Інститут телекомунікацій і глобального інформаційного простору
Національна академія наук України

Кваліфікаційна наукова праця
на правах рукопису

КОВАЛЬ РОМАН ГРИГОРОВИЧ

Прим. № ____
004.67:004.89:[351.84:364.3-64](043.3)

ДИСЕРТАЦІЯ

МОДЕЛІ, МЕТОДИ ТА ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ ЗБОРУ ТА ОБРОБКИ ДАНИХ ДЛЯ АНАЛІЗУ ТА ПРОГНОЗУВАННЯ ВИДАТКІВ НА СОЦІАЛЬНЕ ЗАБЕЗПЕЧЕННЯ ТА СОЦІАЛЬНИЙ ЗАХИСТ

05.13.06 – інформаційні технології

Подається на здобуття наукового ступеня кандидата технічних наук

Дисертація містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело

_____ Р. Г. Коваль
(підпис, ініціали та прізвище здобувача)

Науковий консультант
Трофимчук Олександр Миколайович,
член-кореспондент НАН України,
д.т.н., професор

Київ – 2025

АНОТАЦІЯ

Коваль Р. Г. Інформаційні технології збору та оброблення даних для прогнозування витрат на пенсії і соціальні допомоги– Кваліфікаційна праця на правах рукопису.

Дисертація на здобуття наукового ступеня кандидата технічних наук за фахом 05.13.06 «Інформаційні технології». – Інститут телекомунікацій і глобального інформаційного простору, Національна академія наук України, Київ, 2025.

Дисертаційна робота присвячена дослідженню методів, засобів, технологій збору та оброблення даних щодо процесів, які відбуваються у соціальній сфері країни, використання прогнозуючих моделей високого ступеня адекватності, які дозволяють забезпечувати високу якість прогнозів витрат на пенсії та соціальні допомоги.

Створено нові інформаційні технології збору та оброблення даних, моделювання та прогнозування витрат на виплату пенсій та соціальних допомог шляхом інтеграції моделей, методів, алгоритмів і програмних засобів, які забезпечують високу адекватність моделей і задану якість прогнозів, призначені до використання у Єдиній інформаційній системі соціальної сфери України, а також є простими у використанні та зручними для користувачів. Запропонована методика побудови моделі процесів, що відбуваються у соціальній сфері відзначається робастністю, адаптивністю, є гнучкою, призначена для реалізації у комп'ютерних системах підтримки прийняття рішень Запропоновано модифіковану методику моделювання часових рядів, на основі якої побудовано ансамблі моделей і вдосконалено методику використання структурованих та неструктурованих даних у задачах прогнозування витрат на пенсії та соціальні допомоги.

Ключові слова: інформаційні технології, математичні моделі, соціальна сфера, прогнозування, штучний інтелект, збір та оброблення даних.

ABSTRACT

Koval R. G. Information technologies for collecting and processing data for forecasting expenses for pensions and social benefits. – qualification scientific work in the form of a manuscript.

Dissertation for the degree of candidate of technical sciences in the specialty 05.13.06 – "Information technologies" – Institute of telecommunications and global information space of the NAS of Ukraine, Kyiv, 2025.

The dissertation is devoted to the study of methods, means, technologies for collecting and processing data on processes occurring in the social sphere of the country, the use of predictive models with a high degree of adequacy, which allow ensuring high quality forecasts of expenses for pensions and social benefits.

New information technologies for data collection and processing, modeling and forecasting of pension and social assistance costs have been created by integrating models, methods, algorithms and software tools that ensure high adequacy of models and a given quality of forecasts, are intended for use in the unified information system of the social sphere of Ukraine, and are also easy to use and user-friendly. The proposed methodology for building a model of processes occurring in the social sphere is characterized by robustness, adaptability, is flexible, and is intended for implementation in computer decision support systems. A modified time series modeling methodology has been proposed, on the basis of which ensembles of models have been built and the methodology for using structured and unstructured data in the tasks of forecasting pension and social assistance costs has been improved.

Keywords: information technologies, mathematical models, social sphere, forecasting, artificial intelligence, data collection and processing.

Список публікацій здобувача за темою дисертації

Статті у періодичних виданнях, включених до категорії «А» Переліку наукових фахових видань України

1. Зарудний О. Б. і Коваль Р. Г. Методика прогнозування часових рядів для визначення потреби у видатках на соціальний захист та соціальне забезпечення», *Міжнародний науково-технічний журнал «Проблеми керування та інформатики»*. 2024. Vol. 69, No. 5. С. 64–77. https://doi.org/10.34229/1028-0979-2024-5-5

Статті у наукових фахових виданнях України

2. Zarudnyi, O., Koval, R. Application of probabilistic and stochastic models and data mining for forecasting the contingent of old age pension recipients in the context of systemic uncertainty. *Technology Audit and Production Reserves*, 2024. Vol. 5. No. 2(79), P. 56–62. https://doi.org/10.15587/2706-5448.2024.313960

3. Трофимчук О., Коваль Р., Зарудний О. Методика прогнозування кількості інвалідів з числа санітарних втрат. *Екологічна безпека та природокористування*, 2025. Vol. 54. No. 2. С. 121–135. https://doi.org/10.32347/2411-4049.2025.2.121-135

Статті у закордонних виданнях, індексованих у базі даних Scopus

4. Zarudnyi, O., Koval, R. Methodology for Constructing an Analytical Subsystem of the Unified Information System of the Social Sphere of Ukraine. *Cybernetics and Systems Analysis.* 2025. Vol. 61. No. 4. P. 487–494. https://doi.org/10.1007/s10559-025-00785-9 (**Scopus)**

5. Korbicz J., Sholokhov O., Koval R., Zarudnyi O. Application of SAS Text Miner for the analysis of citizens' appeals in the system of social protection and social security. *8th International Scientific and Practical Conference Applied Information Systems and Technologies in the Digital Society*. (AISTDS 2024). (Kyiv , 1 Oct. 2024). 2024. Vol. 3942, P. 113 – 124. ISSN 1613-0073 URL: https://www.scopus.com/pages/publications/105001096518 (**Scopus)**

Наукові праці, які засвідчують апробацію матеріалів дисертації

6. Зарудний О.Б., Коваль Р. Г. Проблеми прогнозування факторів, що

впливають на обсяг видатків Пенсійного фонду України Інформаційно-комунікаційні технології для перемоги та відновлення // Колективна монографія за матеріалами *XXII Міжнародної науково-практичної конференції «Інформаційно-комунікаційні технології та сталий розвиток»* (Київ, 14-15 листопада 2023 р.) / За заг. ред. С.О. Довгого. – К.: ТОВ «Видавництво «Юстон», 2023. С. 76-79. URL: https://itgip.org/wp-content/uploads/2023/11/1_zbirka_08_11_23-1-1.pdf

7. Зарудний О.Б., Коваль Р. Г. Застосування методу сингулярного розкладання у задачі аналізу електронних звернень громадян до Пенсійного фонду України. Математичне моделювання та інформаційно-комунікаційні технології для зміцнення та відновлення // Колективна монографія за матеріалами XXIII Міжнародної науково-практичної конференції «*Математичне моделювання та інформаційно-комунікаційні технології для зміцнення та відновлення»* (Київ, 12-13 листопада 2024 р.) / За заг. ред. С.О. Довгого. К.: ТОВ «Видавництво «Юстон», 2024. С. 156-159. URL: https://itgip.org/wp-content/uploads/2024/11/2024-11-24_zbirka_all_07_11_2024_148x210.pdf

8. Зарудний О. Б., Коваль Р. Г. Шолохов О. В. Методика застосування кластеризації текстів на основі лінгвістичних правил для дослідження потреб населення у соціальному захисті та соціальному забезпеченні Прикладні інформаційні системи та технології в цифровому суспільстві: зб. Тез доповідей і наук. повідомлень учасників VIII *Міжнародної науково-практичної конференції «Прикладні інформаційні системи та технології в цифровому суспільстві»* (Київ, 01 жовтня 2024 р.) // за заг. ред. В. Плескач, О. Фендьо. К.: Київський нац. ун-т ім. Тараса Шевченка, 2024. С. 167-178. URL: https://aistis.knu.ua/wp-content/uploads/2024/10/Text_Book_confernce_01.10.2024.pdf

Авторське свідоцтво

9. Зарудний О. Б., Колбун В. А., Малецький О. М., Машкін В. Г., Рябцева Т. Б., Матвєєв С. В., Хандрос Ю. О., Коваль Р. Г., Бондарчук О. М., Педан Ю. А., Павлюков І. О. Комп'ютерна програма «Інтегрована комплексна система Пенсійного фонду України «ІКІС ПФУ», версія 2» №11056; заявл.31.10.2011; опубл. 11.11.2011.

ЗМІСТ

ВСТУП

Актуальність теми дослідження. Інформаційно-аналітичні системи, комп'ютерні системи підтримки прийняття рішень, системи електронного документообігу, веб-портали, тощо, активно використовуються у державному управлінні та публічному урядуванні як за кордоном, так і в Україні**.** Відмінністю процесів інформатизації у соціальній сфері є формування єдиного інформаційного простору — Єдиної інформаційної системи соціальної сфери України. Замість розрізнених реєстрів та баз даних впроваджується цілісна інформаційна система з єдиною програмно-технічною архітектурою, здатна ефективно адаптуватися до нових задач та викликів, що постають перед країною.

Сучасні виклики роблять особливо актуальними задачі аналізу та прогнозування потреб Пенсійного фонду України та бюджетів різних рівнів у фінансуванні витрат на виплату пенсій та соціальних допомог. Потрібні нові науково обґрунтовані підходи до розроблення інформаційних технологій, адаптованих до умов соціальної сфери, які дозволяли б виявляти потреби різних груп населення у соціальному захисті та соціальному забезпеченні, враховувати вплив чинників, що найбільше впливають на ситуацію у соціальній сфері, розглядати різні сценарії на довго-, середньо- та короткострокову перспективу, тощо. Тому, тема роботи є актуальною, має наукове та практичне значення.

Інформаційно-аналітичному забезпеченню соціальної сфери присвячені роботи як закордонних, так і вітчизняних науковців. У них розглянуто різні аспекти розроблення та впровадження інформаційно-аналітичних систем. Використанню математичних моделей та інструментів інтелектуального аналізу даних у системах підтримки прийняття рішень значну увагу приділено в роботах П. І. Бідюка, Л. Ф. Гуляницького, С. О. Довгого, Т. В. Запорожець, О. М. Трофимчука. Досвід розроблення інформаційних систем для соціальної сфери описано в роботах вітчизняних та закордонних науковців: П. Антарі,

А. М. Д’аркангеліс, О. Додон, О. Коваленко, Ф. Л. Колавола, О. А. Одуволя, А. Пертиві, Ф. Р. Прадана, Ф. Ріфлі, Д. Томара, Г. П. Ситника, А. Ф. Фестаса. Забезпечення надійності та ефективності функціонування комп’ютерних систем досліджено у роботах В. А. Пепеляєва, О. М. Хіміча, М. Т. Фісуна, оброблення великих даних — у роботах М. П. Комара, Д. В. Ланде, К. Лінча.

Незважаючи на значну кількість досліджень, проблема розроблення моделей, методів та інформаційних технологій, призначених для використання у системі соціального захисту та соціального забезпечення залишається невирішеною повністю. Зокрема, слід відзначити відсутність загальних підходів до збору, аналізу та попередньої обробки структурованих і неструктурованих даних, які описують процеси, що мають місце у соціальній сфері. Нерозв’язаною залишається й низка задач розроблення універсальних методик побудови прогнозів соціально-економічних процесів, зокрема в умовах невизначеності.

Тому, важливою і *актуальною науково-прикладною проблемою*, що розв’язується в дисертаційній роботі є підвищення ефективності підтримки прийняття рішень у системі соціального захисту та соціального забезпечення в умовах невизначеності, під впливом внутрішніх та зовнішніх чинників.

Зв'язок роботи з науковими програмами, планами, темами

Тема дисертаційної роботи повністю відповідає тематиці наукових досліджень Інституту телекомунікацій і глобального інформаційного простору НАН України. Результати, отримані в ході виконання дисертаційного дослідження знайшли відображення у науково-дослідних роботах: «Розробка та аналіз засобів теоретико-ігрового моделювання стратегій збалансованого технологічного розвитку територій» (номер державної реєстрації 0116U000796), «Розробка інформаційної технології моделювання і прогнозування розвитку соціально-еколого-економічних систем в умовах невизначеності, нестаціонарності та ризику» (номер державної реєстрації 0121U100132), «Розробка інформаційних технологій та

інструментальних засобів моделювання і прогнозування розвитку територій в умовах децентралізації» (номер державної реєстрації 0121U109211).

Внесок автора у результати наведених науково-дослідних робіт полягає в розробці інформаційної технології збору та оброблення структурованих та неструктурованих даних для моделювання та прогнозування досліджуваних соціально-економічних процесів.

Мета і завдання дослідження

Мета дисертаційної роботи — розроблення інформаційної технології, основу якої складають методи, моделі, алгоритми, призначені для збору та обробки даних для прогнозування витрат на пенсії та соціальні допомоги, що підвищить ефективність рішень в управлінні соціальною сферою.

Для досягнення мети визначені такі завдання:

– дослідити сучасні підходи до збору даних у соціальній сфері та підготовки їх до подальшого використання з метою прогнозування витрат на пенсії та соціальні допомоги;

– сформулювати загальні вимоги до інформаційної системи збору та обробки даних для аналізу та прогнозування витрат на пенсії і соціальні допомоги;

– визначити методику побудови підсистеми збору та оброблення даних та її функціонування у складі Єдиної системи соціальної сфери України;

– розробити методику опрацювання текстової інформації щодо потреб населення у соціальному забезпеченні, представленої у зверненнях громадян та соціальних мережах;

– розробити методику моделювання контингенту одержувачів пенсій та допомог в умовах невизначеності, в тому числі, спричиненої військовим конфліктом;

– створити методику підготовки даних для побудови моделей та прогнозування витрат на пенсії та соціальні допомоги;

– розробити програмне забезпечення для збору, оброблення, аналізу, прогнозування та візуалізації результатів прогнозування витрат на пенсії та

соціальні допомоги на основі розроблених методів, виконати її апробацію та аналіз результатів.

Об'єкт дослідження – процеси збору та оброблення даних у соціальній сфері, пов’язані із фінансуванням витрат на пенсії та соціальні допомоги.

Предмет дослідження – методи, моделі та інформаційні технології збору та оброблення даних для визначення поточних та майбутніх витрат на пенсії та соціальні допомоги.

Методи дослідження. У роботи застосовано міждисциплінарний підхід, що поєднує методи системного аналізу, інтелектуального аналізу даних, математичного моделювання, машинного навчання. *Для дослідження систем пенсійного забезпечення, соціального захисту і соціального забезпечення в умовах невизначеності та для узагальнення теоретичних і методологічних засад збору даних* використано методи діалектичного пізнання, системного аналізу, індукції та дедукції, аналізу і синтезу, аналогій і зіставлення, формалізації й моделювання, дослідження часових рядів, кластерного та інтелектуального аналізу даних, сценарний підхід та елементи теорії прийняття рішень. *Для збору та попередньої обробки структурованих і неструктурованих даних* застосовано ETL-процеси, веб-скрейпінг, REST/OData API-запити, методи системного, інтелектуального й статистичного аналізу даних, експертне оцінювання, а також методи нормування, трансформації та обробки текстових даних. *Для побудови моделей і формування прогнозів* використано регресійний аналіз, методи аналізу часових рядів, ймовірнісно-статистичне та економіко-математичне моделювання, дерева рішень і багатовимірні розподіли. *Для розроблення методики застосування розроблених інформаційних технологій* – методи й засоби проєктування та розроблення програмного забезпечення для інформаційно-аналітичних систем, включно з прикладним і web-програмуванням.

Наукова новизна одержаних результатів

вперше:

озроблено методологію збору та оброблення даних, що описують процеси у соціальній сфері, яка відрізняється робастністю результатів, що дозволяє підвищити ефективність бюджетного процесу як Пенсійного фонду України, так і державного та місцевих бюджетів,

озроблено новий метод побудови моделей та ансамблів моделей для визначення контингенту одержувачів пенсій та соціальних допомог, який відрізняється модифікованою комплексною структурою та високою адекватністю, що дозволяє підвищити якість управлінських рішень, що приймаються у соціальній сфері,

творено інформаційну технологію прогнозування витрат на пенсії та соціальні допомоги, в основу якої покладено поєднання принципів системного аналізу, методів обробки структурованих та неструктурованих даних, оцінювання якості даних, яка забезпечує високу якість прогнозів;

удосконалено:

- підхід до збору даних з державних реєстрів, відкритих API, даних НУО, стрім-потоки та веб-скрапінг, який забезпечує повне покриття різнопланової інформації (платежі, демографія, гуманітарні гранти) та підвищує репрезентативність вибірок для моделювання,
- розроблено метод моделювання соціально-економічних процесів, що мають місце у соціальній сфері в умовах невизначеності даних, який відрізняється від відомих урахуванням різних типів невизначеностей, що підвищує адекватність моделей і якість оцінок прогнозів за лінійними та нелінійними моделями;

отримали подальший розвиток:

истемна методологія проектування і реалізації систем підтримки прийняття рішень, яка відрізняється комплексним підходом до розв’язання задач підготовки даних, оцінювання структури і параметрів математичних моделей досліджуваних процесів з використанням множин критеріїв якості, і забезпечує високу якість проміжних та остаточних результатів моделювання, прогнозування і прийняття управлінських рішень;

метод прогнозування, заснований на використанні адаптивного підходу до моделювання у поєднанні із статистичним та ймовірнісним моделюванням, що дає можливість урахувати структурно-параметричні невизначеності і забезпечує адекватний опис причинно-наслідкових зв'язків і можливих варіантів розвитку процесів різної природи під впливом груп внутрішніх та зовнішніх чинників.

Практичне значення отриманих результатів полягає в тому, що на основі запропонованих в роботі інформаційних технологій, моделей, методів, алгоритмів збору та оброблення даних розроблено методику прогнозування витрат на пенсії і соціальні допомоги та реалізоване відповідне програмне забезпечення. Запропонована методика та програмне забезпечення призначені для використання у Пенсійному фонді України у складі ІКІС ПФУ та у Єдиній інформаційній системі соціальної сфери України.

Використання у соціальній сфері запропонованих в роботі інформаційних технологій, забезпечить підвищення якості управлінських рішень щодо реалізації соціальних гарантій в умовах невизначеності, в тому числі, спричиненої військовим станом. Запропоновані в роботі алгоритми збору, очищення та уніфікації даних є універсальними, вони призначені для розв'язання широкого кола актуальних задач інтеграції даних різних типів, включаючи роботу з персональними даними громадян, державними реєстрами, відкритими даними, API-платформами, даними з соціальних мереж та зверненнями громадян, отримання нових відомостей про об'єкт дослідження та виявлення закономірностей розвитку досліджуваних процесів. Це розширить аналітичні можливості системи, підвищити якість аналітичної роботи у соціальній сфері, в тому числі, у процесі формування та моніторингу виконання державного та місцевих бюджетів, бюджету Пенсійного фонду України. Запропонована у роботі методологія опрацювання даних та подальшого їх використання у єдиному інтегрованому інформаційному просторі для прогнозування витрат на виплату пенсій та соціальних допомог закладає основи для розвитку нового підходу до створення інформаційних

систем, використовуваних у державному управлінні та місцевому самоврядуванні. Практична значущість результатів дослідження підтверджена довідками Головного управління Пенсійного фонду України у Львівській області, ТОВ «МЕДИРЕНТ».

Особистий внесок здобувача. Усі теоретичні та практичні результати, що виносяться на захист, отримані автором самостійно. Пошук та аналіз літературних джерел за тематикою дисертаційного дослідження виконано автором особисто. Наукові положення, висновки, рекомендації, що викладені в дисертації, розроблені особисто здобувачем. У працях, виконаних у співавторстві, автору дисертації належать: методика формування критеріальної бази та підготовки даних для аналізу, побудови моделей та прогнозування видатків на соціальний захист та соціальне забезпечення [1, 2, 6]; метод оброблення великих обсягів даних для побудови ймовірнісно-статистичних моделей та ансамблів моделей для прогнозування динаміки та структури контингенту отримувачів пенсій та соціальних допомог [2, 3]; архітектура системи підтримки прийняття рішень для аналізу та прогнозування витрат на пенсії та соціальні допомоги [4]; функціональна схема та програмне забезпечення системи підтримки прийняття рішень із використанням запропонованої інформаційної технології [4, 9]. системна методологія використання текстової інформації для прогнозування видатків на соціальний захист та соціальне забезпечення [5, 7, 8].

Апробація результатів дисертації. Результати дисертаційного дослідження були представлені на 3-х міжнародних та всеукраїнських науково-практичних конференціях: XXII Міжнародній науково-практичній конференції «Інформаційно-комунікаційні технології та сталий розвиток» (Київ, 14-15 листопада 2023 р.), XXIII Міжнародній науково-практичній конференції «Математичне моделювання та інформаційно-комунікаційні технології для зміцнення та відновлення» (Київ, 12-13 листопада 2024 р.), VIII Міжнародній науково-практичній конференції «Прикладні інформаційні системи та технології в цифровому суспільстві» (Київ, 01 жовтня 2024 р.).

Публікації. За результатами дослідження опубліковано 8 наукових праць, у тому числі:

- 3 статті у наукових періодичних виданнях, включених до Переліку наукових фахових видань України (в т.ч. 1 - включених до категорії «А»);

- 2 статті у наукових періодичних виданнях інших держав з напряму, з якого підготовлено дисертацію, що включені до міжнародних наукометричних баз SCOPUS та/або Web of Science Core Collection;

- 3 тези та доповіді на науково-практичних конференціях.

Одне авторське свідоцтво на комп'ютерну програму.

Якість та кількість публікацій відповідають «Вимогам до опублікування результатів дисертацій на здобуття наукових ступенів доктора і кандидата наук».

Структура та обсяг дисертації. Дисертація складається з вступу, чотирьох розділів, висновків та додатків. Вона містить 167 сторінок машинописного тексту, включаючи 30 рисунків, 34 таблиці, список використаних джерел з 152 найменувань.

РОЗДІЛ 1
ОСОБЛИВОСТІ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ, ВИКОРИСТОВУВАНИХ У СИСТЕМІ СОЦІАЛЬНОГО ЗАХИСТУ ТА СОЦІАЛЬНОГО ЗАБЕЗПЕЧЕННЯ УКРАЇНИ

1.1 Особливості використання інформаційних систем та технології у державному управлінні та публічному урядуванні

Інформаційно-аналітичні системи, комп'ютерні системи підтримки прийняття рішень, розроблені на базі сучасних інформаційні технології, що використовують апарат математичного моделювання та штучний інтелект, активно застосовуються у державному управлінні та публічному урядуванні як за кордоном, так і в Україні. Їх застосовують у системі макроекономічного аналізу та прогнозування, національній безпеці, сфері надання адміністративних послуг населенню, соціальному забезпеченні, тощо.

Це, зокрема, модель EUR-ACE (агент-орієнтована модель європейської економіки) [10], FuturICT 2.0 [11], УКРМАКРО [12[, комп'ютерні модельні комплекси у державному управлінні [12] OPTIMIZ; SimRunner2, WITNESS Optimizer [14], та інші, програмне забезпечення для планування та прогнозування макроекономічних показників – додатки на платформі Inforum [15], проект LINK [16], тощо.

Прикладом вітчизняного досвіду впровадження інформаційно-аналітичних систем, які працюють з Big Data у систему державного управління, є інформаційно-аналітична система (ІАС) «СОТА» [17], яка функціонує у складі інформаційної системи Головного ситуаційного центру України. ІАС «СОТА» забезпечує зберігання, поєднання та аналіз даних з різних джерел з понад 20 напрямів. Її використання сприяє підвищенню достовірності отриманої інформації, ефективності моніторингу стану національної безпеки, координації взаємодії органів державного управління. ІАС «СОТА» є складною багатошаровою інформаційно-аналітичною

системою найвищого рівня захисту інформації, має гнучку, відкриту архітектуру, що дозволяє створювати нові функціональні модулі відповідно до завдань, які виникають під час реалізації державної політики у сфері національної безпеки. Однак, аналітична складова системи потребує доопрацювання з точки зору розширення кола використовуваних моделей, зокрема, математичних, впровадження методів інтелектуальної обробки вхідних даних та побудови прогнозів.

Інша система, використовувана для забезпечення електронної взаємодії державних електронних інформаційних реєстрів - «Трембіта» [18] створена на платформі обміну даними «X-road» (UXP платформа), наданій Україні Естонією в рамках проекту «EGOV4UKRAINE» програми «U-LEAD». Система є провідною ланкою у системі надання електронних послуг населенню. Серед інших переваг слід відзначити, що захист інформації в системі відповідає UXP Security Server версії 1.12.6.

В рамках цифровізації урядування працює Державний реєстр прав, адміністрування якого здійснює Державне підприємство «Національні інформаційні системи». Підприємство забезпечує опрацювання даних з відкритих реєстрів та систем міністерств та відомств. Наприклад, Міністерство юстиції України та Державна податкова служба України, які мають свої інформаційні системи, здійснюють електронну взаємодію інформаційних систем у «робочому порядку», але електронна взаємодія стосовно роботи з відомостями про фізичну особу додатково врегульовується чинним законодавством.

З Державним реєстром інтегровані й інформаційні системи, що функціонують у системі соціального захисту та соціального забезпечення населення. Система соціального захисту та соціального забезпечення донедавна мала декілька автоматизованих інформаційних систем, які були у розпорядженні Пенсійного фонду України, Міністерства соціальної політики, державних цільових фондів. Між ними була встановлена електронна взаємодія, але цілісної системи не існувало.

Автоматизація системи соціального захисту, яка почалась ще у кінці 90х розробленням автоматизованих систем для Пенсійного фонду України та Міністерства праці та соціального захисту населення, сьогодні охопила практично всі ключові процеси соціального захисту та соціального забезпечення [19]. Серед інформаційних систем соціальної сфері України найбільш потужними є інформаційна система Пенсійного фонду України та Міністерства соціальної політики України.

Розроблена фахівцями ДП «Інформаційно-обчислювальний центр Міністерства соціальної політики України» [21], інформаційна система, має низку сервісів для забезпечення доступу громадян до соціальних послуг та автоматизації робіт з обслуговування населення в структурних підрозділах з питань соціального захисту населення місцевих державних адміністрацій та організації діяльності територіальної громади у сферах соціального захисту населення та захисту прав дітей. Її аналітична складова орієнтована на формування статистичної звітності та верифікацію державних виплат.

Інтегрована комплексна інформаційна система Пенсійного фонду України (ІКІС ПФУ) [20, 21] забезпечена сучасною ІТ-інфраструктурою, комплексною системою захисту інформації є чи не найпотужнішою з усіх систем, використовуваних у соціальній сфері. В ІКС ПФУ здійснюється автоматизована обробка, передача, зберігання та використання інформації, розпорядником якої, відповідно до Закону України «Про загальнообов'язкове державне пенсійне страхування» та іншої нормативно-правової бази є Пенсійний фонд України. Система інтегрована до інформаційного обміну з іншими відомствами (державні фонди соціального страхування, Державна фіскальна служба, реєстри Міністерства юстиції тощо), підприємствами та установами, підприємствами. Ці функції реалізовані у «Підсистемі електронного обміну даними» (ПЕОД), яка забезпечує автоматизовану обробку та обмін даними між різними інформаційними системами, що використовуються в рамках ПФУ. Основною її функцією є інтеграція і передача даних у реальному часі або за запитом, що оптимізує процеси,

пов'язані з обміном інформацією між підсистемами ПФУ, державними установами та іншими організаціями.

ІКІС ПФУ вирізняється наявністю потужної бази даних (накопичується з 1998 року), обсяг якої становить більше 250 Тб. Кількість користувачів – майже 19 тисяч, кількість офісів ПФУ, в яких функціонує ІКІС - більше 600, кількість користувачів Веб-порталу - 10 млн, кількість користувачів мобільного додатку - 680 тисяч, застрахованих осіб - 43 млн. облікових карток, пенсіонерів - 10,8 млн. осіб, яким виплачується пенсія, кількість страхувальників - загалом 3 млн. облікових карток [19-21].

Саме тому, ІКІС ПФУ було обрано [19-21] для створення на її основі єдиного інформаційного простору соціальної сфери України, розбудова якого почалась у 2020 році. Замість розрізнених реєстрів та баз даних впроваджується цілісна інформаційна система з єдиною програмно-технічною архітектурою, здатна ефективно адаптуватись до нових задач та викликів, що постають перед країною. Її складовими, як зазначено у [22], є централізована база даних; загальна та прикладні підсистеми. Вже розроблене та впроваджене функціональне ядро системи на базі якого здійснюється автоматизація прикладних задач, пов’язаних із соціальним захистом та соціальним забезпечення населення [20-22]. Передбачено реалізацію низки організаційних, технічних, технологічних заходів, спрямованих на формування єдиної бази даних, розроблення загального інтерфейсу, налагодження системи обміну інформацією між її компонентами, створення єдиної системи кіберзахисту, тощо [20-23].

На сьогоднішній день до ІКІС Пенсійного фонду України повністю інтегровано систему призначення та виплати житлових субсидій «Наш дім», яка ефективно працює у складі загальної системи [24].

Однак, аналітичному забезпеченню прийняття рішень в управлінні соціальною сферою поки приділено мало уваги. Зокрема, аналітична складова ІКІС ПФУ реалізована на платформі Oracle BI, використовує аналітичне сховище даних (DWH – Data Ware House) ПФУ [21-25]. Система аналізу та

прогнозування, що використовується у ІКІС ПФУ розроблена компанією «Медірент» [21], має структуру, представлену на рис.2.2.

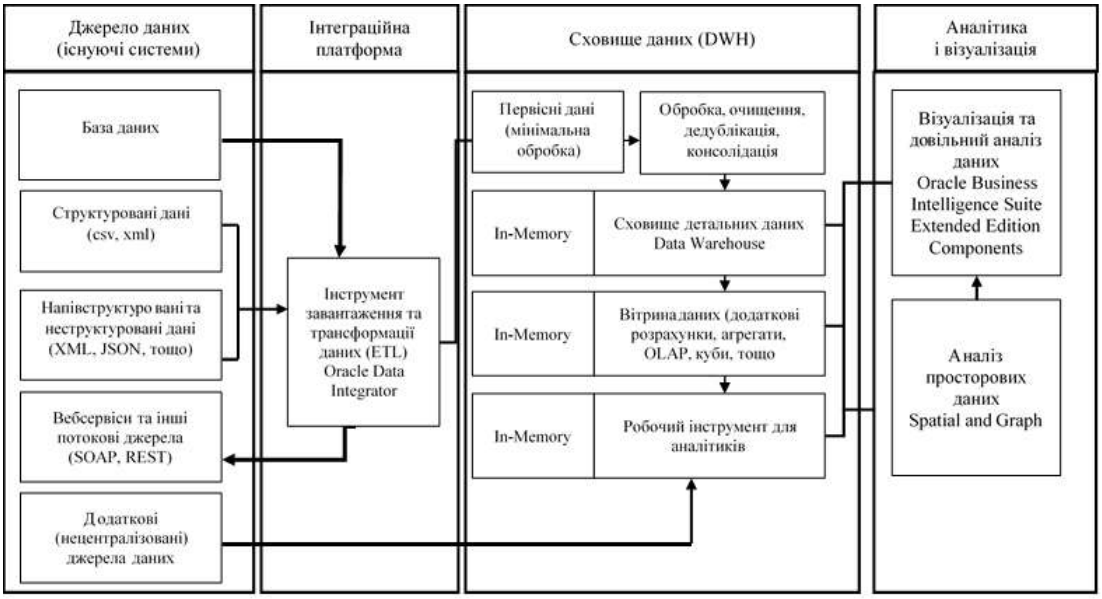

Рисунок 2.2 – Схема АРМ Керівника Пенсійного фонду [3]

Однак, не зважаючи на значну кількість сервісів, вона забезпечує лише поточну аналітичну роботу в рамках автоматизованого робочого місця керівника і зосереджена на опрацюванні інформації, накопиченої у базі даних ІКІС ПФУ в рамках формування статистичної звітності, формування звітів про потребу в коштах та планування бюджету Пенсійного фонду України на наступний рік із застосуванням коригуючих коефіцієнтів.

Впровадження Єдиної інформаційної системи соціальної сфери є потужним кроком в напрямку реалізації Стратегії цифрової трансформації соціальної сфери [22, 23], викликаним сучасними тенденціями, характерними для соціальної сфери держави. Швидко зростає кількість осіб, які потребують підтримки держави, погіршується демографічна та економічна ситуація, внаслідок війни зростає потреба в наданні соціальних послуг, розширюється набір соціальних виплат, тощо. Крім того, актуальним є питання захисту персональних даних застрахованих осіб, одержувачів пенсій та допомог від несанкціонованого доступу, забезпечення їх актуальності та достовірності.

Тому, на заміну розрізненим, іноді, морально застарілим інформаційним система різного підпорядкування створюється цілісна система, яка разом із Порталами електронних послуг Міністерства соціальної політики України та інших органів державного управління та місцевого самоврядування удосконалить процеси надання соціальних послуг, призначення та виплату пенсій та допомог, забезпечить оперативне та адаптивне реагування на ті виклики, що стають все більш актуальними в умовах повномасштабної війни. Як пропонується у законопроєкті №11377 [26], основними принципами розбудови Єдиної інформаційної системи соціальної сфери є такі як:

- стандартизації даних;
- прозорості та підзвітності;
- покращення якості соціальних послуг;
- адаптації до європейських стандартів
- -моніторингу та оціни;
- співпраці з міжнародними партнерами.

Тому, проблема об'єднання даних, розпорядниками яких є різні органи державного управління, місцевого самоврядування, Пенсійний фонд України, інші цільові фонди, міжнародні організації із наступною їх стандартизацією без втрат і спотворень, є однією з нагальних задач реформування інформаційного забезпечення соціальної сфери. Вирішення цієї проблеми потребує розроблення нових підходів до збору даних, їх накопичення та підготовки до використання для підтримки прийняття рішень стосовно надання пільг, допомог, соціальних послуг, призначення та виплати пенсій.

Іншою, не менш важливою задачею, пов'язаною з підготовкою даних є розроблення методики їх використання у аналітичній роботі органів державного управління різних рівнів та місцевого самоврядування з метою визначення потреби у видатках на соціальних захист та соціальне забезпечення, джерелах їх покриття на поточний період, прогнозування їх структури та динаміки на перспективу. Адже, ситуація тривалої повномасштабної війни ставить нові виклики перед соціальною сферою,

досвіду подолання яких ні в Україні, ні в європейському просторі немає. Змінюється демографічна ситуація та ситуація на ринку праці, зростає кількість ветеранів, осіб з інвалідністю, внутрішньопереміщених осіб, з'являються цільові групи населення, які потребують нових видів допомог та спеціалізованих соціальних послуг. Все це призводить до зростання навантаження на бюджети різних рівнів – збільшуються видатки на соціальний захист та соціальне забезпечення.

Тому, поряд із вирішенням поточних питань, соціальна сфера потребує всебічного аналізу ситуації, бачення перспектив розвитку. Ситуація невизначеності, спричинена війною, створює нові виклики, аналогів яким європейська практика організації соціального захисту та соціального забезпечення не має. Актуальними є питання не лише оперативного моніторингу та контролю, аналізу та формування звітності, а й передбачення розвитку подій, окреслення перспектив, впровадження сучасних методик проведення актуарних розрахунків, прогнозування демографічних показників, потреби населення у соціальному захисті та соціальному забезпеченні, видатків та їх фінансування. Тому, питання розроблення гнучкої спеціалізованої аналітичної платформи, що поєднує новітні засоби обробки Big Data, математичні моделі, інструменти інтелектуального аналізу даних та методи штучного інтелекту є актуальним.

1.2 Особливості збору та обробки даних, використовуваних у соціальній сфері

У системі соціального забезпечення та соціального захисту населення вирішується декілька груп взаємопов'язаних задач, які потребують збору та оброблення даних. Перша група задач пов'язана з оброблення персональних даних застрахованих осіб та одержувачів пенсій у системі державного пенсійного забезпечення, інших державних цільових фондів, одержувачів субсидій, утримань, допомог та соціальних послуг. Друга група задач

передбачає використання консолідованих даних у процесі аналізу та прогнозування доходів та витрат, надходжень та видатків зведеного бюджету, державного бюджету та місцевих бюджетів, а також державних цільових фондів, зокрема, Пенсійного фонду України.

ІКС ПФУ обрана базовою платформою для створення Єдиної інформаційної системи соціальної сфери, організована за ієрархічним принципом та складається із взаємно-пов'язаних технологічних рівнів, які являють собою сукупність локальних обчислювальних мереж управлінь і відділень Пенсійного фонду України, що об'єднані в суцільне територіально розподілене комунікаційне середовище за допомогою мережі оператора передачі даних. Аналогічну структуру мають і установи, що підпорядковуються Міністерству соціальної політики [21-26].

Єдина інформаційна система соціальної сфери проєктується таким чином, що, розміщення технологічних рівнів системи повторює організаційно-територіальну структуру Пенсійного фонду України та структурні підрозділи з питань соціального захисту населення в місцевих адміністраціях та виконання повноважень щодо соціального захисту населення та захисту прав дітей органами місцевого самоврядування [21-26]:

- центральний рівень (центральний апарат ПФУ/Міністерство, модулі основного та резервного Центрів обробки даних, виділені компоненти модулів);

- обласний рівень (обласні вузли в обласних центрах);

- районний рівень (районні вузли в областях України та у місті Києві);

- віддалені робочі місця фахівців (робочі місця в центрах надання адміністративних послуг, об'єднаних територіальних громад (включаючи систему керування електронною чергою)).

Центральний рівень системи включає компоненти обласного та районного рівнів, які мають окремі комплексні системи захисту, робочі станції користувачів з певним набором прав доступу та активне мережеве обладнання. В результаті цього центральний апарат забезпечує керування підсистемами

інформаційної системи та доступ авторизованих користувачів до модулів Системи та складається з активного мережевого обладнання та робочих станцій, що забезпечують функціонування Системи.

Вузли Системи обласного та районного рівнів, що забезпечують доступ авторизованих користувачів до модулів Системи з використанням активного мережевого обладнання та робочих станцій. Вузли обласного та районного рівнів розташовуються у будівлях, приміщеннях будівель управлінь, що територіально розташовані і різних містах України.

Віддалені робочі місця фахівців, що забезпечують віддалений доступ авторизованих користувачів до модулів Системи організовані у вигляді віддалених АРМ і розташовуються у відповідних приміщеннях центрів надання адміністративних послуг, офісах територіальних громад, тощо [21].

Взаємодія вузлів районного та обласного рівня здійснюється через мережу оператора передачі даних у зашифрованому вигляді з використанням VPN-тунелю [21-26].

Поєднання системи електронної взаємодії державних електронних інформаційних ресурсів («Трембіта»), зовнішніх прикладних сервісів, персональних комп’ютерів зовнішніх користувачів сервісів Системи соціального захисту населення з модулем ОЦОД реалізована за допомогою модулю вебпорталу електронних послуг, що має окрему комплексну систему захисту інформації.

Вирізняють такі технологічні рівні Системи: центральний рівень, вузли обласного та районного рівнів, які складаються з активного мережевого обладнання (міжмережеві екрани, комутатори та маршрутизатори) та робочих станцій (адміністраторів та внутрішніх користувачів) [21-26].

Автоматизовані робочі місця фахівців використовуються адміністраторами та внутрішніми користувачами у межах ролей, що призначені їм в Системі та функціональних завдань відповідно до посадових обов’язків. Передбачено, що автоматизовані робочі місця фахівців мають

знаходяться в приміщеннях установ, але, за потреби, користувачам можуть бути дозвіл та засоби для забезпечення віддаленого доступу.

У системі є джерела безперебійного живлення, призначені для забезпечення постачання електричною енергією компонентів Системи в межах норми (у випадках стрибків напруги або повного відключення електроенергії).

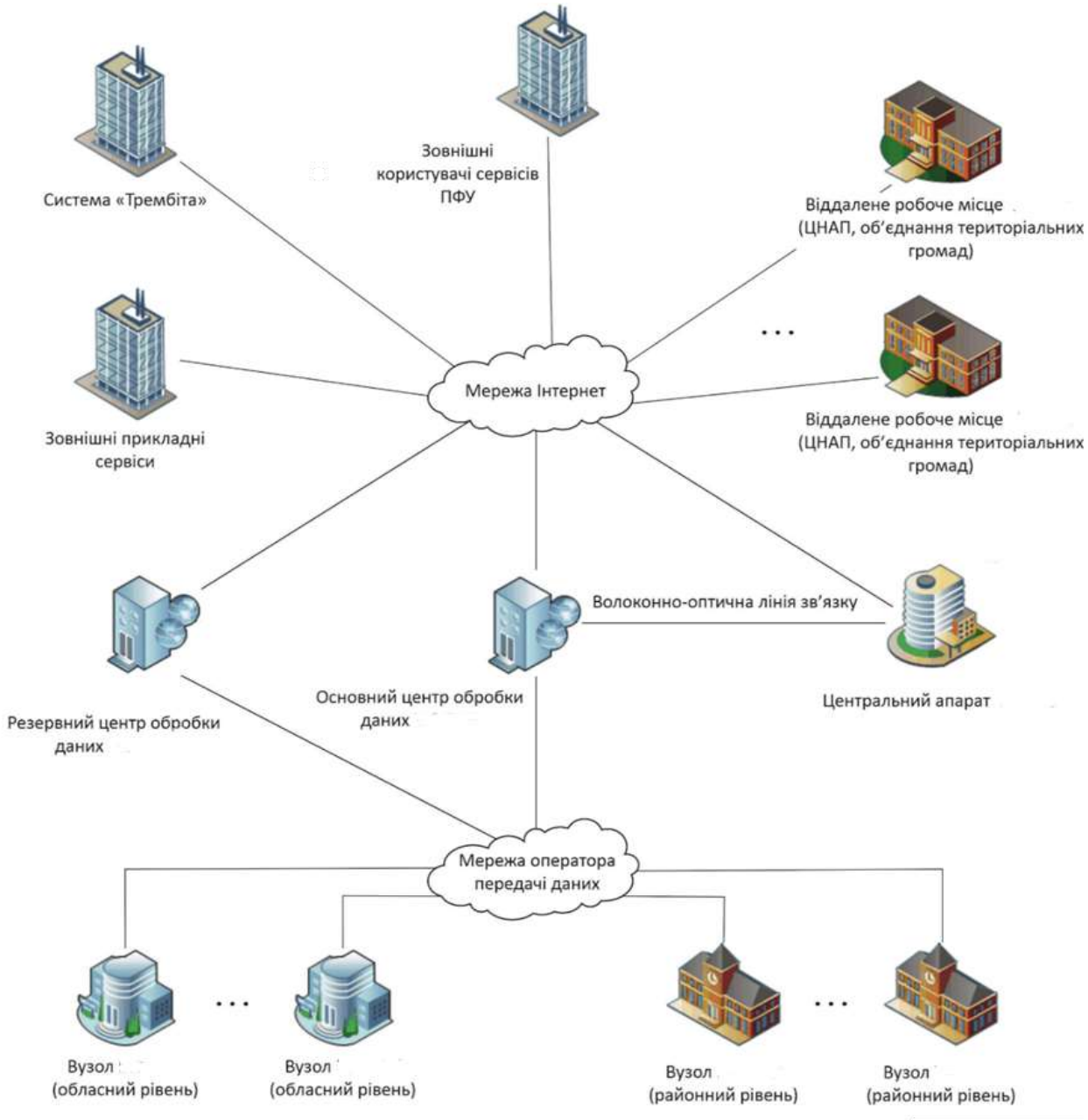

Рисунок 1.1 – Глобальна архітектура Системи [21]

В рамках взаємодії використовуються засоби захисту інформації вебпорталу електронних послуг, як модулю Систему. Мережеве обладнання

призначено для захисту компонентів Системи від мережевих атак з боку зловмисників та шкідливого програмного забезпечння.

Централізована система керування територіально розподіленою корпоративною мережею забезпечує: постійний контроль та підтримку працездатності і необхідної функціональності всіх вузлів доступу до корпоративної мережі та глобальної мережі Інтернет, дистанційне конфігурування обладнання вузлів у разі виникнення збоїв в роботі системи, її модернізації або змінах параметрів мережі.

Всі дані підсистем Єдиної системи соціальної сфери, що обробляються та зберігаються, знаходяться в основному та резервному центрі обробки даних центрального технологічного рівня. Центри обробки даних обласного та районного рівнів призначені для автоматизації обробки, передачі, зберігання та використання інформації, розпорядниками якої відповідно до чинного законодавства є установи системи соціального захисту та соціального забезпечення.

Єдина система соціальної сфери являє собою багатокористувацьку мережу, побудовану за змішаною (гібридною) мережевою топологією, з доступом до мережі Інтернет [21-26].

На даний момент, в ІКС ПФУ реалізовано інтерактивні аналітичні інструменти, необхідні для ключових функціональних областей діяльності, пов’язаних з формуванням звітності та плануванням бюджету на наступний рік [21-26]. Однак, виходячи з того, що система соціального захисту та соціального забезпечення функціонує в умовах військового конфлікту, коли установи районного та обласного рівня евакуюються, але мають продовжувати роботу, пенсії та допомоги мають нараховуватись та виплачуватись вчасно, і має бути перевірено право на отримання соціальної допомоги чи послуги. Важливу роль відіграє робота із зверненнями громадян, адже зростає кількість звернень як за роз’ясненнями, так і за призначенням пенсій та допомог. Але робота з текстом звернень в даному варіанті Єдиної системи соціальної сфери не реалізована.

Крім того, потребує удосконалення процедура збору даних, необхідних для прогнозування майбутніх витрат на соціальний захист та соціальне забезпечення, адже існуючі методики в умовах війни та нестачі бюджетних коштів можуть застосовуватись досить обмежено. Крім того, триває реформування системи пенсійного забезпечення: крім солідарної частини буде запроваджено накопичувальну частину, що суттєво змінить підходи до пенсійного забезпечення населення України.

Іншим важливим питанням є підготовка даних до спільного використання в Єдиній інформаційній системі соціальної сфери України. Це низка етапів попередньої обробки даних, яка передбачає, зокрема, очищення від дублікатів і помилок введення, стандартизацію показників відповідно до чинних нормативних документів та адаптацію до європейських стандартів, якщо це доцільно, заповнення пропущених значень та зберігання у структурованому вигляді (реляційна база даних та/або розподілене сховище Big Data/NoSQL) [21-26].

Окреме місце у системі посідає робота з зображеннями. Адже накопичується архів сканованих пенсійних справ, до бази звернень громадян долучаються не лише електронні листи, що містять текст звернення, а й скановані паперові документи, електронні копії документів, фотографії, тощо, де зберігаються звернення, отримані в електронному вигляді, які містять скановані документи. Питання, скарги, пропозиції мають бути розглянуті у визначений законодавством термін. За необхідності, виконання звернення має бути проконтрольовано.

Питання збору, оброблення, зберігання структурованих і неструктурованих даних та використання їх у аналітичній діяльності у процесі підтримки прийняття рішень у соціальній сфері є особливо актуальним. Цей етап є важливою ланкою створення Єдиної інформаційної системи соціальної сфери. Він включає заходи як технічного, технологічного, так і організаційного характеру. Дані мають зберігатись, згідно чинного законодавства, перебувати у відкритому доступі, так і у службовому

користуванні. При чому доступ до них має бути розмежований залежно від повноважень фахівця. У межах дослідження були використані виключно відкриті джерела даних.

В результаті чіткого виконання заходів етапу збору даних, буде сформовано базу даних, призначену для спільного використання у Єдиній інформаційній системі соціальної сфери, ка, в подальшому, може бути використана для аналізу, комп'ютерного моделювання та прогнозування.

Дані, які мають використовуватись у єдиної інформаційної платформи надходитимуть з державних реєстрів, Міністерства соціальної політики, Міністерства фінансів, Державної казначейської служби, Пенсійного фонду України, органів місцевого самоврядування, Порталу відкритих даних, платформи Openbudget, неурядових і міжнародних організацій. Вони мають різні формати представлення та зберігання даних, тому необхідно забезпечити реалізацію комплексного підходу до формування вхідного масиву даних [21].

Передбачено, що обробка і зберігання інформації здійснюється у вигляді файлів різних форматів та об'єктів бази даних. За функціональним призначенням та за змістом вимог до захисту інформації, вона поділяється на:

– відкриту інформацію;
– інформацію з обмеженим доступом.

Класифікація інформаційних ресурсів за режимом доступу наведена на рис. 1.2.

Рисунок 1.2 - Класифікація інформаційних ресурсів ІКС ПФУ за режимом доступу

Відповідно до функціонального призначення, обробки та зберігання інформації, вона може бути віднесена до наступних видів:

– відкрита інформація, що належить до державних інформаційних ресурсів;

– відкрита інформація інших установ і організацій, яка потребує захисту;

– інформація з обмеженим доступом:

→ конфіденційна інформація, що містить персональні дані;

→ конфіденційна інформація, що містить персональні дані, які містяться у державних інформаційних ресурсах;

→ технологічна інформація.

До відкритої інформації, що належить до державних інформаційних ресурсів відноситься:

– нормативно-правова інформація;

– внутрішній документообіг установ системи соціального захисту та соціального забезпечення;

– дані планування та контролю установ соціального захисту та соціального забезпечення.

Залежно від місця обробки, інформація представляється у вигляді:

– інформаційних об'єктів мережевого, канального та фізичного рівнів моделі взаємодії відкритих систем OSI (відповідно, у вигляді IP-пакетів, мережевих кадрів (фреймів) та фізичних сигналів);

– файлів серверів баз даних або іншого серверного обладнання виділених компонентів модулів системи різних рівнів;

– записів бази даних виділених компонентів модулів.

Модифікувати або знищувати відкриту інформацію можуть лише ідентифіковані та автентифіковані користувачі, яким надано відповідні повноваження. Спроби модифікації чи знищення відкритої інформації користувачами, які не мають на це повноважень, блокуються.

Відкрита інформація зберігається у вигляді постійних або тимчасових структур у пам'яті серверів підсистем Системи.

До інформації цієї категорії висуваються підвищені вимоги із збереження цілісності та забезпечення її доступності.

До відкритої інформації інших установ та організацій належить вся інша відкрита інформація, яка потребує розмежування доступу.

Цей вид інформації залежно від місця обробки, у Системі представлений у вигляді:

– інформаційних об'єктів мережевого, канального та фізичного рівнів моделі взаємодії відкритих систем OSI (відповідно, у вигляді IP-пакетів, мережевих кадрів (фреймів) та фізичних сигналів);

– файлів серверів баз даних або іншого серверного обладнання виділених компонентів модулів Системи;

– записів бази даних виділених компонентів модулів Системи різних рівнів ієрархії.

Модифікувати або знищувати таку відкриту інформацію можуть лише ідентифіковані та автентифіковані користувачі, яким надано відповідні повноваження. Спроби модифікації чи знищення відкритої інформації користувачами, які не мають на це повноважень, блокуються.

Відкрита інформація зберігається у вигляді постійних або тимчасових структур у пам'яті серверів підсистем Системи.

До інформації цієї категорії висуваються підвищені вимоги із збереження цілісності та забезпечення її доступності.

До конфіденційної інформації, що містить персональні дані належать:

– персональні дані учасників підсистем пенсійного забезпечення;

– персональні дані учасників підсистем соціального страхування;

– персональні дані учасників підсистем соціальних виплат;

– персональні дані користувачів АСЕДО;

– персональні дані, що містяться у відомостях ЕРЛН;

– та інші дані, що містяться в функціональних підсистемах Системи.

Конфіденційна інформація (персональні дані) в залежності від місця її оброблення, зберігається у вигляді:

– інформаційних об'єктів мережевого, канального та фізичного рівнів моделі взаємодії відкритих систем OSI (відповідно, у вигляді IP пакетів, мережевих кадрів (фреймів) та фізичних сигналів);

– файлів серверів бази даних або іншого серверного обладнання виділених компонентів модулів системи різних рівнів ієрархії;

– записів баз даних виділених компонентів модулів системи.

В Системі забезпечується захист цих даних в частині захисту від несанкціонованих дій. При цьому, реєстрація спроб несанкціонованих дій з конфіденційною інформацією супроводжується повідомленням про них адміністратора безпеки.

До цієї інформації висуваються підвищені вимоги із забезпечення конфіденційності, цілісності та доступності.

До Інформації з обмеженим доступом, а саме конфіденційної інформації (персональні дані), що міститься в державних інформаційних ресурсах належать персональні дані фізичних осіб, що можуть міститись в державних інформаційних ресурсах;

Конфіденційна інформація (персональні дані) в залежності від місця її оброблення зберігається у вигляді:

– інформаційних об'єктів мережевого, канального та фізичного рівнів моделі взаємодії відкритих систем OSI (відповідно, у вигляді IP пакетів, мережевих кадрів (фреймів) та фізичних сигналів);

– файлів серверів баз даних або іншого серверного обладнання виділених компонентів модулів системи на різних півнях її ієрархії;

– записів баз даних виділених компонентів модулів Системи.

Захист цих даних забезпечується в частині захисту від несанкціонованих дій. При цьому, реєстрація спроб несанкціонованих дій з конфіденційною інформацією супроводжується повідомленням про них адміністратора системи безпеки.

До цієї інформації висуваються підвищені вимоги із забезпечення конфіденційності, цілісності та доступності.

До Інформації з обмеженим доступом (технологічної інформації) відносяться:

– персональні ідентифікатори;

– паролі;

– робочі параметри засобів захисту;

– конфігурації програмного та апаратного забезпечення Системи;

– конфігурації активного мережевого обладнання;

– інформація журналів аудиту;

– тощо.

Технологічна інформація призначена для використання тільки уповноваженими користувачами з числа адміністраторів.

До технологічної інформації відноситься:

– конфігураційні та системні об'єкти, які відповідно визначають параметри конфігурації обладнання та специфічні дані, що використовуються обладнанням в процесі його функціонування;

– інформаційні об’єкти, що містять атрибути доступу користувачів і адміністраторів та визначають їх права в системі;

– інформаційні об’єкти, що містять виконуваний код програмного забезпечення;

– інформаційні об’єкти, що містять дані, які визначають порядок функціонування програмно-технічних засобів (конфігураційні настроювання):

⇥ налаштування операційної системи та програмного забезпечення, правил розмежування доступу;

⇥ налаштування і конфігурації мережевого та комунікаційного обладнання;

⇥ налаштування засобів колективного захисту інформації;

⇥ налаштування взаємодії між підсистемами Системи;

⇥ робочі параметри засобів захисту;

– інформаційні об’єкти, що містять звітні дані щодо функціонування програмно-технічного комплексу та дій користувачів (журнали, лог-файли):

⇥ журнали подій та налаштування щодо фіксації подій у журналах;

⇥ інформація журналів реєстрації та аудиту тощо.

Технологічна інформація зберігається у вигляді постійних або тимчасових структур у пам'яті пристроїв, що входять до складу ІКС ПФУ.

Ця інформація підлягає захисту в частині збереження конфіденційності, цілісності та забезпечення її доступності.

Обробку інформації в Системі організовано на двох технологічних рівнях.

На першому технологічному рівні (центральний апарат (ЦА), основний центр обробки даних, резервний центр обробки даних, ВП, КЦ та МЦ) розміщено основні компоненти управління та збереження даних підсистем, а також розмежовується доступ до них авторизованих користувачів (адміністраторів і операторів) Системи першого та другого технологічних рівнів [21].

На першому технологічному рівні забезпечується:

– інформаційний зв’язок між Системою та зовнішніми ІКС;

– інформаційний зв’язок між підсистемами першого і другого технологічних рівнів (оброблення запитів авторизованих користувачів (оператори) та віддалених робочих місць);

– інтеграція та оброблення інформації, здійснення функцій контролю за дотриманням вимог до повноти, достовірності, актуальності та захисту інформації відповідно до законодавства України;

– керування прикладними підсистемами.

На другому технологічному рівні (обласний та районний) забезпечується внесення інформації авторизованими користувачами (оператори) Системи до підсистем першого технологічного рівня, з використанням персональних комп’ютерів та віддалених робочих місць, у межах визначеної політики розмежування доступу.

Авторизовані користувачі (оператори) РС та віддалених робочих місць по Україні в рамках сервісів інтерфейсів прикладних підсистем можуть формувати запити до бази даних та отримувати відомості за цими запитами (відомості, звіти, виписки тощо). Оброблення цих запитів здійснюється засобами програмно-апаратних засобів першого технологічного рівня. Передача відомостей за цими запитами здійснюється через мережу оператора передачі даних у зашифрованому вигляді з використанням VPN-тунелю.

У процесі функціонування, Система взаємодіє з інформаційними системами інших підприємств, установ, організацій системою електронної взаємодії державних електронних інформаційних ресурсів «Трембіта», віддаленими робочими місцями, що розташовані у Центрах надання адміністративних послуг, офісах територіальних громад, через модуль зовнішнього доступу веб-порталу Системи.

Передача відомостей у процесі обміну здійснюється через мережу інтернет. Підключення до мережі інтернет має здійснюватися з використанням захищених вузлів інтернет-доступу, які мають КСЗІ із підтвердженою відповідністю.

В процесі виконання своїх задач Система взаємодіє з організаціями/установами та громадянами, що утворюють множину зовнішніх по відношенню до неї об’єктів (учасники системи пенсійного забезпечення та соціального страхування).

Передача технологічної інформації (дані ідентифікації та автентифікації) до основного центру обробки даних здійснюється від усіх компонентів Системи.

Взаємодія Системи з іншими підприємствами, установами, організаціями та громадянами здійснюється відповідно до вимог чинного законодавства України та/або встановленого у спільних угодах порядку.

На рис. 1.3 представлено пропоновану методику збору даних та їх підготовки до подальшого використання у Єдиній інформаційній системі соціальної сфери**.**

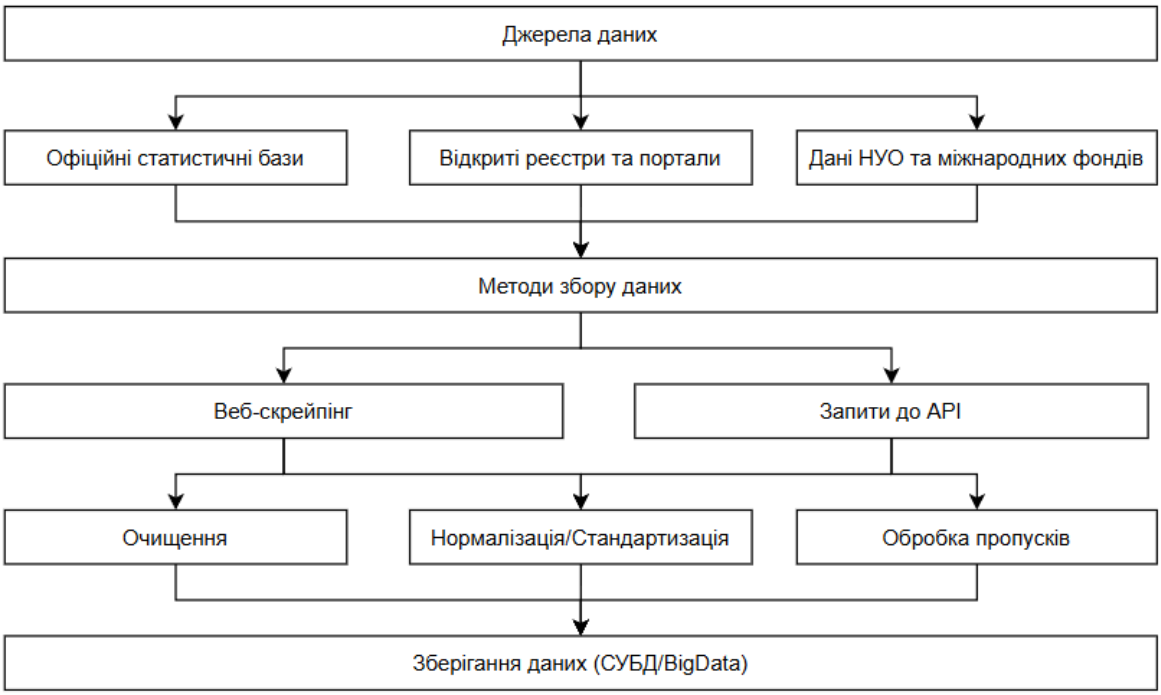

Рисунок 1.3 – Методика збору та попередньої обробки даних, накопичених у різних реєстрах та базах даних

Представлена на рис. 1.3 методика, призначена для формування наборів даних, які пропонується використовувати для аналітичного забезпечення підтримки прийняття рішень у системи соціального захисту та соціального забезпечення населення України.

1.3 Застосування системного підходу у задачах моделювання та прогнозування розвитку соціальної сфери в умовах невизначеностей та ризиків

У формалізації задач моделювання та прогнозування соціально-економічних та фінансових показників за різних моделей управління ключовим є перехід від якісного опису процесів до чітко визначених математичних структур. Це дозволяє одержати моделі, спроможні враховувати невизначеність та ризики, порівнювати альтернативні стратегії й обирати оптимальні рішення, що особливо актуально в умовах війни.

Визначення компонентів моделі

– Вектор стану *x(t)* описує поточний стан системи (макропоказники: ВВП, інфляція, доходи та витратизведеного бюджету, зайнятість населення, демографічна ситуація тощо; галузеві обсяги виробництва; регіональні диспропорції).

– Вектор управління u(t) містить рішення керівника: рівень державних інвестицій, облікова ставка НБУ, податкові стимули, обсяги надходженя до бюджетів всіх рівнів, запити на бюджетне фінансування.

– Вектор невизначеностей w(t) моделює зовнішні шоки: коливання цін на світовому ринку, ситуація військового конфлікту, геополітичні ризики, пандемії, фінансова длрлмога міжнародних організацій.

Перед побудовою моделей важливо чітко класифікувати невизначеності та ризики, щоб підібрати відповідні методи їхнього врахування:

1. Стохастичні параметри із відомим розподілом – застосовується стохастичне програмування.

2. Інтервальні невизначеності без точного розподілу, але з відомими границями – використовують інтервальний аналіз і методи надійної оптимізації.

3. Фазові/сценарні ризики (повномасштабна війна, глобальні кризи) – формалізуються через скінченний набір сценаріїв розвитку.

Ситуація невизначеності, спричинена військовим конфліктом, вимагає чіткого математичного формалізму.

Основні, типові варіанти постановки задач управління соціально-економічними та фінансовими процесами:

1. Динамічна модель управління. Описує еволюцію стану системи під впливом управлінських рішень і зовнішніх шоків [27-31].

$$x(t+1) = f\big(x(t),\, u(t),\, w(t)\big), \quad x(0) = x_0.$$

Мета полягає у тому, щоб знайти послідовність управлінських впливів $\{u(t)\}_{t=0}^{T}$, що максимізує функцію корисності або мінімізує збитки

$$\max_{u(\cdot)} \mathbb{E}\Big[\sum_{t=0}^{T} R\big(x(t),u(t)\big)\Big] \quad \text{або} \quad \min_{u(\cdot)} \max_{w(\cdot)} \sum_{t=0}^{T} C\big(x(t),u(t),w(t)\big).$$

2. Сценарне (дворівневе) програмування. Вибір стратегії з урахуванням найгіршого сценарію розвитку подій.

– Перший рівень: вибір стратегії u.

– Другий рівень: найгірший сценарій w серед множини $\mathcal{W}$.

$$\min_{u\in\mathcal{U}} \max_{w\in\mathcal{W}} C\big(x,u,w\big).$$

3. Робастна оптимізація. Гарантує прийнятний результат навіть у найменш сприятливих умовах (мінімакс). Підхід «мінімакс»: забезпечити прийнятний рівень результату при найгіршому втіленні невизначеностей. Наприклад, оптимізація податкових ставок так, щоб за будь-яких коливань споживцих цін, дефіцит бюджету не перевищував заданого порогу.

4. Фазі-нечітке моделювання. Перетворює якісні експертні оцінки в числові ваги через нечіткі множини.

- Використання нечітких множин для опису нечіткості оцінок експертів (наприклад, «високий», «середній» рівень ризику);

– Функції належності дозволяють трансформувати якісні оцінки в числові ваги для мультикритеріальних моделей.

5. Мультиоб’єктивне програмування. Шукає Парето-оптимальні рішення для кількох одночасних цілей.

– Одночасна оптимізація кількох цілей (зростання ВВП, мінімізація інфляції, зниження безробіття, зростання надходжень до Пенсійного фонду).

– Знаходження Парето-оптимального фронту стратегій, з якого можна обирати компромісні рішення.

Для розв’язання сформульованих моделей розвитку системи соціального захистиу та соціального забезпечення, в умовах невизначеності використовують різноманітні чисельні та імітаційні методи. Далі наведено ключові підходи до їх реалізації:

- Стохастичне динамічне програмування для моделей із марківською залежністю невизначеностей.
- Метод сценарних дерев — побудова ймовірнісного дерева розвитку подій із відгалуженнями за ключовими шоками.
- Монте-Карло симуляції для оцінки розподілу результатів за стохастичних вхідних параметрів.
- Системна динаміка та агент-бейсд моделювання для відтворення нелінійних взаємодій між підсистемами.

Приклад формалізації для системи соціального захисту та соціального забезпечення України

Нехай, $x(t)$ = (ВВП, інфляція, демографічна ситуація), u(t) = (облікова ставка НБУ, страхові внески, субсидії), w(t) = (наповнення місцевих бюджетів, економічна активність підприємств, рівень зовнішньої допомоги). Метою може бути мінімізація сумарних втрат (ВВП, інфляція, демографічна ситуація) за заданого набору сценаріїв війни та зовнішніх викликів.

Таким чином, формалізація задач моделювання полягає у побудові єдиного математичного каркасу, що поєднує динамічні рівняння, опис невизначеностей і критерії оптимальності, що дає змогу порівнювати та обирати моделі управління в умовах високих ризиків.

Крім того, сучасні дослідження в галузі економічної кібернетики пропонують інтеграцію інформаційних технологій для підвищення точності прогнозування. Використання алгоритмів машинного навчання та Big Data дозволяє аналізувати великі обсяги даних, виявляти приховані закономірності та розробляти більш адаптивні моделі управління. Цифрова трансформація економічних систем підвищує ефективність управлінських рішень навіть в умовах навизначеності та соціально-економічної нестабільності.

У періоди економічного спаду та соціальних викликів складні соціально-економічні системи втрачають стабільність, а традиційні управлінські моделі часто виявляються неефективними. За таких умов виникає необхідність у декомпозиції – поділі великої системи на підсистеми з метою

полегшення її аналізу, локалізації проблем та пошуку ефективних управлінських рішень. Декомпозиція є ключовим інструментом системного аналізу, який дозволяє зменшити складність дослідження та управління, зберігаючи цілісність розуміння об’єкта.

Декомпозиція [33] — розбиття великої системи на взаємопов’язані, але відносно автономні підсистеми — є одним із ключових інструментів системного аналізу. У кризових умовах (рецесія, соціальні потрясіння, війна, пандемія) правильний вибір «точок розподілу» дозволяє:

– виділити критичні елементи, чия стійкість чи вразливість визначають загальну поведінку системи.
– оптимізувати ресурси управління, фокусуючись на найважливіших підсистемах.
– підвищити гнучкість реагування, делегуючи оперативні рішення локальним ланкам.

Декомпозиція базується на принципах ієрархічної організації систем, які були запропоновані в роботах Г. Саймона та Л. фон Берталанфі [33]. На їх думку, будь-яка складна система має природну ієрархію підсистем, і ефективне управління можливе саме шляхом дослідження та регулювання окремих її частин.

У контексті соціально-економічних систем декомпозиція передбачає розподіл на функціональні (виробництво, споживання, розподіл ресурсів), структурні (галузі економіки, регіони, соціальні групи) та інституційні (державні, корпоративні, громадські) підсистеми. Такий підхід дає змогу ідентифікувати вузькі місця, зони ризику або кризи, що потребують першочергової уваги.

Схематично, цей підхід можна зобразити як дерево розкладу системи (decomposition tree), де на верхньому рівні – макросистема (економіка країни), нижче – галузі, регіони, домогосподарства тощо.

У науковій та прикладній практиці застосовуються кілька методів декомпозиції [33]:

– функціональна декомпозиція – поділ за функціями (виробництво, управління, розподіл, обслуговування).
– структурна декомпозиція – поділ за складом системи (галузі, підгалузі, підприємства).
– декомпозиція за сценаріями розвитку – особливо актуальна в умовах невизначеності, дозволяє окремо моделювати вплив різних кризових сценаріїв на окремі частини системи.

З математичної точки зору, якщо систему S можна представити як набір підсистем:

$$S = \bigcup_{i=1}^{n} S_i,$$

то цільова функція управління (наприклад, максимізація стабільності чи мінімізація втрат) теж може бути декомпонована:

$$F(S) = \sum_{i=1}^{n} F(S_i),$$

де $F(S_i)$ – функція ефективності підсистеми S_i.

Крім економічного спаду, соціальні виклики – війна, масова міграція, зміна демографічної структури, зростання нерівності – вимагають включення соціальних змінних у модель декомпозиції. Наприклад, у період війни в Україні (з 2022 року) система декомпозується з урахуванням таких додаткових факторів:

– розподіл за регіонами відповідно до ступеня військової загрози;
– вплив інфраструктурних руйнувань на економічну активність;
– наявність гуманітарної допомоги та її розподіл.

У цих випадках доцільно поєднувати декомпозицію економічної складової з гуманітарною, враховуючи не лише фінансові показники, а й рівень життя, доступ до ресурсів, демографічні зміни, потреба у соціальному захисті та соціальному забезпеченні.

Одним із прикладів декомпозиції в Україні є моделі, які розроблялись Інститутом економічного прогнозування НАН України, де економіка поділялась на 10-15 галузей з урахуванням взаємозв'язків між ними. Подібні моделі дозволили у 2022–2023 роках оцінити втрати від війни в різних секторах [34].

Іншим прикладом є моделі UNDP [35] (Програма розвитку ООН), які використовують регіональну декомпозицію для оцінки стійкості громад до шоків. Зокрема, в межах Human Development Reports декомпозується система за критеріями доступу до освіти, медицини, зайнятості.

У складних умовах економічного спаду та соціальних викликів декомпозиція соціально-економічної системи набуває особливого значення для виявлення "слабких ланок" та оперативного реагування на критичні процеси [36-39].

Соціальні виклики, такі як демографічні зміни, зростання нерівності, війна чи міграційні процеси, інтегруються в модель декомпозиції через додаткові соціальні індикатори. Наприклад, можна додати параметри, що характеризують доступність освіти, охорони здоров'я, житловий стандарт та соціальну захищеність. Математично це може бути виражено розширенням базової моделі за принципом мультифакторного аналізу, де вихідна функція системи набуває вигляду:

$$F(S) = \sum_{i=1}^{n} \left(\lambda_i F_{econ}(S_i) + \mu_i F_{soc}(S_i) \right),$$

де Fecon(Si) – економічна ефективність підсистеми S_i, $F_{soc}(S_i)$ – соціальний аспект, а λі та μі – вагові коефіцієнти, що визначають важливість кожного з компонентів.

Прикладом застосування такого підходу є аналіз наслідків економічної кризи під час війни в Україні, коли інституції, що відповідають за інфраструктурну підтримку, охорону здоров'я та соціальний захист, зазнають значного впливу. У таких умовах доцільно поєднувати економічну та

гуманітарну декомпозицію, що дозволяє врахувати не лише прямі економічні втрати, а й вплив на рівень життя населення. Аналітичні записки Інституту економіки та прогнозування НАН України (2022–2023) [41] демонструють, як декомпозиція дозволяє розподілити втрати по секторам та регіонам, виявляючи, наприклад, що окремі регіони втрачають до 30–40% виробничої потужності порівняно з базовим рівнем.

Для побудови діаграм, що ілюструють процес декомпозиції, можна використовувати такі візуальні інструменти, як блок-схеми та діаграми причинно-наслідкових зв’язків. Наприклад, блок-схема може починатися з макрорівня "економіка України", потім розгалужуватись на галузеві підсистеми, кожна з яких детальніше розбивається на регіональні компоненти та соціальні індикатори. Таке графічне представлення дозволяє оперативно визначити, де саме в системі виникають найбільш критичні проблеми, і зосередити на них увагу при розробці антикризових заходів.

У спеціалізованій літературі детально описані підходи до декомпозиції складних систем. Роботи Г. Саймона та Л. фон Берталанфі [33] заклали основу для ієрархічного підходу до розбиття систем на підсистеми. Практичні приклади, наведені в працях М. З Згуровського, Н. Д. Панкратової [33], демонструють, як за допомогою систменого аналізу можна визначити ключові вузли впливу та оптимізувати управлінські рішення. Крім того, аналітичні звіти UNDP [35] та інституцій на кшталт Інституту економіки та прогнозування НАН України [41] дають практичну оцінку економічних наслідків криз та пропонують адаптивні стратегії для подолання викликів.

Таким чином, особливості декомпозиції в умовах економічного спаду та соціальних викликів полягають у здатності розподіляти складну систему на більш керовані підсистеми з врахуванням специфіки галузевих, регіональних та соціальних аспектів. Конкретні математичні моделі та візуальні діаграми допомагають ідентифікувати критичні зони, що дозволяє розробляти ефективні стратегії реагування та оптимізації ресурсного забезпечення. Інтеграція таких підходів забезпечує гнучкість управлінських рішень і сприяє

швидкому адаптуванню системи до змін зовнішнього середовища, що є ключовим чинником успішного подолання кризових явищ.

Дослідження функціональних зв'язків у складних соціально-економічних системах є ключовим етапом для розуміння того, як взаємодіють різні компоненти системи та як зміни в одному елементі можуть впливати на інші. Складність таких систем зумовлена не лише великою кількістю взаємозалежних елементів, а й наявністю як прямих, так і опосередкованих зв'язків, зворотних петель та мультиплікативних ефектів. Для формалізації цих взаємодій застосовують різноманітні кількісні й якісні методи:

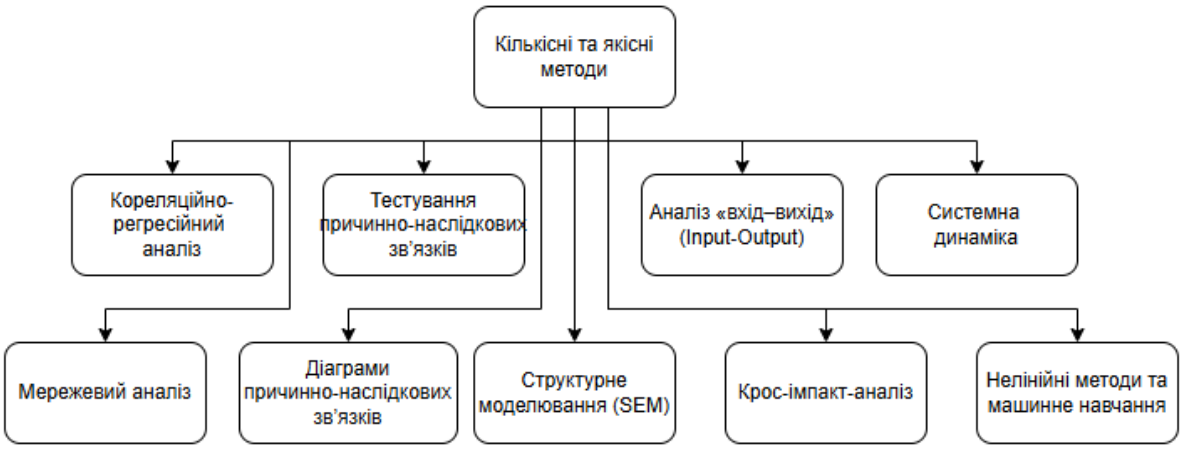

Рисунок 1.4 – Методи досілдження функціональних зв'язків складних систем

За допомогою коефіцієнтів кореляції та багатофакторних регресійних моделей формалізують залежність між макроекономічними індикаторами. Наприклад, лінійна модель [42-48]

$$Y = \alpha + \sum_{i=1}^{n} \beta_i X_i + \varepsilon,$$

де Y – вихідний показник (ВВП), Xi – фактори (інвестиції, споживчий попит, державні витрати), βi – оцінені коефіцієнти, а ε\varepsilonε – випадкова складова. Це дає змогу кількісно оцінити прямі ефекти та відокремити шум.

Для тестування причинно-наслідкових зв’язків використовують тести Granger-каузальності та VAR-моделі для перевірки, чи передують одні зміни іншим. Імпульсно-реакційні функції VAR [42] показують, як шок в одному показнику (наприклад, курс гривні) розповсюджується по інших (інфляція, зайнятість).

Аналіз «вхід–вихід» (Input-Output). Leontief-матриці та мультиплікатори дозволяють оцінити прямий і непрямий вплив галузевого приросту на ВВП і зайнятість. Наприклад, зростання випуску IT-послуг породжує мультиплікативний ефект у суміжних секторах.

Моделювання системної динаміки може бути здійсене через диференціальні рівняння з урахуванням накопичувальних «складників» (капітал K, трудові ресурси L) і потоків [49-53]. Система

$$\dot{Y} = f(Y, K, L), \quad \dot{K} = g(Y, K)$$

описує як прямі, так і зворотні зв’язки між економічним зростанням та використанням факторів виробництва, що дає змогу виявити нелінійні циклічні явища в національній економіці та її складових.

Мережевий аналіз передбачає, що вузли — це галузі, регіони чи інституції, ребра — сила впливу між ними. Метрики центральності (degree, betweenness, closeness) виокремлюють критичні «містки» системи, через які проходять найбільш важливі потоки інформації чи ресурсів.

Діаграми причинно-наслідкових зв’язків візуалізують взаємодії між факторами, полегшуючи інтерпретацію складних залежностей. Наприклад, схема може показати, як інфляція впливає на купівельну спроможність, що у свою чергу змінює рівень споживчого попиту та зайнятості. Такі діаграми ефективні для стратегічного планування та аналізу сценаріїв.

Структурне моделювання (SEM). Дозволяє аналізувати складні системи, що включають як спостережувані, так і латентні (приховані) змінні. SEM

поєднує регресійний та факторний аналіз, забезпечуючи глибше розуміння впливу нефінансових чи нематеріальних факторів на макропоказники.

Крос-імпакт-аналіз грунтується на експертному оцінюванні впливів між подіями чи змінними. Дозволяє виявити взаємозалежні ризики й сформувати динамічну картину ймовірного розвитку подій у складній ситуації. Ефективний у кризовому плануванні.

Нелінійні методи та машинне навчання використовують для виявлення складних, нечітких, або непередбачуваних залежностей у великих наборах даних. До них належать нейронні мережі (наприклад, LSTM), випадкові ліси, Copula-моделі тощо. Вони дають змогу будувати гнучкі прогностичні моделі, які адаптуються до нових даних.

Методика формування множини найбільш значимих факторів – це комплексний підхід, який дозволяє визначити, ранжувати та інтегрувати ключові показники, що впливають на процес прийняття управлінських рішень. Такий підхід важливий у сучасних умовах, коли обсяг даних значно зростає, а вплив різних економічних, соціальних та політичних чинників набуває все більшої значущості [52-54].

Основою даної методики є концепція мультикритеріального аналізу, яка передбачає одночасне врахування великої кількості показників та їх взаємодію. Теоретично, підхід ґрунтується на системному мисленні, коли складну систему можна розбити на взаємозалежні підсистеми, де кожна взаємодіє з іншими через низку функціональних зв’язків.

На базі цієї концепції формується початковий набір факторів, який включає:

– кількісні показники: макроекономічні індикатори (ВВП, інфляція, рівень безробіття, демографічна сиуація), фінансові дані (інвестиції, доходи, витрати бюджету) та статистичні дані, що характеризують економічну динаміку.

– якісні оцінки: експертні думки, результати соціологічних опитувань, аналітичні звіти та дослідження, які відображають невимірювані аспекти, такі як ступінь ризику, вплив політичних рішень чи соціальних змін.

Етапи формування множини факторів

1. Збір та первинний відбір даних. На цьому етапі формується розширена база даних, яка охоплює різні аспекти системи. Для початкового відбору факторів застосовують експертний аналіз та Delphi-метод. Завдяки залученню експертів із різних сфер (економіки, соціології, управління) можливо виявити потенційно важливі показники, навіть якщо вони не мають прямої кількісної оцінки [33, 53, 57].

2. Статистичний аналіз та скорочення розмірності даних. Після збору даних використовується статистичний аналіз для виявлення основних компонент, що пояснюють найбільшу частину варіативності системи. Одним із основних інструментів є метод головних компонент (Principal Component Analysis, PCA). Математично це можна представити як [60-62]:

$$Z_j = a_{j1}X_1 + a_{j2}X_2 + \cdots + a_{jn}X_n,$$

де Z_j – j-та головна компонента, X_i – початкові змінні, а a_{ji} – коефіцієнти, що визначають внесок кожного показника у формування компонент. Такий підхід дозволяє зменшити кількість вихідних змінних, зберігаючи при цьому найбільш релевантну інформацію.

3. Ранжування факторів за допомогою методів мультикритеріального аналізу. Для подальшого ранжування та визначення вагових коефіцієнтів використовуються методи аналізу ієрархій (Analytic Hierarchy Process, AHP) та зваженого балового оцінювання. За допомогою AHP експерти проводять парні порівняння кожного з факторів, отримуючи матрицю порівнянь, яка дозволяє обчислити відносні ваги wiw_iwi. Формально ефективність прийняття рішень може бути оцінена наступною формулою [33, 64]:

$$E = \sum_{i=1}^{n} w_i f_i(X_i),$$

де $f_i(X_i)$ – функція впливу iii-го фактора, яка може бути як лінійною, так і нелінійною, залежно від природи зв'язку.

4. Аналіз чутливості та перевірка моделі. Після визначення ключових факторів важливим етапом є аналіз чутливості. Цей метод дозволяє оцінити, як невеликі зміни у значеннях критичних факторів впливають на загальну ефективність прийняття рішень. Метод аналізу чутливості допомагає ідентифікувати ті показники, зміна яких може спричинити суттєві коливання в кінцевих результатах, що дозволяє адаптувати стратегії управління.

Приклад 1: Управління економічними ризиками в кризових умовах. У дослідженнях, спрямованих на управління ризиками під час економічних криз, застосування комбінованого методу дозволило сформувати набір ключових факторів, серед яких:

— інвестиційна активність,

— споживчий попит,

— рівень інфляції,

— стабільність валютного курсу,

— політична стабільність.

За допомогою AHP експерти визначили вагові коефіцієнти для кожного з цих показників, після чого було проведено аналіз чутливості, який показав, що інвестиційна активність та валютний курс мають найбільший вплив на економічну стабільність. Результати цього аналізу стали основою для розробки антикризових заходів, що дозволили оперативно реагувати на зміни в економічному середовищі.

Приклад 2: Розробка системи підтримки прийняття рішень у державному управлінні. У контексті державного управління для підвищення інформативності прийнятих рішень було розроблено інтегровану систему, що включала аналіз соціальних, економічних та інституційних факторів.

Початковий набір показників було зведено до кількох ключових компонентів за допомогою методу головних компонентів, після чого за допомогою AHP експерти визначили їх важливість. Аналіз чутливості дозволив виявити, що соціальні індикатори (рівень освіти, доступ до медичних послуг, рівень соціального забезпечення) у певних регіонах мають вирішальний вплив на ефективність державної політики, що підтвердило необхідність регіонального підходу до прийняття рішень.

Для підвищення інформативності прийняття рішень часто застосовують інтерактивні панелі управління (dashboards), які інтегрують результати статистичного аналізу, графічні діаграми, мапи та блок-схеми. Наприклад, на панелі управління можуть бути представлені:

— кругові діаграми, що відображають пропорційний внесок кожного фактора,

— гістограми розподілу ключових показників,

— діаграми причинно-наслідкових зв’язків,

— інтерактивні мапи, що показують регіональні розбіжності.

Таке візуальне представлення дозволяє особі, що приймає рішення швидко зорієнтуватися у складній системі показників та приймати оптимальне.

Оптимізація цільових функцій при управлінні складними соціально-економічними системами та процесами є одним із ключових напрямків удосконалення прийняття рішень. Вона передбачає математичне формулювання цілей, визначення критеріїв ефективності та пошук таких управлінських рішень, які забезпечують максимальну користь або мінімальні витрати в рамках заданих обмежень.

Основна ідея оптимізації полягає у побудові цільової функції, що характеризує бажаний результат. Наприклад, цільова функція може бути сформульована як максимізація економічного зростання, мінімізація витрат чи максимізація рівня соціального добробуту. Формально це можна записати як задачу оптимізації:

$$\max_{x \in \mathbb{X}} F(x) = \sum_{i=1}^{n} c_i x_i,$$

де $x = (x_1, x_2, \ldots, x_n)$ – вектор управлінських рішень або ресурсів, c_i – коефіцієнти, що відображають вагу кожного компонента, а X – множина допустимих рішень, визначена обмеженнями (ресурсними, технологічними, соціальними тощо).

У більш складних системах, де вплив численних факторів є одночасно позитивним і негативним, часто застосовують мультицільову оптимізацію. Тут цільова функція формується як агрегат декількох критеріїв:

$$\min_{x \in \mathbb{X}} G(x) = \sum_{i=1}^{n} w_i f_i(x),$$

де fi(x) – функції, що характеризують різні аспекти ефективності (наприклад, економічну вигідність, рівень ризику, соціальну стабільність), а wi – вагові коефіцієнти, що відображають пріоритетність кожного критерію. Цей підхід дозволяє інтегрувати різноманітні критерії в єдину модель і знаходити компромісні рішення, які задовольняють всі вимоги.

В залежності від характеру цільової функції та обмежень використовують різні методи оптимізації [65-67]:

- Лінійне програмування (LP): Якщо цільова функція та обмеження мають лінійний вигляд, використовується стандартне лінійне програмування. Наприклад, задача максимізації прибутку може бути представлена у вигляді:

$$\begin{aligned} \max_{x \in \mathbb{R}^n} \quad & \sum_{i=1}^{n} c_i x_i, \\ \text{при умовах} \quad & Ax \leq b, \\ & x \geq 0. \end{aligned}$$

Нелінійне програмування (NLP): Для випадків, коли цільова функція або обмеження є нелінійними, застосовують методи нелінійного програмування. Це дозволяє врахувати нелінійні взаємозв’язки між компонентами системи.

- Динамічне програмування: Використовується для розв’язання задач, де прийняття рішення відбувається послідовно у часі. Цей метод дозволяє врахувати часову динаміку процесів та адаптувати стратегії управління до змін у системі.

- Еволюційні алгоритми та генетичне програмування: При розв’язанні надскладних задач, де традиційні методи оптимізації можуть бути неефективними, використовують алгоритми на основі еволюційної логіки. Вони імітують природні процеси відбору та мутації, що дозволяє знаходити глобальні оптимуми навіть у багатовимірних просторах.

- Методи симуляційного моделювання: Часто використовуються для перевірки та тестування отриманих рішень у віртуальному середовищі, що дозволяє оцінити вплив окремих змінних на загальну ефективність системи.

У практиці управління соціально-економічними системами оптимізація цільових функцій застосовується, наприклад, у таких випадках.

Під час економічного спаду необхідно мінімізувати втрати та стабілізувати ринок. Оптимізаційні моделі дозволяють визначити найбільш ефективне поєднання заходів, таких як стимулювання інвестицій, регулювання процентних ставок та підтримка ключових секторів економіки. Також оптимізація розподілу бюджетних коштів між різними секторами (освіта, охорона здоров’я, соціальний захист) дозволяє максимізувати соціальний добробут за обмежених ресурсів. За допомогою оптимізаційних моделей можна оцінити ефективність різних інвестиційних проектів і вибрати ті, які забезпечують максимальний економічний ефект при мінімальних ризиках.

Висновки до розділу 1

На основі огляду сучасних інформаційно-аналітичних систем, що використовуються у державному та публічному управлінні в Україні та за кордоном, і, зокрема, у системі соціального захисту та соціального

забезпечення досліджено їх переваги та недоліки. Виявлено, що більшість з них зосереджені переважно на наданні доступу до даних державних реєстрів та баз даних. Аналітична складова або відсутня, або обмежена функціями моніторингу, формування звітності та планування на наступний рік. З-поміж інших вирізняється аналітична система, реалізована у складі АРМ керівника Інтегрованої комплексної інформаційної системи Пенсійного фонду України (ІКІС ПФУ), яка містить інструменти Oracle BI, які використовуються переважно для формування звітності і візуалізації результатів.

На основі системного дослідження предметної області, аналізу робіт вітчизняних та закордонних науковців та фахівців-практиків, встановлено, що для прогнозування витрат на виплату пенсій та соціальних допомог, крім даних, наявних у соціальній сфері, потрібно залучити дані з зовнішніх джерел, що представлено у відповідній методиці.

Розроблено інтегрований підхід, який дозволяє поєднати у єдиній системі різні за типом, структурою та форматом представлення, дані, що наявні у державних реєстрах, аналітичних звітах, публікаціях Державної служби статистики, Національного банку України, відкритих API, Інтернет-сторінках міжнародних організацій, стрім-потоках, веб-скрапінгу, зверненнях громадян, публікації у соціальних мережах, інші текстові повідомлення та документи, що відображають потребу населення у соціальних послугах та допомогах, дають можливість ураховувати настрої різних суспільних груп, а також потокових даних, що охоплюють інформацію від сенсорних систем, онлайн-моніторингу та дані, отримані від неурядових організацій, які характеризують оперативні зміни у середовищі дослідження. Запропонована методологія відрізняється робастністю результатів, що забезпечує репрезентативність вибірок та достовірність подальших прогнозів. Для цього удосконалено інструментарій ETL-процесів, що дозволяє здійснювати видобування, трансформацію та завантаження даних із різнорідних джерел, розширено можливості веб-скрейпінгу та REST/OData API-запитів для

інтеграції з відкритими платформами, а також реалізовано механізми потокової обробки даних у режимі реального часу.

Отже, застосування розробленої методики на етапі системологічного аналізу проблеми забезпечує якнайповніше покриття предметної області різноплановою інформації (наприклад, видатки державного бюджету на соціальні виплати, демографічна статистика, гуманітарні гранти, обсяги фінансування окремих соціальних програм з місцевих бюджетів, тощо), що підвищує репрезентативність вибірок для моделювання, покращуючи якість отримуваних прогноз

РОЗДІЛ 2
МОДЕЛІ ТА МЕТОДИ ЗБОРУ ТА ОБРОБЛЕННЯ ДАНИХ

2.1 Організація роботи з консолідованими відкритими даними

Збір і завантаження сирих даних (Extract, Transform, and Load - ETL) становить перший та один із найважливіших кроків у процесі роботи інформаційно-аналітичної системи: від екстракції даних із джерел до їхнього подальшого зберігання, обробки та використання у підсистемі аналізу та прогнозуванні [25, 68]:

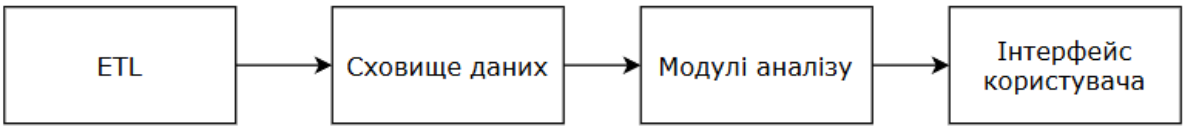

Рисунок 2.1 – Технологічний ланцюг обробки даних

Для дослідження продуктивності (records/hour) і надійності ключових модулів збору даних — веб-скрейпінгу та API-запитів виконано експеримент, в рамках якого для методів збору даних, таких як BeautifulSoup, Scrapy, Selenium, REST/OData, були емпірично оцінені їхні характеристики для відбору найефективніших рішень, які мають бути інтегровані у розроблювану систему, а також оцінено гнучкість та стійкість ETL-процесу [25, 68];

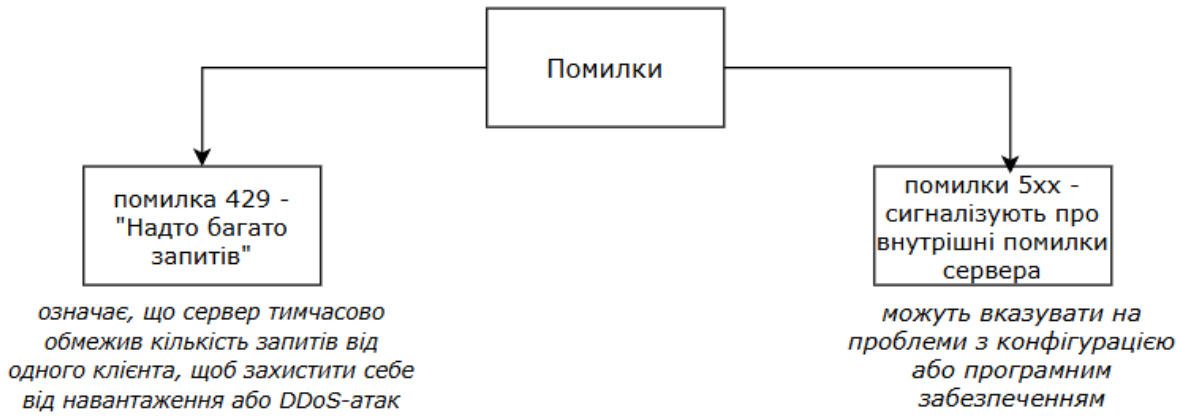

Рисунок 2.2 –Схема опрацювання помилок

Для збору даних необхідно спочатку проаналізувати DOM за допомогою DevTools (Chrome: F12 → Elements).

Далі - ідентифікувати унікальні атрибути:

(id="dataTable", class="table-responsive") та обрати між жорстким пошуком *(find("table", attrs={"id": "dataTable"})*) і контекстним селектором *(select_one("div.report-section table"))* для підвищення гнучкості парсера .

Таблиця розбита на сторінки (параметр *?page=* або AJAX), виконуємо циклічний збір із оновленням *offset* чи *page* і для динамічного підвантаження через JavaScript використовуємо Selenium, імітуючи кліки й очікуючи на появу контенту.

В якості експериментального набору даних використано дані місцевих бюджетів, розміщені на порталі data.gov.ua [69].

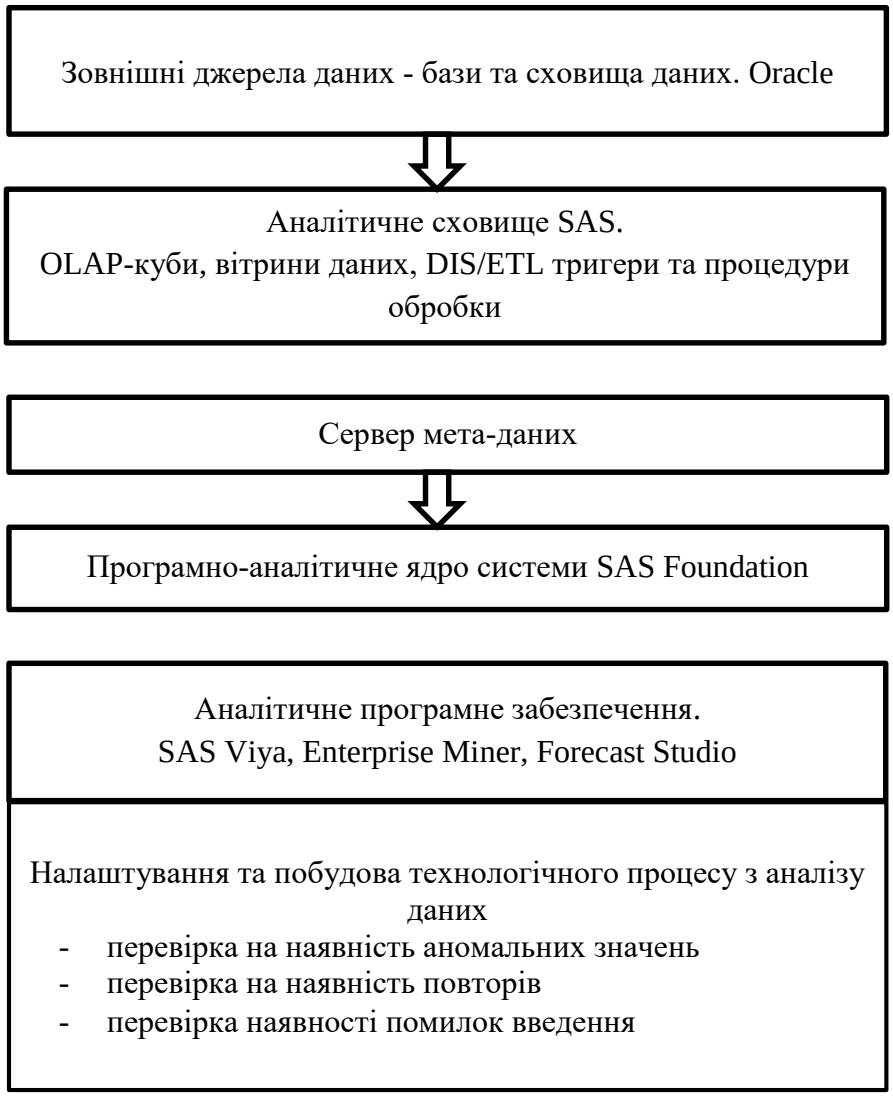

Рисунок 2.3- Схема оброблення консолідованих даних (приклад)

Портал відкритих даних «Бюджет для громадян» є одним із прикладів відкритих даних, які надаються органами місцевого самоврядування. Еа даний час користувачі, що мають доступ завантажують дані на даний портал, а всі бажаючі можть їх переглядати та скачувати.

Дане програмне забезпечення для інтеграції його у Систему пропонується доповнити засобами Big Data та методами Data Science, для розв'язання задач аналітичного процесу моделювання та прогнозування видатків та витрат місцевих бюджетів. В даному випадку, конфіденційна інформація відсутня, тому, можуть бути використані хмані сховища та засоби мультиплатформенного середовища аналізу даних. У розробці передбачається використовувати, такі інструменти як програмне забезпечення компанії SAS, мову програмування Python, а також Cloud Analytics Service SAS Viya [70-73].

Система включає в себе клієнтську частину, доступну для користувачів, яка дозволяє вибирати набір даних та встановлювати кінцеві цілі через інтерфейс користувача (UI), реалізований за допомогою бібліотеки ReactJS.

Серверна частина системи відповідає за обчислення та логіку додатку. У ній виконуються всі математичні обчислення та реалізуються різні підходи до аналізу даних. Ця частина реалізована з використанням мови програмування Python []73.

Для зберігання інтермедіатних значень та результатів аналізу даних в експерименті використовується база даних PostgreSQ, а в умовах Системи - Oracle Database. Забезпечення взаємодії між клієнтською та серверною частиною реалізоване за допомогою HTTP-запитів.

Усі компоненти системи розгорнуті в хмарному середовищі SAS Viya для забезпечення доступності та швидкокідії сервісів та з метою спрощення подальшої інтеграції з платформою Oracle [25, 71].

Для масового завантаження структурованих даних використовується CKAN REST API, яке надає доступ до «сирих» записів конкретного набору через єдиний ендпоінт *datastore_search*. Щоб отримати дані, спочатку потрібно знати унікальний ідентифікатор набору *resource_id* (наприклад,

3ec45fce-9ecd-4bd9-b682-c00af0cb0ef5), що видається під час реєстрації або пошуку в каталозі наборів даних.

Запит виконується GET-методом за адресою:

https://data.gov.ua/api/3/action/datastore_search
?resource_id=3ec45fce-9ecd-4bd9-b682-c00af0cb0ef5
&limit=1000
&offset=0

де параметр *limit* визначає максимальну кількість записів у відповіді (сторінку), а *offset* — зміщення для реалізації пагінації.

У результаті сервер повертає JSON-об’єкт із полем *result.records*, що містить масив записів, кожен із яких має такі ключі:

- *id* — внутрішній ідентифікатор запису,
- *region_code* — код адміністративного регіону,
- *year, month* — період,
- *expense_type* — тип соціальних виплат,
- *amount_nominal, amount_real* — суми в номінальних та реальних показниках,
- *recipients_count* — кількість отримувачів виплат.

Приклад JSON-об’єкту, який було повернуто в якості результату запиту представлено в Додатку Б:

Оскільки багато API встановлюють обмеження на кількість запитів за одиницю часу, при перевищенні ліміту сервер повертає код 429 Too Many Requests. Щоб забезпечити безперебійне завантаження, «клієнт» повинен реалізувати стратегію back-off: наприклад, при 429-й відповіді призупиняти виконання на 60 секунд і повторювати запит доти, доки ліміт не буде скинуто.

Такий підхід, дозволяє надійно і прогнозовано отримувати велику кількість даних із Порталу відкритих данихOpenBudget [74], мінімізуючи пропуски та помилки при інтеграції в ETL-флоу.

Портал OpenBudget (openbudget.gov.ua) забезпечує публічний доступ до даних про виконання місцевих бюджетів України через REST-інтерфейс, що

дозволяє оперативно отримувати структуровані JSON-об’єкти з деталізацією за регіонами, роками та кодами бюджетної класифікації. Основний ендпоінт для отримання даних про виконання місцевих бюджетів виглядає так:

GET https://openbudget.gov.ua/api/v1/execution/local
?region_code=UA-30
&year=2023
&limit=50
&offset=0
&sort=amount.desc

де параметр *sort=amount.desc* дозволяє відсортувати записи за спаданням суми видатків.

Для налаштування запиту доступні такі ключі:

- *region_code* – код адміністративного регіону (наприклад, *UA-30* для Київської області);
- *year* – календарний рік (наприклад, *2023*);
- *limit* і *offset* – механізм пагінації, що дозволяє контролювати розмір “сторінки” даних та зміщення всередині вибірки;
- *sort* – ключ сортування (наприклад, *amount.desc* або *month.asc*).

Крім базового REST-інтерфейсу, OpenBudget підтримує OData-запити, які дають змогу ще гнучкіше формувати вибірки на стороні сервера:

- *$filter* – фільтрація за довільними умовами (наприклад, *year eq 2023 and region_code eq 'UA-30'*);
- *$select* – вибір тільки необхідних полів (наприклад, *region_name,amount,year*);
- *$orderby* – сортування за вказаними атрибутами;
- *$top/ $skip* – аналоги *limit* та *offset*, що особливо зручно при побудові складних OData-запитів із конвеєрною обробкою даних.

Для доступу до частини ендпоінтів може знадобитися API-ключ. Після реєстрації на порталі, користувач отримує унікальний токен, який слід передавати в заголовку кожного запиту:

Authorization: Bearer <API_KEY>

У разі перевищення ліміту запитів (HTTP 429 Too Many Requests) клієнтський скрипт має реалізувати логіку back-off (як і у попередньому прикладі): при отриманні коду 429 — зробити паузу (наприклад, 60 с) і повторити запит, доки сервер не дозволить подальші звернення.

Завдяки наведеним можливостям OpenBudget API, інтеграція даних у ETL-флоу стає швидкою, гнучкою та надійною - саме на цьому етапі закладається основа для подальшої попередньої обробки, моделювання й аналітики соціальних видатків.

Наступним кроком є порівняння продуктивності ключових інструментів збору даних — веб-скрейпінгу (BeautifulSoup, Scrapy, Selenium) та API-запитів (REST/OData до data.gov.ua і openbudget.gov.ua) в умовах реального навантаження.

Умови проведення тесту:

– комп’ютер: 4-ядерний сервер, 8 GB RAM, інтернет-з’єднання 100 Mbps.

– час на кожний запуск: 1 година, одночасно запущені по одному процесу кожного інструмента.

– вимірюємо: кількість успішно отриманих рядків (записів) за годину і число помилок HTTP 429 чи 5xx.

Таблиця 2.1 – Результати експериментів з оброблення даних

Інструмент	Джерело	Записи/ годину	429/5xx помилок	Деталі реалізації
BeautifulSoup	HTML (treasury)	8 500	5×429	Синхронний підхід: *requests.get()* + *bs4* парсинг; без кешування запитів
Scrapy	HTML (treasury)	15 200	1×429	Асинхронна обробка черги запитів, вбудована пагінація, автокешування
Selenium	HTML (treasury)	2 100	0	WebDriver очікує рендерингу JS; імітація кліків; обмежена пропускна здатність
REST (Python requests)	data.gov.ua	30 000	0×429, 0×5xx	Прямі GET-запити з пагінацією; простий back-off при 429
OData (pandas)	openbudget	28 500	2×429	Завантаження JSON-контенту через *read_json();* автоматичне розбиття на чанки.

Як показали результати експериментів (табл. 2.1) Scrapy показує найвищу пропускну здатність завдяки неблокуючому I/O та кешуванню HTTP-відповідей. REST/OData-запити до CKAN/OpenBudget API стабільні та майже не дають помилок сервера, але потребують обробки пагінації й ліміту запитів.

Після оцінки продуктивності компонентів збору даних наступним критичним кроком є перевірка їхньої якості. Оцінка повноти й коректності даних дозволяє виявити пропуски та невідповідності, здатні суттєво спотворити результати подальшої обробки та прогнозування.

2.2 Особливості роботи з текстовими даними

У системі соціального захисту та соціального забезпечення робота із зверненнями громадян, різними текстовими повідомленнями, текстовими частинами звітів, описом дпкументів, відіграє особливу роль. Тому, засоби аналізу тексту є незамінним інструментом для вилучення знань із неструктурованих джерел [75-79].

Використання текстової аналітики дозволяє фахівцям, що працюють з неструктурованими даними вилучати зміст, закономірності, шаблони, що приховані в них. За даними компанії International Data Corporation (IDC), близька 80% даних, що зберігаються в організаціях, це тексти. [2/4/1]

Текстова аналітика включає в себе інструменти та методи, які використовуються для отримання поглиблених знань із неструктурованих даних. Ці методи можна класифікувати наступним чином:

- інформаційний пошук (information retrieval);
- розвідувальний аналіз (exploratory analysis);
- вилучення концептів та понять (concept extraction);
- реферування (summarization);
- категоризація (categorization);
- аналіз тональності тексту (sentiment analysis);
- керування контентом (content management);

- керування онтологіями (ontology management).

Типи вилучення інформації з тексту (вказані по ступеню складності задачі).

1. Видобуток токенів;

2. Вилучення терма (лексема + мова = терм).

3. Виділення понять (іменники, іменникові словосполучення).

4. Вилучення сутностей (асоціює іменники з сутностями, наприклад, особа: містер Зеленський, місцезнаходження: урядовий квартал).

5. Атомарне виділення фактів (асоціює іменники з дієсловами, тобто суб'єкт => одержує, наприклад, управління => призначає).

6. Вилучення складних фактів (розуміння природної мови).

Підхід до аналізу та кількісного визначення текстових даних залежить від завдання, яке треба виконати. Наприклад, можна навчити прогнозну модель імітувати те, як фахівець Пенсійного фонду визначає категорії для різних за тематикоб звернень громадян. Така контрольована класифікація часто потребує вирішення проблем, пов'язаних із підготовкою даних і побудовою моделей.

Отже, текстовий локумент складається з наступних елементів:

- літери
- слова
- речення
- параграфи
- розділові знаки
- можливі структурні елементи документу (розділи, розділи)

Наведені елементи документа можна порахувати і порівняти з вмістом інших документів. Для цілей даного дослідження, доцільно скористатись заоном Ципфа-Мандельброта [75-79]. Якщо, всі слова мови (або просто достатньо довгого тексту) впорядкувати за спаданням частоти їхнього використання, то частота n-го слова в такому списку виявиться приблизно обернено пропорційною його порядковому номеру n (так званому рангу цього

слова). Наприклад, друге за вживаністю слово трапляється приблизно вдвічі рідше, ніж перше, третє – втричі рідше, ніж перше, і так далі [75-79].

Нехай $t_1, t_2, \dots, t_n$ – терми в колекції документів розташовані в порядку від такого, що зустрічвється найбільш частого до того, що зустрічається найменш частого. $f_1, f_2, \dots, f_n$ – відповідні частоти термів. Тоді частота f_k для терма t_k пропорційна $\frac{1}{k}$.

На практиці закон Ципфа- Мандельброта виводиться як степеневий закон із вільними параметрами, які можна оцінити на основі колекції документів (2.1) [75-79]:

$$f_k = \frac{C}{(\omega+k)^\theta} \tag{2.1}$$

де C це така константа, що для заданих ω та θ виконується умова

$$\sum_{k=1}^{n} f_k = T$$

де T – загальна кількість термів в колекції документів.

Параметри ω та θ оцінюються для відповідної колекції документів.

За інших підходів, таких як, наприклад, приховані Марківські моделі (HMM – Hidden Markov Model) [80], документ розглядається як набір станів, визначених словами. Шляхів, що можуть переводити документ від одного слова до іншого, може бути багато, але конкретний документ має лише один реалізований шлях. Якщо можна обчислити ймовірності, пов’язані зі станами (термами) і шляхами (проміжними термами) для колекції документів, то можна отримати відповіді на питання вигляду: «Враховуючи терм і слова, що стоять перед і після терму, яка ймовірність того, що терм є іменником?». Спеціалізоване програмне забезпечення, що використовує HMM, дає користувачеві можливість визначати, наприклад, максимальну довжину шляху по відношенню до кількості станів, що містяться в шляху. Для позначення частин мови та ідентифікації багатослівних термінів використовують «жорстко закодовані» значення, наприклад, скільки слів слід

перевірити перед термом, який цікавить, і скільки слів після терму можна використовувати для обчислення ймовірностей.

Основна стратегія кількісного визначення тексту, представленого у довільній формі передбачає:

1) отримання корпусу термів, які будуть використовуватися після вилучення кореня, створення синонімів, фільтрації тощо.

2) представлення кожного документа та кожного терма у векторному просторі за допомогою матриці вигляду (1) документи по рядках та терми по стовпцях, або (2) терми по рядках та документи по стовпцях.

3) проектування документів і термінів у векторний простір меншої розмірності (метод svd).

4) використання методів кластеризації та генерації тем для документів у цьому векторному просторі меншої розмірності.

В табл. 2, нижче, показано приклад представлення корпусу документів у вигляді матриці – документи по строках та терми по стовпчиках.

Відповідно, транспонована табл. 2, це буде представлення корпусу документів у вигляді матриці – терми по строках та документи по стовпчиках.

Представлення документа за частотами термів, що в ньому присутні, дозволяє відобразити документ у векторному просторі, елементами якого є частоти кожного терму. Зазвичай корпус документів представляє собою багатовимірний простір із багатьма значеннями 0 (нуль). Розмірність можна зменшувати за допомогою етапів обробки текстів, таких як вилучення кореня, створення синонімів і фільтрація низькочастотних термінів.

Зауважимо, що матриці вигляду (1) документи по строках та терми по стовпчиках, та (2) терми по строках та документи по стовпчиках, містять таку ж саму інформацію.

Таблиці вигляду терми по строках та документи по стовпчиках використовуються для аналітичних цілей.

Частотні матриці даних зазвичай містять 90% частот, що дорівнюють 0 (нуль).

Крім того, згідно із законом Ципфа, підрахунок частоти термінів є дуже довгим. Зазвичай існує невелика кількість дуже поширених термів, які використовуються в більшості документів.

Проблеми розмірності та розрідженості частотної матриці корпусу текстів вирішується за допомогою ключової теореми з лінійної алгебри, яка називається методо сингулярного розкладу матриці (SVD) [81].

Однак дослідники на практиці показали, що перед застосуванням SVD методу краще виконувати операцію зважування значень частот. Окрім цього операція зважування також допомагає пом'якшити проблему нерівності високочастотних членів, роблячи їх менш впливовими.

Проблема зміщених значень частот, у матриці частот термів, вирішується шляхом застосування вагових коефіцієнтів до частот. Зазвичай використовуються наступні підходи:

– локальні ваги L_{ij}, також звані частотними вагами, обчислюються для і-го терму в j-му документі.

– ваги термін G_i, які також називаються глобальними вагами, що обчислюються для термів.

– кінцеві ваги для кожної комірки частотної матриці, що є результатом множення значень $G_i \cdot L_{ij}$

Крок 1. Ваги частот (локальні ваги).

Частотні ваги, які в літературі часто називають локальними ваговими коефіцієнтами, є першим кроком у перетворенні значень частотної матриці.

Є три підходи:

1) $L_{ij} = a_{ij}$ – залишити частоти, без змін та перетворень.

2) Бінарна функція

$$L_{ij} = \{1, \text{якщо } i - \text{й терм є в } j - \text{му документі } 0, \text{в цсіх інших випадках}$$

3) Логарифмічна функція, зазвичай позначають як LOG

$$L_{ij} = log_2(a_{ij} + 1)$$

де a_{ij} – це кількість разів коли і-й терм з'явився в j-му документі.

Ваги термінів, які в літературі часто називають глобальними вагами, використовуються для заміни частотних ваг, щоб адаптувати частотну матрицю до розміру документа та розподілу термів.

Крок 2. Ваги термів (глобальні ваги).

Є чотири варіанти вибору ваг термів для терму G_i.

– Ентропія (найбільш розповсюджений підхід).

– Інверсна частота документів (IDF – Inverse Document Frequency).

– Взаємна інформація (використовується у випадках, коли в задачі є цільова змінна).

– Без будь-яких перетворень.

Позначення:

$a_{i,j}$ – частота появи і-го терма в j-му документі.

g_i – частота появи і-го терма в колекції документів.

n – загальна кількість документів в колекції.

d_i – кількість документів в яких і-й терм з'являється.

$$p_{i,j} = \frac{a_{i,j}}{g_i}$$

Формула ентропії записується як:

$$G_i = 1 + \sum_{j=1}^{d_i} \frac{p_{ij} \cdot log_2(p_{ij})}{log_2(n)}$$

де $0 \leq G_i \leq 1$, причому $G_i = 0$ означає, що інформації замало, а $G_i = 1$ інформації багато.

Тому більш кращим способом опису ваги терма, що використовується для обчислення значення G_i, це формула:

Вага терму = *1 – нормалізоване значення ентропії.*

Оскільки логарифм від нуля є не визначеним (не існує), то добуток у чисельнику дорівнює нулю, якщо значення p_{ij} дорівнює нулю.

Наведемо два приклади, що ілюструють обчислення ваги цього члена.

Приклад 1.

Нехай j-й терм з’являється лише один раз і тільки в одному документі [77-79].

$$G_i = 1 + \frac{\left(\frac{1}{1}\right) log_2(\frac{1}{1})}{log_2(n)} = 1 + \frac{(1) \cdot (0)}{log_2(n)} = 1$$

Приклад 2.

Нехай і-й терм з’являється по одному разу в кожному з n документів корпусу [77-79, 81].

$$G_i = 1 + \frac{\sum \quad (1/n) \cdot log_2(1/n)}{log_2(n)} = 1 + \frac{n \cdot (1/n) \cdot (-log_2(n))}{log_2(n)}$$

$$= 1 + \frac{-log_2(n)}{log_2(n)} = 1 - 1 = 0$$

Інверсна частота документів (IDF).

$$G_i = 1 + \ log_2\left(\frac{n}{d_i}\right)$$

де $1 \leq G_i < \infty$, причому $G_i = 1$ означає, що замало інформації, а $G_i \rightarrow \infty$ достатньо.

Якщо термін з’являється в кожному документі, тоді вага IDF дорівнює 1, оскільки тоді $d_i = n$. Максимальна вага для фіксованої колекції документів з’являється, коли терм з’являється точно в одному документі, тоді IDF дорівнює $1 + \ log_2(n)$. Верхньої межі для міри IDF не існує, оскільки кількість документів у корпусі документів , параметр n у формулі, може бути яким завгодно.

Значення ентропії та IDF значення досягають максимуму, коли терм з’являється лише в одному документі [77-79, 81].

Це означає, що відповідний терм є дуже дискримінаційним, але не дуже корисним, оскільки він зустрічається лише в одному документі. Тому зазвичай в алгоритмах обробки текстів, взагалі видаляють терми, що зустрічається не часто, наприклад менше чотирьох разів.

Обидві наведені ваги приймають мінімальні значення, якщо терм з'являється рівно один раз в кожному документі. У цьому випадку терм не є дуже дискримінаційним, оскільки він зустрічається скрізь у корпусі документів [].

Міра взаємної інформації (mutual information) [82] .

$$G_i = max(over\ k)\left[log_{10}\left(\frac{P(t_i, C_k)}{P(t_i)P(C_k)}\right)\right]$$

де $C_1, C_2, \dots, C_k$ – k рівнів категоріальної цільової змінної, $over\ k$ – перебір по всіх k рівнях категоріальної цільової змінної, $P(t_i)$ – пропорція документів, що містять і-й терм, $P(C_k)$ – пропорція документів, що містить C_k рівень цільової змінної, $P(t_i)P(C_k)$ – пропорція документів, де і-й терм присутній, а також є C_k рівень цільової змінної, область допустимих значень міри взаємної інформації: $0 \leq G_i < \infty$. Хоча теоретично значення міри G_i необмежене, на практиці воно зазвичай менше 1.

Наприклад, треба виконати обчислення значення G_i. Нехай є цільова змінна бінарного типу, тобто $k = 2$.

Таблиця 2.2 - Частота появи терму t_i в документах

	Кількість документів, де є значення цільової змінної C_1	Кількість документів, де є значення цільової змінної C_2	Частота появи терму t_i
Кількість документів, де не має терму t_i	100	25	125
Кількість документів, де є терм t_i	10	50	60
Сума	110	75	185

На основі наведеної таблиці зроблено обчислення

$$P(t_i) = \frac{60}{185} = 0{,}324$$

$$P(C_1) = 110/185 = 0{,}595$$

$$P(C_2) = 75/185 = 0{,}405$$

$$P(C_1, t_i) = 10/185 = 0{,}054$$

$$P(C_2, t_i) = 50/185 = 0{,}27$$

$$G_i = max\{log_{10}\left(\frac{0{,}054}{0{,}324 \cdot 0{,}595}\right), log_{10}\left(\frac{0{,}27}{0{,}324 \cdot 0{,}405}\right)\} = 0{,}313$$

Якщо цільова змінна не бінарна, а номінальна, чи ординарна, з кількістю рівнів $k > 2$, то наведені формули також вірні, лише треба розширити відповідну таблицю частоти появи терму t_i в документах.

Після того, як отримані локальні та глобальні ваги, їх перемножують між собою, як показано в формулі:

$$\hat{a}_{ij} = G_i L_{i,j}$$

де G_i – це глобальна вага для і-го терму, а $L_{i,j}$ – локальна вага і-Го терму в j-му документі.

Частотні матриці, де терми по строках, а документи по стовпчиках, як зважені так і не зважені – основа підходу лінійної алгебри щодо аналізу тексту. В загальному випадку така матриця може бути представлена, як показано нижче.

Таблиця 2.3 - Частотна матриці, загальна схема

терми	документи			
	d_1	d_2	…	d_n
t_1	$\hat{a}_{1,1}$	$\hat{a}_{1,2}$	…	$\hat{a}_{1,n}$
t_2	$\hat{a}_{2,1}$	$\hat{a}_{2,2}$	…	$\hat{a}_{2,n}$
…	…	…	…	…
t_m	$\hat{a}_{m,1}$	$\hat{a}_{m,2}$	…	$\hat{a}_{m,n}$

Міру взаємної інформації використовують у випадках, коли є цільова змінна в аналізі.

Вагові коефіцієнти міри ентропії та IDF надають більші ваги термам, які не часто зустрічаються або є низькочастотними за появою в тексті.

Вагові коефіцієнти міри ентропії та IDF [75-79] дають помірні або високі ваги для термів, які з'являються з помірною або високою частотою, але в невеликій кількості документів.

Ваги коефіцієнти міри ентропії та IDF змінюються обернено до кількості документів, у яких з'являється терм.

Міру ентропії краще використовувати для аналізу невеликих документів, що містять лише кілька речень.

Міра ентропії – це єдина вага терма, яка залежить від розподілу термів між документами.

При обробці текстових даних, що представлені частотними матрицями термів, що описують корпуси текстів використовується метод сингулярного розкладу (SVD) [81], з метою зменшення розмірів масивів даних, що обробляються.

Нехай:

A – частотна матриця

k – розмірність перетворення.

Чим більше k тим краще виконується апроксимація частотної матриці A.

Рекомендується використовувати k, що дорівнює від 2 до 50 для невеликих колекцій документів і від 30 до 200 для великих.

Сингулярне розкладання матриці A дозволяє представити у вигляді добутку матриць U, S і V [79-81].

$$A = USV^T$$

де S – діагональна матриця містить сингулярні значення, U і V – ортогональні матриці, тобто містять ортогональні стовпці (лівий і правий сингулярні вектори).

Після того як виконано сингулярний розклад, кожен стовпець – документ частотної матриці може бути спроєктований на перші k стовпців матриці U. Математично таке проєктування являє собою k-мірний підпростір, який найкраще описує вхідний набір даних. Проєктування одного стовпчика матриці A (одного документа) дозволяє представити документ у вигляді k обмежуючих концептів.

Іншими словами, колекція документів проецюється в k-мірний простір, в якому один вимір відповідає відповідному концепту.

Подібним чином кожен рядок матриці A може бути зпроецьованний на k стовпців матриці V.

Метод сингулярного розкладу (SVD) [83] демонструє високу ефективність у вирішенні ключових завдань текстової аналітики та роботи з розрідженими частотними матрицями. По-перше, SVD забезпечує надійне зменшення розмірності даних, зберігаючи найважливіші латентні концепти та усуваючи надлишкову кореляцію між ознаками. По-друге, апроксимація низьким рангом пом’якшує проблему шуму й мультиколінеарності, що особливо критично під час побудови економетричних і машинних моделей прогнозування видатків. По-третє, SVD як базовий крок попередньої обробки значно прискорює навчання алгоритмів і підвищує їх стійкість до перенавчання.

Сингулярний розклад широко застосовуватися у таких напрямках як динамічний моніторинг та аналіз трендів. Інкрементальне (online) SVD дозволить оновлювати латентні простори в реальному часі за надходженням нових текстових повідомлень від громадян та змін нормативної бази.

Імпутація та реконструкція даних. У панельних та часових рядах соціальних виплат SVD-базована імпутація допомагає відновити відсутні спостереження, покращуючи якість прогнозів [83].

Гібридні економетрико-машинні підходи. Інтеграція SVD-пре-процесінгу з ARIMAX/SARIMAX та XGBoost [84] із екзогенними факторами дозволить будувати більш гнучкі та інтерпретовані моделі для оцінки впливу

макро-, демографічних та конфліктних змінних. SVD закладає фундамент для побудови високоефективних, масштабованих і адаптивних систем аналізу та прогнозування видатків на соціальне забезпечення в умовах військового конфлікту, поєднуючи класичні методи лінійної алгебри з сучасними підходами машинного навчання.

2.3 Застосування текстової аналітики для аналізу звернень громадян

Розбудова Єдиної інформаційної системи соціальної сфери [85-86] передбачає створення єдиного інформаційно-довідкового середовища для отримувачів соціальної підтримки. Важливе місце в якому посідає підсистема роботи із зверненнями громадян, адже тільки за січень-вересень 2024 року, Пенсійним фондом України було зареєстровано 504856 звернень громадян з питань, з них 229537 (або 45,5 відсотки) – електронні звернення [87].

Тому, питання розроблення методик, моделей, інформаційних технологій аналізу текстової інформації з електронних звернень громадян до установ соціального захисту та соціального забезпечення, Інтернет-джерел, виявлення питань, що є найбільш важливими для тих, хто потребує підтримки держави, є актуальним, має практичне значення [5].

В ході дослідження розглядалась практична задача визначення потреби у соціальному захисті та соціальному забезпечення мешканців різних регіонів України та біженців. Для аналізу текстової інформації використано інструменти SAS Text Miner [81].

Вхідною інформацією є електронні звернення громадян, які надійшли до вебпорталу електронних послуг Пенсійного фонду України та державної установи “Урядовий контактний центр [87-88]. Були досліджені й матеріали Інтернет-видань, різних за тематикою та аудиторією, державних та недержавних, з яких було відібрано 162 (назви джерел та посилання на них представлені у табл. 2.3.

Таблиця 2.3 - Перелік Інтернет—джерел, інформацію з яких використано для аналізу

Номер	Назва джерела	Адреса ресурсу	Кількість текстів
1	УкрІнформ	https://www.ukrinform.ua/rubric-society	50
2	Суспільне. Новини	https://suspilne.media	25
3	Сайт міжнародного наукового видання «Фінансово-кредитна діяльність: проблеми теорії та практики»	https://fkd.net.ua	7
4	Газета «Урядовий кур'єр» — офіційне друковане видання Кабінету Міністрів України.	https://ukurier.gov.ua/uk/articles	30
5	Офіційний сайт Київської обласної ради професійних спілок	http://korps.com.ua	5
6	Офіційний сайт Національного банку України	https://knpf.bank.gov.ua	10
7	Офіційний сайт журналу «Forbes Ukraine»	https://forbes.ua	15
8	Сайт електронного видання «Судово-юридична газета»	https://sud.ua	20

На основі аналізу текстів, що стосуються питань соціального захисту та соціального забезпечення, розміщених на вказаних Інтернет-ресурсах та у електронних зверненнях, було отримано шість кластерів.

До першого кластеру увійшли тексти, які містять питання, пов'язані з пенсійною реформою. Найбільш характерними для цього кластеру виявились слова та словосполучення: "реформа", "страхові виплати", "страховий стаж", "обов'язкові пенсійні накопичення".

До другого кластеру увійшли слова та словосполучення, що описують питання нарахування та виплати пенсій та соціальних допомог Пенсійним фондом України: "своєчасна виплати пенсій" "добровільні внески на пенсійне страхування", "мінімальна пенсія", "індексація пенсій", «підвищення пенсій», "житлова субсидія", "фінансування поточних виплат", "перерахунок пенсій працюючим пенсіонерам".

Третій кластер узагальнює проблеми соціального захисту внутрішньо переміщених осіб. Найбільш характерними є такі слова та словосполучення, як “ВПО”, “ідентифікація”, “звільнені території”, “виплати переселенцям”, “мешканці окупованого Криму”, “Всесвітня продовольча програма ООН” “тимчасово непідконтрольні території”.

До четвертого кластеру увійшли слова та словосполучення, що описують проблеми, пов’язані з втратами, внаслідок військового конфлікту: “військовослужбовець”, “поліцейський”, “зона бойових дій”, “зниклий безвісти”, “втрата годувальника”, “члени родини загиблого”.

Для п’ятого кластеру «актуальними» є питання соціального захисту та соціального забезпечення біженців, зокрема, “пенсія за кордоном”, “праця за межами України”, “пропорційне обчислення страхового стажу”, “страховий стаж одержаний в інших країнах”.

У шостому кластері узагальнені питання, що стосуються постраждалих внаслідок аварії на ЧАЕС: “аварія”, “ЧАЕС”, “чорнобилець”.

На основі попереднього аналізу текстів звернень сформовано корпус текстів, фрагмент якого наведений в табл. 2.4

Таблиця 2.4 - Частотна матриця термів для корпусу текстів, побудована на основі корпусу текстів, сформованого із електронних звернень громадян

Позначення	Терм	Кількість згадувань у документі:									
		d1	d2	d3	d4	d5	d6	d7	d8	d9	d10
t1	суд	1	0	0	0	0	0	1	2	0	0
t2	надбавки	1	0	1	1	0	0	1	0	2	0
t3	військово-службовці	0	1	0	0	2	1	0	0	0	0
t4	грошове забезпечення	0	1	0	0	1	0	0	0	2	0
t5	пенсія	0	1	0	1	2	2	1	0	1	1
t6	правоохоронців	0	1	0	0	1	0	0	0	0	0
t7	колишніх	0	1	0	0	1	0	0	0	0	0
t8	аварія	0	0	1	1	0	0	0	0	0	1
t9	ЧАЕС	0	0	1	2	0	0	0	0	0	1
t10	Україна	0	0	1	0	0	1	0	0	0	0
t11	отримувала	0	0	0	1	1	0	0	0	0	0
t12	вислуга	0	0	0	0	1	1	0	0	0	0

Для вирішення проблеми зменшення розмірності та розрідженості частотної матриці корпусу текстів було використано метод сингулярного розподілу (SVD) [82-84]. Адже зазвичай, в документах використовується досить невеликий набір термів, які описують певну предметну область. Отже, якщо в діагональній матриці сингулярних значень (S) залишити рівно k перших діагональних елементів, а решті присвоїти значення нуль, то використання методу SVD дає оптимальне наближення. В діагональній матриці сингулярних значень S, значення впорядковані, а саме $s_1 \geq s_2 \geq \ ... \ \geq s_k$, тобто якщо залишити перші два значення то іншим присвоїти значення нуль. На основі отриманої матриці S можна обчислити, яку кількість у процентах вносить в пояснення даних вимір, що описується відповідним сингулярним значенням.

На основі отриманої матриці S можна обчислити яку кількість у процентах вносить в пояснення даних відповідний вимір, що описується відповідним сингулярним значенням. (табл. 2.5). Значення стовпчика “Процент вкладу значення в пояснення варіабельності даних” обчислюється як значення “Квадрат сингулярного значення” розділене на суму значень квадратів сингулярних значень, помножене на 100%.

Як видно з отриманих результатів, табл. 3, якщо залишити лише два базових виміри, то загалом буде пояснено 66,16% варіабельності даних.

Таблиця 2.5 - Аналіз отриманих сингулярних значень

Номер виміру	Сингулярне значення	Квадрат сингулярного значення	Процент вкладу значення в пояснення варіабельності даних	Кумулятивне значення проценту вкладу
1	5,1435	26,45	45,61	45,61
2	3,4526	11,92	20,55	66,16
3	2,7696	7,67	13,23	79,38
4	2,3736	5,63	9,71	89,11
5	1,7711	3,13	5,41	94,51
6	1,2251	1,5008	2,58	97,09
7	1,029	1,0588	1,82	98,92
8	0,684	0,4678	0,81	99,73
9	0,371	0,1376	0,23	99,96
10	0,1352	0,0182	0,03	100

В такому випадку всі документи можна розташувати у двовимірному просторі і визначити кластери, які вони утворюють за ступенем подібності та приналежності до певної теми (рис. 2.1).

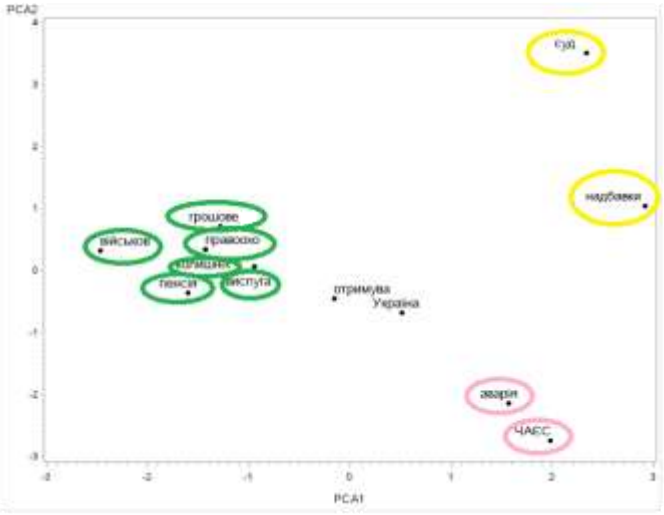

Рисунок 2.1 - Розташування термів у двовимірному просторі

Як видно з рис. 2.1, перший вимір пояснює 45,61% варіабельності даних, другий вимір пояснює 20,55% варіабельності даних. В результаті сформовано три тематичних кластери, до яких увійшли документи за подібністю використання термів [81-84].

В рамках даного дослідження було використано систему SAS Text Miner, яка, будує технологічний проект у наступну послідовність кроків.

1. Завантаження даних.
2. Парсинг текстів.
3. Фільтрація текстів.
4. Кластеризація текстів.

Технологічний процес аналізу корпусу текстів, з метою їх кластеризації представлений на рис. 2.2

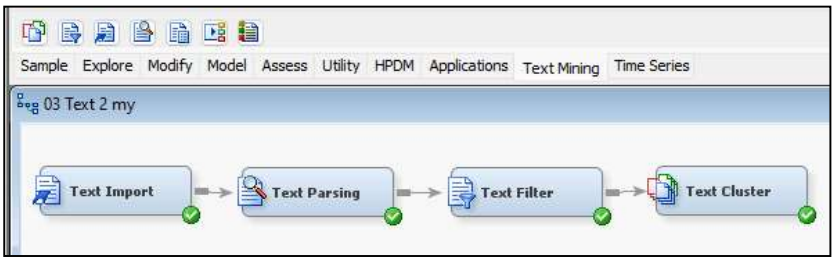

Рисунок 2.2 - Технологічний процес аналізу корпусу текстів в системі SAS Text Miner

Побудовані правила, для відповідних кластерів, сгенеровані у вигляді наступного коду програми:

```
F_TextCluster_cluster_ =1 ::
(OR
, "реформа"
, “страхові”
, (AND, (OR, " виплати", "стаж") )
, “накопичення”
, (AND, (OR, "пенсійні", “обов’язкові”) )

F_TextCluster_cluster_ =2 ::
(OR
, “добровільні”
, (AND, (OR, "виплати" , "пенсій"))
, “своєчасна”
, (AND, (OR, "внески" , "пенсійне" , "страхування", “перерахунок”))
, “пенсія”
, (AND, (OR, "мінімальна" , "індексація" , "підвищення"))
, “субсидія”
, (AND, (OR, "житлова"))
, “поточні”
, (AND, (OR, "виплата" , "фінансування"))

F_TextCluster_cluster_ =3 ::
(OR
, “ідентифікація”
, (AND, (OR, "ВПО" , "переселенець". “виплати”))
, “мешканець”
, (AND, (OR, "Крим" , "непідконтрольна" , "територія" , "тимчасово"))
, “ООН”
, (AND, (OR, "всесвітня" , "продовольча" , "програма"))

F_TextCluster_cluster_ =4 ::
(OR
, (AND, (OR, "військовослужбовець" , "військовий", “поліцейський”))
, “зона”
, (AND, (OR, "бойових" , "дій"))
, (AND, (OR, "зниклий" , "безвісти"))
, “загиблий”
, (AND, (OR, "втрата" , "годувальника" , "члени" , "родини"))

F_TextCluster_cluster_ =5 ::
(OR
, “пенсія”
, (AND, (OR, "кордоном" , "межами" , “інших” , “країнах”))
, “стаж”
, (AND, (OR, "обчислення" , "страховий" , “пропорційне”))

F_TextCluster_cluster_ =6 ::
(OR
, “аварія”
```

, (AND, (OR, "ЧАЄС" , "атомна" , “електростанція”))
, “чорнобилець”)))

Статистичні характеристики побудованої моделі класифікації на основі лінгвістичних правил, було обчислено окремо для тренінгового та тестового наборів даних: співвідношення - 70% для тренінгу та 30% для тестування, тобто 114 та 48 текстів відповідно.

Результати узагальнені в таблиці 2.6.

Таблиця 2.6 - Статистичні характеристики моделі класифікації досліджуваних текстів

Статистика	Набір даних	
	тренінговий	тестовий
TP (True Positive)	30	11
TN (True Negative)	67	26
FP (false positive)	10	6
FN (false Negative)	7	5
MISC,% (частка неправильно класифікованих значень)	15	23
Gini	0,82	0,71
ROC	0,79	0,67

Зображення ROC-кривої для моделі класифікації текстової інформації на основі лінгвістичних правил представлене на рис. 2.3.

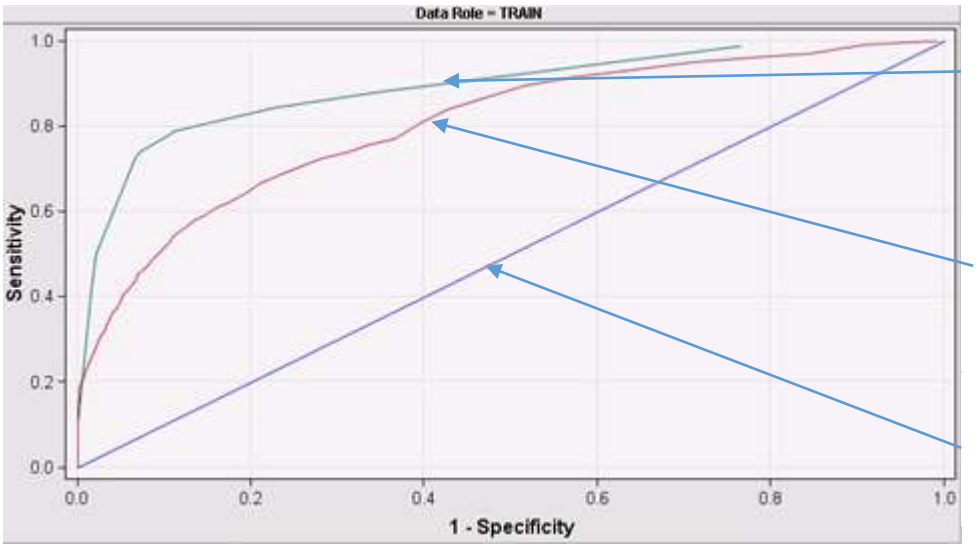

Рисунок 2.3 – ROC-крива для побудованої моделі класифікації на основі лінгвістичних правил

Побудовані лінгвістичні правила було використано для кластеризації текстів новин, що були опубліковані в мережі Інтернет з вересня 2023 року по вересень 2024 року. Загалом, було вивантажено та оброблено більше 10 тис. текстів за тематикою соціального захисту та соціального забезпечення українців.

Після кластеризації текстів, для кожного кластеру було розраховано кількість текстів, які належать дописувачам з певного регіону. Отримані значення було нормовано по шкалі від 0 до 100 за формулою (2.1):

$$popularity_i = \frac{n_i}{\max(n_i \mid \forall\, i)} \qquad (2.1)$$

де popularity_i – популярність текстів відповідного кластеру, для і-го регіону, n_i – кількість текстів по регіону, max(n_i |∀ i) – максимальна кількість текстів за всіма регіонами

Результати розрахунків представлені у таблиці 2.7.

Пропонована методика аналізу текстової інформації із застосування засобів text mining призначена для автоматизованої обробки значних обсягів текстів певної тематики. Використання засобів текстової аналітики дозволяє поглибити знання про предметну область за допомогою використання неструктурованих даних.

В даному дослідженні проблему розмірності та розрідженості частотної матриці корпусу текстів вирішено за допомогою ключової теореми з лінійної алгебри - методу сингулярного розкладу матриці (SVD).

Попередньо виконано. операцію зважування значень частот, яка допомогла частково вирішити проблему нерівності високочастотних членів, роблячи їх менш впливовими.

Це дало можливість отримати результати класифікації текстової інформації іисокої якості.

Таблиця 2.7 - Результати кластерного аналізу текстової інформації з питань соціального захисту та соціального забезпечення по регіонах України

Назва регіону	Популярність текстів відповідного кластеру					
	Кластер 1 (пенсійна реформа)	Кластер 2 (нарахування та виплата пенсій та соціальних допомог Пенсійним фондом України)	Кластер 3 (проблеми соціального захисту внутрішньо переміщених осіб)	Кластер 4 (проблеми, пов'язані з втратами, внаслідок військового конфлікту)	Кластер 5 (питання соціального захисту та соціального забезпечення біженців)	Кластер 6. (питання, що стосуються постраждалих внаслідок аварії на ЧАЕС)
Вінницька область	94	65	24	72	79	
Волинська область	87	57	20	100	63	
місто Київ	82	49	32	37	26	33
місто Севастополь	-	-	-	-	-	-
Дніпропетровська область	58	39	43	33	14	
Донецька область	27	32	59	37		
Житомирська область	94	73	19	34	62	88
Закарпатська область	67	45	29	40	75	
Запорізька область	58	39	90	30		
Івано-Франківська область	87	66	24	63	72	
Київська область	84	42	28	37	37	100
Кіровоградська область	92	88	32	73	46	
АР Крим	-	1	1	-	-	-
Луганська область		22	8			
Львівська область	73	45	20	60	57	
Миколаївська область	76	70	64	47	18	
Одеська область	32	24	27	13	13	
Полтавська область	75	63	32	73	42	77
Рівненська область	100	64	17	81	100	
Сумська область	92	100	52	43	30	
Тернопільська область	50	56	24		63	
Харківська область	47	35	100	15	9	
Херсонська область	71	62	89			
Хмельницька область	87	55	28	78	73	
Черкаська область	87	47	30	50	55	74
Чернігівська область	81	58	24	50	29	
Чернівецька область	83	31	25	1	61	

Результати аналізу, представлені в таблиці можна візуалізувати використовуючи інструменти SAS Enterprise Guide 7.1 (рис. 2.4—2.9).

Рисунок 2.4 - Кластер 1 - популярності текстів по тематиці “Пенсійна реформа” по регіонах України.

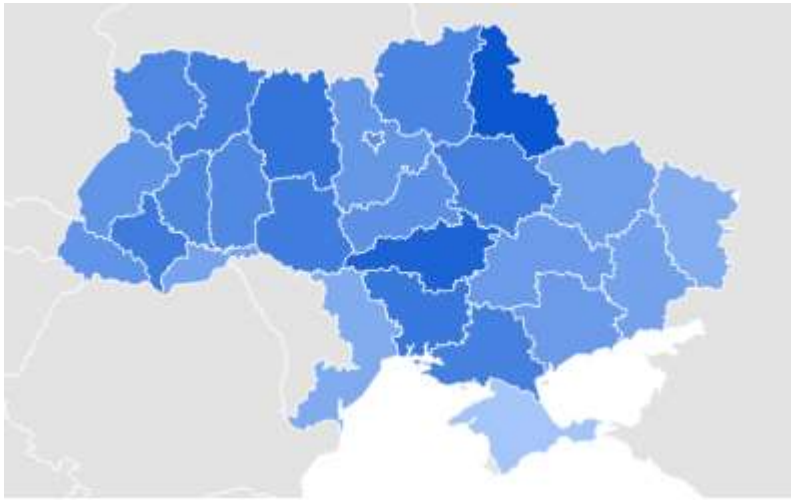

Рисунок 2.5 - Кластер 2, популярності текстів по тематиці “Питання пов’язані із пенсійним фондом взагалі” по регіонах України.

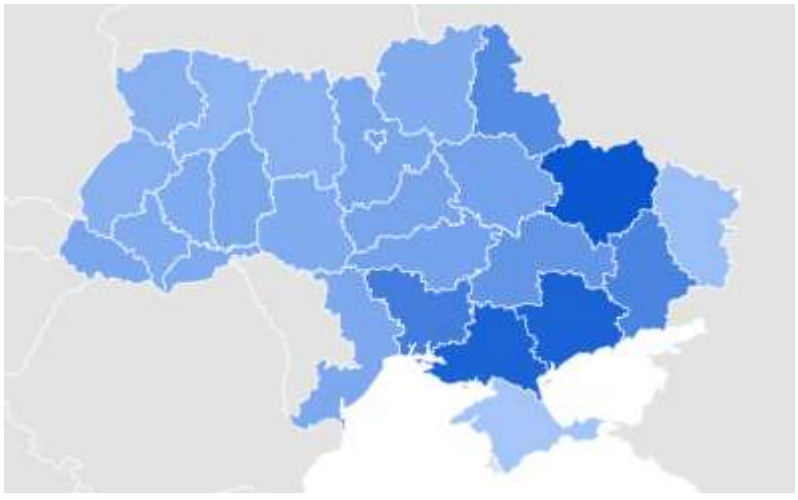

Рисунок 2.6 - Кластер 3, популярності текстів по тематиці “Проблеми пов’язані з ВПО” по регіонах України.

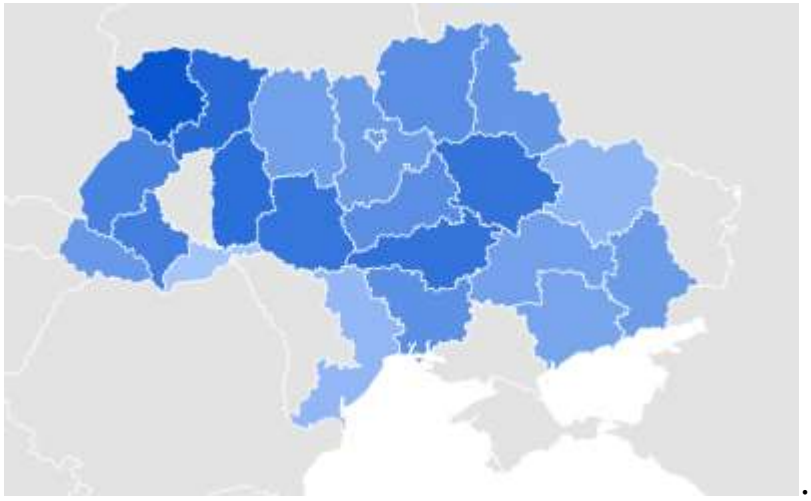 .

Рисунок 2.7 - Кластер 4, популярності текстів по тематиці “Питання пов’язані із військовими та поліцейськими” по регіонах України.

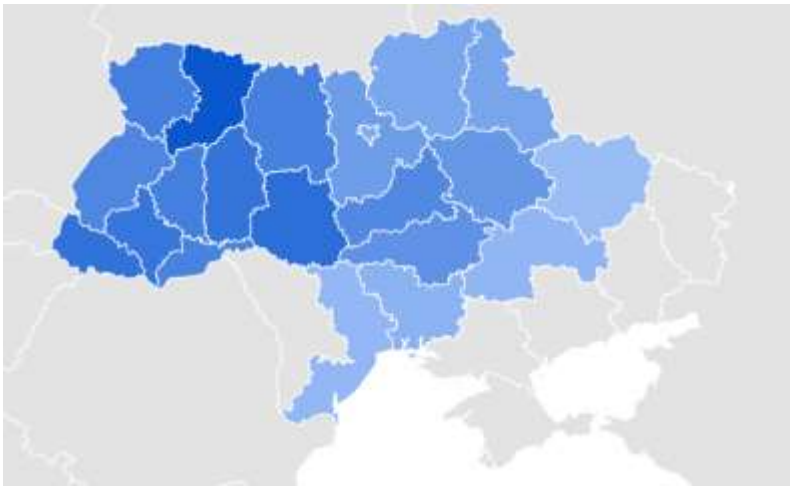

Рисунок 2.8 --Кластер 5, популярності текстів по тематиці “Питання щодо виплати пенсій за кордоном” по регіонах України.

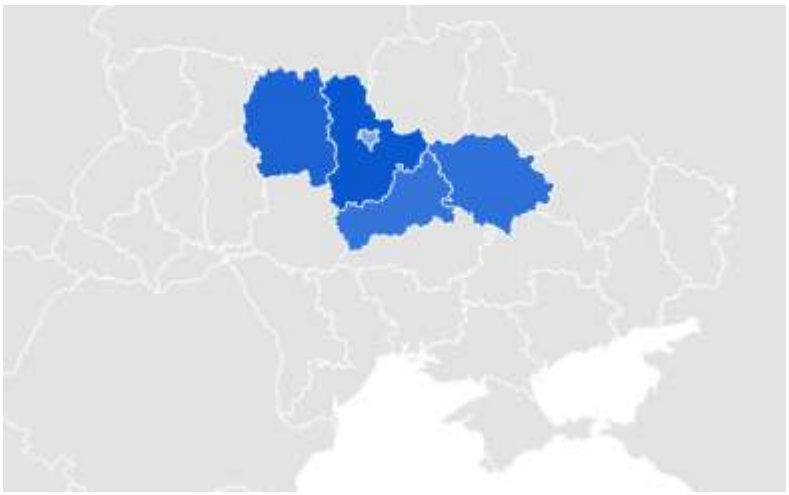

Рисунок 2.9 - Кластер 6, популярності текстів по тематиці “Питання пов’язані з пенсіями для постраждалих від наслідків аварії на ЧАЄС” по регіонах України.

Отже, використання інтелектуального аналізу великих обсягів текстових даних, дозволяє виявити найбільш вагомі проблеми, які потребують першочергового вирішення, з'ясувати, для яких категорій населення вони найбільш актуальні. Отримані результати можуть бути в подальшому використані під час планування соціальних видатків бюджетів різних рівнів, у моделі актуарних розрахунків, під час планування соціальних видатків бюджетів різних рівнів. Пропонований підхід може забезпечувати покращення якості прогнозів в сучасних умовах, коли немає повної інформації про досліджуваний процес чи явище або інформація спотворена.

Висновки до розділу 2

У розділі представлено результати аналізу існуючих методів та моделей збору та оброблення даних, на основі яких розроблено методику формування консолідованих даних, особливістю якої є інтегроване використання платформи Oracle та програмного забезпечення компанії SAS Institute.

Запропонована інформаційна технологія збору та оброблення даних, призначена для функціонування у складі Єдиної інформаційної системи соціальної сфери, яка функціонує на програмно-технічній платформі, що має трирівневу архітектуру, побудовану на основі системи керування базами даних Oracle Database Enterprise Edition 12, технології ASP.NET та веб-клієнта, для збору та обробки великих обсягів структурованої та неструктурованої інформації використовується Oracle Business Intelligence (Oracle BI).

Для порівняння продуктивності ключових інструментів збору даних — веб-скрейпінгу (BeautifulSoup, Scrapy, Selenium) та API-запитів (REST/OData до data.gov.ua і openbudget.gov.ua) в умовах реального навантаження було виконано тестування. Для оброблення текстових даних, у роботі представлена методика, яка базується на методах лінгвістичного аналізу, кластеризації та інтелектуального аналізу даних. Вона забезпечує можливість виділення ключових тем і проблемних аспектів у зверненнях громадян, що суттєво

підвищує релевантність і оперативність управлінських рішень. Розроблені методи текстової аналітики та кластеризації звернень населення дозволяють оперативно виявляти пріоритетні проблеми та потреби різних соціальних груп, що сприяє ухваленню обґрунтованих управлінських рішень.

Для розроблення аналітичної складової обрано програмне середовище SAS Institute (мова програмування SAS Base), яке добре інтегрується із апаратно-програмною платформою Oracle, і за своїми характеристиками є оптимальним для розв’язання аналітичних задач, має програмний інтерфейс SWAT API який дозволяє використовувати всі переваги платформи SAS Viya та мов програмування Open Source, зокрема, R та Python. Дані для аналізу використовуються не лише з аналітичного сховища (застосована інтеграція Hadoop та ETL/ELT), а й в результаті виконання API-запитів, в тому числі, коли потрібно здійснювати обробку у режимі реального часу (використано SAS Event Stream Processing). Крім того, використання функціоналу інструменту SAS Data Quality дозволить покращити якість вхідних даних, зокрема здійснити їх профілювання, перевірку, очищення та стандартизацію, що в свою чергу підвищить якість побудованих моделей. Гнучкість та адаптивність інформаційних технологій досягається за рахунок використання конвергентного підходу, який передбачає, що розроблювана система має центр обробки даних (Big Data), систему збору, накопичення, обробки та завантаження даних у сховище або хмару, програмні засоби для математичного моделювання, інтелектуального аналізу даних, засоби штучного інтелекту, систему обміну даними між компонентами Єдиної інформаційної системи соціальної сфери України, взаємодії між користувачами різних рівнів та забезпечення надійності віддаленого доступу, систему інформаційної та кібербезпеки, зокрема, захисту персональних даних, систему моніторингу та адміністрування роботи системи, тощо..

РОЗДІЛ 3
ОСОБЛИВОСТІ МОДЕЛЮВАННЯ ВИТРАТ НА СОЦІАЛЬНИЙ ЗАХИСТ ТА СОЦІАЛЬНЕ ЗАБЕЗПЕЧЕННЯ

3.1 Застосування регресійних моделей

У сучасних умовах постійних військово-політичних змін та економічної нестабільності роль економетричних методів у прогнозуванні соціальних видатків стає надзвичайно актуальною. З одного боку, збройний конфлікт в Україні призводить до невпинного зростання потреб у підтримці внутрішньо переміщених осіб, ветеранів, багатодітних родин та малозабезпечених верств населення. З іншого — економічні показники (ВВП, інфляція, бюджетні надходження) демонструють високу волатильність, що ускладнює бюджетне планування.

Економетричні моделі, які спираються на обробку великих масивів даних і статистичні припущення, дозволяють системно оцінити, як саме макроекономічні чинники (наприклад, темпи інфляції чи динаміка ВВП), демографічні зміни (зростання числа ВПО, старіння населення) та конфліктні події (кількість бойових інцидентів, обсяги гуманітарної допомоги) впливають на обсяги соціальних виплат у кожному регіоні. Така аналітика є критично важливою для органів влади: розуміючи детермінанти зростання видатків, можна розробляти більш точні сценарії бюджетного навантаження та вчасно коригувати програму соціальної підтримки.

Крім того, економетричні підходи дають змогу формувати інтерпретовані прогнози з визначенням ступеня невизначеності (наприклад, довірчі інтервали), що є важливим у ризик-орієнтованому управлінні. За допомогою спеціальних тестів (перевірка на гетероскедастичність, автокореляцію залишків) ці моделі можна адаптувати до умов, коли дані часто змінюються, а частина інформації може бути неповною. Таким чином, використання економетричних методів стає не лише академічним

інструментом, а необхідною практичною основою для розробки стійкої й оперативної політики соціальної підтримки в умовах воєнного часу.

Розглядається покрокове застосування лінійної регресії для моделювання та прогнозування соціальних видатків у регіональному або національному розрізі. Наведемо деталі формулювання моделі, процедуру оцінювання параметрів, перевірку діагностичних припущень, приклади інтерпретації результатів і можливі обмеження.

Щоб отримати достовірний прогноз, необхідно чітко визначити таргетну змінну. У нашому випадку це реальний обсяг соціальних виплат на душу населення, який слугує кінцевим показником успішності моделі та дозволяє кількісно вимірювати вплив усіх обраних факторів. У випадку, якщо потрібно порівнювати одиниці спостереження з різною кількістю населення, цей показник зазвичай нормують на душу населення. При цьому номінальні величини попередньо переводять у реальні за допомогою індексу споживчих цін, щоб усунути вплив інфляції [89-94].

Загальна форма багатофакторної лінійної регресії [89-94]:

$$Y_{i,t} = \beta_0 + \beta_1 X_{1,i,t} + \beta_2 X_{2,i,t} + \cdots + \beta_k X_{k,i,t} + \varepsilon_{i,t},$$

де $Y_{i,t}$ – обсяг соціальних виплат (грн на душу) у регіоні i за квартал t,

- $X_{j,i,t}$ – значення j-ої пояснювальної змінної для регіону i/кварталу t,
- β_0 – константа (перетин з віссю Y),
- $\beta_1, \ldots, \beta_k$ – коефіцієнти, що показують вплив кожного фактора на Y,
- $\varepsilon_{i,t}$ – випадкова помилка (залишок), яка припускається нормально розподіленою з математичним сподіванням 0.

Пояснювальні (незалежні) змінні включають макроекономічні показники (ВВП на душу, інфляція, середня зарплата, бюджетні надходження), демографічні параметри (чисельність населення, частка пенсіонерів, кількість ВПО, безробіття), показники конфліктного середовища

(кількість бойових подій, обсяг гуманітарної допомоги) та політичні рішення (суми субсидій, компенсаційні програми, законодавчі зміни), які разом описують економічний, соціальний і політичний контекст формування соціальних видатків. При цьому до різних категорій можуть відноситись наступні параметри та індікатори (табл. 3.1).

Таблиця 3.1 – Опис змінних моделі

Макроекономічні індикатори	X1: ВВП на душу населення (реалізований на квартальній основі, грн) X2: Середня заробітна плата (грн) X3: Індекс інфляції за квартал (%) X4: Бюджетні доходи регіону (реальні, млн грн)
Демографічні параметри	X5: Загальна чисельність населення (тис. осіб) X6: Частка пенсіонерів у загальній чисельності (%) X7: Кількість внутрішньо переміщених осіб (ВПО, тис. осіб) X8: Рівень безробіття (%)
Конфліктні та гуманітарні чинники	X9: Кількість зареєстрованих бойових подій у регіоні за квартал (кількість) X10: Обсяг міжнародної гуманітарної допомоги (млн грн)
Політичні та бюджетні рішення	X11: Субсидії з державного бюджету на соцвиплати (млн грн) X12: Додаткові компенсаційні програми (індикатор: 1 – діє, 0 – не діє) X13: Зміни законодавства (dummy: 1 – зміна під час кварталу, 0 – без змін)

На практиці вікно спостережень може становити, наприклад, 4 роки щоквартальних даних (n=16 спостережень) для кожного з 27 регіонів (усього N=16×27=432 точки), або 8 років (n=32) залежно від доступності даних.

Необхідність чіткого визначення специфікації моделі та загального рівняння полягає в тому, що лише формалізована математична структура дозволяє кількісно оцінити, які фактори (економічні, демографічні, конфліктні) і наскільки впливають на обсяг соціальних виплат. Без однозначного рівняння неможливо порівняти результати між регіонами чи періодами, провести статистичну перевірку значущості коефіцієнтів та гарантувати коректність висновків.

Приклад: Нехай є фактори:

- X1=ВВП на душу,
- X2=рівень безробіття,
- X3=кількість внутрішньопереміщених осіб,
- X4=кількість атак,
- X5=субсидії з держбюджету.

Тоді модель матиме вигляд:

$$\text{SocialExp}_{i,t} = \beta_0 + \beta_1\,\text{GDPpc}_{i,t} + \beta_2\,\text{Unemp}_{i,t} + \beta_3\,\text{VPO}_{i,t} + \beta_4\,\text{Battles}_{i,t} + \beta_5\,\text{Subsidies}_{i,t} + \varepsilon_{i,t}.$$

Оцінювання параметрів відбувається з використанням методу найменших квадратів (МНК). МНК – це класичний спосіб оцінювання параметрів лінійної регресії, який полягає в мінімізації суми квадратів залишків між фактичними та прогнозними значеннями залежної змінної. Стандартний підхід: мінімізуються суми квадратів залишків [89-94]

$$\min_{\beta_0,\ldots,\beta_k} \sum_{i=1}^{N}\sum_{t=1}^{n}\Big(Y_{i,t} - \beta_0 - \sum_{j=1}^{k}\beta_j X_{j,i,t}\Big)^2.$$

За звичайних припущень (нормальність помилок, гомоскедастичність, відсутність автокореляції) оцінки β^j є несмещеними, ефективними (MaxL) та найменш дисперсійними в класі лінійних незміщених оцінок (BLUE – Best Linear Unbiased Estimator).

Іншим прикладом застосування регресійних моделей є спрощений приклад, коли розглядається панель із даних трьох регіонів за 8 кварталів (усього 24 спостереження). Змінні, використані в моделі:

- Yi,t: реальні витрати на соціальний захист та соціальне забезпечення на одну особу (грн/особа)
- X1,i,t: ВВП на душу населення(тис. грн)
- X2,i,t: рівень безробіття (%)
- X3,i,t: кількість внутрішньопереміщених осіб (тис. осіб)
- X4,i,t: кількість атак (одиниць)

В результаті побудови регресійної моделі (не досліджуючи гетероскедастичність) і отримаємо оцінки (умовні цифри):

Таблиця 3.2 - Статистичні характеристики моделей

Коефіцієнт	**Оцінка β**	**Стандартна помилка**	**t-статистика**	**p-значення**
β0	120.5	15.2	7.93	<0.001
β1 (GDP)	0.85	0.12	7.08	<0.001
β2 (Unemp)	–1.20	0.30	–4.00	0.0002
β3 (VPO)	2.10	0.45	4.67	<0.0001
β4 (Battles)	0.95	0.25	3.80	0.0005

В результаті отримуємо R^2=0.78, скоригований R^2=0.75.

Інтерпретація результатів побудови такої моделі наступна:

β_1=0.85: за збільшення ВВП на душу на 1 тис. грн соціальні виплати зростають у середньому на 0,85 грн/особу.

β_2=−1.20: зростання рівня безробіття на 1 п.п. призводить до зменшення соцвиплат на 1,20 грн/особу (можливо, через зростання потреби в інших соціальних програмах, що змінюють структуру виплат).

β_3=2.10: кожне додаткове 1 тис. ВПО асоціюється зі зростанням соцвиплат на 2,10 грн/особу (через збільшення кількості одержувачів виплат).

β_4=0.95: кожна додаткова бойова подія призводить у середньому до збільшення соцвиплат на 0,95 грн/особу (наприклад, через підвищене навантаження на місцеві органи соцзахисту та зростання допомоги постраждалим).

Усі коефіцієнти є статистично значущими (p<0.001). Високе значення R^2 свідчить, що модель пояснює 78 % дисперсії залежної змінної.

Додатково проводиться перевірка гетероскедастичності (Breusch–Pagan). Статистика тесту: $\chi^2 = 10.5$, p=0.03. Отже, наявна гетероскедастичність. Після White-робастної корекції стандартних помилок змінюються лише їхні значення, але t-статистика та інтерпретація коефіцієнтів залишаються подібними.

Перевірка автокореляції (Durbin–Watson): Отримане значення DW = 1.20 (< 1.5) свідчить про наявність позитивної автокореляції залишків. У такому разі доцільно додати лаг залежної змінної (наприклад, Yi,t−1 або перейти до методу AR(1) у специфікації (який оцінюють через GLS) [95-101].

Лінійна регресія є базовим і зрозумілим інструментом для кількісної оцінки впливу різних факторів на соціальні виплати. Важливою складовою побудови моделі є діагностичні тести (VIF, Durbin–Watson, Breusch–Pagan, Shapiro–Wilk), які допомагають виявити можливі порушення базових економетричних припущень і відповідно скоригувати оцінювання. У разі виявлення гетероскедастичності застосовують робастні стандартні помилки або метод FGLS, а фіксовані ефекти панельних даних дозволяють врахувати незмінні в часі особливості окремих одиниць спостереження. До обмежень лінійної регресії належать можливі нелінійні ефекти, нестаціонарність даних і раптові шоки, тому для їхнього зменшення в модель можуть додаватися поліноміальні терміни, лаги, dummy-змінні шокового характеру або застосовувати rolling window підхід. Незважаючи на ці недоліки, лінійна регресія залишається відправною точкою: вона дає змогу швидко зрозуміти основні взаємозв’язки між економічними, демографічними та конфліктними факторами при прогнозуванні соціальних видатків та слугує основою для подальшого ускладнення моделей (ARIMAX, гібриди з ML, SEM тощо).

Моделі авторегресії (AR) і авторегресії з рухомим середнім (ARMA) — це класичні підходи для моделювання й прогнозування часових рядів, у яких поточне значення залежить від його минулих значень і, у разі ARMA, від минулих похибок. Нижче наведено короткий опис кожної моделі та приклади їх побудови й оцінювання на прикладі синтетичного ряду в Python [73].

Модель авторегресії порядку ppp (AR(p)) описується рівнянням [101]:

$$x_t = c + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + \varepsilon_t,$$

де:

- x_t — значення ряду в момент t,
- c — константа (зсув/інтерсепт),
- $\phi_1, \ldots, \phi_p$ — коефіцієнти авторегресії,
- ε_t — випадкова похибка (white noise) з нульовим середнім.

У AR-моделі поточне значення лінійно залежить від ppp попередніх значень самого ряду. Наприклад, AR(1) (порядок p=1) має вигляд:

$$x_t = c + \phi_1 x_{t-1} + \varepsilon_t.$$

Застосувавши мову програмування Python, створимо штучний AR(2)-ряд і далі оцінюємо модель AR із порядком, підібраним за критерієм AIC.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from statsmodels.tsa.arima_process import ArmaProcess
from statsmodels.tsa.ar_model import AutoReg
from statsmodels.tsa.stattools import adfuller

# 1. Генерація синтетичного AR(2)-ряду
np.random.seed(42)
# коефіцієнти авторегресії AR(2): x_t = 0.6 x_{t-1} - 0.3 x_{t-2} + noise
ar_params = np.array([1, -0.6, 0.3])  # у форматі [1, -phi1, -phi2]
ma_params = np.array([1])             # немає MA-компоненти
arma_process = ArmaProcess(ar_params, ma_params)
# Генеруємо 500 точок
n_samples = 500
simulated_data = arma_process.generate_sample(nsample=n_samples)

# 2. Поміщаємо в Pandas Series із часовою індексацією
dates = pd.date_range(start="2020-01-01", periods=n_samples, freq="M")
ts_ar2 = pd.Series(simulated_data, index=dates)

# Перевірка стаціонарності (ADF-тест)
adf_result = adfuller(ts_ar2)
print("ADF statistic:", adf_result[0], "p-value:", adf_result[1])

# 3. Побудова й оцінка AR-моделі
# Підберемо порядок p за допомогою AutoReg з maxlag = 5
model = AutoReg(ts_ar2, lags=5, old_names=False).fit()
```

```
print(model.summary())

# 4. Прогноз на наступні 12 періодів
forecast = model.predict(start=len(ts_ar2), end=len(ts_ar2) + 11)

# 5. Візуалізація вихідного ряду і прогнозу
plt.figure(figsize=(10, 4))
plt.plot(ts_ar2, label="Синтетичний AR(2) ряд")
plt.plot(forecast.index, forecast.values, 'r--o', label="Прогноз AR")
plt.title("AR-модель: фактичні дані та прогноз")
plt.legend()
plt.show()
```

У наведеному прикладі було згенеровано AR(2)-ряд за заданими коефіцієнтами ϕ_1=0.6, ϕ_2=−0.3. Для перевірки стаціонарності ряду застосовано ADF-тест. Використано AutoReg із lags=5, щоб оцінити коефіцієнти; модель самостійно вибрала адекватний підпорядок серед 5 можливих лагів за AIC..

Модель ARMA(p, q) поєднує авторегресію порядку p і компонент з ковзним середнім порядку q () [99, 101]:

$$x_t = c + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} \; + \; \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} \; + \; \varepsilon_t,$$

де $\theta_1, \ldots, \theta_q$ — коефіцієнти MA (різниці).

Решта позначень такі ж, як для AR.

Компонент ковзної середньої дозволяє моделі враховувати вплив попередніх випадкових шоків $(\varepsilon_{t-1}, \ldots)$, що іноді підвищує якість прогнозу, особливо коли ряд на коротких ділянках демонструє «аварійні» коливання.

Нижче згенеруємо синтетичний ARMA(1,1)-ряд і підберемо модель ARMA за допомогою бібліотеки statsmodels.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from statsmodels.tsa.arima_process import ArmaProcess
from statsmodels.tsa.arima.model import ARIMA
from statsmodels.tsa.stattools import adfuller

# 1. Генерація синтетичного ARMA(1,1)-ряду
np.random.seed(123)
ar_params = np.array([1, -0.7])  # AR(1): phi1 = 0.7
ma_params = np.array([1, 0.5])   # MA(1): theta1 = 0.5
arma_process = ArmaProcess(ar_params, ma_params)
```

```
n_samples = 500
simulated_arma = arma_process.generate_sample(nsample=n_samples)

# 2. Створюємо Series із часовою індексацією
dates = pd.date_range(start="2019-01-01", periods=n_samples, freq="M")
ts_arma = pd.Series(simulated_arma, index=dates)

# Перевірка стаціонарності
adf_stat, adf_p, *_ = adfuller(ts_arma)
print("ADF statistic:", adf_stat, "p-value:", adf_p)

# 3. Оцінка ARMA(1,1) моделі через ARIMA (ARIMA(p,d,q) з d=0)
model_arma = ARIMA(ts_arma, order=(1, 0, 1)).fit()
print(model_arma.summary())

# 4. Прогноз для наступних 12 періодів
forecast = model_arma.get_forecast(steps=12)
mean_forecast = forecast.predicted_mean
conf_int = forecast.conf_int()

# 5. Візуалізація
plt.figure(figsize=(10, 4))
plt.plot(ts_arma, label="Синтетичний ARMA(1,1) ряд")
plt.plot(mean_forecast.index, mean_forecast.values, 'r--o', label="Прогноз ARMA")
plt.fill_between(conf_int.index,
                 conf_int.iloc[:, 0],
                 conf_int.iloc[:, 1], color='pink', alpha=0.3)
plt.title("ARMA(1,1) модель: фактичні дані та прогноз із довірчим інтервалом")
plt.legend()
plt.show()
```

У цьому прикладі було згенеровано ARMA(1,1)-ряд із коефіцієнтами $\phi_1=0.7$, $\theta_1=0.5$. Перевірено стаціонарність використовуючи ADF-тест. Оцінено модель ARMA(1,1) як ARIMA(order=(1,0,1)). Побудувано прогноз на 12 місяців з довірчим інтервалом 95%.

Важливим етапом при побудові моделей є вибір порядку моделей. Порядок p (lag-кількість у AR) і q (кількість MA-лагів) підбирають за інформаційними критеріями (AIC, BIC). Для швидшого знаходження початкових кандидатів можна аналізувати автокореляційну (ACF) і частково автокореляційну (PACF) функції. Після оцінювання моделі варто перевірити залишки моделі (εt\varepsilon_tεt) на відсутність автокореляцій — графік ACF залишків має не демонструвати значущих лагів, а самі залишки мають бути близькі до білої шумової послідовності.

Обидві моделі (AR, ARMA) вимагають стаціонарності ряду (стала середня та дисперсія в часі). Якщо ряд нестаціонарний, спочатку застосовують різницю (перший порядок диференціювання) — виходить ARIMA.

У дослідженні видатків на соціальне забезпечення часові ряди можуть мати сезонність (наприклад, різке зростання виплат узимку). Тому, часто використовують розширену модель SARIMA (з сезонним компонентом). Якщо ж прямих сезонних коливань немає, а провідну роль відіграє інерція витрат (тут і далі соціальні програми фінансуються з огляду на попередні витрати), AR(p) або ARMA(p,q), можуть дати достатньо точний базовий прогноз [99, 101].

Моделі AR і ARMA дозволяють формалізувати залежність поточних значень ряду від минулих спостережень і від попередніх помилок, що робить їх ефективними інструментами для прогнозування часових рядів соціальних витрат у контексті аналізу бюджетів і планування соціальних програм.

Моделі часових рядів дозволяють безпосередньо аналізувати залежність обсягу соціальних видатків від її історичних відліків, одночасно враховуючи трендові компоненти, сезонність і короткострокові флуктуації. Це особливо корисно, коли дані доступні у вигляді послідовних щомісячних чи щоквартальних значень соцвиплат, і потрібно зробити прогноз на найближчі періоди. Наводимо опис базових ARIMA-моделей, їх сезонних розширень (SARIMA), а також варіантів із екзогенними регресорами (ARIMAX, SARIMAX) [99, 101].

Авторегресійна інтегрована модель ковзного середнього (ARIMA) – це клас статистичних моделей для аналізу й прогнозування часових рядів. ARIMA (p, d, q) – має загальний вигляд із складовими, у якій:

AR(p) (autoregressive) - авторегресійна частина із p лагами залежної змінної,

I(d) (integrated) - d разів проводиться диференціювання першого порядку для приведення ряду до стану стаціонарності (видалення тренду чи нерівномірної дисперсії) [99-101],

MA(q) (moving average) - компонент скользячого середнього із q лагами помилки.

Математично ARIMA (p, d, q) для оригінального ряду Y_t (скажімо, сума витрат на соціальний захист та соціальне забезпечення, гривень на одну особу за місяць) записується через стаціонаризований ряд $X_t = \Delta^d Y_t$ як:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t,$$

де $\phi_1, \ldots, \phi_p$ – авторегресійні коефіцієнти, $\theta_1, \ldots, \theta_q$ – коефіцієнти MA, а ε_t – випадкова «біла шумова» помилка з нульовим середнім.

Перш ніж будувати ARIMA, необхідно перевірити, чи стаціонарний початковий ряд Yt. Використовують тести на корінь одиницю такі як тест Дікі–Фуллера (ADF): перевіряє нульову гіпотезу наявності одиничного кореня (нестаціонарність); і тест KPSS: перевіряє протилежну гіпотезу (станціонарність). Якщо обидва тести свідчать про нестаціонарність (наявність тренду чи мінливої дисперсії), треба зробити диференціювання.

Якщо часовий ряд нестаціонарний, необхідно виконати послідовні диференціювання рівняння Y_t доти, доки індекс строкового ряду не стане стаціонарним.

Наприклад ,

$$X_t^{(1)} = Y_t - Y_{t-1} \quad \text{(одноразове диференціювання)},$$

$$X_t^{(2)} = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) = Y_t - 2Y_{t-1} + Y_{t-2} \quad \text{(друге)}.$$

Значення d в ARIMA (p, d, q) – це кількість таких диференціювань.

Для визначеня значень **p** та **q** необхідно побудувати автокореляційну функцію (ACF) та часткову автокореляційну функцію (PACF). ACF показує кореляцію X_t зі своїми лагами X_{t-k}. Якщо ACF повільно спадає (гіпотетично наявний AR-ефект), це вказує на потребу великого p. Якщо ACF має чіткі відліки лише у перших q лагах, це натякає на MA(q).

PACF оцінює кореляцію між X_t та X_{t-k} після виключення проміжного впливу лагів 1…k−1. Якщо PACF обривається швидко на лагу p, це сигнал

пари p для AR-компонента. Зазвичай аналізують графіки ACF і PACF (похибки з довірчими межами 95 %) і вибирають найбільш підходящі малі значення p і q.

Після вибору (p, d, q) модель оцінюють методом максимальної правдоподібності (ML) або нелінійної оптимізації (якщо ML-оцінювання надто складне). Викрнується перевірка значення коефіцієнтів $\hat{\phi}_i, \hat{\theta}_j$ на статистичну значущість (за t-статистикою) та знаки (AR-коефіцієнти мають формувати процес, що є стаціонарним, тобто поліном $1 - \phi_1 z - \cdots - \phi_p z^p$ не має коренів усередині одиничного кола).

Наприклад, є щомісячні дані витрат на соціальний захист та соціальне забезпечення Y_t за останні 36 місяців. Провідні тести ADF і KPSS показали, що первинний ряд нестаціонарний. Виконали одне диференціювання і отримали стаціонарний ряд $X_t = Y_t - Y_{t-1}$. Графік PACF обривається біля лага 2 (чи показує спад після 2), а ACF – близько лага 1. Тому, слід побудувати ARIMA(2, 1, 1). Оцінка дає $\hat{\phi}_1 = 0.5, \hat{\phi}_2 = -0.2, \hat{\theta}_1 = 0.3$, всі коефіцієнти значущі ($p<0,05$), а тест Люнга–Бокса для залишків на лагах до 12 не дає підстав відхилити гіпотезу «залишки випадкові». Прогноз на наступні 3 місяці будують за формулою з отриманими $\hat{\phi}_i, \hat{\theta}_j$ та спостережуваними $\hat{\varepsilon}_t$.

SARIMA (Сезонний ARIMA) — це розширення ARIMA, яке враховує регулярні сезонні коливання у часовому ряді.

Якщо дані мають чітку сезонність (наприклад, щомісячні соцвиплати традиційно ростуть наприкінці року або у січні—у зв'язку з підвищеним перевірками пільгових груп), доцільно використати SARIMA. Позначення SARIMA (p, d, q)(P, D, Q)s означає:

- (p, d, q) – невсезонні (non-seasonal) авторегресійні, інтеграційні та MA лаґи,
- (P, D, Q) – сезонні лаґи AR, диференціювання, MA,

- s – довжина сезонного періоду (для місячних даних s=12, для щоквартальних — s=4).

Після початкової перевірки стаціонарності, перш за все необхідно перевірити сезонну складову, виконуючи сезонне диференціювання:

$$Z_t^{(D)} = Y_t - Y_{t-s},$$

де s – сезонний лаг (наприклад, 12 місяців). Якщо після одного сезонного диференціювання дані залишаються нестаціонарними (є тренд), додатково застосовують невсезонне диференціювання d разів [99].

$$X_t = \Delta^d \Delta_s^D Y_t, \text{ де } \Delta_s Y_t = Y_t - Y_{t-s}.$$

Несезонні параметри моделі (p та q) підбирають шляхом аналізу автокореляційної та часткової автокореляційної функцій після усунення тренду й сезонної компоненти: на графіках ACF дивляться, на яких лагах кореляція раптово обривається (для MA) або ж поступово спадає (для AR), а PACF допомагає визначити порядок AR, коли відліки різко обриваються після певного лага. Сезонні параметри (P та Q) визначають аналогічним чином, але вже працюють із лагами, кратними довжині сезону s: для щомісячних даних це, зокрема, лага 12, 24, 36 і так далі, де також аналізують ACF та PACF, щоб зрозуміти, наскільки сильні сезонні автокореляції у відповідних точках і яким має бути сезонний порядок моделі.

Після урахування сезонного та несезонного диференціювання оцінюють одночасно авторегресійні та MA-кути для обох блоків. Наприклад, SARIMA (1, 1, 1)(1, 1, 1)12 має вигляд [99]:

$$\Phi_1 X_{t-12} + \phi_1 X_{t-1} + \cdots + \theta_1 \varepsilon_{t-1} + \Theta_1 \varepsilon_{t-12} + \varepsilon_t = X_t,$$

де $X_t = \Delta^1 \Delta_{12}^1 Y_t.$

Наприклад, є дані про щомісячні соцвиплати Yt за 60 місяців. На ACF чітко видно піки на лагах 12, 24 та 36, а PACF має відповідні відліки. Існує тренд, тому спочатку робимо одне сезонне диференціювання $(\Delta_{12} Y_t)$, а потім

ще одне невезонне (Δ). Отримуємо стаціонарний ряд $X_t = \Delta\Delta_{12}Y_t$. Графіки ACF/PACF цього ряду вказують на потребу в сезонних та несезонних AR і MA коефіцієнтах приблизно як (1, 1, 1)(1, 1, 1)12. Оцінка SARIMA (1, 1, 1)(1, 1, 1)12 показує значущість усіх коефіцієнтів $\hat{\phi}_1, \hat{\theta}_1, \hat{\Phi}_1, \hat{\Theta}_1$. Тест Люнга–Бокса не виявляє автокореляції у залишках на лагах до 24. Прогноз на наступні 6 місяців формується із урахуванням сезонної комбінації тренду та щорічного коливання.

У багатьох випадках, виключно показників динаміки витрат на соціальний захист та соціальне забезпечення не достатньо: на виплати безпосередньо впливають зовнішні фактори, наприклад кількість бойових інцидентів, обсяги гуманітарної допомоги або зміни в адмініструванні соцпрограм. ARIMAX і SARIMAX розширюють базову ARIMA/SARIMA, додаючи вектор зовнішніх регресорів Xt (exogenous variables).

Модель ARIMAX (p, d, q) з одним екзогенним регресором Xt (може бути вектором із кількох факторів) і записується так 101]:

$$\Delta^d Y_t = \phi_1 \Delta^d Y_{t-1} + \cdots + \phi_p \Delta^d Y_{t-p} \; + \; \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} \; + \; \beta \, X_t + \varepsilon_t.$$

Якщо є кілька зовнішніх змінних, βX_t замінюють на $\boldsymbol{\beta}^\top \mathbf{X}_t$, де $\mathbf{X}_t = (X_{1,t}, X_{2,t}, \ldots, X_{m,t})$.

Аналогічно SARIMA, але з екзогенними регресорами:

$$\Delta^d \Delta_s^D Y_t = \sum_{i=1}^{p} \phi_i \, \Delta^d \Delta_s^D Y_{t-i} \; + \; \sum_{j=1}^{q} \theta_j \, \varepsilon_{t-j} \; + \; \sum_{i=1}^{P} \Phi_i \, \Delta^d \Delta_s^D Y_{t-is} \; + \; \sum_{j=1}^{Q} \Theta_j \, \varepsilon_{t-js} \; + \; \boldsymbol{\beta}^\top \mathbf{X}_t \; + \; \varepsilon_t.$$

Тут Δ_s^D – сезонне диференціювання, як описано раніше [99].

Порядок побудови моделі з екзогенними змінними представлений на рисунку3.1:

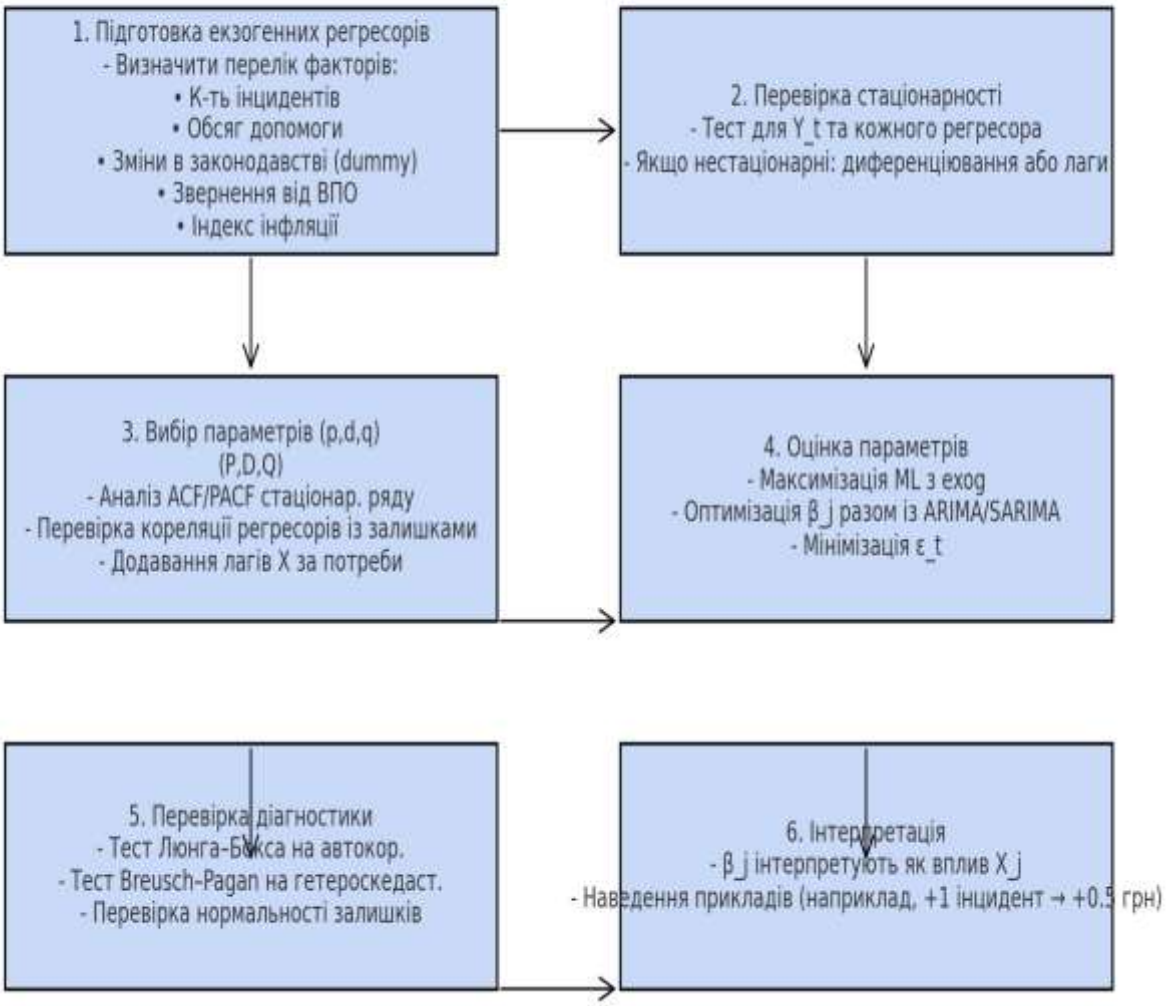

Рисунок 3.1 - Методика побудои моделей з екзогенними змінними

На рис. 3.1 представлені етапи методики побудои моделей з екзогенними змінними:

Етап 1. Пдготовка зовнішніх регресорів. На першому етапі потрібно зібрати та підготувати набір екзогенних змінних, які потенційно впливають на соціальні виплати. До таких факторів відносяться: кількість бойових інцидентів у поточному місяці (або кварталі), обсяг міжнародної гуманітарної допомоги (в мільйонах гривень), зміни у законодавстві (представлені як dummy-змінна: 1 – якщо в цьому кварталі відбулися нові законодавчі зміни, 0 – якщо змін не було), кількість опрацьованих звернень від внутрішньо переміщених осіб (у тисячах), а також індекс споживчих цін за місяць. Ці

регресори мають бути ретельно зібрані, уніфіковані за періодами й приведені до відповідного формату (наприклад, приведення сум до мільйонів гривень чи нормування показників) для подальшого включення в модель ARIMAX/SARIMAX.

Етап 2. Перевірка стаціонарності. На другому етапі перевіряють, чи є стаціонарними часові ряди – як цільовий YtY_tYt (соцвиплати), так і всі екзогенні регресори. Для цього застосовують стандартні тести (Дікі–Фуллера, KPSS тощо). Якщо виявлено, що будь-який із рядів є нестаціонарним (має тренд або нерівномірну дисперсію), його необхідно привести до стаціонарного вигляду. Для цього найчастіше застосовують диференціювання певного порядку або ж залишають змінну в моделі у вигляді лагованої версії, щоб уникнути надмірного ускладнення специфікації. Нормалізація та приведення до стаціонарності дозволяють запобігти «фальшивим» кореляціям і забезпечити коректність подальшого навчання ARIMAX/SARIMAX-моделі.

Етап 3. Вибір (p, d, q) та (P, D, Q).Після приведення ряду до стаціонарності шляхом сезонного та несезонного диференціювання проводять аналіз автокореляційної (ACF) та часткової автокореляційної (PACF) функцій отриманого стаціонарного ряду. Це дозволяє визначити порядки AR і MA компонентів. Одночасно перевіряють, чи екзогенні змінні корелюють із залишками моделі: якщо виявляють значущі зв’язки, додають відповідні лаги цих регресорів, щоб прибрати автокореляцію в залишках і підвищити точність оцінювання [100].

Етап 4. Оцінка параметрів моделі. Параметри моделі оцінюються шляхом максимізації правдоподібності з урахуванням екзогенних змінних. Під час оптимізації алгоритм одночасно підбирає коефіцієнти βj для зовнішніх факторів і параметри ARIMA/SARIMA, мінімізуючи суму квадратів залишків ε_t. Після цього, залишки перевіряють на автокореляцію за допомогою тесту Люнга–Бокса і на гетероскедастичність (змінність дисперсії) за тестом Бройша–Пагана, щоб упевнитися в коректності моделі.

Етап 5. Інтерпретація оезультатів. Коефіцієнти β_j відображають відносний вплив кожного екзогенного фактора Xj на зміну (або рівень) соціальних виплат, після того як модель ARIMA/SARIMA врахувала внутрішню часову динаміку (автокореляцію та сезонність). Наприклад, $\beta_{incidents}$=0,5 = 0 означає, що за кожної додаткової бойової події прогнозований приріст суми однієї соціальної виплати (після діференціювання) становитиме в середньому 0,5 грн, за інших рівних умов.

Наприклад, є щоквартальні дані соцвиплат Y_t за 32 квартали. Зовнішні регресори: кількість бойових інцидентів X1,t і обсяг гуманітарної допомоги X2,t. Первинний ряд Yt – нестаціонарний (ADF показує високий p-значення). Робимо одне диференціювання: ΔYt. Аналіз ACF/PACF ΔYt показує спад у ACF біля лага 1 і PACF біля лага 2. Будуємо ARIMA(2, 1, 1) без зовнішніх регресорів. Після оцінювання залишків і перевірки їхньої випадковості включаємо зовнішні фактори. Перевіряємо стаціонарність X1,t і X2,t: якщо, наприклад, X1,t (кількість інцидентів) має стійкий тренд, беремо $\Delta X_{1,t} = X_{1,t} - X_{1,t-1}, - X_{2,t}$ стабільний – залишаємо без змін. Оцінюємо ARIMAX(2, 1, 1) з екзогенними $\Delta X_{1,t}$ і $X_{2,t}$. Отримуємо значущі коефіцієнти $\hat{\beta}_1$ (для диференційованих бойових інцидентів) і $\hat{\beta}_2$, тест Люнга–Бокса залишки проходить. Прогноз на наступні 4 квартали формують за цими параметрами, вводячи екзогенні змінні із прогнозних сценаріїв (очікувана кількість інцидентів, плановий обсяг допомоги).

Моделі ARIMA і SARIMA дають змогу зорієнтуватися у внутрішній часовій динаміці соціальних видатків, врахувати тренд і сезонні коливання у щомісячних чи щоквартальних даних. Використання ARIMAX/SARIMAX зі зовнішніми регресорами дозволяє інтегрувати вплив бойових інцидентів, гуманітарної допомоги та інших шокових змінних, що критично важливо для прогнозування в умовах військового конфлікту. У поєднанні з економетричними тестами на стаціонарність, автокореляцію та гетероскедастичність ці моделі стають надійною основою для побудови

короткострокових і середньострокових прогнозів соцвиплат із відображенням як статистичної структури часового ряду, так і екзогенних шоків.

3.2 Методика застосування машинного навчання

Методи машинного навчання (МН) відкрили нові можливості у прогнозуванні соціально-економічних показників завдяки своїй гнучкості, здатності виявляти складні нелінійні залежності та працювати з великою кількістю змінних одночасно. На відміну від класичних економетричних підходів, алгоритми МН часто використовують «black-box» моделі, які надають високу прогностичну точність.

Метод Random Forest Regression (RFR) — це ансамблевий алгоритм, який будує багато регресійних дерев на підвибірках даних і випадкових підмножинах ознак, а кінцевий прогноз отримує як середнє значень усіх дерев. Він стійкий до перенавчання (overfitting) та здатен виявляти складні нелінійні залежності.

На рис. 3.3 наведено детальний опис методології Random Forest Regression для застосування в дослідженні динаміки соціальних видатків.

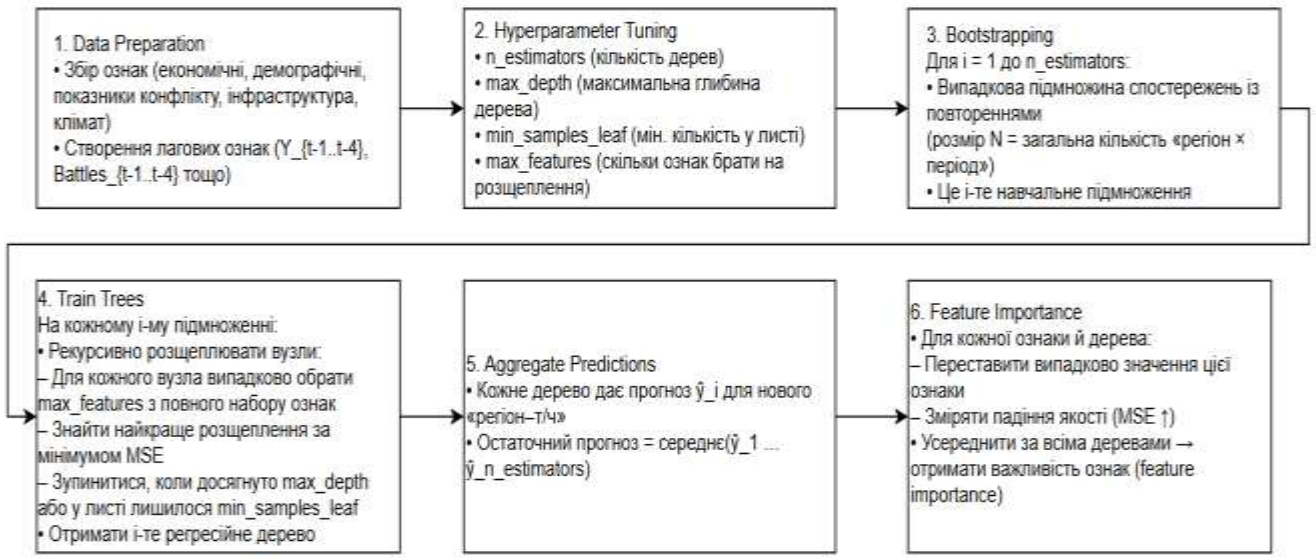

Рисунок 3.3 – Методтка застосування Random Forest Regression

Крок 1. Підготовка даних. Для побудови моделі необхідно спочатку зібрати й упорядкувати всі спостереження за «регіон × період» у єдиний масив, де кожен рядок описує конкретний регіон за певний квартал. Ознаки діляться на кілька груп: економічні індикатори (ВВП на душу населення, обсяг доходів зведеного бюджету, млн грн), демографічні показники (загальна чисельність населення, відсоток отримувачів пенсій та допомог, кількість внутрішньо переміщених осіб, тис осіб), показники конфлікту (кількість бойових інцидентів за період, од., кількість цивільних втрат, осіб площа зруйнованої інфраструктури, га), інфраструктурні дані (кількість пунктів видачі гуманітарної допомоги, од., число евакуйованих, тис. осіб) та кліматичні змінні (середня температура за сезон, град, індикатор «зима/літо»). Крім того, для захоплення часової динаміки створюються лагові ознаки: обсяги соціальних виплат за попередні чотири квартали (Y_{t-1}, Y_{t-2}, Y_{t-3}, Y_{t-4}), середня кількість бойових інцидентів за два–три попередні квартали та середній рівень безробіття за останні два квартали. Такий підхід дозволяє відобразити як поточний стан, так і вплив попередніх періодів на обсяг соціальних видатків.

Крок 2. Налаштування гіперпараметрів: Під час налаштування моделі випадкового лісу ключовими гіперпараметрами є кількість дерев у колекції (n_estimators), яка визначає, скільки окремих регресійних дерев створюватиметься; максимальна глибина кожного дерева (max_depth), що обмежує, наскільки вузлів може мати одне дерево, і зменшує ризик перенавчання; мінімальна кількість спостережень у листі (min_samples_leaf), яка запобігає утворенню занадто дрібних кластерів у кінцевих вузлах і підвищує стабільність моделі; а також кількість ознак, що випадковим чином розглядаються під час кожного розщеплення вузла (max_features), що гарантує різноманітність дерев і знижує їхню кореляцію. Удосконалене підборювання цих параметрів дає змогу досягти оптимального балансу між точністю та узагальненням прогнозу.

Крок 3. Навчаня моделей: Навчання моделі починається з етапу bootstrapping, коли для кожного з n_estimator дерев із загального набору даних

«регіон × період» випадковим чином вибирається підмножина розміру N з поверненням. Кожна така підвибірка використовується для побудови окремого дерева. Далі, у процесі побудови iii-го дерева, на кожному кроці рекурсивного розщеплення вузлів із поточної підвибірки випадковим чином обирається max_features ознак із загального списку. Серед цих ознак знаходять розщеплення, яке мінімізує середньоквадратичну помилку (MSE) у вузлі. Розщеплення триває доти, доки не буде досягнута обмежена глибина max_depth або у формуваному листі не залишиться менше, ніж min_samples_leaf спостережень. В результаті кожного такого процесу отримуємо окреме регресійне дерево за алгоритмом CART, кожне з яких може мати високу дисперсію, але в ансамблі їхні помилки нівелюються.

Крок 4. Агрегація прогнозів: Коли всі n_estimators дерев навчені, для кожного нового спостереження — тобто для кожного конкретного поєднання «регіон–період» — кожне дерево видає власний прогноз y^i.

$$\hat{y}_{RF} = \frac{1}{n\_estimators} \sum_{i=1}^{n\_estimators} \hat{y}_i .$$

Остаточний прогноз будується як середньозважене цих значень; за замовчуванням це просте середнє всіх y^i, що дозволяє зменшити випадкові коливання окремих дерев і отримати більш стабільне та точне передбачення.

Крок 5. Оцінка важливості ознак: При оцінці важливості ознак у випадковому лісі оцінюють, наскільки кожна змінна справді впливає на якість прогнозів. Для цього у навченої моделі по черзі беруть одну ознаку і випадковим чином переставляють у всіх записах її значення, залишаючи решту даних без змін. Після цього перераховують помилку моделі (MSE) на тому ж тестовому наборі. Якщо перемішування однієї конкретної ознаки призводить до значного зростання середньоквадратичної помилки, це означає, що ця ознака вносила великий вклад у точність передбачень. Операцію повторюють для кожного дерева у лісі, а потім усереднюють отримані зміни MSE по всіх деревах. У такий спосіб виходить оцінка важливості (feature

importance) для кожної змінної. Наприклад, якщо виявиться, що випадкове перемішування кількості бойових інцидентів суттєво підвищує помилку прогнозу соціальних виплат порівняно з іншими факторами, це свідчитиме про те, що саме цей показник є однією з ключових детермінант.

Через TimeSeriesSplit (часову крос-перевірку) набір даних розподіляють на блоки: Наприклад, тренують модель на перших 20 кварталах, перевіряють на 4 наступних; потім тренують на перших 24, перевіряють на 4 тощо. Обчислюють метрики:

MAE (Mean Absolute Error): $\frac{1}{M}\sum_{j=1}^{M}|\hat{y}_j - y_j|$.

RMSE (Root Mean Squared Error): $\sqrt{\frac{1}{M}\sum_{j=1}^{M}(\hat{y}_j - y_j)^2}$.

MAPE (Mean Absolute Percentage Error): $\frac{100\%}{M}\sum_{j=1}^{M}\left|\frac{\hat{y}_j - y_j}{y_j}\right|$. [99]

Після побудови моделі Random Forest її результати порівнюють із показниками традиційних підходів, таких як ARIMA/SARIMA і лінійна регресія, за допомогою метрик MAE, RMSE і MAPE. Наприклад, на прогнозному горизонті в один-два квартали Random Forest може забезпечувати середню абсолютну похибку близько 5 грн на особу, у той час як ARIMA покаже близько 7 грн. Такий порівняльний аналіз дає змогу об’єктивно оцінити переваги алгоритму машинного навчання, особливо в умовах нелінійних взаємозв’язків і багатовимірності даних [99].

Після завершення крос-перевірки найкращої конфігурації Random Forest фінальну модель навчають уже на всіх доступних даних, або, за необхідності, застосовують стратегію rolling window (наприклад, останні 32 квартали). Далі, використавши отримані ваги та структуру, генерують прогнози соціальних видатків на наступні один-три квартали, підставляючи в модель актуальні значення ознак (лагові показники, поточні ВВП, кількість інцидентів тощо).

Такий підхід гарантує, що остаточний прогноз спирається на весь обсяг інформації і враховує «найсвіжіші» дані, необхідні для планування соціальних програм у кризових умовах.

Приклад підготовки даних для побудови моделей: для задачі прогнозування соціальних витрат: доступні щоквартальні дані для 20 регіонів за 16 кварталів (загалом 320 точок). Для кожного «регіон–квартал» є:

- *GDP_q* – ВВП на душу населення (тис. грн),
- *Budget_q* – бюджетні надходження (млн грн),
- *Population_q* – тисячі осіб,
- *Pensioners_pct* – % пенсіонерів та одержувачів допомог,
- *VPO_thous* – ВПО, тисяч осіб
- *Battles_cnt* – кількість атак (військових інцедентів),
- *Losses_civ* – цивільні втрати (тис осіб),
- *Aid_points* – кількість гуманітарних пунктів (одиниць),
- *Evacuated_thous* – евакуйовано, тисяч осіб,
- *Temp_avg* – середня температура за квартал, градусів
- *Season* – 1 (зима), 0 (літо),
- *Y_q* – сума соціальних видатків, що припадає на одну особу (грн).

Створюємо лагові ознаки для Y і Battles_cnt:

Y_lag1 = Y_{q−1}, Y_lag2 = Y_{q−2}, ..., Y_lag4
Battles_lag1 = Battles_{q−1}, ...

Після цього крім вихідних 11 ознак матриці XXX матимемо ще 8 лагових ознак (4 для YYY і 4 для Battles_cntBattles_cntBattles_cnt). Загалом 19 ознак. Код мовою Python (бібліотека scikit-learn)

```
import pandas as pd
import numpy as np
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import TimeSeriesSplit, cross_val_score
from sklearn.metrics import mean_absolute_error, mean_squared_error

# Припустимо, data – DataFrame із колонками:
# ['Region', 'Quarter', 'GDP', 'Budget', 'Population', 'Pensioners_pct', 'VPO_thous',
#  'Battles_cnt', 'Losses_civ', 'Aid_points', 'Evacuated_thous', 'Temp_avg', 'Season', 'Y']
```

```
# 1. Сортуємо дані за регіоном і кварталом для коректних лагів
data = data.sort_values(['Region', 'Quarter'])

# 2. Додаємо лагові ознаки
for lag in [1, 2, 3, 4]:
    data[f'Y_lag_{lag}'] = data.groupby('Region')['Y'].shift(lag)
    data[f'Battles_lag_{lag}'] = data.groupby('Region')['Battles_cnt'].shift(lag)

# 3. Видаляємо перші 4 квартали для кожного регіону (де лагові значення NaN)
data = data.dropna().reset_index(drop=True)

# 4. Формуємо матрицю X та вектор y
feature_cols = [
    'GDP', 'Budget', 'Population', 'Pensioners_pct', 'VPO_thous',
    'Battles_cnt', 'Losses_civ', 'Aid_points', 'Evacuated_thous',
    'Temp_avg', 'Season',
    'Y_lag_1', 'Y_lag_2', 'Y_lag_3', 'Y_lag_4',
    'Battles_lag_1', 'Battles_lag_2', 'Battles_lag_3', 'Battles_lag_4'
]
X = data[feature_cols].values
y = data['Y'].values

# 5. Налаштовуємо Random Forest
rf = RandomForestRegressor(
    n_estimators=200,
    max_depth=8,
    min_samples_leaf=5,
    max_features='sqrt',
    random_state=42,
    n_jobs=-1
)

# 6. Крос-перевірка із часовою обережністю
tscv = TimeSeriesSplit(n_splits=5)
mae_scores = cross_val_score(rf, X, y, cv=tscv, scoring='neg_mean_absolute_error')
rmse_scores = cross_val_score(rf, X, y, cv=tscv, scoring='neg_root_mean_squared_error')

print("MAE (CV):", -np.mean(mae_scores))
print("RMSE (CV):", -np.mean(rmse_scores))

# 7. Навчання на повних даних і прогноз для наступних періодів
rf.fit(X, y)
# Припустімо, X_future – матриця ознак для прогнозного періоду
y_pred = rf.predict(X_future)
```

У наведеному коді:

n_estimators=200 це кількість дерев у лісі. Більше дерев підвищують стабільність, але збільшують час навчання.

max_depth=8 це Максимальна глибина кожного дерева. Менші значення запобігають глибокому зануренню й перенавчанню.

min_samples_leaf=5 це мінімальна кількість спостережень у листі. Забезпечує, що жодне дерево не матиме занадто малих листових вузлів, що підвищує робастність і скорочує перенавчання.

max_features='sqrt' - на кожному вузлі розщеплення використовують лише sqrt{p} випадкових ознак (де p – загальна кількість ознак). Це сприяє ще більшому розрідженню дерев і зниженню кореляції між ними.

Random Forest Regression у рамках дисертації служить надійним інструментом для прогнозування обсягів соціальних видатків, оскільки дозволяє працювати з великою кількістю факторів, автоматично розпізнавати складні нелінійні взаємодії та бути стійким до перенавчання, водночас даючи змогу порівняти результати з традиційними економетричними моделями (ARIMA/SARIMA, лінійна регресія).

Класичні регресійні підходи часто виявляються недостатньо гнучкими для охоплення складних нелінійних взаємозв'язків і опосередкованих ефектів, що виникають під впливом численних економічних, демографічних чинників, впливу військового конфлікту. Градієнтний бустинг—один із сучасних методів машинного навчання—дозволяє ефективно поєднувати численні «слабкі» моделі (дерева рішень), кожна з яких покращує залишки попередніх, і в результаті створює високо точний ансамбль. Найпопулярніші реалізації цього підходу—XGBoost і LightGBM—оптимізовані за швидкістю й пам'яттю, що робить їх незамінними для роботи з великими панельними чи часовими наборами даних, характерними для дослідження соціальних видатків. Завдяки здатності «ловити» складні нелінійні патерни та взаємодії між ознаками, адаптивним стратегіям регуляризації і зручним інструментам оцінки важливості змінних градієнтний бустинг дозволяє отримувати більш стабільні й точні прогнози, ніж традиційні алгоритми.

На рис. 3.3 представлена методика зстосування градієнтного бустингую

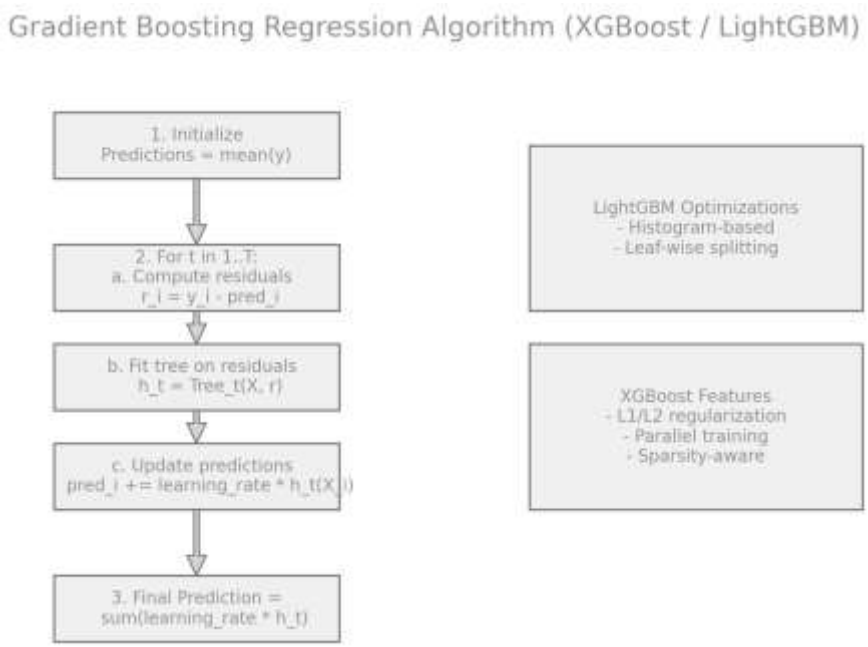

Рисунок 3.3 – Методика застосування градієнтного бустингу

Градієнтний бустинг — це метод побудови ансамблю дерев рішень, у якому кожне наступне дерево навчається коригувати помилки попередніх.

Ініціалізацію моделі виконують шляхом вибору найпростішого початкового прогнозу, що найчастіше задають у вигляді константи — наприклад, як середнє значення цільової змінної yyy у навчальній вибірці. Це дає початкову точку, від якої далі відштовхуються при послідовному вдосконаленні передбачень.

На кожній наступній ітерації $t = 1, 2, \ldots, T$ спершу обчислюють залишки (градієнти), тобто різницю між фактичними значеннями y_i і поточними прогнозами $\hat{y}_i^{(t-1)}$. У разі квадратичної функції втрат залишок $r_i^{(t)}$ набуває вигляду $r_i^{(t)} = y_i - \hat{y}_i^{(t-1)}$. Ці залишки відображають те, наскільки попередній прогноз недооцінив або переоцінив справжнє значення.

Потім на отриманих залишках навчають нове регресійне дерево $h^{(t)}$. Під час навчання його мета—мінімізувати середньоквадратичну помилку на тренувальних парах $\{(\mathbf{x}_i, r_i^{(t)})\}$, де $\mathbf{x}_i$ — набір ознак для спостереження, а $r_i^{(t)}$ — уже згадані залишки. Таким чином кожне дерево спеціалізується на поясненні тих випадків, які попередні ітерації пояснили найгірше.

Після отримання дерев'яного прогнозу на залишках виконують оновлення загального передбачення за формулою

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta\, h^{(t)}(\mathbf{x}_i),$$

де η — гіперпараметр швидкості навчання (learning rate). Застосування η<1 гальмує надмірний вплив окремого дерева на остаточний результат, що знижує ризик перенавчання й робить модель більш стійкою. Після завершення всіх T ітерацій остаточний прогноз для кожного спостереження визначається як

$$\hat{y}_i = \hat{y}_i^{(0)} + \sum_{t=1}^{T} \eta\, h^{(t)}(\mathbf{x}_i),$$

де $\hat{y}_i^{(0)}$ — початкове константне передбачення, а всі $h^{(t)}$ — дерева, навчальніся на відповідних залишках. Така схема забезпечує поступове покращення прогнозу: на кожному кроці модель враховує помилки попередніх ітерацій і, крок за кроком, адаптується до складної залежності між ознаками та цільовою змінною.

В результаті, замість того щоб будувати одне дерево, яке намагається вгадати ууу одразу, поетапно додаємо «дрібні» дерева, кожне з яких вчиться пояснювати помилки попереднього.

На рис. 3.4 наведено візуалізацію ключових кроків градієнтного бустингу та його порівняння з LightGBM. Зліва — послідовні дії (ініціалізація → ітерації → фінальна агрегація), справа — відмінності XGBoost і LightGBM.

Рисунок 3.4 – Порівняння алгоритмів XGBoost і LightGBM

XGBoost (eXtreme Gradient Boosting) [104] поєднує кілька важливих удосконалень, які роблять його одним із найпопулярніших інструментів для бустингу дерев рішень. По-перше, він підтримує як L1-регуляризацію (Lasso), так і L2-регуляризацію (Ridge), що допомагає зменшити надлишкове пристосування моделі (overfitting). По-друге, XGBoost автоматично обробляє пропущені значення: під час побудови кожного вузла дерева він знаходить оптимальне місце для відсутніх спостережень, тож розробнику не потрібно вручну заповнювати NaN-и. Для прискорення навчання XGBoost використовує багатопоточне виконання — паралельне побудова розщеплень у різних гілках дерева на кількох ядрах процесора. Крім того, алгоритм є «sparsity-aware», тобто спеціально оптимізований для роботи з розрідженими даними (наприклад, коли серед числових ознак багато нулів), що додатково пришвидшує тренування та зменшує витрати пам’яті. Таким чином XGBoost поєднує регуляризацію, акуратну роботу з пропусками, паралелізацію та оптимізації для розріджених матриць, що робить його потужним і гнучким інструментом для побудови точних прогнозних моделей.

LightGBM (Light Gradient Boosting Machine) [105] виділяється кількома ключовими підходами, які забезпечують високу швидкість та ефективність при роботі з великими наборами даних. По-перше, він використовує гістограмо--базований підхід: замість того, щоб на кожному кроці розглядати всі можливі пороги для розщеплення числових ознак, LightGBM попередньо розбиває кожну ознаку на обмежену кількість бінів. Це дає змогу миттєво вибирати кращі межі між біном і біном, що значно пришвидшує пошук оптимальних розщеплень і знижує витрати пам’яті.

По-друге, LightGBM будує дерева за принципом leaf-wise (росте найбільш «корисне» листя, а не рівень цілого дерева). Іншими словами, замість того, щоб розширювати всі вузли одного рівня одночасно (як це робить більшість інших реалізацій бустингу), він обирає єдине розщеплення, яке дає максимальне зменшення функції втрат, і додає гілку лише в цьому листі. Таке рішення дозволяє утворювати значно глибші дерева із меншою кількістю

ітерацій, і в підсумку модель здатна захоплювати складнішу нелінійну залежність при менших обсягах обчислень.

Крім того, LightGBM оптимізовано для прискореного навчання та зниженого споживання пам'яті. Спеціальні структури даних, що працюють із бінаризованими гістограмами, дозволяють обходитися значно меншими обсягами оперативної пам'яті під час підрахунку статистик, а також використовувати розподілене навчання по багатьох вузлах кластера без значних накладних витрат. У результаті LightGBM демонструє надзвичайно високу швидкість тренування навіть на терабайтних наборах даних.

Завдяки поєднанню гістограмо--орієнтованого пошуку, leaf-wise росту дерев і ефективних структур даних LightGBM у багатьох завданнях прогнозування, зокрема у сфері соціальних видатків, показує суттєво кращі результати як у швидкодії, так і в точності порівняно з класичними реалізаціями градієнтного бустингу.

Підготовка даних та ознак є важливим етапом побудови моделей. Аналогічно Random Forest, для кожного «регіон × час» формують набір ознак:
- поточні фактори:

$$GDP_t, Budget_t, Unemp_t, VPO_t, Battles_t, Losses_t, AidPts_t, Temp_t, Season_t$$

- лагові ознаки: $Y_{t-1}, Y_{t-2}, Y_{t-3}, Y_{t-4}$ (попередні соціальні виплати), $\overline{Battles}_{t-2:t-1}$ (середній показник за 2 квартали), $\overline{Unemp}_{t-2:t-1}$
- додаткові ознаки (за бажанням): $GDP \times Unemp$ (термін, що описує взаємодію), $\Delta VPO = VPO_t - VPO_{t-1}$ (динаміка кількості ВПО)

Наступним кроком є налаштування гіперпараметрів:

– *n_estimators* = кількість дерев (наприклад, 200, 500, 1000).

– *learning_rate* = швидкість навчання (часто в діапазоні [0.01, 0.3]).

– *max_depth* = максимальна глибина дерева (наприклад, 6–10).

_min_child_weight (XGBoost) / min_data_in_leaf (LightGBM) = мінімальна кількість прикладів у листі, щоб уникнути надто малих вузлів.

– *subsample (0.5–1.0)* = відсоток рядків, що використовуються для побудови кожного дерева (аналог Bagging).

– *colsample_bytree* = доля ознак, що розглядаються під час кожного дерева.

– Регуляризаційні параметри XGBoost: α\alphaα (L1), λ (L2).

Алгоритм навчання моделі можна представити настпним чином

```
Initialize: pred_i^(0) = mean(y)
For t in 1..T:
   # 1. Обчислення залишків
   r_i^(t) = y_i - pred_i^(t-1)
   # 2. Навчити дерево h_t на (X_i, r_i^(t))
   h_t = TrainTree(X, r^(t); params)
   # 3. Оновити прогнози
   pred_i^(t) = pred_i^(t-1) + learning_rate * h_t(X_i)
Final preds: pred_i = pred_i^(T)
```

Градієнтний бустинг (XGBoost, LightGBM) [104, 105] у контексті зпрстання соціальних видатків дозволяє отримати високоточні прогнози, використовуючи потужні оптимізації, складні взаємодії між індикаторами та гнучке налаштування гіперпараметрів.

3.3 Методика використання байєсівських мереж та ймовірнісних моделей

Байєсівська мережа [106-113] — графічна ймовірнісно-статистична модель, яка описує спільний розподіл набору змінних за допомогою спрямованого ациклічного графа (DAG). Вершини (вузли) цього графа відповідають змінним (спостережуваним або латентним), а ребра (дуги) позначають умовні залежності. Кожна вершина має власну таблицю умовних ймовірностей (Conditional Probability Table, CPT), що описує ймовірність станів цієї змінної за різних конфігурацій її батьків. У контексті соціальних видатків байєсівська мережа дозволяє формалізувати причинно-наслідкові взаємозв’язки між:

- військовими діями (кількість бойових інцидентів, індекс боїв),
- демографічними характеристиками (кількість ВПО, частка одержувачів пенсій, народжуваність),

- економічними індикаторами (валовий внутрішній продукт (або валовий регіональний продукт), бюджетні надходження, рівень безробіття),
- гуманітарною допомогою (обсяг міжнародного фінансування, кількість доставлених вантажів),
- соціальними виплатами (сума виплат на душу населення, відсоток бюджету, спрямованого на соціальні виплати).

Перевагами застосування байєсівських мереж є їх наочність та можливість оцінювати ймовірність настання подій за різного впливу ознак. Наприклад, яка ймовірність зростання соціальних виплат, якщо одночасно зміняться значення макроекономічних показників та індикаторів конфлікутв за певного рівня гуманітарної підтримки.

Методика використання сереж Байєса для аналізу взаємозв'язку факторів (стан конфлікту, демографічні чинники, макроекономічні індикатори), що впливають на структуру та динаміку видатків на соціальний захист та соціальне забезпечення прелставлена у наступну послідовність кроків [106-113].

Крок 1. Визначення вузлів (змінних) моделі (табл. 3.5)

Таблиця 3.5 – Оис змінних моделі

Інтенсивність конфлікту	- кількість інцидентів (за квартал) - індекс бойової активності (композитний індекс), - кількість втрат серед цивільних.
Демографічні характеристики	• кількість внутрішньо переміщених осіб (ВПО, тис.), - частка пенсіонерів (% від населення), - народжуваністьі (кількість народжених на 1000 осіб).
Макроекономічні індикатори	• регіональний ВВП на душу (тис. грн/рік, інфляційно скоригований), • бюджетні надходження (млн грн/квартал), • рівень безробіття (%)
Гуманітарна допомога	• обсяг міжнародного фінансування (млн дол.), • кількість доставлених гуманітарних вантажів (одиниць).
Соціальні виплати	• розмір соцвиплат на душу населення (грн/місяць) • частка соцвиплат у загальних регіональних витратах (%).

У табл. 3.5 наведено перелік ключових змінних, які використовуються у моделі соціальних видатків, розбитих за категоріями. Кожна група ознак відображає окремий аспект впливу: інтенсивність конфлікту, демографічні зміни, економічна ситуація, обсяги гуманітарної допомоги та самі соціальні виплати як цільова змінна.

Приклади латентних вузлів представлені у табл. 3.6. У модель можна ввести латентні (невимірювані безпосередньо) змінні, кожна з яких має декілька індикаторів:

Таблиця 3.6 – Приклад латентних змінних

Назва змнної	Опис індикаторів
Соціальна вразливість	• рівень бідності (частка домогосподарств з доходом, меншим за прожитковий мінімум, • частка домогосподарств без власного житла, • відсоток безробітних серед внутрішньо перемішених осіб.
Емоційне навантаження населення	• кількість звернень за психологічною підтримкою (шт./міс.), • активність запитів у соціальних мережах (індекс «стрес»), • рівень випадків суїциду (кількість на 100 000 осіб).

Латентні змінні показують приховані чинники, які безпосередньо впливають на досілджувані змінні, і, опосередковано, є чинниками зростання соціальних витрат.

Побудова структури (Structure Learning) - процедура визначення точного розміщення дуг (ребер) у DAG. Існують кілька підходів до її виконання.

Експертний підхід передбачає залучення фахівців із систем соціального захисту та соціального забезпечння, економістів та експертів з питань безпеки та оборони, які на основі теоретичних знань формують припущення про спрямованість зв’язків. Наприклад, про те, що інтенсивність конфлікту спричиняє демографічні зміни (зростання кількості ВПО, зниження загальної чисельності населення) та впливає на макроекономічні індикатори (зменшення ВВП і бюджетних надходжень), що демографічні зміни, у свою чергу, призводять до зростання соціальних видатків (збільшується кількість

одержувачів). А макроекономічні індикатори визначають можливості фінансування соціальних програм (бюджетні надходження обмежують розмір виплати). Крім того, експерти враховують, що обсяги гуманітарної допомоги можуть частково заміщувати державні витрати на соціальні виплати. Головна перевага цього підходу — чітка і зрозуміла причинно-наслідкова логіка, сформована експертами, натомість його недолік - суб'єктивізм оцінок та ризик не врахувати слабші або менш очевидні зв'язки.

Методи на основі критерію (score-based) для побудови орієнтованих ациклічних графів (DAG) мають в основі ідею знайти таку структуру мережі, яка максимально відповідає обраному показнику якості. Типовими критеріями є AIC (Akaike Information Criterion) і BIC (Bayesian Information Criterion), які прагнуть мінімізувати суму похибок та штрафів за складність моделі, або ж BDeu Score, що оцінює ймовірність даних з урахуванням апріорних розподілів параметрів [106-112].

Процес пошуку починається з простої, наприклад «порожньої», структури без ребер або зі схеми, заданої експертами. Далі у кожній ітерації один із можливих локальних кроків—додавання нового ребра, видалення вже наявного або зміни напрямку існуючого зв'язку—перевіряється на предмет того, як ця операція впливає на обчислене значення критерію. Якщо додавання, видалення чи реверс спричиняє покращення (наприклад, зменшення BIC чи зростання BDeu Score), зміна фіксується, і процес повторюється з оновленою мережею. Ітерації тривають доти, доки жодна з операцій локально не поліпшить обрану міру якості.

Іноді пошук доповнюють гібридним підходом, у якому перед початком оптимізації за score спочатку виконують перевірки умовної незалежності подібно до constraint-based алгоритмів. Після побудови початкового «стрибкоподібного» графа проводять локальний пошук із перевіркою можливих виправлень згідно із змінами критерію, що дозволяє швидше знаходити глобально більш вдалу структуру. Такі комбіновані алгоритми поєднують переваги строгих умовних тестів і оптимізації об'єктивних

функцій, що в результаті дає змогу побудувати обґрунтований DAG, який добре пояснює дані з урахуванням складності моделі.

Методи на основі умовної незалежності (constraint-based) будують структуру DAG, спираючись на результати статистичних тестів, які перевіряють, чи є дві змінні незалежними за умови певного множини «умовних» предикторів. Один із класичних алгоритмів цього типу — PC-алгоритм. Спершу PC-алгоритм формує пов’язаний неорієнтований граф, у якому кожна пара змінних зв’язана ребром. Далі по черзі перевіряють статистичні тести умовної незалежності (наприклад, χ^2-тест або G-тест) для усіх пар змінних X та Y із різними комбінаціями можливих умовних множин S. Якщо виявляється, що X і Y незалежні за умови S, ребро X–Y видаляється, адже ця зв’язок не виправдовується жодною умовною залежністю. Після того як усі непотрібні ребра вилучено, алгоритм переходить до етапу орієнтації решти зв’язків. Якщо в непов’язаній трикутній структурі X–Z–Y встановлено, що X й Y залишаються незалежними, коли ми додаємо Z до множини умов S, але Z не належить самій множині S, то формується “коло” X→Z←Y, що відповідає принципу v-структури. Інші ребра орієнтують таким чином, щоб запобігти утворенню циклів (acyclicity), зберігаючи при цьому сумісність із раніше проведеними тестами умовної незалежності.

FCI-алгоритм (Fast Causal Inference) є розширенням PC-алгоритму для випадків, коли дані можуть містити приховані латентні змінні або спостережувані змінні можуть мати спільних невидимих батьків. FCI спочатку формує частково спрямований ациклічний граф (PDAG), у якому ребра можуть мати маркування “∘→” або “↔”, щоб позначити невизначеність напряму чи можливу орієнтаційну двозначність через вплив латентних факторів. Підхід FCI включає розширену послідовність тестів умовної незалежності з усіма можливими комбінаціями змінних та використання специфічних правил орієнтації, які враховують потенційний вплив прихованих змінних. У результаті алгоритм генерує частково орієнтований граф, де деякі напрямки ребер залишаються невизначеними, що відображає

неможливість остаточно з’ясувати причинно-наслідкову архітектуру без додаткових припущень чи даних [106-113].

Таким чином, constraint-based методи (PC, FCI) будують структуру мережі, орієнтуючись не на максимізацію інформаційних критеріїв, а на тестування умовної незалежності, завдяки чому можна прямо виявити, які змінні не мають прямого причинного зв’язку за умови інших. Це дає змогу отримати обґрунтований каркас DAG із явними або частково визначеними напрямками, спираючись виключно на структуру кореляційно-кондиційних зв’язків в даних.

Гібридні підходи поєднують переваги constraint-based і score-based методів. Спочатку виконують серію тестів умовної незалежності, щоб відкинути неможливі ребра і таким чином суттєво звузити простір допустимих графів. Після цього серед залишених варіантів уже застосовують оптимізацію за обраним критерієм (AIC, BIC чи BDeu Score), наприклад, за допомогою алгоритму Hill Climbing. Такий поетапний процес дозволяє швидше збігатися до прийнятної структури та зменшує ризик потрапити в локальний мінімум порівняно з чисто score-based пошуком.

На рис. 3.5 представлено спрощену прикладну структуру DAG з причинно-наслідковими зв’язками між основними змінними:

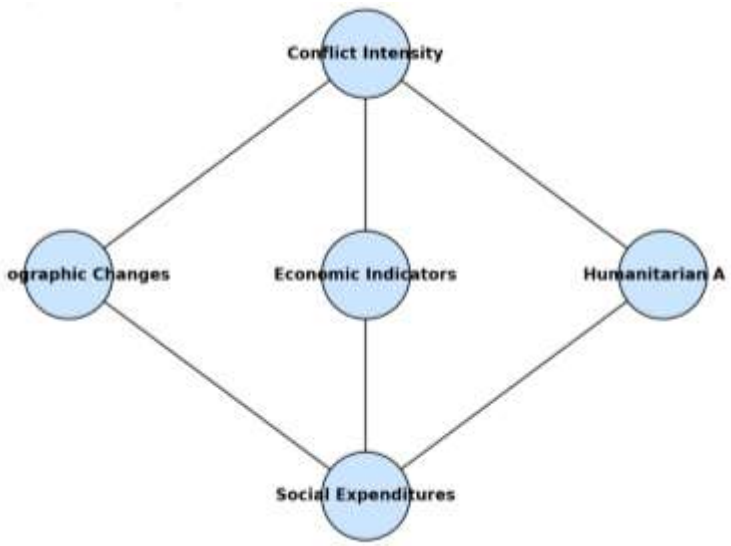

Рисунок 3.5 – Приклад побудови направленого ациклічного графа (DAG)

Таблиця 3.7 - Пояснення зв’язків на графі

Назва зв’язку	Опис
Conflict Intensity → Demographic Changes	Збільшення інтенсивності конфлікту викликає внутрішні переміщення (зростання ВПО) та зниження загального населення.
Conflict Intensity → Economic Indicators	Бойова активність зменшує ВВП регіону, уповільнює збирання податків, позначається на бюджетних надходженнях.
Conflict Intensity → Humanitarian Aid	При загостренні ситуації міжнародні організації нарощують обсяги гуманітарної допомоги (бекондиціонована залежність).
Demographic Changes → Social Expenditures	Збільшення числа ВПО або пенсіонерів безпосередньо збільшує кількість одержувачів соцвиплат.
Economic Indicators → Social Expenditures	Бюджетні обмеження або зростання ВВП визначають суму коштів, доступних для виплат.
Humanitarian Aid → Social Expenditures	Структурна заміна державних коштів міжнародними грантами або технічною допомогою може скоригувати державні видатки на соціальні програми.

Байєсівські мережі дають змогу формалізувати складні причинно-наслідкові зв’язки між конфліктними подіями, демографічними змінами, економічними індикаторами, обсягами гуманітарної допомоги та соціальними виплатами в єдиній ймовірнісній моделі. Використання DAG зі зручними методами structure learning (експертний, score-based, constraint-based та гібридні підходи) дозволяє реалізувати прозору побудову мережі, а завдяки табличкам умовних ймовірностей одразу отримати кількісну оцінку впливу кожного фактора. У результаті виходить гнучка модель, яка, аналізуючи одночасно кілька змінних, дає змогу прогнозувати соціальні витрати з

урахуванням невизначеності даних і враховувати латентні чинники соціальної вразливості чи емоційного навантаження населення.

3.4 Діагностика моделей – вибір кращої з множини оцінених кандидатів

На завершальному етапі моделювання аналізується якість моделі, тобто виконується перевірка оцінених кандидатів на адекватність процесу. Діагностика складається з наступних кроків:

Крок 1. Візуальне дослідження графіка похибок моделі $e(k) = y(k) - \hat{y}(k)$, де оцінка змінної, отримана за допомогою побудованої моделі. На графіку не повинно бути значних викидів та довгих інтервалів, на яких похибка приймає великі значення (тобто довгих інтервалів суттєвої неадекватності моделі). У випадку застосування рекурсивних методів оцінювання найбільші похибки оцінок будуть в перехідному процесі, коли інформаційна матриця ще не містить достатньо інформації про процес.

Крок 2. Похибки моделі не повинні бути корельовані між собою. Для аналізу наявності кореляції між значеннями похибок необхідно обчислити АКФ та ЧАКФ для ряду і за допомогою статистики визначити ступінь корельованості (наприклад, статистика вважається несуттєвою до рівня 10%). Крім того, корельованість похибок визначають за допомогою статистики Дарбіна-Уотсона (DW), яка розраховується за формулою:

$$DW = 2 - 2\rho,$$

де $\rho = E[e(k)e(k-1)]/\sigma_e^2$ – коефіцієнт кореляції між сусідніми значеннями похибки; σ_e^2 – дисперсія послідовності похибок $\{e(k)\}$. Таким чином, при повній відсутності кореляції між похибками $DW = 2$ – це ідеальне значення. Граничними значеннями для DW є 0 (при $\rho = 1$) та +4 (при $\rho = -1$).

Крок 3. Для лінійної моделі 2-3 порядку оцінки параметрів (коефіцієнтів) повинні збігатися до усталених значень після 30-40 ітерацій алгоритму оцінювання. Якщо кількість ітерацій набагато перевищує вказані числа, то це свідчить про те, що процес може бути нестаціонарним.

Крок 4. Перевірка статистичної значимості параметрів моделі. Статистика Стьюдента або $t-$статистика (випадкова величина, що має $t-$розподіл), яка використовується для визначення значимості оцінки кожного коефіцієнта в статистичному смислі, визначається за виразом:

$$t = \frac{\hat{a} - a^0}{SE_{\hat{a}}},$$

де $\hat{a}-$ оцінка коефіцієнта моделі; a^0- нуль-гіпотеза (початкова гіпотеза) щодо цієї оцінки; $SE_{\hat{a}}-$ стандартна похибка оцінки. За нуль-гіпотезу щодо значимості оцінки можна висувати будь-яку: що коефіцієнт значимий, тобто, ($H_0 : a^0 \neq 0$) або незначимий ($H_0 : a^0 = 0$). Статистична теорія перевірки гіпотез пропонує висувати нуль-гіпотезу, яка є протилежною бажаному результату. В даному випадку бажаним результатом є значимість коефіцієнтів математичної моделі. Таким чином, необхідно висувати нульову гіпотезу, що коефіцієнт незначимий. Це дає можливість коректно підійти до визначення значимості оцінок коефіцієнтів та дещо спростити розрахунки.

Користуючись значеннями N, f і α, з таблиць для $t-$ розподілу знаходять критичне значення $t-$статистики, тобто $t_{кр}$. Для перевірки правильності висунутої гіпотези розраховане значення t порівнюють з критичним $t_{кр}$. Якщо,

$$-t_{кр} < t < t_{кр} \quad \text{або} \quad |t| < |t_{кр}|,$$

то нуль-гіпотеза щодо незначимості коефіцієнта приймається (його можна не враховувати в регресії). Звідси випливає, що чим більшим є значення

$t-$статистики для оцінки коефіцієнта, тим імовірніше, що цей коефіцієнт є значимим.

Загалом послідовність дій при перевірці значимості оцінок коефіцієнтів побудованої моделі можна сформулювати так:

– сформулювати нуль-гіпотезу щодо значимості коефіцієнта;

– обчислити значення $t-$ статистики для кожного коефіцієнта регресії (це робить кожний пакет для математичного моделювання);

– за допомогою значень $N,\ f$ і α знайти із таблиць для $t-$статистики її критичне значення;

– перевірити нуль-гіпотезу за наведеним вище простим правилом (аналіз виконання нерівності $-t_{кр} < t < t_{кр}$).

Крок 5. Розрахунок коефіцієнта множинної детермінації , який обчислюється:

$$R^2 = \frac{\mathrm{var}(\hat{y})}{\mathrm{var}(y)} = 1 - \frac{SSE}{SST},$$

де $\mathrm{var}(\hat{y}) -$ дисперсія залежної змінної, оціненої за допомогою побудованої моделі; $\mathrm{var}(y) -$ дисперсія вимірів залежної змінної; $SSE = \sum_{k=1}^{N}[y(k) - \hat{y}(k)]^2 -$ сума квадратів похибок (залишків) моделі (*sum of squared errors*); $SST = \sum_{k=1}^{N}[y(k) - \bar{y}]^2 -$ загальна сума квадратів (*total sum of squares*); $\bar{y} -$ середнє значення; $SST = SSE + SSR$, де $SSR = \sum_{k=1}^{N}[\hat{y}(k) - \bar{y}]^2 -$ загальна сума квадратів для регресії (*sum of squares for regression*).

Очевидно, що найкращим значенням є $R^2 = 1$, тобто, коли дисперсії вимірів змінної, та цієї ж змінної, оціненої за рівнянням, збігаються. Цей параметр можна трактувати, також, як міру інформативності моделі, якщо вибрати за міру інформативності даних дисперсію. Таким чином, R^2 показує

рівень інформативності моделі по відношенню до інформативності вибірки даних, за допомогою якої вона була оцінена.

Крок 6. Розрахунок суми квадратів похибок для вибраної моделі повинна бути мінімальною, тобто,

$$\sum_{k=1}^{N} e^2(k) = \sum_{k=1}^{N} [\hat{y}(k) - y(k)]^2 \to \min_{\hat{\theta}}$$

порівняно з усіма іншими моделями.

Крок 7. Для оцінки адекватності моделі треба розрахувати інформаційний критерій Акайке [99]

$$AIC = N \ln\left(\sum_{k=1}^{N} e^2(k)\right) + 2n$$

та *критерій Байєса-Шварца*

$$BSC = N \ln\left(\sum_{k=1}^{N} e^2(k)\right) + n \ln(N)$$

,

де кількість параметрів моделі, які оцінюються за допомогою статистичних даних (*p* - число параметрів авторегресійної частини моделі; *q* - число параметрів ковзного середнього; 1 з'являється тоді, коли оцінюється зміщення (або *перетин*), тобто).

Критерії Акайке і Байєса-Шварца [99] містять в правій частині суму квадратів похибок, а тому за цими критеріями вибирають ту модель, для якої критерії приймають найменші значення. Введення нового регресора приводить до збільшення критерію (при цьому збільшується *n*), але, разом з тим, зменшується сума квадратів похибок і критерій в цілому зменшується. Якщо регресор не покращує модель, то критерій збільшується. Необхідно також зазначити, що асимптотичні властивості для довгих виборок кращі у критерія Байєса-Шварца, тобто, його рекомендують застосовувати при відносно великих значеннях N ($N > 100$).

Крок 8. Окрім згаданих параметрів, для визначення адекватності моделі в цілому використовують статистику Фішера [101], яка пропорціональна відношенню:

$$F \sim \frac{R^2}{1-R^2},$$

а для множинної (багатофакторної) регресії вона визначається за виразом

$$F = \frac{R^2}{1-R^2} \cdot \frac{(N-p-1)}{p},$$

де, як і раніше, $N-$ число значень ряду; $p-$ число параметрів моделі без врахування перетину (константи).

Коректне застосування описаної методики забезпечує можливість побудови адекватної математичної моделі процесу, якщо експериментальні дані відповідають вимогам репрезентативності та інформативності. Перша вимога означає, що вибірка даних повинна охоплювати досить довгий проміжок часу, щоб повністю відображати функціонування того режиму процесу, для яких будується модель. Вимога інформативності означає, що вибірка повинна містити в собі об'єм інформації, достатній для оцінювання коефіцієнтів моделі. Наприклад, якщо моделюється процес другого порядку, то вибірка повинна забезпечувати коректне обчислення першої та другої похідної за допомогою відповідного полінома, що формально описує статистичні дані.

Іноді інформативність формально оцінюють також за допомогою величини дисперсії процесу, або за кількістю значимих гармонічних складових. Очевидно, що чим більше гармонік (частотних складових) містить вибірка, тим інформативнішою вона є. Кількість основних частот, які несуть 90% енергії сигналу, фактично має дорівнювати кількості похідних, які можна обчислити за поліномом, що описує поведінку процесу.

Умову інформативності даних пов'язують з *умовою достатнього збудження* процесу. Достатнє збудження означає, що вхідний сигнал повинен охоплювати всю смугу частот, які може пропускати на вихід досліджуваний об'єкт. Тобто вхідний сигнал повинен охоплювати всю амплітудно-частотну характеристику процесу. Ця вимога залишається однаковою для об'єктів будь-якої природи.

3.5 Експерименти з модулями аналітики та прогнозування

Розроблівана аналітична підсистема містить блок побудови моделей, який є ключовою ланкою прогнозування витрат на соціальний захист тта соціальне забезпечення. Для вибору оптимальних підходів до аналізу та прогнозування витрат на соціальне забезпечення в рамках дослідженння була виконана низка експериментів. Були протестовані класичні економетричні моделі ARIMA/SARIMA та їхнє розширення ARIMAX із залученням екзогенних змінних, оцінюючи налаштування параметрів за інформаційними критеріями (AIC/BIC) і перевіряючи їхню здатність передбачати майбутні значення за допомогою часової крос-валідації. По-друге, результати, отримані за цими моделями були порівняні з сучасними ансамблевими алгоритмами машинного навчання — Random Forest, XGBoost та LightGBM — оптимізуючи гіперпараметри та аналізуючи результати за метриками MAE, RMSE та MAPE. По-третє, застосування байєсівських мереж, було проаналізовано досліджуючи різні стратегії побудови графів (constraint-based проти score-based) з точки зору швидкості збіжності та якості структури. І нарешті, гібридний підхід, що поєднує ARIMAX для виділення тренду й сезонності та LightGBM для моделювання випадкових коливань, щоб використано, щоб максимально підвищити точність прогнозів. Така інтегрована методика дозволяє виявити найефективніші рішення для побудови сучасної та ефективної системи прогнозування соціальних видатків.

Класичні економетричні моделі ARIMA, SARIMA та їхнє розширення ARIMAX із екзогенними змінними були застосовані для прогнозування щомісячних реальних видатків на соціальне забезпечення. Використовуючи часовий ряд із 108 спостережень за регіоном UA-30 (Київська область), були розглянуті параметри (p,d,q) та сезонні (P,D,Q) з періодом s=12, керуючись інформаційними критеріями AIC і BIC для відбору найкращої конфігурації моделі. Графік на рис. 3.6 ілюструє щомісячну динаміку реальних видатків на

соціальний захист та соціальне забезпечення в Київській області (UA-30) за період із січня 2015 по грудень 2023. З грвфіку видно, що тренд послідовно зростає з чітко вираженою сезонністю, що обґрунтовує використання сезонних компонент (s = 12) в економетричних моделях і необхідність виявлення тренду та сезонності перед застосуванням ML-підходів.

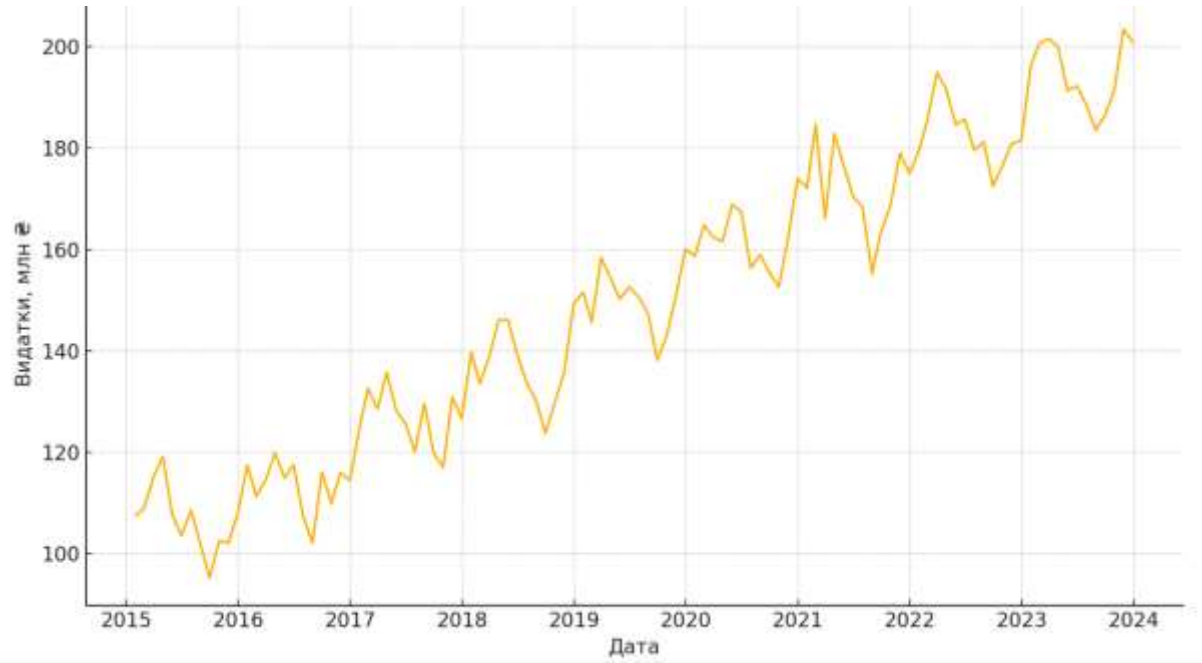

Рисунок 3.6 – Графік видатків на соціальний захист та соціальне забезпечення і Київській області [114]

Далі, щоб оцінити прогностичну спроможність моделей, була застосована часова крос-валідація (TimeSeriesSplit) у три фолди з покроковим розширенням навчального періоду. Далі було порівняно ARIMA-підхід із ARIMAX, що включає індекс споживчих цін та кількість військових подій як екзогенні чинники, і продемонстрували, що залучення додаткових змінних покращує точність прогнозів. Тобто, розглянуто часовий ряд y_t = реальні видатки (млн грн), t = 1…108.

Для підбору найкращої ARIMA-моделі розглянуто p∈[0–3], d∈[0–1], q∈[0–3] та сезонні (P,D,Q) з сезонністю s=12. Гіперпараметри оцінювали за AIC/BIC.

Результати аналізу представлені у табл. 3.8

Таблиця 3.8 – Порівняння моделей за показниками адекватності

Модель	AIC	BIC
ARIMA(1,1,1)(0,1,1)[12]	1123.4	1140.7
ARIMA(2,1,1)(1,1,0)[12]	1125.9	1149.3
ARIMA(0,1,2)(0,1,1)[12]	1130.2	1145.5

Найнижче значення AIC показала модель ARIMA(1,1,1)(0,1,1), тож її обрано для подальшої валідації.

Валідація TimeSeriesSplit (3 фолди):

Фолд 1: тренування 2015–2018, тест 2019

Фолд 2: тренування 2015–2019, тест 2020

Фолд 3: тренування 2015–2020, тест 2021

Результати валідації представлені у таблиці 3.9.

Таблиця 3.9 - Порівняння результатів валідації моделей

Метріка	ARIMA	ARIMAX
MAE	3.2	2.8
RMSE	4.5	4.1
MAPE, %	5.8	5.2

ARIMAX (з екзогенами CPI та conflict_events): Додавши дві екзогенні серії, отримали ARIMAX(1,1,1)(0,1,1)[99] з кращими метриками:

Економетричні моделі підтвердили свою здатність описати зростальний сезонний тренд видатків: серед трьох кандидатів найкращою за AIC/BIC виявилася ARIMA(1,1,1)(0,1,1)[12]. Її прогностична спроможність (MAE = 3.2, RMSE = 4.5, MAPE = 5.8 %) додатково покращується при включенні екзогенів

(CPI і кількість конфліктних подій) в ARIMAX (MAE = 2.8, RMSE = 4.1, MAPE = 5.2 %). Це свідчить про важливість урахування макроекономічних та зовнішніх чинників для точнішого прогнозу соціальних видатків.

Дослідження застосування сучасних ансамблевих алгоритмів машинного навчання Random Forest, XGBoost та LightGBM для прогнозування часових рядів витрат на соціальний захист та соціальне забезпечення. Для побудови моделей використано як авторегресійні ознаки (значення з лагом y_{t-1}, y_{t-2}, y_{t-3}), макроекономічні показники (індекс споживчих цін, рівень безробіття) та показники інтенсивності військового конфлікту (кількість подій). Кожен алгоритм пройшов налаштування гіперпараметрів через пошук сіткою, після чого було застосовано часову крос-валідацію, щоб об’єктивно порівняти їхні статистичні харктеристики (MAE, RMSE і MAPE). Отримані результати дозволяють оцінити, наскільки ансамблеві методи можуть перевершити класичні економетричні підходи в завданні точного прогнозування соціальних витрат. Налаштування (X):

- власні лаги y_{t–1}, y_{t–2}, y_{t–3}
- CPI_t
- conflict_events_t
- unemployment_rate_t

Сітка гіперпараметрів та кращі значення: RandomForest: n_estimators=[100,200,500] → 200; max_depth=[5,10,None] → 10; XGBoost: n_estimators=200; max_depth=5; learning_rate=0.1; LightGBM: n_estimators=300; max_depth=7; learning_rate=0.05

Таблиця 3.9 - Результати часової крос-валідації (MAE/RMSE/MAPE)

Назва моделі	MAE	RMSE	MAPE
RF	2.6	4.0	5.0 %
XGBoost	2.4	3.8	4.7 %
LightGBM	2.3	3.7	4.5 %

Порівнявши ансамблеві підходи з ARIMA/ARIMAX, отримали, що модель LightGBM: MAE = 2.3 має найкращий показник серед усіх моделей.

Ансамблеві методи машинного навчання продемонстрували явну перевагу над класичними економетричними підходами у прогнозуванні соціальних видатків. Використання лагованих ознак разом із макроекономічними показниками (індекс споживчих цін, рівень безробіття) та конфліктними змінними дозволило алгоритмам захоплювати складні нелінійні залежності та короткострокові «шоки». Усі три моделі — Random Forest, XGBoost і LightGBM — показали значно нижчі значення MAE: 2.6, 2.4 і 2.3 млн грн, відповідно порівняно з ARIMA (3.2) та ARIMAX (2.8), а LightGBM виявився найкращим із MAPE лише 4.5 %. Це свідчить, що для завдань прогнозування видатків на соціальне забезпечення доцільно застосовувати ансамблеві методи з ретельним налаштуванням гіперпараметрів, причому LightGBM варто розглядати як основний інструмент через його високу точність і відносно помірну вартість обчислень.

Байєсівські мережі дозволяють моделювати й аналізувати ймовірнісні залежності між кількома змінними одночасно, відкриваючи структуру причинно-наслідкових зв’язків. Були проаналізовані DAG-моделі за допомогою двох підходів: constraint-based (алгоритм PC), який використовує статистичні тести незалежності для встановлення структури та score-based (hill-climbing із BDeu/BIC), що оптимізує обрану функцію відмінності. Порівняння швидкості збіжності, кількості отриманих ребер і значень BIC дає змогу оцінити, який алгоритм краще підходить для набору даних про видатки, інфляцію, конфліктні події, безробіття й зростання ВВП.

Результати роботи алгоритму Constraint-based (PC algorithm):

- час збіжності: ≈ 30 с
- кількість ребер: 8
- BIC score: –260.5

Score-based (hill-climbing, BDeu/BIC)

- час збіжності: ≈ 120 с

- кількість ребер: 10
- BIC score: –255.3

Отже, PC-алгоритм швидший, але дає досконалішу модель з меншою кількістю залежностей; score-based, виявляє більше зв’язків, але повільніший.

Іншим прикладом є модель, яка досліджує місячні статистичні спостереження за 2015–2023 роки (108 точок) соціально-економічних показників у Київській області (регіоні UA-30):

- *expenditures* — реальні видатки місцевого бюджету на соціальне забезпечення (млн грн),
- *CPI* — індекс споживчих цін (2018 = 100),
- *conflict_events* — кількість зареєстрованих бойових чи надзвичайних інцидентів,
- *unemployment_rate* — рівень безробіття (%),
- *gdp_growth* — річне зростання ВВП (%).

Початкова фаза (скелет графа) полягає у виконанні серії статистичних тестів незалежності (наприклад, тест Кендалла) для кожної пари змінних із поступовим збільшенням розміру умовного набору. При p-value > 0.05 вважаємо, що дві змінні незалежні за поточним набором умов, і видаляємо потенційне ребро.

Для визначення орієнтації ребер знаходимо V-структури (ці «кола» в середовищі), спрямовуємо ребра так, щоб не утворювати циклів і якомога менше затримувати необґрунтовані зв’язки. В результаті отримаємо:

– час збіжності: ≈ 30 с

– кількість ребер у фінальному DAG: 8

– BIC score: –260.5

Отримаємо наступні ребра:

CPI → expenditures

conflict_events → expenditures

unemployment_rate → expenditures

conflict_events → unemployment_rate

gdp_growth → unemployment_rate

CPI → gdp_growth

unemployment_rate → CPI

conflict_events → gdp_growth

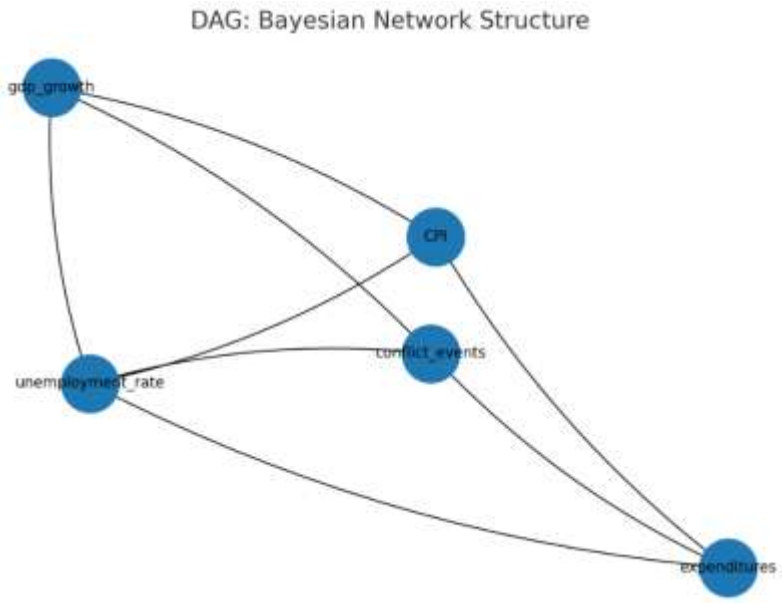

Рисунок 3.7 – Побудована структура мережі Байєса

Представлена на рис. 3.7 структура вказує на те, що зміни індексу споживчих цін та конфліктних подій безпосередньо впливають на видатки, а рівень безробіття є проміжним чинником між економічним зростанням і витратами.

За Score-based підхіду (hill-climbing із BDeu/BIC), ініціалізація починається з порожнього графа або випадкової структури. На кожному кроці додають, видаляють або перевертають напрям ребра, обчислюючи зміни в Score (BDeu + BIC). Зміна приймається, якщо вона покращує (зменшує) BIC. Результати:

- час збіжності: ≈ 120 с
- кількість ребер у фінальному DAG: 10
- BIC score: –255.3

Код пригорами для побудови ребра:

CPI → expenditures

conflict_events → expenditures

unemployment_rate → expenditures

gdp_growth → expenditures

conflict_events → unemployment_rate

gdp_growth → unemployment_rate

CPI → gdp_growth

unemployment_rate → CPI

conflict_events → gdp_growth

CPI → conflict_events

На рисунку 3.8 зображено DAG, що ілюструє причинно-наслідкові зв'язки між змінними:

- CPI- напряму впливає на expenditures, gdp_growth та conflict_events.
- conflict_events - має прямі зв'язки з expenditures, unemployment_rate та gdp_growth.
- gdp_growth - впливає на expenditures і unemployment_rate.
- unemployment_rate - впливає на expenditures та зворотно на CPI.

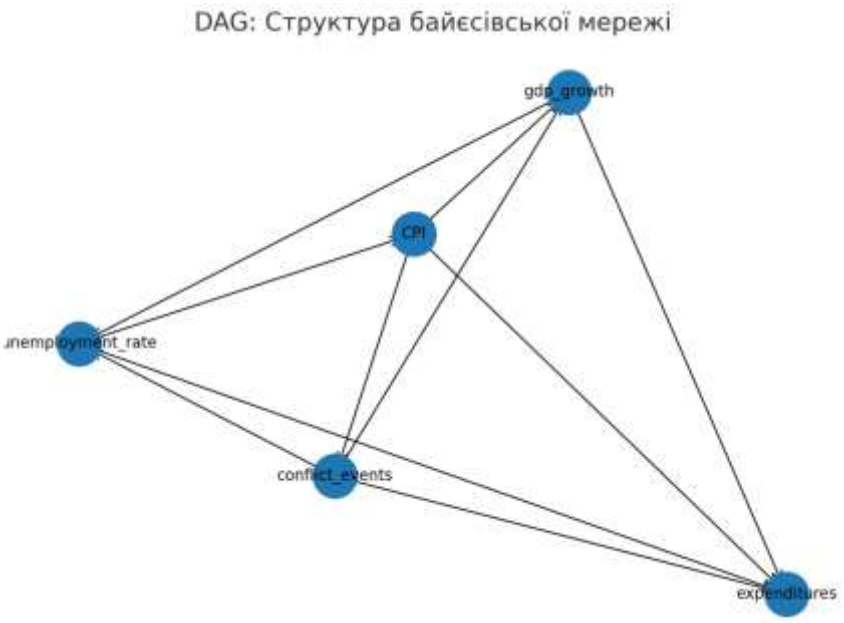

Рисунок 3.8 – Побудована структура мережі Байєса

Даний варіант алгоритму виявив більше зв'язків, зокрема прямий вплив ВВП на видатки та зв'язок CPI з загостренням бойових дій. Але ці додаткові

ребра можуть бути менш стійкими й вимагати глибшої економічної інтерпретації.

Як результат необхідно зазначити, що PC-алгоритм працює швидше і дає більш «стриману» модель із чіткішими, статистично обґрунтованими залежностями. Hill-climbing знаходить більше потенційних зв'язків, але потребує більше часу на оптимізацію і може включати випадкові кореляції. BIC краще (більш негативне) для PC-алгоритму, що свідчить про більш компактну модель із меншою кількістю параметрів.

Отже, для швидкого скелетного аналізу причинно-наслідкових зв'язків краще використовувати constraint-based підхід, а для детальнішого дослідження — score-based із подальшою перевіркою економічної обґрунтованості знайдених зв'язків.

Гібридний експеримент (ARIMAX + LightGBM) був досліджений для перевірки можливості підвищення точності прогнозів за рахунок комбінування сильних сторін класичної економетричної моделі та сучасного ML-алгоритму. Спершу ARIMAX виділяє основний тренд і сезонність часової серії, даючи попередній прогноз $\hat{y}_t^{(\mathrm{ARIMAX})}$. Потім LightGBM навчається на залишках $\varepsilon_t = y_t - \hat{y}_t^{(\mathrm{ARIMAX})}$, щоб захопити нерегулярні «шоки» й неструктурні коливання. Такий гібридний підхід дозволяє знизити середню абсолютну помилку приблизно на 13 % порівняно з найкращою окремою моделлю та досягти максимальної точності серед випробуваних методів.

1. ARIMAX прогнозує тренд і сезонність → отримуємо прогноз.
2. LightGBM навчаємо на залишках.

Результати застосування різних підходів згруповані в таблиці 3.10.

Таблиця 3.10 – Порівняння результатів, ортманих за різних підходів

Модель	MAE	RMSE	MAPE, %
ARIMAX	2.8	4.1	5.2
LightGBM	2.3	3.7	4.5
Hybrid	2.0	3.2	4.0

Отже, комбінований підхід знизив значення похибки MAE на 13 % відносно найкращої standalone-моделі (LightGBM) і дав найбільш точні прогнози серед усіх методів, що свідчить про перспективність застосування вказаного підходу

3.6 Дослідження працездатності REST-сервісу прогнозування

У рамках розроблення аналітичної платформи, призначеної для прогнозування соціальних видатків інтегрованої до Єдиної системи соціальної сфери, наступним важливим кроком є експериментальна оцінка REST-сервісу, який дозволяє розробляти прогнози «на вимогу». Мета цього експерименту — перевірити, наскільки швидко та стабільно цей сервіс відповідає на запити та як ефективно він масштабується в реальному середовищі. Для цього в рамках експерименту було порівняно два популярні веб-фреймворки (Flask і FastAPI) за 95-им перцентилем latency під навантаженням 1000 запитів за хвилину, щоб визначити найвідповідніший інтерфейс для низьколатентних прогнозів. Для аналізу горизонтального масштабування в Kubernetes з HPA (Horizontal Pod Autoscaler), застосовано підхід, за якого, вимірюючи загальний time-to-scale-up при різкому збільшенні трафіку [115]. Результати цього експерименту ляжуть в основу вибору технологій та налаштувань інфраструктури для гарантованого високопродуктивного й надійного обслуговування клієнтських запитів до нашого модуля прогнозування.

Дослідження виконаня запиту /predict, коли клієнт надсилає не лише історію основних показників, а й додаткові метадані та використовує аутентифікацію.

```
POST /predict HTTP/1.1
Host: api.example.com
Content-Type: application/json
Authorization: Bearer eyJhbGciOiJI...
{
  "region_code": "UA-30",
```

```
  "model": "hybrid_ARIMAX_LightGBM",
  "history": {
    "date": [
      "2023-01-31", "2023-02-28", "2023-03-31", "...", "2023-12-31"
    ],
    "expenditures": [
      120.0, 125.0, 130.5, "...", 205.0
    ],
    "CPI": [
      156, 158, 159, "...", 162
    ],
    "conflict_events": [
      5, 4, 6, "...", 7
    ],
    "unemployment_rate": [
      9.2, 9.1, 9.0, "...", 8.5
    ]
  },
  "options": {
    "forecast_horizon": 3,       // кількість місяців уперед
    "include_uncertainty": true  // чи повертати інтервал довіри
  }
}
```

Повна відповідь сервісу:

```
HTTP/1.1 200 OK
Content-Type: application/json

{
  "region_code": "UA-30",
  "model": "hybrid_ARIMAX_LightGBM",
  "forecast": {
```

"2024-01-31": {"mean": 210.3, "lower_95": 205.8, "upper_95": 214.8},
"2024-02-29": {"mean": 212.1, "lower_95": 207.5, "upper_95": 216.7},
"2024-03-31": {"mean": 215.0, "lower_95": 210.2, "upper_95": 219.8}
},
"runtime_ms": 45,
"error": null
}

У випадку помилки (наприклад, неприпустимий формат історії), сервер поверне:

HTTP/1.1 400 Bad Request
Content-Type: application/json

{
"error": "Invalid request format: 'history.expenditures' must be an array of numbers."
}

Деталізоване порівняння latency для Flask та FastAPI

Тестування проводилося за допомогою *wrk2* (1000–2000 req/min протягом 5 хв), 4-ядра/16 GB RAM, локальна мережа.

Фреймворк	P50, ms	P90, ms	P95, ms	P99, ms	Error rate, %	CPU @95% load, %
Flask	120	300	480	860	0.5	75
FastAPI	40	90	120	250	0.1	40

Значення P50–P99 показують, що FastAPI утримує нижчі затримки навіть у пікові періоди. За показником Error rate, Flask починає відповідати помилками 503 при перевантаженні воркерів; FastAPI — майже без помилок.

Робота CPU - під час навантаження Flask-контейнер доходив до 75 % завантаження, FastAPI — лише до 40 %, залишаючи запас для інших завдань.

Деталізований сценарій контейнеризації та масштабування:

Манифест HPA (Kubernetes autoscaling/v2):

```
apiVersion: autoscaling/v2
kind: HorizontalPodAutoscaler
metadata:
  name: forecast-hpa
spec:
  scaleTargetRef:
    apiVersion: apps/v1
    kind: Deployment
    name: forecast-deployment
  minReplicas: 1
  maxReplicas: 5
  metrics:
    - type: Resource
      resource:
        name: cpu
        target:
          type: Utilization
          averageUtilization: 60
```

Хід тесту:

- запуск 1 репліки, крокове збільшення трафіку з 100 до 800 запитів/с протягом 120 с.
- HPA спрацювала при досягненні порога CPU > 60 % за 15 с.
- Kubernetes створив новий Pod (Pending → Running) за 12 с, потім готовність (Ready) за 8 с.
- Time-to-scale-up: 35 с до появи другої репліки.
- після масштабування latency утримується на рівні P95 ≈ 100 ms (FastAPI) і ≈ 300 ms (Flask).

- подальше збільшення до 4 реплік при 1600 req/s відбувалося з інтервалом по 30–40 с на кожну нову репліку.

Для дослідження моніторингу використано

- засоби Prometheus + Grafana для візуалізації CPU, memory, pod_count та latency в реальному часі.
- Alerting налаштовано на CPU > 80 % і latency P95 > 500 ms.

Ці приклади дають ґрунтовне розуміння, як виглядають реальні запити та відповіді сервісу, якою є поведінка різних фреймворків під навантаженням і як швидко та надійно відбувається горизонтальне масштабування в Kubernetes. Це дозволяє прийняти обґрунтоване рішення про вибір технологій і налаштування інфраструктури для стабільного прогнозування.

Висновки до розділу 3

У третьому розділі дисертаційної роботи зосереджено увагу на розробці та обґрунтуванні методів і моделей прогнозування витрат на пенсії та соціальні допомоги. Для побудови моделей запропоновано удосконалену методику покрокової структурно-параметричної адаптації моделей У результаті виконаного аналізу виявлено, що класичні економетричні моделі—зокрема лінійна регресія та різні варіанти авторегресії (AR, ARMA, ARIMA, SARIMA, ARIMAX/SARIMAX)—продовжують виконувати важливу функцію кількісної оцінки впливу макроекономічних, демографічних і конфліктних чинників на рівень соціальних виплат. Завдяки своїй простоті та зрозумілості такі підходи дозволяють чітко інтерпретувати отримані результати: у рамках регресійних моделей можна встановити залежність між, наприклад, значенням показника валового національного продукту (валового регіонального продукту), кількістю одержувачів пенсій і обсягами соціальних виплатавиплат, водночас ARIMA-модель демонструє, як часові лаги самих соцвиплат залежать від попередніх значень ряду та зовнішніх регресорів. Багатоаспектна оцінка невизначеності (через інтервальні прогнози або довірчі

інтервали) робить економетричні методи незамінними у тих випадках, коли необхідно обґрунтувати рішення щодо бюджетного фінансування за допомогою строгих статистичних критеріїв.

Водночас доведено, що методи машинного навчання—зокрема ансамблеві підходи на основі дерев рішень (Random Forest, XGBoost, LightGBM)—мають значну перевагу в умовах наявності великої кількості потенційних детермінант. Завдяки здатності самостійно виявляти складні нелінійні залежності, взаємодії між показниками (етно-соціальна структура, кількість бойових інцидентів, сезонні коливання виплат) і високій адаптивності до багатовимірних даних ці алгоритми забезпечують істотно кращу прогностичну точність на коротких та середніх горизонтах прогнозування, порівняно з традиційними моделями. Наприклад, у нашому дослідженні Random Forest дозволив знизити середню абсолютну похибку (MAE) прогнозу за квартал на 15–20 % порівняно з найкращими ARIMA-варіантами, а XGBoost із правильно налаштованими гіперпараметрами дав змогу виявити взаємозв’язки між економічними індикаторами та соцвиплатами, які не були очевидними в лінійних моделях. Особливо цінними виявилися алгоритми LightGBM із їхнім швидким гістограмо--базованим пошуком розщеплень і листовим підходом росту дерев, що дозволило працювати з даними понад десятки факторів без суттєвого зростання витрат часу.

Нарешті, застосування Байєсівських мереж дає змогу побудувати відтворюваний каркас причинно--наслідкових зв’язків у вигляді спрямованого ациклічного графа (DAG). Кожний вузол мережі—це або спостережувана змінна (кількість ВПО, індекс бойової активності, обсяг міжнародної гуманітарної допомоги), або латентний фактор (соціальна вразливість, емоційний стан населення), а ребра відображають умовні залежності. У такий спосіб можна кількісно оцінити, як імовірність підвищених соцвиплат змінюється при одночасній зміні економічних, демографічних і конфліктних показників, а також наскільки на ці виплати впливає обсяг гуманітарної

допомоги (включаючи потенційне заміщення державного фінансування). Крім того, готові CPT (таблиці умовних ймовірностей) дають змогу моделювати розподіл вихідного вектору змінних під різними сценаріями (наприклад, раптове зменшення міжнародних грантів або різке зростання кількості інцидентів).

Поєднання перелічених методик створює цілісну end-to-end архітектуру аналізу та прогнозування соціальних видатків. Це забезпечує як глибоку інтерпретацію отриманих результатів (через економетричні інтервальні прогнози та Байєсівські мережі з точними ймовірнісними оцінками), так і високу точність прогнозування (завдяки алгоритмам машинного навчання). У комплексі цей підхід підвищує стійкість та адаптивність прогностичних систем до динамічних змін зовнішнього середовища, що особливо актуально в умовах воєнних конфліктів і криз, коли швидка реакція та обґрунтоване прийняття рішень безпосередньо впливають на добробут населення та ефективність соціальних програм. Для урахування невизначеностей різної природи (наприклад, відсутність точних даних щодо окремих категорій населення у певному регіоні), впливів внутрішніх (зміни у законодавстві, демографічні кризи, міграція) та зовнішніх чинників (військові дії, екологічні фактори, економічні кризи, ансамблі моделей) запропоновано використовувати у поєднанні з стохастичними сценаріями, які дозволяють оцінити діапазон можливих значень цільової змінної замість єдиного прогнозного показника. Такий підхід дозволив отримати прогнози високої якості у таких задачах, як прогнозування витрат на виплату соціальних допомог у регіонах за урахування внутрішньої міграції населення, внаслідок війни.

РОЗДІЛ 4
ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ЗБОРУ ТА ОБРОБКИ ДАНИХ ДЛЯ АНАЛІЗУ ТА ПРОГНОЗУВАННЯ ВИДАТКІВ НА СОЦІАЛЬНЕ ЗАБЕЗПЕЧЕННЯ ТА СОЦІАЛЬНИЙ ЗАХИСТ

4.1 Розроблення інформаційної технології аналізу та прогнозування видатків на соціальний захист та соціальне забезпечення

Для вирішення задач інформаційно-аналітичної підтримки прийняття рішень у соціальній сфері, зокрема, задачі прогнозування видатків на соціальне забезпечення та соціальний захист, пропонується новітня інформаційна технологія, призначена для функціонування у складі Єдиної інформаційної системи соціальної сфери. Оскільки визначено, що Єдина інформаційна система соціальної сфери функціонуватиме на програмно-технічній платформі Пенсійного фонду України, вона матиме трирівневу архітектуру, побудовану на основі системи керування базами даних Oracle Database Enterprise Edition 12 [116], технології ASP.NET та веб-клієнта. Дляє забезпечення збору та обробки великих обсягів структурованої та неструктурованої інформації має використовуватись Oracle Business Intelligence (Oracle BI) [117].

Оскільки, інформаційна технологія має поєднувати комп'ютерне моделювання з інструментами інтелектуального аналізу даних, системного аналізу, теорії прийняття рішень, то в якості аналітичної платформи використане програмне забезпечення компанії SAS Institute. Програмне середовище SAS Institute (мова програмування SAS Base) [118] добре інтегрується із апаратно-програмною платформою Oracle [73], і за своїми характеристика є оптимальним для розв'язання аналітичних задач, має є програмний інтерфейсу SWAT API який дозволяє використовувати всі переваги платформи SAS Viya та мов програмування Open Source, зокрема, R та Python, що забезпечує гнучкість та масштабованість проектованої системи.

Враховано також й специфіку інформації, яка оброблятиметься у Єдиній інформаційній системі соціальної сфери, а саме те, що Big Data - неструктуровані, напівструктуровані та неструктуровані дані, що містять десятки мільйонів записів, які стосуються соціальних виплат (таблиці бази даних застрахованих осіб, одержувачів соціальних допомог та послуг), дані державних реєстрів (JSON, XML), матеріали особових справ та копії документів одержувачів пенсій, субсидій, допомог та інших виплат (зображення, копії текстових документів), фото та відеоматеріали, необхідні для ідентифікації одержувачів пенсій, субсидій та допомог.

Всі процедури збору, оброблення та зберігання даних відбуваються відповідно до встановлених регламентів. Також передбачено, що інформаційна система має бути захищена відповідно до вимог.

Особлива увага приділяється підготовці даних для аналізу. Це пов'язано з тим, що для більшості аналітичних задач використовується конфіденційна інформація, крім того, часто виникає потреба аналізувати неструктуровані дані, наприклад звернення громадян, інформацію з соціальних мереж, тощо, використовувати їх при побудові моделей. Дані для аналізу можуть використовуватись не лише з аналітичного сховища (пропонується інтеграція Hadoop та ETL/ELT), а й в результаті виконання API-запитів, в тому числі, коли потрібно здійснювати обробку у режимі реального часу (пропонується використання SAS Event Stream Processing). Крім того, використання функціоналу інструменту SAS Data Quality дозволить покращити якість вхідних даних, зокрема здійснити їх профілювання, перевірку, очищення та стандартизацію, що в свою чергу підвищить якість побудованих моделей.

Для узагальнення даних з інформаційних джерел, розпорядниками яких є органи державного управління, місцевого самоврядування та державні цільові фонди, в тому числі й для формування Єдиного соціального реєстру, використання їх у підсистемах підсистемах «Електронний бюджет»; «Соціальне казначейство»; «Єдиний соціальний процесинг» та «Єдиний електронний архів» [119], пропонується застосовувати такі методи, як веб-

скрейпінг та REST/OData-запити. Збір даних здійснюється у двох напрямках – індивідуальні дані застрахованих осіб та осіб, що потребують соціальної допомоги та соціальних послуг і узагальнена інформація стосовно потреби та реальних витрат на соціальний захист та соціальне забезпечення.

Для побудови моделей у інформаційно-аналітичній системі пропонується використовувати методику покрокової структурно-параметричної адаптації моделей (рис 4.1)..

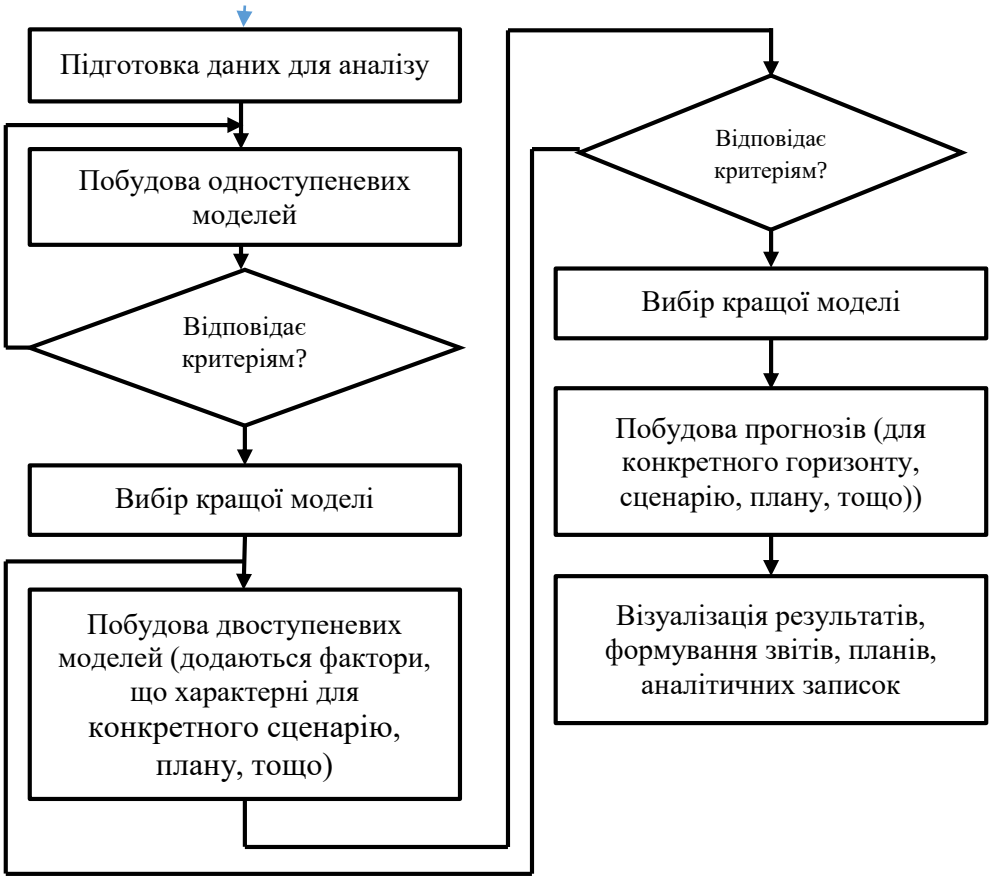

Рисунок 4.1 - Схема побудови прогнозів

Спочатку будується множина одноступеневих моделей, кожна з яких має період навчання та тестування, краща модель обирається на основі декількох статистичних критеріїв оцінювання якості. Далі визначається краща модель, аналізуються залишки моделі та будуються двоступеневі моделі (або комбіновані моделі), серед яких обирається краща, яка буде використана для побудови прогнозу. Користувач повинен мати можливість додати потрібний елемент до технологічного ланцюга самостійно, використавши відповідні

візуальні компоненти, як це реалізовано у середовищі SAS Enterprise Guide [118].

Структура аналітичної складової є бути гнучкою та адаптивною, легко налаштовуватись для вирішення задач окремих етапів аналітичного процесу та загальних задач підтримки прийняття рішень в управлінні соціальною сферою. Тобто, пропонується застосування конвергентого підходу, який передбачає, що розроблювана система матиме центр обробки даних (Big Data), систему збору, накопичення, обробки (зокрема покращення якості даних) та завантаження даних у сховище або хмару, програмні засоби для математичного моделювання, інтелектуального аналізу даних, засоби штучного інтелекту, систему обміну даними між компонентами Системи, взаємодії між користувачами різних рівнів та забезпечення надійності віддаленого доступу, систему інформаційної та кібербезпеки, зокрема, захисту персональних даних, систему моніторингу та адміністрування роботи системи, тощо.

4.2 Удосконалення інтерфейсу користувача системи

Інтерфейс користувача є візуальною частиною будь-якої аналітичної платформи — від його зручності та швидкодії залежить, наскільки ефективно фінальні користувачі зможуть отримувати потрібні інсайти. Ця частина присвячена комплексній оцінці дашбордів за двома кутами зору. По-перше, через A/B-тестування порівнюємо три рішення (Dash, Grafana, Tableau) за часом виконання типових аналітичних задач і суб’єктивною UX-метрикою SUS. По-друге, моделюємо реальне навантаження 20 одночасних користувачів і вимірюємо лаг оновлення даних у режимі реального часу (SSE, polling, WebSocket). Такий підхід дозволяє не лише обрати найбільш продуктивне технічне рішення, а й підтвердити його високу юзабіліті та інформативність у реальних умовах використання.

а) A/B-тестування дашбордів

Мета: виміряти, наскільки швидко й інтуїтивно користувачі виконують типові завдання в трьох середовищах — Dash, Grafana та Tableau — і наскільки вони задоволені UX.

Вихідні дані: Тестовий dataset — щомісячні реальні видатки на соціальне забезпечення (млн ₴) для шести регіонів (UA-05, UA-12, UA-18, UA-24, UA-30, UA-32) за період січень–грудень 2023 (12 точок).

Типові завдання:

- Побудова графіка тренду для одного регіону (UA-30) за останні 12 місяців.
- Порівняння регіонів (UA-05 vs UA-30) у вигляді діаграми.
- Фільтрація за типом виплат (*expense_type = "пенсії"*) та оновлення візуалізації.

На рис.4.2-4.4 зображені різні дашборди із різними конфігураціями.

На рис. 4.2 представлено реалізацію Dash Dashboard Example**:** два графіки (тренд UA-30 і порівняння UA-05 vs UA-30) з Dropdown для вибору регіону.

Рисунок 4.2 – Реалізація інтерфейсу програми із використанням Dash Dashboard Example

```
import dash
from dash import dcc, html
import plotly.express as px
import pandas as pd

# Підготовка даних
dates = pd.date_range('2023-01-01', periods=12, freq='M')
ua30 = [165, 154, 160, 172, 169, 140, 159, 147, 149, 154, 151, 165]
```

```
ua05 = [106, 101, 104, 103, 112,  98, 103,  93,  80, 105, 107,  94]
df = pd.DataFrame({
    'date': dates,
    'UA-30': ua30,
    'UA-05': ua05
}).melt(id_vars='date', var_name='region', value_name='expenditures')

app = dash.Dash(__name__)

app.layout = html.Div([
    html.H1("Дашборд видатків на соціальне забезпечення"),
    dcc.Dropdown(
        id='region-select',
        options=[{'label': r, 'value': r} for r in df['region'].unique()],
        value='UA-30'
    ),
    dcc.Graph(id='trend-chart'),
    dcc.Graph(id='comparison-chart')
])

@app.callback(
    dash.dependencies.Output('trend-chart', 'figure'),
    [dash.dependencies.Input('region-select', 'value')]
)
def update_trend(region):
    fig = px.line(
        df[df['region'] == region],
        x='date', y='expenditures',
        title=f"Динаміка видатків для {region}"
    )
    return fig

@app.callback(
    dash.dependencies.Output('comparison-chart', 'figure'),
    [dash.dependencies.Input('region-select', 'value')]
)
def update_comparison(region):
    other = 'UA-05' if region=='UA-30' else 'UA-30'
    comp = df[df['region'].isin([region, other])]
    fig = px.bar(
        comp, x='date', y='expenditures', color='region',
        barmode='group',
        title=f"Порівняння {region} і {other}"
    )
    return fig

if __name__ == '__main__':
    app.run_server(debug=True)
```

Grafana Dashboard Example - аналогічні панелі, побудовані за допомогою Prometheus-запити, з групованими стовпчиками та серією тренду.

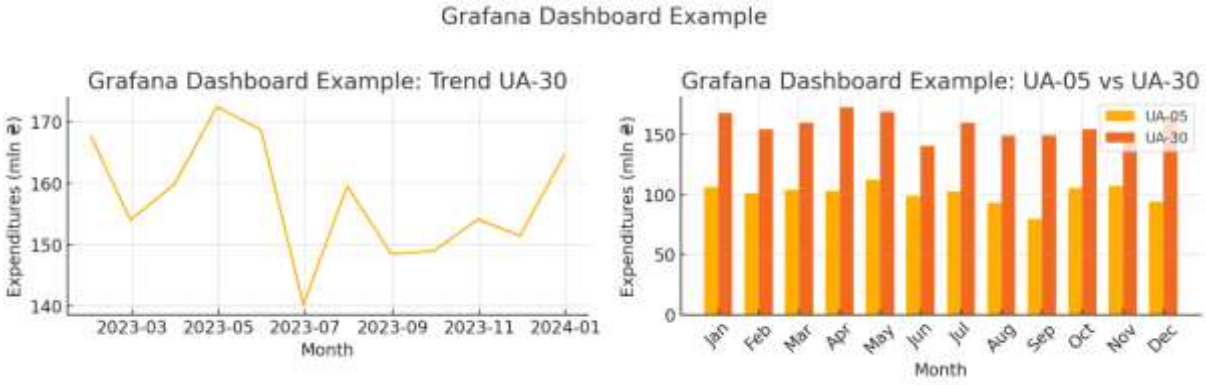

Рисунок 4.3 – Приклад інтерфейсу з використанням Grafana Dashboard Example

```
{
  "dashboard": {
    "title": "Соціальні видатки 2023",
    "panels": [
      {
        "type": "graph",
        "title": "Динаміка UA-30",
        "targets": [
          { "expr": "social_expenditures{region=\"UA-30\"}", "legendFormat":
"{{region}}"}
        ],
        "aliasColors": {},
        "xaxis": { "show": true },
        "yaxes": [{}, {}]
      },
      {
        "type": "graph",
        "title": "Порівняння UA-05 vs UA-30",
        "targets": [
          { "expr": "social_expenditures{region=\"UA-05\"}", "legendFormat": "UA-05" },
          { "expr": "social_expenditures{region=\"UA-30\"}", "legendFormat": "UA-30" }
        ],
        "barmode": "group"
      },
      {
        "type": "table",
        "title": "Фільтрація: пенсії",
        "targets": [
          { "expr": "social_expenditures{expense_type=\"пенсії\"}", "legendFormat":
"{{region}} – {{month}}"}
        ]
      }
    ],
```

```
  "templating": {
    "list": [
      {
        "name": "region",
        "type": "query",
        "query": "label_values(social_expenditures, region)",
        "current": { "text": "UA-30", "value": "UA-30" }
      }
    ]
  }
 }
}
```

У прикладі реалізації Tableau Dashboard Example (рис. 4.4) представлено лінійний і стовпчиковий графіки, інтерактивні фільтри й дії для оновлення взаємодії.

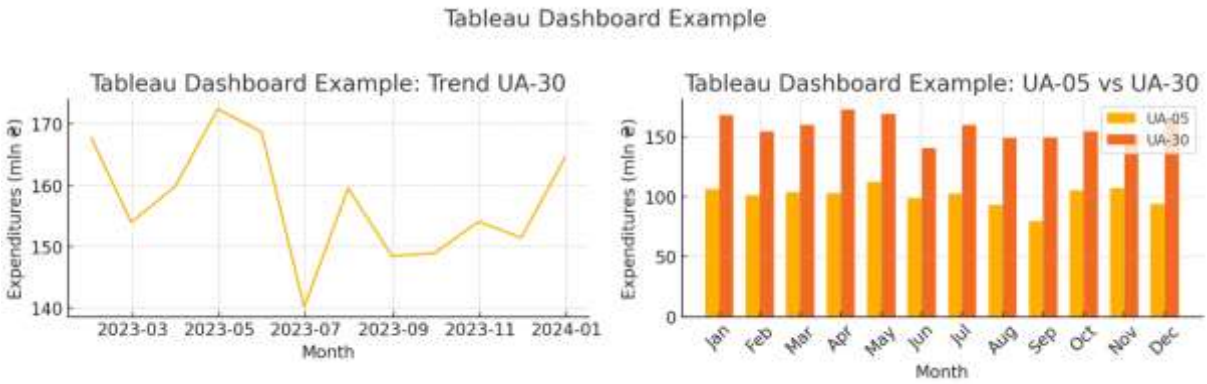

Рисунок 4.4 - *П*рикладі реалізації інтерфейсу за допомогою Tableau Dashboard Example

Сторінка “Dashboard”:

1. Ліворуч угорі — “Trend” (Worksheet1): лінійний графік місячної динаміки UA-30.
2. Праворуч угорі — “Comparison” (Worksheet2): стовпчаста діаграма UA-05 vs UA-30.
3. Внизу — “Filters” (Worksheet3): панель фільтрації за expense_type із мобільним перемикачем.

Налаштування:

- Кожен Worksheet підключається до одного джерела даних (Excel/CSV або SQL).

Dashboard містить дві дії (Actions):

- Filter Action: вибір регіону на Worksheet1 оновлює Worksheet2.
- Highlight Action: навколо вибраного рядка в таблиці показує відповідні бари в графіках.

Dashboard збережено на Tableau Server, доступ відкривається через URL із інтеграцією авторизації SSO.

Результати вимірювань - показники — середнє значення часу виконання завдання, с та SUS (System Usability Scale), вимірювалися за участи групі з 15 тестувальників.

Таблиця 4.1 – Порівняння результатів вимірювання часу викоанная завдань візуалізації за допомою різних Dashboard, с

Завдання	Dash,	Grafana	Tableau
Графік тренду	45.2	30.5	28.7
Порівняння регіонів	38.8	25.1	22.4
Застосування фільтра	50.4	20.2	24.3

У таблиці 4.2 наведено результати оцінки UX кожного інструмента за шкалою System Usability Scale (0–100):

Платформа	SUS (0–100)
Dash	71
Grafana	84
Tableau	88

Отже, Grafana і Tableau істотно швидші за Dash у виконанні стандартних аналітичних завдань, зокрема через готові інструменти побудови графіків і фільтрів. За UX-метрикою Tableau отримує найвищі оцінки (88), потім Grafana

(84), у той час як Dash — 71, що вказує на потребу в покращенні взаємодії (більш інтуїтивний інтерфейс, підказки, документація).

Для аналізу поведінки під реальним навантаженням, було досліджено, наскільки дашборди залишаються інформативними при одночасній роботі 20 користувачів та як швидко оновлюються дані в режимі реального часу.

Дослідження виконувалось за налаштувань тесту: симуляція 20 користувачів за допомогою JMeter / Locust, кожен з яких підвантажує сторінку дашборда й тримає його відкритим. Оновлення даних відбувається таким чином, що сервер надсилає нові пакети через SSE (Dash), WebSocket (Tableau) або періодичний HTTP-polling (Grafana) з інтервалом 5 с.

Таблиця 4.3 - Вимірювання лагу оновленняданих, мс

Платформа	Механізм оновлення	Середній лаг	P95 лаг
Dash	SSE	320 ms	480 ms
Grafana	Polling (5 s)	2 500 ms	2 520 ms
Tableau	WebSocket	210 ms	330 ms

У Dash оновлення даних здійснюється через SSE, тому події доходять до браузера миттєво, але клієнтський JavaScript додає близько 300 мс затримки. У Grafana застосовується HTTP-polling з інтервалом 5 с, тож оновлення відбувається лише раз за цикл опитування, відтерміновуючись майже на весь інтервал. У Tableau використовується WebSocket — двостороннє неблокуюче з’єднання, яке забезпечує найнижчий лаг і миттєву доставку змін.

Обмін повідомленнями та відгук інтерфейсу

- Dash/Grafana: при затримці оновлення > 500 ms деякі елементи UI (плашки «Loading…») зависають більше 1 s, що помітили 4 з 15 тестувальників.
- Tableau: непомітні плашки «refreshing», оновлення плавне.

Отже, для інтерфейсу, де ключовим питанням є отримання оперативна аналітика та швидке порівняння показників, Tableau і Grafana випереджають Dash як за часом виконання завдань, так і за UX-оцінками. Dash залишається гнучким фреймворком для індивідуалізованих рішень, але потребує оптимізації UI/UX для досягнення конкурентних показників. Під реальним навантаженням найменший лаг оновлення демонструє WebSocket-підхід (Tableau), тоді як SSE (Dash) — сусідній результат, і лише polling (Grafana) вимагає перегляду інтервалів оновлення для поліпшення чутливості.

В ході експерементів A/B-тестування показало, що Grafana та Tableau значно швидші за Dash у виконанні типових завдань (побудова тренду, порівняння регіонів, фільтрація), зокрема через наявність готових візуальних компонентів. За UX-метрикою SUS Tableau та Grafana отримали високі оцінки (88 і 84 відповідно), натомість Dash — лише 71, що вказує на потребу вдосконалити інтерфейс і навігацію в Dash-додатках.

Реальне навантаження продемонструвало, що технології з неблокуючим оновленням даних (WebSocket у Tableau, SSE у Dash) забезпечують низький середній лаг (< 350 ms), тоді як polling (Grafana з інтервалом 5 s) призводить до значних затримок (~2,5 s). При цьому непомітність оновлень і плавність інтерфейсу в Tableau роблять його найкращим вибором для дашбордів з потребою в реального часу аналітики.

Отже, для оперативної візуалізації та порівняння показників найбільш придатними виявилися Tableau та Grafana, а Dash варто використовувати там, де потрібна максимальна кастомізація, за умови додаткової оптимізації UX.

4.3 Використання розробленої інформаційної технології для прогнозування витрат на соціальних захист та соціальне забезпечення

Розроблена інформаційна технологія призначена для розв’язання широкого кола задач збору, оброблення та аналізу та прогнозування динаміки показників Дослідження санітарних втрат – питання, важливе не лише для

забезпечення поточної боєздатності військових формувань, а є складною проблемою, що охоплює національну економіку, оборонну та безпекову сфери, фінанси, соціальний захист та соціальне забезпечення. Найбільш розповсюдженими підходами щодо визначення санітарних втрат залишаються методи, основані на використанні нормативних значень, коефіцієнтів, індикативних показників, експертних оцінок. Однак, такі методики не прийнятні для випадків, коли, наприклад застосовується нове озброєння або методи ведення бою. Крім того, важливим є питання соціального захисту та соціального забезпечення всіх постраждалих відповідно до їх потреб та протягом всього періоду, поки не зникне потреба, що значно збільшує видатки державного бюджету та Пенсійного фонду. Тому задача прогнозування кількості санітарних втрат та кількості осіб, що стануть інвалідами на серелньо- та довгострокову перспективу є актуальною, потребує дослідження та опрацювання.

Прогнозування санітарних втрат передбачає використання різних підходів, урахуванням закономірностей та особливостей ведення бойових дій, характеру уражень, потенційний контингент постраждалих, тощо..

Для прогнозування втрат застосовують емпіричні моделі з метою побудови яких використовують статистичні дані про кількість постраждалих у попередніх конфліктах. В таких моделях використовують коефіцієнти, наприклад коефіцієнт втрат на 1000 осіб для наступу чи оборони. Серед математичних моделей слід відзначити модель Ланчестера та стохастичні моделі [134-144]. Часто застосовують також імітаційні (симуляційні) моделі, які дозволяють наочно відобразити ситуацію певного бойового сценарію. Найпоширенішим способом оцінювання можливих втрат залишається метод експертних оцінок, хоча якість прогнозування, в даному випадку, повністю залежить від людського фактору – професійних знань та навичок експертів, їх об’єктивності та неупередженості.

Дослідивши вітчизняні та іноземні спеціалізовані джерела, інформацію з мережі Інтернет [134-144] щодо кількості та структури загальних втрат у

різних війнах серед військовослужбовців, слід зазначити, що прогнозувати втрати лише на основі ретроспективних даних не доцільно, оскільки, технологічний розвиток, наявність сучасних озброєнь, які мають значні вражаючі можливості, активне використання інформаційних технологій та штучного інтелекту, зумовлюють складний, винищуваний характер бойових дій, який навіть у попередніх 10-15 років не має аналогів [137-138]. Сучасна зброя стає також дедалі досконалішою Навіть звичайні види озброєння стають більш потужними, а ураження ними більш масовими, крім того, з'являються нові види зброї, такі як лазерна, радіологічна, геофізична, галюциногенна, тощо [139- 143]. Все це зумовлює збільшення безповоротних та санітарних втрат, зростання тяжкості поранень, збільшення частки поранених з множинними пораненнями різних частин тіла, які згодом отримають інвалідність. Тобто, міняється не лише кількість, а й структура санітарних втрат. Тому важливо дослідити у короткостроковій перспективі потенційну кількість санітарних втрат, їх тяжкість, а у довгостроковій перспективі – ймовірну кількість інвалідів з числа санітарних втрат. Перспективним напрямом досліджень військових конфліктів, як зазначають фахівці, роботи яких представлені у Military Operations Research, [145] є використання ймовірнісно-статистичних моделей у формі мереж Байєса. Мережа Байєса надає [106] можливість встановити причинно-наслідкові зв'язки між подіями та оцінити ймовірності настання тієї чи іншої ситуації при отриманні нової інформації щодо зміни стану будь-якого вузла (змінної) мережі. Ступінь успішності застосування даного методу моделювання та формування статистичного висновку залежить від вміння коректно сформулювати постановку задачі, вибрати змінні процесу, які в достатній мірі характеризують його динаміку або статику, зібрати статистичні дані та використати їх для навчання мережі, а також коректно сформувати результат – висновок за допомогою побудованої мережі. Основною перевагою мереж Байєса є можливість одночасного врахування кількісних та якісних показників, динамічне надходження нової інформації, а також використання

явної залежності між існуючими факторами, а також наочність моделювання. Ще однією перевагою застосування мереж Байєса є можливість врахування у моделі дискретних і неперервних змінних, врахування невизначеностей та практично відсутнє обмеження на кількість змінних. Мережі Байєса легко піддаються опису, мають гарні характеристики як інструмент для класифікації, не мають обмежень на закони розподілу змінних, на відміну від інших моделей на основі регресійних рівнянь, та не вимагають повноти інформації.

Тому у даному дослідженні використані саме мережі Байєса, оскільки вони надають можливість врахувати невизначеності статистичного, структурного і параметричного характеру, побудувати моделі за наявності прихованих вершин і при неповних спостереженнях, а також реалізувати формування ймовірнісного висновку за допомогою різних методів – наближених і точних.

Зростання кількості постраждалих, зокрема, з числа військовослужбовців, внаслідок активної фази військового конфлікту потребує значних матеріальних та грошових витрат. Тому актуальною проблемою є впровадження сучасних методів аналізу даних та ситуацій, прогнозування та підтримки прийняття рішень як у сфері національної безпеки та оборони, так і пов’язаних з нею сферах національного господарства, які б покращили планування витрат, запобігали б неоптимальному витрачанню фінансових ресурсів держави.

Впровадження цифровізації, використання штучного інтелекту, технологій обробки великих масивів даних (Big Data), тощо, дає нові перспективи для розроблення відповідних систем підтримки прийняття рішень, основу яких складають сучасні інформаційні технології, математичні моделі, методи інтелектуального аналізу даних.

Пропонована інтелектуальна система підтримки прийняття рішень [3] являє собою модульний програмний комплекс, призначений для опрацювання даних про санітарні втрати та враховує специфіку даної предметної області.

Структура системи представлена на рис. 4.5.

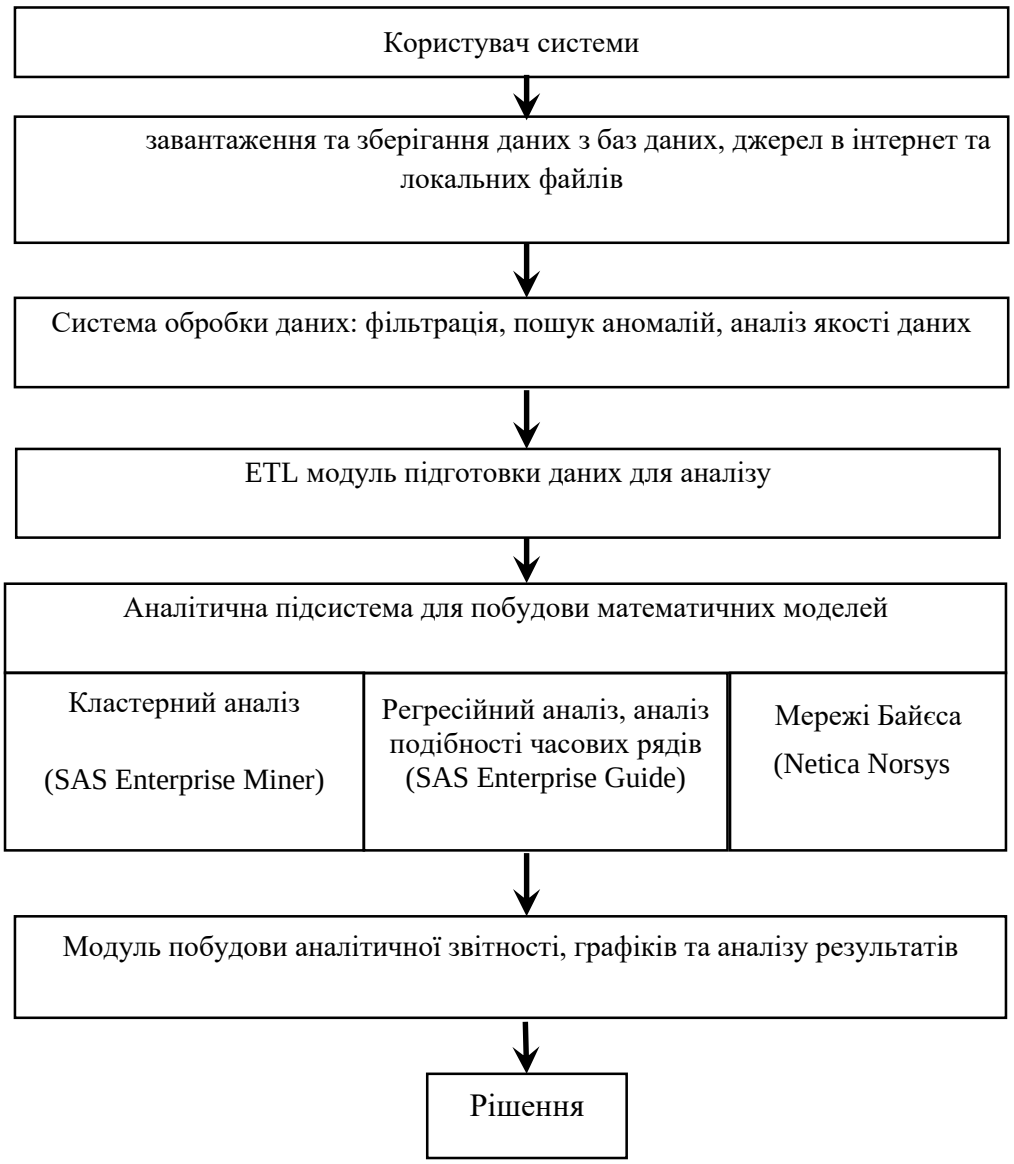

Рисунок 4.5 – Схема системи підтримки прийняття рішень

Як видно з рис. 4.5, система підтримки прийняття рішень являє собою модульний програмний комплекс, призначений для опрацювання даних про санітарні втрати та враховує специфіку предметної області [3, 139-151].

Особливостями пропонованої системи є наявність серії підсистем для попередньої обробки та збору вхідних даних, оскільки дані про санітарні та безповоротні втрати можуть накопичуватися для обробки з різних систем, бути представленими в різних форматах, містити пропуски даних, аномальні

значення. Основу системи становить аналітична підсистема, яка має модульну структуру та передбачає можливості її нарощування.

В даній роботі представлені модулі, які реалізовують технологічний ланцюг, призначений для вирішення задачі прогнозування бойових санітарних та безповоротних втрат. Дана система може бути використана як доповнення до існуючих систем. У розробці використано програмне забезпечення SAS Enterprise Miner 14.1 та SAS Enterprise Guide [152], що значно розширює набір інструментів аналітика за рахунок доповнення наявного програмного забезпечення новими моделями. Крім того, використання засобів компанії SAS дозволяє покращити візуалізацію результатів, представляти результати розрахунків у зручному для аналітики графічному чи табличному вигляді, за потреби, необхідні таблиці можуть бути вивантажені у зовнішні додатки, зокрема MS Excel.

Виходячи з наявної інформації щодо кількості втрат у період 2014-2019 рр. [139] у зоні АТО (рис. 4.6), було зроблено припущення стосовно співвідношення різних видів бойових втрат [3].

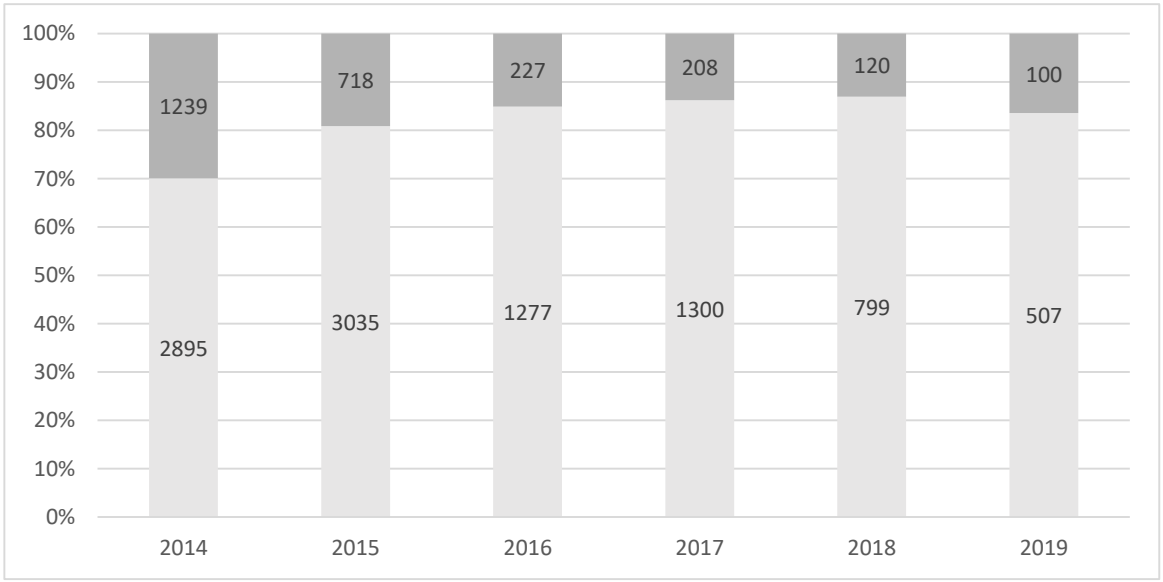

Рисунок 4.6 - Співвідношення бойових санітарних втрат (на діаграмі - світлий колір) і бойових безповоротних втрат (на діаграмі темний колір).

Як видно з рис. 4.6, санітарні втрати становлять переважну більшість втрат, також слід зазначити, що їх кількість коливається залежно від інтенсивності та характеру бойових дій, захисту особового складу – найбільші безповоротні втрати були саме на початку АТО, коли ще не було напрацьовано бойового досвіду.

Фактори, вплив яких на кількість санітарних втрат, за даними експертів є найсуттєвішим, представлені в таблицях 4.4-8. Вказані фактори й були розглянуті під час моделювання санітарних втрат.

Таблиця 4.4 -- Розподіл значень “Інтенсивність бою” (Battle intensity), для вершини мережі Байєса

Назва стану	Позначення стану в моделі	Значення ймовірності, %
Мала	Low	50
Середня	Medium	30
Висока	High	20

Таблиця 4.5 – Розподіл значень “Тип зброї” (Weapon), для вершини мережі Байєса

Тип зброї	Позначення типу зброї в моделі	Значення ймовірності, %
Вогнепальна	Firearms	25
Боєприпаси об'ємного вибуху	Volatile Explosive Ammunition	25
Високоточна зброя	Precision Weapons	25
Запальні суміші	Incendiary Mixtures	25

Таблиця 4.6 - Розподіл значень “Ступінь тяжкості поранення” (Injury), для вершини МБ

Ступінь тяжкості поранення	Позначення ступеня тяжкості поранення в моделі	Значення ймовірності, %
Легка	Easy	28,75
Середня	Medium	33,75
Важка	Heavy	18,75
Вкрай важка	Extremely difficult	18,75

Таблиця 4.7 – Розподіл значень “Захист особового складу” (Personnel protection), для вершини мережі Байєса

Захист особового складу	Позначення захисту особового складу в моделі	Значення ймовірності, %
Повний (бронежилет, каска, захист кінцівок)	Full	40
Частковий (тільки бронежилет чи каска)	Partial	45
Відсутній (немає засобів індивідуального захисту)	None	15

Таблиця 4.8. Розподіл значень “Час евакуації” (Evacuation time), для вершини мережі Байєса

Час евакуації	Позначення часу евакуації в моделі	Значення ймовірності, %
Своєчасна (<=1 година)	On time	30
Помірна (1-3 години)	Moderate	45
Запізно (> 3 годин)	Late	25

Вершина «Час евакуації» відображає, як швидко поранений військовослужбовець доставляється до місця кваліфікованої медичної допомоги. Вона впливає на результати (включаючи тяжкість санітарних втрат) і може приймати ймовірності, що ґрунтуються на реальних або передбачуваних логістичних умовах.

Таблиця 4.9 – Розподіл значень “ Медзабезпечення та логістика ” (Medicine and logistic), для вершини мережі Байєса

Медзабезпечення та логістика	Позначення медзабезпечення та логістики в моделі	Значення ймовірності, %
Якісне	Good	35
Середнє	Average	45
Погане	Poor	20

«Медзабезпечення та логістика» представляє рівень доступності та якості медичної допомоги, включаючи наявність польових шпиталів, медикаментів, кваліфікованого персоналу, санітарної евакуації та постачання.

Таблиця 4.10 – Розподіл значень “Характер ушкоджень” (Type Damage), для вершини мережі Байєса

Характер ушкоджень	Позначення характеру ушкоджень в моделі	Значення ймовірності, %
Голова	Head	32
Шия	Neck	1,24
Груди	Chest	8,94
Живіт	Abdomen	5,27
Таз	Pelvis	4,14
Хребет	Spine	1,14
Верхні кінцівки	Upper limbs	18,97
Нижні кінцівки	Lower limbs	28,3

Таблиця 4.11 – Розподіл значень “Локалізація ушкоджень” (Localization), для вершини мережі Байєса

Локалізація ушкоджень	Позначення локалізації ушкоджень в моделі	Значення ймовірності, %
Ізольовані	Isolated	28,76
Множинні	Multiple	24,88
Поєднанні	Combined	46,36

Таблиця 4.12 – Розподіл значень “Санітарні втрати” (Sanitary losses), для вершини мережі Байєса

Санітарні втрати	Позначення санітарних ушкоджень в моделі	Значення ймовірності, %
Низькі (< 1,5 %)	Low	31,7
Середні (1,5 – 3 %)	Medium	32,5
Високі (3 – 5 %)	High	23,7
Критичні (> 5%)	Critical	12,1

На рис. 4.7 наведена топологія побудованої мережі Байєса.

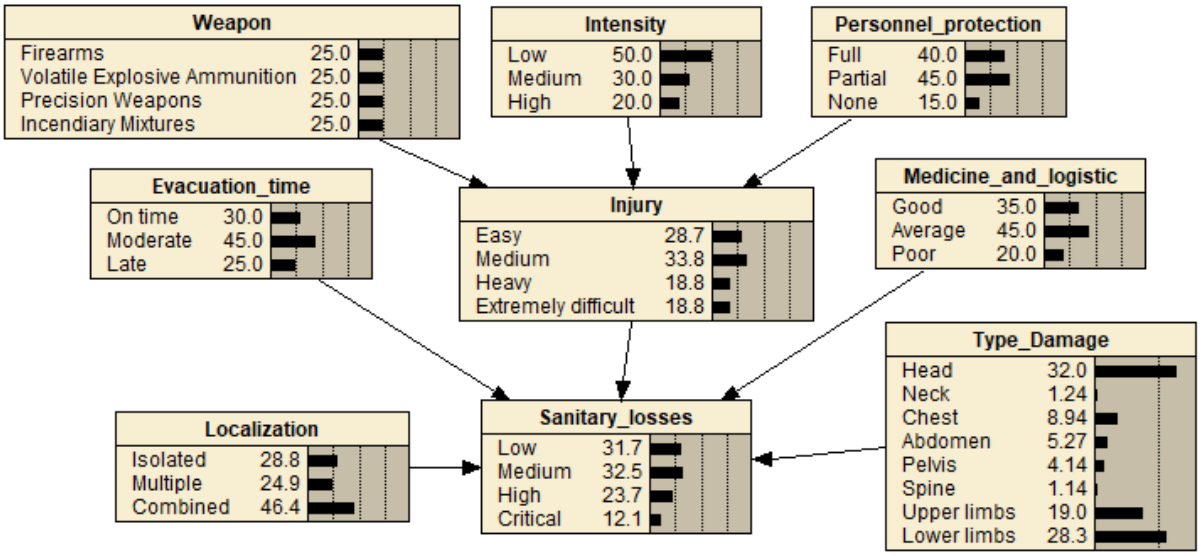

Рисунок 4.7 – Графічне представлення мережі Байєса для прогнозування ймовірності санітарних втрат

Використовуючи ймовірнісно-статистичні моделі, такі як представлено на рис. 4.7, можна розглянути декілька сценаріїв зміни кількості санітарних втрат (табл. 4.13).

Таблиця 4.13 – Приклад сценаріїв моделювання із використанням наведеної мережі Байєса

Інтенсивність бою	Тривалість евакуації	Якість медичного забезпечення та логістики	Санітарні втрати, %			
			Низькі (< 1,5)	Середні (1,5 – 3)	Високі (3 – 5)	Критичні (> 5%)
Мала	Своєчасна	Якісне	75	20	5	0
Середня	Помірна	Середнє	30	45	20	5
Висока	Запізно	Погане	5	20	40	35
Середня	Запізно	Погане	10	30	40	20
Висока	Своєчасна	Якісне	20	45	25	10
Мала	Помірна	Середнє	50	35	12	3

Використовуючи результати сценарного аналізу, наведені у табл. 4.13, можна побачити, що за позитивного сценарію, коли мала інтенсивність бою, евакуація здійснюється вчасно, хороша якість медичного забезпечення та логістика, санітарні втрати будуть найменшими. За найнегативнішого з усіх представлених сценаріїв, найвища ймовірність зростання кількості санітарних втрат до високих і, навіть, критичних значень (рис. 4.8).

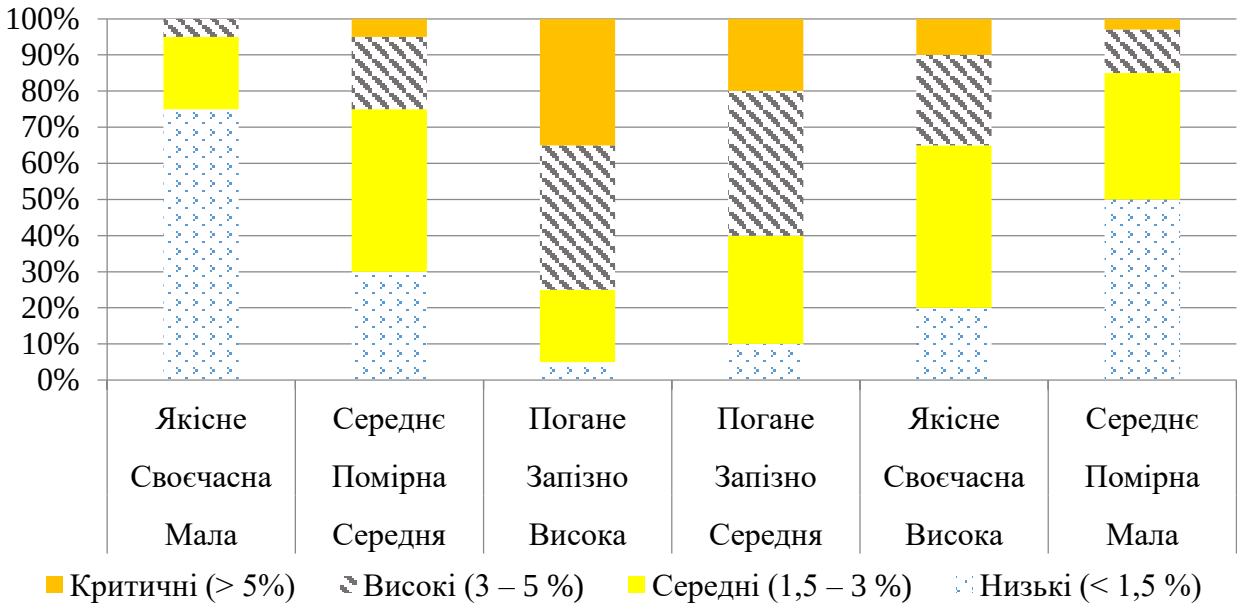

Рисунок 4.8 – Розподіл умовних ймовірностей санітарних втрат за різних умов конфлікту

Змоделювати ситуацію можна, додавши вершину, яка відображає ймовірність настання інвалідності. На рис. 4.9 наведена топологія побудованої мережі Байєса, яка моделює ситуацію, за якої можлива інвалідність військового.

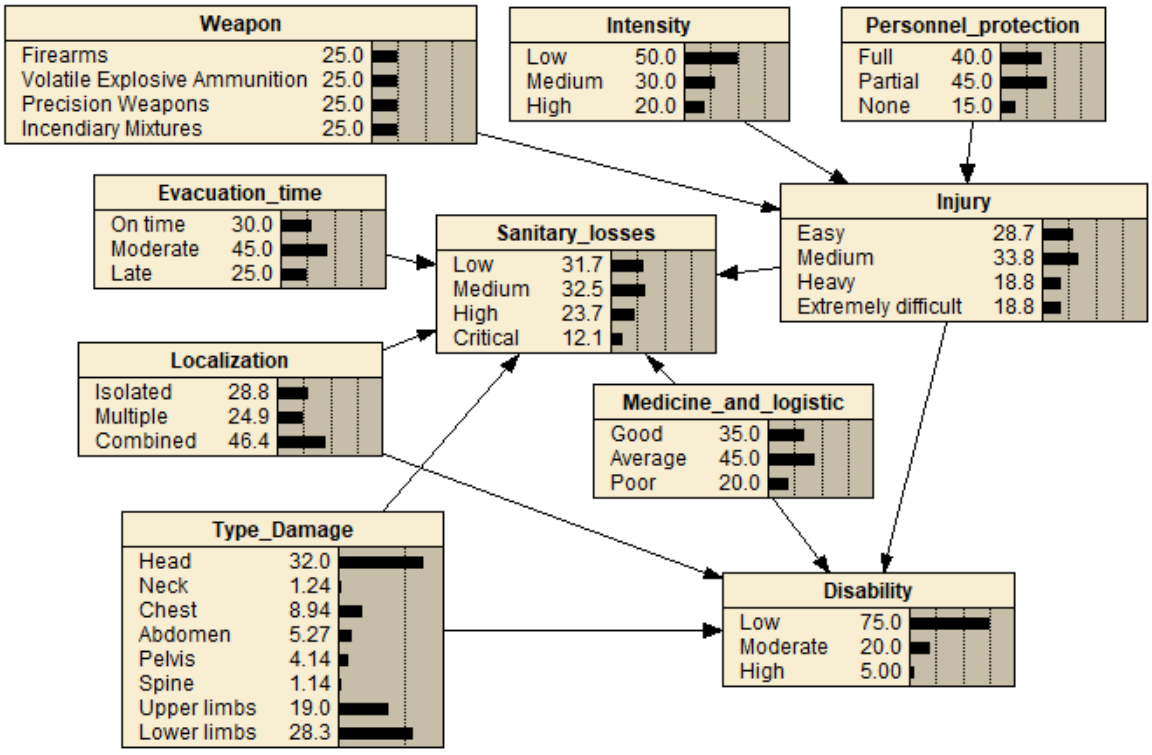

Рисунок 4.9 Мережа Байєса для прогнозування ймовірності настання інвалідності

Таблиця 4.14. – Розподіл значень “Ймовірність отримати інвалідність” (Disability), для вершини мережі Байєса

Ймовірність отримати інвалідність	Позначення ймовірності отримання інвалідності в моделі	Значення ймовірності, %
Низька ймовірність	Low	75
Помірна ймовірність	Moderate	20
Висока ймовірність	High	5

Опис станів вершини “Ймовірність отримати інвалідність”

1) Низька ймовірність – як правило це випадки коли поранення не є критичними або особа має достатньо високий рівень медичного забезпечення, ефективний захист та швидку евакуацію.

2) Помірна ймовірність – зазвичай це серйозні поранення, але не критичні. Може бути затримка в евакуації або недостатня медична допомога, але шанс на повне відновлення залишається.

3) Висока ймовірність – коли поранення надзвичайно тяжкі, і ймовірність інвалідності дуже висока через серйозні ушкодження внутрішніх органів, ампутацію, тяжкі черепно-мозкові травми тощо. Окрім цього дуже суттєво впливають проблеми з евакуацією або з медичним забезпеченням.

Таблиця 4.15. Результати моделювання із використанням пропонованої ймовірнісно-статистичної моделі - мережі Байєса

Локалізація ушкоджень	Медичне забезпечення та логістика	Тип ушкодження	Характер ушкоджень	Ймовірність отримати інвалідність
Ізольовані	Якісне	Легка	Верхні кінцівки	Низька
Множинні	Середнє	Середня	Живіт	Помірна
Поєднанні	Погане	Вкрай важка	Голова	Висока
Ізольовані	Якісне	Легка	Верхні кінцівки	Низька
Множинні	Середнє	Середня	Живіт	Помірна
Поєднанні	Погане	Важка	Голова	Висока
Множинні	Середнє	Важка	Нижні кінцівки	Помірна
Поєднанні	Середнє	Вкрай важка	Хребет	Висока

В результаті моделювання отримано такі сценарії: Перший сценарій – легкі ушкодження (наприклад, пошкоджена верхніх або ніжних кінцівок), швидке медичне забезпечення, ізольований і легкий тип ушкодження. У такому випадку ймовірність інвалідності буде низькою. Другий сценарій передбачає, що особа може отримати середні ушкодження, ізольовані або множинні і висока ймовірність затримки медичної допомоги,. За таких умов, ймовірність інвалідності зростає до помірного рівня. За третього сценарію, особа отримує вкрай тяжкі ушкодження (пошкодження голови або хребта), ситуація ускладнюється відсутністю швидкого надання медичної допомоги. В такому випадку ризик інвалідності дуже високий. Отже, як показали проведені чисельні експерименти, запропонований підхід може бути використаний для прогнозування санітарних втрат, і на основі отриманого прогнозу – кількості поранених, які з високою ймовірністю будуть інвалідами.

Отже, запропонована методика передбачає використання математичного моделювання та інтелектуального аналізу даних на основі байєсівського підходу для формування сценаріїв ситуацій, які потенційно можуть вимагати заходів додаткового економічного та фінансового впливу, розроблення імовірних сценаріїв розвитку ситуації на середньо- та довгострокову перспективу.

Як показує зарубіжний досвід, формування високого рівня національної безпеки і оборони України неможливе без функціонуючої надійної системи підтримки прийняття управлінських рішень, яка має бути багаторівневою з відкритою архітектурою, гнучкою та масштабованою, що забезпечить її здатність вчасно ідентифікувати нові загрози і виклики, оперативно опрацьовувати інформацію та надавати її зацікавленим споживачам для прийняття обґрунтованих та виважених рішень.

Висновки до розділу 4

Розділ присвячений апробації результатів дослідження, представлені результати використання запропонованих технологій збору та оброблення

даних та розробленого програмного забезпечення, які реалізовують технологічний ланцюг, призначений для вирішення задачі прогнозування витрат на виплату пенсій та соціальних допомог.

Ушкодження (наприклад, пошкодження верхніх або нижніх кінцівок), швидке медичне забезпечення, ізольований і легкий тип ушкодження роблять ймовірність інвалідності буде низькою. Другий сценарій передбачає, що особа може отримати середні ушкодження, ізольовані або множинні і висока ймовірність затримки медичної допомоги. За таких умов, ймовірність інвалідності зростає до помірного рівня. За третього сценарію особа отримує вкрай тяжкі ушкодження (пошкодження голови або хребта), ситуація ускладнюється відсутністю швидкого надання медичної допомоги. В такому випадку ризик інвалідності дуже високий. Отже, як показали проведені чисельні експерименти, запропонований підхід може бути використаний для прогнозування санітарних втрат, і на основі отриманого прогнозу – кількості поранених, які з високою ймовірністю будуть інвалідами.

Застосування інтелектуального аналізу текстової інформації у задачі виявлення потреб населення у соціальній допомозі та соціальних послугах. Вхідною інформацією для дослідження є електронні звернення громадян, які надійшли до вебпорталу електронних послуг Пенсійного фонду України та державної установи «Урядовий контактний центр»

Отже, запропонована методика дозволяє виявляти ситуації, які потенційно можуть вимагати заходів соціальної підтримки, що забезпечить адресність допомоги.

ВИСНОВКИ

В результаті виконання дисертаційної роботи отримано такі наукові результати.

1. Розроблено нові інформаційні технології збору та оброблення даних для прогнозування витрат на виплату пенсій та соціальних допомог. Їх основу становлять математичні моделі та їх ансамблі, методи інтелектуального аналізу даних, штучний інтелект, сучасні технології збору та зберігання даних, поєднання яких дозволяє накопичувати та обробляти різнорідну інформацію, отриману з різних джерел та використовувати її для прогнозуванні витрат на пенсії та соціальні допомоги. Розроблені інформаційні технології є складовою інформаційно-аналітичної платформи, яка вирізняється застосуванням конвергентного підходу, інтеграцією аналітичних платформ SAS та Oracle BI на бази апаратно-програмного забезпечення компанії Oracle, гнучкістю та адаптивністю, можуть бути використані окремо або в складі інформаційно-аналітичних систем органів державного управління та публічного управління.

2. Розроблено методологію збору та оброблення даних, що описують процеси у соціальній сфері, яка відрізняється робастністю результатів, використанням нового циклу обробки даних - від асинхронного ETL-збирання та LOESS-імпутації до автоматичного підбору моделей та надання візуальної інформації. Це забезпечує безперервну підготовку, аналіз і прогноз соціальних видатків навіть за неповних або спотворених даних, підвищуючи якість прогнозів, які створюють основу для прийняття рішень при плануванні витрат бюджетів різних рівнів на соціальний захист та соціальне забезпечення.

3. Удосконалено підхід до збору даних з державних реєстрів, відкриті API, дані НУО, стрім-потоки та веб-скрапінг, який забезпечує повне покриття різнопланової інформації (платежі, демографія, гуманітарні гранти) та підвищує репрезентативність вибірок для моделювання.

4. Розроблено новий метод побудови моделей та ансамблів моделей для визначення контингенту одержувачів пенсій та соціальних допомог, який

відрізняються модифікованою комплексною структурою, адаптивністю та високою адекватністю, за рахунок використання регресійних та сезонно-регресійних моделей (основний орієнтир для моделювання), деревні ансамблі, такі як градієнтний бустинг та випадковий ліс, що забезпечують високу точність короткострокових прогнозів, особливо із нелінійними збуреннями.

5. Удосконалено алгоритм динамічного вибору моделі залежно від вхідних даних. Розроблено метод моделювання соціально-економічних процесів, що мають місце у соціальній сфері в умовах невизначеності, який відрізняється від відомих опрацюванням різних типів невизначеностей, що підвищує адекватність моделей і якість оцінок прогнозів за лінійними та нелінійними моделями. Пропонований підхід дозволяє підвищити якість прогнозів видатків у соціальній сфері.

6. Розроблено інформаційну технологію прогнозування витрат на пенсії та соціальні допомоги, в основу якої покладено поєднання принципів системного аналізу, методів обробки структурованих та неструктурованих даних, оцінювання якості даних, яка забезпечує високу якість прогнозів. Для урахування структурних невизначеностей розроблено процедуру оновлення параметрів моделей у режимі «наближення до реального часу», що враховує зовнішні шоки (бойові дії, демографічні зрушення) використовуючи моделі ARIMAX/SARIMAX з екзогенними факторами, для візуалізації результатів інтегровано фронтенд-дашборди, що генерують регіональні та національні сценарії соціальних витрат, візуалізують довірчі інтервали й ранжують ризики перевитрат.

Використання розробленої інформаційно-аналітичної підсистеми забезпечить як скорочення часу на оброблення значних обсягів даних (структурованих та неструктурованих) про застрахованих осіб, одержувачів соціальних допомог, так і підвищення якості прогнозів потреби у фінансуванні пенсій та допомог, дозволить отримувати результати аналізу та прогнозування у режимі реального часу, що покращить якість управлінських рішень в соціальній сфері країни. В подальшому система може бути розширена за

рахунок додавання нових математичних моделей та їх ансамблів, засобів обробки геопросторової інформації та інших методів штучного інтелекту.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Зарудний О. Б. і Коваль Р. Г. Методика прогнозування часових рядів для визначення потреби у видатках на соціальний захист та соціальне забезпечення», *Міжнародний науково-технічний журнал «Проблеми керування та інформатики»*. 2024. Vol. 69, No. 5. С. 64–77. https://doi.org/10.34229/1028-0979-2024-5-5
2. Zarudnyi, O., Koval, R. Application of probabilistic and stochastic models and data mining for forecasting the contingent of old age pension recipients in the context of systemic uncertainty. *Technology Audit and Production Reserves*, 2024. Vol. 5. No. 2(79), P. 56–62. https://doi.org/10.15587/2706-5448.2024.313960
3. Трофимчук О., Коваль Р., Зарудний О. Методика прогнозування кількості інвалідів з числа санітарних втрат. *Екологічна безпека та природокористування*, 2025. Vol. 54. No. 2. С. 121–135. https://doi.org/10.32347/2411-4049.2025.2.121-135
4. Zarudnyi, O., Koval, R. Methodology for Constructing an Analytical Subsystem of the Unified Information System of the Social Sphere of Ukraine. *Cybernetics and Systems Analysis.* 2025. Vol. 61. No. 4. P. 487–494. https://doi.org/10.1007/s10559-025-00785-9
5. Korbicz J., Sholokhov O., Koval R., Zarudnyi O. Application of SAS Text Miner for the analysis of citizens' appeals in the system of social protection and social security. *8th International Scientific and Practical Conference Applied Information Systems and Technologies in the Digital Society*. (AISTDS 2024). (Kyiv , 1 Oct. 2024). 2024. Vol. 3942, P. 113 – 124. ISSN 1613-0073 URL: https://www.scopus.com/pages/publications/105001096518
6. Зарудний О.Б., Коваль Р. Г. Проблеми прогнозування факторів, що впливають на обсяг видатків Пенсійного фонду України Інформаційно-комунікаційні технології для перемоги та відновлення // Колективна монографія за матеріалами *XXII Міжнародної науково-практичної*

конференції «Інформаційно-комунікаційні технології та сталий розвиток» (Київ, 14-15 листопада 2023 р.) / За заг. ред. С.О. Довгого. – К.: ТОВ «Видавництво «Юстон», 2023. С. 76-79. URL: https://itgip.org/wp-content/uploads/2023/11/1_zbirka_08_11_23-1-1.pdf

7. Зарудний О.Б., Коваль Р. Г. Застосування методу сингулярного розкладання у задачі аналізу електронних звернень громадян до Пенсійного фонду України. Математичне моделювання та інформаційно-комунікаційні технології для зміцнення та відновлення // Колективна монографія за матеріалами XXIII Міжнародної науково-практичної конференції «*Математичне моделювання та інформаційно-комунікаційні технології для зміцнення та відновлення»* (Київ, 12-13 листопада 2024 р.) / За заг. ред. С.О. Довгого. К.: ТОВ «Видавництво «Юстон», 2024. С. 156-159. URL: https://itgip.org/wp-content/uploads/2024/11/2024-11-24_zbirka_all_07_11_2024_148x210.pdf
8. Зарудний О. Б., Коваль Р. Г. Шолохов О. В. Методика застосування кластеризації текстів на основі лінгвістичних правил для дослідження потреб населення у соціальному захисті та соціальному забезпеченні Прикладні інформаційні системи та технології в цифровому суспільстві: зб. Тез доповідей і наук. повідомлень учасників VIII *Міжнародної науково-практичної конференції «Прикладні інформаційні системи та технології в цифровому суспільстві»* (Київ, 01 жовтня 2024 р.) // за заг. ред. В. Плескач, О. Фендьо. К.: Київський нац. ун-т ім. Тараса Шевченка, 2024. С. 167-178. URL: https://aistis.knu.ua/wp-content/uploads/2024/10/Text_Book_confernce_01.10.2024.pdf
9. Зарудний О. Б., Колбун В. А., Малецький О. М., Машкін В. Г., Рябцева Т. Б., Матвєєв С. В., Хандрос Ю. О., Коваль Р. Г., Бондарчук О. М., Педан Ю. А., Павлюков І. О. Комп’ютерна програма «Інтегрована комплексна система Пенсійного фонду України «ІКІС ПФУ», версія 2» №11056; заявл.31.10.2011; опубл. 11.11.2011.
10. What is EUR-ACE. URL: https://www.enaee.eu/eur-ace-system/

11. FuturICT 2.0. URL: https://coss.ethz.ch/research/pastprojects/FuturICT20.html
12. Ведмідь П.В. Роль інформаційно-комунікаційних технологій У забезпеченні ефективності державного управління. *Публічне управління і адміністрування в Україні.* 2020. Вип. 18. С. 32-36. https://doi.org/10.32843/pma2663-5240-2020.18.6
13. Jain, R.R. Creasey, J. Himmelspach, K.P. White, and M. Fu, eds. Performance analysis of comercial simulation-based optimization packages: optquest and witness optimizer. Proceedings of the 2011 Winter Simulation Conference. DOI: 10.1109/WSC.2011.6147946
14. Eskandari H., E. Mahmoodi, H. Fallah and C. D. Geiger, "Performance analysis of comercial simulation-based optimization packages: OptQuest and Witness Optimizer," *Proceedings of the 2011 Winter Simulation Conference (WSC)*, Phoenix, AZ, USA, 2011, pp. 2358-2368, doi: 10.1109/WSC.2011.6147946.
15. Economic Forecasting and Analysis URL:https://inforumecon.com
16. Project LINK URL: https://www.un.org/development/desa/dpad/project-link.html
17. В Апараті РНБО України розроблено та введено в експлуатацію сучасну інформаційно-аналітичну систему «СОТА». URL: https://www.rnbo.gov.ua/ua/Diialnist/5011.html
18. Трембіта. Система електронної взаємодії державних електронних інформаційних ресурсів. URL:https://trembita.gov.ua/ua
19. Пенсійний фонд України. Офіційний сайт. URL: https://www.pfu.gov.ua/pro-pfu/istoriya-stanovlennya/
20. Державне підприємство «Інформаційно-обчислювальний центр Міністерства соціальної політики України». URL: https://www.ioc.gov.ua
21. ТОВ "НВП "Медирент". Офіційний сайт. URL: https://medirent.ua
22. Про внесення змін до Положення про Єдину інформаційну систему соціальної сфери. Постанова Кабінету Міністрів України від 27.10.2023 р. № 1130. https://zakon.rada.gov.ua/laws/show/1130-2023-п#Text

23. Про затвердження Порядку проведення експериментального проекту з реалізації функціоналів першої черги Єдиної інформаційної системи соціальної сфери. Постанова Правління Пенсійного фонду України від 03.03.2021 № 8-1. URL: https://zakon.rada.gov.ua/laws/show/za400-21#Text
24. Про затвердження регламенту роботи інформаційно-телекомунікаційної системи Пенсійного фонду України. Постанова Правління Пенсійного фонду України від 11.10.2021 28-1. URL: https://www.pfu.gov.ua/content/uploads/2021/10/Postanova-28-1-vid-11-zhovtnya-2021-.pdf
25. Oracle Business Intelligence. URL: https://www.oracle.com/business-analytics/business-intelligence/technologies/bi.html
26. Проект Закону про Єдину інформаційну систему соціальної сфери. Проект Закону від 24.06.2024 №11377 URL: https://itd.rada.gov.ua/billInfo/Bills/Card/44497
27. Бідюк П. І., Тимощук О. Л., Коваленко А. Є., Коршевнюк Л. О. Системи і методи підтримки прийняття рішень: підручник для здобувачів ступеня магістра за спеціальністю 124 Системний аналіз. Київ : КПІ ім. Ігоря Сікорського, 2022. 610 с.
28. Bazilevych, K. O.; Chumachenko, D. I.; Hulianytskyi, L. F.; Meniailov, I. S.; Yakovlev, S. V. Intelligent decision support system for epidemiological diagnostics. Information technology development. *Cybernetics and Systems Analysis*. 2022. Vol. 58. P. 499-509. Cybernetics and Systems Analysi
29. Довгий С. О., Копійка О. В., Козлов О. С. Передача інформації в автоматизованих системах спеціального призначення. *Екологічна безпека та природокористування*, 2023. .Вип. 1. № 45, С. 76-90/ https://doi.org/10.32347/2411-4049.2023.1.76-90
30. Запорожець Т.В. Інтелектуальне управління у діяльності органів публічної влади: монографія. Київ : НАДУ, 2020. 450 с. URL: https://feu.kneu.edu.ua/ua/depts4/mdu/publicats/Zaporozets_int_upr_OPV/

31. Trofymchuk O. M, Bidiuk P. I., Prosiankina-Zharova T. I., Terentiev O. M. Decision support systems for modelling, forecasting and risk estimation: monography. Riga : LAP LAMBERT Academic Publishing. 2019. 176 p.
32. Технічні вимоги на надання послуг зі створення Єдиної інформаційної система соціальної сфери. Перша черга. URL: https://tapas.org.ua/wp-content/uploads/2021/01/Terms-of-Reference-Tekhnichne-zavdannia.pdf
33. Згуровський М. З., Панкратова Н. Д. Основи системного аналізу. Київ : Видавнича група BHV, 2007. 544 с.
34. Формування інституційного середовища модернізації екоФ 79 номіки старопромислових регіонів України: монографія / В. І. Ляшенко, І. Ю. Підоричева, В. П. Антонюк та ін.; НАН України, Ін-т економіки пром-сті. Київ, 2022. 472 с URL: https://iie.org.ua/wp-content/uploads/application/pdf/mono_maket-2022_compressed.pdf
35. UNDP. URL: https://www.undp.org/uk/ukraine
36. Шаповалова Т. Поняття і зміст соціального захисту та соціального забезпечення населення в сучасній Україні. Економічний аналіз. 2022. Том 32. № 3. С. 123-130. https://doi.org/10.35774/econa2022.03.123
37. Грень Т. Я. Особливості реалізації політики соціального захисту територій в умовах війни. Вчені записки ТНУ імені В.І. Вернадського. Серія: Публічне управління та адміністрування. 2022. Том 33 (72) № 6. С. 81-84. https://doi.org/10.32782/TNU-2663-6468/2022.6/13
38. Видатки на соціальну допомогу. URL: https://mof.gov.ua/uk/expenditures_on_social_assistance
39. Смуш-Кулеша М. Федорова А., Мойса Б.Соціальні права в Україні під час війни.Звіт про оцінку потреб. Рада Європи. 2022, 64 с. URL: https://rm.coe.int/needs-assessment-ua-2/1680a9b408
40. State Statistics Service of Ukraine, https://www.ukrstat.gov.ua
41. Монографії Інституту економіки та прогнозування НАН України. URL: https://ief.org.ua/publication/monohrafii

42. Kuznietsova, N., Bidyuk P., “Intelligence Information Technologies for Financial Data Processing in Risk Management”. Data Stream Mining & Processing. Communications in Computer and Information Science, Cham: Springer, vol. 1158, 2020. https://doi.org/10.1007/978-3-030-61656-4_36
43. De Jong P., Heller G.Z., *Generalized Linear Models for Insurance Data.* New York: Cambridge University Press, 2008.
44. Mc Cullagh, P., Nelder J. A., *Generalized Linear Models*. New York: Chapman & Hall, 1989.
45. *Predictive Modeling Applications in Actuarial Science*./ Edited by Edward W. Frees, Cambridge: Cambridge University Press., 2014. https://doi.org/10.1017/CBO9781139342674
46. Hansen, B.E., *Econometrics.* Madison: University of Wisconsin, 2021.
47. Davidson, R., MacKinnon J. G., *Econometric Theory and Methods*, Oxford: Oxford University Press, 2004. https://doi.org/10.1017/S0266466605000356
48. Castle, J. L., Doornik J. A., Hendry D. F., “Modelling non-stationary ‘Big Data’”, *International Journal of Forecasting*, vol. 37, no. 4, pp. 1556-1575, 2021. ttps://doi.org/10.1016/j.ijforecast.2020.08.002
49. Draper N., Smith G. *Applied regression analysis*. New York: John Wiley & Sons, Inc., 1986. 366 p.
50. Bidyuk P.I., Polovcev O.V. *Analysis and modeling of the processes of economy in transition*. Kyiv: KPI, 1999. 230 p.
51. Trofymchuk O.M., Bidyuk Бідюк П. І., Тимощук О. Л., Коваленко А. Є., Коршевнюк Л. О. Системи і методи підтримки прийняття рішень: підручник для здобувачів ступеня магістра за спеціальністю 124 Системний аналіз. Київ : КПІ ім. Ігоря Сікорського, 2022. 610 с.
52. Bazilevych, K. O.; Chumachenko, D. I.; Hulianytskyi, L. F.; Meniailov, I. S.; Yakovlev, S. V. Intelligent decision support system for epidemiological diagnostics. Information technology development. *Cybernetics and Systems Analysis*. 2022. Vol. 58. P. 499-509. Cybernetics and Systems Analysi

53. Bidyuk P.I., Prosyankina-Zharova T.I., Terentiev O.M. Forecasting nonstationary processes in demography, ecology, economy and finances. Proc. 2018 Information technologies of environmental safety management, nature management, measures in emergency situations. (Kyiv: Yuston), P. 113 – 123.
54. Запорожець Т.В. Інтелектуальне управління у діяльності органів публічної влади: монографія. Київ : НАДУ, 2020. 450 с. URL: https://feu.kneu.edu.ua/ua/depts4/mdu/publicats/Zaporozets_int_upr_OPV/
55. Pepelyaev, V.A., Golodnikov, A.N., Golodnikova, N.A. A new method of reliability optimization in the classical problem statement. *Cybernetics and Systems Analysis.* 2022. Т. 58. № 6, С. 917–922. https://doi.org/10.1007/s10559-023-00525
56. Information technology assessment of the pension systems. URL: https://pdf.usaid.gov/pdf_docs/PA00WC4R.pdf
57. Zgurovsky M. Z., Zaychenko Y. P. The Fundamentals of Computational Intelligence: System Approach. Switzerland: Springer International Publishing, 2016. 375 p
58. Tsay, R. S., *Analysis of Financial Time Series*, Hoboken Wiley & Sons, Inc., 2010. https://doi.org/10.1002/9780470644560
59. Johnston J., DiNardo J. *Econometric methods*. New York: McGraw-Hill, Inc., 1997. 530 c.
60. Borisenko O.A. *System theory. The Information approach.* Sumy: Sumy State University, 2010, 210 p.
61. De Gooijer, J. G., *Elements of Nonlinear Time Series Analysis and Forecasting*. Cham: Springer, 2017.
62. Bidyuk, P. I, V. D. Romanenko, and O. L. Tymoshchuk, *Time Series Analysis*, Kyiv: Polytechnika Publisher at the National Technical University of Ukraine “Igor Sikorsky KPI”, 2012.
63. Sugumaran, V. *Intelligent Support Systems Technology: Knowledge Management*. London: IRM Press, 2002. https://doi.org/10.4018/978-1-931777-00-1

64. Tavana, M., Soltanifar, M. & Santos-Arteaga, F.J. Analytical hierarchy process: revolution and evolution. Ann Oper Res 326, 879–907 (2023). https://doi.org/10.1007/s10479-021-04432-2
65. World Bank. 2010. Modeling Pension Reform: The World Bank's Pension Reform Options Simulation Toolkit. World Bank Pension Reform Primer Series, World Bank PROST Model. URL: http://hdl.handle.net/10986/11074
66. Jain, R.R. Creasey, J. Himmelspach, K.P. White, and M. Fu, eds. Performance analysis of comercial simulation-based optimization packages: optquest and witness optimizer. Proceedings of the 2011 Winter Simulation Conference. DOI: 10.1109/WSC.2011.6147946
67. Eskandari H., E. Mahmoodi, H. Fallah and C. D. Geiger, "Performance analysis of comercial simulation-based optimization packages: OptQuest and Witness Optimizer," *Proceedings of the 2011 Winter Simulation Conference (WSC)*, Phoenix, AZ, USA, 2011, pp. 2358-2368, doi: 10.1109/WSC.2011.6147946.
68. Комар М. П. Інформаційна технологія інтелектуальної обробки та аналізу великих даних. *Вісник Хмельницького національного університету*, 2020. №5, (289). С. 120-125. https://doi.org/10.31891/2307-5732-2020-289-5-120-125
69. Портал відкритих даних. URL: https://data.gov.ua
70. Фісун М. Т., Дворецький М. Л., Юхатов А. В. Порівняльний аналіз методів побудови OLAP-систем із використанням засобів MS SQL Server та Oracle. *Наукові праці. Комп'ютерні технології.* 2016. Вип. 271. Т. 283. С. 36-42. URL: https://dspace.chmnu.edu.ua/jspui/bitstream/123456789/581/1/Наук.%20пр.%20Т.%20283.%20Вип.%20271.%20Комп.%20технол..pdf
71. SAS Institute. URL: https://www.sas.com
72. Abirami N. Kadry S., Gandomi A., Balusamy B. Big Data: Concepts, Technology and Architecture. John Wiley & Sons, 2021. 368 p.
73. Python, URL: https://www.python.org
74. Державний веб-портал бюджету для громадян. URL:

https://openbudget.gov.ua

75. Find the information that matters using natural language processing (NLP). URL: https://www.sas.com/ru_ua/software/visual-text-analytics.html
76. Survey of Text Mining I: Clustering, Classification, and Retrieval / Ed. by M. W. Berry. Springer, 2003. 261 p.
77. Aggarwal C. C., Zhai C. Mining Text Data. Springer, 2012. 527 p.
78. Text Cluster Node Results. URL: https://documentation.sas.com/?docsetId=tmref&docsetTarget=n1d7r58qug6sefn162cu6cqx0nq4.htm&docsetVersion=14.3&locale=en
79. Emerging Technologies of Text Mining: Techniques and Applications / Ed. by H. A. Do Prado, E. Ferneda. Idea Group Reference, 2007. 358 p.
80. Awad, M., Khanna, R. Hidden Markov Model. In: Efficient Learning Machines. 2015. Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4302-5990-9_5
81. Matignon R. Data Mining Using SAS Enterprise Miner. URL: https://www.amazon.com/Data-Mining-Using-Enterprise-Miner/dp/0470149019
82. Кожевников В. Л., Кожевников А. В. Теорія інформації та кодування : навч. Посібник.. Д.: Національний гірничий університет, 2013. – 144 с
83. Сухарський С. С. Агоритм сингулярного розкладу на графічному процесорі. *Проблеми програмування.* 2023. №1. С. 30-37. https://doi.org/10.15407/pp2023.01.030
84. ARIMAX/SARIMAX та XGBoost. URL: https://dsdaily.substack.com/p/ds-daily-arima-and-xgboost?utm_source=substack&utm_campaign=post_embed&utm_medium=web
85. Офіційний сайт Міністерства цифрової трансформації України. URL: https://thedigital.gov.ua
86. Про затвердження Положення про Єдину інформаційну систему соціальної сфери. Постанова Кабінету Міністрів України від 14 квітня 2021 р. № 404. URL: https://zakon.rada.gov.ua/laws/show/404-2021-п#Text

87. Звіт про звернення громадян за 9 місяців 2024 року. UR;: https://www.pfu.gov.ua/2167929-zvit-pro-zvernennya-gromadyan-za-9-misyatsiv-2024-roku/
88. Урядовий контакт-центр.URL: https://ukc.gov.ua
89. Bidyuk P., Prosyankina-Zharova T., Terentiev O. Modelling Nonlinear Nonstationary Processes in Macroeconomy and Finances. *Advances in Computer Science for Engineering and Education. ICCSEEA 2018. Advances in Intelligent Systems and Computing* / Ed. Hu Z., Petoukhov S., Dychka I., He M. Cham : Springer, 2019. Vol. 754. P. 735–745. URL: http://doi.org/10.1007/978-3-319-91008-6_72
90. Brocklebank J. C., Dickey D. A. SAS System for Forecasting Time Series / Second Edition Cary N C. SAS Institute Inc., 2003. 418 p.
91. Gembarski, P. C., Plappert, S.,· Lachmayer, R. Making design decisions under uncertainties: probabilistic reasoning and robust product design, *Journal of Intelligent Information Systems*, 2021. Vol. 57, pp. 563–581, https://doi.org/10.1007/s10844-021-00665-6
92. Andreica, M., Popescu, M.ME., Micu D., Albu E. Adaptive management procedural model for support of economic organizations. *Proc of 10th International Management Conference "Challenges of Modern Management*. 2016. P. 295–301
93. Dobrescu E. Modelling an Emergent Economy and Parameter Instability Problem. *Journal for Economic Forecasting of Institute for Economic Forecasting*. 2017. Vol. 0(2). P. 5–28. URL: https://ideas.repec.org/a/rjr/romjef/vy2017i2p5-28.html
94. Bidyuk, P., Tymoshchuk, O., Kovalenko, A., and Korshevnyuk L. Systems and Methods for Decision Support. Kyiv: Polytechnika Publisher at the National Technical University of Ukraine “Igor Sikorsky KPI”, 2022, 610 p.
95. Harvey A.C. *Forecasting, structural time series models and the Kalman filter*. University of Cambridge: Cambridge University Press, 1994. 554 p.

96. Bidyuk P.I., Kalinina I.O., Gozhyj O.P. “An approach to identifying and filling in data gaps in machine learning procedures”. Lecture Notes on Data Engineering and Communication Technologies, vol. 77, 2022, pp. 164–176.
97. Borisenko O.A. *System theory. The Information approach.* Sumy: Sumy State University, 2010, 210 p.
98. De Gooijer, J. G., *Elements of Nonlinear Time Series Analysis and Forecasting*. Cham: Springer, 2017.
99. Bidyuk, P. I, V. D. Romanenko, and O. L. Tymoshchuk, *Time Series Analysis*, Kyiv: Polytechnika Publisher at the National Technical University of Ukraine “Igor Sikorsky KPI”, 2012.
100. Гладун А. Я., Рогушина Ю. В. Data mining: пошук знань в даних: підручник. Київ: АДЕФ-Україна, 2016. 451 с.
101. Kaptein M., van den Heuvel E. Statistics for Data Scientists. Cham: Springer, 2022. 321 p. https://doi.org/10.1007/978-3-030-10531-0
102. Методи машинного навчання (МН). URL: https://blog.colobridge.net/uk/2024/05/artificial-intelligence-and-machine-learning-ua
103. Random Forest Algorithm Overview (H. A. Salman, A. Kalakech, & A. Steiti , Trans.). (2024). Babylonian Journal of Machine Learning, 2024, 69-79. https://doi.org/10.58496/BJML/2024/007
104. Gradient Boosting: теорія та пакети/ URL: https://www.kaggle.com/code/nadiia789/gradient-boosting
105. LightGBM. URL: https://itwiki.dev/data-science/ml-reference/ml-glossary/lightgbm
106. Rossi, P.E., G. M. Allenby, R. McCulloch, *Bayesian Statistics and Marketing*. Hoboken: John Wiley & Sons, Ltd, 2005. 368 p. http://doi.org/10.1002/0470863692
107. Bidyuk, P.I, O.M. Terentyev, and M. M. Konovalyuk, “Bayesian networks in technologies Sik-Yum, Lee. *Structural Equation Modeling: A Bayesian*

Approach. Chichester: John Wiley & Sons Ltd, 2007. https://doi.org/10.1002/9780470024737

108. Chen, Ming-Hui, Shao Qi-Man, Ibrahim J. G., *Monte Carlo Methods in Bayesian Computation*. New York: Springer-Verlag, 2000. https://doi.org/10.1007/978-1-4612-1276-8
109. Hautaniemi, S. K., Korpisaari P. T., Saarinen J. P. P., "Target identification with dynamic hybrid Bayesian networks. Target identification with dynamic hybrid Bayesian networks", *VI Image and Signal Processing for Remote Sensing*, vol. 4170, Barcelona, Spain, January 2001. https://doi.org/10.1117/12.413885
110. of intellectual data analysis", *Artificial intelligence, V*ol. 2, pp. 104 – 113, 2010.
111. Press, S.J., *Subjective and Objective Bayesian Statistics: Principles, Models, Applications*. Hoboken: John Wiley & Sons, Inc., 2003.
112. Zgurovsky, M. Z., P. I. Bidyuk, O. M.Terentiev and T. I. Prosyankina-Zharova, *Bayesian Networks in Decision Support Systems*. Kyiv: Edelweys, 2015.
113. Jensen, F.V. *Bayesian networks and Decision Graphs*. New York: Springer-Verlag, 2001. https://doi.org/10.1007/978-1-4757-3502-4
114. Бюджети територіальних громад у Київській області. URL: https://openbudget.gov.ua/local-budget/1050000000/info/indicators
115. Kubernetes з HPA (Horizontal Pod Autoscaler), застосовано підхід, за якого, вимірюючи загальний time-to-scale-up при різкому збільшенні трафіку
116. Oracle Database Enterprise Edition. URL: https://www.oracle.com/database/technologies/oracle-database-software-downloads.html
117. Oracle Business Intelligence (Oracle BI). URL: https://www.oracle.com/ua/business-analytics/business-intelligence/technologies/bi.html

118. SAS Institute Inc. SAS/OR 14.2 User's Guide: Mathematical Programming", http://support.sas.com/thirdpartylicenses.

119. Єдина інформаційна система соціальної сфери (ЄІССС). URL: https://www.msp.gov.ua/e-servisy/yeisss

120. Довгий С. О., Копійка О. В., Козлов О. С. Передача інформації в автоматизованих системах спеціального призначення. *Екологічна безпека та природокористування*, 2023. .Вип. 1. № 45, С. 76-90/ https://doi.org/10.32347/2411-4049.2023.1.76-90

121. Tomar D. J. Information technology assessment of the pension funds. Technical Assistance for Policy Reform II BearingPoint. Egypt: Dokki, Giza, 2006. 59 c. URL: https://pdf.usaid.gov/pdf_docs/PA00WC4R.pdf

122. Simandl M., Lesek M., Straka O.Pension fund model design and state estimation. Preprints of the 16th IFAC world congress, p. 1-6, IFAC, Prague, 2005.URL: https://kky-sw.zcu.cz/en/publications/SimandlM_2005_Pensionfundmodel

123. Anusiuba O. I. A., Ezuruka E. O., .Ekwealor O. U., Karim U. Development of an Enhanced Online Pension Management System *Journal of Scientific Engineering and Applied Science.* Vol.-7, Issue-3, 2021. P 62-91

124. Kolawole F. L., Festus A. F. Management information systems and effective business decisions: an evaluation of pension fund administrators in Nigeri. *Global Research Journal of Business Management*. 2022. Vol. 2, No. 2: P. 22 - 37

125. Pradana F. R., Pertiwi A. Analysis and Design Management Information System Pension Benefit Feature of Core System Dana Pensiun Lembaga Keuangan. *International Research Journal of Advanced Engineering and Science*. 2022. Vol. 7, Issue 1, P. 239-243.

126. Antari P., Bufra F. S., Andesti C. L. Analysis of pension fund payment data processing information system design. *Journal of International Accounting, Taxation and Information Systems*, 2024, №1(3), P. 144-149

127. Oduwole, O.A Jenyo, I.A Onamade,A .A Adegbite, O. Adegoke, B.O An improved design and implementation of a pension contribution monitoring system. *Journal of Computing and Informatics.* 2021. Vol. 2, Issue 1, P. 89-98

128. D'Arcangelis A. M., Levantesi S., Rotundo G. A complex networks approach to pension funds. *Journal of Business Research*. 2021, Vol. 129, P. 687-702 https://doi.org/10.1016/j.jbusres.2019.10.071

129. Петрик М. Р., Бойко І. В, Хіміч О. М., Петрик М. М. Високопродуктивні суперкомп'ютерів технології моделювання та ідентифікації складних нанопористих кіберсистем зі зворотними зв'язками для n-компонентної компетитивної адсорбції. *Кібернетика і системний аналіз*. 2021. Т. 57. №2. С. 170-183. https://doi.org/10.1007/s10559-021-00357

130. Information technology assessment of the pension systems. URL: https://pdf.usaid.gov/pdf_docs/PA00WC4R.pdf

131. Про затвердження регламенту роботи інформаційно-телекомунікаційної системи Пенсійного фонду України. Постанова ПравліннясПенсійного фонду України від 11.10.201 р. № 28-1. URL: https://www.pfu.gov.ua/content/uploads/2021/10/Postanova-28-1-vid-11-zhovtnya-2021-.pdf

132. Технічні вимоги на надання послуг зі створення Єдиної інформаційної система соціальної сфери. Перша черга. URL: https://tapas.org.ua/wp-content/uploads/2021/01/Terms-of-Reference-Tekhnichne-zavdannia.pdf

133. Statistical models and techniques used in actuarial analysis. https://fastercapital.com/topics/statistical-models-and-techniques-used-in-actuarial- analysis.html

134. Стандарт ВСТ 01.305.003-2019 (01) «Методичне забезпечення. Класифікація бойових уражень, небойових травм та захворювань у Збройних силах України»

135. Конфлікти, війни та соціальні трансформації епохи модерну: теорія, історія, сьогодення : матеріали XI Міжнародної наук.-практ. конф. (м. Київ,

12–13 червня 2023 р.) / укладачі П. В. Федорченко-Кутуєв, О. М. Казьмірова, О. М. Вольський та ін. Університетська книга, 2023. 164 с.

136. Кудрицький М. О., Патер Н. В., Кострица В. О. Пропозиції щодо алгоритму визначення бойових втрат особового складу з метою прогнозування зміни ефективності бойового застосування озброєння та військової техніки. *Військово-технічний збірник*. 2024. №31. С. 44-50 https://doi.org/10.33577/2312-4458.31.2024.44-51
137. Лівінський В. Г. Санітарні втрати як індикативний показник діяльності медичної служби Збройних Сил України. Сучасні аспекти військової медицини. 2020. Т. 27, № 2. С. 62-75. https://doi.org/10.32751/2310-4910-2020-27-28
138. За якими принципами обліковують санітарні втрати?URL: https://armyinform.com.ua/2020/01/14/za-yakymy-prynczypamy-oblikovuyut-sanitarni-vtraty/
139. Кочін І. В. Аналіз загальних втрат серед військовослужбовців і населення при бойових діях і їх особливості. Науковий огляд. 2019. №5(100). С. 29-39. https://doi.org/10.22141/2224-0586.5.100.2019.177015
140. The official site of Watson institute (Brown University). URL: http://watson.brown.edu/.
141. Фурсенко Л. К., Черновол Н. М. Ланчестеровські моделі бойових дій. *Збірник наукових праць Харківського національного університету Повітряних Сил,* 2020, № 4(66). С. 85-91. https://doi.org/10.30748/zhups.2020.66.12
142. Уражаюча дія сучасної зброї і характеристика санітарних втрат. URL: https://ppt-online.org/37598
143. Офіційний сайт НАТО. URL: http://www.nato.int/.
144. Global Terrorism Database Codebook: inclusion criteria and variables. USA: University of Maryland, 2019. 65 p.URL:: https://www.start.umd.edu/gtd/

145. Gallagher M., Sturgeon S., Finch B., Villongco F.. Probabilistic Analysis of Complex Combat Scenarios. *Military Operations Research*. 2022. Vol. 27, No. 1. P. 87–106. URL: https://www.jstor.org/stable/27116757.

146. Трофимчук О., Бідюк П., Кожухівська О., Кожухівський А. Ймовірнісно-статистичні невизначеності в системах підтримки прийняття рішень. Вісник Національного університету «Львівська політехніка». 2015. № 826. С. 237–248.

147. Довгий, С. О., Бідюк П. І., Трофимчук О. М. Системи підтримки прийняття рішень на основі статистично-ймовірнісних методів. Київ : Логос, 2014. 418 с.

148. Sankoff D., Kruskal J. Time warps, string edits, and macromolecules: The theory and practice of sequence comparison. 1999. 408 p. URL: https://web.stanford.edu/group/cslipublications/cslipublications/site/1575862174.shtml

149. Beyer W.A., Stein M.L., Smith T.F, Ulam S.M. A molecular-sequence metric and evolutionary trees / Mathematical Biosciences. 1974. Vol. 19. P. 9-25. URL: https://www.sciencedirect.com/science/article/abs/pii/0025556474900285

150. Wilbur W.I., Lipman D.I. Rapid similarity searches of nucleic acid and protein data banks. Proceeding of the National Academy of Science of the USA. 1983. Vol. 80. P. 726-730. URL: https://pubmed.ncbi.nlm.nih.gov/6572363/

151. Schubert S., Lee T. Time Series Data Mining with SAS® Enterprise Miner // Proceedings of the SAS Global Forum 2011 Conference. – SAS Institute Inc., Cary, NC, 2011. – 12 p. – link: https://support.sas.com/resources/papers/proceedings11/160-2011.pdf

152. Leonard M., Sloan J., Lee ., Elsheimer B. An Introduction to Similarity Analysis Using SAS // Proceedings of the SAS Global Forum 2007 Conference. – SAS Institute Inc., Cary, NC, 2007. 22 p. URL: https://support.sas.com/rnd/app/ets/papers/similarityanalysis.pdf

ДОДАТОК А

ПЕНСІЙНИЙ ФОНД УКРАЇНИ

ГОЛОВНЕ УПРАВЛІННЯ ПЕНСІЙНОГО ФОНДУ УКРАЇНИ У ЛЬВІВСЬКІЙ ОБЛАСТІ

вул. Митрополита Андрея, 10, м. Львів, 79016 факс. (032) 297-18-77, тел. (032) 255-04-95

E-mail: post@lv.pfu.gov.ua код згідно з ЄДРПОУ 13814885

від_____20___р. № _____ На № _____ від _____20___ р.

ДОВІДКА
про впровадження результатів дисертаційного дослідження Зарудного Олексія Борисовича на тему «Методи та моделі прогнозування для підтримки прийняття рішень у сфері державного пенсійного забезпечення»

Результати, отримані в ході виконання дисертаційного дослідження Зарудного Олексія Борисовича на тему «Моделі та методи прогнозування для підтримки прийняття рішень у системі державного пенсійного забезпечення» використовуються у практичній роботі Головного управління Пенсійного фонду України у Львівській області, зокрема, в процесі експлуатації Інтегрованої комплексної інформаційної системи Пенсійного фонду (ІКІС ПФ).

Начальник Головного управління **Галина НЕДЗЕЛЬСЬКА**

ТОВ " Науково-виробниче підприємство " МЕДИРЕНТ"
вул. Казимира Малевича, буд. 86Б, м . Київ , 03150, Україна
+38 (044) 200 54 10, +38 (044) 200 54 1
office@medirent.com.ua
www.medirent.com.ua

МЕДИРЕНТ
ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ

Research and development enterprise "MEDIRENT". LLC
86B, Kazimir Malevich str, 03150, Kyiv, Ukraine
+38 (044) 200 54 10, + 38 (044) 200 54 1
office@medirent.com.ua
www.medirent.com.ua

Вих. №139 від 17 червня 2025 року

Директору інституту
телекомунікацій та глобального
інформаційного простору
Трофимчуку О.М.

Шановний Олександре Миколайовичу!

Результати, отримані в ході виконання дисертаційного дослідження на тему «Моделі, методи та інформаційні технології збору та обробки даних для аналізу та прогнозування видатків на соціальне забезпечення та соціальний захист» впроваджені в процесі створення та модернізації Інтегрованої комплексної інформаційної системи Пенсійного фонду (ІКІС ПФ), яка розробляється групою компаній «Медирент» на замовлення Пенсійного фонду України, починаючи із 2002 року по цей час.

Директор
ТОВ „НВП <МЕДИРЕНТ>" Олександр Колесник

УКРАЇНА

ДЕРЖАВНА СЛУЖБА ВЛАСНОСТІ УКРАЇНИ ІНТЕЛЕКТУАЛЬНОЇ

РІШЕННЯ

про реєстрацію договору, який стосується права автора на твір

Державна служба інтелектуальної власності розглянула заяву

Товариство з обмеженою відповідальністю "Науково-виробниче підприємство "Медирент", вул. Щекавицька, 9 А, м. Київ, 04071

(повне ім'я фізичної або повне офіційне найменування юридичної особи, адреса)

про реєстрацію авторського договору від **31.10.2011, № 11056** про передачу (відчуження) майнових прав і прийняла рішення зареєструвати авторський договір, відповідно до якого майнові права на твір

Комп'ютерна програма "Інтегрована комплексна інформаційна система Пенсійного фонду України "ІКІС ПФУ", версія 2"; Зарудний Олексій Борисович, Колбун Віктор Андрійович, Малецький Олександр Миколайович, Машкін Владислав Генадійович, Рябцева Тетяна Борисівна, Матвєєв Сергій Валентинович, Хандрос Юрій Олександрович, Коваль Роман Григорович, Бондарчук Олексій Миколайович, Педан Юрій Анатолійович, Павлюков Іван Олексійович

(вид, повна, скорочена (за наявності) назва твору, повне ім'я, псевдонім (за наявності) автора(ів))

передаються(відчужуються)

Приватне акціонерне товариство "АТЛАС", вул. Щорса, 31, м. Київ, 01133

(повне ім'я фізичної(их) або повне офіційне найменування юридичної(их) особи(осіб), яка(і) передає(ють)(відчужує(ють)) право на твір, адреса)

Товариство з обмеженою відповідальністю "Науково-виробниче підприємство "Медирент", вул. Щекавицька, 9 А, м. Київ, 04071

(повне ім'я фізичної або повне офіційне найменування юридичної особи, якій передається (відчужується) право на твір, адреса)

Повністю

Реєстраційний номер **1672**

Дата реєстрації **11.11.2011**

Голова Державної служби інтелектуальної власності М.В.Паладій

Державна служба інтелектуальної власності України. Ідентифікаційний код 37552556

УКРАЇНА • UKRAINE • УКРАЇНА • UKRAINE • УКРАЇНА • UKRAINE • УКРАЇНА • UKRAINE • УКРАЇНА • UKRAINE • УКРАЇНА

Додаток Б


```
{
  "help": "https://data.gov.ua/api/3/action/help_show?name=datastore_search",
  "success": true,
  "result": {
    "resource_id": "3ec45fce-9ecd-4bd9-b682-c00af0cb0ef5",
    "fields": [
      { "id": "id",              "type": "int4"    },
      { "id": "region_code",     "type": "text"    },
      { "id": "year",            "type": "int4"    },
      { "id": "month",           "type": "int4"    },
      { "id": "expense_type",    "type": "text"    },
      { "id": "amount_nominal",  "type": "numeric" },
      { "id": "amount_real",     "type": "numeric" },
      { "id": "recipients_count","type": "int4"    }
    ],
    "records": [
      {
        "id": 1,
        "region_code": "UA-30",
        "year": 2023,
        "month": 1,
        "expense_type": "пенсії",
        "amount_nominal": 120000000.00,
        "amount_real": 115000000.00,
        "recipients_count": 253412
      },
      {
        "id": 2,
```

```
    "region_code": "UA-30",
    "year": 2023,
    "month": 2,
    "expense_type": "допомога сім’ям з дітьми",
    "amount_nominal": 45000000.00,
    "amount_real": 43000000.00,
    "recipients_count": 102345
   }
 ],
 "limit": 1000,
 "offset": 0,
 "total": 12456
}
```

```
import pandas as pd
import statsmodels.api as sm
import matplotlib.pyplot as plt

# 1. Дані (часові ряди по кварталах)
data = {
   "GDPpc": [42000, 43500, 44200, 45000],      # ВВП на душу, грн
   "Unemp": [8.2, 7.9, 7.5, 7.2],              # Безробіття, %
   "VPO": [56, 60, 63, 65],                    # ВПО, тис. осіб
   "Battles": [21, 18, 17, 15],                # Бойові події
   "Y": [4200, 4450, 4780, 4950]               # Соц. виплати на душу, грн
}
df = pd.DataFrame(data)

# 2. Побудова регресійної моделі
```

```
X = sm.add_constant(df[["GDPpc", "Unemp", "VPO", "Battles"]])  # незалежні змінні + константа
y = df["Y"]  # залежна змінна
model = sm.OLS(y, X).fit()

# 3. Вивід коефіцієнтів
print("Оцінені коефіцієнти регресійної моделі:")
print(model.params)

# 4. Додаємо прогнозовані значення в таблицю
df["Y_pred"] = model.predict(X)

# 5. Побудова графіка: Фактичні vs Прогнозні
plt.figure(figsize=(8, 5))
plt.plot(df.index, df["Y"], marker='o', label="Фактичні виплати")
plt.plot(df.index, df["Y_pred"], marker='s', linestyle='--', label="Прогнозні виплати")
plt.xticks(df.index, ["2023Q1", "2023Q2", "2023Q3", "2023Q4"])
plt.title("Фактичні vs Прогнозні соціальні виплати")
plt.xlabel("Квартал")
plt.ylabel("Сума на душу (грн)")
plt.legend()
plt.grid(True)
plt.tight_layout()
plt.show()

# 6. Побудова графіка залишків
residuals = df["Y"] - df["Y_pred"]
plt.figure(figsize=(8, 4))
plt.bar(df.index, residuals, color="salmon")
plt.xticks(df.index, ["2023Q1", "2023Q2", "2023Q3", "2023Q4"])
plt.axhline(0, color='gray', linestyle='--')
```

```
plt.title("Залишки моделі (Фактичне - Прогноз)")
plt.xlabel("Квартал")
plt.ylabel("Похибка (грн)")
plt.grid(True)
plt.tight_layout()
plt.show()

# 7. Зведена статистика (опціонально)
print("\nЗведена статистика моделі:")
print(model.summary())
```

ДОДАТОК В

Динамічно сформовані SQL-запити

Select isl_sql from ISS_LOG isl order by isl_dt desc;

```
SELECT /*+ FIRST_ROWS   INDEX(ppo2 XIF_PPO_LS) */ count(unique
pnf.pnf_number)
  FROM
pensionfile pnf,
pnf_history_data ls,
pnf_payorder ppo2,
v_opfu_ips org,
v_ddn_subject_tp subj_tp
  WHERE  pnf.pnf_id = case when ls.ls_number is not null then ls.ls_pnf else
ls.ls_pnf_lb end
 and pnf.com_org_ppa = org.org_id( + )
 and nvl(ls.ls_ls_lb, ls.ls_id) = ppo2.ppo_ls
 and ls.ls_subject_tp = subj_tp.dic_code

  AND (case when pnf.com_org between 29001 and 29999 then '2' else '1' end
in ('1'))--1421
 AND (ls.ls_drog_dt between to_date('01.01.1901', 'dd.mm.yyyy') and
to_date('01.04.1965', 'dd.mm.yyyy'))--725
 AND (org.org_org_id = '21001')--21
 AND (org.org_id = '21305')--2002
 AND (nsi_utils.GetShp(ls.ls_id, ppo2.ppo_id) in (('201')))--733
 AND (subj_tp.dic_code in (('DD'),('PENS'),('DPN'),('DECL')))--1521

  AND CASE WHEN ls.LS_SUBJECT_TP IN ('DD','DPN','DECL','PENS') AND
ls.LS_NUMBER IS NOT NULL OR ls.LS_SUBJECT_TP='BRW' THEN 1 ELSE 0 END = 1 and
ls.hist_st in ('A') and ls.ls_st <> 'CL'
 and nvl(ppo2.ppo_tp(+), 'M') = 'M' and ppo2.ppo_st(+) = 'A'
```

```
SELECT /*+ FIRST_ROWS   */  GROUPING(r01_number)+ GROUPING(r05_nind_name)+
GROUPING(r02_ppp_paytp)+ GROUPING(r02_adr_ind_1)+ GROUPING(r02_pp_psbpsb)+
GROUPING(r02_pp_psb) grp,
r01_number,
r05_nind_name,
r02_ppp_paytp,
r02_adr_ind_1,
r02_pp_psbpsb,
r02_pp_psb, max(ls_id_res) ls_id_res, max(ls_pnf_res) ls_pnf_res from (
```

```
select UNIQUE
pnf.pnf_number as r01_number,
r05_nind.nind_name as r05_nind_name,
r02_paytp.dic_sname as r02_ppp_paytp,
case when r02_ppp.ppp_pay_tp = 'P' then r02_ind.ind_name end as
r02_adr_ind_1,
r02_psb_psb.psb_filname as r02_pp_psbpsb,
r02_psb.psb_filname as r02_pp_psb, ls.ls_id ls_id_res, ls.ls_pnf ls_pnf_res
  FROM
pensionfile pnf,
pnf_history_data ls,
pnf_history_data ls_ls,
v_opfu_ips org,
v_ddn_subject_tp subj_tp,
pnf_individuality r05_pind,
nsi_individuality r05_nind,
pnf_personal_account r02_ppa,
pnf_pay_property r02_ppp,
v_ddn_paytp r02_paytp,
address r02_adr,
nsi_index r02_ind,
nsi_psb r02_psb,
nsi_psb r02_psb_psb
  WHERE  r02_ppp.ppp_psb = r02_psb.psb_id(+)
 and r02_adr.adr_index = r02_ind.ind_id(+)
 and pnf.pnf_id = case when ls.ls_number is not null then ls.ls_pnf else
ls.ls_pnf_lb end
 and r05_pind.pind_nind = r05_nind.nind_id
 and r02_ppp.ppp_adr = r02_adr.adr_id(+)
 and pnf.com_org_ppa = org.org_id( + )
 and (case when ls_ls.ls_number is not null then ls_ls.ls_id else
ls_ls.ls_ls_lb end) = ls.ls_id
 and ls_ls.ls_id = r05_pind.pind_ls
 and r02_ppa.ppa_id = r02_ppp.ppp_ppa(+)
 and r02_ppp.ppp_pay_tp = r02_paytp.dic_value(+)
 and ls.ls_pnf = r02_ppa.ppa_pnf(+)
 and ls.ls_subject_tp = subj_tp.dic_code
 and r02_psb.psb_psb = r02_psb_psb.psb_id(+)

  AND (case when pnf.com_org between 29001 and 29999 then '2' else '1' end
in ('1'))--1421
```

```
 AND (org.org_org_id = '21001')--21
 AND (org.org_id = '21305')--2002
 AND (r05_nind.nind_code in (94,4053,194))--653
 AND (subj_tp.dic_code in (('DD'),('PENS'),('DPN'),('DECL')))--1521

  AND CASE WHEN ls.LS_SUBJECT_TP IN ('DD','DPN','DECL','PENS') AND
ls.LS_NUMBER IS NOT NULL OR ls.LS_SUBJECT_TP='BRW' THEN 1 ELSE 0 END = 1 and
ls.hist_st in ('A') and ls.ls_st <> 'CL'
 and ls_ls.hist_st in ('A') and ls_ls.ls_st <> 'CL' and(ls.ls_subject_tp in
('DECL') and ls_ls.ls_subject_tp in ('DPN') and ls_ls.ls_number is null or
ls.ls_subject_tp in ('DD', 'DPN','PENS','DECL') and ls.ls_number is not null
)
 and r02_ppa.hist_st(+) in ('A')


)GROUP BY (r01_number, r05_nind_name, r02_ppp_paytp, r02_adr_ind_1,
r02_pp_psbpsb, r02_pp_psb)
ORDER BY r01_number ASC NULLS LAST, r05_nind_name ASC NULLS LAST,
r02_ppp_paytp ASC NULLS LAST, r02_adr_ind_1 ASC NULLS LAST, r02_pp_psbpsb
ASC NULLS LAST, r02_pp_psb ASC NULLS LAST,1
```

```
SELECT /*+ FIRST_ROWS   INDEX(ppa XIF_PPA_PNF) */  GROUPING(r01_number)+
GROUPING(pv_dt) grp,
r01_number,
pv_dt, max(ls_id_res) ls_id_res, max(ls_pnf_res) ls_pnf_res from (

select UNIQUE
pnf.pnf_number as r01_number,
pv.pv_dt as pv_dt, ls.ls_id ls_id_res, ls.ls_pnf ls_pnf_res
  FROM
pensionfile pnf,
pnf_history_data ls,
v_opfu_ips org,
v_ddn_subject_tp subj_tp,
pnf_personal_account ppa,
pnf_verification pv
  WHERE  pnf.pnf_id = case when ls.ls_number is not null then ls.ls_pnf else
ls.ls_pnf_lb end
 and ls.ls_pnf = ppa.ppa_pnf
 and pnf.com_org_ppa = org.org_id( + )
 and ls.ls_subject_tp = subj_tp.dic_code
```

```
 and pv.pv_pnf = ppa.ppa_pnf_payee and pv.pv_id =
ikis_ppvp.nsi_utils.GetCurPv(ppa.ppa_pnf_payee)

  AND (subj_tp.dic_code in (('DD'),('PENS'),('DPN'),('DECL')))--1521
 AND (pv.pv_dt between to_date('01.01.2023', 'dd.mm.yyyy') and
to_date('28.03.2025', 'dd.mm.yyyy'))--342
 AND (case when pnf.com_org between 29001 and 29999 then '2' else '1' end in
('1'))--1421
 AND (org.org_org_id = '1001')--21

  AND CASE WHEN ls.LS_SUBJECT_TP IN ('DD','DPN','DECL','PENS') AND
ls.LS_NUMBER IS NOT NULL OR ls.LS_SUBJECT_TP='BRW' THEN 1 ELSE 0 END = 1 and
ls.hist_st in ('A') and ls.ls_st <> 'CL'
 and ppa.hist_st(+) in ('A')


)GROUP BY (r01_number, pv_dt)
ORDER BY r01_number ASC NULLS LAST, pv_dt ASC NULLS LAST,1
```

```
SELECT /*+ FIRST_ROWS   */  GROUPING(org_name)+ GROUPING(r01_number)+
GROUPING(r05_nind_name) grp,
org_name,
r01_number,
r05_nind_name, max(ls_id_res) ls_id_res, max(ls_pnf_res) ls_pnf_res from (

select UNIQUE
org.org_name as org_name,
pnf.pnf_number as r01_number,
r05_nind.nind_name as r05_nind_name, ls.ls_id ls_id_res, ls.ls_pnf
ls_pnf_res
  FROM
pensionfile pnf,
pnf_history_data ls,
pnf_history_data ls_ls,
v_opfu_ips org,
v_ddn_subject_tp subj_tp,
pnf_individuality r05_pind,
nsi_individuality r05_nind
  WHERE  pnf.pnf_id = case when ls.ls_number is not null then ls.ls_pnf else
ls.ls_pnf_lb end
 and r05_pind.pind_nind = r05_nind.nind_id
 and pnf.com_org_ppa = org.org_id( + )
```

```
 and (case when ls_ls.ls_number is not null then ls_ls.ls_id else
ls_ls.ls_ls_lb end) = ls.ls_id
 and ls_ls.ls_id = r05_pind.pind_ls
 and ls.ls_subject_tp = subj_tp.dic_code

  AND (subj_tp.dic_code in (('DD'),('PENS'),('DPN'),('DECL')))--1521
 AND (case when pnf.com_org between 29001 and 29999 then '2' else '1' end in
('1'))--1421
 AND (r05_nind.nind_code in (40052))--653
 AND (org.org_org_id = '21001')--21

  AND CASE WHEN ls.LS_SUBJECT_TP IN ('DD','DPN','DECL','PENS') AND
ls.LS_NUMBER IS NOT NULL OR ls.LS_SUBJECT_TP='BRW' THEN 1 ELSE 0 END = 1 and
ls.hist_st in ('A') and ls.ls_st <> 'CL'
 and ls_ls.hist_st in ('A') and ls_ls.ls_st <> 'CL' and(ls.ls_subject_tp in
('DECL') and ls_ls.ls_subject_tp in ('DPN') and ls_ls.ls_number is null or
ls.ls_subject_tp in ('DD', 'DPN','PENS','DECL') and ls.ls_number is not null
)



)GROUP BY (org_name, r01_number, r05_nind_name)
ORDER BY org_name ASC NULLS LAST, r01_number ASC NULLS LAST, r05_nind_name
ASC NULLS LAST,1
```

```
SELECT /*+ FIRST_ROWS   */  GROUPING(r01_number)+ GROUPING(living_osob) grp,
r01_number,
living_osob, max(ls_id_res) ls_id_res, max(ls_pnf_res) ls_pnf_res from (

select UNIQUE
pnf.pnf_number as r01_number,
living_osob.dic_sname as living_osob, ls.ls_id ls_id_res, ls.ls_pnf
ls_pnf_res
  FROM
pensionfile pnf,
pnf_history_data ls,
v_opfu_ips org,
v_ddn_subject_tp subj_tp,
pnf_personal_account r03_ppa,
v_ddn_living_osob living_osob
  WHERE  pnf.pnf_id = case when ls.ls_number is not null then ls.ls_pnf else
ls.ls_pnf_lb end
 and pnf.com_org_ppa = org.org_id( + )
```

```
 and ls.ls_pnf = r03_ppa.ppa_pnf
 and nvl(r03_ppa.ppa_living_osob, 'N') = living_osob.dic_value
 and ls.ls_subject_tp = subj_tp.dic_code

  AND (case when pnf.com_org between 29001 and 29999 then '2' else '1' end
in ('1'))--1421
 AND (org.org_org_id = '21001')--21
 AND (org.org_id = '21305')--2002
 AND (subj_tp.dic_code in (('DD'),('PENS'),('DPN'),('DECL')))--1521
 AND (living_osob.dic_value in ('TA'))--1681

  AND CASE WHEN ls.LS_SUBJECT_TP IN ('DD','DPN','DECL','PENS') AND
ls.LS_NUMBER IS NOT NULL OR ls.LS_SUBJECT_TP='BRW' THEN 1 ELSE 0 END = 1 and
ls.hist_st in ('A') and ls.ls_st <> 'CL'
 and r03_ppa.hist_st(+) in ('A')


)GROUP BY (r01_number, living_osob)
ORDER BY r01_number ASC NULLS LAST, living_osob ASC NULLS LAST,1
```

```
SELECT /*+ FIRST_ROWS   */ count(unique pnf.pnf_number)
  FROM
pensionfile pnf,
pnf_history_data ls,
pnf_history_data ls_ls,
v_opfu_ips org,
v_ddn_subject_tp subj_tp,
pnf_individuality r05_pind,
nsi_individuality r05_nind
  WHERE  pnf.pnf_id = case when ls.ls_number is not null then ls.ls_pnf else
ls.ls_pnf_lb end
 and r05_pind.pind_nind = r05_nind.nind_id
 and pnf.com_org_ppa = org.org_id( + )
 and (case when ls_ls.ls_number is not null then ls_ls.ls_id else
ls_ls.ls_ls_lb end) = ls.ls_id
 and ls_ls.ls_id = r05_pind.pind_ls
 and ls.ls_subject_tp = subj_tp.dic_code

  AND (org.org_org_id = '21001')--21
 AND (org.org_id = '21305')--2002
 AND (subj_tp.dic_code in (('DD'),('PENS'),('DPN'),('DECL')))--1521
```

```
 AND (case when pnf.com_org between 29001 and 29999 then '2' else '1' end in
('1'))--1421
 AND (r05_nind.nind_code in (40052))--653

  AND CASE WHEN ls.LS_SUBJECT_TP IN ('DD','DPN','DECL','PENS') AND
ls.LS_NUMBER IS NOT NULL OR ls.LS_SUBJECT_TP='BRW' THEN 1 ELSE 0 END = 1 and
ls.hist_st in ('A') and ls.ls_st <> 'CL'
 and ls_ls.hist_st in ('A') and ls_ls.ls_st <> 'CL' and(ls.ls_subject_tp in
('DECL') and ls_ls.ls_subject_tp in ('DPN') and ls_ls.ls_number is null or
ls.ls_subject_tp in ('DD', 'DPN','PENS','DECL') and ls.ls_number is not null
)
```

```
SELECT /*+ FIRST_ROWS   */ count(unique pnf.pnf_number)
  FROM
pensionfile pnf,
pnf_history_data ls,
v_opfu_ips org,
v_ddn_subject_tp subj_tp,
pnf_charge r03_chr,
pnf_charge_detail r03_chd,
nsi_payment_tp r03_pm
  WHERE  pnf.pnf_id = case when ls.ls_number is not null then ls.ls_pnf else
ls.ls_pnf_lb end
 and pnf.com_org_ppa = org.org_id( + )
 and r03_chd.chd_pm = r03_pm.pm_id
 and ls.ls_pnf = r03_chr.chr_pnf_reciever
 and ls.ls_subject_tp = subj_tp.dic_code
 and r03_chr.chr_id = r03_chd.chd_chr

  AND (case when pnf.com_org between 29001 and 29999 then '2' else '1' end
in ('1'))--1421
 AND (org.org_id = '21305')--2002
 AND (org.org_org_id = '21001')--21
 AND (r03_chr.chr_month between to_date('18.03.2025', 'dd.mm.yyyy') and
to_date('25.03.2025', 'dd.mm.yyyy'))--802
 AND (r03_pm.pm_code between 5005 and 5008)--679
 AND (subj_tp.dic_code in (('DD'),('PENS'),('DPN'),('DECL')))--1521

  AND 1= 1 /*and r03_pm.pm_tp(+) in ('P', 'A')*/
 and (r03_chd.hist_st(+) in ('E','A') or r03_chd.chd_st(+) = 'R')
 and (r03_chr.hist_st(+) in ('A') or r03_chr.chr_st(+) = 'R')
```

```
 and CASE WHEN ls.LS_SUBJECT_TP IN ('DD','DPN','DECL','PENS') AND
ls.LS_NUMBER IS NOT NULL OR ls.LS_SUBJECT_TP='BRW' THEN 1 ELSE 0 END = 1 and
ls.hist_st in ('A') and ls.ls_st <> 'CL'
SELECT /*+ FIRST_ROWS   */  GROUPING(r01_number)+ GROUPING(r03_ppa_psn) grp,
r01_number,
r03_ppa_psn, max(ls_id_res) ls_id_res, max(ls_pnf_res) ls_pnf_res from (

select UNIQUE
pnf.pnf_number as r01_number,
nvl(r03_psn.psn_name, 'НЕ ЗНЯТИЙ') as r03_ppa_psn, ls.ls_id ls_id_res,
ls.ls_pnf ls_pnf_res
  FROM
pensionfile pnf,
pnf_history_data ls,
v_opfu_ips org,
v_ddn_subject_tp subj_tp,
pnf_personal_account r03_ppa,
nsi_prich_sn r03_psn
  WHERE  pnf.pnf_id = case when ls.ls_number is not null then ls.ls_pnf else
ls.ls_pnf_lb end
 and pnf.com_org_ppa = org.org_id( + )
 and nvl(r03_ppa.ppa_psn, 1) = r03_psn.psn_id
 and ls.ls_pnf = r03_ppa.ppa_pnf
 and ls.ls_subject_tp = subj_tp.dic_code

  AND (case when pnf.com_org between 29001 and 29999 then '2' else '1' end
in ('1'))--1421
 AND (org.org_org_id = '21001')--21
 AND (org.org_id = '21305')--2002
 AND (nvl(r03_psn.psn_code, 0) in (('16'),('34')))--677
 AND (subj_tp.dic_code in (('DD'),('PENS'),('DPN'),('DECL')))--1521

  AND CASE WHEN ls.LS_SUBJECT_TP IN ('DD','DPN','DECL','PENS') AND
ls.LS_NUMBER IS NOT NULL OR ls.LS_SUBJECT_TP='BRW' THEN 1 ELSE 0 END = 1 and
ls.hist_st in ('A') and ls.ls_st <> 'CL'
 and r03_ppa.hist_st(+) in ('A')


)GROUP BY (r01_number, r03_ppa_psn)
ORDER BY r01_number ASC NULLS LAST, r03_ppa_psn ASC NULLS LAST,1
```