

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

**НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ
«ІНСТИТУТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ В
ЕКОНОМІЦІ»**

Кафедра системного аналізу та кібербезпеки

ОСВІТНЬО-ПРОФЕСІЙНА ПРОГРАМА «СИСТЕМНИЙ АНАЛІЗ»

Галузь знань 12 «Інформаційні технології»

Спеціальність 124 «Системний аналіз»

Форма навчання: очна (денна)

КВАЛІФІКАЦІЙНА БАКАЛАВРСЬКА РОБОТА

на тему «Інтелектуальний аналіз кредитних даних банку»

здобувача Штирхун Христини Олександрівни

Науковий керівник д.е.н., доцент Камінський Олег Євгенович

(підпис)

Робота допущена до захисту перед екзаменаційною комісією

з атестації здобувачів вищої освіти (ЕК)

Завідувач кафедри: д.ф.-м.н., проф. Джалладова І.А.

(підпис)

Київ 2026

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

**НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ
«ІНСТИТУТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ В
ЕКОНОМІЦІ»**

Кафедра системного аналізу та кібербезпеки

**ОСВІТНЬО-ПРОФЕСІЙНА ПРОГРАМА «СИСТЕМНИЙ АНАЛІЗ»
СПЕЦІАЛЬНІСТЬ 124 «СИСТЕМНИЙ АНАЛІЗ»**

ПОГОДЖЕНО

Керівник проектної групи (гарант)
освітньо-професійної програми

_____ І.А. Джалладова
(підпис)
_____ 2026 р.

ЗАТВЕРДЖУЮ:

Завідувач кафедри системного аналізу
та кібербезпеки

_____ І.А. Джалладова
(підпис)
_____ 2026 р.

ІНДИВІДУАЛЬНЕ ЗАВДАННЯ

здобувачу вищої освіти Штирхун Христини Олександрівни

очної (денної) форми навчання

на підготовку кваліфікаційної бакалаврської роботи

на тему «Інтелектуальний аналіз кредитних даних банку»

Тему затверджено наказом ректора Університету від «__» ____ 20__ р. № ____

Кваліфікаційна бакалаврська робота виконується на матеріалах на матеріалах відкритого набору даних Credit Risk Dataset, а також з урахуванням матеріалів і практичного досвіду, отриманого під час проходження переддипломної практики в АТ «КРЕДІ АГРІКОЛЬ БАНК».

План кваліфікаційної бакалаврської роботи

Розділ 1	Теоретичні основи інтелектуального аналізу кредитних даних банку
Розділ 2	Аналіз кредитних даних та існуючих моделей оцінки ризику
Розділ 3	Розробка моделі інтелектуального аналізу кредитних даних

Об'єкт дослідження:	процес аналізу кредитних даних у банківській системі.
Предмет дослідження:	методи та моделі інтелектуального аналізу кредитних даних, що використовуються для їх підготовки, обробки, класифікації та оцінювання результатів.
Мета кваліфікаційної бакалаврської роботи:	дослідження та практичне застосування методів інтелектуального аналізу кредитних даних банку для підготовки інформації, побудови моделей класифікації та інтерпретації отриманих результатів.

Конкретні завдання, які здобувач повинен виконати для досягнення поставленої мети:

У розділі 1 Розкрити сутність кредитного скорингу та аналізу кредитних даних банку; охарактеризувати особливості кредитної інформації як об'єкта аналітичної обробки; розглянути основні методи Data Mining, машинного навчання та штучного інтелекту, що застосовуються у банківській сфері; визначити сучасні підходи до обробки та класифікації кредитних записів.

У розділі 2 Проаналізувати структуру кредитних даних банку та їх основні особливості; розглянути типові проблеми кредитних наборів даних, зокрема пропуски, дублікати, різні масштаби ознак і дисбаланс класів; описати етапи попередньої обробки даних; проаналізувати логістичну регресію та дерево рішень як базові моделі класифікації кредитних записів.

У розділі 3 Обрати інформаційну базу для практичного аналізу; виконати завантаження, первинну перевірку та очищення Credit Risk Dataset; сформулювати вхідні ознаки та цільову змінну; виконати кодування категоріальних ознак і поділ даних на навчальну та тестову вибірки; побудувати моделі логістичної регресії та

дерева рішень; оцінити результати моделей за точністю та матрицею помилок; провести інтерпретацію отриманих результатів і визначити можливості їх практичного використання для попередньої аналітичної обробки кредитної інформації банку.

**Завдання підготував
науковий керівник**

(підпис)

Камінський О. Є.

«_____» _____ 20__ р.

**Завдання одержав
Здобувач**

(підпис)

Штирхун Х. О.

«_____» _____ 20__ р.

Реферат

Кваліфікаційна бакалаврська робота містить 62 сторінок, 14 таблиць, 12 рисунків, 13 лістингів, список використаних джерел із 39 найменувань, 2 додатки.

«Інтелектуальний аналіз кредитних даних банку»

Об'єктом дослідження є процес аналізу кредитних даних у банківській системі.

Предметом дослідження є методи та моделі інтелектуального аналізу кредитних даних, що використовуються для обробки, класифікації та прогнозування у банківській сфері.

Мета кваліфікаційної бакалаврської роботи - дослідження та практична реалізація методів інтелектуального аналізу кредитних даних банку для підвищення ефективності обробки інформації та підтримки прийняття рішень у процесі кредитування.

Завданнями кваліфікаційної бакалаврської роботи є:

- проаналізувати теоретичні основи інтелектуального аналізу кредитних даних;
- дослідити сучасні методи Data Mining та машинного навчання у банківській сфері;
- розглянути структуру кредитних даних банку та особливості їх підготовки;
- проаналізувати моделі обробки та класифікації кредитних даних;
- сформувати набір даних для практичного аналізу;
- реалізувати моделі класифікації кредитних записів у середовищі Python;
- оцінити результати роботи побудованих моделей;
- визначити можливості практичного використання отриманих результатів у банківській діяльності.

Апаратні та програмні засоби, що використовувалися під час виконання роботи: персональний комп'ютер з операційною системою Windows, Microsoft Excel, Visual Studio Code, мова програмування Python, бібліотеки pandas, numpy, matplotlib, scikit-learn. Під час переддипломної практики також враховано досвід роботи з Microsoft Excel та внутрішньою банківською програмою iOpday.

У процесі роботи було розглянуто теоретичні основи кредитного скорингу, структуру кредитних даних банку, методи попередньої обробки даних та моделі класифікації. У практичній частині використано відкритий Credit Risk Dataset, який містить інформацію про позичальників, параметри кредиту, кредитну історію та статус кредиту. Було виконано завантаження, перевірку та очищення даних, кодування категоріальних ознак, поділ вибірки на навчальну і тестову частини, побудову логістичної регресії та дерева рішень.

Результати, досягнуті в процесі роботи: підготовлено очищений набір кредитних даних, побудовано дві моделі класифікації кредитних записів, виконано порівняння їх результатів за точністю та матрицею помилок. За результатами оцінювання дерево рішень показало вищу загальну точність - 0,891, тоді як логістична регресія показала точність 0,807.

Одержані результати можуть бути використані у навчальних і практичних цілях як приклад послідовної роботи з кредитною інформацією: від підготовки даних до побудови моделей і оцінювання результатів. Запропонований підхід може застосовуватися як допоміжний інструмент для попередньої класифікації кредитних записів та підготовки аналітичних висновків у банківській діяльності.

Рік виконання кваліфікаційної бакалаврської роботи: 2026 рік.

Рік захисту кваліфікаційної бакалаврської роботи: 2026 рік.

Ключові слова: кредитні дані, банк, Data Mining, машинне навчання, логістична регресія, дерево рішень

В і д г у к

про кваліфікаційну бакалаврську роботу

здобувача навчально-наукового інституту Інформаційних технологій в економіці

освітньо-професійної програми «Системний аналіз»

Штирхун Христини Олександрівни

на тему «Інтелектуальний аналіз кредитних даних банку»

1. Актуальність теми. Кваліфікаційна бакалаврська робота присвячена інтелектуальному аналізу кредитних даних банку. Зростання обсягів інформації про клієнтів, їхню кредитну історію та фінансову поведінку зумовлює необхідність застосування сучасних аналітичних методів для виявлення закономірностей і класифікації кредитних записів. Тому тема роботи Штирхун Христини Олександрівни є актуальною.

2. Позитивні риси роботи. У роботі комплексно розглянуто особливості кредитних даних і методи їх обробки. На основі відкритого набору Credit Risk Dataset побудовано моделі логістичної регресії та дерева рішень, проведено їх порівняння і сформовано практичні рекомендації. Робота має чітку структуру, логічну послідовність і належне практичне спрямування.

3. Самостійні розробки автора. Автором розроблено підхід до підготовки, попередньої обробки та класифікації кредитних записів, а також сформовано аналітичні матеріали за результатами моделювання.

4. Теоретична та практична цінність. Отримані результати підтверджують можливість застосування запропонованого підходу для обробки кредитної інформації та підтримки аналітичної діяльності банківських установ.

5. Наявність недоліків. Робота містить окремі недоліки, які не впливають на її загальну якість і практичну цінність.

6. Загальна оцінка. Тему роботи розкрито, поставлені завдання виконано, а висновки достатньо обґрунтовано. Кваліфікаційна бакалаврська робота Штирхун Христини Олександрівни на тему «Інтелектуальний аналіз кредитних даних банку» може бути допущена до захисту перед ЕК, заслуговує на позитивну оцінку, а її авторка - на присвоєння ступеня бакалавра за спеціальністю «Системний аналіз».

Науковий керівник

д.е.н., професор кафедри системного аналізу та кібербезпеки

Камінський О.Є.

(підпис)

(прізвище, ініціали)

“ ____ ” _____ 20__ р.

Рецензія

на кваліфікаційну бакалаврську роботу

здобувача вищої освіти

Штирхун Христини Олександрівни

Тема:

«Інтелектуальний аналіз кредитних даних банку»

Актуальність теми кваліфікаційної роботи і доцільність її розроблення. Тема роботи є актуальною, оскільки банківська діяльність передбачає обробку значних обсягів інформації про позичальників, кредитні договори, платежі та статус обслуговування кредитів. Застосування методів інтелектуального аналізу дає змогу перевіряти якість даних, структурувати кредитні записи, виявляти закономірності та здійснювати їх подальшу класифікацію.

Якість проведеного дослідження. У роботі послідовно розглянуто теоретичні та практичні аспекти аналізу кредитних даних. Авторкою охарактеризовано структуру кредитної інформації, етапи її попередньої обробки та основні моделі класифікації. Практичну частину виконано мовою Python на основі відкритого набору Credit Risk Dataset. Продемонстровано повний процес роботи з даними - від їх завантаження й очищення до побудови моделей та оцінювання отриманих результатів.

Позитивні риси кваліфікаційної бакалаврської роботи. Перевагою роботи є її практична спрямованість. Авторкою перевірено структуру набору даних, виявлено пропущені значення і дублікати, проведено очищення, кодування категоріальних ознак та поділ вибірки на навчальну і тестову частини. Для класифікації кредитних записів побудовано логістичну регресію та дерево рішень. Результати моделей проаналізовано за показниками точності та матрицями помилок, а також визначено ознаки, що можуть впливати на статус кредитного запису.

Зауваження. У роботі використано відкритий навчально-аналітичний набір даних, а не внутрішню інформацію банку. Це рішення є обґрунтованим через конфіденційність персональних і фінансових даних клієнтів. У подальших дослідженнях доцільно розширити перелік моделей машинного навчання та використати додаткові показники для оцінювання їхньої якості. Зазначені зауваження не знижують загальної цінності роботи.

Практична значущість висновків і рекомендацій. Запропонована послідовність дій може використовуватися для попередньої аналітичної обробки кредитної інформації. Отримані результати демонструють можливості перевірки якості даних, їх підготовки до моделювання, класифікації кредитних записів і порівняння моделей. Розроблений підхід може бути корисним для підготовки аналітичних матеріалів та підтримки внутрішніх процесів банківських установ.

Місце роботи та посада рецензента __АТ «КРЕДІ АГРІКОЛЬ БАНК», Начальник управління супроводження кредитних операцій корпоративного бізнесу Операційного департаменту__

Слободян Н. І.

Підпис засвідчую _ Начальник управління супроводження кредитних операцій корпоративного бізнесу Операційного департаменту __

Місце печатки організації, де працює рецензент

ЗМІСТ

ВСТУП.....	3
РОЗДІЛ 1	6
ТЕОРЕТИЧНІ ОСНОВИ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ КРЕДИТНИХ ДАНИХ БАНКУ	6
1.1 Сутність кредитного скорингу та аналізу кредитних даних	6
1.2 Методи інтелектуального аналізу даних (Data Mining, ML, AI)	10
1.3 Сучасні моделі та технології аналізу кредитних даних	14
РОЗДІЛ 2	17
АНАЛІЗ КРЕДИТНИХ ДАНИХ ТА ІСНУЮЧИХ МОДЕЛЕЙ ОЦІНКИ РИЗИКУ.....	17
2.1 Структура кредитних даних банку та їх особливості.....	17
2.2 Попередня обробка даних (очищення, нормалізація, балансування)	21
2.3 Аналіз існуючих моделей скорингу (логістична регресія, дерева рішень).....	30
РОЗДІЛ 3	36
РОЗРОБКА МОДЕЛІ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ КРЕДИТНИХ ДАНИХ.....	36
3.1 Характеристика інформаційної бази та попередня підготовка кредитних даних	36
3.2 Побудова аналітичної моделі обробки кредитних даних	42
3.3 Візуалізація та інтерпретація результатів інтелектуального аналізу даних	48
ВИСНОВКИ.....	56
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ	58
ДОДАТКИ.....	63

ВСТУП

Сучасна банківська діяльність тісно пов'язана з обробкою великих обсягів інформації. Особливо це стосується кредитування, оскільки під час видачі та супроводження кредитів банки накопичують дані про позичальників, їхні доходи, кредитну історію, платежі, забезпечення, страхування та умови кредитних договорів. Такі дані мають важливе значення для оцінювання клієнтів, контролю кредитних операцій і підготовки аналітичних висновків.

Зі збільшенням обсягів кредитної інформації традиційні способи її аналізу стають менш зручними. Ручна перевірка або використання лише окремих фінансових показників не завжди дозволяє швидко побачити закономірності у даних. Тому виникає потреба у застосуванні методів інтелектуального аналізу даних, які дають змогу обробляти кредитні записи, знаходити зв'язки між ознаками та виконувати попередню класифікацію за статусом обслуговування.

Методи Data Mining і машинного навчання можуть бути корисними для банківської сфери, оскільки вони допомагають працювати з великими таблицями даних і автоматизувати частину аналітичних процедур. За їх допомогою можна підготувати кредитну інформацію до моделювання, класифікувати записи, порівняти результати різних моделей і зробити висновки щодо якості отриманої класифікації. У межах цієї роботи такі методи розглядаються як допоміжний інструмент для обробки кредитної інформації, а не як повна заміна рішення банківського фахівця.

Актуальність теми кваліфікаційної роботи пов'язана з необхідністю вдосконалення підходів до обробки кредитних даних банку. У сучасних умовах важливо не лише зберігати інформацію про кредитні операції, а й уміти її перевіряти, структурувати, аналізувати та використовувати для підтримки прийняття рішень. Застосування методів Data Mining і машинного навчання дозволяє зробити цей процес більш послідовним і зрозумілим.

Об'єктом дослідження є процес аналізу кредитних даних у банківській системі.

Предметом дослідження є методи та моделі інтелектуального аналізу кредитних даних, що використовуються для їх підготовки, обробки, класифікації та оцінювання результатів.

Метою кваліфікаційної роботи є дослідження та практичне застосування методів інтелектуального аналізу кредитних даних банку для підготовки інформації, побудови моделей класифікації та інтерпретації отриманих результатів.

Для досягнення поставленої мети необхідно виконати такі завдання:

- проаналізувати теоретичні основи інтелектуального аналізу кредитних даних;
- розглянути можливості використання методів Data Mining та машинного навчання у банківській сфері;
- визначити структуру кредитних даних банку та їх основні особливості;
- дослідити етапи попередньої обробки кредитної інформації;
- розглянути моделі класифікації кредитних записів, зокрема логістичну регресію та дерево рішень;
- обрати набір даних для практичного аналізу;
- виконати підготовку кредитного датасету у середовищі Python;
- побудувати моделі класифікації кредитних записів;
- оцінити результати роботи моделей і порівняти їх між собою;
- визначити можливості використання отриманих результатів для попередньої аналітичної обробки кредитної інформації.

У процесі виконання роботи використовувалися методи аналізу та узагальнення, порівняльний аналіз, статистична обробка даних, методи попередньої підготовки даних, а також моделі машинного навчання. Практична частина виконувалася із застосуванням мови Python та бібліотек для роботи з табличними даними, побудови моделей і візуалізації результатів.

Практичне значення роботи полягає в тому, що запропонована послідовність дій може бути використана як приклад роботи з кредитною інформацією: від первинної перевірки та очищення даних до побудови моделей і порівняння результатів. Такий підхід може допомагати у попередній класифікації кредитних записів за статусом обслуговування та підготовці аналітичних матеріалів для внутрішніх потреб банку.

Кваліфікаційна робота складається зі вступу, трьох розділів, висновків, списку використаних джерел та додатків. У першому розділі розглянуто теоретичні основи інтелектуального аналізу кредитних даних банку. У другому розділі проаналізовано структуру кредитних даних, особливості їх підготовки та моделі класифікації. У третьому розділі виконано практичну реалізацію аналізу кредитного датасету, побудовано моделі логістичної регресії та дерева рішень, а також проведено оцінювання отриманих результатів.

РОЗДІЛ 1

ТЕОРЕТИЧНІ ОСНОВИ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ КРЕДИТНИХ ДАНИХ БАНКУ

1.1 Сутність кредитного скорингу та аналізу кредитних даних

Банківська діяльність сьогодні значною мірою залежить від якості роботи з даними. Особливо це помітно у сфері кредитування, де банк постійно накопичує інформацію про клієнтів, їхні доходи, платіжну поведінку, кредитні договори, забезпечення, страхування та інші дані, пов'язані з обслуговуванням кредитів. Через це кредитна інформація є важливою частиною банківських інформаційних систем і потребує системного аналізу.

Кредитні дані можна розглядати як сукупність відомостей, які банк використовує під час оцінювання позичальника, прийняття рішення щодо кредиту та подальшого супроводження кредитної операції. До таких даних належать фінансові показники клієнта, інформація про доходи й витрати, кредитна історія, параметри кредитного договору, наявність забезпечення, страхові документи та історія виконання зобов'язань.

Особливість таких даних полягає в тому, що вони мають різний формат і постійно оновлюються. Частина інформації надходить із внутрішніх банківських систем, наприклад систем обліку кредитів, CRM-систем або внутрішньої звітності. Інша частина може формуватися із зовнішніх джерел: кредитних бюро, державних реєстрів, фінансової звітності клієнтів та інших інформаційних ресурсів. Тому для якісного аналізу важливо не лише зібрати дані, а й перевірити їх повноту, правильність і придатність до подальшої обробки.

У сучасних умовах кредитна інформація має ознаки великих даних: вона накопичується у значних обсягах, часто оновлюється та може містити різні типи

показників. Це ускладнює її обробку лише традиційними способами, наприклад ручною перевіркою або простим переглядом таблиць. Саме тому в банківській сфері поступово зростає потреба в методах Data Mining та машинного навчання, які дозволяють структурувати дані, виявляти закономірності та виконувати класифікацію кредитних записів.

Практичну актуальність теми підтверджують статистичні дані Національного банку України. Зростання обсягів кредитування означає, що банки працюють із більшою кількістю записів про позичальників, умови договорів, платежі та статус обслуговування кредитів. У такій ситуації важливо мати інструменти, які допомагають швидше перевіряти інформацію, упорядковувати її та використовувати для аналітичних висновків.

Для кількісного підтвердження актуальності теми в табл. 1.1 наведено окремі статистичні показники, які характеризують динаміку банківського кредитування та пояснюють потребу в якісній обробці кредитної інформації.

Таблиця 1.1 - Статистичні показники, що підтверджують актуальність опрацювання кредитної інформації банку

Показник	Значення / динаміка	Значення для теми роботи
Чисті гривневі кредити бізнесу	зростають на 30 % і більше у річному вимірі	збільшення портфеля посилює потребу в системній обробці кредитної інформації
Бізнес-кредити	річне зростання прискорилося до 35 %	розширення корпоративного кредитування потребує якісної класифікації записів
Роздрібний кредитний портфель	темп зростання близько 33 % у річному вимірі	масові кредитні продукти формують великі масиви однотипних записів для аналізу
Якість корпоративного портфеля	частка NPL бізнесу скорочувалася до 37 % у квітні 2025 р., або до 20 % без старих NPL	для банку важливо відстежувати не лише обсяг, а й статус обслуговування кредитів

Джерело: складено автором за даними Національного банку України [6; 7].

З табл. 1.1 видно, що тема роботи пов'язана не лише з теоретичним використанням методів Data Mining, а й із практичними змінами у банківському секторі. Зростання кредитного портфеля приводить до збільшення обсягу інформації, яку потрібно перевіряти, зберігати та аналізувати. За таких умов ручна

або фрагментарна обробка даних не завжди є достатньою, тому доцільно використовувати більш послідовний підхід: очищення записів, перевірку якості, підготовку ознак і подальшу класифікацію.

Для наочного відображення окремих показників банківського кредитування побудовано рис. 1.1.

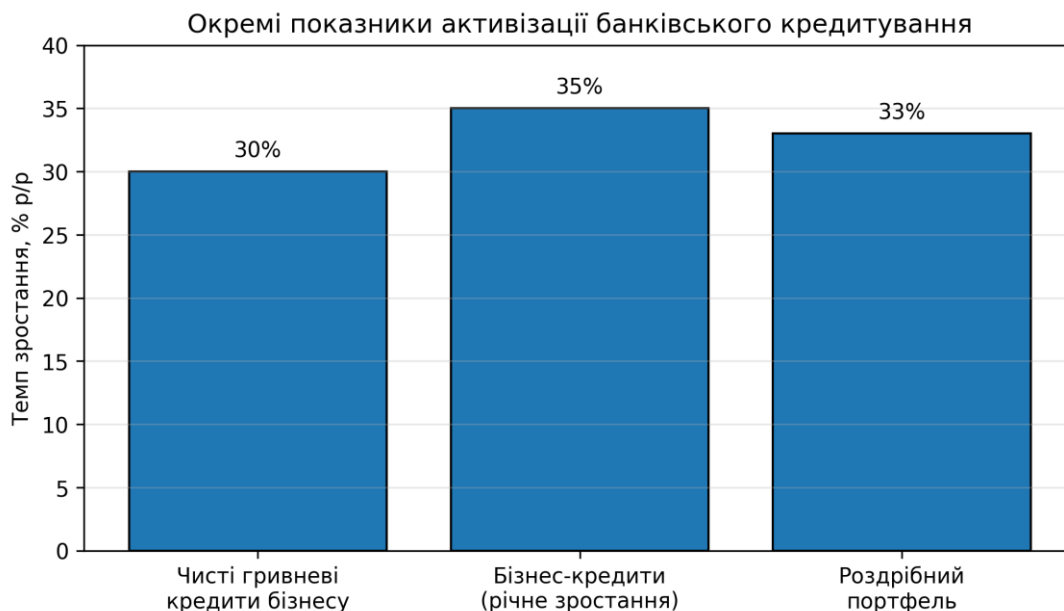


Рисунок 1.1 - Окремі показники активізації банківського кредитування в Україні. Джерело: побудовано автором за даними Національного банку України [6; 7].

З рис. 1.1 видно, що кредитування залишається важливим напрямом банківської діяльності. Для цієї роботи це має значення, оскільки збільшення кількості кредитних операцій супроводжується накопиченням нових даних про позичальників, параметри кредитів, історію платежів і результат обслуговування. Тому застосування методів інтелектуального аналізу є доцільним для впорядкування таких даних та побудови моделей класифікації.

Однією з основних проблем у роботі з кредитною інформацією є її якість. У банківських записах можуть траплятися пропущені значення, дублікати, помилки введення або різні варіанти подання одного й того самого показника. Якщо такі недоліки не усунути, результати аналізу можуть бути неточними. Тому

перед побудовою моделей необхідно виконувати попередню перевірку та підготовку даних.

Окремо потрібно враховувати питання захисту інформації. Кредитні записи містять конфіденційні дані про клієнтів, тому банк має забезпечувати контроль доступу, захист персональних відомостей і дотримання вимог інформаційної безпеки. Це важливо не лише для технічної роботи з даними, а й для збереження довіри клієнтів та стабільності банківських процесів.

Питання безпечної роботи з даними також пов'язане з темою інтелектуального аналізу. У тезах Х. О. Штирхун, підготовлених у межах секції «Смарт-аналітика та кібербезпека», зазначено, що контроль доступу до чутливих інформаційних ресурсів допомагає запобігати несанкціонованому доступу, захищати персональні дані та підтримувати безперервність бізнес-процесів [19]. Для кредитної аналітики це означає, що моделі Data Mining і машинного навчання мають використовуватися з урахуванням правил доступу, автентифікації користувачів і контролю роботи з конфіденційними даними.

Кредитна інформація має не лише облікове, а й аналітичне значення. На її основі банк може оцінювати фінансовий стан клієнта, аналізувати його попередню платіжну поведінку, виявляти типові закономірності та контролювати якість кредитного портфеля. Саме тому такі дані можуть бути основою для побудови моделей, які допомагають виконувати попередню класифікацію кредитних записів.

Інтелектуальний аналіз кредитних даних базується на використанні методів Data Mining і машинного навчання. Такі методи дозволяють не лише зберігати інформацію, а й обробляти її, знаходити зв'язки між ознаками та будувати моделі для класифікації записів. У межах цієї роботи основна увага приділяється не повній автоматизації банківського рішення, а демонстрації того, як підготовлені дані можуть бути використані для аналітичної обробки.

У банківській сфері подібні підходи можуть застосовуватися для кредитного скорингу, виявлення нетипових операцій, сегментації клієнтів, аналізу кредитної історії та підготовки внутрішніх звітів. Їх перевага полягає в тому, що

вони допомагають працювати з великими масивами даних більш швидко та послідовно, ніж звичайна ручна перевірка.

Під час проходження переддипломної практики було встановлено, що значна частина роботи з кредитною інформацією пов'язана з її перевіркою, обробкою та підготовкою матеріалів для внутрішнього використання й аудиту. Робота з таблицями у Microsoft Excel дала змогу побачити, наскільки важливими є правильна структура даних, відсутність помилок і можливість швидко знаходити потрібну інформацію. Саме цей практичний досвід підтвердив доцільність використання аналітичних інструментів у темі дипломної роботи.

Отже, кредитні дані є важливим ресурсом для банку, оскільки вони використовуються не лише для обліку кредитних операцій, а й для аналізу, класифікації та підготовки управлінських висновків. Їх ефективна обробка потребує якісної підготовки даних, дотримання вимог інформаційної безпеки та використання методів, які дозволяють працювати з великими обсягами інформації.

1.2 Методи інтелектуального аналізу даних (Data Mining, ML, AI)

Інтелектуальний аналіз кредитних даних банку пов'язаний із використанням методів Data Mining, машинного навчання та штучного інтелекту. Їх застосовують для того, щоб працювати з великими обсягами фінансової інформації, знаходити зв'язки між показниками та краще розуміти поведінку позичальників. У банківській сфері такі методи можуть допомагати під час підготовки кредитної інформації, класифікації записів, оцінювання статусу обслуговування кредитів і формування аналітичних висновків.

Data Mining можна розглядати як процес пошуку корисної інформації у великих наборах даних. У випадку кредитної аналітики це можуть бути закономірності між доходом позичальника, сумою кредиту, відсотковою ставкою,

кредитною історією та результатом обслуговування кредиту. Тобто головне завдання полягає не просто в накопиченні даних, а в тому, щоб перетворити їх на зрозумілу інформацію для подальшого аналізу.

У межах інтелектуального аналізу кредитних даних можна виділити кілька основних груп методів. Найчастіше використовуються класифікаційні методи, регресійні моделі, кластеризація, асоціативний аналіз і візуальна аналітика. Кожна група методів має своє призначення і може застосовуватися на різних етапах роботи з банківською інформацією.

Для того щоб показати практичне значення цих методів, доцільно пов'язати їх не лише з теоретичним описом, а й із конкретними завданнями обробки кредитної інформації. Відповідність основних методів Data Mining завданням банківської аналітики узагальнено в табл. 1.2.

Таблиця 1.2 - Відповідність методів Data Mining завданням обробки кредитної інформації

Група методів	Сутність застосування	Типове завдання у банківській сфері
Класифікація	віднесення запису до одного із заздалегідь визначених класів	визначення статусу кредиту або групи обслуговування
Регресійні методи	моделювання кількісної залежності між ознаками та результатом	оцінювання впливу доходу, ставки або суми кредиту на підсумковий показник
Кластеризація	групування спостережень без попередньо заданих класів	сегментація позичальників за подібними фінансовими характеристиками
Асоціативний аналіз	виявлення типових комбінацій ознак	пошук повторюваних зв'язків між параметрами кредиту та поведінкою клієнта
Візуальна аналітика	графічне подання структури та якості даних	контроль пропусків, викидів, дисбалансу класів і взаємозв'язків між ознаками

Джерело: сформовано автором на основі узагальнення підходів до Data Mining і машинного навчання в банківській сфері [13].

З табл. 1.2 видно, що інтелектуальний аналіз даних не зводиться лише до побудови моделі. Він охоплює ширший процес: перевірку якості даних, підготовку ознак, вибір методу, отримання результату та його пояснення. Саме така послідовність використовується у практичній частині роботи, де спочатку

виконується підготовка кредитного датасету, а вже потім будуються моделі класифікації.

Загальну послідовність застосування Data Mining до кредитного датасету показано на рисунку 1.2.



Рисунок 1.2 - Узагальнена послідовність застосування Data Mining до кредитного датасету. Джерело: побудовано автором для узагальнення етапів інтелектуального опрацювання кредитної інформації.

З рисунку 1.2 видно, що робота з кредитним датасетом має виконуватися поетапно. Спочатку дані потрібно завантажити, перевірити та очистити. Після цього здійснюється підготовка ознак, вибір методу, побудова моделі та оцінювання результатів. Така логіка є важливою, оскільки навіть якісна модель не дасть коректного результату, якщо початкові дані містять помилки або не підготовлені до аналізу.

Класифікаційні методи є одними з найважливіших для теми цієї роботи. Вони дають змогу віднести кредитний запис до певного класу. У найпростішому випадку йдеться про два класи: кредит без дефолту та кредит із дефолтним статусом. Саме така постановка задачі використовується у практичному розділі, де для класифікації застосовуються логістична регресія та дерево рішень.

Регресійні методи використовують тоді, коли потрібно оцінити кількісний зв'язок між показниками. Наприклад, можна аналізувати, як дохід, сума кредиту, кредитне навантаження або відсоткова ставка пов'язані з підсумковим результатом. У банківських задачах такі методи допомагають краще зрозуміти, які фактори можуть впливати на фінансову поведінку позичальника.

Методи кластеризації застосовуються для групування клієнтів або кредитних записів за схожими характеристиками. На відміну від класифікації, тут заздалегідь не задається готовий клас. Наприклад, позичальників можна поділити на групи за рівнем доходу, кредитним навантаженням, тривалістю кредитної історії або іншими ознаками. Це може бути корисним для сегментації клієнтів і подальшого аналізу їхньої поведінки.

Асоціативний аналіз дає змогу знаходити типові поєднання ознак у даних. У кредитній сфері це може бути корисно для виявлення повторюваних зв'язків між параметрами кредиту та поведінкою клієнта. Наприклад, певні комбінації високої відсоткової ставки, великої суми кредиту та короткої кредитної історії можуть потребувати додаткової уваги під час аналізу.

Машинне навчання охоплює методи, які дозволяють моделі навчатися на історичних даних і використовувати цей досвід для нових записів. У задачах кредитної класифікації найчастіше використовується навчання з учителем, оскільки в датасеті вже є відомий результат - наприклад, статус кредиту. Модель аналізує зв'язок між вхідними ознаками та цим результатом, а потім застосовує отриману логіку до тестових даних.

Штучний інтелект і складніші моделі, зокрема нейронні мережі та ансамблеві алгоритми, також можуть застосовуватися в банківській сфері. Вони здатні працювати з великими наборами даних і знаходити складніші залежності. Проте такі моделі не завжди легко пояснити, тому в бакалаврській роботі доцільно використовувати більш зрозумілі підходи, зокрема логістичну регресію та дерево рішень.

Окреме значення має відбір ознак. Для побудови моделі важливо використовувати ті показники, які справді мають зміст для аналізу. Надлишкові або випадкові змінні можуть ускладнювати роботу моделі та погіршувати інтерпретацію результатів. Тому перед моделюванням потрібно оцінити, які ознаки описують позичальника, умови кредиту та кредитну історію.

Отже, методи інтелектуального аналізу даних дають змогу послідовно працювати з кредитною інформацією: від підготовки даних до побудови моделей і

пояснення результатів. У межах цієї роботи основну увагу приділено тим методам, які можна зрозуміло застосувати на практиці: підготовці кредитного датасету, класифікації записів, порівнянню моделей і тлумаченню отриманих результатів.

1.3 Сучасні моделі та технології аналізу кредитних даних

Сучасний аналіз кредитної інформації банку поєднує статистичні методи, машинне навчання, роботу з великими масивами даних і економічну інтерпретацію результатів. Його головна мета полягає не лише в обробці окремих показників, а й у тому, щоб допомогти банку краще розуміти структуру кредитних записів, виявляти закономірності та використовувати отримані результати для підготовки аналітичних висновків.

Теоретичний аналіз показує, що банки працюють з різними видами інформації. До неї належать фінансові показники клієнтів, кредитна історія, параметри договорів, поведінкові характеристики, дані про платежі та зовнішні економічні умови. Через таку різноманітність даних звичайної ручної перевірки або простого аналізу таблиць часто недостатньо. Саме тому все більшого значення набувають методи інтелектуального аналізу даних.

У роботі з кредитною інформацією можуть використовуватися різні групи методів. Класифікаційні моделі дають змогу віднести запис до певного класу, наприклад до кредитів без дефолту або кредитів із дефолтним статусом. Регресійні підходи допомагають оцінити кількісні залежності між ознаками. Кластеризація використовується для групування схожих клієнтів або кредитних записів. Методи машинного навчання дозволяють навчати модель на історичних даних і застосовувати отриману логіку до нових спостережень.

Окреме місце займають складніші моделі штучного інтелекту, зокрема нейронні мережі та ансамблеві алгоритми. Вони можуть працювати з великими

наборами даних і враховувати складні зв'язки між показниками. Проте такі моделі не завжди зручні для пояснення, що є важливим саме для банківської сфери. У банку результат моделі має бути не лише точним, а й зрозумілим для фахівців, які надалі працюють з кредитною інформацією.

Однією з основних проблем під час аналізу кредитних даних є якість вхідної інформації. Якщо у вибірці є пропуски, дублікати, помилки введення або нетипові значення, це може вплинути на результат моделювання. Тому перед використанням будь-якої моделі потрібно виконати попередню підготовку даних: перевірити записи, очистити їх, привести ознаки до потрібного формату та оцінити структуру вибірки.

Ще однією важливою проблемою є пояснення результатів складних моделей. Наприклад, нейронні мережі або ансамблеві методи можуть показувати високу точність, але їх рішення не завжди легко пояснити. Для банківської аналітики це є суттєвим обмеженням, оскільки результати мають бути придатними для перевірки, обґрунтування та подальшого використання в роботі фахівців.

Також потрібно враховувати, що кредитна інформація часто надходить з різних джерел. Частина даних формується у внутрішніх системах банку, частина може надходити з кредитних бюро, державних реєстрів або інших зовнішніх ресурсів. Через це виникає потреба в узгодженні форматів, перевірці повноти записів і зведенні інформації до єдиної структури.

Отже, сучасний аналіз кредитних даних пов'язаний не тільки з вибором моделі, а й з усією підготовчою роботою: збором, очищенням, перевіркою та структуруванням інформації. Без цього навіть складний алгоритм не зможе дати якісний результат. Тому в цій роботі основна увага приділяється зрозумілій послідовності дій: підготовці даних, вибору моделей класифікації та поясненню отриманих результатів.

У межах бакалаврської роботи доцільно використовувати такі моделі, які можна не лише реалізувати, а й пояснити. Саме тому для практичної частини обрано логістичну регресію та дерево рішень. Логістична регресія дозволяє оцінити ймовірність належності запису до певного класу, а дерево рішень дає

змогу подати класифікацію у вигляді послідовності правил. Обидва підходи є достатньо зрозумілими для інтерпретації та відповідають завданню роботи.

Таким чином, узагальнення теоретичних підходів дозволяє перейти до більш детального аналізу структури кредитних даних, методів їх підготовки та моделей класифікації. Саме ці питання розглядаються у другому розділі, де основну увагу приділено особливостям кредитних записів, попередній обробці даних, логістичній регресії та деревам рішень як базовим інструментам інтелектуального аналізу.

РОЗДІЛ 2

АНАЛІЗ КРЕДИТНИХ ДАНИХ ТА ІСНУЮЧИХ МОДЕЛЕЙ ОЦІНКИ РИЗИКУ

2.1 Структура кредитних даних банку та їх особливості

Кредитні дані банку є важливим джерелом інформації, яке формується під час видачі та супроводження кредитів. Вони використовуються для оцінювання клієнтів, контролю виконання умов кредитних договорів, аналізу платіжної поведінки та підготовки управлінських рішень. У сучасних умовах обсяги такої інформації постійно збільшуються, тому для її обробки доцільно застосовувати не лише стандартні табличні інструменти, а й методи інтелектуального аналізу даних.

У загальному вигляді кредитний набір даних можна подати як таблицю, у якій кожен рядок відповідає окремому позичальнику, а кожен стовпець - певній ознаці. Формально таку структуру можна записати у вигляді матриці спостережень:

$$X = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \dots & x_n^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \dots & x_n^{(2)} \\ \dots & \dots & \dots & \dots \\ x_1^{(N)} & x_2^{(N)} & \dots & x_n^{(N)} \end{bmatrix} \quad (2.1)$$

Джерело: розроблено автором на основі узагальнення підходів до подання вхідних ознак у задачах кредитного скорингу [25; 29].

де кожен рядок відповідає окремому клієнту, а кожен стовпець - певній характеристиці позичальника. Таке подання є зручним для подальшої обробки,

оскільки дозволяє застосовувати статистичні методи та моделі машинного навчання.

Кредитна інформація, яка використовується у банківській аналітиці, може мати різний зміст. Частина даних описує самого клієнта: вік, сімейний стан, рівень освіти або інші загальні характеристики. Інша частина пов'язана з фінансовим станом позичальника: доходами, витратами, наявними зобов'язаннями та кредитним навантаженням. Саме ці показники дають змогу оцінити, чи може клієнт виконувати свої кредитні зобов'язання.

Окреме значення мають дані кредитної історії. До них належать відомості про попередні кредити, прострочення, тривалість кредитної історії та дисципліну погашення. Такі показники допомагають зрозуміти, як клієнт поведився у минулих кредитних операціях. Крім того, у банківській практиці можуть використовуватися поведінкові дані, наприклад регулярність надходжень коштів, активність за рахунками та частота використання банківських продуктів.

Практичне розуміння структури кредитної інформації було сформовано під час проходження переддипломної практики в АТ «КРЕДІ АГРІКОЛЬ БАНК». У межах практики кредитні дані розглядалися не лише як набір числових показників, а як частина реального банківського процесу. Особливу увагу було приділено інформації, пов'язаній із супроводженням кредитних операцій корпоративного бізнесу: параметрам кредитів, даним про позичальників, забезпеченню, страхуванню, лізинговим угодам, строкам виконання зобов'язань і матеріалам для внутрішнього контролю.

АТ «КРЕДІ АГРІКОЛЬ БАНК» здійснює обслуговування фізичних та юридичних осіб, зокрема корпоративних клієнтів. Для цієї роботи важливим є саме напрям супроводження кредитних операцій, оскільки в ньому формується, перевіряється та уточнюється значна частина інформації, пов'язаної з кредитними договорами, забезпеченням та виконанням умов обслуговування.

Під час практики робота була пов'язана з Управлінням супроводження кредитних операцій корпоративного бізнесу операційного департаменту. Основні завдання цього напрямку стосувалися перевірки даних щодо кредитів, страхування,

заставного забезпечення, лізингових операцій, а також підготовки матеріалів для внутрішньої звітності й аудиту. Це показало, що кредитна інформація у банку має не тільки фінансовий, а й документальний, контрольний та технічний характер.

Значна частина практичної роботи виконувалася в Microsoft Excel та внутрішній банківській програмі iOrday. Excel використовувався для фільтрації, звірки, пошуку неточностей та впорядкування даних. Програма iOrday застосовувалася для роботи з операційною інформацією банку. Такий досвід показав, що навіть за наявності спеціалізованих систем важливим залишається етап попередньої підготовки даних: перевірка записів, пошук пропусків, виявлення дублювання, узгодження таблиць і підготовка інформації до подальшого аналізу.

Отже, у реальній банківській діяльності кредитні дані не обмежуються лише показниками доходу, суми кредиту або наявності прострочення. Вони також включають дані про забезпечення, страхування, лізингові договори, історію обслуговування та результати внутрішніх перевірок. Тому перед побудовою моделі важливо не тільки обрати алгоритм, а й правильно підготувати вхідну інформацію. Види банківської інформації, що формуються під час супроводження кредитних операцій, узагальнено в табл. 2.1.

Таблиця 2.1 - Види банківської інформації, пов'язаної із супроводженням кредитних операцій

Група інформації	Приклад змісту	Значення для подальшої обробки
Дані про кредитну операцію	сума кредиту, строк договору, графік платежів, статус обслуговування	формують основу для контролю та класифікації записів
Відомості про позичальника	фінансовий стан, ідентифікаційні дані, історія взаємодії з банком	допомагають пов'язати параметри договору з характеристиками клієнта
Дані щодо забезпечення	застава, документи щодо майна, характеристики забезпечення	використовуються для перевірки повноти супроводження договору

Продовження таблиці 2.1

Страхові документи	наявність страхового договору, строк дії, пов'язані документи	дають змогу контролювати виконання умов кредитного супроводження
Лізингові операції	параметри договору, об'єкт лізингу, строки платежів	розширюють інформаційний контекст корпоративного клієнта
Матеріали для аудиту	звірені таблиці, результати перевірок, підготовлені аналітичні матеріали	забезпечують внутрішній контроль і перевірку узгодженості даних
Технічні записи обробки	Excel-файли, фільтри, формули, звірки з iOpday	підтверджують необхідність очищення й структурування даних

Джерело: сформовано автором на основі завдань, виконаних під час переддипломної практики в АТ «КРЕДІ АГРІКОЛЬ БАНК».

З табл. 2.1 видно, що під час супроводження кредитних операцій банк працює з різними групами інформації. Частина з них має фінансовий зміст, частина стосується документів, внутрішнього контролю або технічної обробки. Це підтверджує, що перед використанням таких даних у моделі їх потрібно перевірити, узгодити та привести до єдиної структури.

Інформаційна база кредитування є ширшою, ніж звичайна таблиця з кількома фінансовими показниками. Вона включає відомості про клієнта, умови договору, забезпечення, страхування, історію обслуговування та результати внутрішніх перевірок. Тому якість підготовки вхідних записів безпосередньо впливає на коректність майбутньої класифікації. Спрощений приклад структури кредитного набору даних подано в табл. 2.2.

Таблиця 2.2 - Приклад структури кредитних даних

ID	Вік	Дохід	Кредитний рейтинг	Прострочення	Сума кредиту	Клас
001	35	25000	720	0	500000	0
002	42	18000	640	1	300000	1
003	29	32000	780	0	700000	0

Джерело: розроблено автором на умовному прикладі для демонстрації структури вхідних даних.

Табл. 2.2 показує спрощений приклад кредитного набору даних, який може бути використаний для побудови моделей класифікації. Кожен рядок відповідає окремому позичальнику, а стовпці містять його основні характеристики. У реальних банківських системах таких ознак може бути значно більше, але наведений приклад дозволяє зрозуміти загальну логіку подання даних.

У таблиці ID є ідентифікатором клієнта, який потрібен для обліку записів, але не має аналітичного значення для моделі. Вік і дохід описують загальні та фінансові характеристики позичальника. Кредитний рейтинг і прострочення пов'язані з кредитною історією. Сума кредиту характеризує розмір зобов'язання, а клас є цільовою змінною: 0 означає кредит без дефолту, 1 - кредит із дефолтним статусом або проблемним обслуговуванням.

Наведений приклад не містить реальних банківських записів і використовується лише для пояснення структури даних. У практичній діяльності банку інформація є більш деталізованою та частково конфіденційною. Саме тому робота з такими даними потребує не лише математичної обробки, а й дотримання правил інформаційної безпеки та обмеження доступу до чутливої інформації.

Для подальшого моделювання важливо розділити вхідні ознаки та цільову змінну. У задачі класифікації кредитних записів цільова змінна найчастіше має два значення: 0 - кредит без дефолту, 1 - кредит із дефолтним або проблемним статусом. Така структура дає змогу перейти до побудови моделей класифікації, зокрема логістичної регресії та дерева рішень.

2.2 Попередня обробка даних (очищення, нормалізація, балансування)

Перед побудовою моделі кредитні дані потрібно підготувати до аналізу. Модель машинного навчання не працює безпосередньо з первинними банківськими документами, а використовує впорядковану таблицю ознак. У такій

таблиці кожен рядок відповідає окремому кредитному запису або позичальнику, а кожен стовпець містить певний показник: вік, дохід, суму кредиту, кредитний рейтинг, наявність прострочень або інші характеристики.

Якщо у даних є пропущені значення, дублікати, різні масштаби показників або нерівномірний розподіл класів, модель може працювати неточно. Тому попередня обробка є обов'язковим етапом перед моделюванням. У цій роботі вона розглядається як послідовність зрозумілих дій: перевірка якості даних, усунення пропусків і повторів, кодування категоріальних змінних, стандартизація числових ознак, аналіз співвідношення класів та графічна перевірка структури датасету.

Загальну логіку підготовки кредитних даних до моделювання наведено на рисунку 2.1.

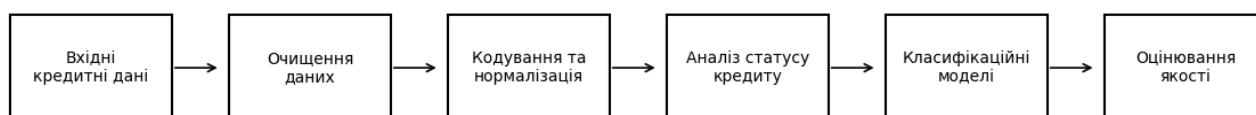


Рисунок 2.1 - Загальна логіка підготовки кредитних даних до моделювання.

Джерело: побудовано автором у Python; код наведено у додатку А, лістинг А.1.

З рисунку 2.1 видно, що побудова моделі є не першим, а одним із завершальних етапів роботи з даними. Перед цим необхідно перевірити записи, усунути технічні помилки, підготувати ознаки та переконатися, що структура датасету придатна для подальшого аналізу.

Очищення даних.

Першим етапом попередньої обробки є перевірка набору даних на пропущені значення, дублікати та очевидні помилки. Наприклад, від'ємний вік клієнта, нульовий дохід за активним кредитом або надто велика сума кредиту можуть свідчити про помилку введення або потребу в додатковій перевірці.

Для числових ознак пропущені значення зазвичай замінюють середнім значенням або медіаною. Середнє значення доцільно використовувати тоді, коли розподіл показника є відносно рівномірним і не має значних викидів. Якщо ж

ознака має асиметричний розподіл, наприклад дохід або сума кредиту, краще застосовувати медіану, оскільки вона менш чутлива до нетипових значень.

Середнє значення числової ознаки визначається за формулою:

$$\bar{x} = (x_1 + x_2 + \dots + x_n) / n, \quad (2.2)$$

де $\bar{x}$ - середнє значення ознаки; $x_1, x_2, \dots, x_n$ - значення ознаки для окремих спостережень;
 n - кількість спостережень.

Джерело формули: складено автором на основі стандартного розрахунку середнього значення, що застосовується у статистичній підготовці ознак [29; 32].

Основні проблеми, які можуть виникати під час попередньої обробки кредитних даних, наведено в табл. 2.3.

Таблиця 2.3 - Типові проблеми кредитних даних та способи їх усунення

Проблема даних	Приклад у кредитному наборі	Можливий спосіб обробки
Пропущені значення	відсутній дохід або кредитний рейтинг	заміна середнім значенням, медіаною або окремою категорією
Дублікати	один кредитний договір випадково внесено двічі	видалення повторів після перевірки ключових ідентифікаторів
Викиди	аномально висока сума кредиту або дохід	квартильний аналіз, перевірка джерела, за потреби обмеження значень
Різні масштаби	вік вимірюється роками, сума кредиту - гривнями	стандартизація числових ознак
Категоріальні ознаки	тип зайнятості, наявність забезпечення	перетворення категорій у числові індикатори

Джерело: розроблено автором для узагальнення етапів попередньої обробки кредитного набору даних.

З табл. 2.3 видно, що перед побудовою моделей потрібно перевіряти не лише пропуски, а й повтори, нетипові значення та формат ознак. Якщо ці проблеми залишити без обробки, модель може неправильно враховувати окремі записи або давати нестійкий результат.

Кодування та стандартизація ознак.

Кредитний набір даних зазвичай містить як числові, так і категоріальні змінні. До числових належать вік, дохід, сума кредиту, строк кредитування або кредитний рейтинг. Категоріальними можуть бути тип зайнятості, рівень освіти, сімейний стан, вид кредитного продукту або наявність забезпечення. Перед використанням у моделі такі категорії потрібно подати у числовому вигляді.

Числові ознаки також потребують додаткової підготовки, оскільки вони можуть мати різні одиниці вимірювання. Наприклад, вік вимірюється роками, кредитний рейтинг - балами, а сума кредиту - гривнями. Якщо не виконати стандартизацію, показники з більшими числовими значеннями можуть мати надмірний вплив на модель.

Стандартизоване значення ознаки обчислюється за формулою:

$$z_i = (x_i - \bar{x}) / s \quad (2.3)$$

де z_i - стандартизоване значення i -го спостереження; x_i - початкове значення ознаки; $\bar{x}$ - середнє значення ознаки; s - стандартне відхилення.

Джерело: розроблено автором на основі використання z -перетворення у задачах попередньої підготовки ознак [29; 32].

Стандартне відхилення для вибірки можна подати так:

$$s = \sqrt{(\sum (x_i - \bar{x})^2 / (n - 1))} \quad (2.4)$$

Джерело: розроблено автором на основі стандартного визначення вибіркового стандартного відхилення [32].

Приклад стандартизації числових ознак подано в табл. 2.4.

Таблиця 2.4 - Приклад стандартизації числових ознак

Ознака	Початкове значення	Середнє значення	Стандартне відхилення	Стандартизоване значення
Дохід, грн	35 000	30 000	10 000	0,50

Продовження таблиці 2.4

Сума кредиту, грн	300 000	250 000	100 000	0,50
Кредитний рейтинг, бали	680	650	60	0,50

Джерело: розраховано автором для демонстрації принципу стандартизації кредитних ознак.

З табл. 2.4 видно, що після стандартизації різні за масштабом показники можна порівнювати між собою. Це важливо для моделей, які чутливі до числових значень ознак, зокрема для логістичної регресії.

Аналіз розподілу класів статусу кредиту.

У задачах класифікації важливо перевірити, як розподілена цільова змінна. У цій роботі клас 0 означає кредит без дефолту, а клас 1 - кредит із дефолтним статусом або проблемним обслуговуванням. На практиці записи з дефолтним статусом часто трапляються рідше, ніж стандартно обслуговувані кредити. Через це модель може краще розпізнавати більший клас і гірше працювати з менш представленим.

Щоб не ускладнювати роботу методами штучного збільшення вибірки, у дослідженні використовується простіший підхід. Спочатку визначається частка кожного класу, потім під час поділу на навчальну та тестову вибірки зберігається їх співвідношення, а в моделях за потреби застосовуються вагові коефіцієнти класів. Такий підхід є зрозумілим і дозволяє не ігнорувати клас, який має меншу кількість спостережень.

Частка об'єктів певного класу визначається за формулою:

$$p_k = n_k / n, \quad (2.5)$$

де p_k - частка об'єктів класу k ; n_k - кількість спостережень класу k ; n - загальна кількість спостережень у вибірці.

Джерело: розроблено автором на основі загального визначення відносної частоти класу [32].

Ваговий коефіцієнт класу за принципом оберненої пропорційності можна подати так:

$$w_k = n / (K \cdot n_k), \quad (2.6)$$

де w_k - вага класу k ; K - кількість класів; n_k - кількість об'єктів класу k .

Джерело: розроблено автором на основі принципу обернено пропорційного зважування класів у задачах класифікації [29; 33].

Якщо клас представлений меншою кількістю спостережень, його вага буде більшою, і модель сильніше враховуватиме помилки саме за цим класом. Приклад розрахунку часток класів і вагових коефіцієнтів для статусу кредиту подано в табл. 2.5.

Таблиця 2.5 - Приклад аналізу співвідношення класів статусу кредиту

Клас	Зміст класу	Кількість спостережень	Частка, %	Ваговий коефіцієнт
0	кредит без дефолту	86	71.7	0.70
1	кредит із дефолтним статусом	34	28.3	1.76
Разом	-	120	100,0	-

Джерело: розраховано автором на умовному прикладі для пояснення дисбалансу класів у кредитному скорингу.

З табл. 2.5 видно, що клас із дефолтним статусом має меншу кількість спостережень, тому для нього ваговий коефіцієнт є вищим. Такий підхід допомагає зменшити ситуацію, коли модель переважно орієнтується на більший клас і гірше розпізнає менш представлений.

Візуалізація даних після попередньої перевірки.

Графічний аналіз потрібен для швидкого огляду структури даних. Деякі проблеми легше побачити саме на графіках, ніж у таблиці. У роботі доцільно використовувати гістограму, діаграму розмаху, стовпчикову діаграму класів і кореляційну матрицю. Наведені графіки побудовано на умовному прикладі

кредитного набору даних, тому вони показують методику аналізу, а не реальні банківські записи.

Для попереднього огляду числових показників побудовано гістограму розподілу доходу позичальників, наведену на рисунку 2.2.

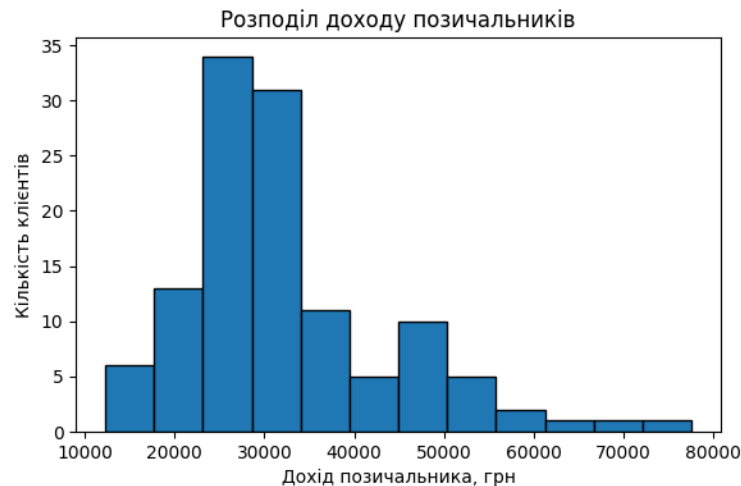


Рисунок 2.2 - Гістограма розподілу доходу позичальників. Джерело: побудовано автором у Python; код наведено в додатку А, лістинг А.2.

З рисунку 2.2 видно, як розподіляються значення доходу в умовному наборі спостережень. Такий графік допомагає побачити, у якому діапазоні зосереджена більшість значень, а також чи є поодинокі дуже високі показники. Це важливо для вибору способу обробки числових ознак.

Для виявлення нетипових значень суми кредиту використано діаграму розмаху, подану на рисунку 2.3.

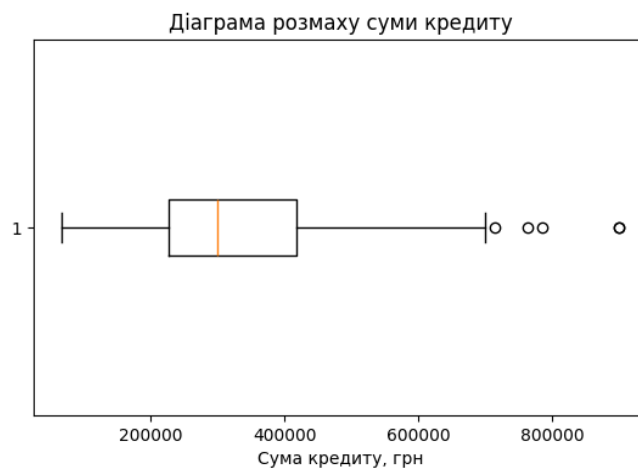


Рисунок 2.3 - Діаграма розмаху для виявлення нетипових значень суми кредиту. Джерело: побудовано автором у Python; код наведено в додатку А, лістинг А.3.

З рисунку 2.3 видно, що діаграма розмаху дозволяє оцінити медіану, квартилі та можливі викиди. У кредитних даних великі значення не завжди є помилками, але вони потребують окремої перевірки, оскільки можуть впливати на загальні характеристики вибірки.

Співвідношення кредитних записів за статусом обслуговування показано на рисунку 2.4.

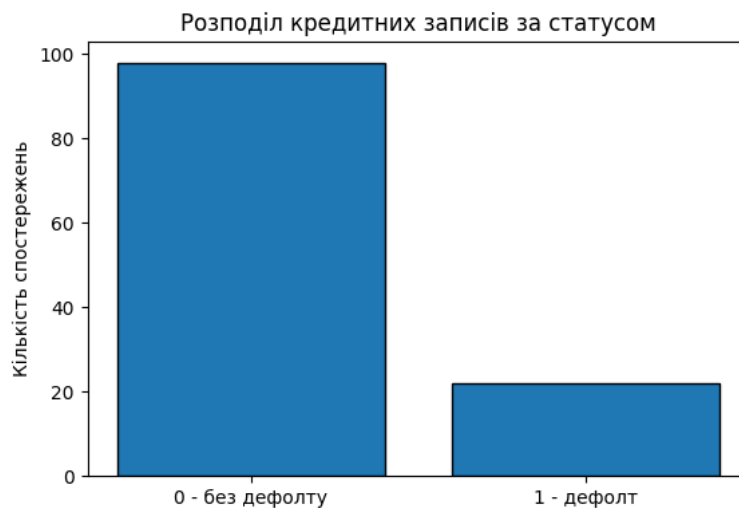


Рисунок 2.4 - Розподіл кредитних записів за статусом обслуговування. Джерело: побудовано автором у Python; код наведено в додатку А, лістинг А.4.

Рисунок 2.4 показує співвідношення між класами цільової змінної. Якщо один клас значно переважає, модель може частіше відносити записи саме до нього. Тому перед побудовою класифікатора важливо оцінити не лише загальну кількість записів, а й баланс між класами.

Взаємозв'язки між основними кредитними ознаками доцільно оцінити за кореляційною матрицею, наведеною на рисунку 2.5.

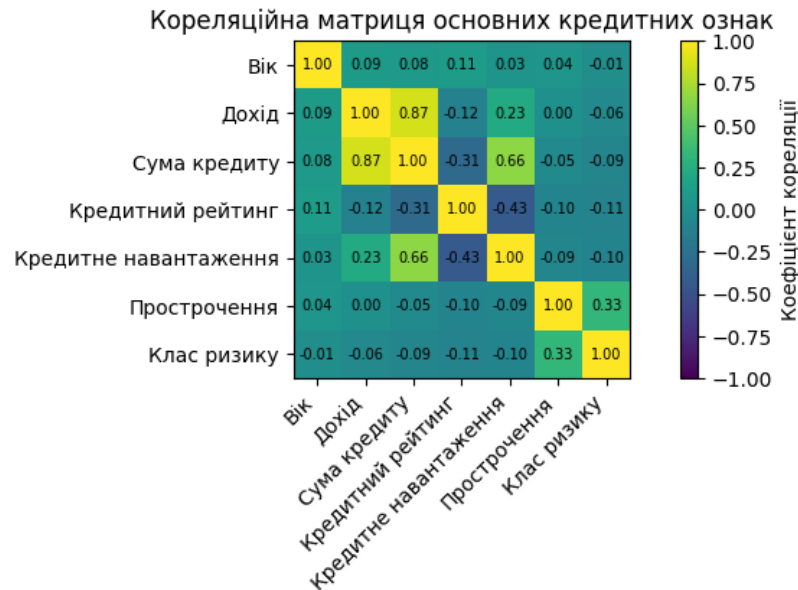


Рисунок 2.5 - Кореляційна матриця основних кредитних ознак. Джерело: побудовано автором у Python; код наведено в додатку А, лістинг А.5.

З рисунку 2.5 можна попередньо оцінити зв'язки між доходом, сумою кредиту, кредитним навантаженням, простроченнями та статусом кредиту. Якщо окремі ознаки мають сильний зв'язок між собою, це може свідчити про дублювання інформації або потребу в додатковому відборі факторів.

Призначення графічних інструментів у попередній обробці кредитних даних узагальнено в табл. 2.6.

Таблиця 2.6 - Призначення графічних інструментів у попередній обробці кредитних даних

Інструмент	Що показує	Як використовується у кредитному аналізі
Гістограма	форму розподілу числової ознаки	дозволяє оцінити асиметрію доходу, суми кредиту або віку
Діаграма розмаху	медіану, квартилі та викиди	допомагає виявити нетипові суми кредиту чи доходу
Стовпчикова діаграма	співвідношення між класами	показує співвідношення кредитів без дефолту та з дефолтним статусом

Продовження таблиці 2.6

Кореляційна матриця	силу зв'язку між ознаками	виявляє взаємозалежності між доходом, кредитним навантаженням і статусом кредиту
---------------------	---------------------------	--

Джерело: розроблено автором для узагальнення інструментів візуального контролю даних.

У табл. 2.6 узагальнено, для чого саме використовуються графічні інструменти під час підготовки кредитних даних. Гістограма допомагає оцінити розподіл числових ознак, діаграма розмаху - побачити нетипові значення, стовпчикова діаграма - перевірити баланс класів, а кореляційна матриця - визначити зв'язки між показниками.

Отже, попередня обробка дає змогу перейти від початкових банківських записів до підготовленого набору даних, який можна використовувати для побудови моделей. Саме на цьому етапі перевіряється якість інформації, усуваються технічні недоліки, приводяться до єдиного вигляду числові та категоріальні ознаки, а також оцінюються співвідношення класів. Це створює основу для подальшого аналізу логістичної регресії та дерева рішень у наступному підрозділі.

2.3 Аналіз існуючих моделей скорингу (логістична регресія, дерева рішень)

У межах цієї роботи кредитний скоринг розглядається як задача класифікації кредитних записів. Тобто на основі даних про позичальника, параметрів кредиту та кредитної історії потрібно визначити, до якого класу належить запис: кредит без дефолту або кредит із дефолтним статусом. Для бакалаврської роботи доцільно використовувати такі моделі, які мають зрозумілу логіку, можуть бути пояснені математично та не потребують надмірно складної

реалізації. Саме тому для подальшого аналізу обрано логістичну регресію та дерево рішень.

Логістична регресія є одним із базових методів бінарної класифікації. У цій роботі вона використовується для оцінювання ймовірності того, що кредитний запис буде віднесено до класу з дефолтним статусом. Перевага цієї моделі полягає в тому, що вона дає не лише готовий клас, а й числову оцінку ймовірності. Це зручно для банківської аналітики, оскільки результат можна не просто отримати, а й пояснити через вплив окремих ознак.

Ймовірність належності запису до класу з дефолтним статусом визначається за формулою:

$$p_i = P(y_i = 1 \mid x_i) = 1 / (1 + e^{(-z_i)}) \quad (2.7)$$

Джерело: розроблено автором на основі стандартної логістичної функції, що використовується у кредитному скорингу [25; 29].

$$z_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_m x_{im}, \quad (2.8)$$

де p_i - прогнозована ймовірність належності i -го запису до класу дефолту; $x_{i1}, x_{i2}, \dots, x_{im}$ - ознаки позичальника та кредиту; β_0 - вільний член; $\beta_1, \beta_2, \dots, \beta_m$ - коефіцієнти моделі.

Джерело: узагальнено автором на основі математичної постановки логістичної регресії у скорингових моделях [25; 27].

Якщо прогнозована ймовірність перевищує встановлений поріг, запис може бути віднесений до класу проблемного обслуговування. Форму логістичної функції, яка використовується для класифікації статусу кредиту, показано на рисунку 2.6.

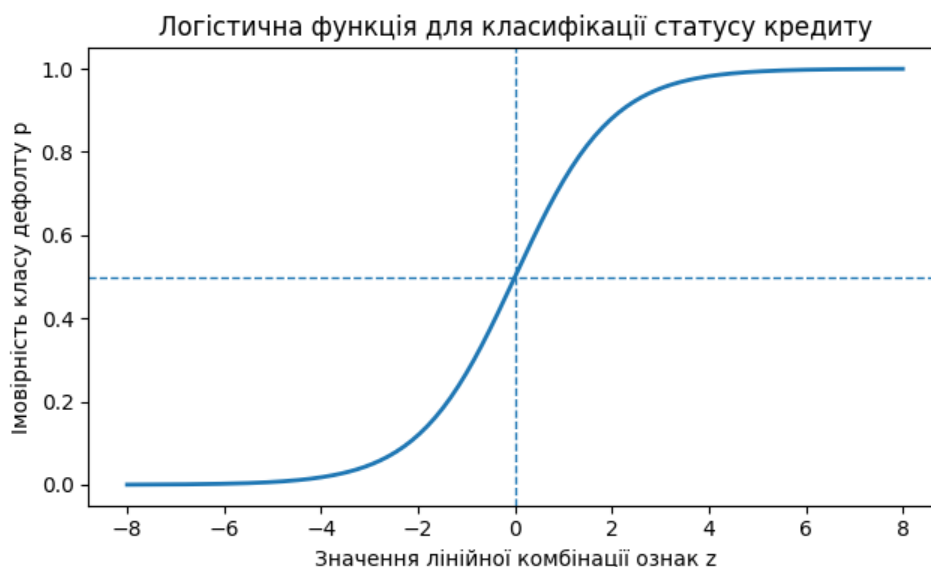


Рисунок 2.6 - Логістична функція для класифікації статусу кредиту. Джерело: побудовано автором у Python; код наведено в додатку А, лістинг А.6.

З рисунку 2.6 видно, що логістична функція переводить значення лінійної комбінації ознак у ймовірність у межах від 0 до 1. Це дає змогу використовувати модель для задачі бінарної класифікації. Наприклад, якщо ймовірність є нижчою за обраний поріг, запис можна віднести до кредитів без дефолту, а якщо вищою - до записів із дефолтним статусом.

Перевагою логістичної регресії є можливість пояснити вплив окремих факторів. Якщо коефіцієнт біля певної ознаки є додатним, то зі збільшенням цієї ознаки ймовірність дефолтного статусу зростає. Якщо коефіцієнт від'ємний, то така ймовірність зменшується. Наприклад, високе кредитне навантаження, наявність прострочення в минулому або низький кредитний рейтинг можуть підвищувати ймовірність проблемного обслуговування. Натомість стабільний дохід і позитивна кредитна історія можуть знижувати таку ймовірність.

Водночас логістична регресія має певні обмеження. Вона краще працює тоді, коли зв'язок між ознаками та результатом можна описати відносно просто. Якщо у даних є складні поєднання факторів, модель може не повністю їх врахувати. Крім того, логістична регресія чутлива до масштабу числових

показників, тому перед її застосуванням потрібно виконувати стандартизацію або нормалізацію ознак.

Альтернативним підходом є дерево рішень. Ця модель будує результат через послідовність правил типу “якщо - то”. Така логіка є зрозумілою для пояснення, оскільки рішення можна пов’язати з конкретними умовами: рівнем доходу, сумою кредиту, наявністю прострочень, кредитним рейтингом або часткою кредиту від доходу. На відміну від логістичної регресії, дерево рішень може враховувати складніші зв’язки між ознаками.

Спрощену логіку класифікації кредитного запису за допомогою дерева рішень наведено на рисунку 2.7.

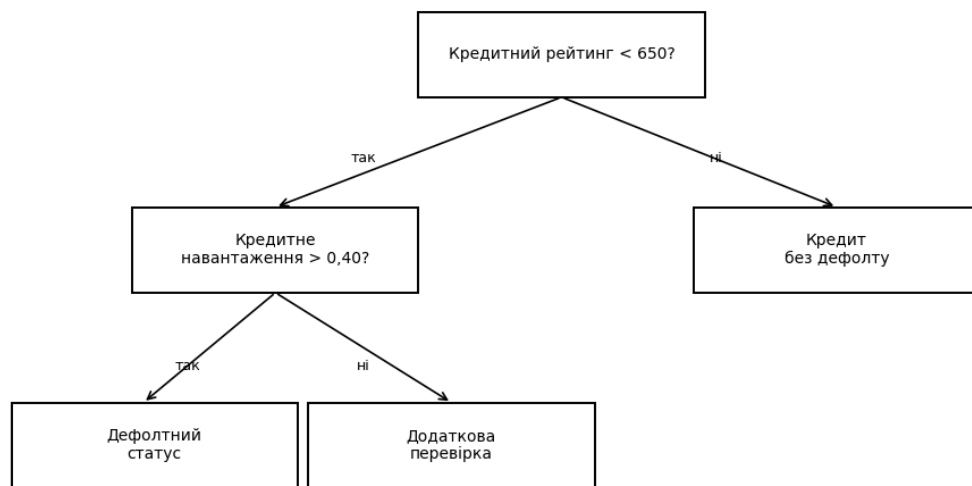


Рисунок 2.7 - Спрощений приклад логіки дерева рішень у кредитному скорингу. Джерело: побудовано автором у Python; код наведено в додатку А, лістинг А.7.

На рисунку 2.7 дерево рішень подано як послідовність перевірок окремих ознак. Модель поступово аналізує умови і в результаті відносить запис до одного з класів. Для цієї роботи такого рівня деталізації достатньо, оскільки головна мета полягає не в поглибленому вивченні всіх критеріїв побудови дерева, а в демонстрації зрозумілої логіки класифікації кредитних записів.

Для порівняння логістичної регресії та дерева рішень за основними характеристиками сформовано табл. 2.7.

Таблиця 2.7 - Порівняння логістичної регресії та дерева рішень

Критерій порівняння	Логістична регресія	Дерево рішень
Основна ідея	Розраховує ймовірність належності запису до класу дефолту	Формує послідовність правил для класифікації кредитних записів
Зручність пояснення	Пояснюється через вплив окремих факторів на статус кредиту	Пояснюється через правила “якщо - то”
Потреба в нормалізації	Бажана, оскільки модель чутлива до масштабу ознак	Не є критичною
Перевага	Дає кількісну оцінку ймовірності дефолтного статусу	Дає наочну логіку прийняття рішення
Обмеження	Гірше враховує складні поєднання ознак	Може давати нестабільні правила при малій кількості даних

Джерело: сформовано автором на основі узагальнення підходів до побудови моделей класифікації кредитних записів [25; 27; 29]

З табл. 2.7 видно, що обидві моделі можуть застосовуватися для однієї задачі, але мають різну логіку. Логістична регресія зручна тоді, коли потрібно отримати ймовірність належності запису до певного класу. Дерево рішень є більш наочним, оскільки його результат можна пояснити через конкретні правила. Для цієї роботи доцільно використати обидва підходи: логістичну регресію як базову модель, а дерево рішень - як модель із зрозумілою послідовністю умов.

Під час вибору моделі для банківської аналітики важливо враховувати не лише точність, а й можливість пояснити отриманий результат. У банківській сфері це має особливе значення, оскільки рішення повинні бути зрозумілими для фахівців і придатними для подальшої перевірки. Тому складніші алгоритми можуть давати кращу якість, але не завжди є зручними для пояснення.

Узагальнення груп моделей, що можуть застосовуватися для обробки кредитної інформації банку, наведено в табл. 2.8.

Таблиця 2.8 - Узагальнення практичних характеристик моделей машинного навчання у кредитній аналітиці

Група моделей	Приклад	Сильна сторона	Обмеження для банківського застосування
Класичні статистичні моделі	логістична регресія	простота реалізації та зрозуміла інтерпретація результатів	гірше враховують складні нелінійні зв'язки між ознаками
Дерева рішень	Decision Tree	наочні правила класифікації, які легко пояснити	можуть бути нестабільними при зміні вибірки
Ансамблеві методи	Random Forest, XGBoost	часто забезпечують вищу якість класифікації	складніші для пояснення та налаштування
Нейронні мережі	багатошарові моделі	можуть працювати зі складними залежностями та великими масивами даних	потребують значного обсягу даних і мають нижчу прозорість результатів

Джерело: сформовано автором на основі узагальнення порівняльних характеристик моделей машинного навчання у банківській сфері [39].

З табл. 2.8 видно, що в кредитній аналітиці можуть використовуватися різні групи моделей. Складні алгоритми, наприклад ансамблеві методи або нейронні мережі, можуть давати високу якість класифікації, але їх складніше пояснювати та налаштовувати. Для цієї бакалаврської роботи більш доцільними є логістична регресія та дерево рішень, оскільки вони достатньо прості для реалізації, зрозумілі для інтерпретації та добре підходять для демонстрації основних етапів класифікації кредитних записів.

РОЗДІЛ 3

РОЗРОБКА МОДЕЛІ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ КРЕДИТНИХ ДАНИХ

3.1 Характеристика інформаційної бази та попередня підготовка кредитних даних

Практичний розділ кваліфікаційної роботи присвячено застосуванню методів інтелектуального аналізу до кредитного набору даних. Основна увага приділяється тому, як із початкового датасету отримати підготовлену таблицю, придатну для подальшої побудови моделей класифікації. У цьому підрозділі розглянуто джерело даних, їх структуру, первинну перевірку та очищення.

Під час проходження переддипломної практики в АТ «КРЕДІ АГРІКОЛЬ БАНК» було отримано досвід роботи з інформацією, пов'язаною з кредитними операціями корпоративного бізнесу. Зокрема, практична робота стосувалася перевірки даних щодо кредитів, страхування, заставного забезпечення, лізингових угод, а також підготовки матеріалів для внутрішнього аналізу й аудиту. Значна частина завдань виконувалася у Microsoft Excel та внутрішній банківській програмі iOrday. Це дало змогу зрозуміти, що перед будь-яким аналізом кредитну інформацію потрібно перевірити, очистити й привести до зручної структури.

У практичній частині роботи кожен запис розглядається як окреме спостереження, а його характеристики - як вхідні ознаки для моделі. Такий підхід відповідає логіці табличної обробки даних: рядки описують окремі кредитні записи, а стовпці містять показники, які можуть бути використані для подальшого аналізу.

Для практичного дослідження обрано відкритий набір даних Credit Risk Dataset, розміщений на платформі Kaggle. Він містить інформацію про

позичальників, параметри кредиту, кредитну історію та статус кредиту. Цей датасет є доцільним для теми роботи, оскільки дозволяє показати повний процес роботи з кредитною інформацією: від завантаження та первинної перевірки до очищення, підготовки ознак і побудови моделей класифікації.

Використання відкритого набору даних є важливим з погляду інформаційної безпеки. Реальні банківські дані містять персональні відомості клієнтів, умови договорів, інформацію про забезпечення, страхування та інші конфіденційні дані. Тому у роботі використано публічний датасет, структура якого наближена до типових задач кредитної аналітики, але не містить внутрішньої інформації конкретного банку.

Початковий набір даних складається з 32 581 спостереження та 12 змінних. Цільовою змінною є `loan_status`, де значення 0 відповідає кредиту без дефолту, а значення 1 - кредиту з дефолтним статусом. За початковим розподілом у датасеті переважають кредити без дефолту: їх кількість становить 25 473 записи. Кредитів із дефолтним статусом менше - 7 108 записів. Це свідчить про нерівномірний розподіл класів, який потрібно враховувати під час подальшого аналізу результатів моделей.

Практична реалізація виконувалася мовою Python у середовищі Visual Studio Code. Такий вибір є зручним, оскільки Python дозволяє працювати з табличними даними, виконувати очищення, будувати графіки, створювати моделі та повторно запускати код у разі потреби. Visual Studio Code використано як середовище для написання та перевірки програмного коду.

На першому етапі було виконано завантаження даних у програмне середовище для подальшої обробки.

Лістинг 3.1 - Завантаження `damacemy Credit Risk Dataset` у Python

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
url = "https://raw.githubusercontent.com/sangramdedge/Credit-Risk-Analysis-Prediction-Dashboard/main/credit_risk_dataset.csv"
df = pd.read_csv(url)
print("Розмір датасету:", df.shape)
```

```
print(df.head())
df.info()
```

Після виконання лістингу 3.1 було завантажено початковий набір кредитних даних. Він містить 32 581 запис і 12 змінних, що є достатнім обсягом для практичного аналізу. Датасет поєднує числові та категоріальні ознаки. До числових належать, наприклад, вік позичальника, річний дохід, сума кредиту, відсоткова ставка та тривалість кредитної історії. До категоріальних належать тип володіння житлом, мета кредиту, кредитний клас і наявність попереднього дефолту.

Для обробки табличних даних використано бібліотеку `pandas`, для числових операцій - `numpy`, для побудови графіків - `matplotlib`, а для побудови моделей - інструменти бібліотеки `scikit-learn`. Характеристику змінних Credit Risk Dataset подано в табл. 3.1.

Таблиця 3.1 - Характеристика змінних Credit Risk Dataset

Назва змінної	Зміст показника	Тип змінної	Роль у моделі
person_age	вік позичальника	числова	вхідна ознака
person_income	річний дохід позичальника	числова	вхідна ознака
person_home_ownership	тип володіння житлом	категоріальна	вхідна ознака
person_emp_length	стаж роботи в роках	числова	вхідна ознака
loan_intent	мета отримання кредиту	категоріальна	вхідна ознака
loan_grade	кредитний клас	категоріальна	вхідна ознака
loan_amnt	сума кредиту	числова	вхідна ознака
loan_int_rate	відсоткова ставка за кредитом	числова	вхідна ознака
loan_percent_income	частка кредиту від доходу	числова	вхідна ознака
cb_person_default_on_file	наявність попереднього дефолту	категоріальна	вхідна ознака
cb_person_cred_hist_length	тривалість кредитної історії	числова	вхідна ознака
loan_status	статус кредиту	бінарна	цільова змінна

Джерело: сформовано автором на основі структури Credit Risk Dataset.

З табл. 3.1 видно, що датасет містить три основні групи інформації: характеристики позичальника, параметри кредиту та елементи кредитної історії.

Для подальшого моделювання найважливішою є змінна `loan_status`, оскільки саме вона використовується як результат, який модель має навчитися визначати.

Наступним етапом була первинна перевірка даних. Вона потрібна для того, щоб виявити пропущені значення, дублікати та оцінити розподіл цільової змінної перед очищенням датасету.

Лістинг 3.2 - Первинна перевірка даних датасету Credit Risk Dataset у Python

```
print("Кількість пропущених значень:")
print(df.isnull().sum())
print("\nКількість дублікатів:", df.duplicated().sum())
print("\nРозподіл цільової змінної loan_status:")
print(df["loan_status"].value_counts())
print("\nЧастки класів, %:")
print((df["loan_status"].value_counts(normalize=True) * 100).round(2))
```

Результат виконання коду:

Кількість пропущених значень:

```
person_age          0
person_income       0
person_home_ownership  0
person_emp_length   895
loan_intent          0
loan_grade          0
loan_amnt           0
loan_int_rate       3116
loan_status         0
loan_percent_income  0
cb_person_default_on_file  0
cb_person_cred_hist_length  0
dtype: int64
```

Кількість дублікатів: 165

Розподіл цільової змінної `loan_status`:

```
loan_status
```

```
0    25473
```

```
1     7108
```

```
Name: count, dtype: int64
```

Частки класів, %:

```
loan_status
```

```
0    78.18
```

```
1    21.82
```

```
Name: proportion, dtype: float64
```

Результати виконання лістингу 3.2 показали, що у датасеті є пропущені значення у двох змінних: `person_emp_length` - 895 пропусків і `loan_int_rate` - 3116 пропусків. Це означає, що перед побудовою моделей ці показники потрібно обробити, оскільки алгоритми машинного навчання не можуть коректно працювати з порожніми значеннями. Також було знайдено 165 дубльованих записів, які доцільно видалити, щоб однакові спостереження не впливали на результат моделювання.

Розподіл цільової змінної `loan_status` показав, що у початковому наборі 25 473 записи належать до класу без дефолту, а 7 108 - до класу з дефолтним статусом. У відсотковому співвідношенні це становить 78,18 % та 21,82 % відповідно. Отже, у датасеті переважають кредити без дефолту. Такий розподіл потрібно враховувати під час оцінювання моделей, оскільки висока загальна точність не завжди означає якісне розпізнавання менш представленого класу.

Після первинної перевірки було виконано очищення даних. На цьому етапі видалено дублікати, заповнено пропущені значення медіаною та вилучено очевидно нелогічні значення віку й стажу роботи. Така підготовка потрібна для того, щоб подальший аналіз виконувався на повному й логічно коректному наборі спостережень.

Лістинг 3.3 - Очищення даних `damacemy Credit Risk Dataset` у Python

```
df_clean = df.copy()
# Видалення дублікатів
df_clean = df_clean.drop_duplicates()
# Заповнення пропусків медіаною
```

```
df_clean["person_emp_length"] =
df_clean["person_emp_length"].fillna(df_clean["person_emp_length"].median())
df_clean["loan_int_rate"] =
df_clean["loan_int_rate"].fillna(df_clean["loan_int_rate"].median())
# Видалення очевидно нелогічних значень
df_clean = df_clean[df_clean["person_age"] <= 100]
df_clean = df_clean[df_clean["person_emp_length"] <= 60]
print("Розмір очищеного датасету:", df_clean.shape)
print("Пропуски після очищення:")
print(df_clean.isnull().sum())
```

Результат виконання коду:

Розмір очищеного датасету: (32409, 12)

Пропуски після очищення:

```
person_age          0
person_income       0
person_home_ownership  0
person_emp_length   0
loan_intent         0
loan_grade         0
loan_amnt          0
loan_int_rate      0
loan_status        0
loan_percent_income  0
cb_person_default_on_file  0
cb_person_cred_hist_length  0
dtype: int64
```

Після виконання лістингу 3.3 розмір очищеного датасету становив 32 409 записів і 12 змінних. Порівняно з початковим набором кількість записів зменшилася з 32 581 до 32 409. Це пов'язано з видаленням дублікатів та очевидно нелогічних значень. При цьому кількість змінних залишилася незмінною, тобто структура датасету була збережена.

Після заповнення пропусків у змінних `person_emp_length` та `loan_int_rate` у датасеті не залишилося порожніх значень. Для заповнення було використано медіану, оскільки вона менш чутлива до викидів, ніж середнє арифметичне. Це

особливо важливо для фінансових показників, де можуть траплятися нетипово високі або низькі значення.

У результаті очищення було отримано підготовлений набір даних, який можна використовувати для наступних етапів роботи: кодування категоріальних ознак, стандартизації числових змінних, поділу вибірки та побудови моделей класифікації. Результати первинної перевірки та очищення датасету узагальнено в табл. 3.2.

Таблиця 3.2 - Результати первинної перевірки та очищення датасету

Показник	До очищення	Після очищення	Висновок
Кількість записів	32 581	32 409	частину записів видалено через дублікати та нелогічні значення
Кількість змінних	12	12	структура датасету збережена
Пропуски в person_emp_length	895	0	пропуски заповнено медіаною
Пропуски в loan_int_rate	3116	0	пропуски заповнено медіаною
Кількість дублікатів	165	0	повторювані записи видалено
Частка кредитів без дефолту	78,18 %	-	виявлено переважання надійного класу
Частка кредитів із дефолтом	21,82 %	-	ризиковий клас є менш представленим

Джерело: розраховано автором на основі проведених етапів підготовки даних у Python.

Дані табл. 3.2 показують, що після очищення датасет став придатним для подальшого аналізу. Було усунуто пропущені значення у змінних, що описують стаж роботи та відсоткову ставку, видалено дублікати й збережено початкову структуру змінних. Водночас розподіл класів показав, що записи з дефолтним статусом становлять меншу частину вибірки. Тому під час подальшого оцінювання моделей потрібно звертати увагу не лише на загальну точність, а й на те, як модель визначає саме цей клас.

3.2 Побудова аналітичної моделі обробки кредитних даних

Після очищення датасету можна перейти до підготовки даних для побудови моделей класифікації. У підрозділі 3.1 було сформовано очищений набір `df_clean`, у якому видалено дублікати, заповнено пропущені значення та вилучено очевидно нелогічні записи. Тому наступним кроком є формування вхідних ознак, визначення цільової змінної, поділ вибірки та побудова моделей.

Під час переддипломної практики в АТ «КРЕДІ АГРІКОЛЬ БАНК» робота з даними виконувалася переважно в Microsoft Excel та внутрішній банківській програмі iOpday. Це дало змогу зрозуміти, що перед аналізом будь-яку інформацію потрібно перевірити, очистити та подати у зручному вигляді. У практичній частині дипломної роботи цей самий підхід реалізовано засобами Python.

Для моделювання обрано дві моделі, які були розглянуті у підрозділі 2.3: логістичну регресію та дерево рішень. Логістична регресія використовується як базова модель бінарної класифікації. Дерево рішень застосовується як модель, що дозволяє подати результат через послідовність зрозумілих умов. Обидві моделі використовуються для класифікації кредитних записів за статусом обслуговування.

Спочатку потрібно відокремити ознаки, на основі яких модель буде навчатися, від результату, який потрібно прогнозувати. До вхідних ознак належать характеристики позичальника, параметри кредиту та елементи кредитної історії. Цільовою змінною є `loan_status`, яка показує статус обслуговування кредиту: 0 - кредит без дефолту, 1 - кредит із дефолтним статусом.

Лістинг 3.4 - Формування вхідних ознак і цільової змінної

```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
X = df_clean.drop("loan_status", axis=1)
y = df_clean["loan_status"]
data_structure = pd.DataFrame({
    "Елемент даних": [
        "Очищений датасет",
        "Матриця вхідних ознак X",
```

```

"Цільова змінна y"
],
"Кількість записів": [
df_clean.shape[0],
X.shape[0],
y.shape[0]
],
"Кількість змінних": [
df_clean.shape[1],
X.shape[1],
1
],
"Призначення": [
"повний набір даних після очищення",
"характеристики позичальника та кредиту",
"статус обслуговування кредиту"
]
})
print(data_structure.to_string(index=False))

```

Результат виконання коду:

Елемент даних	Кількість записів	Кількість змінних	Призначення
Очищений датасет	32409	12	повний набір даних після очищення
Матриця вхідних ознак X	32409	11	характеристики позичальника та кредиту
Цільова змінна y	32409	1	статус обслуговування кредиту

У результаті виконання лістингу 3.4 очищений датасет було поділено на матрицю вхідних ознак X та цільову змінну y. Повний очищений набір містить 32 409 записів і 12 змінних. Після відокремлення змінної `loan_status` у матриці X залишилося 11 ознак, які описують позичальника, умови кредиту та кредитну історію. Цільова змінна y містить 32 409 значень і використовується як результат, який має визначати модель.

Отже, одна змінна з датасету використовується як цільова, а решта - як вхідні характеристики для навчання. Оскільки серед цих характеристик є не лише числові, а й текстові значення, наступним кроком є перетворення категоріальних змінних у числовий формат.

Лістинг 3.5 - Кодування категоріальних змінних

```
X_encoded = pd.get_dummies(X, drop_first=True)
print("Розмір таблиці після кодування ознак:")
print(X_encoded.shape)
```

Результат виконання коду:

Розмір таблиці після кодування ознак:

(32409, 22)

У лістингу 3.5 категоріальні змінні було перетворено у числовий вигляд за допомогою функції `pd.get_dummies()`. Це потрібно тому, що моделі не можуть напряму працювати з текстовими значеннями. Наприклад, тип володіння житлом, мета кредиту, кредитний клас або наявність попереднього дефолту мають бути подані у вигляді числових індикаторів.

Після кодування кількість ознак збільшилася з 11 до 22. Це є нормальним результатом, оскільки текстові категорії були розкладені на окремі числові стовпці. Після цього таблиця `X_encoded` стала придатною для подальшого використання у моделях. Наступним етапом є поділ підготовлених даних на навчальну та тестову вибірки.

Лістинг 3.6 - Поділ датасету на навчальну та тестову вибірки

```
X_train, X_test, y_train, y_test = train_test_split(
X_encoded,
y,
test_size=0.2,
random_state=42,
stratify=y
)
print("Розмір навчальної вибірки:")
print(X_train.shape)
print("Розмір тестової вибірки:")
print(X_test.shape)
```

Результат виконання коду:

Розмір навчальної вибірки:

(25927, 22)

Розмір тестової вибірки:

(6482, 22)

Після виконання лістингу 3.6 дані було поділено на дві частини. Навчальна вибірка містить 25 927 записів і використовується для побудови моделей. Тестова вибірка містить 6 482 записи та потрібна для перевірки того, як моделі працюють із даними, які не брали участі у навчанні.

У коді використано параметр `stratify=y`. Це зроблено для того, щоб у навчальній і тестовій вибірках зберегти приблизно однакове співвідношення кредитів без дефолту та кредитів із дефолтним статусом. Такий підхід є важливим, оскільки у підрозділі 3.1 було встановлено, що клас із дефолтним статусом представлений меншою кількістю записів. Після поділу вибірки можна переходити до побудови моделей.

Лістинг 3.7 - Побудова логістичної регресії та дерева рішень

```
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
log_model = LogisticRegression(
    max_iter=1000,
    class_weight="balanced",
    random_state=42
)
log_model.fit(X_train_scaled, y_train)
y_pred_log = log_model.predict(X_test_scaled)
tree_model = DecisionTreeClassifier(
    max_depth=4,
    class_weight="balanced",
    random_state=42
)
tree_model.fit(X_train, y_train)
y_pred_tree = tree_model.predict(X_test)
```

У лістингу 3.7 було побудовано дві моделі: логістичну регресію та дерево рішень. Перед навчанням логістичної регресії виконано стандартизацію ознак, оскільки ця модель чутлива до масштабу числових значень. Наприклад, дохід і сума кредиту мають значно більші числові значення, ніж відсоткова ставка або частка кредиту від доходу, тому їх потрібно привести до порівнюваного вигляду.

Дерево рішень побудовано без стандартизації, оскільки для цього алгоритму масштаб ознак не має такого значення. Після навчання обидві моделі сформуvalи прогнози для тестової вибірки: `y_pred_log` - для логістичної регресії, `y_pred_tree` - для дерева рішень. Щоб попередньо побачити, як моделі розподілили записи між класами, далі сформовано узагальнення прогнозів.

Лістинг 3.8 - Узагальнення сформованих прогнозів

```
prediction_summary = pd.DataFrame({
    "Модель": [
        "Логістична регресія",
        "Дерево рішень"
    ],
    "Кількість прогнозів": [
        len(y_pred_log),
        len(y_pred_tree)
    ],
    "Прогноз: кредит без дефолту": [
        (y_pred_log == 0).sum(),
        (y_pred_tree == 0).sum()
    ],
    "Прогноз: кредит із дефолтним статусом": [
        (y_pred_log == 1).sum(),
        (y_pred_tree == 1).sum()
    ]
})
print(prediction_summary.to_string(index=False))
```

Розподіл прогнозів, отриманих логістичною регресією та деревом рішень, подано в таблиці 3.3.

Таблиця 3.3 - Узагальнення прогнозів моделей класифікації

Модель	Кількість прогнозів	Без дефолту	З дефолтним статусом
Логістична регресія	6 482	4 430	2 052
Дерево рішень	6 482	5 096	1 386

Джерело: розраховано автором у Python.

З табл. 3.3 видно, що обидві моделі сформуvalи прогнози для всіх 6 482 записів тестової вибірки. Логістична регресія частіше відносить записи до класу з

дефолтним статусом: 2 052 записи проти 1 386 у дерева рішень. Натомість дерево рішень частіше класифікує записи як кредити без дефолту: 5 096 записів проти 4 430 у логістичної регресії.

Ця таблиця показує лише розподіл прогнозів, але ще не дозволяє визначити, яка модель працює краще. Для цього потрібно порівняти прогнозні значення з фактичними значеннями `loan_status`. Таке порівняння буде виконано в підрозділі 3.3 за допомогою матриці помилок і показників якості моделей.

Отже, у підрозділі 3.2 було виконано перехід від очищеного датасету до побудови моделей класифікації. Для цього сформовано вхідні ознаки та цільову змінну, перетворено категоріальні показники у числовий формат, поділено дані на навчальну й тестову вибірки, а також побудовано логістичну регресію та дерево рішень. Отримані прогнози стали основою для наступного етапу - оцінювання якості моделей та інтерпретації результатів у підрозділі 3.3.

3.3 Візуалізація та інтерпретація результатів інтелектуального аналізу даних

Після побудови логістичної регресії та дерева рішень потрібно перевірити, наскільки правильно ці моделі класифікують записи тестової вибірки. У підрозділі 3.2 було отримано прогнози `y_pred_log` та `y_pred_tree`, але самі прогнози ще не показують якість моделей. Щоб оцінити результат, їх потрібно порівняти з фактичними значеннями `y_test`.

Для оцінювання використано два основні інструменти: загальну точність і матрицю помилок. Загальна точність показує, яку частку записів модель визначила правильно. Матриця помилок дає змогу детальніше побачити, які саме класи модель розпізнає краще, а де допускає помилки.

Лістинг 3.9 - Розрахунок загальної точності та матриць помилок

```

from sklearn.metrics import accuracy_score
from sklearn.metrics import confusion_matrix
from sklearn.metrics import ConfusionMatrixDisplay
log_accuracy = accuracy_score(y_test, y_pred_log)
tree_accuracy = accuracy_score(y_test, y_pred_tree)
log_cm = confusion_matrix(y_test, y_pred_log)
tree_cm = confusion_matrix(y_test, y_pred_tree)
print("Точність логістичної регресії:", round(log_accuracy, 3))
print("Матриця помилок логістичної регресії:")
print(log_cm)
print("Точність дерева рішень:", round(tree_accuracy, 3))
print("Матриця помилок дерева рішень:")
print(tree_cm)

```

Результат виконання коду:

Точність логістичної регресії: 0.807

Матриця помилок логістичної регресії:

```
[[4120 944]
```

```
[ 310 1108]]
```

Точність дерева рішень: 0.891

Матриця помилок дерева рішень:

```
[[4728 336]
```

```
[ 368 1050]]
```

У результаті виконання лістингу 3.9 було розраховано точність і матриці помилок для двох моделей. Логістична регресія показала точність 0,807, а дерево рішень - 0,891. Отже, за загальною точністю кращий результат отримано для дерева рішень. Водночас логістична регресія правильно визначила трохи більше записів із дефолтним статусом: 1 108 проти 1 050 у дерева рішень. Тому для порівняння моделей важливо враховувати не лише загальну точність, а й окремі типи помилок.

Щоб результати було зручно порівняти, основні значення з матриць помилок зведено в одну таблицю. Формування такої таблиці показано в лістингу 3.10.

Лістинг 3.10 - Формування підсумкової таблиці результатів

```
results_table = pd.DataFrame({
```

```

"Модель": ["Логістична регресія", "Дерево рішень"],
"Точність": [round(log_accuracy, 3), round(tree_accuracy, 3)],
"Без дефолту (вірно)": [log_cm[0, 0], tree_cm[0, 0]],
"Без дефолту -> дефолт": [log_cm[0, 1], tree_cm[0, 1]],
"Дефолт -> без дефолту": [log_cm[1, 0], tree_cm[1, 0]],
"Дефолт (вірно)": [log_cm[1, 1], tree_cm[1, 1]]
})
print(results_table.to_string(index=False))

```

Після запуску лістингу 3.10 результати логістичної регресії та дерева рішень подано в однаковому форматі. Це дозволяє порівняти моделі не лише за точністю, а й за тим, які саме помилки вони допускають. Порівняння результатів двох моделей наведено в табл. 3.4.

Таблиця 3.4 - Порівняння результатів логістичної регресії та дерева рішень

Модель	Загальна точність	Правильно визначено без дефолту	Помилково віднесено до дефолтних	Пропущено дефолтних	Правильно визначено дефолтних
Логістична регресія	0,807	4 120	944	310	1 108
Дерево рішень	0,891	4 728	336	368	1 050

Джерело: розраховано автором у Python.

З табл. 3.4 видно, що дерево рішень має вищу загальну точність і краще визначає кредити без дефолту. Воно правильно класифікувало 4 728 таких записів, тоді як логістична регресія - 4 120. Також дерево рішень значно рідше помилково відносило кредити без дефолту до дефолтних.

Разом з тим логістична регресія трохи краще розпізнала записи з дефолтним статусом: 1 108 правильних визначень проти 1 050 у дерева рішень. Це означає, що дерево рішень краще працює за загальною точністю, але для аналізу дефолтного класу результати логістичної регресії також мають практичне значення.

Для кращого пояснення результатів моделей важливо враховувати не тільки точність, а й те, за рахунок яких показників формується класифікація кредитних записів. Тобто потрібно пояснити, які характеристики можуть впливати на те, чи буде запис віднесено до класу без дефолту або до класу з дефолтним статусом. У

використаному Credit Risk Dataset для цього найбільше значення мають показники, що описують дохід позичальника, умови кредиту та його кредитну історію.

Одним із важливих показників є `loan_percent_income`, який відображає частку кредиту від доходу позичальника. Якщо ця частка є високою, це означає, що кредит займає значну частину доходу клієнта. У такому випадку обслуговувати кредит може бути складніше, тому запис частіше може бути віднесений моделлю до класу з дефолтним статусом.

Також важливим є показник `person_income`, тобто річний дохід позичальника. Вищий дохід може свідчити про більші можливості клієнта виконувати кредитні зобов'язання. Проте дохід не можна розглядати окремо від інших показників, тому що навіть за високого доходу велика сума кредиту або висока відсоткова ставка можуть збільшувати фінансове навантаження.

Показники `loan_amnt` та `loan_int_rate` характеризують умови самого кредиту. `loan_amnt` показує суму кредиту, а `loan_int_rate` - відсоткову ставку. Чим більшою є сума кредиту та чим вищою є ставка, тим більшим може бути навантаження на позичальника. Тому ці ознаки є важливими для пояснення того, чому окремі записи модель відносить до проблемного класу.

Окремо слід враховувати показники кредитної історії: `cb_person_default_on_file` та `cb_person_cred_hist_length`. Перший показник відображає наявність попереднього дефолту, тобто минулий досвід проблемного обслуговування кредитів. Другий показник характеризує тривалість кредитної історії. Якщо історія довша, банк має більше інформації про поведінку клієнта. Якщо ж вона коротка, оцінювання може бути менш повним.

Отже, результати моделей потрібно розглядати не тільки через точність, а й через зміст самих ознак. Найбільш важливими для пояснення класифікації є кредитне навантаження, дохід позичальника, сума кредиту, відсоткова ставка, кредитна історія та наявність попереднього дефолту. Саме ці показники допомагають зрозуміти, чому певний кредитний запис може бути віднесений до одного з двох класів.

Лістинг 3.11 - Візуалізація матриць помилок логістичної регресії

```
plt.figure(figsize=(5, 4))
plt.imshow(log_cm)
plt.title("Матриця помилок: логістична регресія")
plt.xlabel("Прогноз")
plt.ylabel("Фактичний клас")
plt.xticks([0, 1], ["Без дефолту", "Дефолт"])
plt.yticks([0, 1], ["Без дефолту", "Дефолт"])
for i in range(2):
    for j in range(2):
        plt.text(j, i, log_cm[i, j], ha="center", va="center")
plt.colorbar()
plt.tight_layout()
plt.show()
```

Для наочного подання помилок логістичної регресії побудовано матрицю помилок, наведену на рисунку 3.1.



Рисунок 3.1 - Матриця помилок: логістична регресія. Джерело: розроблено автором у Python.

З рис. 3.1 видно, що логістична регресія правильно визначила значну частину кредитів без дефолту та більшість записів із дефолтним статусом. Водночас модель допускає помилки в обох напрямках: частину кредитів без дефолту вона відносить до дефолтних, а частину дефолтних записів не виявляє. Це означає, що модель може бути корисною для попередньої класифікації, але її результати потрібно аналізувати разом із матрицею помилок.

Лістинг 3.12 - Візуалізація матриць помилок дерева рішень

```
plt.figure(figsize=(5, 4))
plt.imshow(tree_cm)
plt.title("Матриця помилок: дерево рішень")
plt.xlabel("Прогноз")
plt.ylabel("Фактичний клас")
plt.xticks([0, 1], ["Без дефолту", "Дефолт"])
plt.yticks([0, 1], ["Без дефолту", "Дефолт"])
for i in range(2):
    for j in range(2):
        plt.text(j, i, tree_cm[i, j], ha="center", va="center")
plt.colorbar()
plt.tight_layout()
plt.show()
```

Для наочного подання помилок помилок дерева рішень для тестової вибірки побудовано матрицю, наведену на рисунку 3.2.

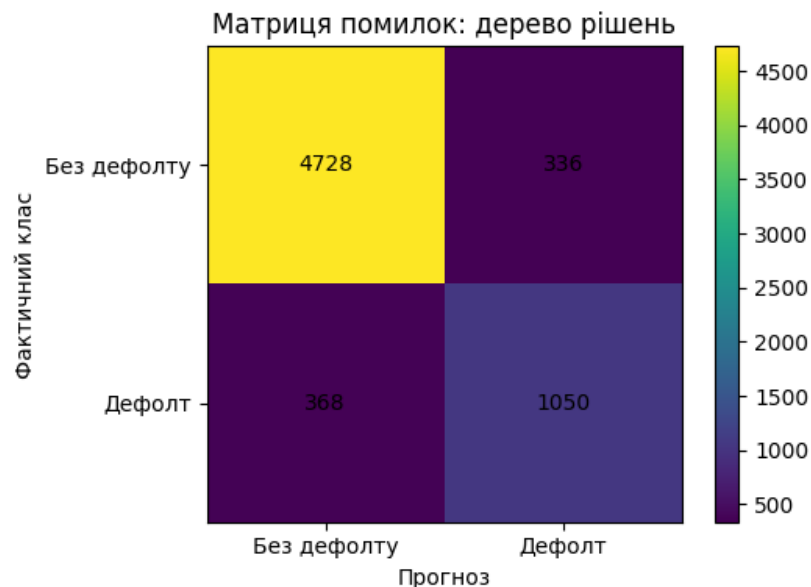


Рисунок 3.2 - Матриця помилок: дерево рішень. Джерело: розроблено автором у Python.

З рис. 3.2 видно, що дерево рішень краще розпізнає кредити без дефолту та робить менше помилок при віднесенні таких записів до дефолтного класу. Разом з тим ця модель пропускає трохи більше дефолтних записів, ніж логістична регресія. Тому дерево рішень має кращий загальний результат, але під час аналізу дефолтного класу потрібно враховувати і його помилки.

Лістинг 3.13 - Візуалізація матриць помилок дерева рішень

```
models = ["Логістична регресія", "Дерево рішень"]  
accuracy_values = [0.807, 0.891]  
plt.figure(figsize=(7, 5))  
plt.bar(models, accuracy_values)  
plt.title("Порівняння моделей за загальною точністю")  
plt.ylabel("Загальна точність")  
plt.ylim(0, 1)  
plt.grid(axis="y", alpha=0.3)  
plt.tight_layout()  
plt.show()
```

Порівняння моделей за загальною точністю показано на рисунку 3.3.

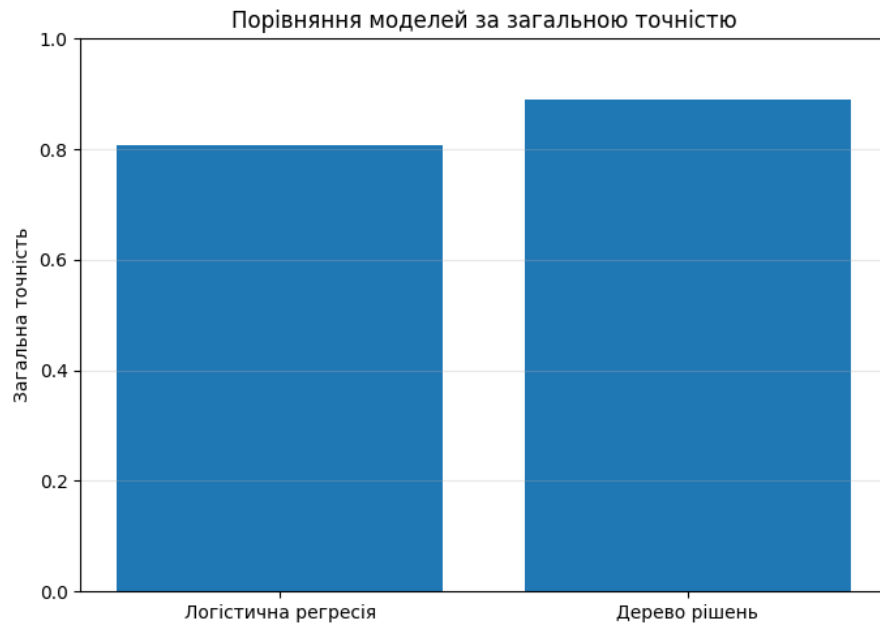


Рисунок 3.3 - Порівняння моделей за загальною точністю. Джерело: розроблено автором у Python.

З рис. 3.3 видно, що дерево рішень має вищу загальну точність порівняно з логістичною регресією. Це означає, що на тестовій вибірці дерево рішень частіше давало правильний результат. Однак остаточну оцінку моделі не можна робити лише за цим показником, оскільки важливо також розуміти, як модель працює з кожним класом окремо.

У третьому розділі було виконано практичну частину роботи: завантажено Credit Risk Dataset, перевірено його структуру, виявлено пропуски, дублікати та

розподіл цільової змінної. Після очищення датасет містив 32 409 записів і 12 змінних, що дозволило перейти до підготовки даних для моделювання.

На наступному етапі було відокремлено вхідні ознаки від цільової змінної `loan_status`, виконано кодування категоріальних показників і поділ даних на навчальну та тестову вибірки. Для класифікації кредитних записів побудовано дві моделі: логістичну регресію та дерево рішень.

Оцінювання моделей показало, що дерево рішень має вищу загальну точність - 0,891, тоді як логістична регресія показала точність 0,807. Водночас логістична регресія трохи краще визначила записи з дефолтним статусом. Отже, дерево рішень можна вважати кращим за загальною точністю, але результати обох моделей є корисними для попередньої аналітичної обробки кредитної інформації.

Таким чином, у розділі 3 було показано повний шлях роботи з даними: від очищення датасету до побудови моделей і оцінювання їх результатів. Побудовані моделі не замінюють рішення банківського фахівця, але можуть використовуватися як допоміжний інструмент для класифікації кредитних записів і підготовки подальших висновків.

ВИСНОВКИ

У кваліфікаційній бакалаврській роботі було досліджено підходи до інтелектуального аналізу кредитних даних банку. Встановлено, що кредитна інформація є важливим ресурсом банківської діяльності, оскільки використовується для аналізу позичальників, супроводження кредитних операцій, контролю якості даних і підготовки аналітичних висновків.

У теоретичній частині роботи розглянуто сутність кредитного скорингу, особливості кредитних даних та основні методи їх обробки. Визначено, що кредитні набори даних зазвичай містять числові й категоріальні ознаки, тому перед побудовою моделей їх потрібно перевірити, очистити та привести до придатного для аналізу вигляду.

У роботі також було враховано досвід переддипломної практики в АТ «КРЕДІ АГРІКОЛЬ БАНК». Практика показала, що робота з банківською інформацією починається з перевірки, упорядкування та підготовки даних. Саме тому в дослідженні значну увагу приділено етапу попередньої обробки: роботі з пропусками, дублікатами, кодуванню ознак, стандартизації та аналізу розподілу класів.

У практичній частині використано відкритий Credit Risk Dataset, що дозволило виконати дослідження без використання конфіденційної банківської інформації. Після завантаження датасету було перевірено його структуру, виявлено пропущені значення та дублікати, після чого виконано очищення даних. Підготовлений набір містив 32 409 записів і 12 змінних.

Для класифікації кредитних записів було побудовано дві моделі: логістичну регресію та дерево рішень. За результатами оцінювання дерево рішень показало вищу загальну точність - 0,891, тоді як точність логістичної регресії становила 0,807. Водночас логістична регресія трохи краще виявила записи з дефолтним статусом, тому під час порівняння моделей важливо враховувати не лише точність, а й матрицю помилок.

Практичне значення роботи полягає в тому, що запропонована послідовність дій може бути використана як приклад обробки кредитної інформації: від первинної перевірки та очищення даних до побудови моделей і аналізу їх результатів. Такий підхід може бути корисним для попередньої класифікації кредитних записів і підготовки аналітичних матеріалів.

Отже, мету роботи досягнуто. У роботі розглянуто теоретичні основи інтелектуального аналізу кредитних даних, проаналізовано методи їх підготовки та практично реалізовано моделі класифікації у Python. Отримані результати підтверджують, що методи машинного навчання можуть бути корисними як допоміжний інструмент для обробки та аналізу кредитної інформації банку.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Про банки і банківську діяльність : Закон України від 07.12.2000 № 2121-III (чинна редакція станом на 2026 р.). URL:
<https://zakon.rada.gov.ua/laws/show/2121-14>
2. Про захист персональних даних : Закон України від 01.06.2010 № 2297-VI (чинна редакція станом на 2026 р.). URL:
<https://zakon.rada.gov.ua/laws/show/2297-17>
3. Про організацію формування та обігу кредитних історій : Закон України від 23.06.2005 № 2704-IV (чинна редакція станом на 2026 р.). URL:
<https://zakon.rada.gov.ua/laws/show/2704-15>
4. Національний банк України. Про затвердження Положення про організацію системи управління ризиками в банках України та банківських групах : Постанова Правління НБУ від 11.06.2018 № 64 (зі змінами станом на 2026 р.). URL: <https://bank.gov.ua>
5. Національний банк України. Про затвердження Положення про визначення банками України розміру кредитного ризику за активними банківськими операціями : Постанова Правління НБУ від 30.06.2016 № 351 (зі змінами станом на 2026 р.). URL:
https://bank.gov.ua/ua/legislation/Resolution_30062016_351
6. Національний банк України. Стратегія з розвитку кредитування. Схвалена Радою з фінансової стабільності у 2024 р., актуалізована у 2025 р. URL:
<https://bank.gov.ua/ua/about/lending-strategy>
7. Національний банк України. Звіт про фінансову стабільність. 2025. URL:
<https://bank.gov.ua/ua/stability/report>
8. Національний банк України. Оцінка кредитного ризику корпоративних боржників: матеріали семінару. 2024. URL:
https://bank.gov.ua/admin_uploads/article/CreditRisk_seminar_11-2024.pdf

9. Національний банк України. Про затвердження Положення про регулювання діяльності бюро кредитних історій : Постанова Правління НБУ від 06.10.2023 № 127. URL: <https://bank.gov.ua>

10. Національний банк України. Про затвердження Положення про здійснення контролю за діяльністю бюро кредитних історій : Постанова Правління НБУ від 27.12.2025 № 162. URL: <https://zakon.rada.gov.ua/go/v0162500-25>

11. Національний банк України. Про затвердження Положення про організацію заходів із забезпечення інформаційної безпеки в банківській системі України : Постанова Правління НБУ від 28.09.2017 № 95 (зі змінами станом на 2026 р.). URL: <https://bank.gov.ua>

12. ДСТУ ISO/IEC 27001:2023. Інформаційна безпека, кібербезпека та захист приватності. Системи керування інформаційною безпекою. Вимоги. Київ: ДП «УкрНДНЦ», 2023.

13. Бобиль В. В., Погорєлова Т. Ю. Використання технологій Data Mining у банківській діяльності. Економіка та суспільство. 2023. № 52.

14. Коваленко В. В., Нечипоренко А. В. Інтелектуальний аналіз даних у системах банківського кредитування. Інфраструктура ринку. 2022. № 67. С. 214-220.

15. Дацко М., Друшак Н. Моделювання кредитного скорингу методами машинного навчання. Вісник Львівського університету. Серія економічна. 2022. Вип. 63. С. 56-66. DOI: 10.30970/ves.2022.63.0.6305. URL: <https://publications.lnu.edu.ua/bulletins/index.php/economics/article/view/11898>

16. Кушнірова К. С. Моделювання кредитного скорингу методами інтелектуального аналізу даних. Проблеми системного підходу в економіці. 2025.

17. Ключка О. Активізація ринку банківського кредитування в Україні. Економіка та суспільство. 2025.

18. Оцінювання кредитних ризиків у системі ризик-менеджменту банку. Бізнес Інформ. 2024. № 1. С. 320-332.

19. Колєчка Л. М., Штирхун Х. О. Оптимізація впровадження системи контролю доступу до чутливих даних. Секція 2 «Смарт-аналітика та кібербезпека». Київ: КНЕУ імені Вадима Гетьмана, 2026. Тези доповіді.
20. Дьякова Л. Штучний інтелект як інструмент оцінки кредитоспроможності позичальників. Проблеми і перспективи економіки та управління. 2025.
21. Злобін М. Систематичний огляд глибокого та машинного навчання у фінансовій сфері. Технічні науки та технології. 2025.
22. Штерма Т. В. Застосування бізнес-аналітики в прогнозуванні фінансових процесів. Academy Vision. 2025.
23. Солошенко О. М. Інноваційні практики та інноваційний потенціал методології кредитного скорингу компаній мікрокредитування. Економіка та управління. 2025.
24. Недождій М. Scoring models in solvency evaluation. Kyiv-Mohyla Academy Repository. 2022.
25. Dumitrescu E., Hue S., Hurlin C., Tokpavi S. Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. European Journal of Operational Research. 2022. Vol. 297, No. 3. P. 1178-1192. DOI: 10.1016/j.ejor.2021.06.053.
26. Chang V., et al. Credit Risk Prediction Using Machine Learning and Deep Learning: A Review and Comparative Study. Risks. 2024. Vol. 12, No. 11. Article 174. DOI: 10.3390/risks12110174.
27. El Khair Ghoujdam M. Consumer credit risk analysis through artificial intelligence. Cogent Economics & Finance. 2024. DOI: 10.1080/23322039.2024.2414926.
28. Ayari H., Ben Amor N., Feki J. Machine learning powered financial credit scoring: A systematic review. Artificial Intelligence Review. 2025. DOI: 10.1007/s10462-025-11416-2.

29. Wu Y., Pan Y. Application Analysis of Credit Scoring of Financial Institutions Based on Machine Learning Model. Complexity. 2021. Article ID 9222617. DOI: 10.1155/2021/9222617.

30. Surianto S. J., Chairunisah. Comparison of Logistic Regression and XGBoost Model Performance in Predicting Credit Scores. Journal of Artificial Intelligence and Engineering Applications. 2026. Vol. 5, No. 2. P. 2427-2434. DOI: 10.59934/jaiea.v5i2.1877. URL: <https://ioinformatic.org/index.php/JAIEA/article/view/1877>

31. European Banking Authority. Discussion Paper on Machine Learning for IRB Models. 2021. URL: <https://www.eba.europa.eu>

32. Basel Committee on Banking Supervision. Newsletter on artificial intelligence and machine learning. Bank for International Settlements. 2022. URL: <https://www.bis.org>

33. National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0). 2023. URL: <https://www.nist.gov/itl/ai-risk-management-framework>

34. ISO/IEC 23894:2023. Information technology - Artificial intelligence - Guidance on risk management. International Organization for Standardization, 2023.

35. OECD. Artificial Intelligence, Machine Learning and Big Data in Finance: Opportunities, Challenges, and Implications for Policy Makers. 2021. URL: <https://www.oecd.org/finance/>

36. Credit Risk Dataset. Kaggle. URL: <https://www.kaggle.com/datasets/laotse/credit-risk-dataset>

37. sangramdedge. Credit-Risk-Analysis-Prediction-Dashboard: credit_risk_dataset.csv. GitHub. URL: https://github.com/sangramdedge/Credit-Risk-Analysis-Prediction-Dashboard/blob/main/credit_risk_dataset.csv

38. How Data Quality Affects Machine Learning Models for Credit Risk Modelling. arXiv, 2025. URL: <https://arxiv.org/html/2511.10964v1>

39. Ali S., Ashraf H. Machine Learning and Artificial Intelligence in Credit Risk Prediction for Commercial Banks: A Review. International Journal of Social Sciences

Bulletin. 2025. Vol. 3, No. 12. P. 1377-1394. URL:
<https://ijssbulletin.com/index.php/IJSSB/article/view/2106>

ДОДАТКИ

Додаток А.

Код Python для побудови рисунків розділу 2

У додатку наведено фрагменти коду, що відповідають рисункам розділу 2. Дані у прикладах є умовними та використовуються лише для демонстрації методики візуалізації кредитних даних.

Лістинг А.0. Формування умовного набору кредитних даних для побудови рисунків

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

# 1. Формування умовного кредитного набору даних
np.random.seed(42)
n = 120
income = np.concatenate([
    np.random.normal(28000, 6000, 90),
    np.random.normal(48000, 12000, 30)
]).clip(9000, 85000)
loan = (
    income * np.random.uniform(6, 14, n) +
    np.random.normal(0, 35000, n)
).clip(50000, 900000)
age = np.random.randint(22, 61, n)
credit_rating = (
    700 - (loan / income) * 8 +
    np.random.normal(0, 35, n)
).clip(450, 850)
dti = (loan / (income * 36)).clip(0.05, 0.95)
prev_delay = (
    np.random.random(n) < (0.12 + dti * 0.25)
```

```

).astype(int)

logit = (
    -2.2 +
    1.7 * dti +
    1.1 * prev_delay -
    0.006 * (credit_rating - 650)
)

prob = 1 / (1 + np.exp(-logit))

risk = (np.random.random(n) < prob).astype(int)

df = pd.DataFrame({
    "Вік": age,
    "Дохід": income.round(0),
    "Сума кредиту": loan.round(0),
    "Кредитний рейтинг": credit_rating.round(0),
    "Кредитне навантаження": dti.round(3),
    "Прострочення": prev_delay,
    "Клас ризику": risk
})

# 2. Виведення перших рядків таблиці

print("Перші 10 рядків умовного кредитного набору даних:")

print(df.head(10))

print("\nОписова статистика:")

print(df.describe())

print("\nРозподіл клієнтів за класами ризику:")

print(df["Клас ризику"].value_counts())

```

Лістинг А.1. Код для побудови: Рисунок 2.1 - Загальна логіка підготовки кредитних даних до моделювання

```

import matplotlib.pyplot as plt
from matplotlib.patches import Rectangle, FancyArrowPatch

# Етапи, які відображаються на рисунку 2.1

steps = [
    "Вхідні\нкредитні дані",
    "Очищення\нданих",
    "Кодування та\нормалізація",
    "Аналіз класів\нризиків",

```

```

    "Скорингові\пмоделі",
    "Оцінювання\пякості"
]
# Створення області рисунка
fig, ax = plt.subplots(figsize=(12, 2.8))
# Координати блоків
x_start = 0.2
y = 0.45
box_width = 1.55
box_height = 0.55
gap = 0.55
# Побудова прямокутників та стрілок
for i, step in enumerate(steps):
    x = x_start + i * (box_width + gap)
    rect = Rectangle(
        (x, y),
        box_width,
        box_height,
        fill=False,
        linewidth=1.5
    )
    ax.add_patch(rect)
    ax.text(
        x + box_width / 2,
        y + box_height / 2,
        step,
        ha="center",
        va="center",
        fontsize=10
    )
    if i < len(steps) - 1:
        arrow = FancyArrowPatch(
            (x + box_width, y + box_height / 2),
            (x + box_width + gap * 0.85, y + box_height / 2),
            arrowstyle="->",
            mutation_scale=12,
            linewidth=1.2
        )
        ax.add_patch(arrow)
# Прибирання осей
ax.set_xlim(0, x_start + len(steps) * (box_width + gap))

```

```

ax.set_ylim(0, 1.5)
ax.axis("off")

# Збереження рисунка
plt.tight_layout()
plt.savefig("Рисунок_2_1_Логіка_підготовки_кредитних_даних.png", dpi=300)
plt.show()

```

Лістинг А.2. Код для побудови: Рисунок 2.2 - Гістограма розподілу доходу позичальників

```

import pandas as pd
import matplotlib.pyplot as plt
plt.figure(figsize=(6.5, 4))
plt.hist(df["Дохід"], bins=12, edgecolor="black")
plt.xlabel("Дохід позичальника, грн")
plt.ylabel("Кількість клієнтів")
plt.title("Розподіл доходу позичальників")
plt.savefig("fig_2_2_income_hist.png", dpi=180, bbox_inches="tight")
plt.show()

```

Лістинг А.3. Код для побудови: Рисунок 2.3 - Діаграма розмаху суми кредиту

```

import matplotlib.pyplot as plt
plt.figure(figsize=(6.5, 4))
plt.boxplot(df["Сума кредиту"], vert=False)
plt.xlabel("Сума кредиту, грн")
plt.title("Діаграма розмаху суми кредиту")
plt.savefig("fig_2_3_loan_box.png", dpi=180, bbox_inches="tight")
plt.show()

```

Лістинг А.4. Код для побудови: Рисунок 2.4 - Розподіл кредитних записів за статусом обслуговування

```

import matplotlib.pyplot as plt
class_counts = df["Статус кредиту"].value_counts().sort_index()
plt.figure(figsize=(5.8, 4))
plt.bar(["0 - без дефолту", "1 - дефолт"], class_counts.values, edgecolor="black")
plt.ylabel("Кількість спостережень")
plt.title("Розподіл кредитних записів за статусом")
plt.savefig("fig_2_4_class_dist.png", dpi=180, bbox_inches="tight")
plt.show()

```

Лістинг А.5. Код для побудови: Рисунок 2.5 - Кореляційна матриця основних кредитних ознак

```

import matplotlib.pyplot as plt
corr = df[["Вік", "Дохід", "Сума кредиту", "Кредитний рейтинг",
          "Кредитне навантаження", "Прострочення", "Клас ризику"]].corr()
plt.figure(figsize=(7, 5.8))
plt.imshow(corr, vmin=-1, vmax=1)
plt.colorbar(label="Коефіцієнт кореляції")
plt.xticks(range(len(corr.columns)), corr.columns, rotation=45, ha="right")
plt.yticks(range(len(corr.columns)), corr.columns)
for i in range(len(corr.columns)):
    for j in range(len(corr.columns)):
        plt.text(j, i, f"{corr.iloc[i, j]:.2f}", ha="center", va="center", fontsize=7)
plt.title("Кореляційна матриця основних кредитних ознак")

```

```
plt.savefig("fig_2_5_corr.png", dpi=180, bbox_inches="tight")
plt.show()
```

Лістинг А.6. Код для побудови: Рисунок 2.6 - Логістична функція

```
import numpy as np
import matplotlib.pyplot as plt
z = np.linspace(-8, 8, 200)
p = 1 / (1 + np.exp(-z))
plt.figure(figsize=(6.5, 4))
plt.plot(z, p, linewidth=2)
plt.axhline(0.5, linestyle="--", linewidth=1)
plt.axvline(0, linestyle="--", linewidth=1)
plt.xlabel("Значення лінійної комбінації ознак z")
plt.ylabel("Ймовірність класу дефолту p")
plt.title("Логістична функція для класифікації статусу кредиту")
plt.savefig("fig_2_6_logistic.png", dpi=180, bbox_inches="tight")
plt.show()
```

Лістинг А.7. Код для побудови: Рисунок 2.7 - Схема дерева рішень

```
import matplotlib.pyplot as plt
from matplotlib.patches import Rectangle, FancyArrowPatch
# Створення області рисунка
fig, ax = plt.subplots(figsize=(10, 6))
# Функція для побудови прямокутного блока
def add_box(x, y, text, width=2.6, height=0.7):
    rect = Rectangle(
        (x, y),
        width,
        height,
        fill=False,
        linewidth=1.4
    )
    ax.add_patch(rect)
    ax.text(
        x + width / 2,
        y + height / 2,
        text,
        ha="center",
        va="center",
        fontsize=10
    )
# Функція для побудови стрілки між блоками
def add_arrow(x1, y1, x2, y2, label=None, label_x=None, label_y=None):
    arrow = FancyArrowPatch(
        (x1, y1),
        (x2, y2),
        arrowstyle="->",
        mutation_scale=12,
        linewidth=1.2
    )
    ax.add_patch(arrow)
    if label is not None:
        ax.text(
```

```

        label_x,
        label_y,
        label,
        fontsize=9,
        ha="center",
        va="center"
    )

# Побудова вузлів дерева рішень
add_box(3.7, 5.0, "Кредитний рейтинг < 650?")
add_box(1.1, 3.4, "Кредитне\пнавантаження > 0,40?")
add_box(6.2, 3.4, "Кредит\пбез дефолту")
add_box(0.0, 1.8, "Дефолтний\пстатус")
add_box(2.7, 1.8, "Додаткова\пперевірка")
# Побудова стрілок між вузлами
add_arrow(5.0, 5.0, 2.4, 4.1, label="так", label_x=3.2, label_y=4.5)
add_arrow(5.0, 5.0, 7.5, 4.1, label="ні", label_x=6.4, label_y=4.5)
add_arrow(2.4, 3.4, 1.2, 2.5, label="так", label_x=1.6, label_y=2.8)
add_arrow(2.4, 3.4, 4.0, 2.5, label="ні", label_x=3.2, label_y=2.8)
# Налаштування вигляду рисунка
ax.set_xlim(-0.5, 9.5)
ax.set_ylim(1.0, 6.3)
ax.axis("off")

# Збереження рисунка
plt.tight_layout()
plt.savefig("Рисунок_2_7_Схема_дерева_рішень.png", dpi=300)
plt.show()

```

Копії тез доповідей та статей

Секція 2 «Смарт-аналітика та кібербезпека»

Колєчкіна Л. М.

доктор фізико-математичних наук, професор кафедри системного аналізу та кібербезпеки, Київський Національний університет ім. Вадима Гетьмана, м. Київ, Україна

ORCID ID: <https://orcid.org/0000-0002-4079-1201>

Штирхун Х. О.

студентка, бакалавр, Київський Національний університет ім. Вадима Гетьмана, провідний економіст управління супроводження кредитних операцій корпоративного бізнесу в банку Credit Agricole, м. Київ, Україна

ОПТИМІЗАЦІЯ ВПРОВАДЖЕННЯ СИСТЕМИ КОНТРОЛЮ ДОСТУПУ ДО ЧУТЛИВИХ ДАНИХ

Контроль доступу до інформаційних ресурсів є ключовим елементом інформаційної безпеки в сучасних організаціях [1]. Він забезпечує запобігання несанкціонованому доступу, захист персональних даних та підтримку безпечного функціонування бізнес-процесів. Одним із сучасних підходів до вибору ефективних систем контролю доступу є метод аналізу ієрархій (MAI), що дозволяє формалізувати процес прийняття рішень у багатокритеріальних задачах [2].

Метод MAI передбачає побудову ієрархії, що складається з трьох рівнів: мета, критерії та альтернативи. Критерії можуть включати безпеку, вартість впровадження, зручність для користувачів, масштабованість та відповідність нормативним вимогам [3]. Застосування методу передбачає попарне порівняння критеріїв, нормалізацію матриці та обчислення ваг кожного критерію, що дозволяє кількісно оцінити їхню важливість [4].

Основні кроки алгоритму MAI включають:

1. Визначення мети та критеріїв оцінки;
2. Визначення альтернатив для оцінки;
3. Попарне порівняння критеріїв і альтернатив за шкалою 1–9;
4. Нормалізація матриць попарних порівнянь та обчислення ваг;

$$w_i = \frac{\sum_{j=1}^n a_{ij} / \sum_{i=1}^n a_{ij}}{n}, \quad \text{для } i = 1, 2, \dots, n$$

де a_{ij} — елемент матриці попарних порівнянь, n — кількість критеріїв, w_i — вага критерію [4].

5. Розрахунок інтегральних рейтингів альтернатив;

$$S_k = \sum_{i=1}^n w_i \cdot c_{ik}, \quad k = 1, 2, \dots, m$$

Короткий звіт подібності



Звіт не був оцінений

Метадані

ДОКУМЕНТ

Заголовок

БКР_Штирхун Христини

Автор

Штирхун Христина

Науковий керівник / Експерт

Камінський О.Є.

ІД документу

334334432

ОРГАНІЗАЦІЯ

Назва організації

Kyiv National Economic University named after Vadym Hetman
KNEU

підрозділ

кафедра системного аналізу та кібербезпеки

ЗВІТ

Дата звіту

6/16/2026

Дата редагування

Обсяг знайдених подібностей

Коефіцієнт подібності визначає, який відсоток тексту по відношенню до загального обсягу тексту було знайдено в різних джерелах. Зверніть увагу, що високі значення коефіцієнта не автоматично означають плагіат. Звіт має аналізувати компетентна / уповноважена особа.



КП 1

25

Довжина фрази для коефіцієнта подібності 2



КП 2

10975

Кількість слів



КЦ

83268

Кількість символів

Тривога

У цьому розділі ви знайдете інформацію щодо текстових спотворень. Ці спотворення в тексті можуть говорити про МОЖЛИВІ маніпуляції в тексті. Спотворення в тексті можуть мати навмисний характер, але частіше характер технічних помилок при конвертації документа та його збереженні, тому ми рекомендуємо вам підходити до аналізу цього модуля відповідально. У разі виникнення запитань, просимо звертатися до нашої служби підтримки.

Заміна букв



9

Інтервали



0

Мікропробіли



0

Білі знаки



0

Парафрази (SmartMarks)



8

Джерела

Нижче наведений список джерел. В цьому списку є джерела із різних баз даних. Колір тексту означає в якому джерелі він був знайдений. Ці джерела і значення Коефіцієнту Подібності не відображають прямого плагіату. Необхідно відкрити кожне джерело і проаналізувати зміст і правильність оформлення джерела.

10 найдовших фраз

Колір тексту

#	НАЗВА ТА АДРЕСА ДЖЕРЕЛА URL (НАЗВА БАЗИ)	КІЛЬКІСТЬ ІДЕНТИЧНИХ СЛІВ (ФРАГМЕНТІВ)
1	https://neerth.com/uk/blogs/programming-insights/streamlining-data-preparation-with-nvthon-	17 (0.15 %)