МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Західноукраїнський національний університет
Факультет комп'ютерних інформаційних технологій
Кафедра інформаційно-обчислювальних систем і управління

СУКМАНОВСЬКИЙ Олесь Васильович

Інтелектуальний метод визначення шахрайських інтернет-магазинів / Intelligent Method for Detecting Fraudulent Online Stores

спеціальність: 122 - Комп'ютерні науки
освітньо-професійна програма - Комп'ютерні науки

Кваліфікаційна робота

Виконав студент групи КНм-21
О. В. Сукмановський
_____

Науковий керівник:
к.т.н., доцент, Ліп'яніна-Гончаренко Х.В.
_____

Кваліфікаційну роботу
допущено до захисту:
«___» _____ 20___ р.
Завідувач кафедри
_____ М.П. Комар

ТЕРНОПІЛЬ - 2023

Факультет комп'ютерних інформаційних технологій
Кафедра інформаційно-обчислювальних систем і управління
Освітній ступінь «магістр»
спеціальність: 122 – Комп'ютерні науки
освітньо-професійна програма – Комп'ютерні науки

ЗАТВЕРДЖУЮ
Завідувач кафедри
_____ М.П. Комар
«____» _____20__р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ СТУДЕНТУ
Сукмановському Олесю Васильовичу

(прізвище, ім'я, по батькові)

1. Тема кваліфікаційної роботи
Інтелектуальний метод визначення шахрайських інтернет-магазинів / Intelligent Method for Detecting Fraudulent Online Stores
керівник роботи к.т.н., доцент, Ліп'яніна-Гончаренко Х.В.
затверджені наказом по університету від 8 грудня 2022 року № 491.

2. Строк подання студентом закінченої кваліфікаційної роботи 1 грудня 2023 р.

3. Вихідні дані до кваліфікаційної роботи: завдання на кваліфікаційну роботу студента, наукові статті, технічна література.

4. Основні питання, які потрібно розробити
1. Вивчення суті та особливостей поняття шахрайських інтернет-магазинів.
2. Аналіз існуючих підходів до виявлення шахрайських інтернет-магазинів.
3. Визначення ідентифікаторів потенційної небезпеки в інтернет-торгівлі.
4. Розробка проекту системи визначення шахрайських інтернет-магазинів
5. Опис методів класифікації на основі машинного навчання
6. Розробка інтелектуального методу визначення шахрайських інтернет-магазинів
7. Опис реалізації запропонованого методу.
8. Розвідувальний аналіз даних.
9. Реалізація запропонованого методу.
10. Створення прототипу системи визначення шахрайських інтернет-магазинів.

5. Перелік графічного матеріалу у роботі
Модель системи визначення шахрайських інтернет-магазинів
Структура інтелектуального методу визначення шахрайських інтернет-магазинів
Результати моделювання

6. Консультанти розділів кваліфікаційної роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		Завдання видав	Завдання прийняв

7. Дата видачі завдання 8 грудня 2022 р.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів кваліфікаційної роботи	Строк виконання етапів кваліфікаційної роботи	Примітка
1	Сучасний стан предметної області і постановка задачі дослідження	12.2022 р. – 03.2023 р.	
2	Визначення шахрайських інтернет-магазинів	03.2023 р. – 05.2023 р.	
3	Реалізація визначення шахрайських інтернет-магазинів	05.2023 р. – 11.2023 р.	
4	Повне завершення та представлення кваліфікаційної роботи на кафедру	01.12.2023 р.	

Студент _____ О. В. Сукмановський
підпис

Керівник роботи _____ к.т.н., доцент, Ліп’яніна-Гончаренко Х.В.
підпис

РЕЗЮМЕ

Кваліфікаційна робота на тему "Інтелектуальний метод визначення шахрайських інтернет-магазинів" написана для отримання ступеня вищої освіти "Магістр" зі спеціальності 122 «Комп’ютерні науки» освітньо-професійної програми «Комп’ютерні науки» та включає в себе 96 сторінок тексту, 13 рисунків, 2 таблиці та список літератури із 23 джерелами.

Метою даної роботи є розробка та застосування інтелектуального методу для виявлення шахрайських інтернет-магазинів. У роботі використовуються методи машинного навчання та аналізу даних для покращення ефективності виявлення шахрайських сайтів.

Результати дослідження показують, що розроблений метод має високий потенціал у виявленні шахрайських інтернет-магазинів та сприяє підвищенню довіри до електронної комерції. Ця робота може бути корисною для боротьби з онлайн шахрайством та захисту споживачів в інтернеті.

Ключові слова: ШАХРАЙСЬКІ ІНТЕРНЕТ-МАГАЗИНИ, МАШИННЕ НАВЧАННЯ, АНАЛІЗ ДАНИХ, ВИЯВЛЕННЯ ШАХРАЙСТВА.

RESUME

The qualification work on the topic " Intelligent Method for Detecting Fraudulent Online Stores" was written to obtain a Master's degree in the field of Computer Science, specialization 122, within the educational-professional program of Computer Science. The work comprises 96 pages of text, 13 figures, 2 tables, and a bibliography with 23 sources.

The objective of this research is to develop and apply an intelligent method for the detection of fraudulent internet stores. The study utilizes machine learning methods and data analysis to enhance the efficiency of identifying fraudulent websites.

The research results demonstrate that the developed method holds significant potential in detecting fraudulent internet stores and contributes to increasing trust in e-commerce. This work can be valuable in combating online fraud and safeguarding consumers on the internet.

Keywords: FRAUDULENT INTERNET STORES, MACHINE LEARNING, DATA ANALYSIS, FRAUD DETECTION.

.

ЗМІСТ

ВСТУП

Тема інтелектуального методу визначення шахрайських інтернет-магазинів є надзвичайно актуальною в сучасному цифровому світі. Зростання обсягів онлайн-торгівлі та використанням електронних платежів призводить до збільшення кількості шахрайських інтернет-магазинів, які намагаються обманювати споживачів і заволодіти їхніми фінансовими ресурсами. Одночасно з цим з'являється ріст занепокоєння щодо безпеки та надійності онлайн-покупок.

Онлайн-торгівля набуває все більшого поширення, що викликає зростання кількості інтернет-магазинів. Ця динаміка створює ідеальні умови для зростання числа шахрайських сайтів, які намагаються вигідно використовувати цей ринок.

Шахрайські інтернет-магазини можуть завдати серйозної фінансової шкоди споживачам, обманюючи їх і вимагаючи оплату за товари або послуги, які фактично не існують. Це може вплинути на довіру споживачів до онлайн-торгівлі та зменшити їхню активність в цьому сегменті ринку.

Шахрайство в електронній комерції може відобразитися на загальному довір'ї споживачів до інтернету та цифрової економіки. Важливо захищати споживачів від обману та забезпечувати їхню безпеку в онлайн-середовищі.

Сучасні методи машинного навчання та обробки даних надають інструменти для розробки інтелектуальних систем виявлення шахрайських магазинів з високою точністю та ефективністю.

Інтелектуальні методи можуть значно поліпшити виявлення шахрайських інтернет-магазинів, зменшуючи ризик для споживачів та підвищуючи довіру до електронної комерції.

У зв'язку з цим, можна вважати, що розробка і впровадження інтелектуального методу виявлення шахрайських інтернет-магазинів є критично важливим завданням для забезпечення безпеки та надійності онлайн-торгівлі.

Мета даної роботи полягає у розробці та впровадженні інтелектуального методу для визначення шахрайських інтернет-магазинів. Цей метод має на меті використання сучасних технологій машинного навчання та аналізу даних для ефективного ідентифікування та класифікації потенційно небезпечних онлайн-торгових платформ. Основна увага зосереджується на розробці надійної та точної системи, яка здатна автоматично аналізувати різноманітні характеристики веб-сайтів та визначати їхню надійність.

Досягнення цієї мети зумовило потребу теоретичних розробок, визначення та послідовного вирішення таких завдань:

1. Вивчення суті та особливостей поняття шахрайських інтернет-магазинів.
2. Аналіз існуючих підходів до виявлення шахрайських інтернет-магазинів.
3. Визначення ідентифікаторів потенційної небезпеки в інтернет-торгівлі.
4. Розробка проекту системи визначення шахрайських інтернет-магазинів
5. Опис методів класифікації на основі машинного навчання

6. Розробка інтелектуального методу визначення шахрайських інтернет-магазинів
7. Опис реалізації запропонованого методу.
8. Розвідувальний аналіз даних.
9. Реалізація запропонованого методу.
10. Створення прототипу системи визначення шахрайських інтернет-магазинів.

Об'єктом дослідження є процес виявлення шахрайських інтернет-магазинів.

Предметом дослідження є розробка та застосування інтелектуального методу для ефективного виявлення шахрайських інтернет-магазинів.

Методи дослідження: у роботі використовувалися різні методи, такі як аналіз даних, машинне навчання, класифікація на основі різних алгоритмів (DecisionTree, Logistic Regression, Random Forest, KNN, SVM, Naive Bayes) та методи обробки незбалансованих даних, такі як Imbalanced, Undersampling, Oversampling, SMOTE та ADASYN.

Наукова новизна отриманих результатів полягає в розробці та застосуванні методу ідентифікації шахрайських інтернет-магазинів, який використовує інтелектуальний підхід, а також методи машинного навчання та обробки незбалансованих даних. Цей метод дозволяє досягнути високої точності та надійності при виявленні шахрайських дій в електронній комерції. Результати дослідження також показують, що певні підходи, такі як ADASYN та Oversampling, можуть досягати надзвичайно високих показників ефективності у роботі з незбалансованими даними. Ці результати важливі для подальшого

розвитку та вдосконалення систем виявлення шахрайських інтернет-магазинів та захисту користувачів від онлайн шахрайства.

Практичне значення одержаних результатів полягає в створенні інтелектуального методу ідентифікації шахрайських інтернет-магазинів, який може бути використаний для покращення безпеки та довіри в електронній комерції. Цей метод дозволяє користувачам більш ефективно виявляти потенційно шахрайські сайти, зменшуючи ризик фінансових втрат та інших негативних наслідків. Використання розробленого методу може покращити якість обслуговування в електронній комерції та сприяти зростанню довіри споживачів до онлайн покупок.

Структура та обсяг роботи. Дипломна робота включає в себе вступ, три розділи, висновки, список літератури та додатки. Загальний обсяг роботи складає 96 сторінок тексту, який містить 13 рисунків та 2 таблиці. У роботі наведено 23 джерела літератури.

Апробація результатів дослідження. Основні теоретичні положення роботи й практичні результати дослідження доповідалися й обговорювалися: IV Всеукраїнської мультидисциплінарної студентської наукової конференції «Експериментальні та теоретичні дослідження в контексті сучасної науки», яка відбулася 29 вересня 2023 року у місті Чернігів, Україна; X Міжнародній науково-практичній конференції "Актуальні питання розвитку науки та освіти" у м. Львів, 9-10 грудня 2023 року.

1 СУЧАСНИЙ СТАН ПРЕДМЕТНОЇ ОБЛАСТІ І ПОСТАНОВКА ЗАДАЧІ ДОСЛІДЖЕННЯ

1.1 Суть та особливості поняття шахрайських інтернет-магазинів

Шахрайство в Інтернеті є широким та багатогранним явищем, що охоплює різні види діяльності, від крадіжки особистих даних до фінансових обманів. Розглянемо ключові аспекти інтернет-шахрайства та чому шахрайські інтернет-магазини є особливо значущою його частиною:

1. Різноманітність Інтернет-шахрайств.
2. Шахрайські інтернет-магазини як ключовий елемент
3. Фінансові та особисті втрати.
4. Вплив на довіру до електронної комерції.
5. Складність виявлення та протидії.

Особлива небезпека шахрайських інтернет-магазинів полягає у їх безпосередньому впливі на кінцевих користувачів і споживачів, які часто не підозрюють про можливість обману до моменту, коли втрати вже сталися. З цієї причини боротьба з такими шахрайствами та освіта споживачів щодо потенційних ризиків є критично важливими.

Інтернет-шахрайство охоплює різні форми обману в мережі, кожна з яких використовує унікальні методи для введення в оману користувачів. Одним з найпоширеніших видів є фішинг, де шахраї, маскуючись під легітимні компанії, виманюють чутливі дані, такі як

паролі чи банківські дані, за допомогою обманних електронних листів або повідомлень.

Іншим розповсюдженим методом є використання шкідливого програмного забезпечення, яке може бути непомітно встановлено на комп'ютерах або мобільних пристроях. Це програмне забезпечення здатне красти інформацію, стежити за активністю користувачів або навіть блокувати доступ до важливих даних, часто з метою викупу. Крадіжка ідентичності також є серйозною загрозою, оскільки шахраї можуть використовувати чужу особисту інформацію для відкриття рахунків або здійснення покупок.

Крім того, інтернет-шахрайство включає махінації з криптовалютами, де шахраї пропонують фальшиві інвестиційні можливості або крадуть криптовалюту. Також поширені інші обманні схеми, як-от шахрайства на аукціонах, романтичні шахрайства, чи обман на сайтах знайомств. У світлі цієї різноманітності та складності, користувачам Інтернету важливо захищати свої особисті дані та бути обережними під час онлайн-діяльності.

Відповідно до чинного Українського законодавства інтернет-магазин це засіб для представлення або реалізації товару, роботи чи послуги шляхом вчинення електронного правочину. Відповідно ж, реалізація товару дистанційним способом це укладення електронного договору на підставі ознайомлення покупця з описом товару, наданим продавцем у порядку, визначеному діючими нормами, шляхом забезпечення доступу до каталогів, проспектів, буклетів, фотографій тощо з використанням інформаційно-телекомунікаційних систем, телевізійним, поштовим, радіозв’язком або в інший спосіб, що

виключає можливість безпосереднього ознайомлення покупця з товаром або із зразками товару під час укладення такого договору.

Здійснюючи покупку в Інтернеті, користувачу пропонують різноманітні можливості оплати. Розрахунки у сфері електронної комерції проводять відповідно до Закону України «Про електронну комерцію», Закону України «Про платіжні системи та переказ коштів в Україні», Закону України «Про фінансові послуги та державне регулювання ринків фінансових послуг» тощо. Такі розрахунки можна здійснювати з використанням платіжних інструментів, електронних грошей, шляхом переказу коштів або оплати готівкою з дотриманням вимог законодавства щодо оформлення готівкових і безготівкових розрахунків, а також в інший спосіб, передбачений законодавством України.

Звичайно, можливість вчинення шахрайських дій створюється сприятливими умовами, які безпосередньо складають основу для таких дій і на які можуть впливати дві групи факторів: зовнішні (формують насамперед матеріальні передумови та породжують наміри вчинити шахрайські дії, здійснюючи об’єктивний вплив на потенційних шахраїв) та внутрішні (сприяють здійсненню шахрайських намірів, виступаючи своєрідним «каталізатором») [1].

Варто відмітити, що в сучасних умовах спостерігається значна консолідація ІТ-шахраїв у групи, що дозволяє швидко обмінюватися інформацією про обманні об'єкти шляхом надання бот-мереж, обміну техніками та методами вчинення злочинів, способів перекладу електронних готівкових грошей, тощо [2]. Так, лише у 2018 році співробітники кіберполіції України зупинили діяльність 18 злочинних

груп, які представили свою діяльність як роботу різних фінансових проектів.

Проте часто визначення шахрайського сайту займає велику кількість часу, тому у зв’язку з цим розробка інтелектуального методу визначення шахрайських інтернет-магазинів, дасть можливість автоматизувати процес виявленню шахрайський інтернет-магазинів.

Отже, шахрайські інтернет-магазини становлять значну загрозу в сучасному цифровому світі, де вони не тільки спричиняють фінансові та особисті втрати для споживачів, але й підривають довіру до електронної комерції. Ці шахрайства варіюються від фішингу та використання шкідливого програмного забезпечення до крадіжки ідентичності та махінацій з криптовалютами. Українське законодавство визначає інтернет-магазин як засіб для реалізації товарів чи послуг через електронні правочини, але це також відкриває двері для шахрайських дій. З огляду на складність виявлення та протидії таким шахрайствам, існує важлива потреба в освіті споживачів та розробці інтелектуальних методів для автоматизації процесу виявлення шахрайських інтернет-магазинів. Консолідація ІТ-шахраїв у групи та їх здатність швидко обмінюватися інформацією та методами злочинів ще більше підсилює необхідність ефективних контрзаходів у цій сфері.

1.2 Аналіз існуючих підходів

У статті [13] розглянуто й проаналізовано різноманітні стратегії протистояння шахрайству в онлайн середовищі.

У дослідженні, описаному в статті [7], висвітлено успішні методи запобігання шахрайству в мережі Інтернет, які використовують

перевірку баз даних шахрайських веб-ресурсів та спеціалізовані служби перевірки веб-сайтів, що ведуть нечесну діяльність в Інтернеті.

У роботі [8] розроблено й запропоновано ефективний метод виявлення фішингових веб-сайтів. Це досягнуто за допомогою застосування методу дерева рішень C4.S та аналізу особливостей URL-адрес.

У дослідженні, описаному у роботі [9], розглядається метод, що ґрунтується на використанні ансамблю для класифікації рекламних показів в Інтернеті як потенційно шахрайських чи нормальних. Цей підхід спрямований на визначення кожного показу реклами з точки зору його легітимності чи підозрілості.

У дослідженні [10] пропонується узагальнена модель для аналізу даних про шахрайство в часі, яка базується на виявленні аномалій через використання ймовірностей або врахування часових аспектів у моделюванні часових рядів. Ця модель спроектована для розпізнавання несправедливих дій на основі відбитків часу, таких як періоди в хвилинах чи годинах.

У роботі [11] запропоновано ефективну модель класифікації шахрайства, яка оперує змінними клацаннями в онлайн середовищі. Цей підхід використовує нейронні мережі зворотного поширення, що комбінуються із новим алгоритмом, відомим як "Штучна колонія бджіл", для виявлення шахрайських дій у мережі за допомогою взаємодії з динамічними змінами у кліканні.

У дослідженні [12] пропонується новий метод класифікації, відомий як класифікатор k-найближчих сусідів з використанням методу Quad Division (QDPSKNN). Цей метод впроваджує підхід поділу даних

на чотири частини для ефективної обробки нерівномірного розподілу класів. Основна мета цього підходу - розпізнавання шахрайства в мобільній рекламі (FDMA).

Згадані раніше дослідження не фокусувалися на конкретному виявленні шахрайських інтернет-магазинів через машинне навчання. Отже, мета цієї конкретної статті полягає у розробці інтелектуального методу для ідентифікації шахрайських інтернет-магазинів, з метою автоматизації процесу їх виявлення.

Відмінність цього дослідження від попередніх досліджень, зокрема згаданих у вказаних статтях [7, 8, 13], полягає в тому, що вони, хоч і аналізують сферу шахрайства в Інтернеті, проте не досліджують використання машинного навчання для виявлення шахрайських інтернет-магазинів та розробки відповідного програмного забезпечення.

Таким чином, дане дослідження відзначається своєю спрямованістю на розвиток інтелектуального підходу до виявлення шахрайських інтернет-магазинів, що може привести до автоматизації процесу виявлення та захисту від шахрайства в онлайн середовищі за допомогою передових методів машинного навчання.

Отже, аналіз існуючих підходів до протидії шахрайству в онлайн середовищі, представлений у різних дослідженнях, відкриває широкий спектр методів та стратегій, від використання баз даних шахрайських веб-ресурсів до розробки інноваційних алгоритмів машинного навчання. Особливо важливим є розвиток методів для виявлення фішингових сайтів, класифікації рекламних показів, аналізу даних про шахрайство в часі, та визначення шахрайських дій у мережі. Ці підходи

включають застосування дерев рішень, ансамблів, нейронних мереж, та інших передових технік. Однак, важливо відзначити, що більшість існуючих досліджень не зосереджувалися безпосередньо на виявленні шахрайських інтернет-магазинів через машинне навчання. Тому, розробка спеціалізованого інтелектуального методу для ідентифікації таких магазинів стає ключовим напрямком, який може значно підвищити ефективність виявлення та захисту від шахрайства в онлайн середовищі. Це дослідження вносить важливий внесок у цю область, пропонуючи новітній підхід, який відрізняється від попередніх робіт своєю спеціалізацією на використанні машинного навчання для конкретної мети виявлення шахрайських інтернет-магазинів.

1.3 Ідентифікатори потенційної небезпеки в інтернет-торгівлі

Використовуючи «спеціальне програмне забезпечення, яке створює ілюзію торгівлі на реальних фінансових ринках, можливості SIP-телефонії, що дозволяють замінювати номери, шахраї імітують роботу офісу за кордоном. Крім того, для залучення іноземних інвесторів шахраї вербують іноземців, які можуть спілкуватися з клієнтами рідною мовою (зазвичай це іноземні студенти, яким нібито пропонують легальний заробіток) [3]. Повідомлення про шахрайство в Інтернеті становлять 80% усіх скарг громадян. Найбільш поширеними видами шахрайства в кіберпросторі є продаж неіснуючих товарів, а також фішинг в Інтернет-магазинах. Загалом з початку цього року кіберполіція зареєструвала понад 32 000 звернень громадян [4].

Проблема полягає в тому, що через «анонімність, можливість шахраїв, які використовують системи зв'язку з вільним доступом, купують необмежену кількість сим-карт, створюють акаунти на різних ресурсах, сервери яких знаходяться та належать різним країнам, суттєво ускладнила практичну сторону боротьби з Інтернет-шахрайством. Крім того, наявність низки бюрократичних компонентів під час виконання запитів для розслідування фактів цього виду злочину зводить нанівець роботу правоохоронних органів через довгі терміни» [2]. Важливо підкреслити, що, говорячи про дії інтернет-шахраїв, необхідно розуміти масштаби цієї загрози, оскільки зловмисники винаходять більш досконалі способи та схеми (незважаючи на те, що щоденні банки та платіжні служби, засновані на захищених Exchange- Інтернет-обмін інформацією про підозрілі електронні платежі з ознаками шахрайства), а рівень культури платежів українців залишає бажати кращого. Так, лише за період з 1 серпня по 26 вересня 2017 року платіжні служби зафіксували 12 416 підозр у шахрайських операціях на загальну суму 3409 000 гривень, а в таких операціях брали участь 7390 банківських карток 135 банків із 53 країн (67 українських банків). .Шахраї намагалися вивести кошти на 975 мобільних номерів [5].

За даними Європолу, кіберзлочинність зачіпає всі держави-члени ЄС без винятку і пов’язана насамперед із фінансовими шахрайствами. Згідно з дослідженням Європейської комісії, «8% користувачів Інтернету в ЄС постраждали від крадіжки особистих даних, а 12% - від шахрайства в Інтернеті. Крім того, зловмисне програмне забезпечення завдало шкоди мільйонам сімей, а загальна кількість банківських шахрайств, скоєних за допомогою Інтернету, щороку зростає. Більше

того, згідно з доповіддю держав-членів ЄС про кіберзлочинність, лише близько 30% від загальної кількості кіберзлочинів, таких як шахрайство, викрадення особистих даних (як підготовка до шахрайства), жертви повідомляють правоохоронним органам» [6].

Для визначення шахрайських сайтів пропонується інтелектуальний метод (див.2.3), що базується на наступних параметрах, що представленні в таблиці 1.

Таблиця 2.1 - Параметри

№	**Опис**	**Позначення**
1	Відсутність інформації про призначення товару	V1
2	Відсутність інформації про склад матеріалу, з якого зроблений товар	V2
3	Відсутність інформації про продукцію державною мовою (на українських сайтах).	V3
4	Приховування недоліків товарів (при продажі товарів, що були в користуванні)	V4
5	Відсутність інформації про стан зношеності товарів, якщо вони були у користуванні	V5
6	Відсутність інформації про виробника товару.	V6
7	Відсутність інформації про місце найближчого сервісного обслуговування товару.	V7
8	Відсутність інформації про гарантійний строк товару.	V8
9	Відсутність адреси, за якою покупець може повернути товар протягом чотирнадцяти днів з моменту купівлі.	V9
10	Відсутність інформації про виробника товару.	V10
11	Відсутність інформації про продавця товару (що унеможливлює звернення до суду з позовом).	V11

Кожен з перерахованих параметрів є важливим індикатором, що може свідчити про потенційну ненадійність інтернет-магазину. Детальніше про кожен з них:

1. Відсутність інформації про призначення товару: Це вказує на неповну або недостатню описову інформацію про товар. Покупці повинні отримати чітке уявлення про те, для чого призначений товар, як він використовується та які його функції.
2. Відсутність інформації про склад матеріалу, з якого зроблений товар: Повна інформація про матеріали, з яких виготовлено товар, є важливою з точки зору безпеки та якості. Це також може впливати на алергічні реакції чи особливості догляду за товаром.
3. Відсутність інформації про продукцію державною мовою (на українських сайтах): Для українських сайтів важливо, щоб інформація про товар була доступна українською мовою. Це забезпечує прозорість та зручність для місцевих споживачів.
4. Приховування недоліків товарів (при продажі товарів, що були в користуванні): Чесність щодо стану товару, особливо вживаного, є критичною. Покупці мають право знати про будь-які дефекти або недоліки товару.
5. Відсутність інформації про стан зношеності товарів, якщо вони були у користуванні: Це важливо для оцінки якості та цінності товару, особливо якщо він був у використанні.
6. Відсутність інформації про виробника товару: Вказівка виробника допомагає споживачам перевірити автентичність продукту і забезпечує довіру до якості товару.
7. Відсутність інформації про місце найближчого сервісного обслуговування товару: Ця інформація критично важлива для

вирішення будь-яких питань післяпродажного обслуговування або ремонту товару.

8. Відсутність інформації про гарантійний строк товару: Наявність гарантійного періоду та умов його дії є важливим аспектом захисту прав споживача.
9. Відсутність адреси, за якою покупець може повернути товар протягом чотирнадцяти днів з моменту купівлі: Надання адреси для повернення товару є обов'язковим, щоб забезпечити право споживача на повернення або обмін товару.
10. Відсутність інформації про продавця товару: Інформація про продавця повинна бути доступна для забезпечення можливості контактування та звернення до суду в разі конфліктних ситуацій.

Загалом, наявність цих інформаційних елементів на сайті інтернет-магазину є важливою для забезпечення прозорості, безпеки покупки та захисту прав споживачів. Відсутність такої інформації може бути сигналом для додаткової обережності при роботі з таким сайтом.

Отже, ідентифікація потенційної небезпеки в інтернет-торгівлі є ключовим аспектом у боротьбі з шахрайством в онлайн середовищі. Використання спеціального програмного забезпечення, SIP-телефонії та інших технологій шахраями для імітації легітимної діяльності значно ускладнює виявлення шахрайських дій. Важливість цього питання підкреслюється високим відсотком скарг на шахрайство в Інтернеті та зростаючою кількістю фінансових шахрайств. Проблема ускладнюється анонімністю та бюрократичними перешкодами у розслідуванні таких

злочинів. Визначення шахрайських сайтів за допомогою інтелектуального методу, що базується на аналізі ключових параметрів, представлених у таблиці 2.1, є важливим кроком у напрямку забезпечення безпеки споживачів. Відсутність інформації про товар, його виробника, гарантійні умови, а також інші важливі аспекти можуть бути індикаторами потенційної небезпеки. Таким чином, уважне вивчення цих параметрів може допомогти споживачам уникнути шахрайства та забезпечити безпечні та прозорі онлайн-покупки.

1.4 Постановка задачі

Актуальність даного дослідження полягає у зростаючій потребі в надійних інструментах та методах для ідентифікації та визначення шахрайських інтернет-магазинів. У сучасному цифровому світі, де електронна комерція стрімко розвивається, збільшується кількість інтернет-магазинів, що, в свою чергу, призводить до зростання шахрайських схем та обманів. Це створює значні ризики для споживачів, які можуть втратити гроші, стати жертвами крадіжки особистих даних або придбати неякісні товари.

Дослідження, спрямоване на розробку ефективних методів виявлення та аналізу потенційно шахрайських сайтів, є важливим для забезпечення безпеки онлайн-покупок. Воно допомагає розробити надійні алгоритми та інструменти, які можуть автоматично аналізувати та оцінювати веб-сайти на предмет їхньої надійності, допомагаючи користувачам уникнути взаємодії з шахрайськими ресурсами.

Крім того, це дослідження має велике значення для підвищення обізнаності споживачів щодо цифрової безпеки та розвитку культури безпечних онлайн-покупок. Воно сприяє розробці освітніх матеріалів та

рекомендацій, які допомагають користувачам розпізнавати та уникати шахрайських схем в інтернеті.

Таким чином, актуальність даного дослідження обумовлена не тільки технічними аспектами розробки безпечних інструментів для електронної комерції, але й соціальною важливістю підвищення рівня цифрової безпеки серед споживачів.

Мета даної роботи полягає у розробці та впровадженні інтелектуального методу для визначення шахрайських інтернет-магазинів. Цей метод має на меті використання сучасних технологій машинного навчання та аналізу даних для ефективного ідентифікування та класифікації потенційно небезпечних онлайн-торгових платформ. Основна увага зосереджується на розробці надійної та точної системи, яка здатна автоматично аналізувати різноманітні характеристики веб-сайтів та визначати їхню надійність.

Досягнення цієї мети зумовило потребу теоретичних розробок, визначення та послідовного вирішення таких завдань:

1. Вивчення суті та особливостей поняття шахрайських інтернет-магазинів допоможе зрозуміти ключові характеристики та методи, які використовують шахраї в онлайн-просторі.
2. Аналіз існуючих підходів до виявлення шахрайських інтернет-магазинів забезпечує огляд та оцінку ефективності наявних методів та інструментів.
3. Визначення ідентифікаторів потенційної небезпеки в інтернет-торгівлі сприяє розробці критеріїв для оцінки та класифікації веб-сайтів.
4. Розробка проекту системи визначення шахрайських інтернет-магазинів включає створення концептуальної моделі та архітектури системи.

5. Опис методів класифікації на основі машинного навчання передбачає аналіз різних алгоритмів та їх придатності для виявлення шахрайства.
6. Розробка інтелектуального методу визначення шахрайських інтернет-магазинів полягає у створенні нових, більш ефективних підходів та алгоритмів.
7. Опис реалізації запропонованого методу включає детальний опис процесів та технік, які використовуються у системі.
8. Розвідувальний аналіз даних дозволяє здійснити попередній аналіз та обробку даних, що використовуються для тренування моделей.
9. Реалізація запропонованого методу включає практичне впровадження розроблених алгоритмів та їх тестування.
10. Створення прототипу системи визначення шахрайських інтернет-магазинів демонструє практичне застосування розробленої системи в реальних умовах.

Завдання 1 розглянуто в параграфах 1.1, 1.2 та 1.3 відповідно, а виконання завдань 4-7 представлені у параграфах 2.1-3.4.

Висновки до розділу 1:

1. Шахрайські інтернет-магазини представляють серйозну загрозу в епоху цифровізації, викликаючи не тільки фінансові та особисті втрати для споживачів, але й негативно впливаючи на довіру до онлайн торгівлі. Різноманіття шахрайських схем, включаючи фішинг, використання шкідливого ПЗ, крадіжку особистих даних та махінації з криптовалютами, вимагає комплексного підходу до протидії. Українське законодавство визначає інтернет-магазини як механізм для здійснення електронних правочинів, що відкриває потенціал для шахрайських дій. Тому важливо зосередитися на освіті споживачів та розробці інноваційних методів для виявлення та запобігання шахрайству в інтернеті.
2. Аналізуючи існуючі методи протидії шахрайству в онлайн середовищі, можна відзначити широкий спектр стратегій та технологій, від баз даних шахрайських веб-ресурсів до розробки алгоритмів машинного навчання. Особливо значущими є методи виявлення фішингових сайтів, класифікації рекламних показів та аналізу шахрайських дій. Важливо підкреслити, що багато досліджень не зосереджувались безпосередньо на виявленні шахрайських інтернет-магазинів через машинне навчання, тому розробка спеціалізованих інтелектуальних методів для цієї мети є актуальною та може значно підвищити ефективність виявлення та протидії шахрайству в цифровому просторі.
3. Ідентифікація потенційної небезпеки в інтернет-торгівлі відіграє ключову роль у запобіганні шахрайству в онлайн середовищі. Використання різноманітних технологій шахраями для створення ілюзії легітимної діяльності ускладнює процес виявлення шахрайських дій. Високий рівень скарг на шахрайство в Інтернеті

та зростання фінансових шахрайств підкреслюють важливість цієї проблеми. Використання інтелектуальних методів для аналізу ключових параметрів, які можуть вказувати на шахрайство, є важливим кроком у забезпеченні безпеки споживачів. Уважне вивчення цих параметрів допоможе споживачам уникнути ризиків та забезпечити безпечні та прозорі онлайн-покупки.

2 ВИЗНАЧЕННЯ ШАХРАЙСЬКИХ ІНТЕРНЕТ-МАГАЗИНІВ

2.1 Проектування системи визначення шахрайських інтернет-магазинів

2.1.1 Модель системи

Діаграма (рис.2.1), представляє модель для системи визначення шахрайських інтернет-магазинів, використовуючи інтелектуальний метод. Ось детальний опис її компонентів:

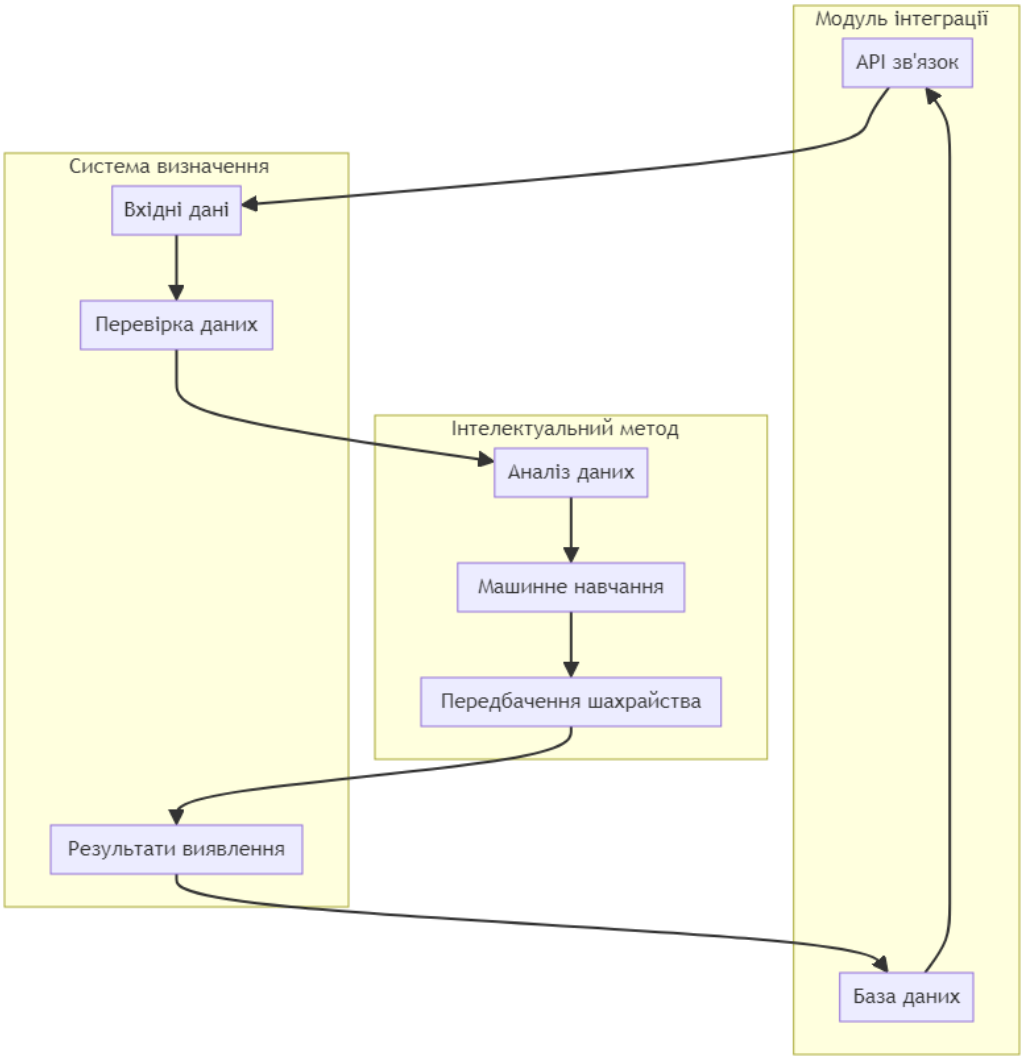

Рисунок 2.1 - Модель системи визначення шахрайських інтернет-магазинів

1. Інтелектуальний метод (IM): Ця частина діаграми представляє основний метод, який використовується для визначення шахрайських інтернет-магазинів. Вона включає в себе наступні підкомпоненти:
    - Аналіз даних (AI): Початковий етап, де збираються та аналізуються дані про інтернет-магазини.
    - Машинне навчання (ML): На цьому етапі використовуються алгоритми машинного навчання для ідентифікації потенційних шахрайських патернів.
    - Передбачення шахрайства (PA): Кінцевий етап інтелектуального методу, де робиться передбачення щодо того, чи є інтернет-магазин шахрайським.
2. Система визначення (DS): Ця частина діаграми відображає процес визначення шахрайських інтернет-магазинів. Вона складається з:
    - Вхідні дані (IN): Початковий етап, де система отримує дані для аналізу.
    - Перевірка даних (VD): Етап, на якому відбувається перевірка та обробка вхідних даних.
    - Результати виявлення (RD): Виведення результатів процесу визначення, які показують, чи є магазин шахрайським.
3. Модуль інтеграції (IMC): Ця частина забезпечує інтеграцію між різними компонентами системи. Вона включає:
    - API зв'язок (API): Інтерфейс для зв'язку між різними частинами системи.
    - База даних (DB): Місце зберігання даних, які використовуються та генеруються системою.

Кожен компонент на діаграмі з'єднаний стрілками, які показують потік даних та процесів між різними частинами системи. Наприклад,

вхідні дані спочатку перевіряються, потім передаються до інтелектуального методу для аналізу, машинного навчання та передбачення. Результати потім зберігаються в базі даних та використовуються для подальшої інтеграції та аналізу.

2.1.2 Діаграма відношень сутностей

Розглянемо структуру компонента "API зв'язок" (API) у системі визначення шахрайських інтернет-магазинів:

1. Інтерфейс API (Application Programming Interface): Це основний компонент, який дозволяє різним частинам системи взаємодіяти між собою. Інтерфейс API виконує наступні функції:
    - Прийом запитів: API приймає запити від зовнішніх джерел (наприклад, веб-інтерфейсу користувача або інших систем) для отримання або відправки даних.
    - Обробка запитів: Після отримання запиту, API обробляє його, визначаючи необхідні дії, такі як доступ до бази даних або взаємодія з іншими компонентами системи.
    - Відправка відповідей: Після обробки запиту, API формує відповідь та відправляє її назад до запитувача.
2. Безпека API: Це критично важливий аспект, який забезпечує, що лише авторизовані користувачі мають доступ до API. Включає в себе:
    - Аутентифікація та авторизація: Перевірка ідентичності користувача та визначення його прав на доступ до певних даних або функцій.
    - Шифрування: Захист даних, які передаються через API, для запобігання несанкціонованому доступу або витоку інформації.

3. Інтеграція з іншими компонентами: API служить мостом між різними частинами системи, такими як:
    - База даних (DB): API забезпечує доступ до бази даних для зберігання або отримання даних.
    - Інтелектуальний метод (IM): API може використовуватися для передачі даних до компонентів інтелектуального методу для аналізу та обробки.
4. Масштабування та ефективність: API повинен бути спроектований таким чином, щоб він міг ефективно обробляти велику кількість запитів та масштабуватися відповідно до зростаючих потреб системи.

Ця структура API є ключовою для забезпечення ефективної та безпечної взаємодії між різними компонентами системи визначення шахрайських інтернет-магазинів.

Далі представимо діаграму відношень сутностей (рис.2.2), яка ілюструє розширену структуру бази даних з додатковими компонентами системи .

На цій діаграмі (див.рис.2.2) представлені три основні сутності:

1. PRODUCT_INFORMATION: Ця сутність містить атрибути, які відображають різні аспекти інформації про товари, що можуть вказувати на потенційні шахрайські інтернет-магазини.
2. INTEGRATION_MODULE: Ця сутність представляє модуль інтеграції системи, який включає в себе компоненти "API зв'язок" та "База даних".
3. FRAUD_DETECTION_SYSTEM: Ця сутність представляє систему виявлення шахрайства, яка включає в себе компоненти "Аналіз даних", "Машинне навчання" та "Передбачення шахрайства".

PRODUCT_INFORMATION		
string	V1	Відсутність інформації про призначення товару
string	V2	Відсутність інформації про склад матеріалу, з якого зроблений товар
string	V3	Відсутність інформації про продукцію державною мовою (на українських сайтах)
string	V4	Приховування недоліків товарів (при продажі товарів, що були в користуванні)
string	V5	Відсутність інформації про стан зношеності товарів, якщо вони були у користуванні
string	V6	Відсутність інформації про виробника товару
string	V7	Відсутність інформації про місце найближчого сервісного обслуговування товару
string	V8	Відсутність інформації про гарантійний строк товару
string	V9	Відсутність адреси, за якою покупець може повернути товар протягом чотирнадцяти днів з моменту купівлі
string	V10	Відсутність інформації про виробника товару
string	V11	Відсутність інформації про продавця товару (що унеможливлює звернення до суду з позовом)

використовує

INTEGRATION_MODULE		
string	API	API зв'язок
string	DB	База даних

інтегрується з

FRAUD_DETECTION_SYSTEM		
string	Data_Analysis	Аналіз даних
string	Machine_Learning	Машинне навчання
string	Fraud_Prediction	Передбачення шахрайства

Рисунок 2.2 – Діаграма відношень сутностей

Зв'язки між цими сутностями показують, як модуль інтеграції використовує дані з сутності "PRODUCT_INFORMATION" та інтегрується з системою виявлення шахрайства. Це демонструє, як інформація про товари може бути аналізована та використана для виявлення потенційних шахрайських дій.

2.1.3 Структура клієнтського інтерфейсу

Діаграма, представлена на Рис.2.3, детально відображає процес взаємодії користувача з системою визначення шахрайських інтернет-магазинів. Ця діаграма є ключовою для розуміння того, як користувачі взаємодіють з різними функціями системи, починаючи від автентифікації та закінчуючи аналізом та обробкою даних. Вона включає три основні етапи: вхід в систему, аналіз інтернет-магазину та взаємодію з результатами. Кожен етап складається з декількох кроків, які дозволяють користувачу ефективно використовувати систему для ідентифікації потенційно шахрайських магазинів.

Діаграма (Рис.2.3) описує наступні етапи взаємодії користувача з системою:

1. Вхід в систему:
    - Ввід логіна та пароля (Користувач)
    - Перевірка автентичності (Система)
    - Доступ до головного інтерфейсу (Користувач)
2. Аналіз інтернет-магазину:
    - Введення URL магазину (Користувач)
    - Запуск аналізу даних (Система)
    - Відображення результатів аналізу (Система)
3. Взаємодія з результатами:
    - Перегляд детальної інформації (Користувач)
    - Застосування фільтрів (Користувач)
    - Збереження або експорт результатів (Користувач)

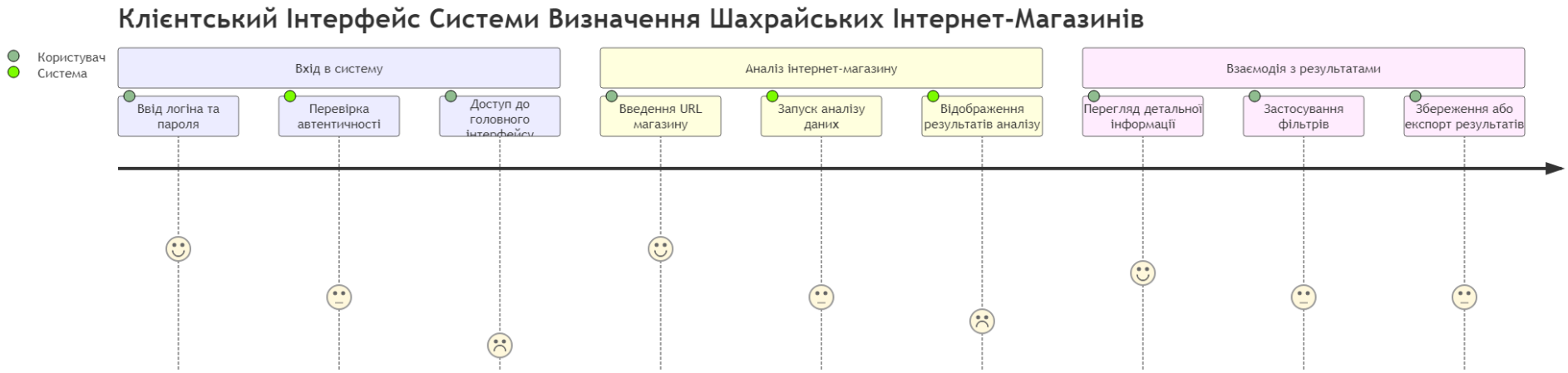

Рисунок 2.3 - Структура клієнтського інтерфейсу

Ця діаграма (рис.2.3) демонструє, як користувач взаємодіє з системою від моменту входу до аналізу інтернет-магазинів та подальшої роботи з отриманими результатами.

Отже, проектування системи для визначення шахрайських інтернет-магазинів, представлене у цьому параграфі 2.1, є комплексним підходом, який інтегрує різні технології та методики для ефективного виявлення потенційних шахрайств. Основою системи є інтелектуальний метод, що включає аналіз даних, машинне навчання та передбачення шахрайства, що дозволяє точно ідентифікувати шахрайські патерни. Система визначення обробляє вхідні дані, перевіряє їх та виводить результати, що допомагає користувачам у виявленні шахрайських магазинів. Модуль інтеграції, включаючи API зв'язок та базу даних, забезпечує ефективну взаємодію між різними компонентами системи.

Діаграма відношень сутностей демонструє взаємозв'язок між ключовими компонентами системи, включаючи інформацію про товари, модуль інтеграції та систему виявлення шахрайства. Це дозволяє зрозуміти, як дані про товари аналізуються та використовуються для виявлення шахрайських дій.

Структура клієнтського інтерфейсу, представлена на діаграмі, показує, як користувачі можуть взаємодіяти з системою, від автентифікації до аналізу інтернет-магазинів та роботи з результатами. Це забезпечує зручний та інтуїтивно зрозумілий спосіб використання системи для ідентифікації шахрайських магазинів.

Загалом, цей параграф 2.1. представляє детальне проектування системи, яка може значно підвищити ефективність виявлення шахрайських інтернет-магазинів, забезпечуючи користувачам надійний інструмент для захисту від онлайн шахрайства.

2.2 Опис методів класифікації на основі машинного навчання

Класифікація - це процес віднесення об'єктів або явищ до певних категорій на основі спільних ознак. Існує кілька підходів до класифікації, які використовуються в різних областях науки та техніки. Далі розглянемо детальніше основні підходи методів класифікації на основі машинного навчання.

2.2.1 Логістична регресія

Логістична регресія (Logistic Regression) - це статистичний метод, який використовується для аналізу даних, де залежна змінна (ціль) є категоріальною. Це тип регресійного аналізу, який часто використовується для прогнозування імовірності певної події.

В логістичній регресії, на відміну від лінійної регресії, де вихід може бути будь-яким числом, вихід обмежується відрізком [0,1] і інтерпретується як імовірність. Це досягається за допомогою логістичної (або сигмоїдної) функції.

Сигмоїдна функція, яка є основою логістичної регресії, визначається як:

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

де *e* - основа натурального логарифма, а *z* - лінійна комбінація вхідних змінних, $z = b + w_1x_1 + w_2x_2 + ... + w_nx_n$, де *b* є перехідним значенням (bias), а *wi* - ваги асоційовані з кожною вхідною змінною *xi*.

Імовірність того, що даний вхідний набір даних *X* належить до класу 1 (позитивного класу), може бути записана як:

$$P(Y = 1|X) = \sigma(b + w_1x_1 + w_2x_2 + ... + w_nx_n)$$

де *Y* - цільова змінна.

Для підбору параметрів *b* та *wi* використовується метод максимальної вірогідності. Цей метод намагається максимізувати функцію вірогідності, таким чином оптимізуючи параметри моделі, щоб максимально точно прогнозувати реальні дані.

В логістичній регресії часто використовується функція втрат від логістичної вірогідності (log-loss), яка визначається як:

$$L(Y, P) = -\sum(Y \log(P) + (1 - Y) \log(1 - P))$$

де *Y* - істинні значення, а *P* - передбачені імовірності.

2.2.2 Випадковий ліс

Випадковий ліс (Random Forest) - це ансамбльовий метод машинного навчання, який використовує множину дерев рішень для досягнення кращої продуктивності та точності прогнозування. Ось детальний опис із математичними деталями:

1. Дерева рішень: Випадковий ліс складається з багатьох дерев рішень. Кожне дерево рішень - це модель прогнозування, яка розділяє дані на різні сегменти на основі певних ознак.

2. Бутстрепова вибірка (Bootstrap sampling): Для кожного дерева рішень випадковий ліс використовує різну вибірку даних, отриману методом бутстрепінгу (вибірка з поверненням із вихідного набору даних).
3. Випадковість при виборі ознак: При кожному розділенні в дереві рішень випадковий ліс вибирає випадковий піднабір ознак, серед яких шукається найкращий критерій розділення.

Нехай у нас є набір тренувальних даних $D = \{(x_1, y_1), ..., (x_N, y_N)\}$, де *xi* - це вектори ознак, а *yi* - цільові мітки.

1. Створення Дерев Рішень:

   Для k=1,...,K (де K - кількість дерев у лісі):

   a. Виберіть випадкову бутстрепову вибірку Dk із D.

   b. Побудуйте дерево рішень Tk на основі Dk:

   - При кожному розділенні використовуйте випадковий піднабір ознак.
   - Розділяйте вузли до тих пір, поки не досягнете максимальної глибини або не будуть виконані інші критерії зупинки.

2. Прогнозування:

Для класифікації: Прогноз для нового вектора ознак x визначається як найчастіше передбачуваний клас серед усіх дерев:

$$\hat{y} = \text{mode}\{T_1(x), T_2(x), ..., T_K(x)\}$$

Для регресії: Прогноз для x - це середнє значення передбачень усіх дерев:

$$\hat{y} = \frac{1}{K} \sum_{k=1}^{K} T_k(x)$$

Використання бутстрепових вибірок та випадковості у виборі ознак зменшує ризик перенавчання моделі на тренувальних даних.

Випадковий ліс може оцінювати важливість ознак на основі того, наскільки ефективно ознаки покращують продуктивність прогнозування.

Випадковий ліс може бути використаний для як класифікації, так і регресії.

2.2.3 K-найближчих сусідів

K-найближчих сусідів (KNN) - це простий, але потужний алгоритм машинного навчання, який використовується для класифікації та регресії.

KNN працює на основі простої ідеї: об'єкти, які є подібними, часто знаходяться поруч один з одним. Таким чином, класифікація або прогнозування для невідомого об'єкту може бути виконане шляхом визначення "найближчих" (тобто найбільш схожих) відомих об'єктів.

Кроки Алгоритму KNN:

1. Вибір числа *K*: Визначте *K*, кількість найближчих сусідів.
2. Обчислення відстаней: Для кожного невідомого об'єкта обчисліть відстань до кожного з тренувальних об'єктів. Найчастіше використовуються:

Евклідова відстань: $d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$

Манхеттенська відстань: $d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$

Мінковського відстань: $d(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p}$

3. Знаходження K найближчих сусідів: Виберіть K об'єктів з тренувального набору, які мають найменшу відстань до невідомого об'єкта.
4. Голосування для класифікації / усереднення для регресії: Для класифікації, використовуйте "голосування" серед сусідів для визначення найпопулярнішої мітки класу. Для регресії, обчисліть середнє значення цільових змінних сусідів.

2.2.4 Наївний Байєсівський класифікатор

Наївний Байєсівський класифікатор (Naive Bayes) - це простий, але ефективний алгоритм машинного навчання, заснований на теоремі Байєса з наївним припущенням про незалежність між ознаками. Наївний Байєсівський класифікатор спирається на теорему Байєса, яка є фундаментальною в теорії ймовірності. Теорема Байєса визначає ймовірність події, засновану на апріорних знаннях, що можуть вплинути на цю ймовірність. Математично теорема виглядає так:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

де:

$P(A|B)$ - апостеріорна ймовірність події A за умови B.

$P(B|A)$ - ймовірність спостерігати B, якщо відбулася подія A.

$P(A)$ - апріорна ймовірність події A.

$P(B)$ - апріорна ймовірність події B.

В контексті класифікації, Наївний Байєсівський класифікатор передбачає ймовірність того, що точка даних належить до певного класу, засновуючись на ознаках цієї точки. Основне припущення тут - це наївне припущення про незалежність ознак.

Нехай $X=(x1,x2,...,xn)$ - вектор ознак, і потрібно оцінити ймовірність того, що X належить до класу Ck. Згідно з наївним Байєсівським підходом:

$$P(C_k|X) = \frac{P(X|C_k) \times P(C_k)}{P(X)}$$

де:

$P(Ck|X)$ - ймовірність того, що X належить до класу Ck.

$P(X|Ck)$ - ймовірність спостерігати вектор ознак X в класі Ck, яка розкладається на добуток ймовірностей за наївним припущенням:

$$P(X|C_k) = \prod_{i=1}^{n} P(x_i|C_k)$$

$P(Ck)$ - апріорна ймовірність класу Ck.

$P(X)$ - ймовірність спостерігати вектор ознак X, яка є константою для всіх класів і часто ігнорується при прогнозуванні, оскільки шукаємо клас з найвищою ймовірністю.

2.2.5 Класифікатор на основі опорних векторів

Класифікатор на основі опорних векторів (Support Vector Classifier, часто відомий як SVM - Support Vector Machine) - це потужний і гнучкий алгоритм машинного навчання, використовуваний

для класифікації. Він працює шляхом визначення гіперплощини або набору гіперплощин у високовимірному просторі, який може бути використаний для класифікації або регресії.

Основна мета SVM - знайти гіперплощину, яка найкраще розділяє два класи точок даних. У двовимірному просторі ця гіперплощина є просто лінією.

Гіперплощина визначається рівнянням:

$$w \cdot x - b = 0$$

де *w* - вектор ваг, *x* - вектор вхідних ознак, і *b* - зсув (bias).

Важливим поняттям у SVM є маржа, яка є відстанню між гіперплощиною і найближчими точками даних з кожного класу, які називаються опорними векторами. SVM намагається максимізувати цю маржу, щоб збільшити загальну точність класифікатора.

Мета полягає у мінімізації наступного виразу:

$$\min_{w,b} \frac{1}{2}\|w\|^2$$

за умови, що:

$$y_i(w \cdot x_i - b) \geq 1, \quad \forall i$$

де *yi* - це мітки класів (зазвичай +1 або -1), *xi* - вектори вхідних даних.

У випадках, коли дані не є лінійно роздільними, SVM використовує так звані ядрові функції для перетворення вхідних даних у простір вищої вимірності, де вони можуть бути розділені лінійно. Найпоширеніші ядрові функції включають:

Лінійне ядро: $K(x, x') = x \cdot x'$
Поліноміальне ядро: $K(x, x') = (1 + x \cdot x')^d$
Радіально-базисна функція (RBF) або Гауссове ядро: $K(x, x') = e^{-\gamma \|x - x'\|^2}$
Сигмоїдне ядро: $K(x, x') = \tanh(\alpha x \cdot x' + r)$

2.2.6 Дерево рішень

Дерево рішень (DecisionTree Classifier) - це графічне представлення всіх можливих рішень та їх потенційних наслідків. Воно складається з вузлів, гілок і листя. Кожен внутрішній вузол представляє ознаку або атрибут, кожна гілка представляє правило рішення, а кожен лист - кінцевий результат або клас.

Алгоритм побудови дерева:

1. Вибір ознаки: При побудові дерева рішень, алгоритм вибирає ознаки, що найкраще розділяють дані на класи. Вибір ознаки здійснюється на основі критеріїв чистоти, таких як індекс Джині, ентропія або помилка класифікації.
2. Критерії розділення:

   Ентропія: Міра розподілу категорій у наборі даних. Визначається як:

   $$H(T) = -\sum_i p_i \log_2(p_i)$$

   де *pi* - це ймовірність класу *i* у наборі даних *T*.

   Індекс Джині: Міра ступеня невизначеності або чистоти. Визначається як:

   $$G(T) = 1 - \sum_i p_i^2$$

Помилка класифікації: Це просто частка неправильно класифікованих елементів у вузлі.

3. Рекурсивне розділення: Алгоритм продовжує розділяти вузли на підвузли, використовуючи кращі ознаки, до тих пір, поки не буде досягнуто певних критеріїв зупинки, таких як максимальна глибина дерева або мінімальна кількість елементів у вузлі.

Отже, у параграфі 2.2 розглядаються різні методи класифікації на основі машинного навчання, що є ключовими для розробки ефективних систем виявлення шахрайських інтернет-магазинів. Логістична регресія використовується для прогнозування імовірності подій, використовуючи логістичну функцію для обмеження виходу від 0 до 1. Випадковий ліс, як ансамбльовий метод, використовує множину дерев рішень для підвищення точності прогнозування, використовуючи бутстрепову вибірку та випадковість у виборі ознак. K-найближчих сусідів (KNN) використовує просту ідею про те, що подібні об'єкти часто знаходяться поруч, для класифікації або прогнозування. Наївний Байєсівський класифікатор базується на теоремі Байєса та наївному припущенні про незалежність між ознаками. Класифікатор на основі опорних векторів (SVM) використовує гіперплощини для розділення класів, максимізуючи маржу між ними. Дерево рішень використовує графічне представлення можливих рішень та їх наслідків, вибираючи ознаки на основі критеріїв чистоти. Кожен з цих методів має свої унікальні характеристики та застосування, і їх вибір залежить від конкретних вимог та характеристик даних у задачах класифікації.

2.3 Поліпшення ефективності моделей машинного навчання при роботі з незбалансованими даними

В даному розділі розглядається проблема незбалансованих даних, яка виникає, коли кількість прикладів одного класу значно відстає від кількості прикладів іншого класу. Незбалансованість даних може виникати в різних задачах, таких як виявлення рідкісних подій, наприклад, виявлення шахрайства, де більшість прикладів належить до класу "нормально", а лише дуже маленька частина - до класу "шахрайство". Для покращення ефективності моделей машинного навчання при роботі з незбалансованими даними розглядаються різні методи, такі як підгенерація (undersampling), надгенерація (oversampling), SMOTE (Synthetic Minority Over-sampling Technique) та ADASYN (Adaptive Synthetic Sampling). Кожен з цих методів має свої особливості і використовується для досягнення більшого балансу в розподілі класів у навчальних даних.

Imbalanced (Незбалансований) [20]: Це вказує на ситуацію, коли кількість прикладів одного класу в наборі даних суттєво вища або нижча, ніж кількість прикладів іншого класу. Наприклад, у задачах виявлення рідкісних подій, таких як шахрайство, може виникати незбалансованість, коли більшість прикладів належать до класу "нормально", а лише дуже невелика частина - до класу "шахрайство".

Припустимо, що маємо набір даних для бінарної класифікації, де маємо два класи: "Клас 0" і "Клас 1". Позначимо кількість прикладів класу 0 як N0 і кількість прикладів класу 1 як N1.

Незбалансований набір даних може мати різний ступінь нерівності між кількістю прикладів обох класів, де N0 може бути набагато більше або менше за N1. Це можна представити математично наступним чином:

1. Якщо N0 >> N1 (кількість прикладів класу 0 суттєво більше за кількість прикладів класу 1):

   N0 >> N1

   Де N0 - кількість прикладів класу 0, і N1 - кількість прикладів класу 1. У цьому випадку набір даних є незбалансованим через перевагу класу 0.

2. Якщо N0 << N1 (кількість прикладів класу 0 суттєво менше за кількість прикладів класу 1):

   N0 << N1

   Де N0 - кількість прикладів класу 0, і N1 - кількість прикладів класу 1. У цьому випадку набір даних також є незбалансованим через перевагу класу 1.

Найчастіше виникає ситуація, коли маємо незбалансований набір даних, де один клас має велику кількість прикладів (N0 більше), а інший клас має меншу кількість прикладів (N1 менше). Рішення щодо роботи з незбалансованими даними включає в себе використання різних методів для покращення результатів моделі, таких як методи підгенерації або надгенерації даних, що були згадані раніше.

Undersampling (Підгенерація) [21]: Цей метод включає в себе зменшення кількості прикладів більшого класу (зазвичай, класу з більшою кількістю екземплярів) так, щоб вирівняти розподіл класів. Наприклад, видалення деяких випадків із класу "нормально" для підгенерації. Метод підгенерації (undersampling) полягає в тому, щоб

зменшити кількість прикладів більшого класу (зазвичай, класу з більшою кількістю екземплярів) так, щоб вирівняти розподіл класів. Математично цей метод може бути описаний наступним чином:

- N0 - кількість прикладів класу 0 (менший клас).
- N1 - кількість прикладів класу 1 (більший клас).
- p - обрана частка (відсоток) прикладів, які будуть видалені з більшого класу.

Тоді, кількість прикладів, які залишаться в більшому класі після застосування методу підгенерації, буде дорівнювати:

N1_new = (1 - p) * N1

Це означає, що потрібно видаляти певну частку (p) прикладів з більшого класу, зменшуючи його кількість до нового значення N1_new, таким чином вирівнюючи розподіл класів.

Oversampling (Надгенерація) [22]: Цей метод передбачає збільшення кількості прикладів меншого класу (зазвичай, менш представленого класу) для досягнення більшого балансу. Це може бути досягнуто шляхом додавання дубльованих або модифікованих прикладів до меншого класу.

Метод надгенерації (oversampling) передбачає збільшення кількості прикладів меншого класу (зазвичай, менш представленого класу) для досягнення більшого балансу. Математично цей метод може бути описаний наступним чином:

- N0 - кількість прикладів класу 0 (менший клас).
- N1 - кількість прикладів класу 1 (більший клас).
- f - обрана частка (відсоток), на яку потрібно збільшити кількість прикладів меншого класу.

Тоді, кількість нових прикладів, які додано до меншого класу, буде дорівнювати:

N0_new = (1 + f) * N0

Це означає, що збільшується кількість прикладів в меншому класі на певну частку (f), щоб збалансувати розподіл класів.

SMOTE (Synthetic Minority Over-sampling Technique) [23]: Це конкретний метод надгенерації, який створює штучні екземпляри меншого класу на основі існуючих прикладів цього класу та їхніх найближчих сусідів у просторі ознак.

Метод надгенерації SMOTE (Synthetic Minority Over-sampling Technique) [23] створює штучні екземпляри меншого класу на основі існуючих прикладів цього класу та їхніх найближчих сусідів у просторі ознак. Математично цей метод може бути описаний наступним чином:

- X_minority - набір прикладів меншого класу, де кожен приклад описується вектором ознак X_minority[i], де i = 1, 2, ..., N_minority, де N_minority - кількість прикладів меншого класу.
- N_synthetic - кількість штучних (синтетичних) прикладів, які потрібно створити.

Для кожного прикладу X_minority[i], SMOTE вибирає k найближчих сусідів цього прикладу (де k - це параметр SMOTE, зазвичай задається користувачем).

Потім для кожного прикладу X_minority[i], SMOTE генерує N_synthetic / N_minority штучних прикладів наступним чином:

1. Випадково вибирається один з k найближчих сусідів, скажімо X_neighbor.
2. Випадковим чином обирається число від 0 до 1, скажімо lambda.

3. Штучний приклад генерується як:

X_synthetic = X_minority[i] + lambda * (X_neighbor - X_minority[i])

Де "+" в цьому контексті означає поелементне додавання векторів ознак.

Цей процес повторюється для кожного прикладу X_minority[i], і на виході отримуємо N_synthetic нових штучних прикладів, які можна додати до набору даних меншого класу для досягнення більшого балансу. SMOTE допомагає збільшити кількість прикладів меншого класу, при цьому враховуючи його структуру в просторі ознак.

ADASYN (Adaptive Synthetic Sampling) [24]: Цей метод є покращенням SMOTE та розглядає нерівномірність в розподілі класів для кожного прикладу, змінюючи вагу створених штучних екземплярів в залежності від їхнього рівня переваги або недостатності.

Метод надгенерації ADASYN (Adaptive Synthetic Sampling) [24] є покращенням SMOTE та враховує нерівномірність в розподілі класів для кожного прикладу, змінюючи вагу створених штучних екземплярів в залежності від їхнього рівня переваги або недостатності. Математично ADASYN може бути описаний так:

- X_minority - набір прикладів меншого класу, де кожен приклад описується вектором ознак X_minority[i], де i = 1, 2, ..., N_minority, де N_minority - кількість прикладів меншого класу.
- X_majority - набір прикладів більшого класу, де кожен приклад описується вектором ознак X_majority[j], де j = 1, 2, ..., N_majority, де N_majority - кількість прикладів більшого класу.
- N_synthetic - кількість штучних (синтетичних) прикладів, які потрібно створити для меншого класу.

Крім того, для кожного прикладу X_minority[i], визначаються наступні параметри:

- G_i - кількість сусідів прикладу X_minority[i] з класу X_majority.
- N_i - кількість сусідів прикладу X_minority[i] з класу X_minority (включаючи самого себе).

Кроки ADASYN:

1. Визначаємо вагу кожного прикладу X_minority[i] як: Weight_i = G_i / (N_i + ε)

   Де ε - додатковий параметр для уникнення ділення на нуль.
2. Визначаємо кількість штучних прикладів, які необхідно створити для кожного X_minority[i] як: N_synthetic_i = round(Weight_i * N_synthetic)
3. Для кожного прикладу X_minority[i] генеруємо N_synthetic_i штучних прикладів наступним чином, подібно до SMOTE, з використанням найближчих сусідів з класу X_majority.

Цей процес повторюється для кожного прикладу меншого класу X_minority[i], і на виході отримуємо N_synthetic нових штучних прикладів з урахуванням нерівномірності в розподілі класів та ваги кожного прикладу. ADASYN дозволяє більш ефективно балансувати набір даних за рахунок адаптації до конкретного розподілу класів.

У цьому розділі було розглянуто проблему незбалансованих даних, яка може виникати в різних задачах машинного навчання, особливо в контексті виявлення рідкісних подій, наприклад, шахрайства в електронній комерції. Виявлення і робота з незбалансованими даними є важливою складовою у забезпеченні ефективної та точної роботи моделей машинного навчання. Для

вирішення цієї проблеми було представлено різні методи, такі як підгенерація, надгенерація, SMOTE та ADASYN, кожен з яких має свої переваги та може бути застосований в залежності від конкретного завдання та характеристик даних.

Використання цих методів допомагає досягнути більшого балансу в розподілі класів у навчальних даних, що покращує якість та надійність моделей машинного навчання при роботі з незбалансованими даними. Важливо враховувати, що вибір конкретного методу може залежати від специфіки задачі та особливостей даних. Таким чином, використання відповідного підходу до обробки незбалансованих даних може підвищити результативність та ефективність моделей машинного навчання у виявленні рідкісних подій та інших задачах з незбалансованими даними.

2.4 Інтелектуальний метод визначення шахрайських інтернет-магазинів

У сучасному світі інтернет-торгівлі, зростання числа шахрайських інтернет-магазинів становить серйозну загрозу для користувачів. Важливість визначення та попередження шахрайства набуває ключового значення в цій сфері. Для цього розроблено спеціалізований інтелектуальний метод, який дозволяє точно ідентифікувати потенційно шахрайські сайти. Цей метод базується на декількох кроках, які включають вибір підозрілого сайту, збір і аналіз даних з нього, а також використання найсучасніших алгоритмів машинного навчання для класифікації сайтів. Кожен крок цього процесу є важливим у визначенні

надійності інтернет-магазину. Детальний опис кожного кроку, включно з використаними методами класифікації та аналізом результатів, дозволяє ефективно розпізнати та уникнути потенційних шахрайських пасток у цифровому світі. Далі представлено інтелектуальний метод визначення шахрайських інтернет-магазинів кроками та структурно (рис. 2.2):

Крок 1. Вибір сайту, що може бути шахрайським.

Крок 2. Збір даних з сайту. Всі параметри (таблиця 1) парсимо з зазначеного сайту та при наявності інформації система прописує 0, при відсутності ставить 1.

Крок3. Для параметру V4 (за наявності 1), проводимо аналіз відгуків користувачів, інтелектуальним методом вибору конкурентного товару на основі емоційного забарвлення відзивів [14]. За наявності позитивного відзиву проставляємо 0, негативного 1.

Крок 4. Проводимо класифікацію чотирьома найпопулярнішими методами класифікації.

Крок 4.1. Logistic Regression [17].

Крок 4.2. Random Forest [15].

Крок 4.3. KNN [16].

Крок 4.4. Naive Bayes [19].

Крок 4.5. Support Vector Classifier [16].

Крок 4.6. DecisionTree Classifier [18].

Крок 5. Вивід отриманих результатів.

Крок 6. Визначення чи сайт шахрайський чи ні, по більшості класифікованих результатів.

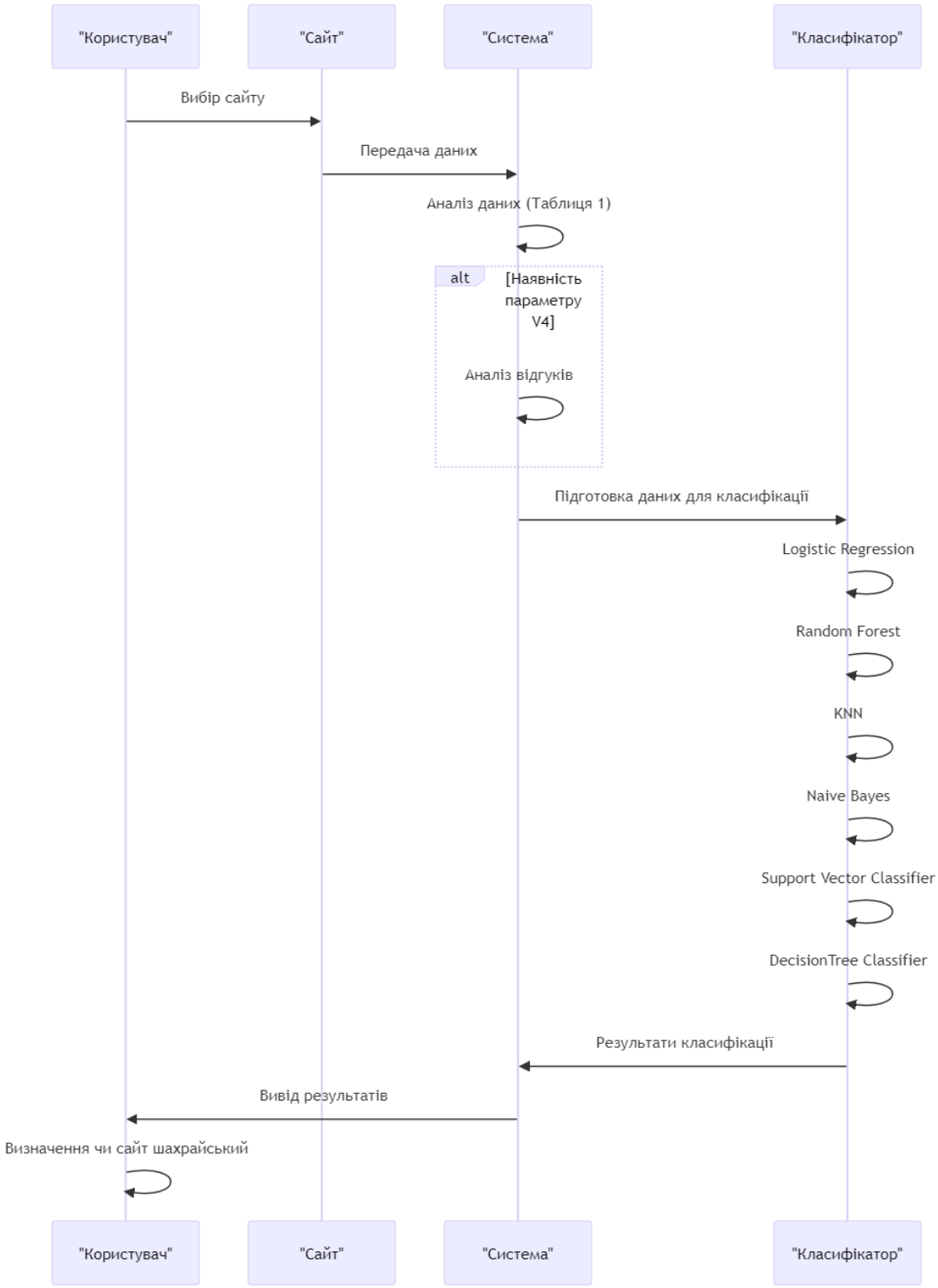

Рисунок 2.2 – Структура інтелектуального методу визначення шахрайських інтернет-магазинів

Цей метод, розроблений для захисту користувачів від недобросовісних онлайн продавців, включає в себе кілька ключових етапів, кожен з яких відіграє важливу роль у виявленні потенційних шахрайських сайтів.

На першому етапі відбувається вибір підозрілого сайту, що потребує подальшого аналізу. Другий крок полягає у зборі та аналізі даних з цього сайту, використовуючи різні параметри для оцінки його надійності. Третій крок включає аналіз відгуків користувачів, що дозволяє отримати додаткову інформацію про репутацію сайту.

Наступні кроки (4.1 - 4.6) передбачають застосування різних методів класифікації, таких як логістична регресія, випадковий ліс, KNN, наївний Байєсівський класифікатор, класифікатор на основі опорних векторів та дерево рішень. Кожен з цих методів вносить свій вклад у загальний процес аналізу, дозволяючи системі точно визначати потенційно шахрайські сайти на основі комплексного аналізу даних.

Останній крок полягає у виведенні отриманих результатів та визначенні, чи є сайт шахрайським, на основі більшості класифікованих результатів. Цей підхід дозволяє не тільки ефективно ідентифікувати шахрайські сайти, але й забезпечує високий рівень точності та надійності виявлення.

Загалом, розроблений інтелектуальний метод є важливим інструментом у боротьбі з шахрайством в інтернет-торгівлі, забезпечуючи користувачам додатковий рівень захисту та допомагаючи уникнути потенційних фінансових втрат та інших негативних наслідків.

Висновки до розділу 2:

1. Розробка системи для ідентифікації шахрайських інтернет-магазинів, описана у параграфі 2.1, представляє собою всеосяжний підхід, що об'єднує різноманітні технології та методології для ефективного виявлення шахрайства. Центральним елементом системи є інтелектуальний метод, який інтегрує аналіз даних, машинне навчання та прогнозування для точного виявлення шахрайських дій. Система обробляє вхідні дані, перевіряє їх та надає результати, допомагаючи користувачам виявляти ненадійні магазини. Модуль інтеграції, що включає API та базу даних, забезпечує злагоджену взаємодію між компонентами системи.
2. У параграфі 2.2 детально розглядаються методи класифікації на основі машинного навчання, які є фундаментальними для створення ефективних систем виявлення шахрайських інтернет-магазинів. Різноманітність методів, включаючи логістичну регресію, випадковий ліс, KNN, наївний Байєсівський класифікатор, SVM та дерево рішень, демонструє глибину та гнучкість підходів до класифікації. Кожен метод вносить свій унікальний вклад у процес аналізу, дозволяючи системі точно ідентифікувати потенційно шахрайські сайти.
3. У параграфі 2.3 розглянута проблема незбалансованих даних, що може виникати у різних завданнях машинного навчання, зокрема в області виявлення рідкісних подій, наприклад, шахрайства в електронній комерції. Маніпулювання та вирішення цієї проблеми є суттєвими для досягнення ефективності та точності моделей машинного навчання. У цьому контексті було представлено різні методи, такі як підгенерація, надгенерація, SMOTE та ADASYN, кожен з яких має свої переваги та може бути застосований в

залежності від конкретної задачі та властивостей даних. Використання цих методів сприяє досягненню більшої рівноваги в розподілі класів у навчальних даних, що підвищує якість та надійність моделей машинного навчання в умовах незбалансованих даних. Важливо враховувати, що вибір конкретного методу повинен базуватися на конкретних вимогах завдання та особливостях даних, і відповідний підхід до обробки незбалансованих даних може значно поліпшити результати та продуктивність моделей машинного навчання в роботі з рідкісними подіями та іншими завданнями з незбалансованими даними.

4. Розроблений інтелектуальний метод, описаний у розділі 2.4, є ключовим інструментом у боротьбі з онлайн шахрайством. Він включає в себе кроки від вибору підозрілого сайту до збору та аналізу даних, а також використання передових методів класифікації для оцінки надійності сайтів. Кінцевий етап системи полягає у виведенні результатів та визначенні статусу сайту, забезпечуючи користувачам високу точність та надійність у виявленні шахрайських магазинів. Цей комплексний підхід не тільки ефективно виявляє шахрайські сайти, але й надає користувачам додатковий захист, допомагаючи уникнути потенційних фінансових втрат та інших негативних наслідків.

3 РЕАЛІЗАЦІЯ ВИЗНАЧЕННЯ ШАХРАЙСЬКИХ ІНТЕРНЕТ-МАГАЗИНІВ

3.1 Опис реалізації запропонованого методу

Для втілення намічених кроків потрібно написати скрипт на Python, який має кілька етапів: отримання даних із веб-сайту, аналіз відгуків, використання моделей машинного навчання для класифікації та оцінки результатів. Проте важливо зазначити, що для виконання цих кроків потрібен доступ до Інтернету, специфічних API для аналізу відгуків і додаткових даних та ресурсів, яких немає у моєму поточному середовищі.

Щодо вибору сайту, цей етап передбачає вручний вибір сайту, який ви підозрюєте у шахрайстві.

Для збору даних з сайту можна скористатися бібліотеками, такими як requests та BeautifulSoup. Нижче наведено приклад коду, який можна використати для отримання HTML сторінки:

```
import requests
from bs4 import BeautifulSoup
url = "https://example.com"
response = requests.get(url)
soup = BeautifulSoup(response.content, 'html.parser')
```

Перед тим, як переходити до використання моделей машинного навчання, необхідно провести підготовку даних для аналізу відгуків. Цей процес включає застосування природно-мовних обробок для

розуміння емоційного забарвлення відгуків. Для цього можна використовувати бібліотеки, такі як TextBlob або nltk.

```
X = data.drop('Target', axis=1)  # Видаляємо цільову змінну з набору даних
y = data['Target']  # Це наша цільова змінна
```

У цьому відрізку коду працюємо з двома основними наборами даних: X та y. X містить особливості чи незалежні змінні, які модель буде використовувати для навчання, тоді як y це цільова змінна, яку намагаємося передбачити.

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)
```

Для оцінки ефективності моделі розділяємо наші дані на навчальний і тестовий набори. X_train та y_train використовуються для навчання моделі, тоді як X_test і y_test використовуються для перевірки, наскільки добре модель справляється з прогнозуванням не бачених раніше даних.

Параметр test_size=0.2 означає, що 20% вихідних даних використовується для тестування, тоді як решта (80%) використовується для навчання моделі.

У даному відрізку коду проводиться ініціалізація різних моделей машинного навчання. Кожна з цих моделей має свої унікальні характеристики та відповідності, що роблять їх придатними для різних типів даних і завдань машинного навчання.

```
models = {
    "Logistic Regression": LogisticRegression(),
    "Random Forest": RandomForestClassifier(),
    "KNN": KNeighborsClassifier(),
    "Naive Bayes": GaussianNB(),
    "SVC": SVC(),
    "Decision Tree": DecisionTreeClassifier()
}
```

У цьому циклі коду проводимо навчання кожної моделі, використовуючи метод model.fit(X_train, y_train). Цей процес включає підгонку моделі до навчальних даних, де вона "вчиться" на основі наданих особливостей (X_train) та відповідних цільових значень (y_train).

Після цього використовуємо навчені моделі для передбачення значень на тестових даних за допомогою методу model.predict(X_test). Це дозволяє нам оцінити, наскільки добре модель узгоджується з реальними даними, які вона не бачила під час навчання.

```
for name, model in models.items():
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    print(f"{name}:\n{classification_report(y_test, y_pred)}\n")
```

Виведення звіту про класифікацію для кожної моделі, що включає точність, повноту та інші метрики, є важливим етапом для визначення того, яка модель працює найкраще з наданими даними. Кожна модель має свої переваги та недоліки і може показувати кращі або гірші результати в залежності від конкретного набору даних та завдання.

Підтвердження того, чи сайт є шахрайським, включає аналіз результатів класифікації та прийняття рішення на підставі більшості класифікаційних результатів. Це означає, що, спираючись на визначені моделями класи, вирішується, чи можна вважати сайт шахрайським.

Важливо зауважити, що реалізація цих кроків потребує більш детальних та індивідуальних налаштувань, особливо у випадку парсингу веб-сайтів та аналізу відгуків, оскільки ці процеси залежать від конкретної структури сайту та доступних даних.

Отже, у цьому параграфі описано розробку скрипта на Python, який є ключовим елементом у виявленні шахрайських інтернет-магазинів. Процес включає кілька етапів: від отримання даних з веб-сайту до аналізу відгуків та використання моделей машинного навчання для класифікації та оцінки результатів. Однак, важливо врахувати, що для ефективного виконання цих кроків потрібен доступ до Інтернету, специфічних API та додаткових ресурсів, які можуть бути недоступні в поточному середовищі.

Вибір підозрілого сайту для аналізу здійснюється вручну, після чого використовуються бібліотеки, такі як requests та BeautifulSoup, для збору даних. Підготовка даних перед аналізом відгуків включає застосування природно-мовних обробок, що дозволяє визначити емоційне забарвлення відгуків. Використання різних моделей машинного навчання, таких як логістична регресія, випадковий ліс, KNN, наївний Байєсівський класифікатор та інші, дозволяє провести глибокий аналіз та класифікацію даних.

Останнім етапом є оцінка ефективності моделі та виведення звіту про класифікацію, що включає ключові метрики, такі як точність та

повнота. Це дозволяє визначити, яка модель найкраще підходить для конкретного набору даних. Важливо підкреслити, що остаточне рішення про шахрайський характер сайту базується на аналізі результатів класифікації та враховує більшість класифікаційних результатів. Такий підхід дозволяє точно ідентифікувати потенційно шахрайські сайти, забезпечуючи користувачам надійний інструмент для захисту від онлайн шахрайства. Однак, слід зазначити, що успішність цього процесу залежить від детальних налаштувань та специфіки веб-сайтів, з яких збираються дані.

3.2 Розвідувальний аналіз даних

Для проведення експериментальних досліджень був створений набір даних на основі веб-сайтів, які функціонують в Україні. У цьому наборі даних міститься інформація про 67 сайтів, приблизно 45% з яких вважаються шахрайськими.

Датасет складається з декількох колонок, більшість з яких, схоже, містять числові дані. Основні з них:

• Fit: Ймовірно, це бінарна змінна зі значеннями 0 або 1, що, можливо, показує відповідь на певний критерій чи умову

. • V1 - V11: Ці колонки містять числові дані, ймовірно, якісь параметри або характеристики, що характеризують кожен сайт.

• Sait: Ця колонка містить текст, що, ймовірно, представляє URL-адреси веб-сайтів.

Ці візуалізації надають інформацію про розподіл та зв'язки між даними (Рис.3.1):

• Розподіл змінної Fit: Змінна Fit демонструє приблизно рівномірний розподіл між значеннями 0 та 1. Це означає, що в даних майже однакова кількість спостережень для обох категорій.

• Розподіл інших числових змінних (V1 - V11): Аналогічно змінній Fit, ці змінні також мають приблизно рівномірний розподіл між 0 та 1.

• Кореляційна матриця: Більшість змінних мають низький або помірний рівень кореляції між собою. Найвища взаємозв'язок спостерігається між парами змінних, але він все ще не є дуже високим (менше 0.5).

Слабкі взаємозв'язки між бінарними змінними у вашому наборі даних підтверджують, що кожна змінна несе унікальну інформацію і має слабку взаємодію з іншими змінними.

У відношенні до моделювання чи аналізу, це свідчить про те, що видалення будь-якої з цих змінних може призвести до втрати важливої інформації. Ці змінні не є взаємозамінними або дублюючими одна одну, тож кожна з них має свою унікальну вагомість у сприйнятті та аналізі даних.

У розділі 3.2 представлено розвідувальний аналіз даних, здійснений на основі набору даних, що включає інформацію про 67 українських веб-сайтів, з яких близько 45% вважаються шахрайськими. Аналіз охоплює декілька ключових аспектів, включаючи розподіл бінарної змінної Fit та інших числових змінних (V1 - V11), а також кореляційні зв'язки між цими змінними. Результати показують приблизно рівномірний розподіл значень 0 та 1 для змінної Fit, а також для інших числових змінних, що свідчить про збалансованість даних.

Кореляційна матриця, представлена на Рисунку 3.1, виявляє слабкі або помірні взаємозв'язки між змінними, з найвищими кореляціями менше 0.5. Це вказує на те, що кожна змінна вносить унікальний вклад у набір даних і має обмежену взаємодію з іншими змінними. Така характеристика даних є важливою для моделювання та аналізу, оскільки видалення будь-якої з цих змінних може призвести до втрати значущої інформації. Змінні не дублюють одна одну і кожна з них має свою значущість у контексті аналізу даних.

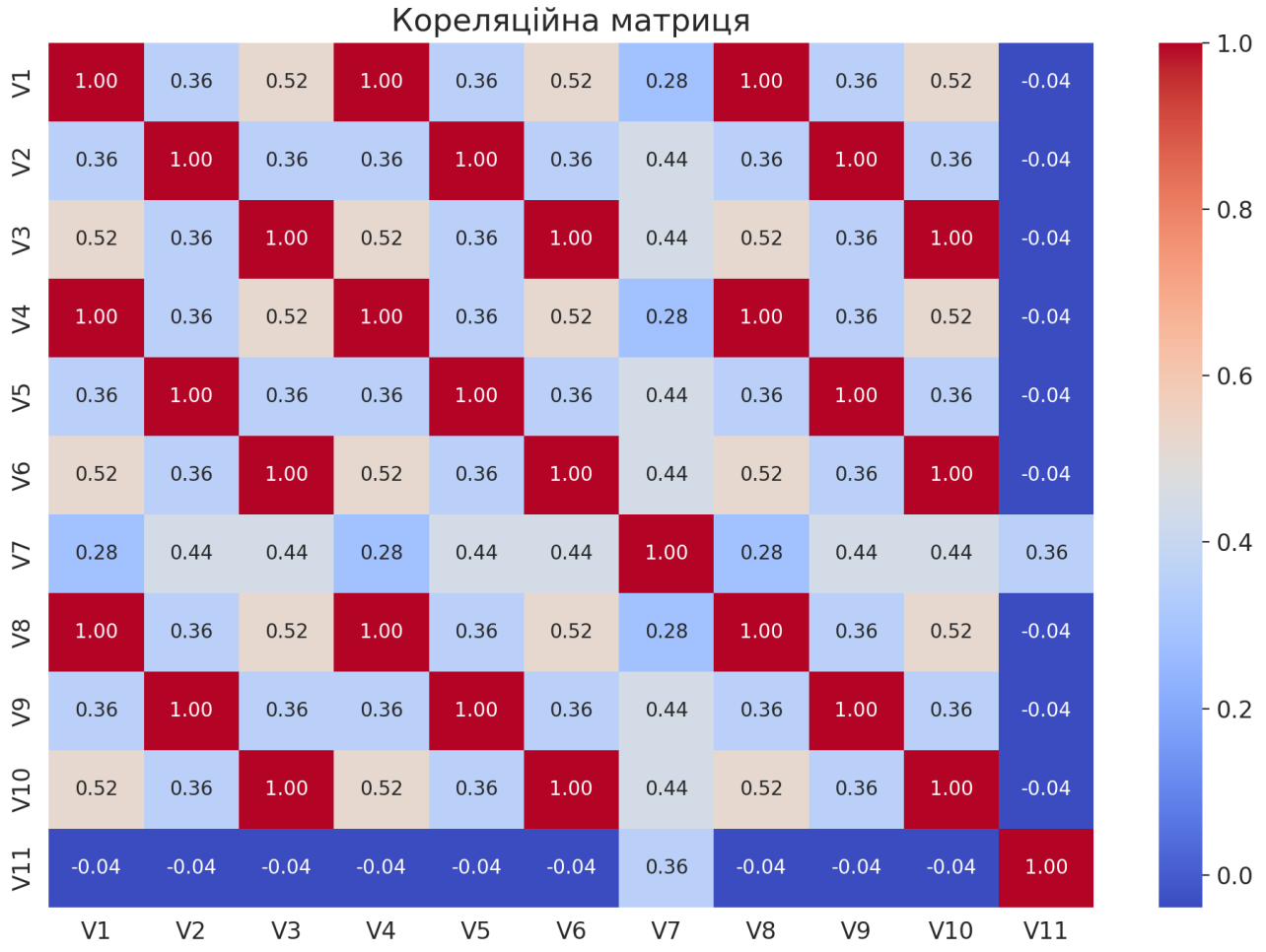

Рисунок 3.1 – Кореляційна матриця

Загалом, розвідувальний аналіз даних, представлений у цьому розділі, є фундаментальним для розуміння структури та взаємозв'язків у наборі даних, що є критично важливим для подальшого моделювання та виявлення шахрайських інтернет-магазинів. Він надає цінну інформацію, яка допомагає у формуванні ефективних стратегій аналізу та класифікації в контексті боротьби з онлайн шахрайством.

3.3 Реалізація запропонованого методу

Проведена класифікація даних методами Logistic Regression (LR), Random Forest (RF), KNN, Naive Bayes (NB), Support Vector (SVM) та DecisionTree (DT). Авторами проведено навчання цих класифікаторів для подальшого вирішення, який класифікатор буде більш ефективним у виявленні шахрайських сайтів. Також кожен метод класифікації, промоделюємо різними підходами, а саме:

- Imbalanced [20];
- Undersampling [21];
- Oversampling [22];
- SMOTE [23];
- ADASYN [24].

Отже, проведемо оцінку класифікації за допомогою ROC кривої та AUC_Test, кожного підходу окремо.

Далі проаналізовано результати класифікації методом DecisionTree (рисунок 3.2). Перший підхід, який розглядаємо, це Imbalanced, Undersampling та SMOTE, і всі вони виявились не надто

точними, з точністю класифікації на рівні 66%. Це можна вважати досить низьким результатом. З іншого боку, підходи Oversampling та ADASYN продемонстрували вражаючі результати, з точністю класифікації в 90% і навіть 100% відповідно. Ці результати свідчать про те, що методи Oversampling та ADASYN виявилися дуже ефективними у покращенні точності класифікації для даної задачі.

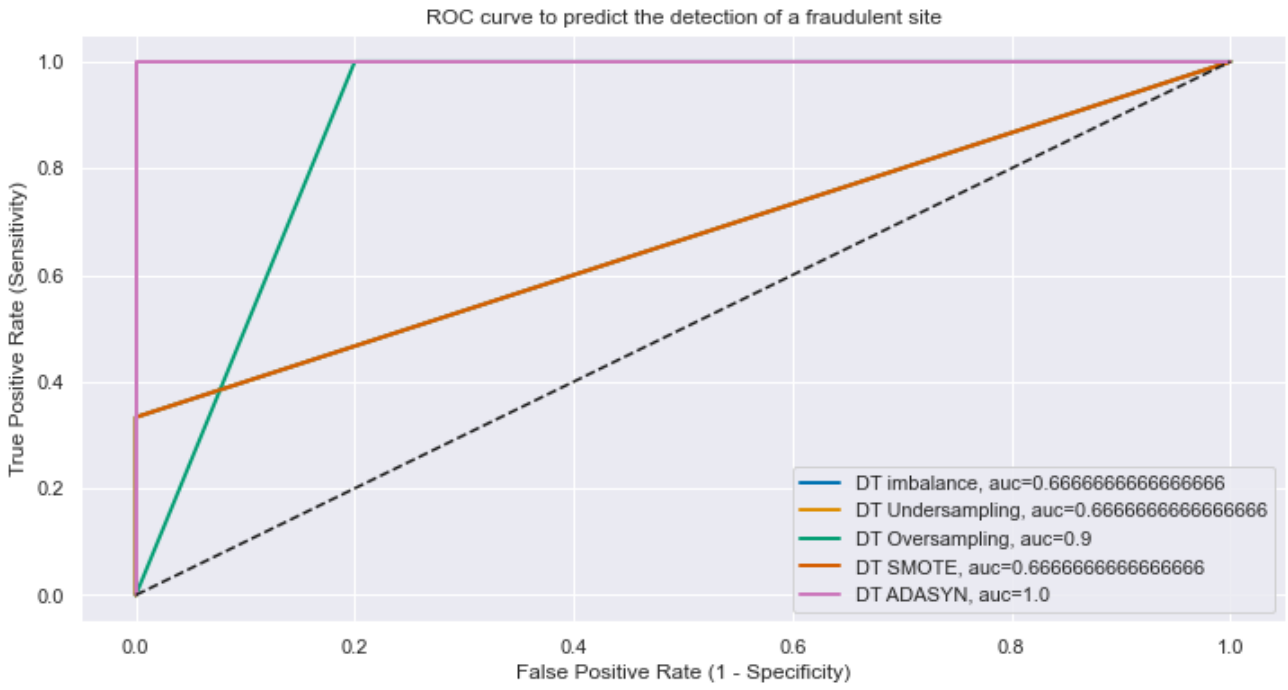

Рисунок 3.2 - Класифікація методом DecisionTree

Далі розглянемо класифікацію методом Logistic Regression (рисунок 3.3). Підходи Imbalanced, SMOTE, ADASYN та Oversampling показали дуже високу точність визначення шахрайських сайтів на рівні 90%. Ці результати свідчать про ефективність цих підходів у вирішенні завдань класифікації шахрайських сайтів. Натомість, підхід Undersampling продемонстрував нижчу точність класифікації на рівні

73%, що можна вважати менш задовільним результатом порівняно з іншими методами.

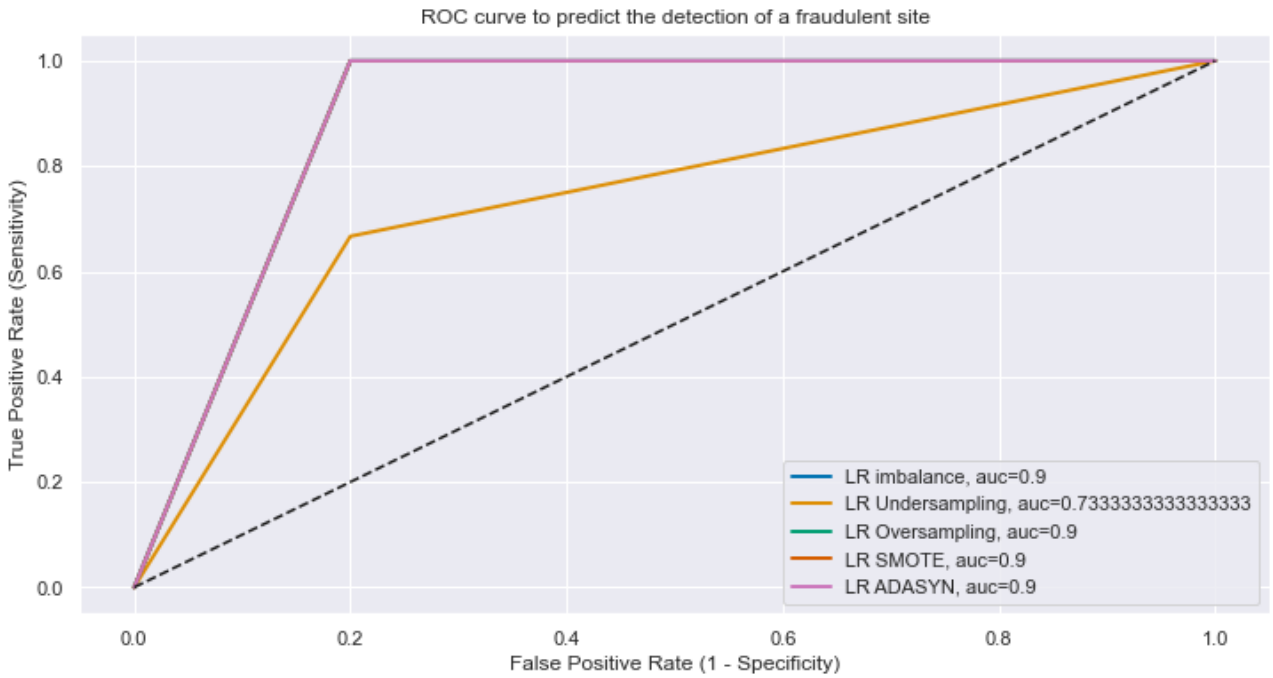

Рисунок 3.3 - Класифікація методом Logistic Regression

При розгляді класифікації методом Random Forest, яка представлена на рисунку 3.4, можна відзначити наступні результати. Підхід Undersampling показав найнижчу точність класифікації на рівні 66%, що вважається дуже низьким показником. Методи Imbalanced, SMOTE та ADASYN продемонстрували подібний рівень точності класифікації, який становить 83%. Незважаючи на те, що ці результати не є найкращими, вони все ж вказують на певний рівень ефективності цих підходів. Найкращим результатом виявився підхід Oversampling, де ROC-крива показала ідеальний результат на рівні 100%, що свідчить про абсолютну точність класифікації.

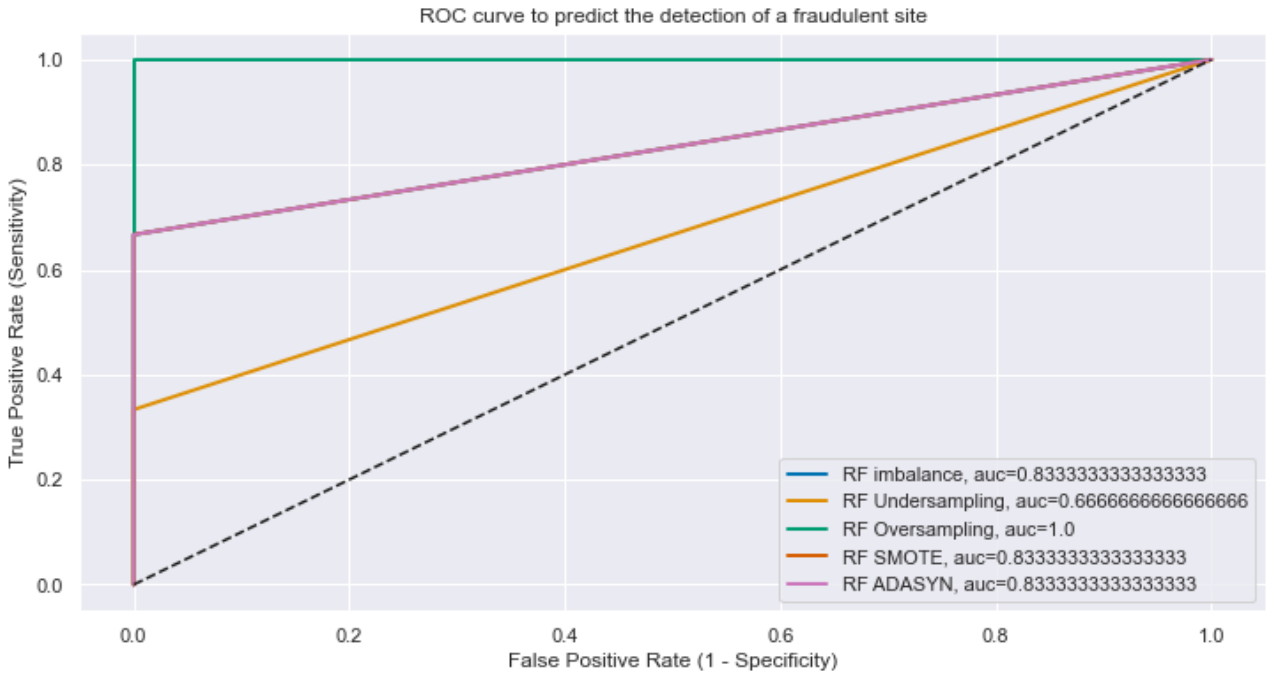

Рисунок 3.4 - Класифікація методом Random Forest

У випадку класифікації методом KNN, зображеної на рисунку 3.5, слід зауважити, що всі різні підходи продемонстрували схожі результати. Кожен з них виявив точність класифікації на рівні 83%, що, на жаль, можна вважати не дуже високим показником успішності моделі. Такий результат вказує на те, що метод KNN не дуже добре підходить для цієї конкретної задачі класифікації шахрайських сайтів, незважаючи на різні підходи до обробки незбалансованих даних. Це може свідчити про те, що для покращення результатів, можливо, слід розглянути інші методи класифікації або додаткову оптимізацію параметрів KNN-моделі.

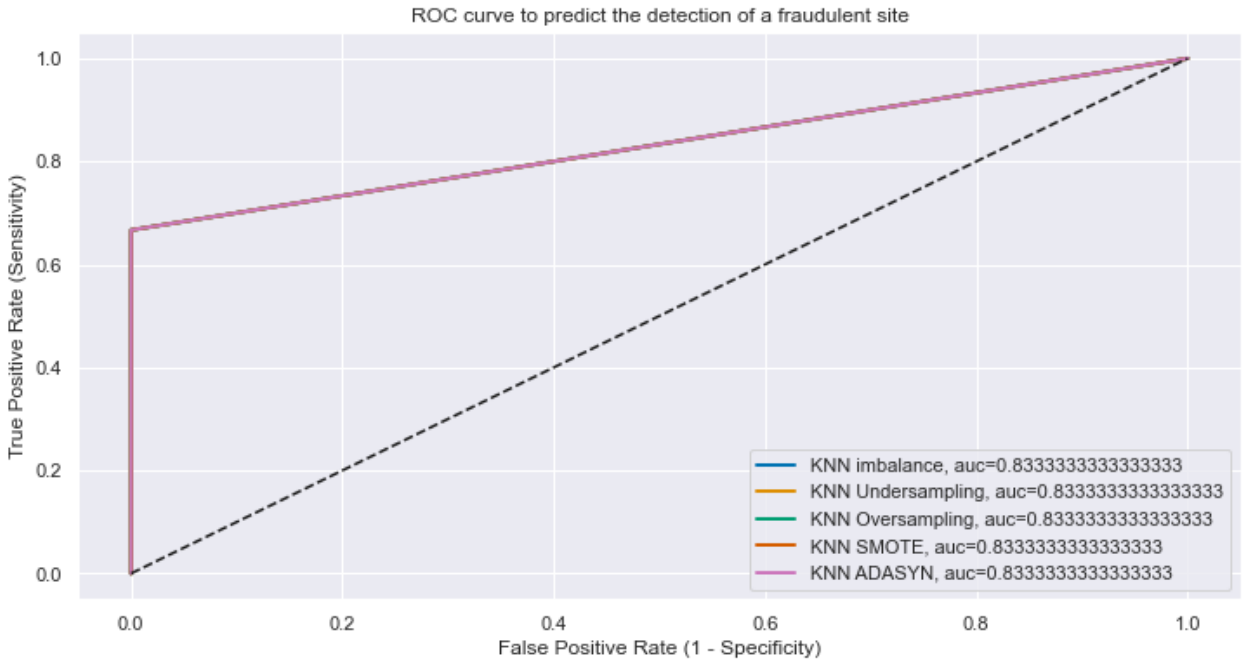

Рисунок 3.5 - Класифікація методом KNN

У випадку класифікації методом SVM, поданої на рисунку 3.6, слід відзначити, що всі різні підходи не досягли дуже високого рівня точності в класифікації. Підхід Undersampling та Oversampling показали найнижчий рівень точності на рівні 66%, що може бути вважено досить поганим результатом. У той час як Imbalanced, SMOTE та ADASYN демонструють незначно кращу точність класифікації на рівні 83%, це також не є дуже високим результатом. Ці результати можуть вказувати на те, що метод SVM може бути менш підходящим для даної задачі класифікації шахрайських сайтів, і можливо, для досягнення кращих результатів, слід розглянути інші методи машинного навчання або подальшу настройку параметрів SVM-моделі.

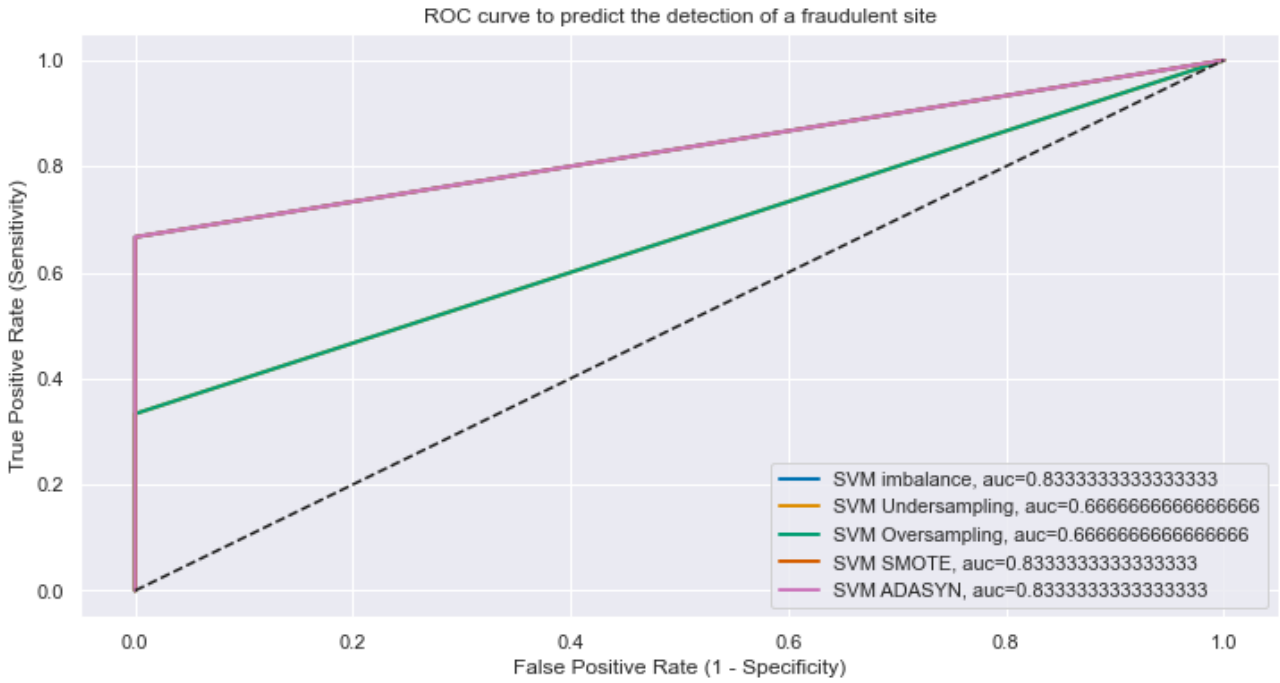

Рисунок 3.6 - Класифікація методом SVM

Результати класифікації методом Naive Bayes, представлені на рисунку 3.7, свідчать про те, що підходи Imbalanced, SMOTE та ADASYN досягли однакового рівня точності класифікації на рівні 83%. Це можна вважати високим результатом, оскільки ці підходи ефективно виявили шахрайські сайти. У той час як підхід Undersampling демонструє найнижчий рівень точності класифікації на рівні 56%, що є досить поганим результатом і може свідчити про обмежену ефективність цього методу в даній задачі. Точність класифікації підходу Oversampling також не вражає, становлячи 73%, що також може вважатися не дуже задовільним результатом. Таким чином, для досягнення високої точності в класифікації шахрайських сайтів, метод Naive Bayes краще застосовувати з підходами, які забезпечують більший баланс класів, такі як Imbalanced, SMOTE та ADASYN..

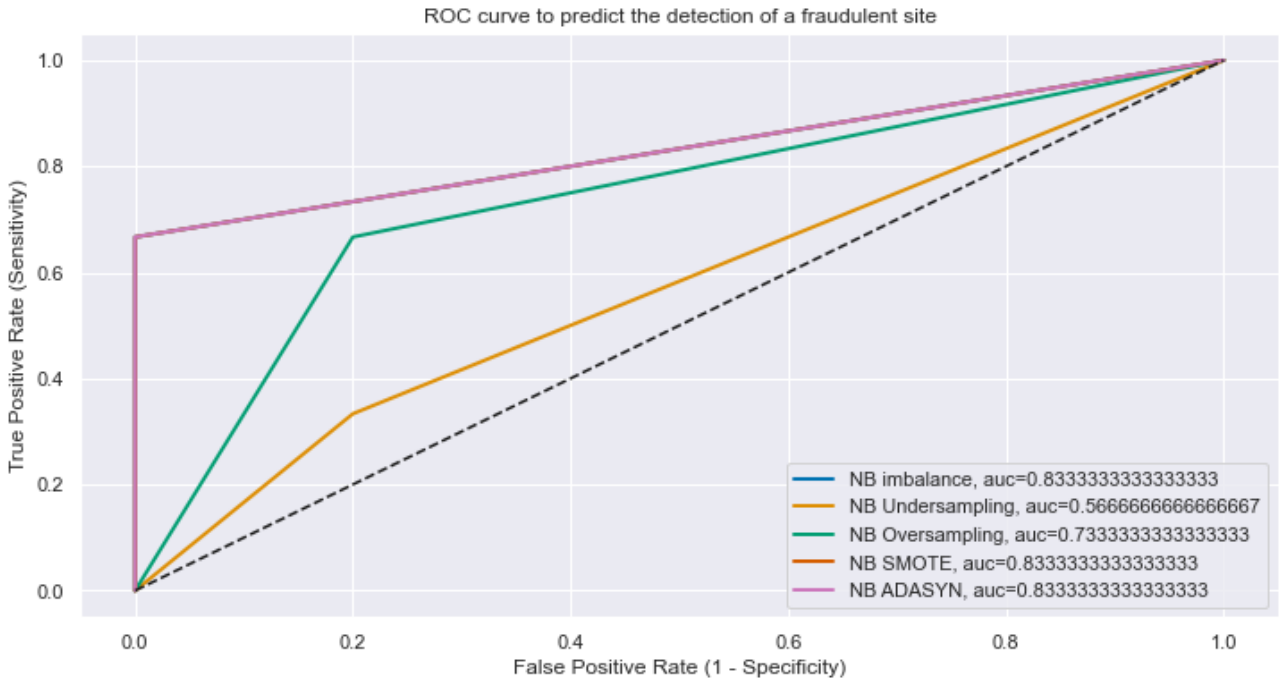

Рисунок 3.7 - Класифікація методом Naive Bayes (NB)

Після узагальнення всіх отриманих результатів (див. Таблиця 3.1), можна виокремити кілька важливих спостережень. З результатів моделювання видно, що найкращі оцінки отримані при використанні методів DecisionTree з підходом ADASYN та Random Forest з підходом Oversampling, обидва методи показали ідеальний результат визначення шахрайських сайтів на рівні 100%. Це свідчить про великий потенціал цих методів для ефективного виявлення шахрайських сайтів і рекомендує їх для подальших досліджень та застосувань.

Також важливо відзначити, що зі збільшенням обсягу доступних даних можуть змінюватися результати класифікації для інших підходів. Тому важливо розглядати всі підходи на основі методів DecisionTree та Random Forest, оскільки вони показали найкращі результати моделювання на поточному обсязі даних. Проте при збільшенні обсягу даних може виникнути необхідність переоцінити підходи та можливо

вибрати інший метод або комбінацію методів для досягнення ще кращої продуктивності та точності виявлення шахрайських сайтів.

Такблиця 3.1 - Результати моделювання

	Model	Accuracy_Test	AUC_Test	PrecisionScore_Test	RecallScore_Test	F1Score_Test
9	DT ADASYN	1.000	1.000000	1.000000	1.000000	1.000000
22	RF Oversampling	1.000	1.000000	1.000000	1.000000	1.000000
0	LR imbalance	0.875	0.900000	0.750000	1.000000	0.857143
2	LR Oversampling	0.875	0.900000	0.750000	1.000000	0.857143
3	LR SMOTE	0.875	0.900000	0.750000	1.000000	0.857143
4	LR ADASYN	0.875	0.900000	0.750000	1.000000	0.857143
7	DT Oversampling	0.875	0.900000	0.750000	1.000000	0.857143
10	KNN imbalance	0.875	0.833333	1.000000	0.666667	0.800000
11	KNN Undersampling	0.875	0.833333	1.000000	0.666667	0.800000
12	KNN Oversampling	0.875	0.833333	1.000000	0.666667	0.800000
13	KNN SMOTE	0.875	0.833333	1.000000	0.666667	0.800000
14	KNN ADASYN	0.875	0.833333	1.000000	0.666667	0.800000
15	SVM imbalance	0.875	0.833333	1.000000	0.666667	0.800000
18	SVM SMOTE	0.875	0.833333	1.000000	0.666667	0.800000
19	SVM ADASYN	0.875	0.833333	1.000000	0.666667	0.800000
20	RF imbalance	0.875	0.833333	1.000000	0.666667	0.800000
23	RF SMOTE	0.875	0.833333	1.000000	0.666667	0.800000
24	RF ADASYN	0.875	0.833333	1.000000	0.666667	0.800000
25	NB imbalance	0.875	0.833333	1.000000	0.666667	0.800000
28	NB SMOTE	0.875	0.833333	1.000000	0.666667	0.800000
29	NB ADASYN	0.875	0.833333	1.000000	0.666667	0.800000
1	LR Undersampling	0.750	0.733333	0.666667	0.666667	0.666667
27	NB Oversampling	0.750	0.733333	0.666667	0.666667	0.666667
5	DT imbalance	0.750	0.666667	1.000000	0.333333	0.500000
6	DT Undersampling	0.750	0.666667	1.000000	0.333333	0.500000
8	DT SMOTE	0.750	0.666667	1.000000	0.333333	0.500000
16	SVM Undersampling	0.750	0.666667	1.000000	0.333333	0.500000
17	SVM Oversampling	0.750	0.666667	1.000000	0.333333	0.500000
21	RF Undersampling	0.750	0.666667	1.000000	0.333333	0.500000
26	NB Undersampling	0.625	0.566667	0.500000	0.333333	0.400000

Отже, представлені результати реалізації запропонованого методу для класифікації шахрайських інтернет-магазинів, використовуючи

різні підходи та алгоритми машинного навчання, включаючи Logistic Regression, Random Forest, KNN, Naive Bayes, Support Vector та DecisionTree. Кожен з цих методів був протестований за допомогою різних стратегій, таких як Imbalanced, Undersampling, Oversampling, SMOTE та ADASYN, для визначення їх ефективності у виявленні шахрайських сайтів. Результати показали, що підходи ADASYN у DecisionTree та Oversampling у Random Forest демонструють найвищу ефективність, досягаючи 100% точності у виявленні шахрайських сайтів. Це свідчить про значний потенціал цих підходів у боротьбі з онлайн шахрайством. Однак, слід врахувати, що результати можуть змінюватися зі збільшенням обсягу даних, тому рекомендується розглядати всі підходи, особливо ті, що базуються на методах DecisionTree та Random Forest, оскільки вони показали найкращі результати у дослідженні.

3.4 Прототип системи визначення шахрайських інтернет-магазинів

Розвиток цифрових технологій відкриває нові горизонти для інтернет-торгівлі, але разом з тим збільшує й ризики взаємодії з шахрайськими платформами. Відповідно до цієї задачі, було розроблено прототип системи для визначення ненадійних інтернет-магазинів. Система створена з використанням сучасних бібліотек Python, що дозволяє ефективно аналізувати та класифікувати веб-ресурси за допомогою інтелектуальних методів. Візуальний інтерфейс платформи, представлений на Рисунку 3.8, демонструє високий рівень

інтуїтивності та зручності використання, а також забезпечує користувачу швидкий доступ до результатів перевірки надійності веб-сайтів. Рисунок 3.9, у свою чергу, показує результати визначення потенційно шахрайського сайту, демонструючи ефективність системи в ідентифікації ризиків, асоційованих з електронною комерцією.

На Рисунку 3.8 зображено інтерфейс веб-платформи, де користувач може перевірити надійність інтернет-магазину. У центрі знаходиться веб-сторінка магазину "rozetka.com.ua", яка виглядає професійною та має чітку структуру. З правого боку інтерфейсу розташована панель з інструментом для визначення, який включає кнопку "START" для запуску процесу перевірки. Під кнопкою розміщено зелений галочку, що позначає "YES", сигналізуючи про те, що сайт є надійним і на ньому можна безпечно робити покупки.

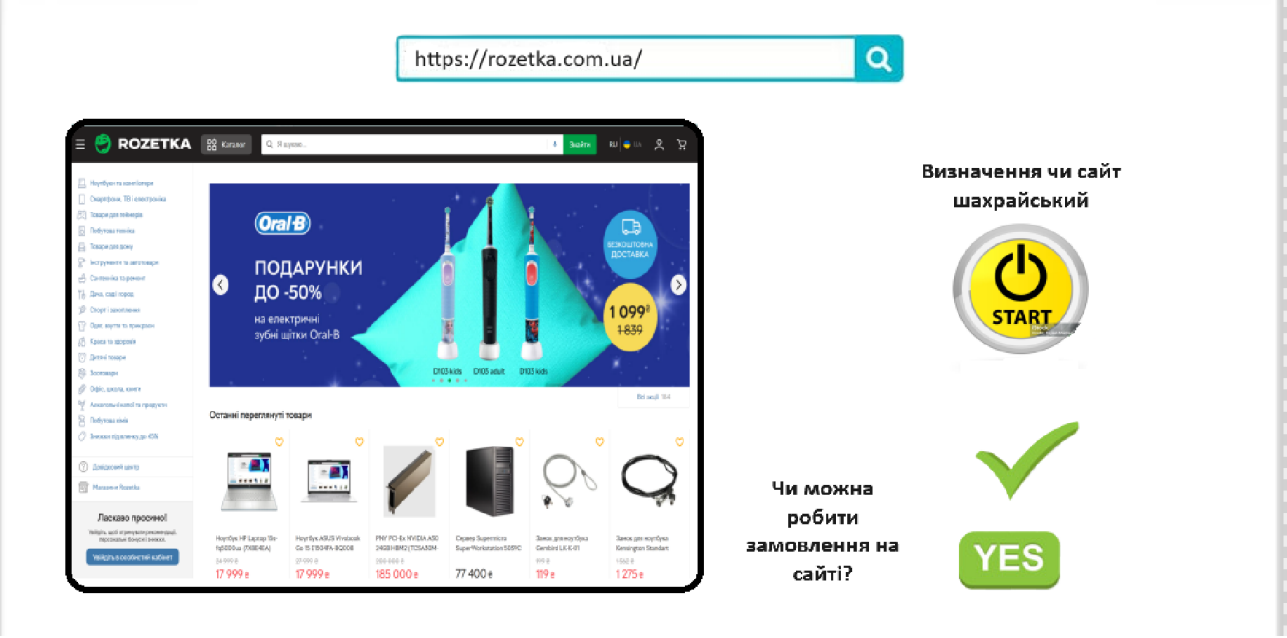

Рисунок 3.8 – Вікно веб-платформи де визначено надійний сайт

Рисунок 3.9 показує інший інтерфейс веб-платформи, де визначено потенційно шахрайський сайт. Центральна частина інтерфейсу відображає іншу веб-сторінку з URL "rosetka.ru". Праворуч також знаходиться панель з інструментом визначення, подібна до попередньої, але з червоним хрестиком під кнопкою "START", що позначає "NO", індикуючи, що сайт може бути шахрайським і користувачам слід уникати здійснення покупок на ньому.

Рисунок 3.9 - Вікно веб-платформи де визначено шахрайський сайт

Отже, розроблений прототип системи для визначення шахрайських інтернет-магазинів, який є важливим кроком у боротьбі з онлайн шахрайством в сучасному світі електронної комерції. Система, розроблена на основі передових бібліотек Python, забезпечує ефективний аналіз та класифікацію веб-ресурсів, використовуючи інтелектуальні методи. Інтуїтивно зрозумілий візуальний інтерфейс, представлений на рисунках 3.8 та 3.9, демонструє зручність

використання системи та забезпечує користувачам швидкий доступ до результатів перевірки сайтів. Ця система не тільки підвищує рівень безпеки користувачів у цифровому просторі, але й сприяє підвищенню довіри до електронної комерції, надаючи засоби для ідентифікації потенційно ненадійних інтернет-магазинів.

Висновки до розділу 3

1. У цьому розділі представлено розробку скрипта на Python, який є ключовим інструментом у виявленні шахрайських інтернет-магазинів. Процес розробки охоплює кілька етапів, від збору даних з веб-сайтів до аналізу відгуків та використання різних методів машинного навчання для класифікації сайтів. Важливим аспектом є необхідність доступу до Інтернету, специфічних API та додаткових ресурсів для ефективної реалізації цих кроків. Вибір підозрілого сайту, збір даних за допомогою бібліотек Python та підготовка даних для аналізу відгуків є важливими етапами у процесі ідентифікації шахрайських сайтів.
2. Розвідувальний аналіз даних, проведений на основі набору даних з 67 українських веб-сайтів, виявив рівномірний розподіл значень у бінарних та числових змінних, а також слабкі взаємозв'язки між цими змінними. Це підкреслює унікальність кожної змінної та її важливість у наборі даних, що є ключовим для подальшого моделювання та аналізу. Кореляційна матриця, що демонструє ці взаємозв'язки, є важливим інструментом для розуміння структури даних.
3. Реалізація запропонованого методу включає класифікацію даних за допомогою різних методів машинного навчання та оцінку їх ефективності. Результати показали, що підходи ADASYN у DecisionTree та Oversampling у Random Forest є найбільш ефективними, досягаючи 100% точності у виявленні шахрайських сайтів. Це вказує на високий потенціал цих методів

у боротьбі з онлайн шахрайством, хоча результати можуть змінюватися зі збільшенням обсягу даних.

4. Прототип системи для визначення шахрайських інтернет-магазинів, розроблений у цьому розділі, демонструє ефективність у ідентифікації ризиків, пов'язаних з електронною комерцією. Інтуїтивно зрозумілий візуальний інтерфейс та швидкий доступ до результатів перевірки сайтів забезпечують користувачам зручний інструмент для захисту від онлайн шахрайства. Система сприяє підвищенню довіри до електронної комерції, надаючи засоби для ідентифікації потенційно ненадійних інтернет-магазинів.

ВИСНОВКИ

У цій роботі розглянуто проблему шахрайських інтернет-магазинів, яка є актуальною в умовах стрімкого розвитку цифрових технологій. Шахрайство в онлайн середовищі не лише призводить до фінансових втрат споживачів, але й підриває довіру до електронної комерції. Українське законодавство визначає інтернет-магазини як засіб для електронних правочинів, що створює потенціал для шахрайських дій. Важливість освіти споживачів та розробки інноваційних методів для виявлення та запобігання шахрайству стає ключовою в цьому контексті.

Аналіз існуючих підходів до протидії шахрайству в онлайн середовищі виявив широкий спектр стратегій, від баз даних шахрайських веб-ресурсів до розробки алгоритмів машинного навчання. Особливо важливими є методи для виявлення фішингових сайтів та класифікації рекламних показів. Незважаючи на існуючі дослідження, багато з них не зосереджувалися безпосередньо на виявленні шахрайських інтернет-магазинів через машинне навчання, тому розробка спеціалізованих інтелектуальних методів є актуальною та може значно підвищити ефективність виявлення та протидії шахрайству.

Ідентифікація потенційної небезпеки в інтернет-торгівлі є ключовим аспектом у боротьбі з шахрайством в онлайн середовищі. Використання інтелектуальних методів для аналізу ключових параметрів, що можуть вказувати на шахрайство, є важливим кроком у забезпеченні безпеки споживачів. Уважне вивчення цих параметрів

може допомогти споживачам уникнути ризиків та забезпечити безпечні та прозорі онлайн-покупки.

У параграфі 2.1 надано огляд розробленої системи для виявлення шахрайських інтернет-магазинів, що відзначається всебічним підходом, об'єднуючи різноманітні технології та методології для ефективного виявлення шахрайства. Центральним складовим елементом цієї системи є інтелектуальний метод, який інтегрує в себе аналіз даних, машинне навчання та прогнозування з метою точного визначення шахрайських дій. Система обробляє вхідні дані, проводить їх перевірку та надає результати, сприяючи користувачам у виявленні ненадійних магазинів. Крім того, модуль інтеграції, що включає API та базу даних, забезпечує гармонійну взаємодію між компонентами цієї системи.

У параграфі 2.2 віддано детальний огляд методів класифікації на основі машинного навчання, які є фундаментальними для створення ефективних систем виявлення шахрайських інтернет-магазинів. Різноманітність цих методів, таких як логістична регресія, випадковий ліс, KNN, наївний Байєсівський класифікатор, SVM та дерево рішень, демонструє глибину та гнучкість підходів до класифікації. Кожен з цих методів приносить унікальний внесок у процес аналізу, що дозволяє системі точно визначати потенційно шахрайські сайти.

У параграфі 2.3 розглядається проблема незбалансованих даних, яка може виникати в різних задачах машинного навчання, зокрема у виявленні рідкісних подій, таких як шахрайства в електронній комерції. Вирішення цієї проблеми є ключовим для досягнення ефективності та точності моделей машинного навчання. У цьому контексті розглядаються різні методи, такі як підгенерація, надгенерація, SMOTE

та ADASYN, кожен із яких має свої переваги та може бути використаний залежно від конкретних вимог задачі та особливостей даних. Використання цих методів сприяє досягненню більшої рівноваги в розподілі класів у навчальних даних, покращуючи якість та надійність моделей машинного навчання у ситуаціях з незбалансованими даними. Важливо враховувати, що вибір конкретного методу повинен базуватися на конкретних вимогах завдання та особливостях даних, і відповідний підхід до обробки незбалансованих даних може значно поліпшити результати та продуктивність моделей машинного навчання у виявленні рідкісних подій та інших ситуаціях з незбалансованими даними.

Розроблений інтелектуальний метод, представлений у параграфі 2.4, є ключовим інструментом у боротьбі з онлайн шахрайством. Цей метод включає в себе кроки від вибору підозрілого сайту до збору та аналізу даних, а також використання передових методів класифікації для оцінки надійності сайтів. Завершальний етап роботи методу полягає у виведенні результатів та визначенні статусу сайту, що допомагає користувачам отримувати високі показники точності та надійності при виявленні шахрайських магазинів. Цей комплексний підхід не лише успішно ідентифікує шахрайські сайти, але й надає користувачам додатковий рівень захисту, допомагаючи уникнути можливих фінансових втрат та інших негативних наслідків.

У параграфі 3.1 описана розробка Python-скрипта, який є ключовим інструментом у виявленні шахрайських інтернет-магазинів. Процес розробки включає кілька етапів, починаючи від збору даних з веб-сайтів і завершуючи аналізом відгуків та використанням різних

методів машинного навчання для класифікації сайтів. Для успішної реалізації цих кроків необхідний доступ до Інтернету, використання спеціалізованих API та додаткових ресурсів. Важливими аспектами є вибір підозрілих сайтів, використання Python-бібліотек для збору даних та підготовки їх для аналізу відгуків.

Розвідувальний аналіз даних, проведений на основі набору даних з 67 українських веб-сайтів, показав, що значення в бінарних та числових змінних розподілені рівномірно, а взаємозв'язки між ними слабкі. Це підкреслює важливість кожної змінної у наборі даних та її унікальність, що є ключовим для подальшого моделювання та аналізу. Кореляційна матриця, яка відображає ці зв'язки, є важливим інструментом для кращого розуміння структури даних.

Реалізація запропонованого методу включає в себе класифікацію даних за допомогою різних методів машинного навчання та оцінку їх ефективності. Результати дослідження показали, що підходи, такі як ADASYN у DecisionTree та Oversampling у Random Forest, є найбільш ефективними, досягаючи 100% точності у виявленні шахрайських сайтів. Це свідчить про високий потенціал цих методів у боротьбі з онлайн шахрайством, хоча результати можуть змінюватися в залежності від обсягу даних.

Прототип системи для визначення шахрайських інтернет-магазинів, розроблений у цьому розділі, продемонстрував свою ефективність у виявленні ризиків, пов'язаних з електронною комерцією. Ця система має інтуїтивно зрозумілий візуальний інтерфейс та надає користувачам швидкий доступ до результатів перевірки веб-сайтів, що забезпечує зручний інструмент для захисту від онлайн шахрайства.

Вона також сприяє підвищенню довіри до електронної комерції, надаючи засоби для ідентифікації потенційно ненадійних інтернет-магазинів.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

[1] Smetanko, O. V. (2015). Udoskonalennya protsesu vnutrishn′oho audytu prychyn shakhraystva v systemi korporatyvnoho upravlinnya. *Ekonomichnyy forum, 3,* 424-430. [in Ukrainian].

[2] Shapochka, S. V. (2014). Do pytannya borot′by z shakhraystvom, yake vchynyayet′sya z vykorystannyam mozhlyvostey merezhi Internet. *Pravova informatyka, 3 (43),* 89-95. [in Ukrainian].

[3] Kiberpolitsiya fiksuye zbil′shennya vypadkiv shakhraystv, vchynenykh pid vyhlyadom investuvannya na finansovykh rynkakh. (15 lyutoho 2019 r.). Retrieved from: link [in Ukrainian].

[4] Z pochatku roku nadiyshlo ponad 25 tysyach zvernen′ shchodo shakhraystva v interneti – kiberpolitsiya. Retrieved from: link [in Ukrainian].

[5] Skachko, Y. E. (2017, September 29). Top-5 novyn v sferi shakhraystva z platizhnymy instrumentamy. Retrieved from: link [in Ukrainian].

[6] EU Serious and Organised Crime Threat Assessment (SOCTA 2013). Retrieved from: link [in English].

[7] Sabadash, I., Dumanskyi, N., & Korobiichuk, I. (2020). Methods and Means of Identifying Fraudulent Websites. In *COAPSN*. ceur-ws.org, pp. 177-186.

[8] Machado, L., & Gadge, J. (2017). Phishing Sites Detection Based on C4.5 Decision Tree Algorithm. In *2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA),* pp. 1-5. doi: 10.1109/ICCUBEA.2017.8463818.

[9] Haider, C. M. R., Iqbal, A., Rahman, A. H., & Rahman, M. S. (2018). An ensemble learning based approach for impression fraud detection in mobile advertising. *Journal of Network and Computer Applications, 112,* 126-141.

[10] Thejas, G. S., et al. (2019). A Multi-time-scale Time Series Analysis for Click Fraud Forecasting using Binary Labeled Imbalanced Dataset. In *2019 4th International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS),* pp. 1-8. doi: 10.1109/CSITSS47250.2019.9031036.

[11] Zhang, X., Liu, X., & Guo, H. (2018). A Click Fraud Detection Scheme based on Cost sensitive BPNN and ABC in Mobile Advertising. In *2018 IEEE 4th International Conference on Computer and Communications (ICCC),* pp. 1360-1365. doi: 10.1109/CompComm.2018.8780941.

[12] Sisodia, D., & Sisodia, D. S. (2021). Quad division prototype selection-based k-nearest neighbor classifier for click fraud detection from highly skewed user click dataset. *Engineering Science and Technology, an International Journal.* doi: 10.1016/j.jestch.2021.05.015.

[13] Gillespie, A.A., & Magor, S. (2020). Tackling online fraud. *ERA Forum, 20,* 439–454. doi: 10.1007/s12027-019-00580-y

[14] Gramyak, R., Lipyanina-Goncharenko, H., Sachenko, A., Lendyuk, T., & Zahorodnia, D. (2021). Intelligent Method of a Competitive Product Choosing based on the Emotional Feedbacks Coloring. In *IntelITSIS (pp. 246-257).*

[15] Lipyanina, H., et al. (2020, August). Assessing the Investment Risk of Virtual IT Company Based on Machine Learning. In *International*

Conference on Data Stream Mining and Processing (pp. 167-187). Springer, Cham.

[16] Savenko, O., et al. (2020). BOTNET DETECTION APPROACH BASED ON THE DISTRIBUTED SYSTEMS. *International Journal of Computing*, *19*(2), 190-198. doi: 10.47839/ijc.19.2.1761

[17] Makki, S., et al. (2019). An Experimental Study With Imbalanced Classification Approaches for Credit Card Fraud Detection. *IEEE Access*, *7*, 93010-93022. doi: 10.1109/ACCESS.2019.2927266.

[18] Kang, Q., et al. (2018). A Distance-Based Weighted Undersampling Scheme for Support Vector Machines and its Application to Imbalanced Classification. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(9), 4152-4165. doi: 10.1109/TNNLS.2017.2755595.

[19] Mathew, J., et al. (2018). Classification of Imbalanced Data by Oversampling in Kernel Space of Support Vector Machines. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(9), 4065-4076. doi: 10.1109/TNNLS.2017.2751612.

[20] Demidova, L., & Klyueva, I. (2017). SVM classification: Optimization with the SMOTE algorithm for the class imbalance problem. *2017 6th Mediterranean Conference on Embedded Computing (MECO)*, pp. 1-4. doi: 10.1109/MECO.2017.7977136.

[21] Dan, W., & Yian, L. (2020). Denoise-Based Over-Sampling for Imbalanced Data Classification. *2020 19th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)*, pp. 1-4. doi: 10.1109/DCABES50732.2020.00078.

[22] Загальні рекомендації з підготовки, оформлення, захисту та оцінювання випускних кваліфікаційних робіт здобувачів вищої освіти

першого «бакалаврського» і другого «магістерського» рівнів / За ред. доц. М.І. Шинкарика. Тернопіль: ТНЕУ, 2018. 67 с.

[23] Комар М.П., Саченко А.О., Васильків Н.М. Методичні рекомендації до виконання кваліфікаційної роботи з освітньо-професійної програми «Комп'ютерні науки» спеціальності 122 «Комп'ютерні науки» за другим (магістерським) рівнем вищої освіти. Тернопіль: ЗУНУ, 2021. 32 с.

ДОДАТОК А

АПРОБАЦІЯ ОТРИМАНИХ РЕЗУЛЬТАТІВ

УДК 082:001
Е 45

Голова оргкомітету: Коренюк І.О.
Верстка: Зрада С.І.
Дизайн: Бондаренко І.В.

Конференцію зареєстровано Державною науковою установою «УкрІНТЕІ» в базі даних науково-технічних заходів України та інформаційному бюлетені «План проведення наукових, науково-технічних заходів в Україні» (Посвідчення №320 від 16.06.2023).

Е 45 **Експериментальні та теоретичні дослідження в контексті сучасної науки**: матеріали IV Всеукраїнської студентської наукової конференції, м. Чернігів, 29 вересня, 2023 рік / ГО «Молодіжна наукова ліга». — Вінниця: ТОВ «УКРЛОГОС Груп», 2023. — 154 с.

ISBN 978-617-8126-63-6
DOI 10.36074/liga-ukr-29.09.2023

Викладено матеріали учасників IV Всеукраїнської мультидисциплінарної студентської наукової конференції «Експериментальні та теоретичні дослідження в контексті сучасної науки», яка відбулася 29 вересня 2023 року у місті Чернігів, Україна.

УДК 082:001

ISBN 978-617-8126-63-6

СЕКЦІЯ 8.
ФІЗИКО-МАТЕМАТИЧНІ НАУКИ

Сукмановський Олесь, здобувач вищої освіти факультету комп'ютерних інформаційних технологій
Західноукраїнський національний університет, Україна

Науковий керівник: Ліп'яніна-Гончаренко Христина Володимирівна, канд. техн. наук, доцент, доцент кафедри інформаційно-обчислювальних систем та управління
Західноукраїнський національний університет, Україна

ІНТЕЛЕКТУАЛЬНИЙ МЕТОД ВИЗНАЧЕННЯ ШАХРАЙСЬКИХ ІНТЕРНЕТ-МАГАЗИНІВ

Сучасний інтернет-світ зазнав раптового зростання кіберзлочинності, зокрема, інтернет-шахрайства, яке стало великою загрозою для інтернет-користувачів та фінансової стійкості. Ця стаття розглядає проблему інтернет-шахрайства та пропонує інтелектуальний метод для визначення шахрайських інтернет-магазинів. Вона також обговорює важливість боротьби з цією формою кіберзлочинності та вплив насильства на фінансовий сектор і довіру споживачів.

Для визначення шахрайських сайтів пропонується інтелектуальний метод, що базується на наступних параметрах, що представленні в таблиці 1.

Таблиця 1.

Параметри

№	Опис	Позначення
1	Відсутність інформації про призначення товару	V1
2	Відсутність інформації про склад матеріалу, з якого зроблений товар	V2
3	Відсутність інформації про продукцію державною мовою (на українських сайтах).	V3
4	Приховування недоліків товарів (при продажі товарів, що були в користуванні)	V4
5	Відсутність інформації про стан зношеності товарів, якщо вони були у користуванні	V5
6	Відсутність інформації про виробника товару.	V6
7	Відсутність інформації про місце найближчого сервісного обслуговування товару.	V7
8	Відсутність інформації про гарантійний строк товару.	V8
9	Відсутність адреси, за якою покупець може повернути товар протягом чотирнадцяти днів з моменту купівлі.	V9
10	Відсутність інформації про виробника товару.	V10
11	Відсутність інформації про продавця товару (що унеможливлює звернення до суду з позовом).	V11

Інтелектуальний метод визначення шахрайських інтернет-магазинів представлено наступними кроками:

Крок 1. Вибір сайту, що може бути шахрайським.

Крок 2. Збір даних з сайту. Всі параметри (таблиця 1) парсимо з зазначеного сайту та при наявності інформації система прописує 0, при відсутності ставить 1.

Крок3. Для параметру V4 (за наявності 1), проводимо аналіз відгуків користувачів, інтелектуальним методом вибору конкурентного товару на основі емоційного забарвлення відзивів [1]. За наявності позитивного відзиву проставляємо 0, негативного 1.

Крок 4. Проводимо класифікацію чотирьома найпопулярнішими методами

класифікації.

Крок 4.1. Logistic Regression [2].

Крок 4.2. Random Forest [3].

Крок 4.3. KNN [3].

Крок 4.4. Naive Bayes [3].

Крок 4.5. Support Vector Classifier [3].

Крок 4.6. DecisionTree Classifier [4].

Крок 5. Вивід отриманих результатів.

Крок 6. Визначення чи сайт шахрайський чи ні, по більшості класифікованих результатів.

Зростаюча кіберзлочинність, зокрема інтернет-шахрайства, є серйозною загрозою для суспільства і економіки. Анонімність та складність відстеження зловмисників ускладнюють завдання правоохоронних органів у боротьбі з цим явищем. До того ж, довгі терміни бюрократичних процедур роблять розслідування ще більш складним завданням. Інтелектуальні методи визначення шахрайських інтернет-магазинів можуть бути корисними для попередження інтернет-шахрайства та захисту користувачів від фінансових втрат. Боротьба з цією загрозою вимагає спільних зусиль як з боку правоохоронних органів, так і з боку інтернет-спільноти, і лише разом ми можемо створити безпечний інтернет-середовище для всіх користувачів.

Список використаних джерел:

1. Gramyak, R., Lipyanina-Goncharenko, H., Sachenko, A., Lendyuk, T., & Zahorodnia, D. (2021). Intelligent Method of a Competitive Product Choosing based on the Emotional Feedbacks Coloring. In IntelITSIS (pp. 246-257).
2. Lipyanina H., Sachenko S., Lendyuk T., Brych V., Yatskiv V., Osolinskiy O. (2021) Method of Detecting a Fictitious Company on the Machine Learning Base. In: Hu Z., Petoukhov S., Dychka I., He M. (eds) Advances in Computer Science for Engineering and Education IV. ICCSEEA 2021. Lecture Notes on Data Engineering and Communications Technologies, vol 83. Springer, Cham. https://doi.org/10.1007/978-3-030-80472-5_12.
3. Lipyanina, H., Maksymovych, V., Sachenko, A., Lendyuk, T., Fomenko, A., & Kit, I. (2020, August). Assessing the Investment Risk of Virtual IT Company Based on Machine Learning. In International Conference on Data Stream Mining and Processing (pp. 167-187). Springer, Cham.
4. Lipyanina, H., Sachenko, S., Lendyuk, T., & Sachenko, A. (2020). Targeting Model of HEI Video Marketing based on Classification Tree. ICTERI Workshops. CEUR Workshop Proceedingsthis link is disabled, 2020, 2732, pp. 487–498.

ЛЬВІВСЬКИЙ НАУКОВИЙ ФОРУМ

МАТЕРІАЛИ

X МІЖНАРОДНОЇ НАУКОВО-ПРАКТИЧНОЇ КОНФЕРЕНЦІЇ

АКТУАЛЬНІ ПИТАННЯ РОЗВИТКУ НАУКИ ТА ОСВІТИ

9-10 грудня 2023 року

УДК 005
ББК 94.3(0)

Актуальні питання розвитку науки та освіти: матеріали X Міжнародної науково-практичної конференції м. Львів, 9-10 грудня 2023 року. – Львів: Львівський науковий форум, 2023. – 308 с.

У даному збірнику представлені тези доповідей учасників Міжнародної науково-практичної конференції «Актуальні питання розвитку науки та освіти», організованої Львівським науковим форумом. Висвітлюються Актуальні питання розвитку науки та освіти на сучасному етапі становлення, розглядаються сучасні наукові дискусії різних наукових напрямів.

Збірник предназначений для студентів, здобувачів наукових ступенів, науковців та практиків.

Всі матеріали представлені в авторській редакції. За повноту та цілісність яких автори безпосередньо несуть відповідальність.

Львівський науковий форум, 2023

РЕАЛІЗАЦІЯ ІНТЕЛЕКТУАЛЬНОГО МЕТОДУ ВИЗНАЧЕННЯ ШАХРАЙСЬКИХ ІНТЕРНЕТ-МАГАЗИНІВ

Сукмановський Олесь
здобувач вищої освіти факультету комп'ютерних інформаційних технологій
Західноукраїнський національний університет, Україна
Науковий керівник: Ліп'яніна-Гончаренко Христина Володимирівна
канд. тех. наук, доцент, доцент кафедри інформаційно-обчислювальних систем та управління
Західноукраїнський національний університет, Україна

Сучасний світ насичений цифровими технологіями, які значно полегшують життя людей, але водночас створюють нові виклики і загрози. Однією з таких загроз є Інтернет-шахрайство, яке активно розвивається і негативно впливає на фінансову безпеку користувачів Інтернету. В цій статті ми дослідимо проблему Інтернет-шахрайства і запропонуємо реалізацію інтелектуального методу для виявлення шахрайських сайтів.

Для проведення експериментальних досліджень сформовано dataset на основі сайтів, що функціонують в Україні. В dataset є 67 сайтів, 45% яких є шахрайськими.

За допомогою мови програмування Phyton проведена класифікація даних методами Logistic Regression (LR), Random Forest (RF), KNN, Naive Bayes (NB), Support Vector (SVM) та DecisionTree (DT). Авторами проведено навчання цих класифікаторів для подальшого вирішення, який класифікатор буде більш ефективним у виявленні шахрайських сайтів. Також кожен метод класифікації, промоделюємо різними підходами, а саме:

- Imbalanced [1];
- Undersampling [2];
- Oversampling [3];
- SMOTE [4];
- ADASYN [5].

Далі підсумуємо всі отримані результати (таблиця 2). З результатів моделювання видно, що найкращі оцінки є на основі DecisionTree підходом ADASYN та Random Forest підходом Oversampling, що демонструє 100% результат визначення шахрайського сайту.

Автори визнають, що при більшій кількості даних результати класифікації можуть змінюватись в сторону покращення для інших підходів, тому варто розглядати всі підходи на основі методів DecisionTree та Random Forest, бо вони показують найкращі результати моделювання.

Таблиця 2. Результати моделювання

	Model	Accuracy_Test	AUC_Test	PrecisionScore_Test	RecallScore_Test	F1Score_Test
9	DT ADASYN	1.000	1.000000	1.000000	1.000000	1.000000
22	RF Oversampling	1.000	1.000000	1.000000	1.000000	1.000000
0	LR imbalance	0.875	0.900000	0.750000	1.000000	0.857143
2	LR Oversampling	0.875	0.900000	0.750000	1.000000	0.857143
3	LR SMOTE	0.875	0.900000	0.750000	1.000000	0.857143
4	LR ADASYN	0.875	0.900000	0.750000	1.000000	0.857143
7	DT Oversampling	0.875	0.900000	0.750000	1.000000	0.857143
10	KNN imbalance	0.875	0.833333	1.000000	0.666667	0.800000
11	KNN Undersampling	0.875	0.833333	1.000000	0.666667	0.800000
12	KNN Oversampling	0.875	0.833333	1.000000	0.666667	0.800000
13	KNN SMOTE	0.875	0.833333	1.000000	0.666667	0.800000
14	KNN ADASYN	0.875	0.833333	1.000000	0.666667	0.800000
15	SVM imbalance	0.875	0.833333	1.000000	0.666667	0.800000
18	SVM SMOTE	0.875	0.833333	1.000000	0.666667	0.800000
19	SVM ADASYN	0.875	0.833333	1.000000	0.666667	0.800000
20	RF imbalance	0.875	0.833333	1.000000	0.666667	0.800000
23	RF SMOTE	0.875	0.833333	1.000000	0.666667	0.800000
24	RF ADASYN	0.875	0.833333	1.000000	0.666667	0.800000
25	NB imbalance	0.875	0.833333	1.000000	0.666667	0.800000
28	NB SMOTE	0.875	0.833333	1.000000	0.666667	0.800000
29	NB ADASYN	0.875	0.833333	1.000000	0.666667	0.800000
1	LR Undersampling	0.750	0.733333	0.666667	0.666667	0.666667
27	NB Oversampling	0.750	0.733333	0.666667	0.666667	0.666667
5	DT imbalance	0.750	0.666667	1.000000	0.333333	0.500000
6	DT Undersampling	0.750	0.666667	1.000000	0.333333	0.500000
8	DT SMOTE	0.750	0.666667	1.000000	0.333333	0.500000
16	SVM Undersampling	0.750	0.666667	1.000000	0.333333	0.500000
17	SVM Oversampling	0.750	0.666667	1.000000	0.333333	0.500000
21	RF Undersampling	0.750	0.666667	1.000000	0.333333	0.500000
26	NB Undersampling	0.625	0.566667	0.500000	0.333333	0.400000

Результати показали, що найкращі оцінки моделювання є на основі DecisionTree підходом ADASYN та Random Forest підходом Oversampling, що демонструє 100% результат визначення шахрайського сайту.

Список використаних джерел:

1. S. Makki, Z. Assaghir, Y. Taher, R. Haque, M. Hacid and H. Zeineddine, "An Experimental Study With Imbalanced Classification Approaches for Credit Card Fraud Detection," in IEEE Access, vol. 7, pp. 93010-93022, 2019, doi: 10.1109/ACCESS.2019.2927266.

1. Q. Kang, L. Shi, M. Zhou, X. Wang, Q. Wu and Z. Wei, "A Distance-Based Weighted Undersampling Scheme for Support Vector Machines and its Application to Imbalanced Classification," in IEEE Transactions on Neural Networks and Learning Systems, vol. 29, no. 9, pp. 4152-4165, Sept. 2018, doi: 10.1109/TNNLS.2017.2755595.
2. J. Mathew, C. K. Pang, M. Luo and W. H. Leong, "Classification of Imbalanced Data by Oversampling in Kernel Space of Support Vector Machines," in IEEE Transactions on Neural Networks and Learning Systems, vol. 29, no. 9, pp. 4065-4076, Sept. 2018, doi: 10.1109/TNNLS.2017.2751612.
3. L. Demidova and I. Klyueva, "SVM classification: Optimization with the SMOTE algorithm for the class imbalance problem," 2017 6th Mediterranean Conference on Embedded Computing (MECO), 2017, pp. 1-4, doi: 10.1109/MECO.2017.7977136.
4. W. Dan and L. Yian, "Denoise-Based Over-Sampling for Imbalanced Data Classification," 2020 19th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES), 2020, pp. 1-4, doi: 10.1109/DCABES50732.2020.00078.