МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ЛЬВІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ВЕТЕРИНАРНОЇ МЕДИЦИНИ ТА БІОТЕХНОЛОГІЙ ІМ.С.З. ГЖИЦЬКОГО

ФАКУЛЬТЕТ МЕХАНІКИ, ЕНЕРГЕТКИ ТА ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

Кафедра інформаційних технологій

КВАЛІФІКАЦІЙНА РОБОТА

першого (бакалаврського) рівня вищої освіти

на тему: «Оцінка методів машинного навчання для виявлення викидів у наборах даних»

Виконав: студент групи КН-41

Спеціальності 122-Комп’ютерні науки

Наконечний Мар’ян Романович

Керівник
доц. Боярчук О.В.
(науковий ступінь, вчене звання, прізвище та ініціали)

Рецензент:
доц. Шарибура А.О.
(науковий ступінь, вчене звання, прізвище та ініціали)

Дубляни-2025

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ЛЬВІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ВЕТЕРЕНАРНОЇ МЕДИЦИНИ ТА БІОТЕХНОЛОГІЙ
ІМ.С.З. ГЖИЦЬКОГО
ФАКУЛЬТЕТ МЕХАНІКИ, ЕНЕРГЕТКИ ТА ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

КАФЕДРА ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

перший (бакалаврський) рівень вищої освіти
ОС «Бакалавр» за спеціальністю – 122– «Комп'ютерні науки»

"ЗАТВЕРДЖУЮ"
Завідувач кафедри_____
д.т.н., проф. А.М. Тригуба
"____"_____2025р.

З А В Д А Н Н Я

НА КВАЛІФІКАЦІЙНУ РОБОТУ СТУДЕНТУ

Наконечний Мар'ян Романович

Тема роботи *«Оцінка методів машинного навчання для виявлення викидів у наборах даних»*

Керівник роботи *к.т.н., доц., Боярчук О.В.*

затверджені наказом по університету 172 /к-с від 25. 02. 2025

Строк подання студентом роботи 16.06. 2025 р.

Вхідні дані до роботи Власні зібрані транзакції, реальні або анонімізовані дані з фінтех-платформ, публічні датасети.

Зміст розрахунково-пояснювальної записки:

1. ДОСЛІДЖЕННЯ ПРЕДМЕТНОЇ ОБЛАСТІ
2. МЕТОДИ МАШИННОГО НАВЧАННЯ ДЛЯ ВИЯВЛЕННЯ ВИКИДІВ
3. РОЗРОБКА ІНСРУМЕНТІВ ДЛЯ ВИЯВЛЕННЯ ВИКИДІВ У ДАНИХ ТА АНАЛІЗ РЕЗУЛЬТАТІВ
4. ОХОРОНА ПРАЦІ ТА БЕЗПЕКА В НАДЗВИЧАЙНИХ СИТУАЦІЯХ
5. СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ
6. ДОДАТОК А

5. Перелік презентаційного матеріалу (з зазначенням обов'язкових елементів):_____
_____
_____
_____
_____
_____
_____

6. Консультанти з розділів:

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
1, 2, 3,	*Боярчук О.В., доцент кафедри інформаційних технологій*		
4	*Городецький І.М., доцент кафедри інженерної механіки*		

7. Дата видачі завдання 25. 02. 2025

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів кваліфікаційної роботи	Терміни виконання роботи	Примітка
1	*Написання першого розділу та означення головних завдань роботи*	*25.02.2025 – 01.04.2025*	
2	*Виконання другого розділу та формування головних показників для розрахунків*	*01.04.2025 – 01.05.2025*	
3.	*Виконання третього розділу та формування початкових даних*	*01.04.2025 – 01.05.2025*	
4.	*Виконання четвертого розділу та узагальнення отриманих результатів бакалаврської роботи*	*01.04.2025 – 01.05.2025*	
5.	*Завершення роботи в цілому*	*01.06.2025 – 09.06.2025*	

Студент _____ Наконечний М.Р.
(підпис)

Керівник роботи _____ Боярчук О.В.
(підпис)

УДК 004.8'1:004.93'1:519.25

Кваліфікаційна робота: 95 с., 25 рис., 7 табл., 2 дод., 14 джерел.

МАШИННЕ НАВЧАННЯ, БІНАРНА КЛАСИФІКАЦІЯ, ЛОГІСТИЧНА РЕГРЕСІЯ, ДЕРЕВА РІШЕНЬ, XGBOOST, SVM.

*Об'єкт дослідження***:** методи та моделі машинного навчання. *Предмет дослідження***:** методи виявлення аномалій та моделі класифікації, призначені для виявлення викидів у транзакційних наборах даних. *Мета дослідження:* розробка та оцінка ефективності моделей машинного навчання для автоматичного виявлення викидів у великих наборах даних, зокрема в контексті транзакційної активності користувачів.

Використані моделі: у програмній реалізації було використано логістичну регресію, дерева рішень, випадковий ліс, XGBoost та SVM — для порівняння їх здатності виявляти нетипові (аномальні) спостереження в даних.

Актуальність роботи зумовлена стрімким зростанням обсягів цифрових даних, у яких виявлення викидів є критично важливим як для підвищення якості аналітики, так і для своєчасного реагування на потенційно небезпечні або помилкові транзакції. Викиди можуть свідчити про збої, шахрайство або нестандартну поведінку системи.

Отримані результати: реалізовано та протестовано декілька моделей машинного навчання, які продемонстрували здатність виявляти викиди у транзакційних даних з високою ефективністю. Було проведено аналіз якості моделей за стандартними метриками.

Подальші напрямки дослідження включають використання глибокого навчання, методів напівасупервізованого навчання, а також розширення вхідних ознак для підвищення точності виявлення аномалій.

ЗМІСТ

ВСТУП

У сучасному цифровому світі обсяг даних зростає надзвичайно швидкими темпами. Великі обсяги інформації генеруються щодня в різних сферах – фінансах, охороні здоров'я, телекомунікаціях, електронній комерції тощо. Одним із ключових завдань при роботі з такими даними є виявлення аномалій або викидів — нетипових записів, що можуть свідчити про технічні збої, шахрайство або інші небажані події. Здатність своєчасно і точно виявляти такі записи є критично важливою для підвищення безпеки, точності аналітики та прийняття рішень.

З традиційними статистичними методами часто важко ефективно виявляти складні, багатовимірні та нетривіальні відхилення. У зв'язку з цим зростає актуальність використання алгоритмів машинного навчання, які здатні адаптивно навчатися на історичних даних і виявляти приховані закономірності. Особливо важливим стає оцінювання ефективності різних моделей у задачах виявлення викидів, оскільки вибір методу безпосередньо впливає на точність і швидкість аналітичної системи.

Метою даної дипломної роботи є дослідження, реалізація та порівняльна оцінка ефективності різних методів машинного навчання для задачі виявлення аномальних записів у великих транзакційних наборах даних. У процесі дослідження застосовуються популярні моделі класифікації, зокрема логістична регресія, дерева рішень, випадковий ліс, XGBoost та SVM, а також проводиться аналіз якості побудованих моделей за допомогою ключових метрик: Precision, Recall, F1-score та ROC AUC.

РОЗДІЛ 1
ДОСЛІДЖЕННЯ ПРЕДМЕТНОЇ ОБЛАСТІ

1.4 **Актуальність проблеми платіжного шахрайства**

З поширенням цифрових фінансових технологій платіжне шахрайство набуло особливої гостроти в умовах сучасного інформаційного простору. Активне використання онлайн-покупок, мобільного банкінгу та електронних гаманців призвело до стрімкого зростання кількості безготівкових транзакцій. Водночас це створило сприятливе середовище для зловмисників, які, використовуючи недоліки платіжної інфраструктури, постійно вдосконалюють свої методи і швидко адаптуються до нових механізмів захисту.

Рисунок 1.1. Візуалізація загрози крадіжки платіжних даних у цифровому середовищі

Платіжне шахрайство призводить до серйозних фінансових втрат як для окремих користувачів, так і для компаній та фінансових організацій. Такі втрати не лише безпосередньо шкодять постраждалим, але й можуть мати масштабні економічні наслідки — зростання витрат для бізнесу та банків,

підвищення цін для кінцевих споживачів і ризик дестабілізації економічних процесів загалом.

Одним із ключових аспектів цієї проблеми є підрив довіри користувачів до платіжної інфраструктури. Клієнти очікують високого рівня безпеки під час здійснення транзакцій, і нездатність своєчасно виявляти та запобігати шахрайським діям може серйозно вплинути на репутацію компаній і банківських установ. Збереження довіри користувачів є критичним чинником для стабільного розвитку цифрової економіки.

Не менш важливо й те, що урядові структури та регулятори багатьох країн приділяють дедалі більше уваги питанням запобігання фінансовим шахрайствам. Вони встановлюють жорсткі вимоги для забезпечення захисту прав споживачів і стабільності фінансової інфраструктури. Відтак, для бізнесу та банків дотримання таких норм стає необхідною умовою для легальної та безперебійної діяльності на міжнародному рівні.

1.5 **Вирішення проблеми методами машинного навчання**

Стрімкий розвиток машинного навчання та штучного інтелекту надав нові інструменти для ефективного виявлення платіжного шахрайства. Сучасні алгоритми здатні оперативно обробляти великі масиви даних, виявляючи атипові шаблони поведінки та формуючи точні прогнози щодо ризиків шахрайства. Завдяки впровадженню моделей машинного навчання стає можливим своєчасне виявлення та запобігання шахрайським транзакціям.

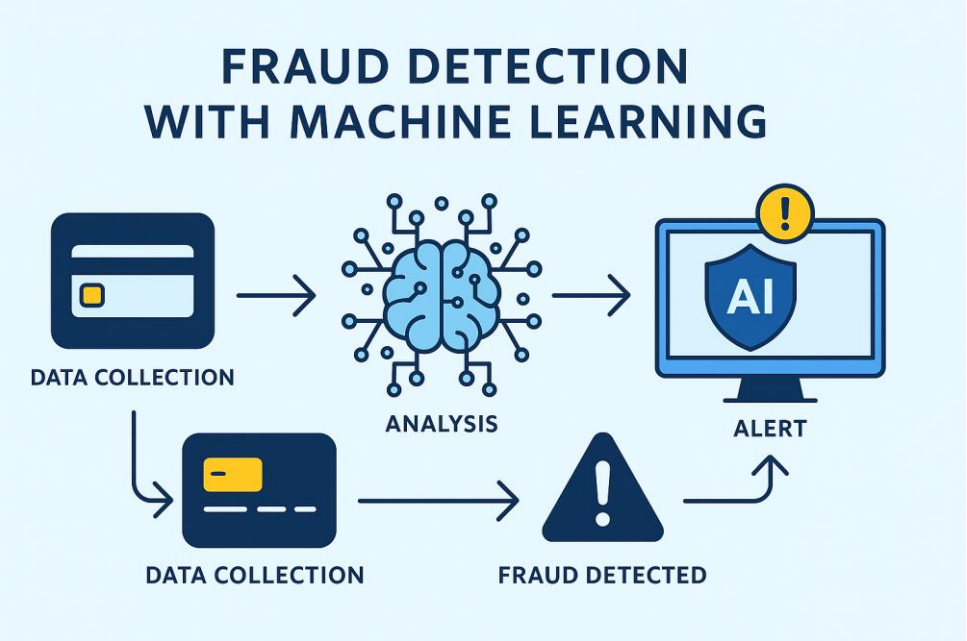


Рисунок 1.2. Схема процесу виявлення платіжного шахрайства за допомогою моделей машинного навчання

Проблема платіжного шахрайства є актуальною для широкого кола учасників — від фінансових організацій і торгових платформ до систем електронних платежів. Важливу роль у вирішенні цієї проблеми відіграє координація зусиль та обмін інформацією між усіма зацікавленими сторонами. Новітні технології, зокрема федеративне навчання, дають змогу співпрацювати без порушення конфіденційності даних, що значно підвищує надійність систем виявлення шахрайства.

Традиційні підходи, що спираються переважно на ручну перевірку, є трудомісткими та схильними до помилок. Натомість алгоритми машинного навчання дозволяють автоматизувати процес виявлення підозрілих транзакцій, скорочуючи витрати часу і ресурсів. Особливо важливою є здатність таких моделей працювати в режимі реального часу, забезпечуючи оперативне реагування на шахрайські дії.

Ще одним вагомим аргументом на користь машинного навчання є динамічний характер шахрайських схем: зловмисники постійно оновлюють свої стратегії, змушуючи системи безпеки адаптуватися. Моделі машинного навчання можуть еволюціонувати разом із новими викликами — навчаючись на

історичних даних і зворотному зв’язку в реальному часі, вони поступово вдосконалюють здатність розпізнавати нові патерни поведінки та оперативно реагувати на потенційні загрози.

1.6 Огляд методів виявлення аномальних транзакцій та викидів даних

Сьогодні існує широкий спектр методів для виявлення шахрайських платіжних транзакцій — від простих рішень із мінімальними математичними вимогами до складних моделей із глибоким аналітичним підґрунтям та високими вимогами до обчислювальних ресурсів.

1. Системи на основі правил працюють за принципом попередньо визначених умов, що сигналізують про підозрілі транзакції. Ці умови можуть стосуватися суми операції, частоти, геолокації тощо. Хоча такі системи легкі у впровадженні, вони часто не здатні виявити складні шахрайські сценарії, що постійно змінюються.

2. Методи навчання з учителем використовують заздалегідь розмічені набори даних, у яких транзакції класифіковано як легітимні або шахрайські. Алгоритми, як-от логістична регресія, дерева рішень чи Random Forest, вивчають закономірності та дозволяють моделі прогнозувати ймовірність шахрайства для нових транзакцій.

3. Безвчительське машинне навчання застосовується у випадках, коли мітки відсутні або є обмеженими. Тут використовуються алгоритми кластеризації (наприклад, K-Means, DBSCAN), які групують подібні транзакції та виявляють відхилення від типових патернів.

4. Гібридні моделі поєднують переваги систем правил і алгоритмів машинного навчання. Правила дають змогу швидко виявляти відомі типи шахрайства, а ML-моделі — адаптуватися до нових, ще невідомих шаблонів.

5. Графова аналітика базується на вивченні взаємозв’язків між учасниками транзакцій. Створення графів дозволяє виявити нетипові взаємодії, зокрема групи пов’язаних облікових записів, що можуть свідчити про скоординовані шахрайські дії.

6. Аналіз текстових даних і методи NLP застосовуються для роботи з неструктурованою інформацією: описами платежів, повідомленнями клієнтів чи електронною перепискою. Вони дозволяють виявляти приховані сигнали шахрайства на основі вмісту текстів.

7. Поведінкова аналітика фокусується на вивченні звичних транзакцій користувача, формуючи для нього "нормальний" профіль поведінки. Відхилення від цієї норми можуть свідчити про викрадення облікового запису або використання особистих даних у шахрайських цілях.

Ефективність кожного підходу залежить від характеру завдання, обсягу й типу доступних даних, технічних можливостей організації та її здатності адаптуватися до нових викликів. Найчастіше використовується комбінація різних методів, що дозволяє досягти вищої точності та знизити кількість хибних спрацювань.

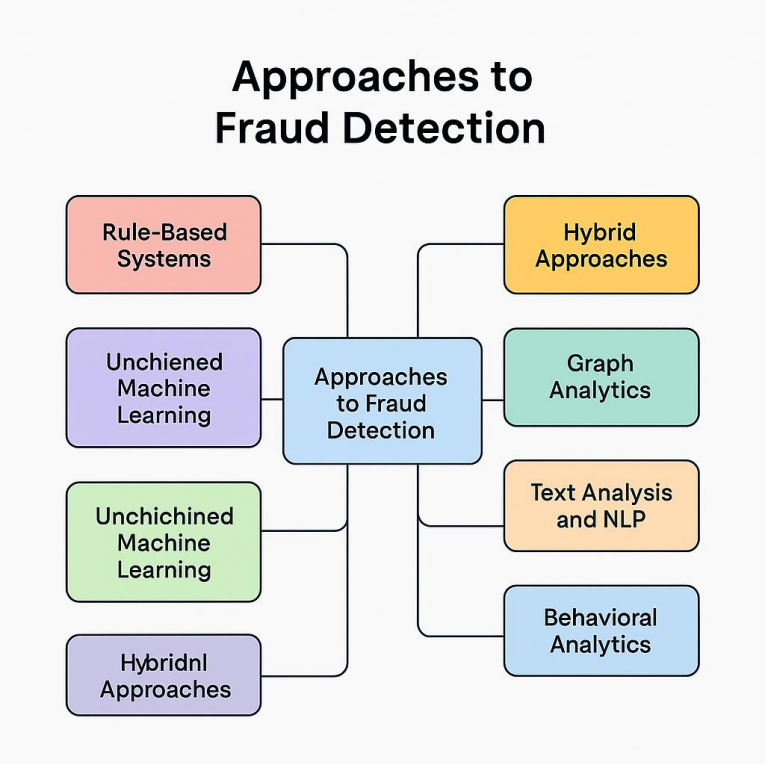


Рисунок 1.3. Блок-схема підходів до виявлення платіжного шахрайства

Як показано на блок-схемі в рисунку 1.3. основні підходи до виявлення платіжного шахрайства поділяються на кілька ключових категорій. Системи на основі правил дозволяють швидко реагувати на вже відомі сценарії шахрайства, але їх гнучкість у розпізнаванні нових загроз є обмеженою. Методи машинного навчання — як керовані, так і некеровані — відкривають можливість виявлення складних і динамічних шахрайських схем завдяки глибокому аналізу великих обсягів даних.

Гібридні рішення поєднують переваги традиційних правил і машинного навчання, що дозволяє досягти вищої точності та адаптивності. Підходи на основі графової аналітики й обробки природної мови (NLP) дають змогу виявляти приховані зв’язки між транзакціями та аналізувати неструктуровану інформацію, яка може містити ознаки шахрайства.

Окремо варто виділити поведінкову аналітику, яка зосереджується на виявленні аномалій у звичній активності користувача. Це дозволяє оперативно фіксувати випадки компрометації облікових записів або інших форм зловмисних дій.

РОЗДІЛ 2
МЕТОДИ МАШИННОГО НАВЧАННЯ ДЛЯ ВИЯВЛЕННЯ ВИКИДІВ

2.4. Логістична регресія для виявлення аномалій

У логістичній регресії сигмоїдна (логістична) функція застосовується для перетворення вихідного результату моделі у значення ймовірності. Це дає змогу визначити, чи є транзакція шахрайською, на основі заданого порогового рівня. Такий підхід особливо добре підходить для задач бінарної класифікації, зокрема при виявленні підозрілих або аномальних фінансових операцій [4]. Графічне представлення сигмоїдної функції наведено на рисунку 2.1.

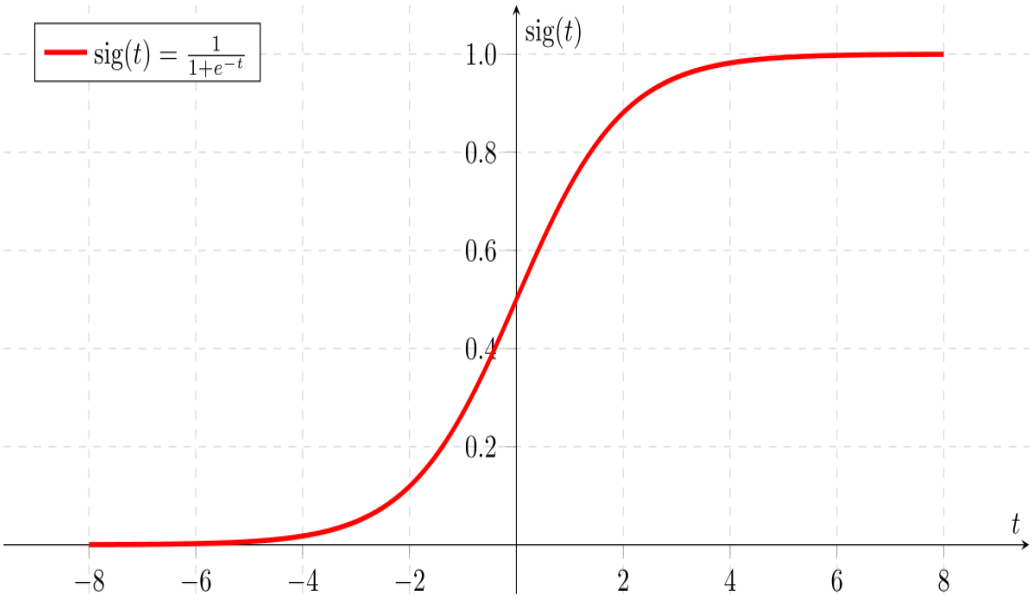


Рисунок 2.1 – Сигмоїдна функція

Наведемо більш формальний опис роботи моделі логістичної регресії.

Нехай вектор незалежних вхідних у модель змінних можна представити у наступному вигляді:

$$X = \left|\left| \overline{x}_1, \overline{x}_2, \ldots, \overline{x}_n \right|\right|, \text{ де } \overline{x}_i - \text{вектор значень } i\text{-ої змінної}$$

А залежна змінна Y має наступне представлення:

$Y \in \{0, 1\}$Для відтворення функціональної залежності між X та Y спочатку вводиться деяка проміжна змінна z:

$$z = w_1 x_1 + w_2 x_2 + \ldots + w_n x_n + b, \text{ де } (w_1, w_2, \ldots, w_n, b) - \text{невідомі}$$

параметри моделі. Інколи кажуть, що w – ваги моделі, b – зміщення.

Зрештою,

$$y = g(z) = \frac{1}{1 + e^{-z}}$$

Формально можна записати,

$$y = g(W^T X + b), \text{ де } g - \text{сигмоїдна функція}$$

Для знаходження параметрів $(w_1, w_2, \ldots, w_n, b)$ використовується задана вибірка і метод максимальної правдоподібності. Даний метод знаходить набір параметрів, які максимізують ймовірність спостереження заданих бінарних результатів для відповідних незалежних змінних.

Оскільки задача максимізації функції правдоподібності еквівалентна задачі максимізації її логарифму, цільова функція моделі приймає вигляд:

$$log(L(W, b)) = \sum_{i=1}^{m} y^{(i)} log(g(W^T x^{(i)} + b)) +$$

2.5. Дерева рішень для виявлення викидів у платіжних транзакціях

Дерева рішень є одним із поширених алгоритмів машинного навчання, який ефективно застосовується як для класифікаційних задач, так і для задач регресії. Суть його роботи полягає у рекурсивному поділі простору даних на підмножини. Структурно дерево рішень являє собою спрямовану ієрархічну модель з початковим вузлом, що називається коренем і не має вхідних зв'язків. Решта вузлів приєднується до дерева через одне вхідне ребро. Вузли, які мають вихідні ребра, називають внутрішніми або вузлами перевірки, а ті, що не мають виходів, — листовими або кінцевими (вузлами прийняття рішень) [5]. Схематичне зображення дерева рішень наведено на рисунку 2.2.

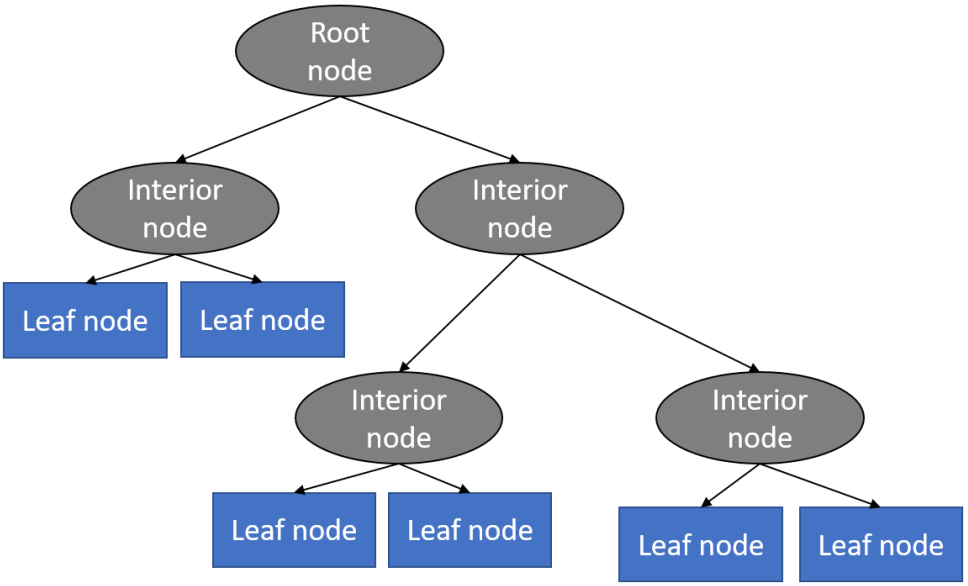


Рисунок 2.2 – Структура дерев рішень

Для кращого розуміння принципів роботи дерева рішень його можна уявити у вигляді послідовності вкладених умов типу *if-else*, подібних до конструкцій, що широко використовуються в програмуванні. Саме ця інтерпретованість робить алгоритм зручним і зрозумілим для користувачів, що є однією з ключових переваг дерева рішень над багатьма іншими моделями машинного навчання.

Далі розглядається математичне обґрунтування оцінки якості поділу вузлів дерева. У більшості випадків внутрішні вузли формуються за значенням лише одного атрибута — тобто розбиття є одновимірним. Алгоритм обирає найінформативніший атрибут, що дозволяє найкращим чином розділити дані. Для цього використовуються різноманітні однофакторні критерії [5].

Одним із найвідоміших є критерій приросту інформації, який базується на понятті ентропії — показника, що відображає ступінь невизначеності або випадковості в наборі даних. На рисунку 2.3 продемонстровано приклад поділу з різним ступенем чистоти.

У застосуванні до виявлення підозрілих транзакцій або аномалій у фінансових операціях, дерева рішень дозволяють поетапно оцінювати ознаки і приймати рішення щодо відхилення від нормальної поведінки, що допомагає своєчасно виявляти потенційно шахрайські дії.

2.6. XGBoost (Extreme Gradient Boosting)

XGBoost (Extreme Gradient Boosting, екстремальний градієнтний бустинг) — це вдосконалений і потужний алгоритм машинного навчання, заснований на принципі градієнтного бустингу з використанням дерев рішень. Він відзначається високою продуктивністю та адаптивністю, що робить його ефективним інструментом для розв’язання широкого спектра задач у сфері машинного навчання.

Цей алгоритм реалізує підхід ансамблевого навчання, в якому кілька слабких моделей (зазвичай дерев рішень) об’єднуються в одну сильну модель з високою точністю. Основні елементи, які забезпечують ефективність XGBoost, включають:

1. Регуляризація**.** XGBoost впроваджує засоби контролю складності моделі за допомогою регуляризації, застосовуючи як L1 (Lasso), так і L2 (Ridge)

методи. Це допомагає зменшити ризик перенавчання та оптимізувати вибір ознак.

2. Градієнтна оптимізація. Алгоритм покладається на градієнтний спуск для поступового поліпшення ансамблю моделей шляхом мінімізації функції втрат. Це дозволяє ефективно моделювати складні нелінійні залежності між вхідними і цільовими змінними.

3. Обрізка дерев (pruning)**.** У процесі побудови дерев XGBoost видаляє ті розгалуження, які не сприяють суттєвому покращенню моделі. Це знижує складність алгоритму та підвищує його здатність до узагальнення.

4. Обробка відсутніх даних. Алгоритм має вбудовані механізми для автоматичної роботи з пропущеними значеннями під час навчання і прогнозування, що значно спрощує попередню підготовку даних.

5. Паралельна обробка. XGBoost підтримує багатопоточність і розподілені обчислення, а також може використовувати графічні процесори, що суттєво прискорює процес навчання і робить алгоритм придатним для обробки великих обсягів даних.

XGBoost функціонує як послідовний ансамбль дерев рішень, де кожне наступне дерево навчається на помилках попередніх. На відміну від випадкового лісу, підсумковий прогноз формується як сума результатів усіх дерев [9].

$$\widehat{y}_i = \sum_{k=1}^{n} f_k(x_i)$$

де $f_k(x_i)$ – результат k-того дерева для прикладу i

Цільова функція алгоритму:

$$Obj(\theta) = L(\theta) + \Omega(\theta),\ L(\theta) = \sum_{i=1}^{n} l(y_i, \widehat{y}_i) - \text{функція втрат}$$

y_i – значення цільової змінної, $\widehat{y}_i$ – прогнозне значення цільової змінної

$\Omega(\theta) = \sum_{k=1}^{K} \Omega(f_k)$ – штрафування складності моделі (регуляризація)

Модель тренується в адитивній манері. Якщо $\widehat{y}_i^{\,t}$ – прогноз моделі для i-того прикладу на ітерації t, то $\widehat{y}_i^{\,t}$ – може бути представлена у вигляді:

$$\widehat{y}_i^{\,t} = \widehat{y}_i^{\,t-1} + f_t(x_i)$$

Архітектуру алгоритму XGBoost ілюстративно представлено на рисунку 2.4.

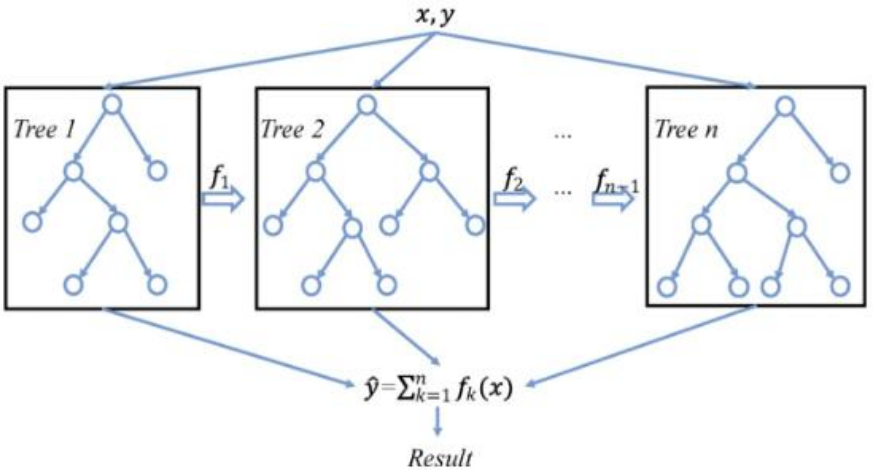


Рисунок 2.4 – Архітектура XGBoost

На рисунку представлено схему роботи ансамблевого підходу, зокрема алгоритму XGBoost, який є одним із найрезультативніших засобів для виявлення викидів у даних.

Принцип дії полягає в поступовому побудуванні серії дерев рішень, де кожне наступне дерево навчається на помилках попередніх, поступово покращуючи точність. У фіналі прогнози всіх дерев об'єднуються, формуючи загальний результат. Така стратегія дозволяє моделі зосереджуватися саме на складних і нетипових випадках, які залишаються непоміченими при використанні простіших методів.

Це має особливе значення у задачах виявлення шахрайських транзакцій або інших аномальних записів у великих обсягах інформації.

РОЗДІЛ 3
РОЗРОБКА ІНСРУМЕНТІВ ДЛЯ ВИЯВЛЕННЯ ВИКИДІВ У ДАНИХ ТА АНАЛІЗ РЕЗУЛЬТАТІВ

3.1. Обґрунтування мови програмування та бібліотек для реалізації продукту

Сьогодні існує безліч мов програмування, фреймворків та бібліотек для реалізації моделей і систем машинного навчання. Проте вже тривалий час провідною мовою в цій сфері залишається Python — як для невеликих завдань аналізу даних, так і для побудови масштабних AI-рішень.

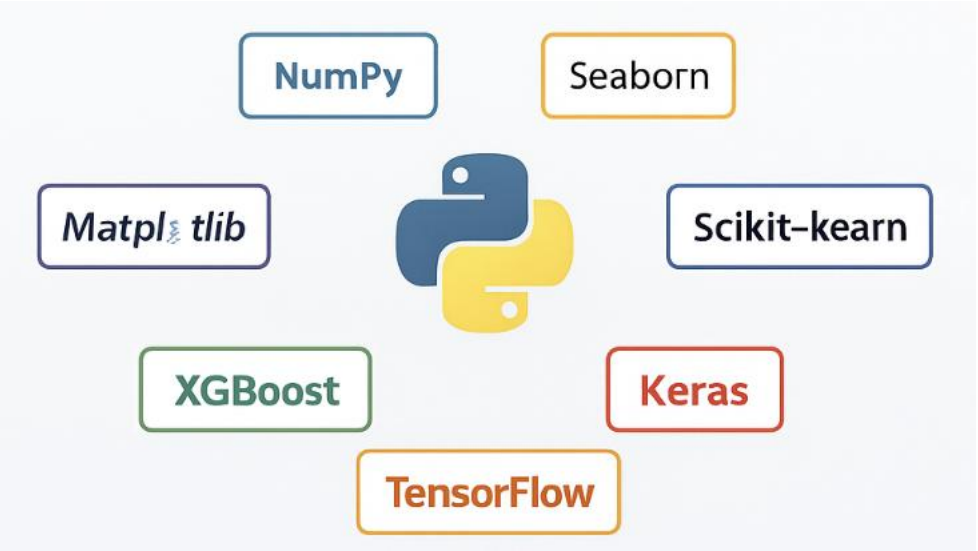


Рисунок 3.1. Інструменти Python, що застосовуються для реалізації систем машинного навчання

Python вирізняється простим і зрозумілим синтаксисом, зручністю у використанні, підтримкою крос-платформності та широким спектром інструментів для обробки, візуалізації та моделювання даних. Усе це робить його найефективнішим і найпопулярнішим вибором як для початківців, так і для досвідчених фахівців у галузі машинного навчання.

У межах даного дослідження для роботи з даними були використані бібліотеки **numpy** та **pandas**, які забезпечують швидке та зручне оперування масивами і таблицями у векторному та матричному форматі.

Для побудови моделей, кодування категоріальних ознак і оцінки якості результатів застосовується бібліотека **scikit-learn**, що є однією з найпоширеніших у сфері машинного навчання. Вона пропонує великий набір інструментів для попередньої обробки, моделювання та аналізу, а також підтримується активним співтовариством розробників, що значно спрощує вирішення практичних питань.

3.2.Опис використовуваних даних

Перш ніж перейти до побудови та випробування моделей для виявлення викидів у даних, слід проаналізувати структуру та основні характеристики використовуваного датасету. У межах дослідження застосовується набір даних, який містить інформацію про транзакції користувачів певної онлайн-платформи. Загальна кількість записів у цьому наборі становить приблизно 405 тисяч.

Головною метою є виявлення аномальних транзакцій (викидів) шляхом аналізу відповідних характеристик.

Набір даних включає такі змінні:

- id_transaction – унікальний ідентифікатор кожної транзакції;
- is_fraud – цільовий показник, який позначає, чи є операція шахрайською;
- date_created – дата і час проведення операції;
- card_brand – бренд платіжної картки;
- card_country – країна, з якої походить картка;
- time_diff_pay_reg – інтервал між реєстрацією користувача та транзакцією;

- is_non_marketing_user – маркер, що відображає канал залучення користувача;
- cnt_payments_before – кількість попередніх платежів;
- time_diff_from_last_failed_payment – час з моменту останньої невдалої спроби оплати;
- avg_time_diff_between_payments – середній час між транзакціями;
- cnt_failed_pays_not_same_card – кількість невдалих оплат з інших карт;
- cnt_failed_pays_same_card – кількість невдалих оплат з цієї ж картки; cnt_fraud_users_of_other_metas_who_used_same_card – кількість підозрілих користувачів, які користувалися тією ж карткою;
- cnt_blocked_users_same_meta_who_paid – число користувачів, які були заблоковані після оплати;
- perc_blocked_users_same_meta_who_paid – відсоток заблокованих акаунтів серед подібних користувачів;
- perc_failed_pays_users_same_meta – частка невдалих транзакцій;
- sum_spent_bonuses – сума витрачених бонусів;
- cnt_out_messages – кількість повідомлень, надісланих користувачем;
- transaction_status – поточний статус транзакції;
- cnt_card_holders_before – кількість інших власників карток, пов’язаних із цим користувачем;
- cnt_error – число певних транзакційних помилок; perc_error – частка таких помилок серед усіх операцій користувача.

Цей датасет використовується для тренування моделей машинного навчання, які повинні виявляти нетипові транзакції як потенційні викиди у загальному наборі даних.

На рисунку 3.1 представлено частину п’яти перших записів у датасеті.

card_brand	card_country	time_diff_pay_reg	is_non_marketing_user	cnt_payments_before	time_diff_from_last_failed_payment
MAESTRO	IRN	69641394	0	1	102.0
VISA	USA	5706136	0	2	56.0
VISA	USA	5879518	0	3	173382.0
VISA	NZL	2190888	0	1	120796.0
MASTERCARD	USA	1076	0	1	98.0

Рисунок 3.1 – Приклад частини змінних для перших 5 записів

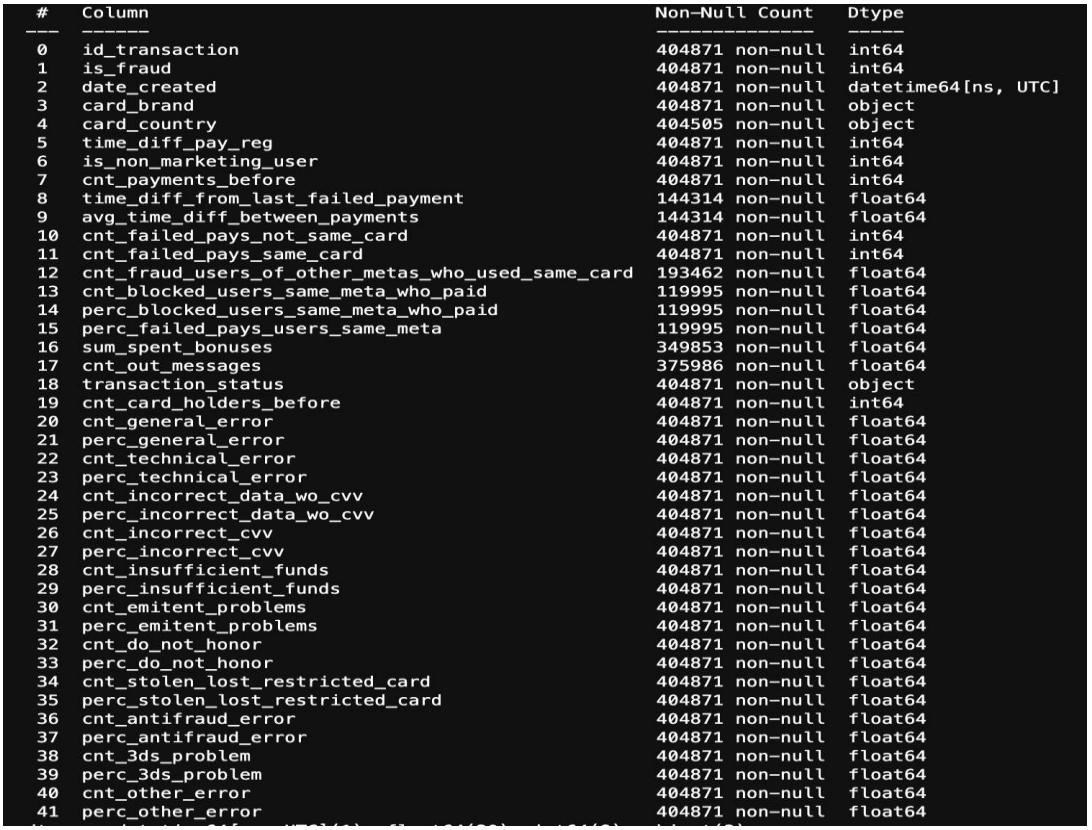
Як уже згадувалося, дані, що використовуються в дослідженні, характеризуються суттєвим дисбалансом між класами, що є типовою ознакою задач з виявлення викидів. Кількість звичайних (нормальних) транзакцій значно перевищує кількість аномальних записів. На рисунку 3.2 наведено візуалізацію цього дисбалансу у вибраному наборі даних.

```
df['is_fraud'].value_counts(normalize=True) * 100

is_fraud
0    72.922486
1    27.077514
```

Рисунок 3.2 – Незбалансованість класів у датасеті

У наборі даних наявні пропущені значення для окремих змінних, що є поширеним явищем під час роботи з реальними даними. Цей факт підтверджується аналізом загальної описової статистики змінних, наведеної на рисунку 3.3.



#	Column	Non-Null Count	Dtype
0	id_transaction	404871 non-null	int64
1	is_fraud	404871 non-null	int64
2	date_created	404871 non-null	datetime64[ns, UTC]
3	card_brand	404871 non-null	object
4	card_country	404505 non-null	object
5	time_diff_pay_reg	404871 non-null	int64
6	is_non_marketing_user	404871 non-null	int64
7	cnt_payments_before	404871 non-null	int64
8	time_diff_from_last_failed_payment	144314 non-null	float64
9	avg_time_diff_between_payments	144314 non-null	float64
10	cnt_failed_pays_not_same_card	404871 non-null	int64
11	cnt_failed_pays_same_card	404871 non-null	int64
12	cnt_fraud_users_of_other_metas_who_used_same_card	193462 non-null	float64
13	cnt_blocked_users_same_meta_who_paid	119995 non-null	float64
14	perc_blocked_users_same_meta_who_paid	119995 non-null	float64
15	perc_failed_pays_users_same_meta	119995 non-null	float64
16	sum_spent_bonuses	349853 non-null	float64
17	cnt_out_messages	375986 non-null	float64
18	transaction_status	404871 non-null	object
19	cnt_card_holders_before	404871 non-null	int64
20	cnt_general_error	404871 non-null	float64
21	perc_general_error	404871 non-null	float64
22	cnt_technical_error	404871 non-null	float64
23	perc_technical_error	404871 non-null	float64
24	cnt_incorrect_data_wo_cvv	404871 non-null	float64
25	perc_incorrect_data_wo_cvv	404871 non-null	float64
26	cnt_incorrect_cvv	404871 non-null	float64
27	perc_incorrect_cvv	404871 non-null	float64
28	cnt_insufficient_funds	404871 non-null	float64
29	perc_insufficient_funds	404871 non-null	float64
30	cnt_emitent_problems	404871 non-null	float64
31	perc_emitent_problems	404871 non-null	float64
32	cnt_do_not_honor	404871 non-null	float64
33	perc_do_not_honor	404871 non-null	float64
34	cnt_stolen_lost_restricted_card	404871 non-null	float64
35	perc_stolen_lost_restricted_card	404871 non-null	float64
36	cnt_antifraud_error	404871 non-null	float64
37	perc_antifraud_error	404871 non-null	float64
38	cnt_3ds_problem	404871 non-null	float64
39	perc_3ds_problem	404871 non-null	float64
40	cnt_other_error	404871 non-null	float64
41	perc_other_error	404871 non-null	float64

Рисунок 3.3 – Загальний опис змінних датасету

З огляду на особливості кожної змінної у наборі даних, пропущені значення були заповнені з урахуванням змістового контексту відповідної ознаки.

Наприклад, для змінної sum_spent_bonuses, яка відображає обсяг витрачених бонусів, відсутні значення було логічно замінено на 0, оскільки їхня відсутність може свідчити про те, що бонуси не використовувались. Натомість для змінної avg_time_diff_between_payments, що описує середній інтервал між транзакціями, встановлення нульового значення є некоректним, адже такий інтервал не може бути рівним нулю. У цьому випадку пропущені значення було замінено на -1 як маркер їх відсутності або невизначеності.

Розподіл змінної sum_spent_bonuses після заповнення наведено на рисунку 3.4.

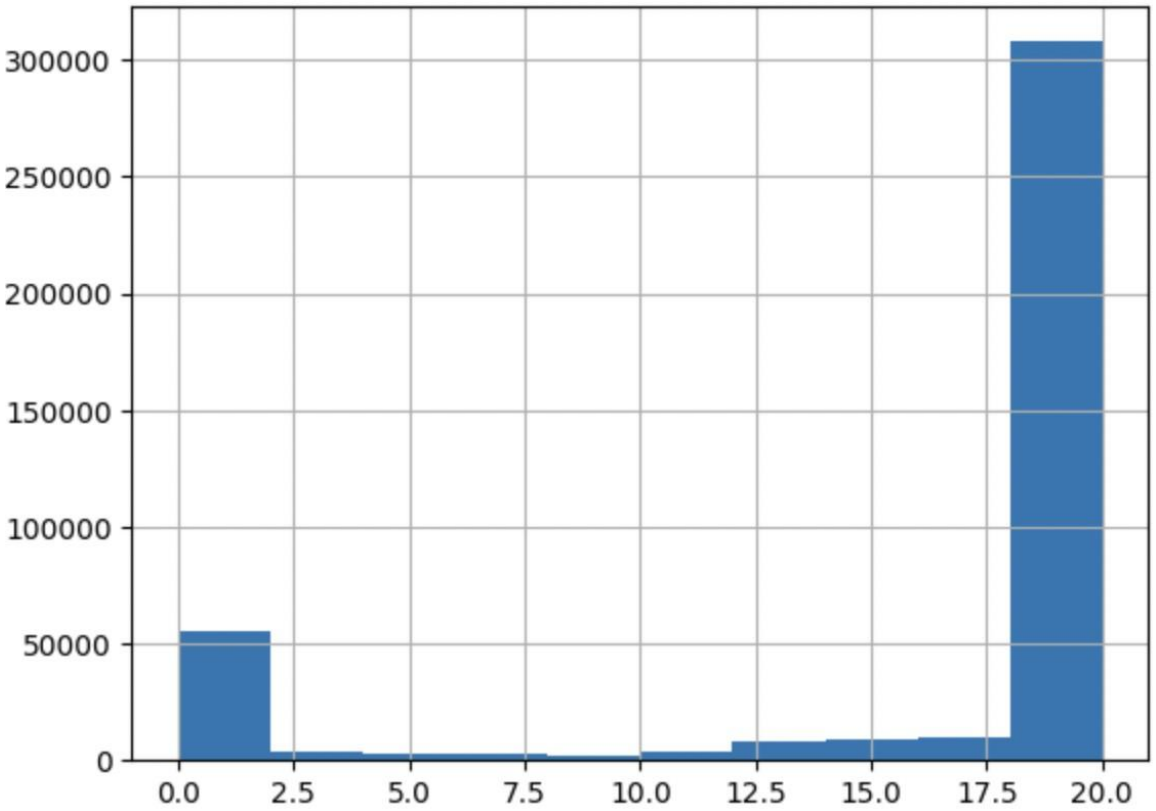


Рисунок 3.4 – Розподіл змінної *sum_spent_bonuses*

Після заповнення пропущених значень категоріальні ознаки були закодовані методом One-hot-encoding. Вибір цього підходу пояснюється відсутністю природного порядку серед значень таких змінних, тому кодування має уникати штучного надання їм ієрархії.

Для подальшого аналізу даних важливо виявити можливі зв’язки між ознаками та цільовою змінною, що дозволяє краще зрозуміти внутрішню структуру набору даних і вплив окремих характеристик. З цією метою була побудована кореляційна матриця, у якій коефіцієнти кореляції між змінними розраховано за методом Пірсона.

Рисунок 3.5 демонструє найбільш інформативну частину кореляційної матриці — стовпець, що містить значення коефіцієнтів кореляції між незалежними змінними та цільовою ознакою.

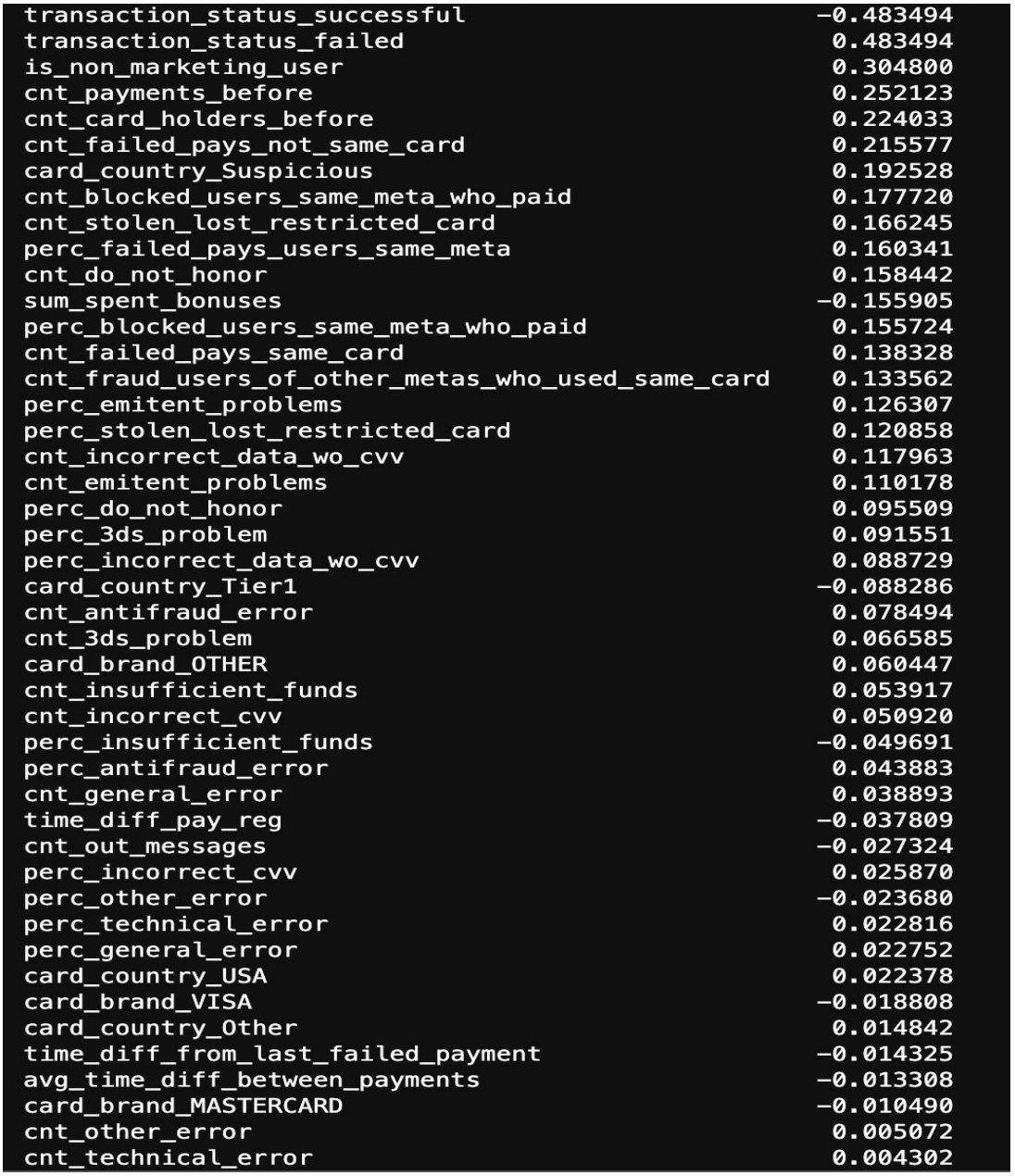


Рисунок 3.5 – Коефіцієнти кореляції між незалежними і цільовою змінною

На рисунку 3.5 наведено коефіцієнти кореляції між незалежними змінними та цільовою змінною is_fraud. Найсильніший зв’язок спостерігається у змінних transaction_status_successful і transaction_status_failed, які мають майже дзеркальні значення кореляції. Це дозволяє визначити, які саме ознаки найбільше впливають на виявлення аномальних транзакцій, і є важливою частиною етапу відбору ознак у побудові моделей машинного навчання.

3.3.Архітектура програмного продукту

Розроблена система являє собою готову модель машинного навчання, призначену для виявлення аномальних записів (викидів) у наборах даних у режимі реального часу з можливістю інтеграції в різноманітні програмні середовища.

Для забезпечення стабільної та ефективної роботи моделі реалізовано повний процес обробки, підготовки та збагачення даних, що дає змогу автоматизувати прийом і обробку нових даних. Виявлені аномалії можуть зберігатися у базі даних для подальшого моніторингу ефективності моделі, аналізу нових типів викидів і, за необхідності, її перенавчання.

У межах дослідження було обрано кілька поширених алгоритмів машинного навчання для порівняння ефективності: логістична регресія, дерево рішень, випадковий ліс, XGBoost та метод опорних векторів (SVM).

Після етапу навчання й тестування оцінка якості кожної моделі проводилася за допомогою таких метрик, як precision, recall, f1-score та AUC, які є найбільш показовими для задач виявлення викидів.

На основі аналізу отриманих результатів було визначено модель, що продемонструвала найвищу ефективність. У таблиці 3.1 представлено порівняння основних показників якості для всіх розглянутих моделей. Таблиця 3.1 – Порівняльна оцінка метрик якості різних алгоритмів.

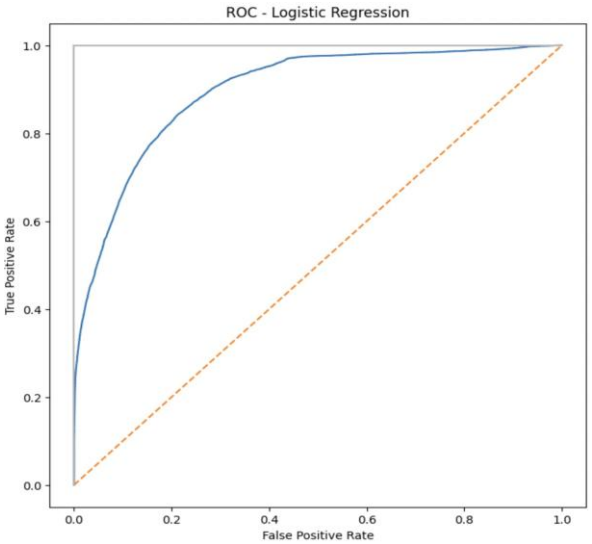
Алгоритм	Precision	Recall	F1-score	AUC
Logistic Regression	0.75	0.582	0.655	0.896
Decision Tree	0.74	0.692	0.715	0.911
Random Forest	0.769	0.687	0.726	0.921
XGBoost	0.801	0.688	0.74	0.93
SVM	0.757	0.637	0.692	0.892

У таблиці наведено порівняння результатів роботи різних алгоритмів машинного навчання за чотирма основними метриками: Precision, Recall, F1-score та AUC. Найвищу точність (Precision = 0.801) та площу під кривою ROC (AUC = 0.93) продемонструвала модель XGBoost, що підтверджує її ефективність для виявлення викидів. Водночас найвищий F1-score (0.74) теж належить XGBoost, що робить її найбільш збалансованою моделлю за всіма критеріями. Це свідчить про доцільність її застосування у завданнях класифікації аномальних транзакцій.

Аналіз значень отриманих метрик свідчить про те, що модель, побудована з використанням алгоритму XGBoost, показала найвищу ефективність серед усіх протестованих підходів — як за метрикою f1-score, так і за показником AUC.

Цей результат є очікуваним, оскільки XGBoost вважається одним із найпотужніших класичних алгоритмів машинного навчання, особливо ефективним у задачах виявлення складних аномаопиши мені ці криві як завершення розділу і дипломної в цілому лій у великих масивах даних.

На рисунках 3.6–3.10 представлено графіки ROC-кривих для кожної з розглянутих моделей.



.Рисунок 3.6 – ROC-крива для Logistic Regression

ROC-крива для моделі Logistic Regression, яка відображає співвідношення між рівнем істинно позитивних результатів (True Positive Rate) і хибно позитивних (False Positive Rate). Крива розміщується вище діагоналі випадкового класифікатора, що свідчить про достатню якість моделі. Чим ближче крива до верхнього лівого кута, тим вища точність класифікації — у цьому випадку логістична регресія демонструє задовільну здатність до виявлення викидів у даних.

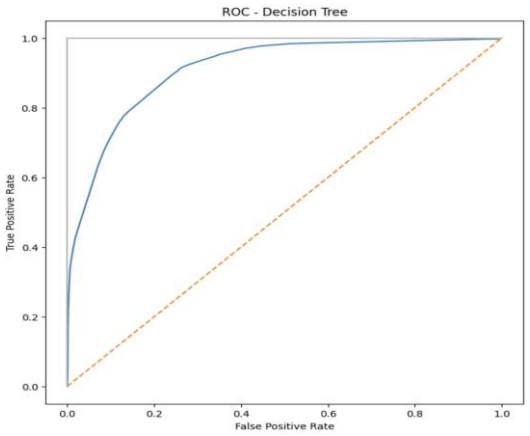


Рисунок 3.7 – ROC-крива для Decision Tree

На графіку показано ROC-криву для моделі Decision Tree, яка демонструє високу ефективність у виявленні викидів. Крива пролягає значно вище лінії випадкового класифікатора, що свідчить про добру здатність моделі відокремлювати нормальні транзакції від аномальних. Це підтверджує, що дерево рішень є ефективним базовим підходом у задачах виявлення аномалій.

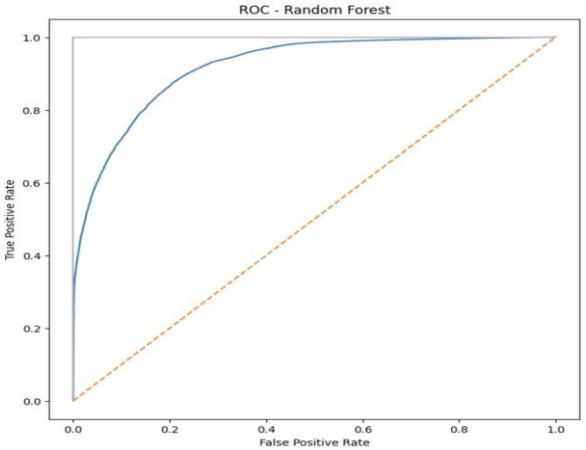


Рисунок 3.8 – ROC-крива для Random Forest

На графіку зображено ROC-криву моделі Random Forest, яка демонструє високі показники класифікації. Крива проходить близько до лівого верхнього кута, що вказує на відмінну здатність моделі розрізняти нормальні та аномальні транзакції. Це робить Random Forest одним із найнадійніших методів для задач виявлення викидів.

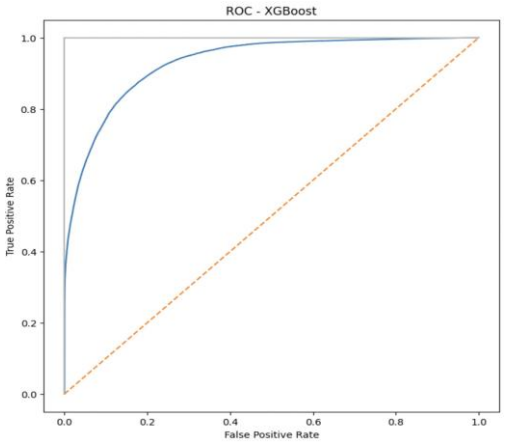


Рисунок 3.9 – ROC-крива для XGBoost

На графіку зображено ROC-криву для моделі XGBoost, яка демонструє високу ефективність класифікації. Крива проходить майже впритул до верхнього лівого кута, що свідчить про відмінну чутливість і специфічність моделі, а отже, її здатність точно виявляти аномальні транзакції. Це підтверджує доцільність застосування XGBoost для виявлення викидів у великих наборах даних.

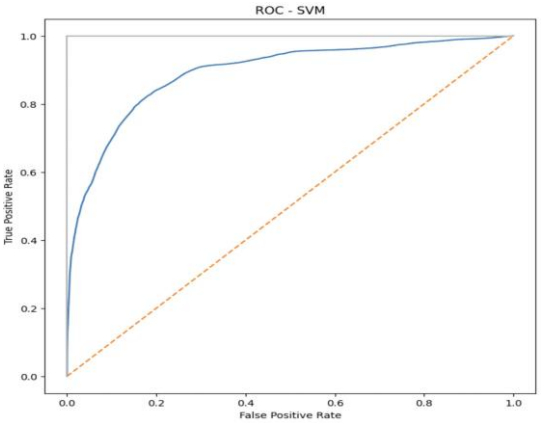


Рисунок 3.10 – ROC-крива для SVM

На рисунку представлено ROC-криву для моделі SVM (Support Vector Machine). Хоча вона демонструє помірну здатність до класифікації, її ефективність дещо поступається таким моделям, як XGBoost чи Random Forest. Це вказує на те, що SVM може бути менш чутливим до виявлення аномалій у транзакційних даних.

Кожна з кривих демонструє здатність відповідної моделі розрізняти нормальні та аномальні записи залежно від обраного порогу класифікації.

З аналізу графіків видно, що всі моделі мають певну дискримінативну здатність, проте найбільшу площу під кривою (AUC) показує алгоритм XGBoost, що вказує на його найкращу ефективність серед протестованих підходів. Це підтверджує, що XGBoost найточніше справляється із завданням виявлення викидів, зберігаючи високу чутливість і специфічність навіть при роботі з дисбалансованими даними.

Таким чином, результати експериментального дослідження дозволили здійснити об'єктивну оцінку якості різних методів машинного навчання для виявлення аномалій. Побудовані моделі пройшли порівняльний аналіз за ключовими метриками, а найефективніша — XGBoost — рекомендована для використання в практичних задачах аналізу транзакційних або інших великих наборів даних. Це підсумовує основну мету дипломної роботи та демонструє практичну значущість застосованого підходу.

РОЗДІЛ 4.
ОХОРОНА ПРАЦІ ТА БЕЗПЕКА В НАДЗВИЧАЙНИХ СИТУАЦІЯХ

4.1. Структурно-функціональний аналіз технологічного процесу

Розробка та вживання ефективних заходів запобігання аварійним і травмонебезпечних ситуаціям можливі лише при завчасному виявлення тих небезпек, з яких починаються процеси їх формування. Оскільки небезпечні умови не завжди завчасно можна виявити, а для вивчення небезпечних дій іноді потрібно багато часу, щоб зібрати статичний матеріал, то і методи виявлення цих небезпек повинні бути відповідно диференційовані (табл. 5.1).

Таблиця 4.1. - Моделі формування та виникнення травмонебезпечних і аварійних ситуацій

Вид робіт, виробничий підрозділ, робоче місце, виробниче обладнання, склад агрегату	Виробнича небезпека			Можливі наслідки	Заходи запобігання небезпечним ситуаціям
	Небезпечна умова (НУ)	Небезпечна дія (НД)	Небезпечна ситуація (НС)		
Виконання робіт із електрообладнанням	Не вимкнено живлення. Відсутність заземлення.	Нехтування правилами ТБ	Ураження струмом	Травма (Т)	Проведення повторного інструктажу з ТБ. Розробка нових способів захисту. Встановлення заземлення.
НД НУ → НС → Т					

Відповідно до аналізу небезпечних умов, які існують у виробничому процесі виокремлено такі наступні за характером дії на працівника їх групи:

- характеризують стан або рівень небезпеки обладнання, які використовуються.

- сприяють виникненню технологічних помилок обслуговуючого персоналу впродовж виробничого процесу;
- створювати умови та можливість проникнення працівника в небезпечну зону;
- приводять до виникнення небезпечних дій (внаслідок низького рівня професійної підготовки працівників та організації навчання з охорони праці).

Моделі формування та виникнення травмонебезпечних і аварійних ситуацій в комп'ютерному кабінеті представлено у вигляді моделі формування та виникнення травмонебезпечних і аварійних ситуацій – табл. 5.1.

4.2. Розрахунок освітлення приміщення комп'ютерного кабінету.

Освітленість виробничих приміщень може бути штучною і природною. Природне освітлення при правильному обладнанні найбільш сприятливе для людини. Основні вимоги для освітлення наступні:

- освітлення повинне бути достатнім для швидкого і легкого розпізнання об'єктів роботи;
- освітлення повинно бути рівномірне без різких тіней;
- джерело світла не повинно осліплювати працівника;
- рівень освітленості не повинен обмежуватись часом.

Природне освітлення забезпечується обладнанням вікон (бокове освітлення) фонарів і світильних покрить приміщень (верхнє освітлення). Природне освітлення нормується коефіцієнтом природної освітленості. Коефіцієнт природної освітленості – це процентне відношення фактичної освітленості Fв в будь-якій точці приміщення до освітленості Fн розсіяної світлом небозводу точки, яка лежить на відкритій місцевості. Розрахунок природного освітлення через бокові вікна по нормам освітленості ведеться для самої дальньої від вікон точки, тобто знаходять мінімальне значення стик

коефіцієнта природної освітленості:

$$e_{\min} = \frac{Fb}{Fн} \cdot 100. \qquad (4.1)$$

Значення коефіцієнта природної освітленості визначається не менше чим в п’яти точках. Значення коефіцієнта природної освітленості для сільськогосподарських виробничих приміщень в даному випадку ремонтній майстерні, беремо $e_{min} = 5\%$.

Розрахунок природного освітлення зводиться до визначення площі світлових променів.

Сумарну площу світлових променів $\sum F_o\,(м^2)$ по коефіцієнту природної освітленості для бокових променів визначаємо по формулі:

$$\sum F_o = \frac{F_н \cdot e_{\min} \cdot r_o \cdot K}{100 \cdot \tau \cdot Г_1},$$

(4.2)

де Fн – площа підлоги; e_{min} – величина мінімального коефіцієнта природного освітлення; τ – загальний коефіцієнт світловикористання віконного отвору із врахуванням його забруднення, τ = 0,25; ro – світлова характеристика вікна, ro – 9,5; Г1 – коефіцієнт, який враховує підвищення освітленості за рахунок світла, яке відбивається від стін і стелі, Г1 = 1,2; К – коефіцієнт, який враховує затінення вікон сусідніми приміщеннями і загорожею, К = 1.

$$\sum F_o = \frac{36 \cdot 0,5 \cdot 9,5 \cdot 1}{100 \cdot 0,25 \cdot 1,2} = 5,7 м^2.$$

Кількість світлових променів визначимо:

$$N = \frac{\sum F_o}{F_o},$$

(4.3)

де Fo – площа вікна згідно стандарту, м2.

$$N = \frac{5,7}{6} = 0,95.$$

Приймаємо кількість вікон – одне вікно.

При розрахунку природного освітлення найбільш поширеним і простим є метод світлового потоку. При цьому методі розраховуємо світловий потік $F_л$ (Лк), який повинна випромінювати кожна лампа (при заданій кількості ламп).

$$F_л = \frac{k \cdot S_п \cdot E}{n_л \cdot \eta \cdot r^2},$$ (4.4)

де k – коефіцієнт запасу, $k = 1,3$; S_n – площа підлоги, м2; $S_n = 36$ м2.

E – нормативна освітленість, Е = 300 Лк; $n_л$ – кількість встановлених ламп, $n_л$=6 од; η – коефіцієнт використання світлового потоку, $\eta = 0{,}25$; r – коефіцієнт нерівномірності освітленості, $r = 0{,}545$.

Коефіцієнт запасу (K) враховує можливість забруднення світильників пилом, що залежить від характеру виробництва.

Розрахунок штучного освітлення починаємо із визначення висоти розташування світильника і їх кількості. Висоту $h_н$ (м) розташування світильників над робочим місцем знаходимо за формулою:

$$h_н = H - (h_1 + h_2), \qquad (4.5)$$

де H – висота приміщення, м; h_1 – віддаль від підлоги до освітлювальної поверхні, м; h_2 – віддаль від стелі до світильника, м.

$$h_н = 4{,}5 - (2{,}2 + 1{,}5) = 0{,}8 \text{ м.}$$

При симетричному розміщенні світильників по вершинах квадратів їх кількість визначається за формулою:

$$n_c = \frac{S_n}{l^2}, \qquad (4.6)$$

де l – віддаль між світильниками, м.

Підставивши значення отримаємо:

$$n_c = \frac{36}{9} = 4 \text{ од.}$$

Тоді світловий потік буде становити

$$F_л = \frac{1{,}3 \cdot 36 \cdot 300}{4 \cdot 0{,}25 \cdot 0{,}545} = 2576{,}2 \text{ Лк.}$$

При світловому потоці 2576,2 Лк для заданої лампи вибираємо тип і потужність.

Вибираємо тип лампи – люмінесцентну, потужністю 40Вт.

4.3. Безпека в надзвичайних ситуаціях

Забезпечення захисту населення і території у разі загрози і виникнення надзвичайних ситуацій є одним з найважливіших завдань держави.

Захист населення є системою загально-державних заходів, які реалізуються центральними і місцевими органами виконавчої влади, виконавчими органами влад, органами управління з питань надзвичайних ситуацій та цивільного захисту населення, підпорядкованими їм системами, та підприємств, що забезпечують виконання організаційних, інженерно – технічних, санітарно – гігієнічних, проти епідемічних та інших заходів у сфері запобігання та ліквідації наслідків надзвичайних ситуацій.

Загрози життєво важливих інтересів громадян, держави, суспільства поділяють на зовнішні та внутрішні, виконують під час надзвичайних ситуацій техногенного та природного характеру та воєнних конфліктах.

Принципи захисту випливають з основних положень Женевської конвенції щодо захисту жертв війни та додаткових протоколів до неї, можливого характеру воєнних дій, реальних можливостей держави щодо створення матеріальної бази захисту. З метою захисту населення, зменшення втрат та шкоди економіці в разі виникнення надзвичайних ситуацій має право проводитись спеціальний комплекс заходів.

Оповіщення та інформування, яке досягається завчасним створенням і підтримкою в постійній готовності загально-державної, територіальних та об’єктивних систем оповіщення населення.

ВИСНОВКИ

У контексті стрімкого зростання обсягів цифрових даних та ускладнення сучасних інформаційних систем завдання виявлення викидів (аномалій) у великих наборах даних набуває особливої актуальності. Зокрема, в транзакційних даних викиди можуть свідчити про технічні помилки, системні збої або потенційні шахрайські дії. Традиційні аналітичні інструменти часто виявляються неефективними у виявленні таких нетипових записів, тому зростає потреба у застосуванні гнучкіших і точніших методів, зокрема машинного навчання.

Моделі машинного навчання здатні виявляти приховані закономірності в поведінці користувачів на основі історичних транзакцій. У рамках цієї дипломної роботи було реалізовано та порівняно кілька популярних алгоритмів класифікації, зокрема логістичну регресію, дерево рішень, випадковий ліс, XGBoost і метод опорних векторів (SVM), з метою виявлення аномальних транзакцій у даних.

Оцінювання ефективності моделей проводилося за ключовими метриками бінарної класифікації: precision, recall, f1-score та AUC. Серед усіх протестованих підходів найкращі результати продемонстрував алгоритм XGBoost, який забезпечив високу точність і добре справлявся з виявленням аномалій у складних, незбалансованих наборах даних.

Важливою особливістю реалізованого підходу стало врахування нетипових характеристик транзакцій — таких як підозра на використання кількох акаунтів, джерела залучення користувачів, історія взаємодії з сервісом та інформація про картку. Це суттєво підвищило якість ідентифікації викидів.

У перспективі розвиток системи може бути спрямований на покращення якості вхідних даних, а також на впровадження методів глибокого навчання або спеціалізованих алгоритмів виявлення аномалій, орієнтованих на рідкісні, але критично важливі випадки у транзакційних наборах.

.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Dalal, S., Seth, B., Radulescu, M., Secara, C., Tolea, C. Predicting Fraud in Financial Payment Services through Optimized Hyper-Parameter-Tuned XGBoost Model // Mathematics. 2022. Vol. 10, No. 24. P. 4679. DOI: 10.3390/math10244679.

2. Lindholm, A., Wahlström, N., Lindsten, F., Schön, T.B. Supervised Machine Learning. Lecture notes for the Statistical Machine Learning course. Department of Information Technology, Uppsala University, 2019. P. 111–112.

3. Sharma, V. Survey of Classification Algorithms and Various Model Selection Methods // Journal of Machine Learning Research. 2000. Vol. 1. P. 1–48.

4. Hyeoun-Ae, P. An Introduction to Logistic Regression: From Basic Concepts to Interpretation with Particular Attention to Nursing Domain // Journal of Korean Academy of Nursing. 2013. Vol. 43, No. 2. P. 154–164. DOI: 10.4040/jkan.2013.43.2.154.

5. Rokach, L., Maimon, O. The Data Mining and Knowledge Discovery Handbook. 2005. P. 165–192. DOI: 10.1007/0-387-25465-X_9.

6. Entropy and Information Gain to Build Decision Trees in Machine Learning [Electronic resource]. – Available at: https://www.section.io/engineering-education/entropy-information-gain-machine-learning/ (Accessed: 20.04.2023).

7. Decision Trees Explained — Entropy, Information Gain, Gini Index, CCP Pruning [Electronic resource]. – Available at: https://towardsdatascience.com/decision-trees-explained-entropy-information-gain-gini-index-ccp-pruning-4d78070db36c(Accessed: 22.04.2023).

8. Cutler, A., Cutler, R.D., Stevens, R.J. Ensemble Machine Learning: Methods and Applications. 2011. P. 157–176. DOI: 10.1007/978-1-4419-9326-7_5.

9. Wang, Y., Pan, Z., Zheng, J., Qian, L., Li, M. A hybrid ensemble method for pulsar candidate classification // Astrophysics and Space Science. 2019. Springer Nature B.V. DOI: 10.1007/s10509-019-3602-4.

10. Is Machine Learning Beneficial in Detecting Fraud? [Electronic resource]. – Available at: https://medium.com/mlearning-ai/how-beneficial-is-detecting-fraud-using-machine-learning-b00089d84930 (Accessed: 28.04.2023).

11. XGBoost: its Genealogy, its Architectural Features, and its Innovation [Electronic resource]. – Available at: https://towardsdatascience.com/xgboost-its-genealogy-its-architectural-features-and-its-innovation-bf32b15b45d2(Accessed: 28.04.2023).

12. 24 Evaluation Metrics for Binary Classification (And When to Use Them) [Electronic resource]. – Available at: https://neptune.ai/blog/evaluation-metrics-binary-classification (Accessed: 01.05.2023).

13. Amaratunga, D., Cabrera, J., Lee, Y.-S. Enriched random forests // Bioinformatics. 2008. Vol. 24, No. 18. P. 2010–2014. DOI: 10.1093/bioinformatics/btn356.

14. Vujovic, D.Z. Classification Model Evaluation Metrics // International Journal of Advanced Computer Science and Applications. 2021. Vol. 12, No. 6.

ДОДАТОК А
ЛІСТИНГ ПРОГРАМНОГО ПРОДУКТУ

```
import pandas as
pd import numpy
as np
import
google.oauth2.credentials
import getpass
import
Levenshtein
import pycountry

from google.cloud import bigquery
from sklearn.preprocessing import OneHotEncoder
from sklearn.cluster import AgglomerativeClustering as HCLUST

import warnings
warnings.filterwarnings('ignore')

token = getpass.getpass()
creds_bq = google.oauth2.credentials.Credentials(token=token)

### Count distinct card holders
LEVENSTEIN_RATIO = 0.2

def levenstein_nunique(card_holders):
    card_holders = [str(x) for x in
    card_holders] if len(card_holders) > 1000:
        return -1
    ds = np.zeros((len(card_holders),
    len(card_holders))) if len(card_holders) < 2:
```

```
        return
    len(card_holders) else:
        for i1 in range(len(card_holders)):
            for i2 in range(i1,
                len(card_holders)): d =
Levenshtein.distance(card_holders[i1].lower(),card_holders[i2].lower())
```