = 90506 \cdot 0,67 = 60639$ грн.

Загальна вартість розробки програмного продукту обчислюється як сума всіх статей витрат:

$$C_{ПП} = C_{ЗП} + C_{ВІД} + C_M + C_H$$

Розглянемо для обох варіантів:

- 1) Варіант 1: $C_{ПП} = 91408 + 20110 + 73234 + 61243 = 245995$ грн;
- 2) Варіант 2: $C_{ПП} = 90506 + 19911 + 72513 + 60639 = 163569$ грн.

Підсумкові результати калькуляції витрат наведено в таблиці 4.7.

Таблиця 4.7 – Порівняння вартості розробки програмного продукту

Стаття витрат	Варіант 1 (Colab), грн	Варіант 2 (PyCharm), грн
Трудовісткість, люд.-год	648,5	642,1
Заробітна плата (основна + додаткова)	91 408	90 506
Відрахування на ЄСВ (22 %)	20 110	19 911
Витрати на машинний час	73 234	72 513
Накладні витрати (67 % від СЗП)	61 243	60 639

Продовження таблиці 4.7

Вартість розробки ПП (СПП)	245 995	243 569
-------------------------------	---------	---------

4.7 Вибір кращого варіанту за техніко-економічним рівнем

Для остаточного вибору оптимального варіанта реалізації розраховується коефіцієнт техніко-економічного рівня ($K_{\text{ТЕР}}$):

$$K_{\text{ТЕР}j} = \frac{K_{\text{К}j}}{C_{\text{Ф}j}}$$

де $K_{\text{К}j}$ — коефіцієнт рівня якості j -го варіанту, $C_{\text{Ф}j}$ — вартість розробки j -го варіанту, грн.

Розрахуємо для кожного варіанту:

- 1) Варіант 1: $K_{\text{ТЕР}1} = 8,16 / 245995 = 3,32 \cdot 10^{-5}$;
- 2) Варіант 2: $K_{\text{ТЕР}2} = 7,36 / 243569 = 3,02 \cdot 10^{-5}$.

Після порівняння отриманих значень бачимо, що $K_{\text{ТЕР}1} > K_{\text{ТЕР}2}$. Отже, оптимальним визначено Варіант 1 — реалізація аналітичної DSS-системи на мові Python з використанням хмарного інтерактивного середовища Jupyter Notebook / Google Colab. Цей варіант демонструє значно вищу якість та ефективність розробки аналітичного коду. Підсумкові техніко-економічні показники розробленого програмного продукту зібрано в таблицю 4.8.

Таблиця 4.8 – Основні техніко-економічні показники проекту

Показник	Одиниці виміру	Варіант 1	Варіант 2
Трудомісткість розробки	люд.-год	648,5	642,1
Обсяг програмного коду (X_i)	рядків	750	850

Продовження таблиці 4.8

Час навчання моделей (X ₂)	хвилини	18	20
Загальний час розробки (X ₃)	дні	40	44
Точність прогнозованої моделі R ² (X ₄)	%	71,6	71,6
Коефіцієнт технічного рівня КК	—	8,16	7,36
Вартість розробки ПП	грн	245 995	243 569
Коефіцієнт техніко-економічного рівня КТЕР	—	$3,32 \cdot 10^{-5}$	$3,02 \cdot 10^{-5}$

4.8 Висновки до розділу 4

У четвертому розділі було проведено техніко-економічне обґрунтування та аналіз доцільності розробки програмного комплексу за допомогою методології функціонально-вартісного аналізу (ФВА).

На основі аналізу функцій та побудованої морфологічної карти було сформовано та допущено до детального розгляду два альтернативні варіанти архітектури системи підтримки прийняття рішень. За допомогою методу попарного порівняння та експертного оцінювання (при коефіцієнті конкордації $W = 0,821$) встановлено коефіцієнти вагомості критеріїв якості, де найбільш пріоритетним визначено точність прогнозування індексу Джині ($K_B = 0,360$).

Розрахований інтегральний коефіцієнт технічного рівня Варіанта 1 становить $K_K = 8,16$, тоді як Варіанта 2 — $K_K = 7,36$. Проведений детальний фінансово-економічний аналіз та калькуляція річних витрат на експлуатацію комп'ютерної техніки встановили, що повна вартість розробки Варіанта 1 становить 245 995 грн, а Варіанта 2 — 243 569 грн.

На основі розрахованих коефіцієнтів техніко-економічного рівня ($K_{\text{ТЕР}1} = 3,32 \cdot 10^{-5} > K_{\text{ТЕР}2} = 3,02 \cdot 10^{-5}$) кращим та оптимальним варіантом визначено Варіант 1. Дана архітектура системи (мова Python + хмарне середовище Google Colab / Jupyter) забезпечує оптимальний баланс вартості та технічної якості і була успішно реалізована.

ВИСНОВКИ

У дипломній роботі вирішено актуальну науково-прикладну задачу — розроблено систему математичного моделювання та прогнозування індексу Джині в умовах глобальних демографічних трансформацій для п'яти європейських країн (Україна, Польща, Німеччина, Італія, Іспанія) за 2000–2024 роки. Актуальність дослідження зумовлена демографічними шоками в Україні (відтік понад 6 млн біженців, коефіцієнт фертильності на рівні 0,897–1,6 дитини на жінку), загальноєвропейськими тенденціями старіння населення (частка осіб 65+ сягає 24,63 % у Німеччині та Італії) та пов'язаною з ними нерівномірністю розподілу доходів. Огляд існуючих рішень (EViews, STATA, Tableau, Power BI) показав їх недостатність для комплексного вирішення поставлених задач, що обґрунтувало вибір аналітичної платформи на основі Python з автоматизованим збором даних через World Bank API.

Сформовано багатовимірну матрицю із семи статистичних індикаторів — індексу Джині, рівня безробіття, частки населення 65+, коефіцієнта фертильності, інфляції, чистої міграції та ВВП на душу населення. Середнє значення індексу Джині у вибірці становить 31,95 % (діапазон 24,0–38,0 %), а ВВП на душу населення варіюється від \$653 до \$56104, що свідчить про суттєву економічну поляризацію між досліджуваними країнами. Кореляційний аналіз виявив помірний позитивний зв'язок між рівнем безробіття та індексом Джині ($r = 0,44$), сильну пряму залежність між ВВП на душу населення та часткою літнього населення ($r = 0,83$) і суттєвий зворотний зв'язок між коефіцієнтом фертильності та часткою населення 65+ ($r = -0,58$). Ці результати математично підтверджують, що зростання безробіття поглиблює соціальне розшарування, а демографічне старіння та спад народжуваності є взаємообумовленими процесами, характерними для всіх досліджуваних країн.

Побудовано та порівняно дві прогностні моделі — множинну лінійну регресію та Random Forest Regressor. Ансамблева модель забезпечила суттєво вищу точність: $R^2 = 0,716$ та $MAPE = 3,56 \%$ проти $R^2 = 0,457$ та $MAPE = 5,49 \%$ у лінійної регресії. Це свідчить про нелінійний характер залежностей між демографічними та економічними показниками, які лінійна модель не здатна адекватно описати. Аналіз важливості ознак у моделі Random Forest показав, що ВВП на душу населення є абсолютним домінуючим драйвером нерівності (важливість 0,722), рівень безробіття посідає друге місце (0,143), тоді як демографічні фактори мають меншу пряму значущість через їхній опосередкований вплив — через кореляцію з ВВП. Це означає, що боротьба з нерівністю потребує передусім економічного зростання та зайнятості, а не лише демографічних інтервенцій.

Для України побудовано середньостроковий прогноз індексу Джині методом ARIMA(1,1,1). ADF-тест підтвердив стаціонарність часового ряду ($p\text{-value} = 0,0000$), а прогностна траєкторія вказує на тенденцію до поступової стабілізації показника за умови збереження поточних макроекономічних тенденцій без урахування екстремальних шоків. Кластеризація методом K-Means виділила три чітко розмежовані групи: Кластер 0 (Іспанія, Італія) — висока нерівність (Джині 33–36 %) з підвищеним безробіттям; Кластер 1 (Україна) — найнижчі значення Джині у вибірці (24–27 %), що відображає специфіку вирівнювання доходів через соціальні трансферти; Кластер 2 (Польща, Німеччина) — стабільний баланс між помірним безробіттям і контрольованою нерівністю (Джині 28–32 %).

Сценарне моделювання для України до 2030 року показало, що інерційний базовий сценарій передбачає стабілізацію індексу Джині в діапазоні 25,5–26,2 %, песимістичний сценарій (без структурних реформ) прогнозує зростання до 29,5–30,1 %, тоді як оптимістичний — за умови демографічної стабілізації та реформ ринку праці — передбачає зниження до 23,1–23,8 %, що наблизить Україну до стандартів соціально стабільних країн

ЄС. Усі прогностні траєкторії та матриці сценарного простору збережено у форматі CSV як основу для аналітичного блоку DSS.

Проведений функціонально-вартісний аналіз підтвердив доцільність обраної архітектури програмного продукту. З двох альтернативних варіантів за коефіцієнтом техніко-економічного рівня ($KTEP_1 = 3,32 \cdot 10^{-5} > KTEP_2 = 3,02 \cdot 10^{-5}$) оптимальним визначено реалізацію на Python у середовищі Jupyter Notebook / Google Colab, що забезпечує найкращий баланс між технічним рівнем системи та вартістю розробки (245 995 грн). Практична значущість роботи полягає у можливості використання розробленої DSS-системи для підтримки прийняття управлінських рішень у сфері соціально-економічної політики, оцінювання ризиків нерівності доходів та формування обґрунтованих сценаріїв розвитку для країн із різними демографічними структурами.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Системні дослідження та інформаційні технології : міжнар. наук.-техн. журнал. URL: <https://journal.iasa.kpi.ua/issue/archive> (дата звернення: 17.05.2026).
2. Наукові вісті КПІ : наук. журнал КПІ ім. Ігоря Сікорського. URL: <https://scinews.kpi.ua/issue/archive> (дата звернення: 17.05.2026).
3. World Bank Open Data. World Development Indicators. URL: <https://data.worldbank.org/> (last accessed: 17.05.2026).
4. OECD Data. URL: <https://www.oecd.org/> (last accessed: 17.05.2026).
5. Meadows D. H. Thinking in Systems: A Primer. White River Junction : Chelsea Green Publishing, 2008. 240 p.
6. Sen A. On Economic Inequality. Oxford : Oxford University Press, 1997. 464 p.
7. Wooldridge J. M. Introductory Econometrics: A Modern Approach. 7th ed. Boston : Cengage Learning, 2020. 912 p.
8. Gujarati D. N., Porter D. C. Basic Econometrics. 5th ed. New York : McGraw-Hill Education, 2017. 944 p.
9. Breiman L. Random Forests. Machine Learning. 2001. Vol. 45, No. 1. P. 5–32. DOI: 10.1023/A:1010933404324.
10. Box G. E. P., Jenkins G. M., Reinsel G. C., Ljung G. M. Time Series Analysis: Forecasting and Control. 5th ed. Hoboken : Wiley, 2015. 712 p.
11. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. New York : Springer, 2017. 745 p. DOI: 10.1007/978-0-387-84858-7.
12. Godet M. Creating Futures: Scenario Planning as a Strategic Management Tool. 2nd ed. London : Economica, 2006. 370 p.
13. Turban E., Sharda R., Delen D. Decision Support and Business Intelligence Systems. 10th ed. Boston : Pearson, 2014. 720 p.

14. World Bank Indicators API Documentation. URL: <https://datahelpdesk.worldbank.org/knowledgebase/topics/125589-developer-information> (last accessed: 01.06.2026).
15. Scikit-learn: Machine Learning in Python. URL: <https://scikit-learn.org/> (last accessed: 01.06.2026).

ДОДАТОК А Лістинг програми

```
import warnings
warnings.filterwarnings("ignore")

import pandas as pd
import numpy as np

import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split

from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestRegressor

from sklearn.metrics import (
    mean_absolute_error,
    mean_squared_error,
    mean_absolute_percentage_error,
    r2_score
)

from sklearn.cluster import KMeans

from statsmodels.tsa.arima.model import ARIMA
from statsmodels.tsa.stattools import adfuller
```

```
import wbdata
```

```
from datetime import datetime
```

```
#%% md
```

Завантаження даних з World Bank API. для роботи візьмемо 5 країн:
україну, польщу, німеччину, італію, іспанію

```
#%%
```

```
countries = [
```

```
    "UKR",
```

```
    "POL",
```

```
    "DEU",
```

```
    "ITA",
```

```
    "ESP"
```

```
]
```

```
# World Bank
```

```
indicators = {
```

```
    "SI.POV.GINI": "gini_index",
```

```
    "SL.UEM.TOTL.ZS": "unemployment_rate",
```

```
    "SP.POP.65UP.TO.ZS": "population_65_plus",
```

```
    "SP.DYN.TFRT.IN": "fertility_rate",
```

```
    "FP.CPI.TOTL.ZG": "inflation",
```

```
    "SM.POP.NETM": "net_migration",
```

```
    "NY.GDP.PCAP.CD": "gdp_per_capita"
```

```
}
```

```
data_date = (
```

```
    datetime(2000, 1, 1),
```

```
    datetime(2024, 1, 1)
```

)

```
df = wbdata.get_dataframe(
    indicators,
    country=countries,
    date=data_date
)
```

індекси

```
df.reset_index(inplace=True)
df.index = df.index + 1
```

```
df.rename(columns={
    "country": "Country",
    "date": "Year"
}, inplace=True)
```

```
print("\n          Розглянемо перші 20 рядків датасету:")
print(df.head(20))
print(df["Country"].unique())
print(df.shape) # розмір датасету
print(df["Country"].value_counts()) # кількість записів по кожній країні
# %% md
```

Збереження даних

```
# %%
df.to_csv(
    "economic_demographic_data.csv",
    index=False
)
```

```

print("\nДані успішно збережено")
#%%% md

#%%%

print("\n    Інформація про датасет")
print(df.info())

print("\n                                Описова статистика")
print(df.describe())

print("\n    Пропуски")
print(df.isnull().sum())
#%%% md
пропуски
#%%%
df.sort_values(
    by=["Country", "Year"],
    inplace=True
)

# заповнення
columns_to_fill = [
    "unemployment_rate",
    "population_65_plus",
    "fertility_rate",
    "inflation",
    "net_migration",
    "gdp_per_capita"
]

```

```

df[columns_to_fill] = df[columns_to_fill].ffill()
df[columns_to_fill] = df[columns_to_fill].bfill()

# видалення
df.dropna(
    subset=["gini_index"],
    inplace=True)

df.reset_index(drop=True, inplace=True)
df.index = df.index + 1

print("\n    Після очищення:")
print(df.isnull().sum())

print("\n                Перші 20 рядків після очищення:")
print(df.head(30))
#%% md
Візуалізуємо дані
#%%
country_colors = {
    "Ukraine": "#0057B7",
    "Germany": "#D00000",
    "Poland": "#6A0DAD",
    "Italy": "#008C45",
    "Spain": "#FFC400"
}

# графік індексу Джині
plt.figure(figsize=(14, 7))

```

```

for country in df["Country"].unique():

    country_df = df[df["Country"] == country]

    plt.plot(
        country_df["Year"],
        country_df["gini_index"],
        label=country,
        color=country_colors[country],
        linewidth=2
    )

    plt.title(
        "Динаміка індексу Джині",
        fontsize=16
    )

    plt.xlabel("Рік")
    plt.ylabel("Індекс Джині")
    plt.grid(True)
    plt.legend()

    plt.show()

# графік безробіття
plt.figure(figsize=(14, 7))

for country in df["Country"].unique():

```

```

country_df = df[df["Country"] == country]

plt.plot(
    country_df["Year"],
    country_df["unemployment_rate"],
    label=country,
    color=country_colors[country],
    linewidth=2
)

plt.title(
    "Динаміка безробіття",
    fontsize=16
)

plt.xlabel("Рік")
plt.ylabel("Безробіття")
plt.grid(True)
plt.legend()

plt.show()

# графік старіння
plt.figure(figsize=(14, 7))

for country in df["Country"].unique():

```

```
country_df = df[df["Country"] == country]
```

```
plt.plot(
    country_df["Year"],
    country_df["population_65_plus"],
    label=country,
    color=country_colors[country],
    linewidth=2
)
```

```
plt.title(
    "Частка населення віком 65+",
    fontsize=16
)
```

```
plt.xlabel("Рік")
plt.ylabel("Населення 65+")
plt.grid(True)
plt.legend()
```

```
plt.show()
```

```
### md
```

Кореляційний аналіз

```
###
```

```
numeric_df = df.select_dtypes(include=np.number)
```

```
correlation_matrix = numeric_df.corr()
```

```
plt.figure(figsize=(12, 10))
```

```
sns.heatmap(  
    correlation_matrix,  
    annot=True,  
    cmap="magma",  
    fmt=".2f"  
)  
  
plt.title(  
    "Кореляційна матриця економічних та демографічних показників",  
    fontsize=15  
)  
  
plt.show()  
#%% md  
Масштабування даних  
#%%  
features = [  
    "unemployment_rate",  
    "population_65_plus",  
    "fertility_rate",  
    "inflation",  
    "net_migration",  
    "gdp_per_capita"  
]  
  
target = "gini_index"  
  
scaler = StandardScaler()
```

```

df_scaled = df.copy()

df_scaled[features] = scaler.fit_transform(
    df_scaled[features]
)

print("\nМасштабування виконано")
#%% md
Регресійний аналіз
#%%
X = df_scaled[features]

y = df_scaled[target]

# розділення вибірки
X_train, X_test, y_train, y_test = train_test_split(
    X,
    y,
    test_size=0.2,
    random_state=42
)

linear_model = LinearRegression()

linear_model.fit(X_train, y_train)

linear_predictions = linear_model.predict(X_test)

# метрики
linear_mae = mean_absolute_error(

```

```

    y_test,
    linear_predictions
)

```

```

linear_rmse = np.sqrt(
    mean_squared_error(
        y_test,
        linear_predictions
    )
)

```

```

linear_mape = mean_absolute_percentage_error(
    y_test,
    linear_predictions
)

```

```

linear_r2 = r2_score(
    y_test,
    linear_predictions
)

```

```

print("\nLinear regression results")
print(f"MAE : {linear_mae:.4f}")
print(f"RMSE : {linear_rmse:.4f}")
print(f"MAPE : {linear_mape:.4f}")
print(f"R2 : {linear_r2:.4f}")
#%%% md

```

Ознаки

```

#%%%

```

```

coefficients = pd.DataFrame({

```

```

    "Feature": features,
    "Coefficient": linear_model.coef_
})

coefficients = coefficients.sort_values(
    by="Coefficient",
    ascending=False
)

print("\n    Вплив факторів")
print(coefficients)

plt.figure(figsize=(10, 6))

sns.barplot(
    data=coefficients,
    x="Coefficient",
    y="Feature",
    color="purple"
)

plt.title(
    "Важливість факторів у регресійній моделі",
    fontsize=15
)

plt.show()
#%% md
Random Forest
#%%

```

```
forest_model = RandomForestRegressor(  
    n_estimators=200,  
    random_state=42  
)  
  
forest_model.fit(X_train, y_train)  
  
forest_predictions = forest_model.predict(X_test)  
  
forest_mae = mean_absolute_error(  
    y_test,  
    forest_predictions  
)  
  
forest_rmse = np.sqrt(  
    mean_squared_error(  
        y_test,  
        forest_predictions  
    )  
)  
  
forest_mape = mean_absolute_percentage_error(  
    y_test,  
    forest_predictions  
)  
  
forest_r2 = r2_score(  
    y_test,  
    forest_predictions  
)
```

```

print("\nRandom forest results")
print(f"MAE : {forest_mae:.4f}")
print(f"RMSE : {forest_rmse:.4f}")
print(f"MAPE : {forest_mape:.4f}")
print(f"R2 : {forest_r2:.4f}")
#%%% md

Feature importance
#%%%

importance_df = pd.DataFrame({
    "Feature": features,
    "Importance": forest_model.feature_importances_
})

importance_df = importance_df.sort_values(
    by="Importance",
    ascending=False
)

print("\n    Feature importance")
print(importance_df)

plt.figure(figsize=(10, 6))

sns.barplot(
    data=importance_df,
    x="Importance",
    y="Feature",
    color="purple"
)

```

```

plt.title(
    "Важливість ознак Random Forest",
    fontsize=15
)

plt.show()
#%% md
Кластеризація країн
#%%
cluster_features = df_scaled[
    [
        "gini_index",
        "unemployment_rate",
        "population_65_plus"
    ]
]

kmeans = KMeans(
    n_clusters=3,
    random_state=42
)

df_scaled["Cluster"] = kmeans.fit_predict(
    cluster_features
)

plt.figure(figsize=(10, 7))

sns.scatterplot(

```

```

data=df_scaled,
x="unemployment_rate",
y="gini_index",
hue="Cluster",
style="Country",
s=100
)

plt.title(
    "Кластеризація країн",
    fontsize=15
)

plt.show()

#%% md
Перевірка стаціонарності
#%%

ukraine_df = df[
    df["Country"] == "Ukraine"
]

ukraine_df = ukraine_df.sort_values(
    by="Year"
)

series = ukraine_df["gini_index"].astype(float)

adf_result = adfuller(series)

print("\nADF TEST")

```

```
print(f"ADF Statistic : {adf_result[0]:.4f}")
print(f"p-value      : {adf_result[1]:.4f}")
```

```
if adf_result[1] < 0.05:
    print("Ряд є стаціонарним")
else:
    print("Ряд є нестаціонарним")
```

```
#%% md
```

Прогнозування arima

```
#%%
```

```
arima_model = ARIMA(
    series,
    order=(1, 1, 1)
)
```

```
arima_fitted = arima_model.fit()
```

```
print(arima_fitted.summary())
```

```
#%% md
```

прогноз на 5 років

```
#%%
```

```
forecast_steps = 5
```

```
forecast = arima_fitted.forecast(
    steps=forecast_steps
)
```

```
future_years = range(
    int(ukraine_df["Year"].max()) + 1,
    int(ukraine_df["Year"].max()) + 6
```

)

```
forecast_df = pd.DataFrame({
    "Year": future_years,
    "Forecasted_Gini_Index": forecast
})
```

```
print("\n    Прогноз індексу Джині")
```

```
print(forecast_df)
```

```
#%% md
```

Візуалізація прогнозу

```
#%%
```

```
ukraine_df["Year"] = pd.to_numeric(ukraine_df["Year"])
```

```
plt.figure(figsize=(14, 7))
```

```
plt.plot(
    ukraine_df["Year"],
    series,
    label="Історичні дані",
    color="purple",
    linewidth=2
```

)

```
plt.plot(
    future_years,
    forecast,
    label="ARIMA прогноз",
    color="orange",
    linestyle="dashed",
    linewidth=3
```

)

```
plt.title(
    "Прогноз індексу Джині України",
    fontsize=16
)
```

```
plt.xlabel("Рік")
plt.ylabel("Індекс Джині")
```

```
plt.grid(True)
plt.legend()
```

```
plt.show()
#%% md
Сценарне моделювання
#%%
# базовий сценарій
baseline = forecast
```

```
# оптимістичний сценарій
optimistic = forecast * 0.90
```

```
# песимістичний сценарій
pessimistic = forecast * 1.15
```

```
scenario_df = pd.DataFrame({
    "Year": future_years,
    "Baseline": baseline,
    "Optimistic": optimistic,
```

```

    "Pessimistic": pessimistic
})

print("\n        Сценарний аналіз")
print(scenario_df)
#%% md
Візуалізація сценаріїв
#%%
plt.figure(figsize=(14, 7))

plt.plot(
    future_years,
    baseline,
    label="Baseline",
    color="blue",
    linewidth=3
)

plt.plot(
    future_years,
    optimistic,
    label="Optimistic",
    color="orange",
    linewidth=3
)

plt.plot(
    future_years,
    pessimistic,
    label="Pessimistic",

```

```

        color="purple",
        linewidth=3
    )

plt.title(
    "Сценарії прогнозування соціально-економічної нерівності",
    fontsize=16
)

plt.xlabel("Рік")
plt.ylabel("Індекс Джині")

plt.grid(True)
plt.legend()

plt.show()

#%%% md
Порівняння моделей
#%%%

comparison_df = pd.DataFrame({
    "Model": [
        "Linear Regression",
        "Random Forest"
    ],
    "MAE": [
        linear_mae,
        forest_mae
    ],
    "RMSE": [
        linear_rmse,

```

```

        forest_rmse
    ],
    "MAPE": [
        linear_mape,
        forest_mape
    ],
    "R2": [
        linear_r2,
        forest_r2
    ]
})

```

```

print("\n          Порівняння моделей")
print(comparison_df)

```

```

#%%% md

```

```

#%%%
forecast_df.to_csv(
    "forecast_results.csv",
    index=False
)

```

```

comparison_df.to_csv(
    "model_comparison.csv",
    index=False
)

```

```

scenario_df.to_csv(
    "scenario_analysis.csv",

```

```
        index=False
    )

    print("\nРезультати успішно збережено")
```

ДОДАТОК Б Презентація

КПІ ім. Ігоря Сікорського · ІПСА · Кафедра ММСА

Київ, 2026

Математичне моделювання та прогнозування показників економічного стану населення в умовах глобальних демографічних трансформацій

Виконала: студентка групи КА-23 Коваленко Олександра Павлівна
Науковий керівник: доцент, к.ф. Гуськова Віра Геннадіївна

● ОБ'ЄКТ ДОСЛІДЖЕННЯ

Соціально-економічні та демографічні процеси в п'яти європейських країнах (Україна, Польща, Німеччина, Італія, Іспанія) в умовах сучасних демографічних трансформацій за 2000–2024 роки.

● ПРЕДМЕТ ДОСЛІДЖЕННЯ

Математичні моделі та методи прогнозування індексу Джині на основі демографічних (частка населення 65+, фертильність, міграція) та макроекономічних (ВВП, безробіття, інфляція) чинників.

● МЕТА РОБОТИ

Розробити систему математичного моделювання та прогнозування індексу Джині, що поєднує множину регресію, Random Forest, ARIMA, K-Means та сценарне моделювання в єдиному аналітичному контурі DSS.

Актуальність теми

01 5 млн +

українців виїхали за кордон —
критичний демографічний шок

02 0,9

дитини на жінку — мінімум
народжуваності у вибірці

03 24,6%

частка населення 65+ у
Німеччині та Італії (максимум)

04 \$653 – \$56 104

діапазон ВВП на душу населення
між країнами вибірки

Аналіз існуючих програмних рішень

Інструмент	Переваги	Недоліки	ML	API
EViews	Сильний економетричний аналіз, ARIMA, VAR	Закритий код, обмежені можливості ML	✗	✗
STATA	Розвинений регресійний аналіз, підтримка панельних даних	Платна ліцензія, складна інтеграція з API	—	✗
Tableau / Power BI	Зручна візуалізація, інтерактивні дашборди	Обмежені можливості прогнозування та кластеризації	—	✓
Python	Повний аналітичний цикл, відкритий код, велика кількість бібліотек	Потребує програмування та налаштування середовища	✓	✓

Математичний апарат. Індекс Джині та регресія

Індекс Джині

Коефіцієнт Джині G — відношення площі між лінією абсолютної рівності та кривою Лоренца $L(x)$ до площі під лінією рівності:

$$G = 1 - 2 \int_0^1 L(x) dx$$

Для дискретного набору даних (n підгруп):

$$G = \sum_i \sum_j |y_i - y_j| / (2n^2 \bar{y})$$

де y_i — дохід i -ї особи (підгрупи), $\bar{y}$ — середній дохід по вибірці.
 $G \in [0, 1]$: 0 — абсолютна рівність, 1 — абсолютна нерівність.
 Діапазон у вибірці: 24,0 % – 38,0 % (середнє 31,95 %)

Множинна лінійна регресія

Матрична форма моделі:

$$Y = X\beta + \varepsilon$$

де Y — вектор спостережень ($n \times 1$); X — матриця плану ($n \times (k+1)$); β — вектор параметрів; ε — вектор похибок.

Оцінка β методом МНК — мінімізація $S(\beta) = \varepsilon^T \varepsilon \rightarrow \min$:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Random Forest — критерій розщеплення у вузлі (MSE):

$$Q = (1/|R_m|) \sum_{i \in R_m} (y_i - \bar{y}_{R_m})^2$$

Математичний апарат. Методологія ARIMA (Бокс–Дженкінс)

01 AR(p)

$$Y_t = c + \sum_{i=1}^p \varphi_i Y_{t-i} + \varepsilon_t$$

φ_i — параметри; ε_t — білий шум; p — порядок авторегресії.

Умова стаціонарності Y_t :

$$E(Y_t) = \mu, D(Y_t) = \sigma^2, \\ y(\tau) = \text{cov}(Y_t, Y_{t-\tau})$$

02 MA(q)

$$Y_t = \mu + \sum_{i=1}^q \theta_i \varepsilon_{t-i} + \varepsilon_t$$

θ_i — параметри; q — порядок ковзного середнього.

03 ARMA(p,q)

$$Y_t = c + \sum_{i=1}^p \varphi_i Y_{t-i} + \sum_{i=1}^q \theta_i \varepsilon_{t-i} + \varepsilon_t$$

Поєднує AR і MA. Порядки p, q — за мін. AIC/BIC.

04 ARIMA(p,d,q)

$$\Phi_p(L)(1-L)^d Y_t = \Theta_q(L) \varepsilon_t$$

L — оператор лагу; d — порядок диференціювання.

05 SARIMA(p,d,q)(P,D,Q)s

$$\Phi_p(L) \Phi_p(L^s)(1-L)^d(1-L^s)^P Y_t = \Theta_q(L) \Theta_q(L^s) \varepsilon_t$$

s — довжина сезонного циклу; (P,D,Q) — сезонні порядки.

Математичний апарат. K-Means та метрики якості

Кластеризація K-Means

Мінімізація внутрішньокластерної дисперсії:

$$J = \sum_{k=1}^K \sum_{x_i \in S_k} \|x_i - \mu_k\|^2 \rightarrow \min$$

де μ_k — центроїд кластера S_k :

$$\mu_k = (1/|S_k|) \sum_{x_i \in S_k} x_i$$

Ознаки кластеризації: індекс Джині, рівень безробіття, частка населення 65+. Обрано $k = 3$ кластери.

Метрики оцінки якості моделей

$$MAE = (1/n) \sum_{i=1}^n |y_i - \hat{y}_i|$$

$$MSE = (1/n) \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$RMSE = \sqrt{(1/n) \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

$$MAPE = (1/n) \sum_{i=1}^n |y_i - \hat{y}_i| / |y_i| \times 100 \%$$

$$R^2 = 1 - \sum (y_i - \hat{y}_i)^2 / \sum (y_i - \bar{y})^2$$

Random Forest — усереднення N дерев

$$\hat{y}(x) = (1/N) \sum_{b=1}^N T_b(x)$$

де $T_b(x)$ — вихід b-го дерева для вектора ознак x .

Усереднення знижує дисперсію без збільшення зсуву.

Архітектура DSS (4 рівні):

Data Layer → Processing Layer →

Modeling Layer → Decision Layer

Вхідні дані та дескриптивний аналіз

Матриця даних: 125 × 9

5 країн · 25 років (2000–2024) · 7 індикаторів + 2 ідентифікатори

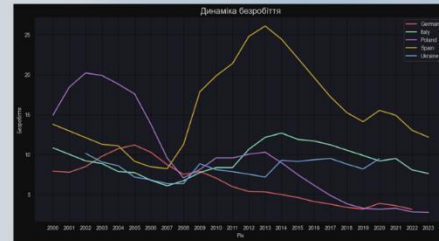
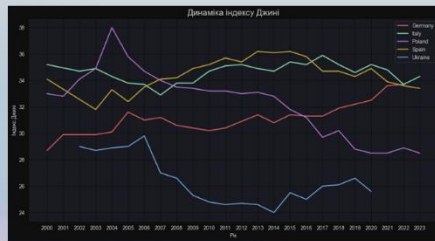
Gini Index	цільова змінна
Unemployment rate	безробіття, %
Population 65+	частка літніх, %
Fertility rate	коеф. народжуваності
Inflation	інфляція, %
Net migration	чиста міграція, млн
GDP per capita	ВВП/особу, USD

Дескриптивні статистики

Показник (метрика)	Мінімум (min)	Середнє (mean)	Максимум (max)
gini_index	24	31.952	38
unemployment_rate	2.743	9.81	26.094
population_65_plus	12.244	18.08	24.625
fertility_rate	0.897	1.326	1.6
inflation	-0.874	4.45	48.699
net_migration	-5,699.445	114,739	1,175,283
gdp_per_capita	653.29	23,874	56,104

- Великий розрив ВВП свідчить про економічну поляризацію між країнами
- Від'ємна міграція (–5,69 млн) відображає шок від повномасштабного вторгнення
- Фертильність 0,897 — демографічна криза, критичний рівень

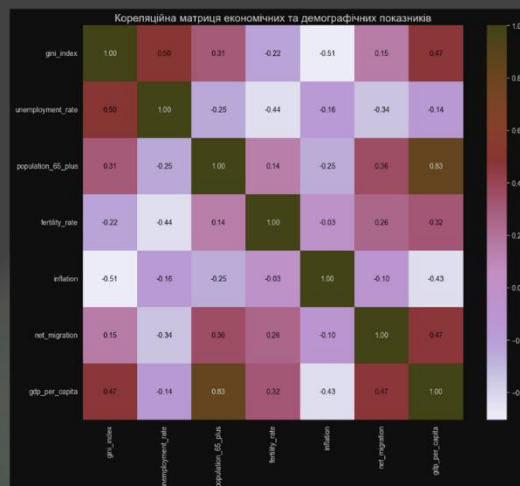
Динаміка ключових показників по країнах (2000–2024)



Ключові спостереження

- Україна — найнижчий Gini (24–27 %), різке зростання безробіття після 2022 р.
- Іспанія та Італія — стабільно високий Gini (33–36 %) і найвищий рівень безробіття
- Всі країни: безперервне старіння населення; Італія сягає 24,6 % літнього населення
- Польща — найдинамічніше зниження безробіття (17,7 % → 3,0 %)

Кореляційний аналіз (матриця Пірсона)



Ключові залежності

$r = +0,50$
Gini ↑ з безробіттям

$r = +0,83$
ВВП ↑ з часткою 65+

$r = -0,51$
інфляція ↓ знижує
Gini

$r = +0,47$
ВВП та Gini
позитивно пов'язані

Побудова та порівняння моделей прогнозування

Модель (Model)	MAE	RMSE	MAPE	R^2 (Коеф. детермінації)
Linear Regression	1.7367	2.0474	0.0549	0.4565
Random Forest Regressor	1.1022	1.4808	0.0356	0.7157

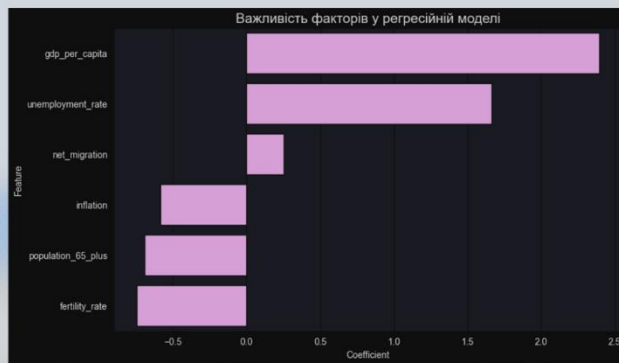
Переваги Random Forest:

- Враховує нелінійні залежності між факторами
- Стійкий до мультиколінеарності
- Автоматичний відбір найважливіших ознак
- Оцінює feature importance (внесок кожного фактора)
- Не вимагає припущень про розподіл залишків

Висновки

- 1) Random Forest перевищує регресію за всіма метриками.
- 2) R^2 зріс на 56,9 % — модель пояснює 71,6 % дисперсії Джині.
- 3) Вирішальний фактор — ВВП / особу (вага 72,2 %): боротьба з нерівністю потребує насамперед економічного зростання.
- 4) Демографічні фактори діють опосередковано через ВВП ($r = 0,83$).

Важливість факторів у лінійній регресійній моделі



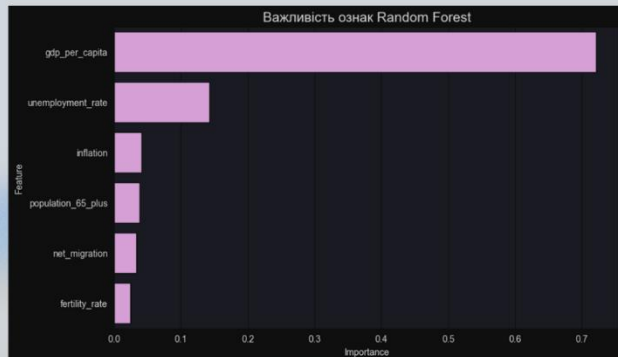
Висновки

Лінійна регресія зафіксувала:

- 1) + gdp_per_capita: +2,39 (зростання ВВП → ↑ Gini)
- 2) + unemployment: +1,66 (безробіття → ↑ Gini)
- 3) - fertility_rate: -0,74 (вища народжуваність → ↓ Gini)
- 4) - population_65+: -0,69 (більше літніх → ↓ Gini)

$R^2 = 0,457$ — модель пояснює 45,6 % дисперсії Gini

Важливість ознак у моделі Random Forest Regressor



Висновки

- 1) ВВП/особу — абсолютний домінуючий фактор: 72,17 % всіх рішень прийнято на основі цього індикатора.
- 2) Безробіття посідає 2-е місце: 14,30 %.
- 3) Усі демографічні фактори разом: лише ~13 % — але діють опосередковано через ВВП ($r = 0,83$ між ВВП і часткою 65+).

$R^2 = 0,716$ — модель пояснює 71,6 % дисперсії Gini.

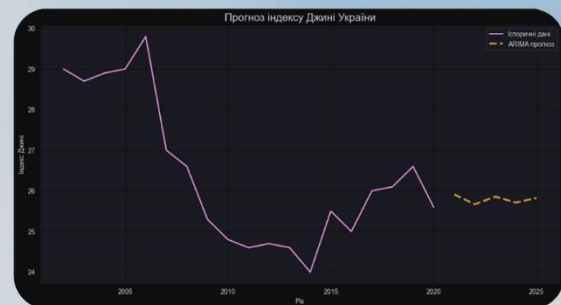
Прогнозування ARIMA та кластеризація K-Means

Результати ARIMA

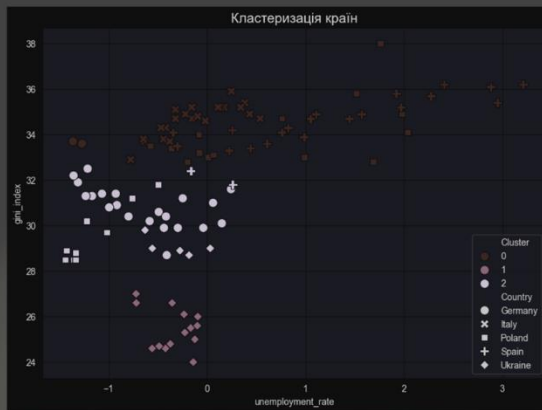
ADF-тест:
ADF = -4,986
p-value = 0,0000
 H_0 відхиляється → ряд
стаціонарний

Прогноз:
Джині стабілізується у
діапазоні 25,5–26,2 %
(базовий сценарій)

Модель:
ARIMA(1,1,1)
(AR + diff + MA)



Кластеризація країн (K-Means, k = 3)



01 Висока нерівність

Іспанія, Італія · Джині 33–36 %
Хронічно високе безробіття (10–22 %) супроводжується стабільно високим рівнем соціального розшарування.

02 Унікальна модель

Україна · Джині 24–27 %
Найнижчий Gini у вибірці при помірному безробітті.
Специфіка — вирівнювання через соціальні трансферти.

03 Стабільний баланс

Польща, Німеччина · Джині 28–35 %
Контрольований рівень нерівності при низькому або помірному безробітті та активному демографічному старінні.

Сценарне моделювання для України (2025–2030)



Базовий

Джині: 25,5–26,2 %
Збереження тенденцій,
помірне відтворення робочої сили

Оптимістичний

Джині: 23,1–23,8 %
Реформи ринку праці,
демографічна стабілізація,
стандарти ЄС

Песимістичний

Джині: 29,5–30,1 %
Відтік населення, зростання
безробіття, поляризація
доходів

Висновки

- 01 Розроблено аналітичну DSS-систему на Python з автоматизованим збором даних через World Bank API (матриця 125×9 , 7 показників).
- 02 Random Forest ($R^2 = 0,716$, MAPE = 3,56 %) перевершив лінійну регресію ($R^2 = 0,457$). Ключовий фактор нерівності — ВВП/особу (вага 72,2 %).
- 03 ADF-тест підтвердив стаціонарність ряду України ($p = 0,0000$). ARIMA (1,1,1) прогнозує стабілізацію Джині на рівні 25,5–26,2 %.
- 04 K-Means виділила 3 кластери: висока нерівність (Іспанія, Італія), ізольована модель (Україна), стабільний баланс (Польща, Німеччина).
- 05 Сценарне моделювання до 2030 р.: базовий 25,5–26,2 %, оптимістичний 23,1–23,8 %, песимістичний 29,5–30,1 %.

Дякую за увагу!