

**МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
УМАНСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ**

Факультет економіки і підприємництва

Кафедра інформаційних технологій

Допущено до захисту

«_____» _____ 2025 р.

Зав. кафедри, к.т.н., доцент

_____ **Ліщук Р.І.**

**КВАЛІФІКАЦІЙНА РОБОТА
БАКАЛАВРА**

**РОЗРОБКА ІНФОРМАЦІЙНОЇ ПІДСИСТЕМИ АНАЛІЗУ
ІНФОРМАЦІЇ ПРО ОПЛАТУ ЗАКЛАДАМ ОХОРОНИ
ЗДОРОВ'Я ЗА ПРОГРАМОЮ МЕДИЧНИХ ГАРАНТІЙ НА
ОСНОВІ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ**

Виконав: студент 4 курсу, групи 22к-кн
Освітньої програми «Комп'ютерні науки»
спеціальності 122 Комп'ютерні науки
Жданов Юрій Олександрович

Науковий керівник - д.т.н., професор
Кулаков Павло Ігорович

Зміст

ПЕРЕЛІК УМОВНИХ ПОЗНАЧОК, СКОРОЧЕНЬ ТА ТЕРМІНІВ..... **Помилка!**

Закладку не визначено.

ВСТУП	3
РОЗДІЛ 1. ДОСЛІДЖЕННЯ ПРЕДМЕТНОЇ ОБЛАСТІ ТА ПОСТАНОВКА ЗАДАЧІ.....	5
1.1. Актуальність теми.....	5
1.2. Аналіз існуючих рішень та публікацій	6
1.3. Методи інтелектуального аналізу даних в охороні здоров'я.....	10
1.4. Постановка задачі кваліфікаційної роботи.....	12
РОЗДІЛ 2. ПРОЄКТУВАННЯ ІНФОРМАЦІЙНОЇ ПІДСИСТЕМИ	14
2.1. Концептуальна модель підсистеми	15
2.2. Функціональні та нефункціональні вимоги	16
2.3. Архітектура підсистеми	18
2.4. Модель даних	19
2.5. Опис діаграми потоків даних.....	20
РОЗДІЛ 3. РОЗРОБКА ТА РЕАЛІЗАЦІЯ ІНФОРМАЦІЙНОЇ ПІДСИСТЕМИ .	23
3.1. Інструментальні засоби розробки	23
3.2. Реалізація модуля інтелектуального аналізу даних (ІАД).....	24
3.3. Реалізація модуля фільтрації даних	28
3.4. Реалізація модуля прогнозування	29
3.5. Реалізація модуля порівняння та оцінки	31
РОЗДІЛ 4. АНАЛІЗ РЕЗУЛЬТАТІВ РОБОТИ ТА ВИСНОВКИ	34
4.1. Аналіз результатів дослідного аналізу даних	34
4.2. Аналіз точності прогнозування	35
4.3. Можливості впровадження та розвитку підсистеми	38
ВИСНОВОК.....	39
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ.....	41
ДОДАТОК А.....	Помилка! Закладку не визначено.

Вступ

Система охорони здоров'я України перебуває в стані постійної трансформації, ключовим елементом якої є Програма медичних гарантій (ПМГ)[2]. Ця програма докорінно змінила принципи фінансування медичних послуг, перейшовши від утримання інфраструктури закладів до оплати за конкретно надані медичні послуги пацієнтам. Національна служба здоров'я України (НСЗУ), як центральний закупівельник медичних послуг, оперує значними обсягами фінансових та медичних даних, що стосуються оплат закладам охорони здоров'я (ЗОЗ).

Ефективне управління цими фінансовими потоками, забезпечення їх прозорості та прогнозування майбутніх витрат є критично важливим для стабільності та розвитку медичної галузі. Великі масиви даних про оплати, що генеруються в рамках ПМГ, містять цінну інформацію, яка може бути використана для аналізу ефективності програми, виявлення тенденцій, аномалій та обґрунтованого планування бюджету. Однак ручний аналіз таких обсягів даних є неефективним та трудомістким.

Саме тут на допомогу приходять методи інтелектуального аналізу даних (ІАД) та машинного навчання, які дозволяють автоматизувати процес виявлення знань із даних, будувати прогнозні моделі та надавати підтримку в прийнятті управлінських рішень. Розробка спеціалізованої інформаційної підсистеми, яка б інтегрувала ці методи для аналізу оплат ЗОЗ за ПМГ, є актуальним та важливим завданням.

Метою даної кваліфікаційної роботи бакалавра є розробка інформаційної підсистеми, призначеної для аналізу даних про оплати закладам охорони здоров'я за Програмою медичних гарантій з використанням методів інтелектуального аналізу даних та побудови прогнозів майбутніх оплат.

Для досягнення поставленої мети було визначено наступні завдання:

1. Дослідити предметну область – фінансування ЗОЗ в рамках ПМГ, діяльність НСЗУ.

2. Проаналізувати існуючі підходи та інформаційні системи для аналізу даних в охороні здоров'я.
3. Обрати та обґрунтувати методи інтелектуального аналізу даних для вирішення поставлених завдань (аналіз, прогнозування).
4. Спроекувати архітектуру та компоненти інформаційної підсистеми.
5. Розробити програмні модулі для реалізації функціоналу підсистеми: попередньої обробки даних, дослідного аналізу, фільтрації, прогнозування та оцінки якості прогнозу.
6. Реалізувати конвеєр обробки даних, що забезпечує послідовне виконання аналітичних задач.
7. Провести аналіз отриманих результатів роботи підсистеми на прикладі реальних даних про оплати.
8. Сформулювати висновки щодо розробленої підсистеми, її можливостей та напрямків подальшого розвитку.

Об'єктом дослідження є процес аналізу та прогнозування оплат закладам охорони здоров'я за Програмою медичних гарантій.

Предметом дослідження є методи інтелектуального аналізу даних та програмні засоби для їх застосування в аналізі фінансових даних ПМГ.

Структурно кваліфікаційна робота бакалавра складається зі вступу, чотирьох розділів, висновків, списку використаних джерел та додатків. У першому розділі розглядається актуальність теми, аналізуються існуючі рішення та методи ІАД, формулюється постановка задачі. Другий розділ присвячений проєктуванню інформаційної підсистеми: її архітектурі, компонентам, моделі даних. У третьому розділі детально описується процес розробки програмних модулів підсистеми та їх інтеграція. Четвертий розділ містить аналіз результатів роботи підсистеми, оцінку точності прогнозування та обговорення можливостей її впровадження. У висновках підсумовуються результати виконаної роботи.

РОЗДІЛ 1. ДОСЛІДЖЕННЯ ПРЕДМЕТНОЇ ОБЛАСТІ ТА ПОСТАНОВКА ЗАДАЧІ

1.1. Актуальність теми

Українська система охорони здоров'я перебуває у процесі глибокої трансформації, ключовим рушієм якої виступає Програма медичних гарантій (ПМГ), запроваджена у 2018 році. Ця програма кардинально змінила підхід до фінансування медичної галузі, здійснивши перехід від утримання інфраструктури закладів до оплати за фактично надані пацієнтам медичні послуги, реалізуючи принцип "гроші ходять за пацієнтом". За цією моделлю, Національна служба здоров'я України (НСЗУ), виступаючи як єдиний стратегічний закупівельник медичних послуг, контракує тисячі закладів охорони здоров'я (ЗОЗ) різних форм власності та здійснює їм оплати на підставі даних, що генеруються та обробляються в Електронній системі охорони здоров'я (ЕСОЗ) [4].

Такий механізм фінансування породжує величезні потоки даних. Кожна взаємодія пацієнта з системою, що покривається ПМГ, фіксується в ЕСОЗ та стає підставою для фінансових транзакцій. Як наслідок, НСЗУ акумулює значні масиви деталізованої інформації: суми та дати платежів, типи оплат, специфікації пакетів медичних послуг, географічна прив'язка, ідентифікатори ЗОЗ та інше. Зважаючи на постійне розширення ПМГ та зростання обсягів фінансування, ефективне управління цими даними стає критично важливим.

Нагальна потреба в розробці спеціалізованої інформаційної підсистеми для аналізу цих даних зумовлена кількома взаємопов'язаними чинниками.

По-перше, існує гостра необхідність в інструментах для оперативного моніторингу та об'єктивної оцінки ефективності використання коштів ПМГ. Це дозволить НСЗУ, Міністерству охорони здоров'я (МОЗ), регіональним органам влади та адміністраціям ЗОЗ не лише відстежувати фінансові потоки, але й виявляти закономірності споживання медичних послуг, ідентифікувати потенційні диспропорції у фінансуванні, а також виявляти аномальні транзакції,

які можуть сигналізувати про помилки чи потребують додаткової перевірки [4, 5].

По-друге, можливість точного прогнозування майбутніх обсягів оплат є ключовою для стабільного функціонування всієї системи. Такі прогнози необхідні для реалістичного бюджетного планування на національному та регіональному рівнях, а також для забезпечення фінансової стійкості самих ЗОЗ, дозволяючи їм ефективніше управляти ресурсами [4, 5].

По-третє, традиційні підходи до аналізу, особливо ручна обробка, стають неефективними та надзвичайно трудомісткими при роботі з такими обсягами даних. Саме тут на перший план виходять методи інтелектуального аналізу даних (ІАД), які дозволяють автоматизувати процес пошуку знань, виявляти складні, неочевидні залежності та патерни, недоступні при поверхневому розгляді.

Нарешті, розробка та впровадження таких аналітичних інструментів сприятиме підвищенню прозорості використання державних коштів та покращенню керованості системи фінансування охорони здоров'я загалом.

Отже, створення інформаційної підсистеми, що інтегрує сучасні методи ІАД для аналізу та прогнозування даних про оплати за ПМГ, є своєчасним та важливим завданням, спрямованим на підтримку реформи та підвищення ефективності медичної галузі України.

1.2. Аналіз існуючих рішень та публікацій

Питання аналізу фінансових потоків в системах охорони здоров'я є предметом уваги в багатьох країнах, і Україна не є винятком, особливо після запуску ПМГ. Однак наявні рішення та джерела даних мають певні обмеження, що не дозволяють повною мірою вирішити завдання глибокого аналізу та прогнозування. Важливо розглянути основні з них для розуміння контексту та обґрунтування необхідності розробки нової підсистеми.

Національна служба здоров'я України (НСЗУ), як ключовий оператор ПМГ, публікує певні дані та аналітику на своєму офіційному веб-сайті та спеціалізованих інформаційних панелях (дашбордах) (Рис 1.1) [4].



Рисунок 1.1 – Дашборди НСЗУ

Ці ресурси зазвичай містять агреговану інформацію про кількість укладених договорів, обсяги фінансування за окремими напрямками та регіонами, статистику по пакетах послуг. Наприклад, НСЗУ регулярно оприлюднює дані про виплати закладам та аналітичні звіти щодо виконання договорів. Ці інструменти є корисними для загального моніторингу програми, проте вони, як правило, надають описову статистику і можуть не пропонувати глибоких інструментів для інтелектуального аналізу, виявлення складних залежностей або побудови прогнозних моделей для зовнішніх дослідників чи управлінців на місцях.

Основою для всіх розрахунків та аналізу є дані Електронної системи охорони здоров'я (ЕСОЗ або eHealth)(Рис 1.2).

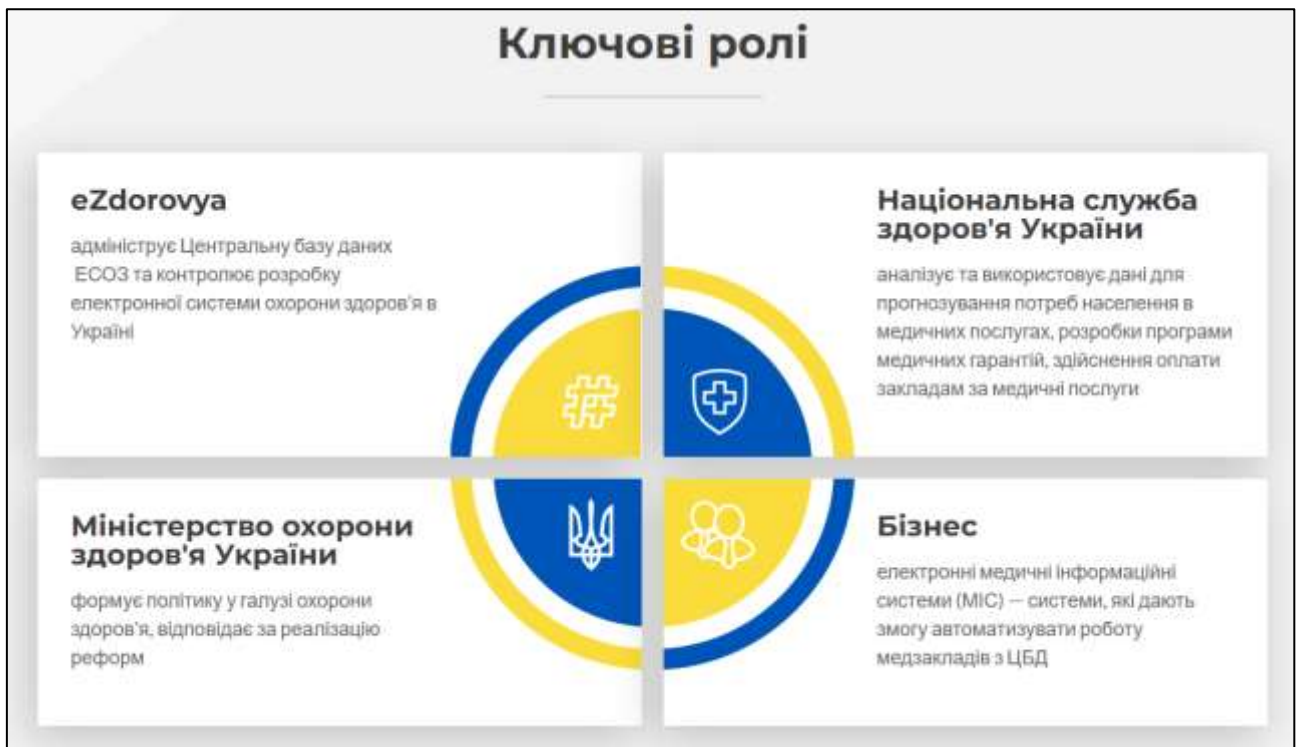


Рисунок 1.2 – Ключові ролі ЕСОЗ

Це централізована система, де зберігається медична інформація, дані про пацієнтів, медичні записи та звіти про надані послуги, на основі яких НСЗУ здійснює оплати. ЕСОЗ є надзвичайно важливим джерелом первинних даних. Однак, її основне призначення – операційна діяльність, забезпечення взаємодії між пацієнтами, лікарями та НСЗУ. Публічні аналітичні модулі ЕСОЗ можуть бути обмеженими або орієнтованими переважно на операційну звітність, а не на фінансовий аналіз чи прогнозування.

Важливим ресурсом є Єдиний державний веб-портал відкритих даних (data.gov.ua), де НСЗУ та МОЗ періодично публікують знеособлені набори даних, включно з даними про оплати за ПМГ (Рис 1.3) [2, 4, 5].

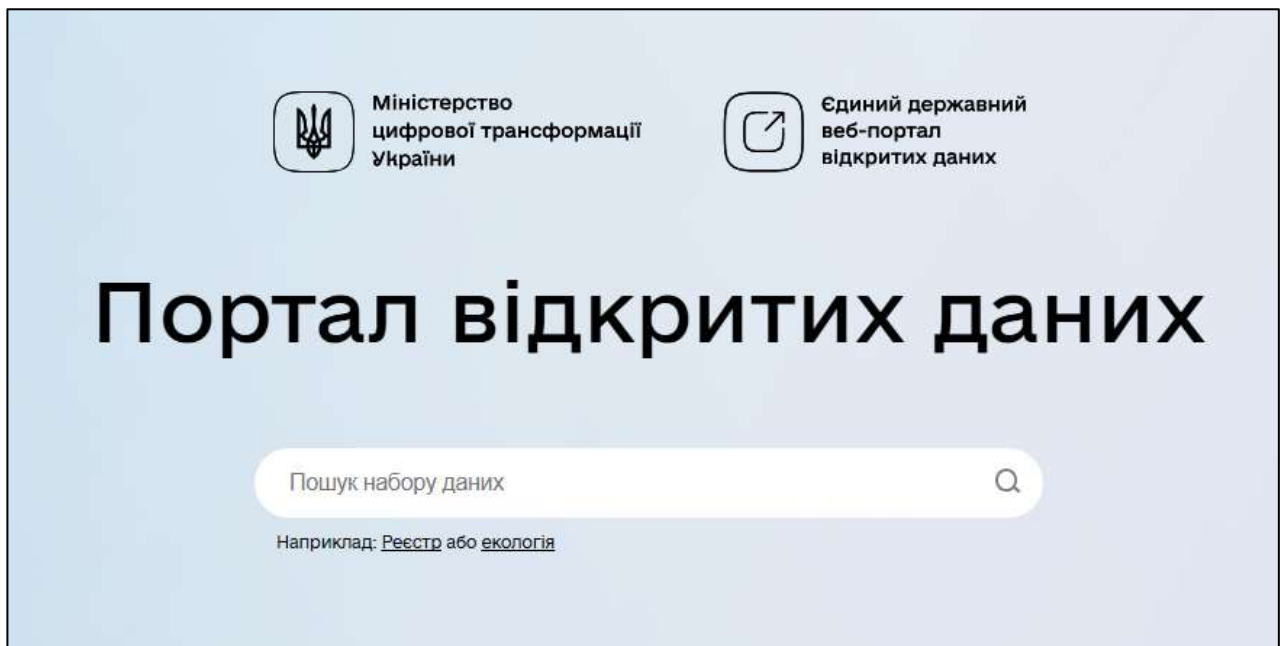


Рисунок 1.3 - Єдиний державний веб-портал відкритих даних

Ці дані відкривають широкі можливості для незалежних досліджень та аналізу громадськістю, аналітиками та розробниками. Проте, робота з "сирими" відкритими даними вимагає від користувачів значних зусиль з їх очищення, обробки, а також наявності власних інструментів та навичок для аналізу та моделювання [2]. Самі портали не пропонують готових аналітичних чи прогнозних підсистем. Саме розробка такого інструменту і є завданням даного проєкту.

Академічні дослідження та публікації. Українські науковці, економісти та фахівці у сфері громадського здоров'я проводять дослідження, аналізуючи різні аспекти ПМГ, включаючи фінансові. Ці дослідження публікуються у наукових журналах, на конференціях та в аналітичних звітах. Вони часто містять глибокий аналіз конкретних питань, але зазвичай є разовими дослідженнями, а не постійно діючими інформаційними системами чи інструментами.

Якщо розглядати міжнародний досвід, то країни з розвиненими системами медичного страхування або державним фінансуванням охорони здоров'я активно використовують інформаційні технології для аналізу фінансових даних. Це можуть бути комплексні сховища даних Data Warehouses, які включають аналітичну базу даних і критичні аналітичні компоненти та процедури. Вони

підтримують спеціальний аналіз і спеціальні звіти, такі як канали даних, запити та бізнес-додатки.

Платформи бізнес-аналітики Business Intelligence, BI — комп'ютерні методи і інструменти для організацій, що забезпечують переклад транзакційної ділової інформації в форму, придатну для бізнес-аналізу, а також засоби для роботи з обробленою таким чином інформацією.

Системи підтримки прийняття рішень. Часто застосовуються методи машинного навчання для прогнозування витрат, виявлення шахрайства (fraud detection), аналізу ефективності різних моделей фінансування та оптимізації розподілу ресурсів.

Таким чином, незважаючи на наявність окремих джерел даних та інструментів базової звітності, в Україні існує чітка потреба у створенні більш гнучких та потужних спеціалізованих інформаційних підсистем. Ці підсистеми мають бути сфокусовані на поглибленому інтелектуальному аналізі великих обсягів даних про оплати за ПМГ та, що особливо важливо, на побудові точних прогнозів майбутніх витрат з використанням сучасних алгоритмів машинного навчання. Розробка саме такої підсистеми, що поєднує аналіз та прогнозування, є центральним завданням цієї роботи.

1.3. Методи інтелектуального аналізу даних в охороні здоров'я

Інтелектуальний аналіз даних (ІАД), відомий також як Data Mining чи Knowledge Discovery in Databases (KDD), являє собою процес застосування автоматизованих методів для виявлення раніше невідомих, нетривіальних, потенційно корисних та інтерпретованих закономірностей (патернів, знань) у великих наборах даних [9]. Це міждисциплінарна галузь, що поєднує підходи зі статистики, машинного навчання, теорії баз даних та штучного інтелекту [9]. В контексті аналізу фінансових даних ПМГ, де обсяги інформації є значними, а залежності можуть бути складними та неочевидними, застосування ІАД надає суттєві переваги порівняно з традиційними методами.

Першим і foundational етапом є Дослідницький аналіз даних (Exploratory Data Analysis, EDA). Його основна мета полягає не стільки в побудові кінцевих

моделей, скільки в глибокому зануренні в дані для розуміння їхньої природи, структури, якості та основних характеристик. EDA передбачає розрахунок описових статистик для оцінки центральних тенденцій та варіативності показників, а також візуалізацію даних за допомогою різноманітних графічних інструментів (гістограми, діаграми розсіювання, коробкові діаграми, графіки часових рядів тощо). Для таких візуалізацій можна використовувати потужні інструменти, що надаються бібліотеками Python, такими як Matplotlib та Seaborn [8]. Також на цьому етапі проводиться аналіз пропущених значень, перевірка гіпотез щодо даних та розрахунок кореляцій. Якісне проведення EDA дозволяє виявити потенційні проблеми (помилки, аномалії), сформулювати інтуїтивне розуміння даних та підготувати їх до подальшого, більш складного моделювання.

Другим є Аналіз часових рядів та Прогнозування. Дані про щомісячні оплати за ПМГ є часовим рядом – послідовністю значень, виміряних у послідовні моменти часу. Аналіз таких рядів дозволяє виявити тренди (довгострокові тенденції до зростання чи спадання), сезонність (регулярні коливання протягом року) та випадкові флуктуації. Основною задачею при роботі з часовими рядами в проєкті є прогнозування – передбачення майбутніх значень оплат на основі історичних даних. Для цього було обрано регресійний підхід із застосуванням алгоритму XGBoost (Extreme Gradient Boosting) [3]. Цей алгоритм належить до сімейства методів градієнтного бустингу дерев рішень і відомий своєю високою точністю та ефективністю на структурованих (табличних) даних [3]. Для прогнозування часових рядів за допомогою XGBoost створюються спеціальні ознаки (features) на основі часової інформації: номер місяця, тригонометричні функції від номера місяця (для моделювання сезонності), лагові значення (значення показника за попередні періоди часу) тощо. Модель навчається знаходити залежність між цими ознаками та цільовою змінною (сумою оплат).

Третім йде Виявлення аномалій (Anomaly Detection). Цей метод спрямований на пошук об'єктів (у нашому випадку – окремих транзакцій або періодів), які суттєво відрізняються від загальної маси даних. Аномалії можуть

свідчити про помилки в даних, нетипові ситуації або навіть шахрайські дії. В рамках дослідного аналізу було використано простий статистичний метод на основі міжквартильного розмаху (Interquartile Range, IQR). Значення, що виходять за межі певного діапазону (наприклад, $Q3 + 1.5 \cdot IQR$), вважаються потенційними аномаліями (викидами) і можуть потребувати додаткового дослідження.

Та останнім є Оцінка якості моделі. Після побудови прогнозної моделі необхідно оцінити, наскільки добре вона працює, тобто наскільки точно її прогнози відповідають реальним даним. Для цього використовуються різні метрики похибки. У проєкті було розраховано:

- MAE (Mean Absolute Error) – середня абсолютна похибка. Показує середнє абсолютне відхилення прогнозу від факту в одиницях вимірювання цільової змінної (гривнях);
- RMSE (Root Mean Squared Error) – корінь із середньоквадратичної похибки. Також вимірюється в одиницях цільової змінної, але штрафує за великі похибки сильніше, ніж MAE;
- MAPE (Mean Absolute Percentage Error) – середня абсолютна відсоткова похибка. Показує середнє відхилення у відсотках, що зручно для порівняння точності моделей на різних даних або для різних показників.

Реалізацію цих та інших метрик можна знайти, наприклад, у бібліотеці Scikit-learn [7].

Вибір цих методів дозволив комплексно підійти до аналізу даних: від первинного дослідження та виявлення закономірностей до побудови та оцінки прогнозної моделі, що відповідає цілям дипломного проєкту.

1.4. Постановка задачі кваліфікаційної роботи

На основі аналізу актуальності теми, існуючих рішень та можливостей методів ІАД, було сформульовано основну задачу даної кваліфікаційної роботи.

Задача полягає у розробці програмної інформаційної підсистеми, яка реалізує конвеєр обробки даних для аналізу та прогнозування фінансових

показників (загальної суми оплати та суми оплати за пакетом) закладам охорони здоров'я в рамках Програми медичних гарантій України.

Підсистема повинна вирішувати наступні конкретні завдання:

1. Здійснювати дослідницький аналіз даних (EDA) на основі вхідного файлу з детальною інформацією про оплати за певний період (рік). Цей аналіз має включати розрахунок описових статистик, візуалізацію розподілів ключових показників, аналіз кореляцій, виявлення потенційних аномалій та аналіз даних за різними групами (наприклад, за пакетами послуг). Цей функціонал має бути реалізований у вигляді окремого модуля (скрипта).

2. Виконувати попередню обробку та фільтрацію даних. Перед прогнозуванням необхідно очистити дані від некоректних значень (наприклад, від'ємних сум оплат) та відібрати лише ті стовпці, які необхідні для побудови моделі прогнозування. Цей етап має готувати дані для подальших кроків.

3. Будувати прогнозну модель на основі історичних даних (за попередні роки) та генерувати прогноз ключових фінансових показників (pay_package, pay_all) на майбутній період (наприклад, наступний рік) на агрегованому місячному рівні. Модель має враховувати часові залежності (тренди, сезонність) і бути достатньо універсальною для застосування до даних різних років. Як основний алгоритм прогнозування було обрано XGBoost.

4. Здійснювати порівняння згенерованого прогнозу з фактичними даними за відповідний період (коли ці дані стають доступними). Необхідно розраховувати метрики точності прогнозу (MAE, RMSE, MAPE) та візуалізувати порівняння прогнозованих і фактичних значень.

5. Забезпечити послідовне виконання етапів фільтрації, прогнозування та порівняння у вигляді єдиного конвеєру обробки даних.

Вимоги до розроблюваної підсистеми:

- Мова програмування: Python.
- Основні бібліотеки: Pandas (для маніпуляції даними), NumPy (для числових операцій), Matplotlib та Seaborn (для візуалізації), Scikit-learn (для метрик та допоміжних функцій), XGBoost (для моделювання).

- Вхідні дані: CSV файли зі стандартизованою структурою, що містять дані про оплати НСЗУ (з колонками period_month, period_year, pay_package, pay_all, pay_date та іншими для EDA).
- Вихідні дані: Очищені та відфільтровані дані (проміжні CSV), файл з прогнозом (CSV), текстовий звіт з метриками точності, візуалізації (графіки).
- Стиль реалізації: Дослідницько-розробний, з фокусом на реалізацію та аналіз алгоритмів ІАД та прогнозування.
- Універсальність: Скрипти прогнозування та порівняння мають бути розраховані на можливість використання з даними за різні роки для прогнозування наступних періодів.
- Представлення результатів: Результати аналізу та прогнозування мають бути представлені у зрозумілому вигляді (таблиці, графіки, метрики).

Вирішення поставленої задачі дозволить створити програмний інструмент для глибшого аналізу фінансування за ПМГ та підтримки прийняття рішень на основі даних.

РОЗДІЛ 2. ПРОЄКТУВАННЯ ІНФОРМАЦІЙНОЇ ПІДСИСТЕМИ

На основі постановки задачі було розроблено проєкт інформаційної підсистеми аналізу інформації про оплати закладам охорони здоров'я за ПМГ. Цей розділ описує концептуальну модель, вимоги, архітектуру та модель даних розробленої підсистеми.

2.1. Концептуальна модель підсистеми

Розроблювана інформаційна підсистема призначена для автоматизації процесів аналізу та прогнозування фінансових потоків у рамках Програми медичних гарантій. Концептуально її можна представити як набір взаємопов'язаних програмних модулів, які послідовно обробляють дані для досягнення кінцевої мети.

Основне призначення підсистеми – надати інструменти для:

1. Глибокого дослідного аналізу історичних даних про оплати за певний період для виявлення закономірностей, аномалій та ключових факторів, що впливають на обсяги фінансування.
2. Побудови прогнозів майбутніх обсягів оплат (загальних та за пакетами) на місячному рівні на основі історичних даних.
3. Оцінки точності побудованих прогнозів шляхом їх порівняння з фактичними даними.

Вхідними даними для підсистеми є CSV файли, що містять детальну інформацію про оплати НСЗУ закладам охорони здоров'я. Передбачається, що ці файли мають певну структуру, яка включає щонайменше ідентифікатори періоду (рік, місяць), дату платежу, суму оплати за пакетом (pay_package), загальну суму оплати (pay_all) та інші атрибути, необхідні для аналізу (ID пакету, тип оплати тощо).

Вихідними даними підсистеми є:

- Результати дослідного аналізу: описові статистики, візуалізації (графіки), виявлені аномалії, кореляції (представлені у вигляді виводу в консоль та графічних файлів/вікон).
- Проміжні файли: очищені та відфільтровані дані у форматі CSV.
- Кінцевий результат прогнозування: CSV файл з прогнозованими значеннями pay_package та pay_all на майбутні місяці.
- Результати оцінки прогнозу: текстовий звіт з розрахованими метриками точності (MAE, RMSE, MAPE) та візуалізація порівняння прогнозу з фактом (графік).

Підсистема складається з чотирьох основних логічних блоків (модулів), реалізованих у вигляді окремих Python скриптів: Модуль ІАД, Модуль Фільтрації, Модуль Прогнозування, Модуль Порівняння.

2.2. Функціональні та нефункціональні вимоги

Функціональні вимоги (що система повинна робити):

1. Завантаження даних: Підсистема повинна мати можливість завантажувати вхідні дані з CSV файлів. Модуль ІАД має підтримувати інтерактивне завантаження (наприклад, через інтерфейс Google Colab). Інші модулі мають працювати з файлами за вказаними шляхами.

2. Дослідницький аналіз даних (EDA): Підсистема повинна виконувати EDA, включаючи:

- Виведення базової інформації про дані (розмірність, типи даних, наявність пропусків).
- Розрахунок та виведення описових статистик.
- Візуалізацію розподілів ключових числових змінних (гістограми, коробкові діаграми).
- Аналіз категоріальних змінних (розподіл за пакетами, типами оплат).
- Аналіз динаміки оплат у часі (графіки часових рядів).
- Аналіз взаємозв'язків між змінними (діаграми розсіювання, розрахунок кореляції).
- Виявлення та виведення потенційних аномалій (на основі IQR).
- Групування даних для аналізу (наприклад, середні оплати за пакетами).

3. Очищення даних: Підсистема повинна мати функціонал для видалення некоректних записів (наприклад, рядків з від'ємними сумами оплат).

4. Фільтрація даних: Підсистема повинна дозволяти відбирати лише необхідний набір стовпців для подальшого прогнозування.

5. Агрегація даних: Підсистема повинна агрегувати детальні дані про оплати до місячного рівня для побудови прогнозів.

6. Інженерія ознак: Перед прогнозуванням підсистема повинна генерувати часові ознаки (індекс місяця, тригонометричні ознаки, лагові значення) для моделі.

7. Тренування прогнозової моделі: Підсистема повинна тренувати модель машинного навчання (XGBoost) на історичних даних для прогнозування `pay_package` та `pay_all`.

8. Генерація прогнозу: Підсистема повинна генерувати прогноз значень `pay_package` та `pay_all` на заданий майбутній період (наприклад, 12 місяців).

9. Збереження прогнозу: Згенерований прогноз має зберігатися у CSV файл.

10. Порівняння прогнозу з фактом: Підсистема повинна завантажувати згенерований прогноз та фактичні дані за той самий період, об'єднувати їх.

11. Розрахунок метрик точності: Мають бути розраховані метрики MAE, RMSE, MAPE для оцінки якості прогнозу.

12. Візуалізація порівняння: Результати порівняння прогнозу та факту мають бути візуалізовані на графіку.

13. Генерація звіту: Має генеруватися текстовий файл зі звітом про результати порівняння та метрики точності.

Нефункціональні вимоги:

1. Універсальність: Скрипти прогнозування та порівняння мають бути спроектовані таким чином, щоб їх можна було використовувати для обробки даних за різні роки та прогнозування на різні майбутні періоди без суттєвої модифікації коду (шляхом зміни вхідних файлів та параметрів).

2. Простота використання: Взаємодія з підсистемою має відбуватися через запуск Python скриптів з командного рядка або в середовищі розробки (наприклад, Google Colab для ІАД). Налаштування (шляхи до файлів) мають бути чітко визначені у коді.

3. Модифікованість та розширюваність: Структура коду має бути модульною, щоб полегшити внесення змін або додавання нового функціоналу (наприклад, інших методів прогнозування, нових видів аналізу).

4. Продуктивність: Хоча висока продуктивність не є критичною вимогою для кваліфікаційної роботи, підсистема має обробляти типові обсяги даних (десятки-сотні тисяч записів) за прийнятний час на сучасному персональному комп'ютері. Використання бібліотек Pandas та NumPy сприяє ефективній обробці даних.

2.3. Архітектура підсистеми

Враховуючи послідовний характер обробки даних від сирих даних до аналізу результатів прогнозу, було обрано конвеєрну (pipeline) архітектуру. Ця архітектура передбачає, що дані проходять через ряд послідовних етапів обробки, де результат одного етапу стає вхідними даними для наступного.

Підсистема складається з чотирьох основних модулів, які реалізують логічні етапи обробки:

Модуль ІАД (IDA Module): Реалізований скриптом код ІАД.txt. Цей модуль працює незалежно від основного конвеєру. Він приймає на вхід вихідний (сирий або попередньо оброблений) CSV файл з даними за певний період і виконує дослідницький аналіз. Результати (статистики, графіки) виводяться користувачеві інтерактивно. Цей модуль використовується для глибокого розуміння даних перед або паралельно з використанням конвеєру прогнозування.

Модуль Фільтрації (Filtering Module): Реалізований скриптом фільтрація.txt. Це перший етап основного конвеєру. Він приймає на вхід вихідний CSV файл (наприклад, pmg_24.csv), виконує очищення (видалення від'ємних значень) та фільтрацію (вибір необхідних стовпців). Результат зберігається у проміжний CSV файл (наприклад, clnd_pmg_24.csv).

Модуль Прогнозування (Forecasting Module): Реалізований скриптом Код прогнозу.txt. Це другий етап конвеєру. Він приймає на вхід відфільтрований CSV файл, виконує агрегацію даних по місяцях, інженерію ознак, тренує моделі XGBoost та генерує прогноз на майбутній період. Результат (прогноз) зберігається у вихідний CSV файл (наприклад, forecast_2024.csv).

Модуль Порівняння (Comparison Module): Реалізований скриптом Код порівняння.txt. Це третій і завершальний етап конвеєру. Він приймає на вхід файл з прогнозом (з попереднього етапу) та файл з фактичними даними за той самий період (наприклад, 5_atr_pmg_24.csv). Модуль об'єднує дані, розраховує метрики точності, генерує візуалізацію порівняння та зберігає текстовий звіт (наприклад, forecast_report.txt).

Таким чином, основний робочий процес (прогнозування та оцінка) має вигляд послідовного запуску скриптів: Фільтрація -> Прогнозування -> Порівняння. Модуль ІАД може використовуватися окремо для дослідження будь-якого з файлів даних (вихідного, відфільтрованого, фактичного).

2.4. Модель даних

Оскільки підсистема працює з даними у форматі CSV та використовує бібліотеку Pandas, основною структурою даних під час обробки є Pandas DataFrame.

Вхідні дані:

Передбачається, що вихідний CSV файл для аналізу та прогнозування містить щонайменше такі стовпці:

- legal_entity_edrrou (Код ЄДРПОУ закладу) - використовується в ІАД.
- period_month (Номер місяця періоду оплати, числовий).
- period_year (Рік періоду оплати, числовий).
- package_id (Ідентифікатор пакету послуг, числовий/категоріальний) - використовується в ІАД.
- pay_package (Сума оплати за пакетом, числова).
- pay_all (Загальна сума оплати за договором у звітному періоді, числова).
- pay_type (Тип оплати, числовий/категоріальний) - використовується в ІАД.
- pay_date (Дата оплати, формат дати).

Проміжні дані:

Після Модуля Фільтрації: CSV файл, що містить лише стовпці: `period_month`, `period_year`, `pay_package`, `pay_all`, `pay_date`. Дані очищені від від'ємних значень.

Внутрішні дані Модуля Прогнозування: `Pandas DataFrame` агрегований по місяцях, з доданими інженерними ознаками (наприклад, `month_index`, `month_sin`, `month_cos`, `lag_1`, `lag_2` тощо для `pay_package` та `pay_all`).

Вихідні дані:

Модуль Прогнозування: CSV файл (напр., `forecast_2024.csv`) зі стовпцями: `period_year`, `period_month`, `predicted_pay_package`, `predicted_pay_all`.

Модуль Порівняння:

- Текстовий файл (`forecast_report.txt`) зі звітом, що містить розраховані метрики MAE, RMSE, MAPE.
- Візуалізація (графік), що показує порівняння `predicted_pay_package` з `actual_pay_package` та `predicted_pay_all` з `actual_pay_all` по місяцях.
- Внутрішній `DataFrame` (`comparison_df`), що об'єднує прогнозні та фактичні значення по місяцях для розрахунків та візуалізації.

Така модель даних забезпечує чіткий потік інформації між модулями підсистеми.

2.5. Опис діаграми потоків даних

Було розроблено та візуалізовано в онлайн редакторі `Mermaid.live`, діаграму потоків даних (`Data Flow Diagram, DFD`) нульового рівня для основного конвеєру підсистеми, для наглядності, діаграму зображено на Рис.1

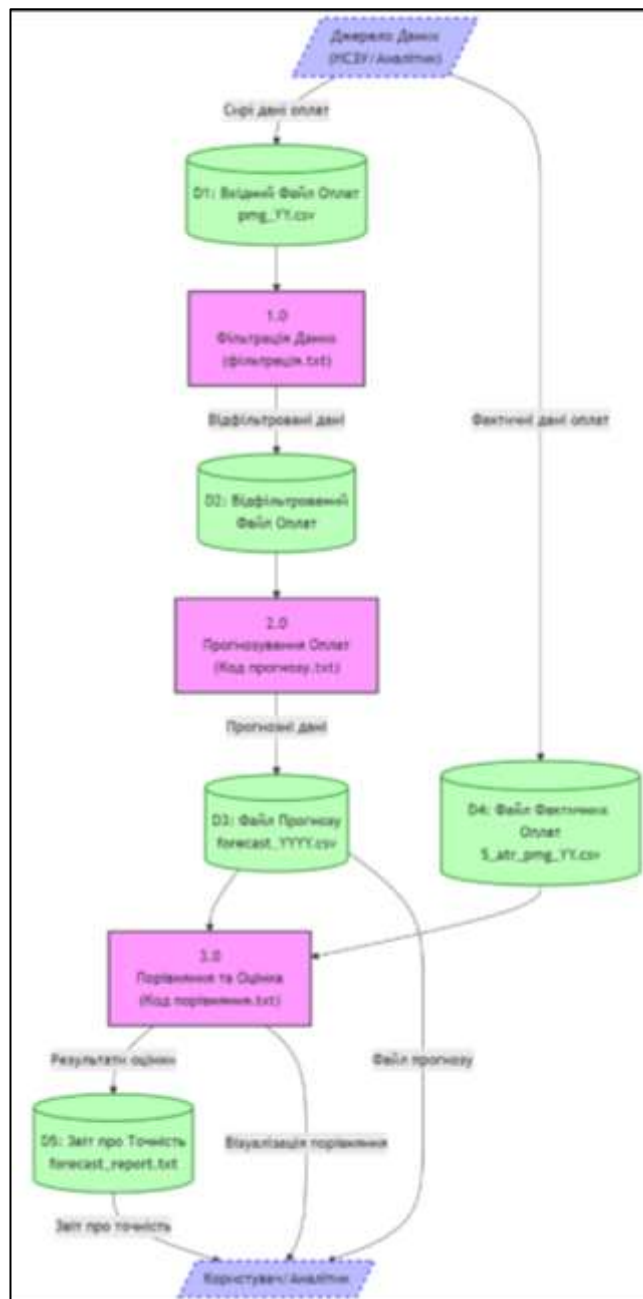


Рисунок 2.1 - DFD0 діаграма підсистеми

Опис діаграми наступний:

Зовнішні сутності:

- Джерело Даних (наприклад, "Сховище даних НСЗУ" або "Аналітик"): Надає вхідні CSV файли (pmg_23.csv, 5_atr_pmg_24.csv).
- Користувач/Аналітик: Отримує кінцеві результати (прогнозований CSV, звіт про точність, візуалізації).

Сховища даних (Файли):

- D1: Вхідний Файл Оплат (наприклад, pmg_23.csv)

- D2: Відфільтрований Файл Оплат (вихід з Процесу 1.0, вхід для Процесу 2.0)
- D3: Файл Прогнозу (вихід з Процесу 2.0, вхід для Процесу 3.0, напр., forecast_2024.csv)
- D4: Файл Фактичних Оплат (наприклад, 5_atr_pmg_24.csv, вхід для Процесу 3.0)
- D5: Звіт про Точність (вихід з Процесу 3.0, напр., forecast_report.txt)

Потоки даних:

- Від "Джерело Даних" до Процесу 1.0: Потік "Сирі дані оплат" (з D1).
- Від Процесу 1.0 до D2: Потік "Відфільтровані дані".
- З D2 до Процесу 2.0: Потік "Відфільтровані дані".
- Від Процесу 2.0 до D3: Потік "Прогнозні дані".
- З D3 до Процесу 3.0: Потік "Прогнозні дані".
- Від "Джерело Даних" до Процесу 3.0: Потік "Фактичні дані оплат" (з D4).
- Від Процесу 3.0 до D5: Потік "Результати оцінки".
- Від Процесу 3.0 до "Користувач/Аналітик": Потік "Візуалізація порівняння" (графік).
- З D3 до "Користувач/Аналітик": Потік "Файл прогнозу".
- З D5 до "Користувач/Аналітик": Потік "Звіт про точність".

Цей опис DFD ілюструє взаємодію між компонентами підсистеми та рух даних через конвеєр обробки, що забезпечує розуміння процесу.

РОЗДІЛ 3. РОЗРОБКА ТА РЕАЛІЗАЦІЯ ІНФОРМАЦІЙНОЇ ПІДСИСТЕМИ

Цей розділ присвячений технічним аспектам створення інформаційної підсистеми. Тут описано інструменти, які були використані, детально розглянуто реалізацію кожного програмного модуля та їх інтеграцію в єдиний робочий процес.

3.1. Інструментальні засоби розробки

Для розробки інформаційної підсистеми, використовувалося середовище Google Colaboratory, що дозволило легко інтегрувати завантаження файлів та візуалізацію, було обрано мову програмування Python завдяки її широким можливостям для аналізу даних, наявності великої кількості спеціалізованих бібліотек та відносній простоті синтаксису. Було використано наступний набір бібліотек:

Pandas: Основна бібліотека для роботи з даними. Використовувалася для читання CSV файлів, маніпуляції даними (фільтрація, агрегація, об'єднання), створення та модифікації DataFrame.

NumPy: Фундаментальна бібліотека для наукових обчислень в Python. Використовувалася для ефективних операцій з масивами, математичних функцій (наприклад, `sqrt` для RMSE, `sin`, `cos` для інженерії ознак).

Matplotlib: Базова бібліотека для створення статичних, анімованих та інтерактивних візуалізацій. Використовувалася для побудови графіків (лінійних, стовпчикових, діаграм розсіювання, коробкових діаграм) у всіх модулях, де потрібна візуалізація.

Seaborn: Надбудова над Matplotlib, яка надає більш високорівневий інтерфейс для створення привабливих та інформативних статистичних графіків. Використовувалася в Модулі ІАД для покращення візуалізацій.

Scikit-learn: Популярна бібліотека для машинного навчання. У даному проєкті використовувалася переважно для розрахунку метрик якості моделі (`mean_absolute_error`, `mean_squared_error`, `mean_absolute_percentage_error` – хоча MAPE було реалізовано окремою функцією в коді порівняння) та потенційно для

інших допоміжних функцій (наприклад, розділення даних, хоча в прогнозі воно не використовувалось явно для фінального тренування).

XGBoost: Високоєфективна реалізація алгоритму градієнтного бустингу. Використовувалася як основний інструмент для побудови регресійних моделей прогнозування в Модулі Прогнозування (XGBRegressor).

Google Colab files module: (Специфічно для Модуля ІАД) Використовувався для інтерактивного завантаження файлу даних у середовищі Google Colaboratory.

io: Стандартний модуль Python, використаний в Модулі ІАД для читання завантаженого файлу з пам'яті.

warnings: Стандартний модуль для керування попередженнями (наприклад, для їх ігнорування при відомих безпечних операціях).

sys: Стандартний модуль для взаємодії з системними параметрами та функціями (наприклад, для завершення скрипта `sys.exit()` у Модулі Порівняння при помилках).

3.2. Реалізація модуля інтелектуального аналізу даних (ІАД)

Модуль (ІАД.ipynb) був розроблений як інструмент для первинного дослідження даних про оплати за ПМГ за певний період (рік). Його основна мета – отримати загальне уявлення про дані, їхню структуру, розподіли та виявити цікаві особливості або потенційні проблеми. Скрипт виконує такі кроки:

Завантаження даних: Використовується модуль `google.colab.files` для інтерактивного завантаження CSV файлу користувачем у середовищі Google Colab. Після завантаження файл зчитується у Pandas DataFrame за допомогою `pd.read_csv` та `io.BytesIO`.

Первинний огляд даних: Виводяться перші 5 рядків DataFrame (`df.head()`) для візуальної перевірки структури та вмісту даних. Використовується метод `df.info()` для отримання інформації про типи даних у кожному стовпці та наявність пропущених значень.

Дослідницький аналіз (EDA) та Візуалізація: Цей етап є основним у модулі.

Описова статистика: Розраховуються та виводяться основні статистичні показники (кількість, середнє, стандартне відхилення, мінімум, максимум,

квантилі) для всіх числових стовпців за допомогою `df.describe()`. Наприклад, для даних за 2024 рік було отримано середню загальну оплату (`pay_all`) близько 3.76 млн грн, але з дуже високим стандартним відхиленням (7.39 млн грн), що вказує на великий розкид значень. Також видно наявність від'ємних значень у `pay_package` та `pay_all`, що потребує очищення на наступних етапах.

Аналіз розподілів: Будуються гістограми та/або діаграми щільності для ключових числових змінних, таких як `pay_all`, щоб візуально оцінити їх розподіл. Як видно з текстового виводу ("Аналіз розподілу сум оплат"), цей розподіл, ймовірно, є сильно скошеним вправо (багато невеликих оплат і невелика кількість дуже великих), що відображено на відповідній візуалізації (Рис. 3.1).

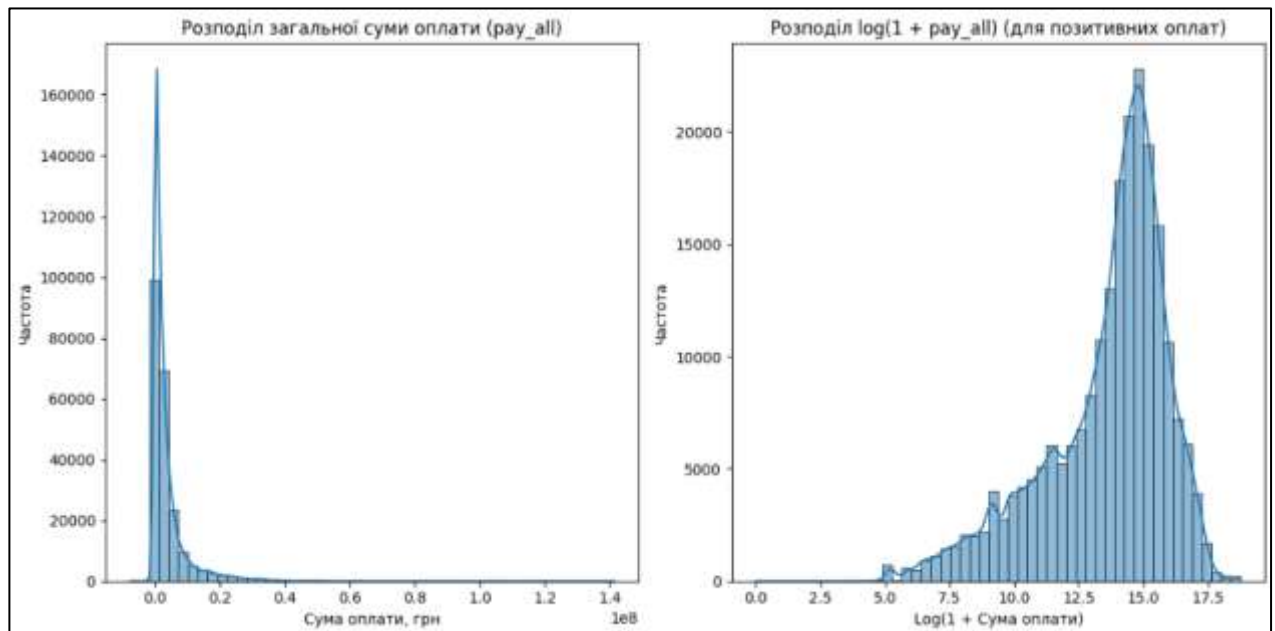


Рисунок 3.1 - Аналіз розподілів

Аналіз категоріальних змінних: Досліджується розподіл даних за категоріями, наприклад, за пакетами послуг (`package_id`) та типами оплат (`pay_type`). Будуються стовпчикові діаграми, що показують частоту або сумарні/середні значення для кожної категорії. Наприклад, аналізуються Топ-20 пакетів за кількістю або сумою оплат (Рис. 3.2,3.3) та розподіл різних типів оплат (Рис. 3.4).

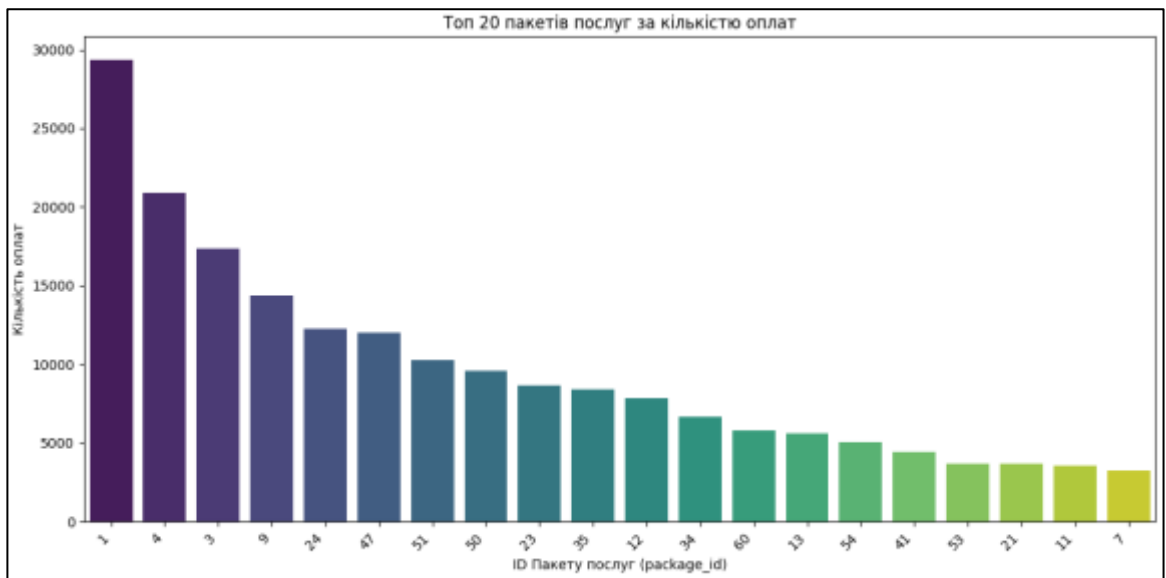


Рисунок 3.2 – Топ 20 пакетів послуг за кількістю оплат

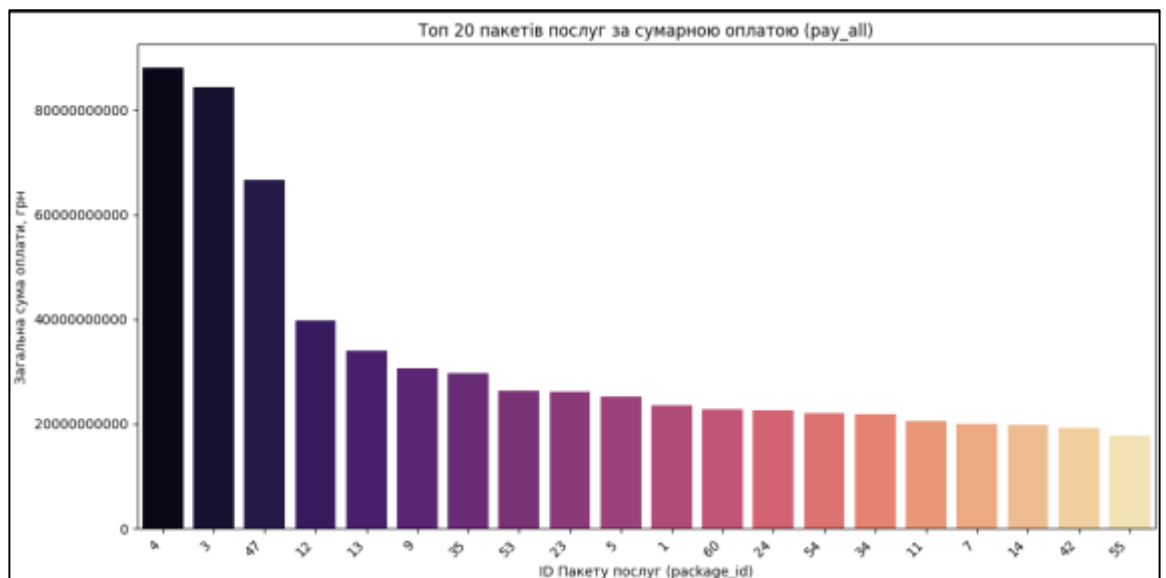


Рисунок 1.3 –Топ 20 пакетів послуг за сумарною оплатою

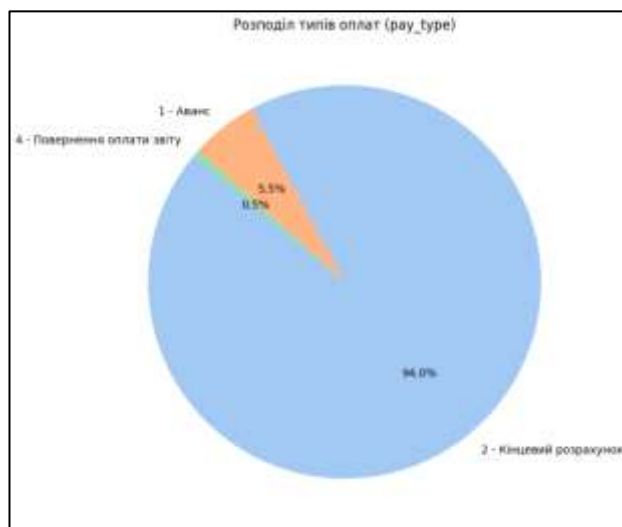


Рисунок 3.4 - Розподіл типів оплат

Аналіз динаміки: Якщо дані охоплюють певний часовий проміжок з достатньою деталізацією (наприклад, по днях чи місяцях), будуються графіки часових рядів для візуалізації змін показників у часі. Наприклад, аналізується загальна сума оплат (pay_all) по місяцях (Рис 3.5).

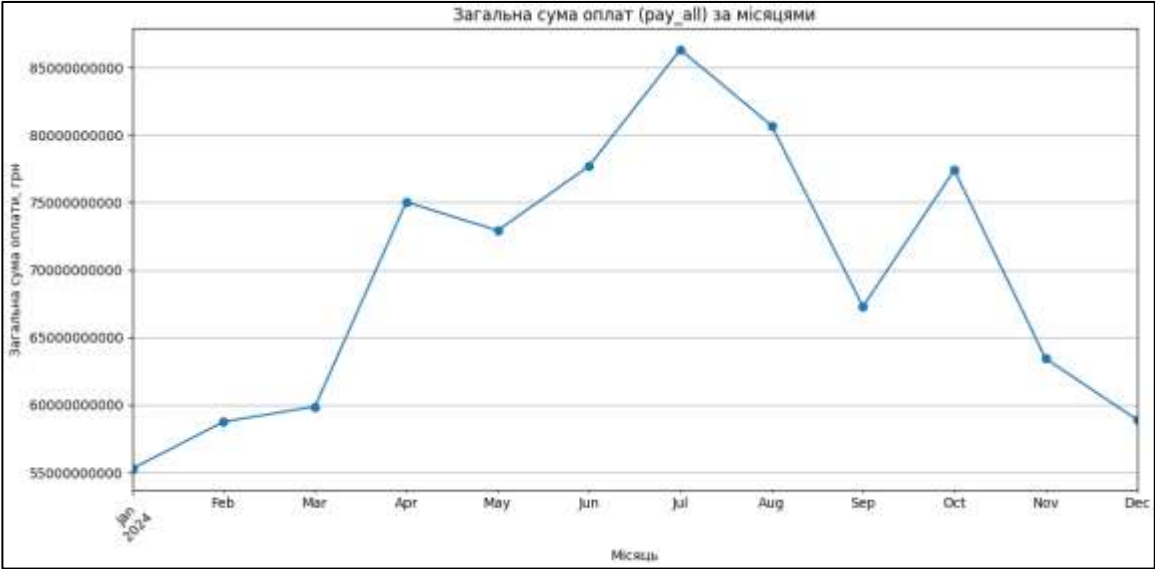


Рисунок 2.5 – Загальна сума оплат за місяцями

Аналіз взаємозв'язків: Будуються діаграми розсіювання (scatter plots) для візуальної оцінки зв'язку між двома числовими змінними, наприклад, між оплатою за пакетом (pay_package) та загальною оплатою (pay_all) (Рис 3.6).

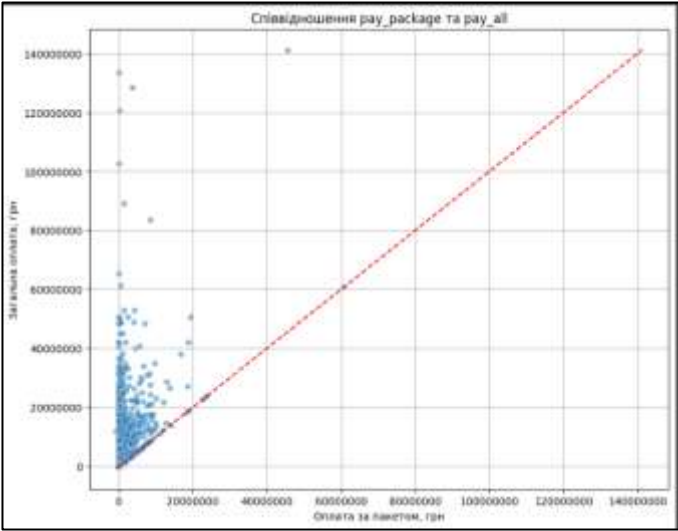


Рисунок 3.6 – Співвідношення сплат пакетів та повних оплат

Розраховується коефіцієнт кореляції Пірсона ($df['pay_package'].corr(df['pay_all'])$) для кількісної оцінки лінійного зв'язку. Для даних 2024 року було отримано помірну позитивну кореляцію 0.3918.

Приклади Інтелектуального Аналізу Даних:

Пошук аномалій: Реалізовано пошук потенційно аномально великих позитивних значень `pay_all` за допомогою методу IQR. Розраховується верхня межа ($Q3 + 1.5 * IQR$). Наприклад, для даних 2024 року ця межа склала 9,643,470.63 грн, і було знайдено 21,253 потенційних викиди вище цієї межі. Виводяться приклади таких аномальних записів.

Аналіз без аномалій: Після виявлення аномалій може бути корисно повторити деякі аналізи (наприклад, діаграму розсіювання `pay_package` vs `pay_all`) на даних без викидів, щоб краще побачити основну тенденцію. У скрипті передбачено побудову такої діаграми з розрізненням за типом оплати (`pay_type`) (Рисунок 1, секція 7). Перерахована кореляція без позитивних аномалій (0.3882) не сильно змінилася, що може свідчити про те, що викиди не спотворювали загальну лінійну залежність кардинально.

Групування: Розраховуються агреговані показники для різних груп. Наприклад, розраховується та візуалізується середня сума оплати (`pay_all`) за кожним пакетом послуг (`package_id`), виводяться Топ-10 пакетів за цим показником (Рисунок 1, секція 8).

Модуль ІАД надає важливу інформацію для розуміння характеру даних та може допомогти у формулюванні гіпотез для подальшого, більш глибокого аналізу або моделювання.

3.3. Реалізація модуля фільтрації даних

Цей модуль виконує підготовку даних для подальшого використання в модулі прогнозування. Він складається з двох основних функцій, які викликаються послідовно:

`remove_negative_rows(input_file, output_file)`: Ця функція читає вхідний CSV файл (`pmg_22,23.csv`), знаходить усі числові стовпці та видаляє ті рядки, де хоча б в одному числовому стовпці міститься від'ємне значення. Це важливий крок очищення, оскільки від'ємні суми оплат є некоректними в контексті аналізу загальних витрат. Очищений `DataFrame` зберігається у проміжний файл (`clnd_pmg_22,23.csv`).

`filter_dataset(input_file, output_file)`: Ця функція читає файл, що пройшов попереднє очищення (`clnd_pmg_22,23.csv`), та відбирає лише ті стовпці, які є необхідними для побудови прогнозної моделі часових рядів: `period_month`, `period_year`, `pay_package`, `pay_all`, `pay_date`. Відкидання зайвих стовпців зменшує обсяг даних та спрощує подальшу обробку. Результат зберігається у фінальний відфільтрований файл (наприклад, знову `output_csv = clnd_pmg_22,23.csv` у наданому коді, але логічно це мав би бути окремий файл, скажімо `filtered_data.csv`, який використовується на вході прогнозу).

Модуль виводить інформаційні повідомлення про збереження файлів та кількість рядків і залишені стовпці у відфільтрованому датасеті. Цей модуль є обов'язковим першим кроком у конвеєрі прогнозування.

3.4. Реалізація модуля прогнозування

Модуль прогнозу є ядром підсистеми, оскільки саме тут відбувається побудова та застосування моделі машинного навчання для прогнозування майбутніх оплат. Робота модуля складається з таких етапів:

Завантаження та підготовка даних (`load_and_prepare_data`):

Функція завантажує відфільтровані дані з CSV файлу, отриманого на попередньому етапі.

Виконується перетворення стовпця `pay_date` у формат `datetime`.

Дані агрегуються на місячному рівні. Розраховуються сумарні `pay_package` та `pay_all` для кожного унікального поєднання `period_year` та `period_month`.

Інженерія ознак (Feature Engineering): Створюються нові стовпці (ознаки), які допоможуть моделі вловити часові залежності:

- `month_index`: Порядковий номер місяця від початку даних.
- `month`: Номер місяця (1-12).
- `year`: Рік.
- `month_sin`, `month_cos`: Синус та косинус від номера місяця, помноженого на $2\pi/12$. Це допомагає моделі зрозуміти циклічність (сезонність) року.
- Лагові ознаки (`lag_1`, `lag_2`, ...): Значення цільової змінної (окремо для `pay_package` та `pay_all`) за попередні місяці. Наприклад, `lag_1` - це

значення за попередній місяць, `lag_12` - значення рік тому. Лаги допомагають моделі враховувати автокореляцію та тренди. Кількість лагів є параметром, який можна налаштовувати.

Тренування моделей (`train_xgboost_model`):

Оскільки потрібно прогнозувати два показники (`pay_package` та `pay_all`), тренуються дві окремі регресійні моделі XGBoost (`XGBRegressor`). Одна модель навчається прогнозувати `pay_package` на основі створених ознак, інша – `pay_all`.

Визначається набір ознак (`features`) для кожної моделі (зазвичай це всі створені часові та лагові ознаки).

Моделі навчаються на всіх наявних історичних даних (`model.fit(X_train, y_train)`).

У результаті отримано метрики MAPE та RMSE, розраховані на тренувальних даних. MAPE = 0.0000 та дуже низькі RMSE (1381.08 для `pay_package` та 6198.51 для `pay_all`) свідчать про те, що моделі дуже добре "запам'ятали" тренувальні дані. Однак, це не є показником реальної точності прогнозування на нових, раніше не бачених даних. Реальна точність буде оцінена в Модулі Порівняння.

Генерація прогнозу (`forecast_future_months`):

Прогнозування на майбутні періоди (наприклад, 12 місяців наступного року) відбувається ітеративно.

Для прогнозу першого майбутнього місяця використовуються останні відомі лагові значення та розраховані часові ознаки для цього місяця.

Прогноз для першого місяця використовується для розрахунку лагових значень, необхідних для прогнозу другого місяця, і так далі.

Генерується `DataFrame` (`forecast_df`) з прогнозованими значеннями `predicted_pay_package` та `predicted_pay_all` для кожного місяця прогнозного періоду.

Результати прогнозу для 2024 року (згідно з виводом) показують очікувані місячні суми, наприклад, близько 11.5 млрд грн для `pay_package` та 52-64 млрд грн для `pay_all`.

Збереження та Візуалізація:

Згенерований DataFrame прогнозу зберігається у CSV файл (forecast_2024.csv).

Будується графік (Рис. 3.7), який візуалізує історичні дані та згенерований прогноз для обох показників, що дозволяє наочно оцінити продовження трендів та сезонності моделлю.

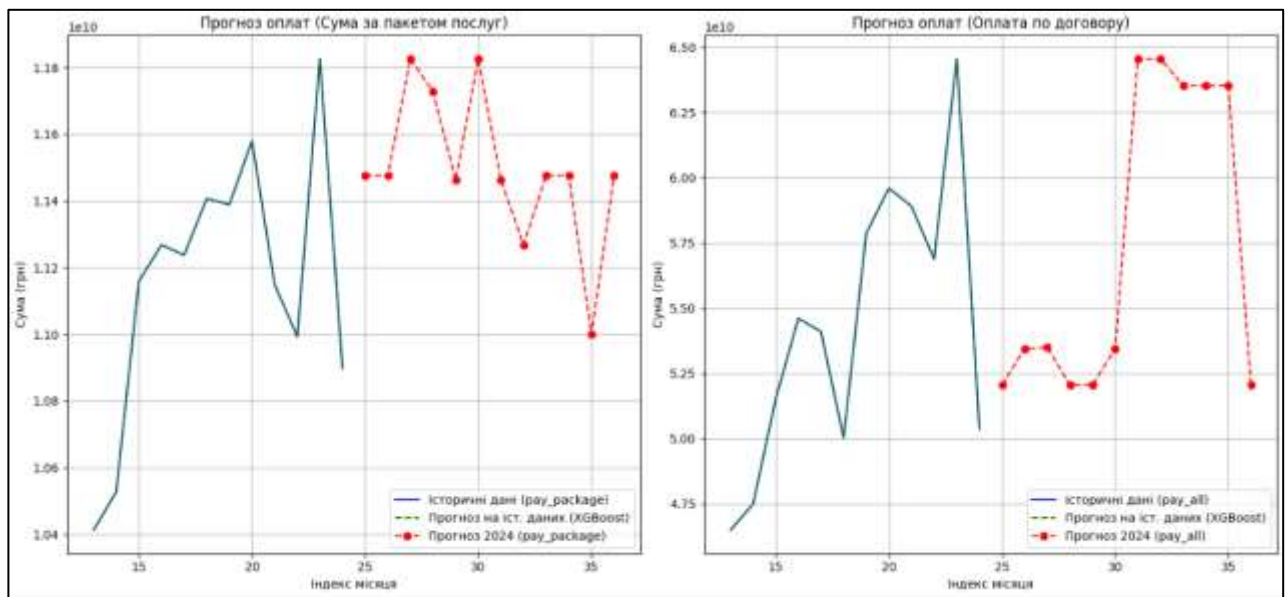


Рисунок 3.7 – Візуалізація прогнозу даних

Цей модуль реалізує основну інтелектуальну частину підсистеми, застосовуючи машинне навчання для вирішення задачі прогнозування.

3.5. Реалізація модуля порівняння та оцінки

Завершальний модуль конвеєру скрипт порівняння призначений для оцінки якості прогнозу, отриманого на попередньому етапі. Він виконує такі дії:

Завантаження даних (load_data):

Завантажується CSV файл з прогнозом (forecast_2024.csv), створений Модулем Прогнозування.

Завантажується CSV файл з фактичними даними про оплати за той самий період, на який робився прогноз (наприклад, 5_atr_pmg_24.csv). Важливо, щоб цей файл містив ті ж самі ключові стовпці, що й дані, які використовувалися для фільтрації та тренування (period_month, period_year, pay_package, pay_all).

Фактичні дані агрегуються по місяцях (розраховуються сумарні `actual_pay_package` та `actual_pay_all` для кожного місяця).

Об'єднання даних (`merge_and_compare`):

Два `DataFrame` (прогнознний та агрегований фактичний) об'єднуються за роком та місяцем.

Розраховуються абсолютна (`diff_package`, `diff_all`) та відсоткова (`pe_package`, `pe_all`) похибки для кожного місяця.

Виводиться приклад об'єднаних даних з розрахованими похибками, що дозволяє побачити відхилення прогнозу від факту для кожного місяця. Наприклад, для січня 2024 року `pay_package` було недопрогнозовано на 7.77%, а `pay_all` - на 5.82%.

Розрахунок метрик точності (`calculate_metrics`):

На основі об'єднаних даних розраховуються загальні метрики якості прогнозу за весь період:

MAPE: Середня абсолютна відсоткова похибка. Для прогнозу на 2024 рік вона склала 8.67% для `pay_package` та 16.27% для `pay_all`. Це означає, що в середньому прогноз відхилявся від факту на ці відсотки.

Як правило, прогнозування `pay_all` виявляється складнішим завданням (з вищою відсотковою помилкою), що може бути зумовлено кількома факторами:

1. Ефект Агрегації: `pay_all` є сукупним показником, що об'єднує оплати за різними пакетами послуг в рамках одного договору за певний період. Хоча це може згладжувати випадкові коливання окремих пакетів, водночас акумулюється різноспрямована динаміка та "шум" від багатьох джерел, що ускладнює моделювання єдиної тенденції.

2. Включення Різних Типів Оплат: Показник `pay_all` може включати не лише оплати за надані послуги в рамках пакетів, але й інші види платежів (наприклад, оплату капітаційної ставки, платежі за виконання специфічних умов договору, оплату кредиторської заборгованості [source: 344]), які можуть мати власну, відмінну від пакетних оплат, логіку та передбачуваність.

3. Вплив Зовнішніх Факторів: Загальна сума оплати може бути більш чутливою до адміністративних рішень, змін у бюджетному процесі НСЗУ або специфіки окремих договорів, які важко врахувати лише на основі історичних часових рядів самих оплат. Оплата за конкретним пакетом (`pay_package`) може бути тісніше пов'язана з обсягом наданих послуг, який може мати більш стабільну динаміку.

MAE: Середня абсолютна похибка (в гривнях). Склала близько 1.11 млрд грн для `pay_package` та 12.02 млрд грн для `pay_all`.

Також розраховується RMSE, яка сильніше штрафує за великі помилки.

Візуалізація результатів (`visualize_comparison`):

Будується лінійний графік (Рисунок 3), на якому одночасно відображаються прогнозовані та фактичні значення (`pay_package` та `pay_all`) по місяцях. Це дозволяє візуально оцінити, наскільки добре модель вловила динаміку реальних оплат.

Генерація звіту (`generate_report`):

Створюється текстовий файл (`forecast_report.txt`), куди записуються розраховані метрики точності (MAPE, MAE, RMSE) та, можливо, інша релевантна інформація про процес порівняння.

Цей модуль надає об'єктивну оцінку ефективності розробленої прогнозної моделі та завершує аналітичний цикл.

РОЗДІЛ 4. АНАЛІЗ РЕЗУЛЬТАТІВ РОБОТИ ТА ВИСНОВКИ

У цьому розділі проводиться аналіз результатів, отриманих за допомогою розробленої інформаційної підсистеми, оцінюється точність прогнозування та обговорюються можливості практичного застосування і подальшого розвитку підсистеми.

4.1. Аналіз результатів дослідного аналізу даних

Модуль інтелектуального аналізу даних (ІАД) був застосований до набору даних про оплати за ПМГ за 2024 рік. Аналіз дозволив зробити низку спостережень щодо характеру цих даних.

По-перше, описова статистика виявила значний діапазон сум оплат. Середня загальна оплата (pay_all) становила близько 3.76 млн грн, проте максимальне значення сягало понад 141 млн грн, а стандартне відхилення було дуже великим (7.39 млн грн). Це свідчить про високу варіативність фінансових потоків та наявність як великої кількості відносно дрібних транзакцій, так і значних платежів окремим закладам або за окремими пакетами. Наявність від'ємних значень підтвердила необхідність етапу очищення даних.

По-друге, аналіз розподілу загальної суми оплати (pay_all), ймовірно, показав (як очікується для фінансових даних) сильно асиметричний розподіл з довгим "хвостом" у бік великих значень, як зображено на рис 3.1. Це означає, що більшість оплат є відносно невеликими, але існують окремі дуже великі платежі, які суттєво впливають на середнє значення.

По-третє, аналіз за пакетами послуг (Рис. 3.2, 3.3) дозволив ідентифікувати пакети, за якими здійснюються найбільші або найчастіші оплати, а також пакети з найвищою середньою сумою оплати. Наприклад, було виявлено Топ-10 пакетів з найвищою середньою сумою pay_all, серед яких пакети з ID 2, 64, 37 мали середні оплати понад 20 млн грн. Ця інформація може бути корисною для розуміння структури витрат ПМГ.

По-четверте, аналіз типів оплат (Рис 3.4) показав їх розподіл, що може вказувати на переважання певних типів розрахунків (наприклад, кінцевих розрахунків над авансами).

По-п'яте, аналіз динаміки (Рис 3.5 дозволив візуалізувати зміни загальних сум оплат протягом року, виявляючи можливі сезонні коливання або тренди.

По-шосте, аналіз взаємозв'язку між оплатою за пакетом (pay_package) та загальною оплатою (pay_all) показав помірну позитивну кореляцію (близько 0.39), як видно з діаграми розсіювання (Рис 3.6) та розрахованого коефіцієнта Пірсона. Це очікувано, але кореляція не є дуже високою, що може свідчити про вплив інших факторів на загальну суму оплати.

По-сьоме, виявлення аномалій за методом IQR ідентифікувало значну кількість (понад 21 тис.) потенційно аномально великих позитивних оплат pay_all (вище 9.6 млн грн). Хоча не кожна така оплата обов'язково є помилковою (це можуть бути великі контракти або специфічні пакети послуг), їх наявність та кількість вказують на необхідність уважного контролю за великими транзакціями та потенційну корисність інструментів виявлення аномалій для НСЗУ. Аналіз діаграми розсіювання без цих аномалій (Рис 3.7) не сильно змінив загальну картину кореляції.

Загалом, результати ІАД підтвердили складність та багатогранність даних про оплати за ПМГ і надали цінну вихідну інформацію для подальшого моделювання та аналізу.

4.2. Аналіз точності прогнозування

Ключовим результатом роботи підсистеми є прогноз майбутніх оплат та оцінка його точності. Модуль прогнозування згенерував місячний прогноз для pay_package та pay_all на 2024 рік, використовуючи модель XGBoost, навчену на історичних даних (імовірно, за 2022-2023 роки, хоча це не вказано явно, але передбачено логікою "прогноз на основі двох попередніх років"). Модуль порівняння зіставив цей прогноз з фактичними даними за 2024 рік.

Основні метрики точності, розраховані Модулем Порівняння:

MAPE (Середня абсолютна відсоткова похибка):

- Для pay_package: 8.67%
- Для pay_all: 16.27%

MAE (Середня абсолютна похибка):

- Для `pay_package`: 1.11 млрд грн
- Для `pay_all`: 12.02 млрд грн

Значення `MARE` показують, що прогноз оплати за конкретним пакетом (`pay_package`) виявився досить точним (похибка менше 10%), що є хорошим результатом для фінансових часових рядів такого масштабу. Прогноз загальної суми оплати (`pay_all`) мав вищу відсоткову похибку (16.27%), що може бути пов'язано з більшою складністю цього показника, який агрегує різні типи платежів та пакети. Абсолютні похибки (`MAE`) є значними в грошовому вимірі, що очікувано, враховуючи мільярдні обсяги щомісячних оплат.

Візуальний аналіз графіків порівняння (Рис 4.1) дозволяє побачити, наскільки добре прогнозна модель відтворила динаміку фактичних даних. Ймовірно, графік показує, що модель змогла вловити загальний тренд та певну сезонність, однак можуть бути помітні місяці, де відхилення прогнозу від факту було суттєвим. Наприклад, аналіз таблиці порівняння у виводі показує, що у квітні та травні 2024 року відсоткова похибка для `pay_all` була особливо великою (близько 30% та 28% відповідно), що вказує на непередбачені моделью фактори або події, які вплинули на оплати в ці місяці.

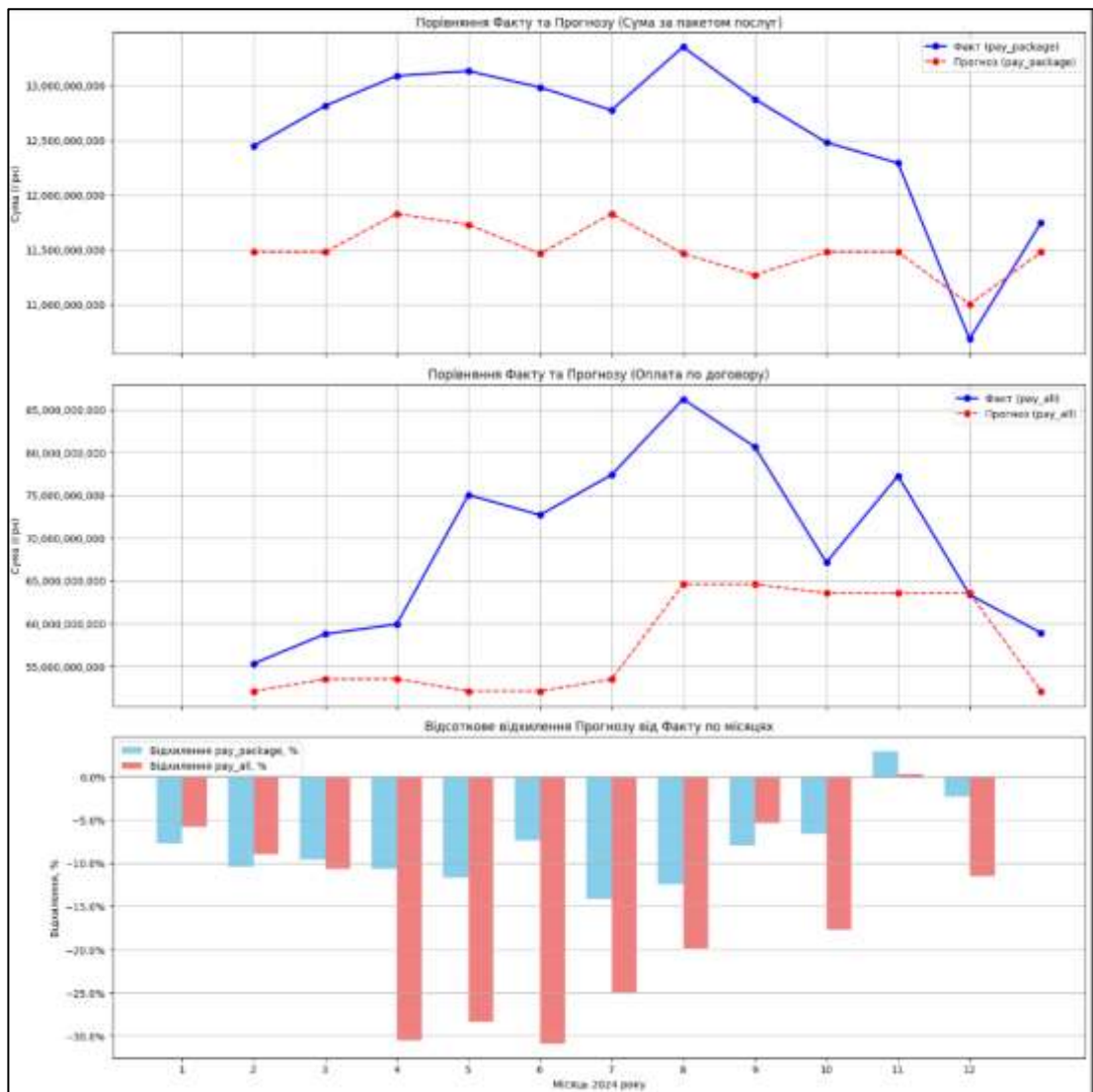


Рисунок 3.1 – Візуалізація порівняння прогнозу з актуальними даними

Загалом, точність прогнозу можна вважати задовільною, особливо для показника `rau_package`. Результат для `rau_all` вказує на потенційні шляхи покращення моделі, можливо, шляхом додавання зовнішніх економічних факторів, більш складної інженерії ознак або використання інших алгоритмів прогнозування. Важливо також пам'ятати, що точність прогнозу може залежати від стабільності фінансових процесів та відсутності значних непередбачуваних змін у системі фінансування, наприклад, змін у законодавстві, тарифній політиці, виникнення кризових ситуацій.

4.3. Можливості впровадження та розвитку підсистеми

Розроблена інформаційна підсистема, незважаючи на її реалізацію у вигляді набору скриптів, має потенціал для практичного застосування та подальшого розвитку.

Можливості впровадження:

Аналітичний інструмент для НСЗУ: Фахівці НСЗУ можуть використовувати підсистему для регулярного аналізу даних про оплати, моніторингу фінансових потоків, виявлення аномалій та, що найважливіше, для коротко- та середньострокового прогнозування бюджетних потреб ПМГ. Це може допомогти у більш ефективному плануванні та розподілі коштів.

Інструмент для ЗОЗ: Керівники та фінансові служби закладів охорони здоров'я, які працюють за договорами з НСЗУ, могли б використовувати подібний інструмент для аналізу власних надходжень та прогнозування майбутніх доходів від НСЗУ, що сприятиме кращому фінансовому плануванню на рівні закладу.

Основа для досліджень: Підсистема може слугувати інструментом для дослідників та аналітиків, які вивчають систему фінансування охорони здоров'я в Україні, дозволяючи їм проводити власний аналіз на основі відкритих даних НСЗУ.

Універсальність підсистеми, закладена в архітектурі, особливо в модулях прогнозування та порівняння, дозволяє використовувати її для аналізу даних за різні роки. Достатньо підготувати вхідні файли з даними за відповідні періоди, наприклад, дані за 2023-2024 роки для прогнозу на 2025 рік, та запустити конвеєр скриптів.

ВИСНОВОК

У ході виконання кваліфікаційної роботи було успішно досягнуто поставленої мети – розроблено інформаційну підсистему для аналізу та прогнозування оплат закладам охорони здоров'я за Програмою медичних гарантій на основі інтелектуального аналізу даних.

Було проведено дослідження предметної області, проаналізовано актуальність задачі в контексті реформи фінансування охорони здоров'я України та діяльності НСЗУ. Визначено основні методи інтелектуального аналізу даних, релевантні для вирішення поставлених завдань: дослідницький аналіз даних, виявлення аномалій, аналіз часових рядів та прогнозування за допомогою регресійних моделей (XGBoost), оцінка точності моделей.

Спроектовано архітектуру підсистеми у вигляді конвеєру обробки даних, що складається з чотирьох основних програмних модулів: Модуль ІАД (для дослідження даних), Модуль Фільтрації (для очищення та підготовки даних), Модуль Прогнозування (для побудови моделі та генерації прогнозу) та Модуль Порівняння (для оцінки точності прогнозу).

Розроблено програмну реалізацію всіх модулів підсистеми на мові Python з використанням сучасних бібліотек для аналізу даних та машинного навчання (Pandas, NumPy, Matplotlib, Scikit-learn, XGBoost). Реалізовано конвеєр послідовної обробки даних від сирого файлу до отримання прогнозу та звіту про його якість.

Проведено апробацію роботи підсистеми на реальних даних про оплати за ПМГ. Результати дослідного аналізу даних виявили ключові характеристики фінансових потоків, наявність аномалій та закономірностей у розподілі оплат. Побудована прогнозна модель XGBoost продемонструвала задовільну точність прогнозування місячних обсягів оплат, особливо для показника `pay_package` (MAPE 8.67%). Точність прогнозу для `pay_all` була дещо нижчою (MAPE 16.27%), що вказує на потенціал для подальшого вдосконалення моделі.

Розроблена підсистема є універсальним інструментом, який може бути використаний для аналізу даних за різні періоди та прогнозування майбутніх

витрат, що становить практичну цінність для НСЗУ, ЗОЗ та дослідників. Визначено конкретні напрямки для подальшого розвитку підсистеми, включаючи створення графічного інтерфейсу, автоматизацію, розширення аналітичного та прогнозного функціоналу.

Таким чином, завдання кваліфікаційної роботи виконано в повному обсязі. Створено працюючу інформаційну підсистему, яка демонструє ефективність застосування методів інтелектуального аналізу даних для аналізу та прогнозування фінансових показників Програми медичних гарантій України.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Методичні вказівки з підготовки та написання кваліфікаційної роботи бакалавра (для студентів освітнього рівня «бакалавр» спеціальності 122 «Комп'ютерні науки»). – Умань: Редакційно-видавничий відділ УНУС. – 2019. – 44 с.
2. Єдиний державний веб-портал відкритих даних. URL: <https://data.gov.ua/>
3. XGBoost documentation. URL: <https://xgboost.readthedocs.io/en/stable/>
4. Національна служба здоров'я України : офіційний вебсайт. URL: <https://nszu.gov.ua/>
5. Міністерство охорони здоров'я України : офіційний вебсайт. URL: <https://moz.gov.ua/uk>
6. Pandas documentation. URL: <https://pandas.pydata.org/docs/>
7. Scikit-learn documentation. URL: <https://scikit-learn.org/stable/documentation.html>
8. Seaborn documentation. URL: <https://seaborn.pydata.org/api.html>
9. Ланде Д.В., Субач І.Ю., Бояринова Ю.Є. Основи теорії і практики інтелектуального аналізу даних у сфері кібербезпеки: навчальний посібник. — К.: ІСЗІ КПІ ім. Ігоря Сікорського», 2018. — 297 с.
10. Міністерство охорони здоров'я України : офіційний вебсайт. URL: <https://moz.gov.ua/>
11. NumPy documentation. URL: <https://numpy.org/doc/>
12. Matplotlib documentation. URL: <https://matplotlib.org/stable/contents.html>
13. Лупан І.В. Інтелектуальний аналіз даних Data Mining: навчально-методичний посібник. – Кропивницький, ФОП Піскова М. А., 2022. – 112 с.
14. Hromady.org. (2024, Лютий 6). Про дані для управління здоров'ям та інформаційні панелі НСЗУ для ОМС: стан, розвиток та потреби. URL: <https://hromady.org/pro-dani-dlya-upravlinnya-zdorovyam-ta-informacijni-paneli-nszu-dlya-oms-stan-rozvitok-ta-potrebi/>
15. Варенко В.М. Інформаційно-аналітична діяльність: Навч. посіб. / В. М. Варенко. – К.: Університет «Україна», 2014. – 417 с.

16. Інформатор НСЗУ. URL: <https://service.e-health.gov.ua/gromadyanam>
17. ПОСІБНИК ПРОГРАМИ МЕДИЧНИХ ГАРАНТІЙ 2024: ВІДНОВЛЕННЯ І РОЗВИТОК. URL: https://vaccine.org.ua/wp-content/uploads/2024/06/posibnyk_pmg_2024.pdf
18. WHO. (2023). Виконання бюджету у сфері охорони здоров'я: Тематичне дослідження. URL: <https://iris.who.int/bitstream/handle/10665/369795/WHO-EURO-2023-7195-46961-69547-ukr.pdf?sequence=1&isAllowed=y>
19. Грицишин, Л.М. (2022). НОРМАТИВНО-ПРАВОВЕ ЗАБЕЗПЕЧЕННЯ РЕФОРМУВАННЯ ОХОРОНИ ЗДОРОВ'Я В УКРАЇНІ В УМОВАХ ДІЇ ВОЄННОГО СТАНУ. URL: http://dspace.wunu.edu.ua/bitstream/316497/44612/1/%D0%93%D1%80%D0%B8%D1%86%D0%B8%D1%88%D0%B8%D0%BD_%D0%9B_%D0%9C%D0%97%D0%9E%D0%97%D0%B7%D0%BC%D1%96-21_original_15022022_111859.doc
20. Рахункова палата. (2021). Звіт про результати аудиту ефективності використання коштів державного бюджету, спрямованих Національній службі здоров'я України на реалізацію програми державних гарантій медичного обслуговування населення. URL: https://rp.gov.ua/upload-files/Activity/Collegium/2021/11-3_2021/14692.pdf
21. Gu, Y., & Li, H. (2022). Research on medical cost prediction for basic medical insurance based on machine learning. Scientific Reports. URL: <https://www.nature.com/articles/s41598-022-24326-y>
22. Al-Gwaiz, R., et al. (2023). "A Machine Learning Approach for Predicting Health Insurance Claim Rejection." *Healthcare*. URL: <https://www.mdpi.com/2227-9032/11/17/2372>
23. Chen, T., & Guestrin, C. (2016). "XGBoost: A Scalable Tree Boosting System." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. URL: <https://dl.acm.org/doi/10.1145/2939672.2939785>
24. Ahmad, S., et al. (2021). "Time Series Forecasting of Daily Hospital Visits Using Data-Driven Models." *IEEE Access*. URL: <https://ieeexplore.ieee.org/document/9360815>

25. Yunpu Wu; Peixuan Yan; Paul Allen (2024) "Enhancing Fault Detection in High-Speed Train Dampers with Auxiliary Task Learning" URL: <https://ieeexplore.ieee.org/abstract/document/10843880>
26. Otexts (Hyndman, R.J., & Athanasopoulos, G.). Forecasting: Principles and Practice (online textbook). URL: <https://otexts.com/fpp3/>
27. Weiskopf, N. G., & Weng, C. (2013). Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. URL (PubMed Central - відкритий дотин): <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3638192/>