Вінницький національний технічний університет

Факультет менеджменту та інформаційної безпеки

Кафедра менеджменту та безпеки інформаційних систем

МАГІСТЕРСЬКА КВАЛІФІКАЦІЙНА РОБОТА

на тему:

Вдосконалений метод інтелектуального аналізу даних з підвищеною достовірністю на основі математичного апарату нейронних мереж

Виконав: ст. 2-го курсу, групи УБ-21м
спеціальності 125– Кібербезпека
Освітня програма – Управління інформаційною безпекою
Панасюк Б.В.
Керівник: д.т.н., професор каф. МБІС
Яремчук Ю.Є.
«15» грудня 2022 р.
Опонент: к.т.н., доц., доцент каф. ОТ
Савицька Л.А.
«15» грудня 2022 р.

Допущено до захисту
Голова секції УБ кафедри МБІС

Юрій ЯРЕМЧУК
“15” грудня 2022 р.

Вінниця ВНТУ - 2022 рік

Вінницький національний технічний університет
Факультет менеджменту та інформаційної безпеки
Кафедра менеджменту та безпеки інформаційних систем

Рівень вищої освіти ІІ-й (магістерський)
Галузь знань 12 – Інформаційні технології
Спеціальність 125 – Кібербезпека
Освітньо-професійна програма – Управління інформаційною безпекою

ЗАТВЕРДЖУЮ
Голова секції УБ, кафедра МБІС

_____ **Юрій ЯРЕМЧУК**
"15" вересня 2022 р.

З А В Д А Н Н Я

на магістерську кваліфікаційну роботу студенту

Панасюку Богдану Васильовичу

1. Тема роботи «Вдосконалений метод інтелектуального аналізу даних з підвищеною достовірністю на основі математичного апарату нейронних мереж»

Керівник роботи Яремчук Ю.Є. д.т.н., професор затверджені наказом вищого навчального закладу від «14» вересня 2022 року № 203

2. Строк подання студентом роботи за тиждень до захисту

3. Вихідні дані до роботи: нормативно-правова база, монографії та сучасні наукові статті по темі, Інтернет-ресурси, стандарти, існуюче ПЗ.

4. Зміст текстової частини: в першому розділі проаналізувати засоби інтелектуального аналізу даних, використання нейронних мереж для здійснення аналізу даних; в другому розділі здійснити вдосконалення методу, провести проектування розробки, розробити алгоритми програмної частини; в третьому розділі здійснити програмну реалізацію розробки та аналіз результатів; в четвертому розділі проаналізувати економічну ефективність розробленого програмного забезпечення.

5. Перелік ілюстративного матеріалу (з точним зазначенням обов`язкових креслень):

– у першому розділі наведено 10 рис., 3 табл.;

– у другому розділі наведено 8 рис., 1 табл.;

– у третьому розділі наведено 25 рис., 1 табл.;

– у четвертому розділі наведено 1 рис. та 6 табл.

6. Консультанти розділів магістерської кваліфікаційної роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
1	д.т.н., проф. каф. МБІС Яремчук Ю.Є.		
2	д.т.н., проф. каф. МБІС Яремчук Ю.Є		
3	д.т.н., проф. каф. МБІС Яремчук Ю.Є		
4	зав. каф. ЕПВМ, к.т.н. Лесько О.Й.		

7.Дата видачі завдання 15 вересня 2022 р.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів МКР	Строк виконання етапів МКР	Примітка
1.	Визначення напрямку МКР, формулювання теми	15.09.2022	30.09.2022
2.	Аналіз предметної області обраної теми	01.10.2022	15.10.2022
3.	Розробка алгоритму роботи	16.10.2022	31.10.2022
4.	Написання МКР на основі розробленої теми	01.11.2022	15.11.2022
5.	Розробка економічної частини	15.11.2022	23.11.2022
6.	Попередній захист МКР	24.11.2022	25.11.2022
7.	Виправлення, уточнення, коригування роботи	26.11.2021	15.12.2022
8.	Захист МКР	19.12.2022	21.12.2022

Студент _____ Панасюк Б.В.

Керівник МКР _____ Яремчук Ю.Є.

АНОТАЦІЯ

УДК 004.056.5

Панасюк Б.В. Вдосконалений метод інтелектуального аналізу даних з підвищеною достовірністю на основі математичного апарату нейронних мереж. Магістерська кваліфікаційна робота зі спеціальності 125 – «Кібербезпека», освітня програма «Управління інформаційною безпекою». Вінниця: ВНТУ, 2022. 104 с.

На укр. мові. Бібліогр.: 33 назв; рис.: 44; табл. 11.

У магістерській кваліфікаційній роботі здійснено вдосконалення методу для виявлення загроз веб-ресурсу на основі інтелектуального аналізу даних з підвищеною достовірністю з використанням математичного апарату нейронних мереж та методу кластеризації.

У першому розділі роботи проаналізовано можливість дослідження та виявлення загроз веб-ресурсу на основі data mining, досліджено застосування нейронних мереж для виявлення загроз веб-ресурсу та розглянуто аналоги програмних засобів data mining, що працюють на основі наведених методів.

У розділі розробки здійснено вдосконалення методу, зокрема описано алгоритм навчання нейромережі, механізм вдосконалення за рахунок аналізу відстаней в кластері та міжкластерних відстаней, розроблено алгоритм роботи програмного засобу на основі вдосконаленого методу.

В третьому розділі роботи здійснено практичну розробку додатку, зокрема описано особливості здійсненої програмної реалізації, інструкцію користувача та проведено результат аналізів інтелектуального аналізу даних, що показали підвищення достовірності на основі вдосконаленого методу на 10%.

В четвертому розділі було виконано оцінювання комерційного потенціалу розробки програмного засобу, яке показало економічну доцільність її впровадження.

Ключові слова: веб-ресурс, загрози, data mining, нейроні мережі, кластеризація.

ABSTRACT

Bohdan Panasiuk. An improved method of intelligent data analysis with increased reliability based on the mathematical apparatus of neural networks. Master's thesis in specialty 125 – «Cyber Security», Education Program «Information Security Management». Vinnitsa: VNTU, 2022. – 104 p.

In Ukrainian language. Bibliographer: 33 titles; fig.: 44; tabl. 11.

In the master's qualification work, the method for detecting web resource threats was improved on the basis of intelligent data analysis with increased reliability using the mathematical apparatus of neural networks and the clustering method.

In the first part of the work, the possibility of researching and detecting web resource threats based on data mining is analyzed, the use of neural networks to detect web resource threats is investigated, and analogues of data mining software tools operating on the basis of the above methods are considered.

In the development section, the method of intelligent data analysis has been improved, in particular, the learning algorithm of the neural network, the mechanism of improvement due to the analysis of distances in the cluster and intercluster distances, and the algorithm of the software based on the improved method have been developed.

In the third section of the work, the practical development of the application was carried out, in particular, the features of the software implementation, the user manual were described, and the results of the intelligent data analysis were conducted, which showed a 10% increase in reliability based on the improved method.

In the fourth chapter, an assessment of the commercial potential of software development was performed, which showed the economic feasibility of its implementation.

Keywords: web resource, threats, data mining, neural networks, clustering.

ЗМІСТ

ВСТУП

Актуальність. Аналіз кіберзагроз – це сукупність аналітичних даних, орієнтованих на забезпечення інформаційної безпеки. Інформація, що ретельно відбирається, дозволяє приймати обґрунтовані рішення про способи захисту користувачів і бізнесу від кіберзагроз. Аналіз загроз продовжує розвиватися та дозволяє отримати інформацію про методи, які використовуються зловмисниками. Це надає можливість створити безпечніші середовища, в яких можна ідентифікувати зловмисників. Дедалі більше організацій державного та приватного секторів починають використовувати аналіз кіберзагроз, зокрема на основі data mining, що дозволяє обробляти масивні бази даних та «витягувати» з них корисну інформацію [1].

Завданням інтелектуального аналізу даних є виявлення правил та закономірностей у наборах даних. Тривалий час основним інструментом інтелектуального аналізу даних була традиційна математична статистика, проте з часом набуло поширення застосування нейронних мереж, враховуючи їх переваги, такі як висока допустимість до зашумлених даних і низький відсоток помилок, постійне вдосконалення та оптимізація алгоритмів навчання мереж, вилучення правил, спрощення мереж, що роблять нейронні мережі все більш і більш перспективним напрямом у data mining [2].

Нейроні мережі широко застосовуються у автоматизації процесів розпізнавання образів, прогнозуванні, адаптивному управлінні, створенні експертних систем, організації асоціативної пам'яті, синтезі та ідентифікації електронних ланцюгів та систем.

Таким чином, можна вважати, що використання нейронних мереж у технології інтелектуального аналізу даних для виявлення загроз є прогресуючим напрямком, який безперервно розвивається, шляхом усунення недоліків, а тому його подальше дослідження є актуальним для даної роботи.

В даній роботі увага приділяється дослідженню нейронних мереж та застосування їх математичного апарату на основі вдосконаленого алгоритму

кластеризації з метою забезпечення захищеності даних, а саме виявлення наявності загроз для веб-ресурсу в результаті аналізу даних, що містять інформацію про взаємодію користувачів з ресурсом.

Мета і задачі дослідження. Метою роботи є розробка вдосконаленого методу інтелектуального аналізу даних з підвищеною достовірністю на основі математичного апарату нейронних мереж для виявлення загроз веб-ресурсу.

Задачами дослідження є:

– дослідити поняття інтелектуального аналізу як методу обробки даних;

– здійснити аналіз методів data mining та огляд існуючих аналогів їх реалізацій;

– проаналізувати можливості застосування нейронних мереж для здійснення інтелектуального аналізу даних та виявлення загроз веб-ресурсу;

– розробити модуль аналізу даних на основі нейромережі та здійснити вдосконалення даної моделі за допомогою використання кластерів;

– розробити алгоритм роботи програми на основі вдосконаленого методу;

– обґрунтувати вибір середовища розробки та мови програмування;

– здійснити програмну реалізацію додатку на основі вдосконаленого методу;

– провести тестування розробки та проаналізувати отримані результати;

– економічно обґрунтувати доцільність здійсненої розробки.

Об’єкт дослідження – метод інтелектуального аналізу даних на основі математичного апарату нейронних мереж.

Предмет дослідження – процес вдосконалення методу інтелектуального аналізу даних на основі математичного апарату нейронних мереж.

Наукова новизна – вдосконалення методу інтелектуального аналізу даних з підвищеною достовірністю на основі математичного апарату нейронних мереж та кластеризації для виявлення загроз веб-ресурсу.

Практична цінність. Розроблено програмний продукт, який реалізує вдосконалений метод інтелектуального аналізу даних з підвищеною достовірністю на основі математичного апарату нейронних мереж.

1 АНАЛІЗ ЗАСТОСУВАННЯ МЕТОДІВ DATA MINING ДЛЯ ВИЯВЛЕННЯ ЗАГРОЗ ВЕБ-РЕСУРСУ

Аналіз загроз – це знання, отримані на основі фактичних даних про існуючу або виникаючу загрозу чи небезпеку для веб-ресурсів, які можуть використовуватися для прийняття рішень про реагування суб'єкта на цю загрозу чи небезпеку. Для здійснення аналізу загроз на сьогодні використовуються різноманітні методи, що дають оцінку наявності загроз, стану системи, заходів безпеки і т.д.

В даному розділі здійснимо аналіз літератури у галузі застосування інтелектуального аналізу даних для виявлення загроз безпеці веб-ресурсів, існуючих методів та аналогів на їх основі. Дослідження даного напрямку спрямоване на здійснення вдосконалення методу інтелектуального аналізу даних та застосування його для виявлення наявності загроз для веб-ресурсу.

1.1 Аналіз загроз безпеки веб-ресурсу на основі інтелектуального аналізу

При розробці веб-додатків розробники більше уваги приділяють, зазвичай, забезпеченню необхідної функціональності. Натомість, питанням безпеки та якості програмного коду приділяється недостатньо уваги. В результаті переважна більшість веб-застосунків містить уразливості різного ступеня критичності та, відповідно, спричиняє різного типу загрози (рис. 1.1).

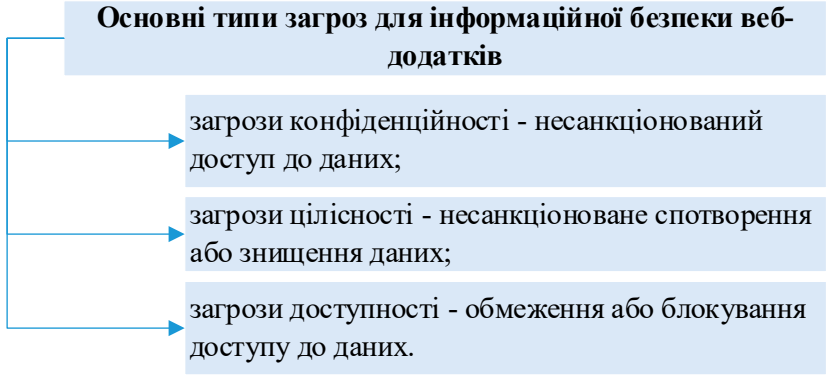


Рисунок 1.1 – Типи загроз для веб-застосунків(за [3])

Крім того, існує безліч різних атак на веб-додатки, і з кожним роком їх стає дедалі більше (рис. 1.2).

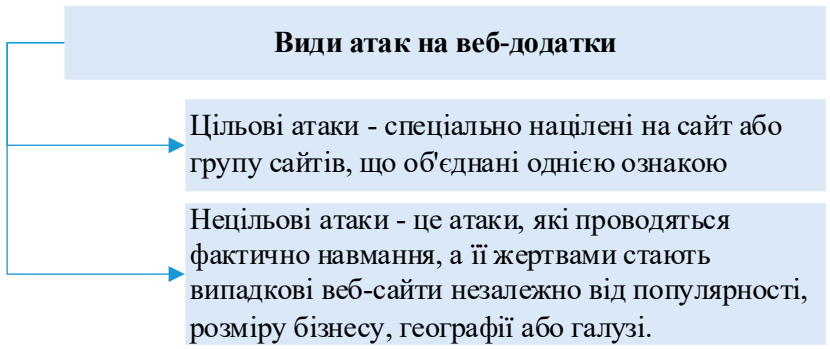


Рисунок 1.2 – Види атак на веб-ресурси (за [4 – 5])

У зв’язку з цим на сьогодні існують різні методи автоматичного виявлення вразливостей додатків, що умовно можна поділити на дві основні групи [6 – 7]:

– методи, що аналізують роботу веб-додатку без звернення до вихідних кодів веб-додатку.

До таких методів належать: метод отримання ідентифікуючої інформації про веб-додаток та виявлення його вразливостей та метод тестування на проникнення).

– методи, що аналізують вихідні коди веб-додатку та конфігураційні налаштування.

До таких методів належать: метод статичного аналізу вихідних кодів веб-додатку та метод динамічного аналізу вихідних кодів веб-додатку.

В даній роботі дослідимо метод, що аналізує роботу веб-ресурсу без звернення до вихідних кодів, оскільки важливим механізмом захисту веб-сайтів від загроз безпеки є виявлення аномальної або нетипової активності, яка може бути наслідком дій зловмисників, спрямованих на порушення цілісності, конфіденційності або доступності інформаційних ресурсів, що у сукупності складають основу поняття інформаційної безпеки.

Метод захисту, що розглядається далі, заснований на застосуванні інтелектуального аналізу даних з використанням нейронних мереж для виявлення на веб-ресурсах нетипової активності з метою підвищення точності

виявлення за умови великої кількості звернень до ресурсів сайту.

Інтелектуальний аналіз даних (ІАД) є одним із прогресивних способів аналізу великих обсягів даних. Це процес виявлення та подальшого застосування знань або раніше невідомої інформації з наявних наборів, основними завданнями якого є класифікація; асоціація; кластеризація; прогнозування; послідовність [6].

У сфері комп'ютерної безпеки методи ІАД тісно пов'язані зі створенням перспективних систем захисту (СЗІ).

Саме методологія ІАД допомагає реалізувати в СЗІ еволюційні властивості адаптації, самоорганізації, навчання, можливості успадкування та подання досвіду експертів інформаційної безпеки у вигляді доступної для аналізу системи нечітких правил.

Застосування методів ІАД для визначення загроз веб-ресурсам є доцільним вибором, оскільки більшість комп'ютерних систем записують події, пов'язані з цією системою, файли журналів [8].

У правильно налаштованому веб-сервері будь-які дії, які є порушенням безпеки, також призводять до появи відповідного запису в одному або кількох файлах журналу. Аналогічним чином можуть функціонувати засоби аудиту безпеки сайту, додаючи до певної таблиці бази даних необхідні відомості про звернення до ресурсів сайту.

Фахівець із захисту інформації, отримуючи результати прогнозування виникнення загрози або вразливості, може оцінити ступінь небезпеки для інформаційних ресурсів, що захищаються, коректність застосовуваної моделі загроз інформаційній безпеці та вжити заходів щодо нейтралізації вразливостей.

У зв’язку з цим, далі в розділі розглянемо особливості інтелектуального аналізу даних та його застосування для вибору оптимального підходу для вирішення поставленої задачі.

Найперше, варто зауважити, що на основі інтелектуального аналізу даних опрацьовуються «сирі» дані, для процесу оброки яких ставляться такі вимоги (рис.1.3) [8].

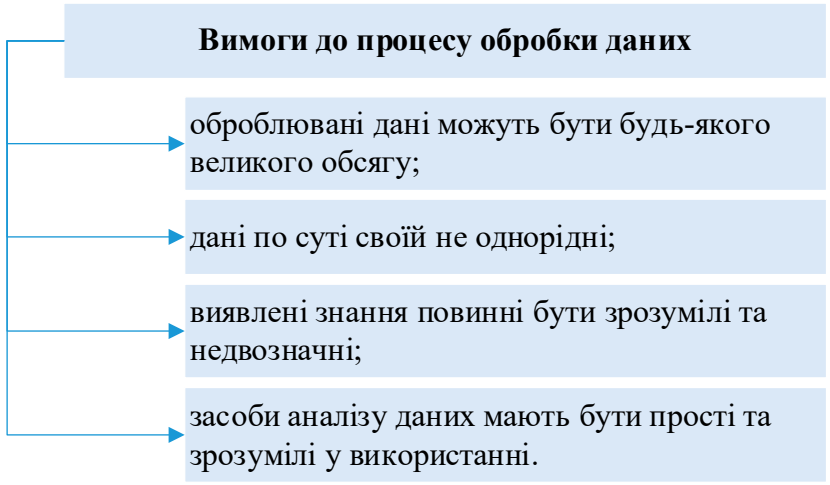


Рисунок 1.3 – Вимоги до процесу обробки сирих даних на корисні знання

Методи Data Mining спрямовані на ті завдання, яких сьогодні найбільше потребують – аналіз великих обсягів даних та пошук прихованих знань усередині цих обсягів. Рівні знань та застосовні до них аналітичні інструменти наведено у таблиці 1.1

Таблиця 1.1 – Рівні застосування аналітичних інструментів [8]

Рівень знань, видобуваних з даних	Аналітичні інструменти
Поверховий	Мова простих запитів
Недеталізований	Оперативна аналітична обробка (OLAP)
Прихований	Інтелектуальний аналіз даних (Data Mining)

Інтелектуальний аналіз даних можна представити у вигляді методу підтримки прийняття рішень, що базується на аналізі виявлених залежностей між наявними даними [8]. Тобто, враховуючи загальність такого формулювання типовий аналіз звітів, що побудований за базою даних, також можемо розглядати як один із різновидів інтелектуального аналізу. Щоб перейти до розгляду більш сучасних технологій, розглянемо, як можна автоматизувати пошук залежностей між даними.

Метою інтелектуального аналізу даних є виявлення неявних закономірностей у наборах даних. І хоча існуючі технології дозволяли, наприклад, швидко знайти у базі даних потрібну інформацію, цього у багатьох випадках було недостатньо. Виникла потреба пошуку взаємозв'язків між

окремими подіями серед великих обсягів даних, для чого знадобилися методи математичної статистики, теорії баз даних, теорії штучного інтелекту та інших областей. Враховуючи різноманітність форм подання даних, використовуваних алгоритмів та сфер застосування, ІАД може проводитись за допомогою програмних продуктів певних класів (рис. 1.4).

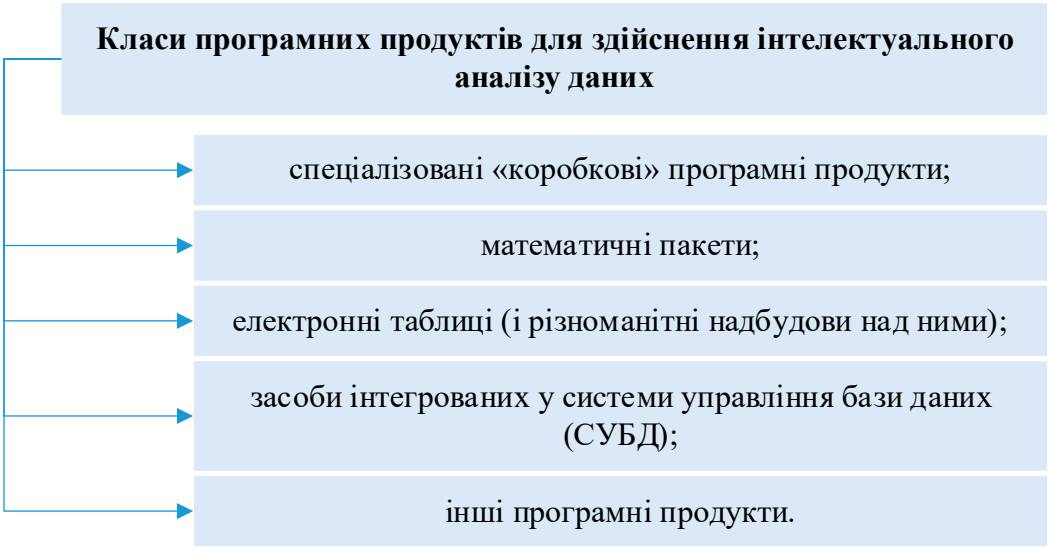


Рисунок 1.4 – Класи програмних продуктів для здійснення інтелектуального аналізу даних (за [9])

У ході проведення інтелектуального аналізу даних проводиться дослідження багатьох об'єктів (або варіантів). У більшості випадків його можна подати у вигляді таблиці, кожен рядок якої відповідає одному з варіантів, а в стовпцях містяться значення параметрів, що його характеризують [10]:

$$p\ align = "justify"$$

Залежна змінна – параметр, значення якого розглядаємо як залежне від інших параметрів (незалежних змінних). Саме дану залежність потрібно прорахувати засобами інтелектуального аналізу даних.

У зв’язку з цим далі у розділі здійснимо аналіз методів та задач Data Mining для вибору підходу до вирішення поставлених задач роботи, оскільки при використанні методів та технологій інтелектуального аналізу даних з'являється дійсна можливість відкривати неочевидні закономірності між накопиченими даними та використовувати отримані знання в системах прийняття рішень.

1.2 Аналіз існуючих методів та задач Data Mining

Головна перевага методів Data Mining полягає в тому, що вони поєднують у собі як математичний інструментарій, так і останні розуміння в галузі IT [10]. У методах інтелектуального аналізу даних комплексно використовуються формалізовані та неформальні способи та техніки аналітики, різні способи аналізу даних [10]. Таким чином, методи інтелектуального аналізу можна поділити на технологічні, статистичні та кібернетичні (рис. 1.5).

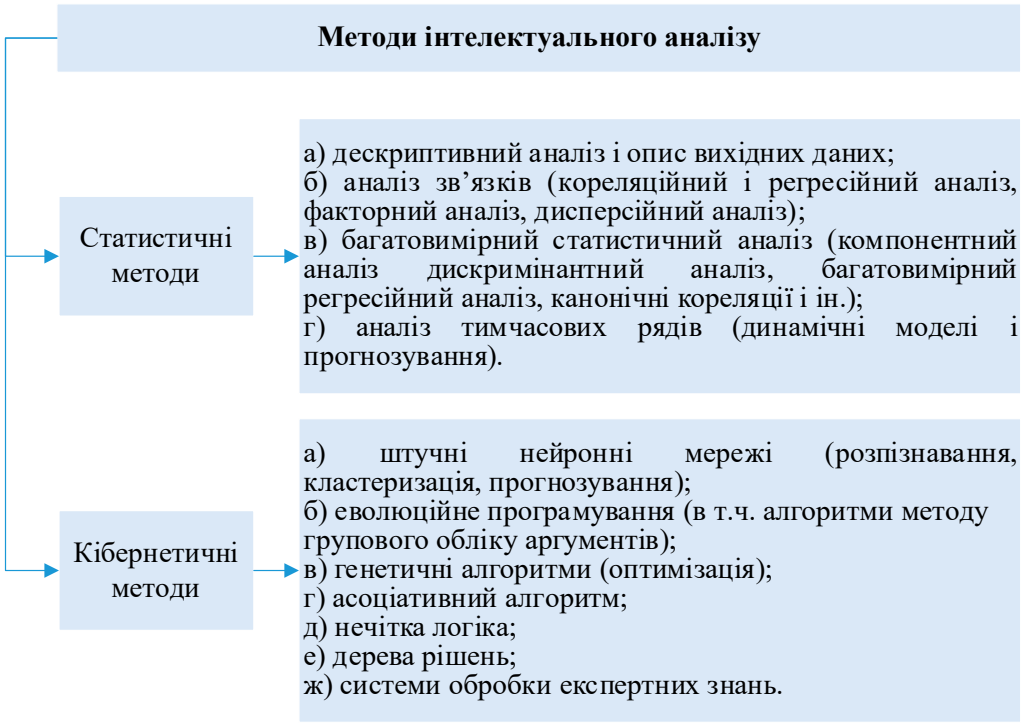


Рисунок 1.5 – Методи інтелектуального аналізу [11]

Можемо підсумувати, що у статистичних методах використовується усереднений досвід за даними, що накопичилися у базі даних за тривалий період. При використанні статистичних методів попередньо аналізується природа статистичних даних, виявляються зв'язки та закономірності, здійснюється багатовимірний статистичний аналіз, будуються динамічні моделі та прогноз на основі часових рядів.

У кібернетичних методах використовуються основи комп'ютерної математики та технології штучного інтелекту. Саме у розрізі вивчення даних методів зародилася популярна на сьогодні технологія нейромережевої обробки

та аналізу даних. Таким чином, можна прийти до твердження, що зараз більшість технологічно реалізованого аналітичного інструментарію базується на принципах, сформульованих у рамках кібернетичного підходу.

Для вирішення задачі виявлення загроз для веб-ресурсу далі в роботі розглянемо завдання інтелектуального аналізу даних та визначимо оптимальне рішення для застосування.

Основні завдання інтелектуального аналізу даних наведено на рис. 1.6).

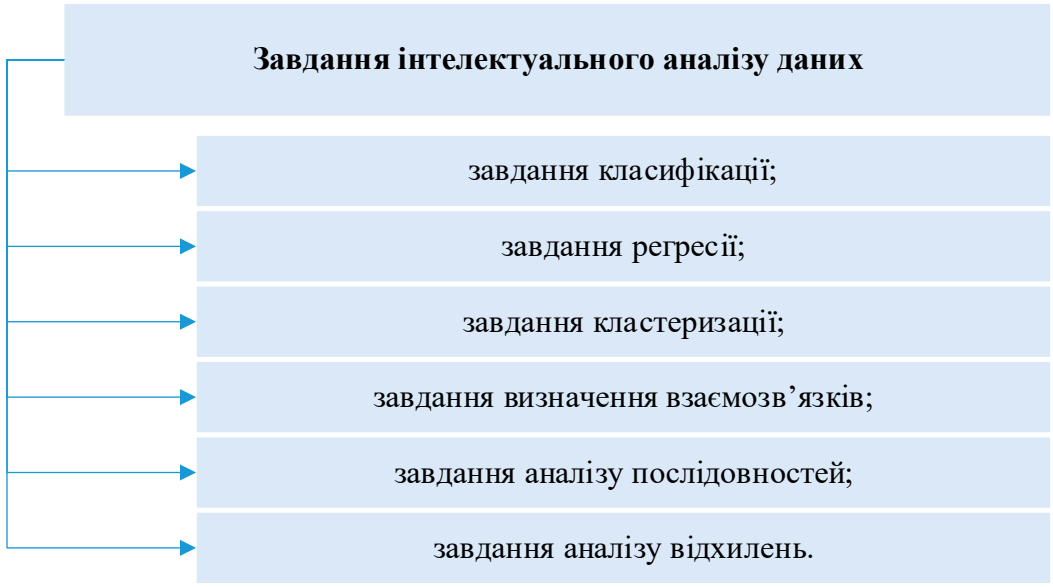


Рисунок 1.6 – Основні завдання інтелектуального аналізу даних (за [11])

Класифікація [11] передбачає, що кожному з варіантів відповідає категорія чи клас, якому він належить. Слід зауважити, що для вирішення даного завдання необхідно, щоб сукупність класів була відома завчасно.

Регресія [11] подібна до класифікації, проте в ході її виконання здійснюється пошук шаблонів для визначення числового значення. Тобто, прогнозований параметр є числом безперервного діапазону.

Прогнозування [11] є окремим завданням аналізу даних та передбачає прогнозування нових значень враховуючи наявні значення числової послідовності. При цьому можуть враховуватися інші фактори.

Кластеризація [11] має на меті розподіл сукупності об'єктів на групи (кластери) за певними подібними параметрами. Кластеризація відноситься до завдань навчання без вчителя, так званих «некерованих» завдань.

Завдання визначення взаємозв'язків [11], або завдання пошуку асоціативних правил, полягає у визначенні найпоширеніших наборів об'єктів серед безлічі подібних наборів.

Аналіз послідовностей [11] або сиквенціальний аналіз одними авторами розглядається як варіант попередньої задачі, іншими – виділяється окремо. Має на меті виявлення закономірностей в послідовностях подій.

Аналіз відхилень [11] надає можливість виявити відхилення, яке може попереджувати про якусь незвичайну подію (нетиповий результат експерименту, шахрайська операція по банківській карті) або, помилку введення даних певним працівником.

У таблиці 1.2 наведено приклади завдань інтелектуального аналізу даних із різних галузей.

Таблиця 1.2 – Приклади застосування інтелектуального аналізу даних (за [11])

	Інформаційні технології	Торгівля	Фінансова сфера
Класифікація			Оцінка кредитоспроможності
Регресія			Оцінка допустимого кредитного ліміту
Прогнозування		Прогнозування продажів	Прогнозування цін акцій
Кластеризація	Виявлення загроз в ІС	Сегментація клієнтів	Сегментація клієнтів
Визначення взаємозв’язків		Аналіз споживчого кошика	
Аналіз послідовностей	Аналіз переходів по сторінкам web-сайту		
Аналіз відхилень	Виявлення вторгнень в ІС		Виявлення шахрайства з банк. картками

Здійснивши аналіз завдань Data Mining, оптимальним є вибір підходу, що ґрунтується на завданні кластеризації, яка розрахована саме для виявлення

наявності загроз в інформаційній системі.

Провівши кластеризацію даних, наступним етапом є проведення аналізу цих даних та прийняття відповідного рішення. Для цього скористаємось засобом кібернетичного аналізу – нейронними мережами, що мають ряд переваг, зокрема високу допустимість до зашумлених даних і низький відсоток помилок, постійне вдосконалення та оптимізацію алгоритмів навчання мереж, вилучення правил, спрощення мереж. Крім того, велика кількість вихідних даних, а також складність їх внутрішньої структури ускладнюють застосування для аналізу відомих статистичних методів, у зв'язку з чим перспективним є метод, що ґрунтується на штучних нейронних мережах. Далі в роботі розглянемо більш детальніше саме метод нейронних мереж у поєднанні з одним із найбільш затребуваним методом кластеризації.

1.3 Аналіз застосування нейромереж для виявлення загроз веб-ресурсу

Штучні нейронні мережі – це математичні моделі, їх програмні та апаратні реалізації, які будуються за принципами організації та функціонування біологічних нейронних мереж [12].

В якості інструменту аналізу даних нейронні були представлені в 1943 році в роботі Уоррена Мак-Каллока і Вальтера Піттса, що містила модель штучного нейрона [13]. Дані автори висунули гіпотезу математичного нейрона, який змоделював нейрон мозку людини. Такий нейрон аналогічно має кілька входів і один вихід, а вихідному сигналу нейрона може надаватись два значення – нуль чи один. Штучні нейронні мережі засновані на таких принципах роботи мозку, як знання і процесор, які не розділені, а перебувають у рівномірно розподіленому стані, неявно існуючи як синаптичні зв'язки. Подібні знання спочатку відсутні та набуваються у процесі навчання.

Нейронні мережі – це сукупність математичних методів, які використовуються для обробки даних, прогнозування та кластеризації [13].

Модель нейронної мережі можна розділити на три типи (рис. 1.5) [14]:

1) мережі прямого поширення (зворотне поширення): застосовується в

таких галузях, як прогнозування та розпізнавання образів;

2) мережі зі зворотним зв'язком: в основному використовується для оптимізації обчислень та асоціативної пам'яті;

3) мережі, що самоорганізуються: включають моделі теорії адаптивного резонансу (АРТ) і моделі Кохонена, які в основному використовуються для кластерного аналізу.

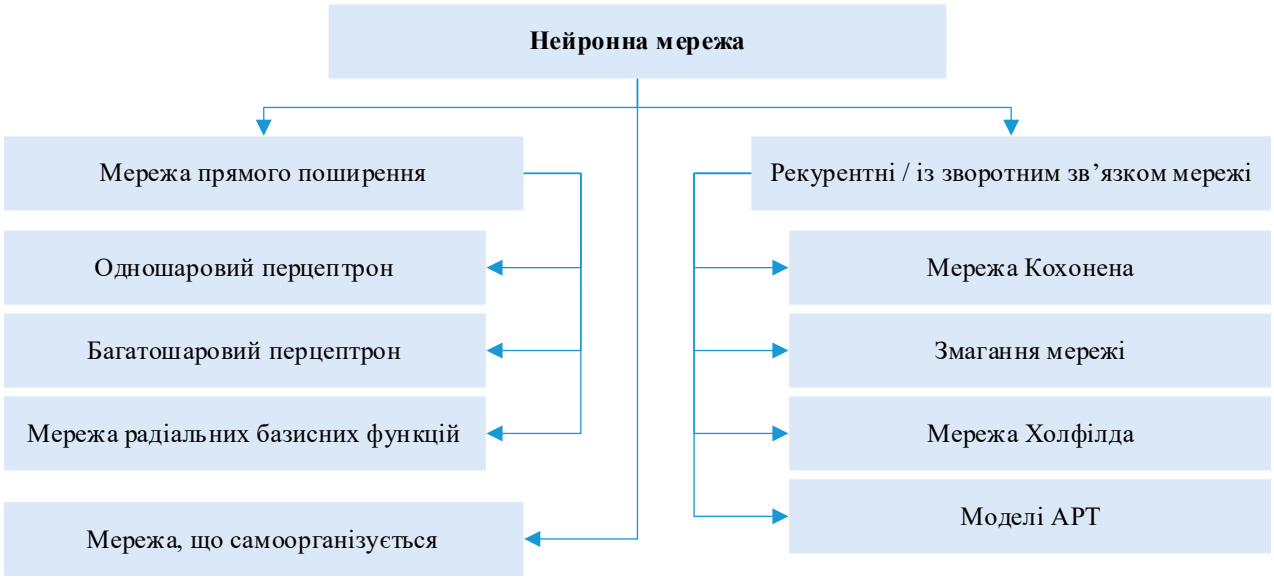


Рисунок 1.7 – Класифікація нейронних мереж [14]

В даний час для аналізу даних використовують нейронні мережі прямого поширення. Штучні нейронні мережі є сферою науки, що активно розвивається, але досі деякі теорії не були повністю сформовані, такі як проблеми збіжності, стійкості, локального мінімуму і коригування параметрів. Для мережі прямого розподілу загальних проблем – навчання повільне, воно може потрапити до локального мінімуму та складно визначити параметри навчання. Через ці проблеми багато хто перейшов на метод комбінування штучних нейронних мереж з генетичними алгоритмами і досяг найкращих результатів [15].

Однією з основних переваг нейронних мереж є можливість апроксимувати будь-яку безперервну функцію, що дозволяє досліднику заздалегідь не приймати жодних гіпотез про модель.

Розглянемо для прикладу теорію навчання нейронних мереж у випадку навчання з учителем. Модель навчання з учителем складається з трьох

взаємопов'язаних компонентів, які в математичних термінах можна описати наступним чином [16]:

1. середовище: характеризується розподілом ймовірностей $P_X(x)$ із випадками x, що випадково випадково і незалежно з'являються;

2. вчитель: генерує бажаний відгук d для кожного з вхідних векторів x, отриманих із зовнішнього середовища, відповідно до умовної функції розподілу $P_X(d|x)$. Ні характеристика середовища $P_X(x)$, ні правило класифікації $P_X(d|x)$ невідомі. Однак відомо, що обидві функції існують, тобто існує спільний розподіл ймовірностей:

$$P_X(d,x) = P(x) \cdot P_X(d|x)$$

Бажаний відгук d та вхідний вектор x пов'язані наступним співвідношенням:

$$d = f(x,v)$$

де $v -$ шум, тобто спочатку передбачається зашумленість даних вчителя.

3. машина, що навчається: нейронна мережа здатна реалізувати безліч функцій відображення вхід-вихід, що описуються співвідношенням

$$y = F(x,w)$$

де $y -$ фактичний відгук, згенерований машиною, що навчається, у відповідь на вхідний сигнал x; $w -$ набір вільних параметрів (синаптичних ваг), вибраних із простору параметрів W.

Завдання навчання з учителем полягає у виборі конкретної функції $F(x,w)$, яка оптимально (у певному статистичному сенсі) апроксимує очікуваний відгук d. Вибір, у свою чергу, ґрунтується на безлічі N незалежних, рівномірно розподілених прикладів навчання:

$$T = \{(d_i, x_i)\}_{i=1}^{N}$$

Кожна пара вибирається машиною, що навчається, з множини T з деякою узагальненою функцією розподілу ймовірності $P_{X,D}(d|x)$, яка, як і інші функції розподілу, фіксована, але невідома.

Розглянемо завдання навчання з учителем як задачу апроксимації, яка полягає у знаходженні функції $F(x,w)$, яка найкраще наближає бажану функцію

$f\ (x)$.

Нехай $L(d, F\ (x, w))$ – міра втрат між бажаним відгуком d, що відповідає вхідному вектору x, і відгуком $F(x, w)$, згенерованим машиною, що навчається.

Тоді очікувана величина втрат визначається функціоналом середнього ризику:

$$R(w)\ =\ \int\ L(d, F(x, w))dP_{X,D}(d|x)$$

Завдання навчання розпізнаванню образів є окремим випадком завдання про мінімізацію середнього ризику: потрібно знайти мінімум по α функціоналу

$$R(\alpha) = \int Q(z, \alpha)dP(z)$$

якщо невідома функція розподілу $P(z)$, але дана випадкова вибірка $z_1, \ldots,\ z_l$.

Особливість полягає в тому, що на функцію $Q(z, \alpha)$ накладено обмеження:

– вектор z задається $n\ +\ 1$ такими координатами: координатою ω та координатами $x_1, \ldots,\ x_n$;

– функцію втрат $Q(z, \alpha)$ задано у вигляді

$$\left(\omega - F\ (x, \alpha)\right)^2,$$

де $F\ (x, \alpha)$ – характеристична функція множин.

Традиційно виділяють три способи вирішення задачі мінімізації функціоналу середнього ризику (рис.1.6).

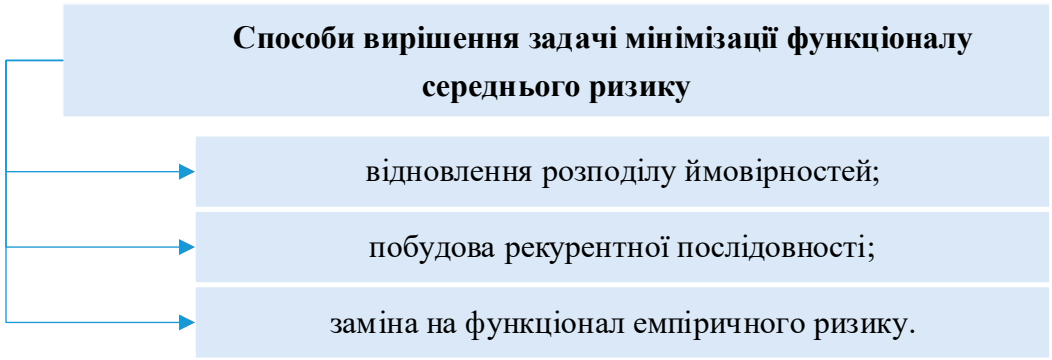


Рисунок 1.8 – Способи вирішення задачі мінімізації функціоналу середнього ризику (за [17])

Кожен із цих способів має свої переваги та недоліки, теорію навчання розпізнавання образів надалі можна застосовувати у всіх трьох напрямках, оскільки вони активно розвиваються.

Побудова класифікатора на основі нейронної мережі містить низку кроків. Розглянемо їх детальніше.

Покровоко даний алгоритм можна описати наступним чином:

Крок 1. Підготовка даних.

1.1 Скласти базу даних із прикладів, характерних для даної задачі

1.2 Розбити всю сукупність даних на дві множини: навчальна та тестова (можливе розбиття на 3 множини: навчальна, тестова та валідаційна).

Крок 2. Передобробка даних.

2.1 Провести вибір ознак, значимих з погляду завдання класифікації.

2.2 Виконати трансформацію та при необхідності очищення даних (нормалізацію, виключення дублікатів та протиріч, придушення викидів тощо).

В результаті бажано отримати лінійно розділений за класами простір безлічі прикладів.

2.3 Вибрати систему кодування вихідних значень (класичне кодування, «2 на 2» - кодування тощо).

Крок 3. Конструювання, навчання та оцінка якості мережі.

3.1 Вибрати топологію мережі: кількість шарів, число нейронів у шарах тощо.

3.2 Вибрати активаційну функцію нейронів.

3.3 Вибрати алгоритм навчання мережі.

3.4 Оцінити якість роботи мережі на основі валідаційної множини, або іншого критерію, оптимізувати архітектуру.

3.5 Зупинитись на варіанті мережі, який забезпечує найкращу здатність до узагальнення та оцінити якість роботи з тестової множини.

Крок 4. Використання та діагностика.

4.1 З'ясувати ступінь впливу різних факторів на рішення (евристичний підхід).

4.2 Переконатись, що мережа забезпечує необхідну точність класифікації (число неправильно розпізнаних прикладів мало).

4.3 При необхідності повернутися на етап 2, змінивши спосіб подання прикладів або змінивши базу даних.

4.4 Практично використовувати мережу для вирішення задачі.

Схематично даний алгоритм наведемо на рис. 1.9.

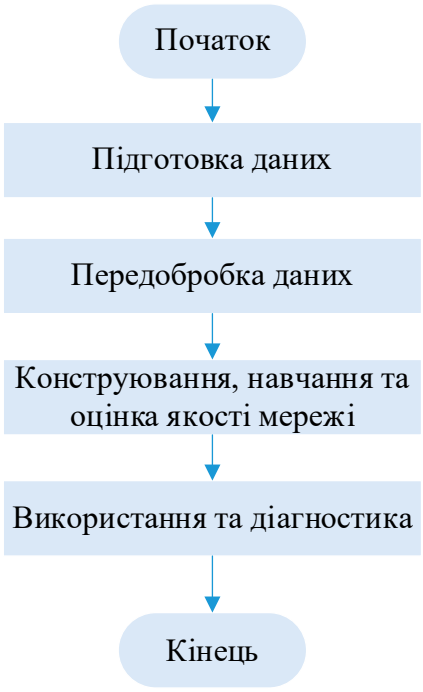


Рисунок 1.9 – Типовий алгоритм класифікатора для нейромережі (за [17])

Таким чином, основною перевагою застосування нейронних мереж є можливість вирішувати різноманітні неформалізовані завдання. При цьому можна просто моделювати різні ситуації, подаючи на вхід мережі різні дані і оцінюючи результат, що видається мережею.

1.4 Аналіз аналогів програмних засобів Data Mining

Ринок програмного забезпечення засобів інтелектуального аналізу даних представлений безліччю варіантів. З кожним роком їхня кількість лише зростає.

Найбільш популярні на ринку програмні засоби для аналізу даних наведені на рис. 1.10.

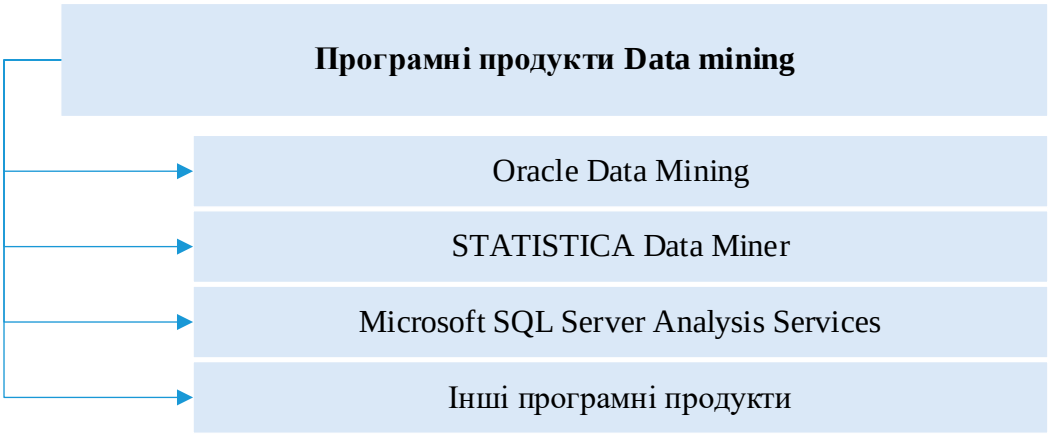


Рисунок 1.10 – Програмні засоби для здійснення інтелектуального аналізу даних

Розглянемо можливості найпопулярніших на сьогоднішній день продуктів.

SAS Enterprise Miner [18]. Програма американської компанії включає широкий інструментарій для інтелектуального аналізу даних, включаючи такі поширені на сьогоднішній день методи інтелектуального аналізу даних, як:

– дерева рішень;

– нейронні мережі;

– регресійний аналіз;

– кластеризацію;

– пошук асоціативних правил;

– секвенсіальний аналіз.

STATISTICA Data Miner [19]. Продукт компанії СтатСофт. За роки свого існування платформа STATISTICA обросла безліччю різноманітних модулів, що реалізують різні методи та технології аналітики та статистики. Важко не загубитися у такому розмаїтті.

Програмний продукт включає роботу з нейронними мережами, побудову та аналіз різноманітних моделей, регресійний аналіз, дерева рішень, методи кластеризації, класифікації, асоціації, широкий набір засобів візуалізації. Основні компоненти програми забезпечують роботу з такими завданнями:

– класифікація;

– моделювання;

– предиктивний аналіз;

– нейронні мережі;

– пошук асоціативних правил;

– побудова різноманітних дерев рішень.

Oracle Data Mining [20]. Oracle є найбільш популярним на ринку послуг та технологій інтелектуального аналізу даних. У Oracle Data Mining реалізовані такі технології, як:

– класифікація;

– кластеризація;

– пошук асоціацій;

– виділення ознак.

KXEN Analytic Framework [21]. KXEN Analytic Framework є платформою для різних модулів, що дозволяють проводити описовий та предиктивний аналіз. Продукт використовує регресійні алгоритми, дозволяє виявити природні кластери у наборі даних, дозволяє проводити бінарну класифікацію, дозволяє проводити предиктивний аналіз.

Microsoft SQL Server Analysis Services [22]. Продукт компанії Microsoft, як і раніше перелічені продукти, включає основні технології для інтелектуального аналізу баз даних на своїй платформі.

Програмні продукти SPSS [23]. Дозволяють будувати різноманітні дерева рішень, проводити кластерний аналіз, знаходити асоціативні правила у базах даних та проводити предиктивний аналіз. Містить більшість технологій, які надаються його прямими конкурентами.

Порівняльний аналіз методів Data Mining, які надаються у найбільш поширених програмних продуктах інтелектуального аналізу даних наведено у таблиці 1.3, із урахування критеріїв використання таких методів: адаптивна Байєвська мережа; аналіз тимчасових рядів; граничні методи; дерева рішень; ієрархічна кластеризація; лінійна регресія; логістична регресія; неієрархічна кластеризація; нейронні мережі; пошук асоціативних правил.

Таблиця 1.3 – Методи Data Mining у програмних продуктах

	SAS Enterprise Miner	STATISTICA	Oracle Data Mining	KXEN Analytic Framework	Microsoft SQL Server Analysis Services	IBM SPSS	Як часто вик-тсья
Адаптивна Байєвська мережа			+				1
Аналіз тимчасових рядів		+		+	+	+	4
Граничні методи		+	+	+			3
Дерева рішень	+	+			+	+	4
Ієрархічна кластеризація	+	+	+	+			4
Лінійна регресія	+	+	+	+	+	+	6
Логістична регресія	+	+	+	+	+	+	6
Неієрархічна кластеризація	+	+	+	+	+		5
Нейронні мережі	+	+			+		3
Пошук асоціативних правил	+	+	+		+		4

Таким чином, аналізуючи таблицю 1.3, можна зробити висновок, що найбільш часто використовуваними є методи регресії та кластеризації. Застосування нейронних мереж є менш вживаним, враховуючи складнощі, що виникають під час реалізації процесу аналізу.

Проте, враховуючи суттєві переваги та можливості нейронних мереж та кластеризації застосуємо їх для вирішення поставлених задач виявлення загроз веб-ресурсів у подальшій роботі.

1.5 Висновки та постановка задач

Отже, в даному розділі було проведено теоретичний аналіз галузі, в якій проводиться розробка. Метою роботи є виявлення загроз для веб-ресурсів на

основі вдосконаленого методу інтелектуального аналізу даних з підвищеною достовірністю на основі математичного апарату нейронних мереж.

У зв’язку з цим здійснено аналіз загроз безпеки веб-ресурсу, досліджено існуючі методи та задачі інтелектуального аналізу даних.

Для розробки та реалізації вдосконаленого методу виявлення загроз веб-ресурсу було обрано нейронні мережі та метод кластеризації за рахунок виявлених переваг.

В результаті проведеного аналізу теоретичного матеріалу, були поставлені наступні задачі подальшої роботи:

– розробити модуль аналізу даних на основі нейромережі та здійснити вдосконалення методу кластеризації;

– розробити алгоритм роботи програми для виявлення загроз веб-ресурсу на основі вдосконаленого методу;

– обґрунтувати вибір середовища розробки та мови програмування;

– програмно реалізувати додаток для виявлення загроз веб-ресурсу на основі вдосконаленого методу;

– провести тестування розробки з метою виявлення загроз веб-ресурсу та проаналізувати отримані результати;

– економічно обґрунтувати доцільність здійсненої розробки.

В результаті виконання поставлених задач, можливо досягти основної початково поставленої мети роботи.

2 РОЗРОБКА ВДОСКОНАЛЕНОГО МЕТОДУ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ ДЛЯ ВИЯВЛЕННЯ ЗАГРОЗ ВЕБ-РЕСУРСУ

В якості запропонованого вдосконалення методу інтелектуального аналізу даних з підвищеною достовірністю на основі математичного апарату нейронних мереж застосуємо кластерний метод [24] аналізу з використанням карт Кохонена [25] та його подальшою модифікацією з метою підвищення якості аналізу, що застосовується для визначення наявності загроз веб-ресурсу.

Вдосконалення методу полягає в тому, що після певної кількості сформованих кластерів, продовжимо аналіз неврахованих кластерів, що не містять нейронів, а тому не були враховані до основної вибірки. Аналіз проведемо за рахунок визначення відстаней за відомою формулою Евкліда [26].

Далі в роботі розглянемо більш детальніше запропонований метод підвищення достовірності проведеного інтелектуального аналізу даних, що розробляється для подальшої реалізації програмного додатку, що визначатиме наявність загроз для веб-ресурсу.

2.1 Алгоритм навчання нейромережі для виявлення загроз веб-ресурсу

Ґрунтуючись на описаних раніше принципах реалізації штучних нейронних мереж, можна створити нейронну мережу для визначення загроз веб-ресурсу. Вхідні вектори штучної нейронної мережі є рядками таблиці, отримані в результаті процесу передобробки вихідних даних. Кожен вектор виражає поведінку користувача за певний період часу та представлений у вигляді проміжних статистичних значень.

Після формування нейронної мережі її можна навчати. При цьому для навчання «нормальній» поведінці використовуються дані, що є в журналі та обраний за період, в якому не було порушень безпеки веб-ресурсу, а для навчання стану наявності загрози – генеруються на основі заданих критеріїв.

За основу реалізації кластеризації даних застосуємо нейронну мережу, що складається з єдиного шару нейронів (шар Кохонена) та без коефіцієнтів усунення (рис. 2.1) [27].

Загальну кількість вагових коефіцієнтів розрахуємо наступним чином:

$$N_w = MK,$$

де $N_w -$ кількість вагових коефіцієнтів;

$M -$ кількість вхідних змінних мережі;

$K -$ кількість нейронів шару.

Мережа організована таким чином, що загальна кількість нейронів позначає кількість кластерів, на основі яких здійснюється початковий розподіл та подальший перерозподіл навчальних зразків [27].

Кількість вхідних змінних нейронної мережі дорівнює числу ознак, які є характерними для об'єкта дослідження та з урахуванням яких відбувається віднесення їх до одного з кластерів.

При самонавчанні дана мережа має чітку фіксовану структуру, тобто певну кількість нейронів, що не піддається змінам протягом «життєвого циклу».

Процес нормалізації вхідних змінних виконується в межах [-1, 1] або [0, 1].

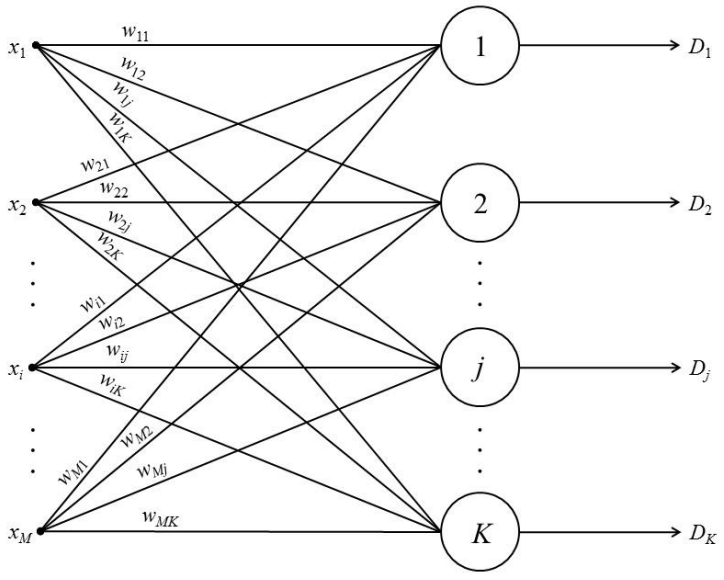


Рисунок 2.1 – Представлення загальної структури нейронної мережі Кохонена

Життєвий цикл нейронної мережі обраної архітектури має три основні стадії життєвого циклу (рис.2.2).

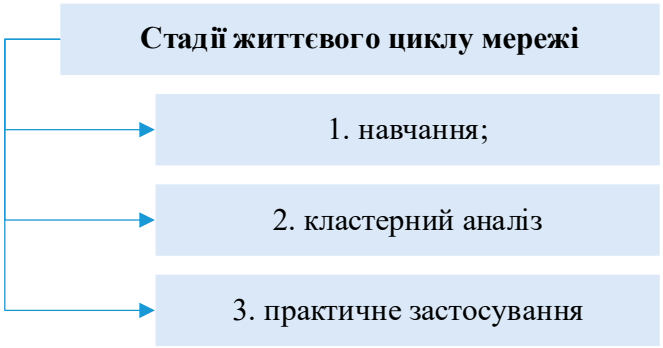


Рисунок 2.2 – Стадії життєвого циклу

Алгоритм навчання для мережі Кохонена складається із певних етапів, склад яких залежить від типу структури: постійної (мережа, що самонавчається) або змінної (мережа, що самоорганізується). В межах даної роботи опишемо алгоритм для самонавчання мережі із врахуванням вагового коефіцієнта.

Крок 1. Визначення структури мережі (к-ть нейронів шару Кохонена) (K).

Крок 2. Випадкова ініціалізація вагових коефіцієнтів значеннями, що будуть задовольняти одне із заданих обмежень:

– за умови нормалізації вихідної вибірки в межах [–1, 1]:

$$|w_{ij}| \leq \frac{1}{\sqrt{M}}$$

– за умови нормалізації вихідної вибірки в межах [0, 1]:

$$0.5 - \frac{1}{\sqrt{M}} \leq w_{ij} \leq 0.5 + \frac{1}{\sqrt{M}}.$$

де w_{ij} – ваговий коефіцієнт; M – кількість вхідних змінних мережі – характеристичних ознак об'єкта дослідження.

Крок 3. Подання на входи мережі випадкового навчального зразку поточної сесії навчання та обчислення відстані Евкліда від вхідного вектора до центрів усіх кластерів:

$$R_j = \sqrt{\sum_{i=1}^{M} (\widetilde{x_i} - w_{ij})^2}$$

Крок 4. По найменшому значенню R_jj обирається нейрон-переможець j, найбільшим чином близький за значеннями з вхідним вектором.

Для обраного нейрона (і тільки для нього) виконується корекція вагових коефіцієнтів:

$$w_{ij}{}^{(q+1)} = w_{ij}{}^{(q)} + v\left(\tilde{x}_i - w_{ij}{}^{(q)}\right),$$

де v – коефіцієнт швидкості навчання.

Крок 5. Цикл повторюватиметься з кроку 3 до виконання однієї чи кількох умов закінчення.

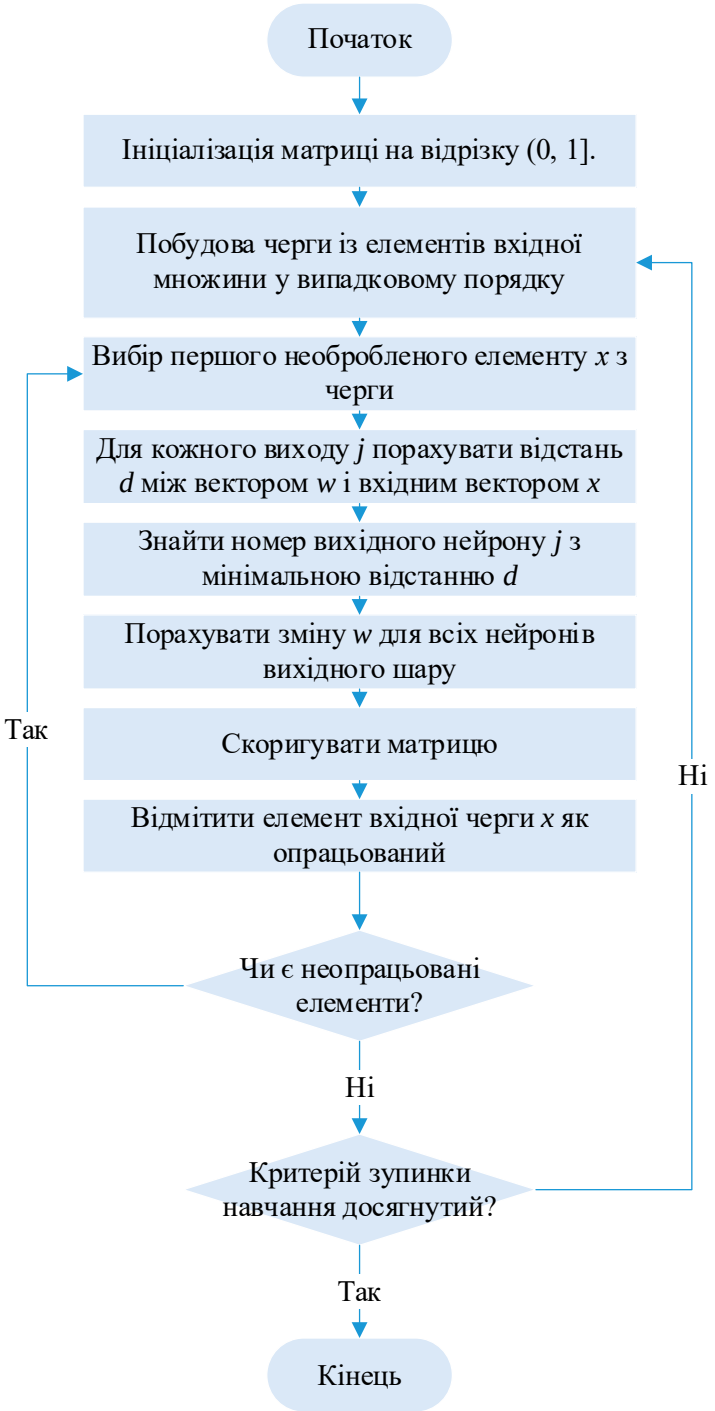


Рисунок 2.3 – Алгоритм процесу навчання нейронної мережі

Умовами для закінчення циклу є:

– вичерпано задану граничну кількість епох навчання;

– не відбулося значної зміни вагових коефіцієнтів у межах заданої точності за час останньої доби навчання;

– вичерпано заданий граничний час навчання.

Коефіцієнт, що визначає швидкість навчання задано постійними межами (0, 1].

Далі до навченої нейронної мережі застосуємо метод кластерного аналізу – процедури опису властивостей кластера на основі аналізу кількісного та якісного складів прикладів, що його сформували на основі яких буде здійснено аналіз наявності загроз веб-ресурсу.

2.2 Вдосконалення методу кластеризації для виявлення загроз веб-ресурсу

Обраний в роботі математичний апарат нейронних мереж на основі карт Кохонена орієнтований на застосування методу кластеризації, що й зумовив його вибір, оскільки застосування кластеризації має ряд переваг (рис. 2.4).

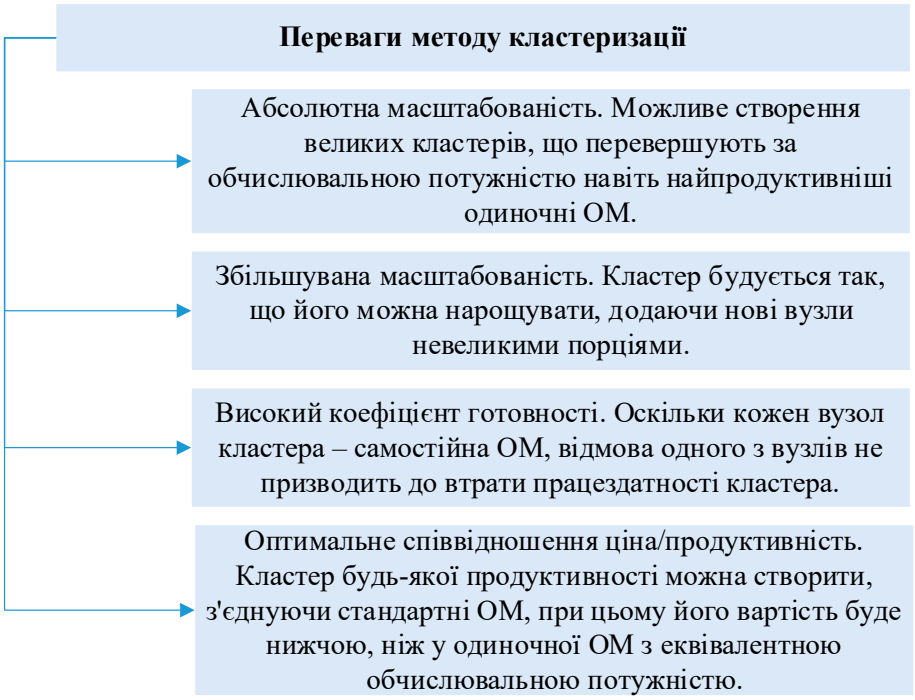


Рисунок 2.4 – Переваги методу кластеризації [28]

Здійснивши аналіз вищенаведеного алгоритму можна відмітити, що недоліком кластеризації на основі карт Кохонена є неможливість розпізнавання складних кластерів. Враховуючи, що при цьому задається точна кількість кластерів, частина даних може бути не врахована, а тому такі недоліки суттєво вплинуть на результат дослідження.

В якості модифікації запропонуємо наступне рішення. Здійснивши навчання на основі карти Кохонена та провівши перший етап кластеризації, продовжимо дослідження за рахунок аналізу відстаней в кластері та міжкластерних відстаней.

Суть ідеї полягає в тому, що після певної кількості сформованих кластерів, продовжимо аналіз неврахованих кластерів, що не містять нейронів, а тому не були враховані до основної вибірки. Аналіз проведемо за рахунок визначення відстаней за одною із метрик.

Говорячи про кластерний аналіз, можна відмітити ряд метрик, що застосовують для визначення відстаней, серед них: відстань Евкліда, міра схожості Хемінга, Манхетинська метрика, відстань Махаланобіса.

Здійснивши аналіз поданих метрик, в межах даної роботи для виконання другого етапу кластеризації вирішено застосовувати відстань Евкліда, враховуючи її основну особливість порівняно з іншими методами: відстань між двома об'єктами не змінюється при введенні в аналіз нового об'єкта, який може виявитися вкидом.

Для обчислення відстані між точками в кластері скористаємось формулою:

$$d_{ik} = \left(\sum_{j=1}^{N} \left(x_{ij} - x_{kj} \right)^2 \right)^{\frac{1}{2}},$$

де d_{ik} – шукана відстань;

x_{ij}і x_{kj} – точки між якими обраховується відстань.

Далі визначимо середнє значення відстані в межах кластера:

$$d_c = \frac{\sum_{i,j}^{t} d\left(x_i, x_j\right)}{t},$$

де $d_c -$ шукана середня відстань,

$t -$ кількість точок в кластері.

Тобто, після того, як на першому етапі сформувалось задане число кластерів відбувається обчислення відстаней між точками в кластері, та відповідно визначення середнього значення відстаней в межах кластера.

Наступним кроком є аналіз сусідніх кластерів, точки яких знаходяться на визначеній середній (або меншій) відстані. Якщо дана відстань менша, ніж найбільша відстань в кластері, що аналізується, то відповідно відбувається уточнення і внесення змін до кластерів.

Таким чином, аналіз даних буде здійснюватися більш детальніше та проходитиме два цикла перевірки. Підвищення точності аналізу даних сприяє підвищенню достовірності отриманих результатів внаслідок здійснення даного аналізу.

2.3 Алгоритм роботи програмного засобу для визначення загроз веб-ресурсу

В даному підрозділі опишемо алгоритм роботи розроблюваного програмного засобу з використанням вдосконаленого методу інтелектуального аналізу даних з підвищеною достовірністю на основі математичного апарату нейронних мереж для визначення стану інформаційної системи на прикладі веб-ресурсу.

Отже, першим кроком розробимо модель інтелектуального аналізу даних, яка складається з наступних етапів:

– визначення мети здійснення інтелектуального аналізу даних;

– вибір засобів для здійснення інтелектуального аналізу даних;

– математична модель обраних засобів;

– підготовка даних для проведення аналізу;

– розробка програмного забезпечення на основі описаного методу;

– представлення на аналіз отриманих результатів.

Схематично дані етапи представимо на рис. 2.5.

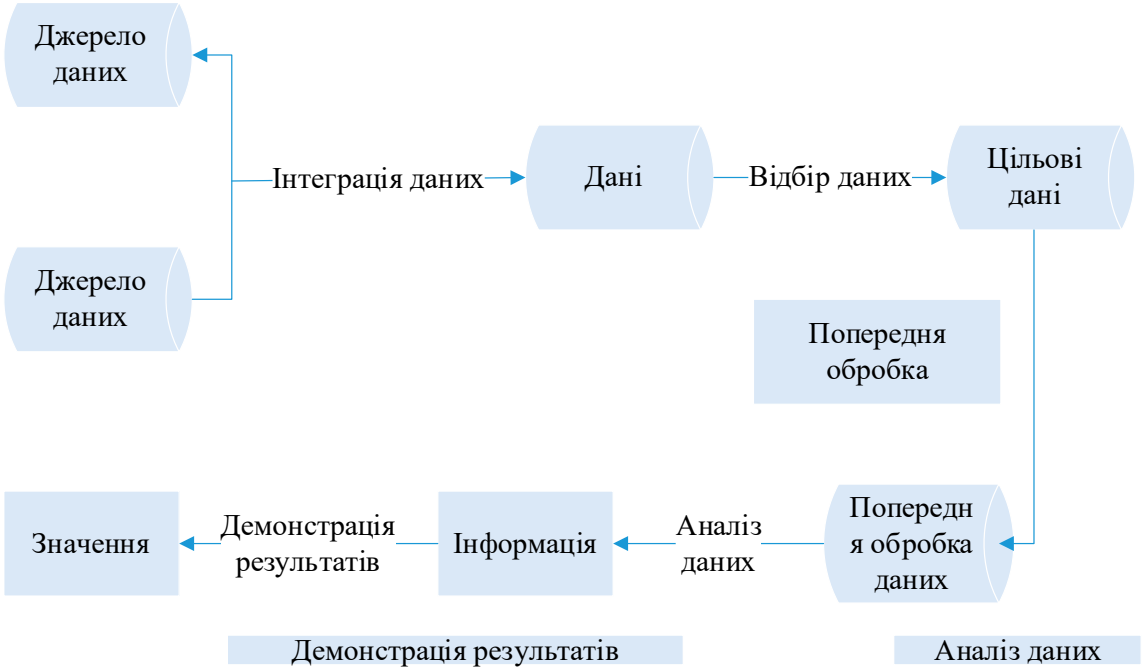


Рисунок 2.5 – Блок-схема етапів моделі інтелектуального аналізу даних

Як зазначалось вище, вдосконалений та описаний в роботі метод інтелектуального аналізу даних призначений для виявлення рівня загроз, стану інформаційної системи на прикладі веб-ресурсу.

Розглянемо чинники, суттєві зміни яких можемо вважати кіберзагрозами та порушеннями стану інформаційної системи:

– країна з якої здійснюється підключення (в більшості випадків кількісний показник звернень користувачів з інших країн до ресурсу орієнтовно відомий та не зазнає суттєвих змін). У випадку, якщо список країн із яких надходять звернення непередбачувано змінюється, або суттєво відрізняється від традиційних показників – можна робити висновок про ризик виникнення кіберзагрози;

– середній час сесії (відвідуючи ресурс користувач переслідує певну мету, з урахуванням того, що дана мета відрізняється для кожного користувача, вираховується середній час сесії). У випадку, якщо зафіксований середній час сесії суттєво відрізняється від типового – можна робити висновок про ризик виникнення кіберзагрози;

– трафік (кількість переданих та отриманих даних також піддається обрахунку середнього показника). У випадку, якщо зафіксоване середнє значення трафіку суттєво відрізняється від типового – можна робити висновок про ризик виникнення кіберзагрози.

Враховуючи, обрані вхідні дані та засоби здійснення інтелектуального аналізу даних, побудуємо відповідну модель (рис. 2.6).

Рисунок 2.6 – Модель виявлення кіберзагроз на основі інтелектуального аналізу даних

Отже, одним із початкових етапів алгоритму за поданою моделлю є визначення змінних. Опишемо їх у табл. 2.1.

Таблиця 2.1 – Характеристики змінних

Змінна	Зміст	Роль	Значення
Result (y)	Виявлення ознак кіберзагроз ресурсу	результуюча (цільова)	$y = 0$ (ознак кіберзагрози не виявлено) $y = 1$ (виявлено ознаки кіберзагрози
Time (x_1)	Середній час сесії користувачів на досліджуваному ресурсі	вхідна	$=$ або $\neq$ середньому значенню (з можливою похибкою)
Country (x_2)	Країна з якої здійснюється підключення до ресурсу	вхідна	UA, USA, FR, DE, Other

Продовження таблиці 2.1

Data (x_3)	Обсяг трафіку, що обробляється під час сесії користувача	вхідна	= або ≠ середньому значенню (з можливою похибкою)

Отже, заданими змінними є:

– результуюча (цільова) змінна $Result\ (y)$, що позначає виявлення ознак кіберзагроз ресурсу, значення змінної отримується в результаті проведеного аналізу та може бути двох типів: $y = 0$ (ознак кіберзагрози не виявлено); $y = 1$ (виявлено ознаки кіберзагрози).

– вхідна зміна Time (x_1), що позначає середній час сесії користувачів на досліджуваному ресурсі, порівнюється із середніми значеннями з можливою похибкою у 10%;

– вхідна зміна Country (x_2), що позначає країну з якої здійснюється підключення до ресурсу, передбачені значення UA (Україна), USA (Сполучені Штати Америки), FR (Франція), DE (Німеччина), Other (інші);

– вхідна зміна Data (x_3), що позначає обсяг трафіку, що обробляється під час сесії користувача, порівнюється із середніми значеннями з можливою похибкою у 10%.

Для описаних змінних сформуємо діаграму класів, що буде застосована при подальшій практичній реалізації.

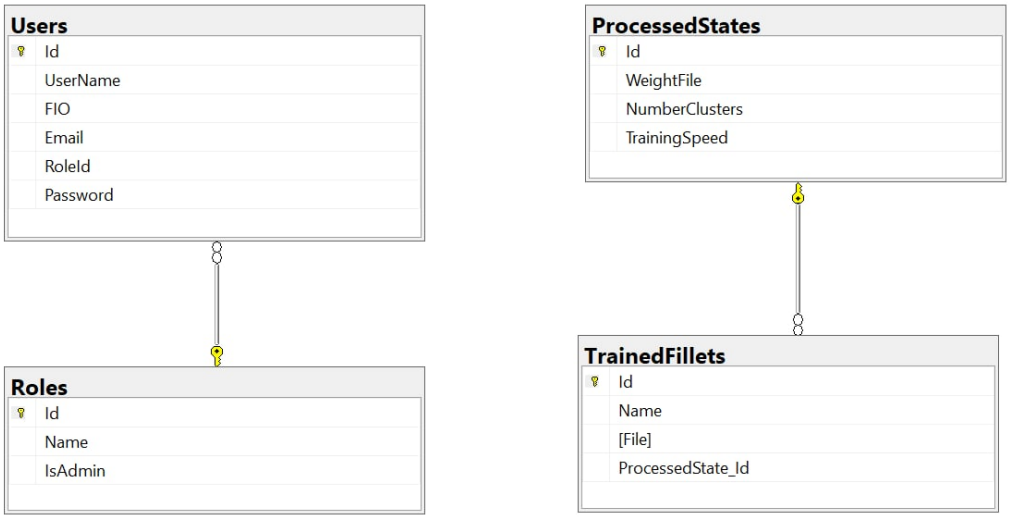


Рисунок 2.7 – Діаграма класів

Наступним кроком розробимо загальний алгоритм роботи програмного засобу для інтелектуального аналізу даних з метою виявлення кіберзагроз для інформаційної системи.

Крок 1. Підключення до програмного додатку.

Для роботи з програмним засобом, що реалізовує інтелектуальний аналіз даних потрібно налаштувати відповідні параметри доступу, зокрема:

– пройти процес авторизації;

– здійснити підключення до відповідних систем з метою надання можливості завантаження даних (СУБД, файли, веб-сервіси).

Крок 2. Завантаження (імпорт даних).

Для здійснення аналізу необхідно реалізувати:

– завантаження файлів для подальшого навчання нейронної мережі;

– дослідження вхідних даних, які дозволяють визначити стан системи відносно до кіберзагроз.

Крок 3. Обробка даних.

Даний крок має на увазі реалізацію можливості:

– навчання нейронної мережі;

– опрацювання завантажених результатів для аналізу;

– формування рішення про наявність кіберзагроз;

– формування графічних даних.

Крок 4. Представлення результатів.

На даному кроці відбувається:

– виведення відповідного сповіщення системою про наявність кіберзагроз;

– представлення графічних даних, що містять детальніші відомості про опрацьовані матеріали.

Крок 5. Вивантаження (експорт) результатів.

Надання користувачеві можливості завантажити опрацьований файл. Дані результати можуть бути використані у подальшій обробці у інших програмах та відповідно слугувати для прийняття подальших рішень.

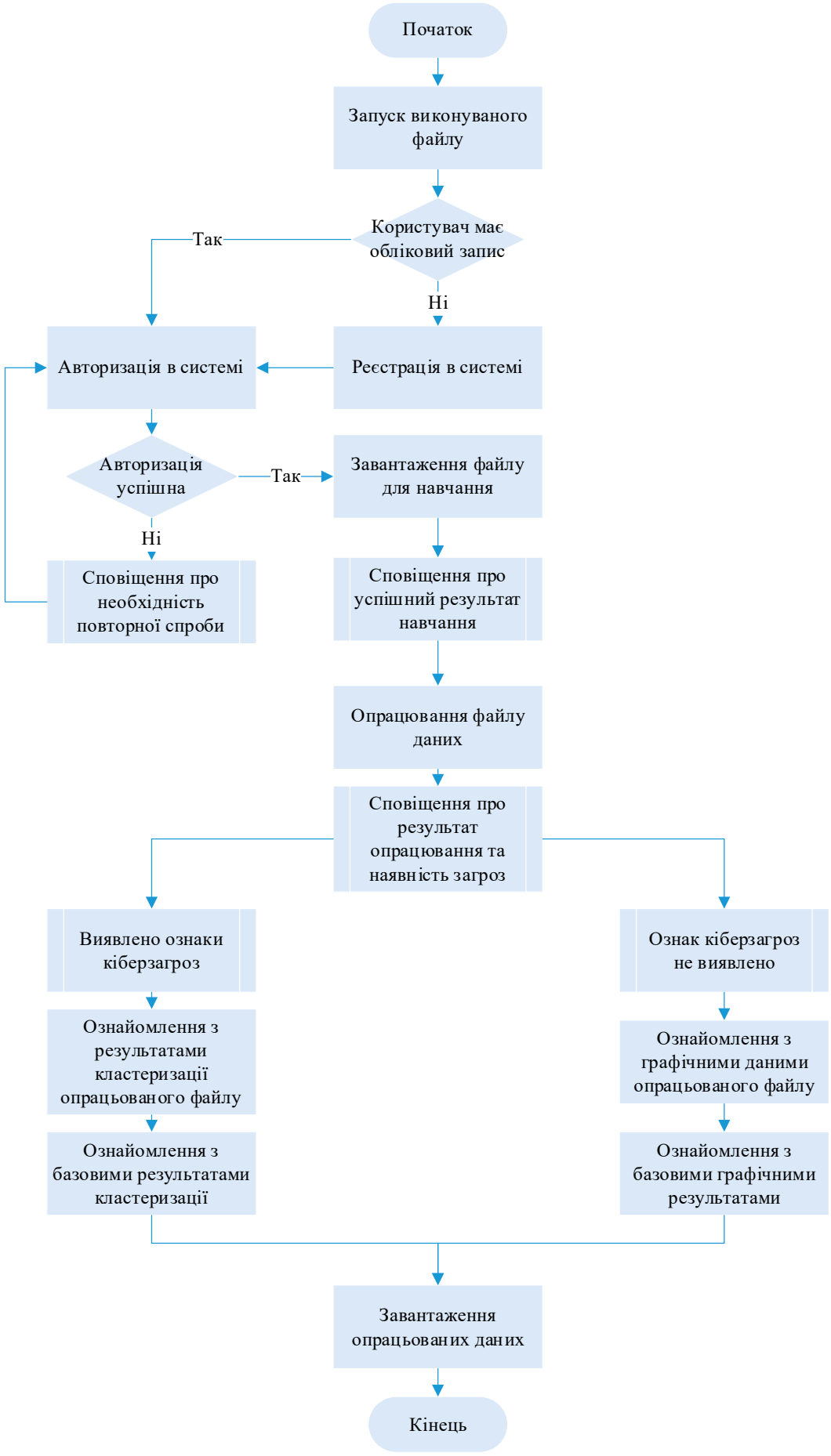


Рисунок 2.8 – Алгоритм роботи програмного засобу

Таким чином, на основі розробленої моделі аналізу даних, обраних засобів аналізу, описаних змінних та загального алгоритму роботи додатку, далі в роботі практично реалізуємо програмний додаток з використанням вдосконаленого методу інтелектуального аналізу даних з підвищеною достовірністю на основі математичного апарату нейронних мереж для визначення ознак кіберзагроз для досліджуваного ресурсу.

2.4 Обґрунтування вибору засобів програмування

Виходячи з особливостей програмного додатку розроблюваного з метою реалізації вдосконаленого методу інтелектуального аналізу даних з підвищеною достовірністю на основі математичного апарату нейронних мереж, було обрано відповідні засоби програмування, а саме:

– мова об’єктно – орієнтованого програмування C# [29];

– середовище програмування Visual Studio [30];

– WPF – як технологію розробки користувацького інтерфейсу [31].

Розглянемо їх детальніше.

Мова C Sharp спочатку була придумана компанією Microsoft для власних цілей та служб [29]. Проте з часом, мова була представлена для загалу та користується попитом серед розробників оскільки має ряд преваг:

– строгу типізацію;

– збереження концепцій об'єктно-орієнтованого програмування;

– функціональність;

– досить потужний інструментарій;

– стабільну роботу через Visual Studio;

– компактний код, що легко читається;

– зрозумілий навіть новачкам синтаксис.

При використанні цієї мови зручно працювати з обробкою винятків, а також наявністю збирача сміття. Тут все продумано так, щоб програмісту було легко писати та зчитувати підсумкові кодифікації.

Синтаксис мови чимось нагадує як C і C++, так і Java. Це робить його доступним для вивчення. Порівняно з іншими мовами програмування, рівень її складності відповідає рівню функцій та можливостей, що вона надає. Тому використання даної мови зумовлене чинником її універсальності та можливості реалізації додатків різної складності.

Visual Studio – середовище розробки, комплексний продукт, представлений з конструкторами графічних інтерфейсів, з можливостями роботи з базами даних, з підтримкою фрагментів коду, з можливостями для перегляду всього проекту в цілому, з переглядом властивостей об'єктів [30].

Хоча названі можливості присутні і в деяких інших IDE-середовищах, Visual Studio має додаткові переваги:

– можливість розробки програм для мобільних пристроїв Windows (Windows Phone);

– можливість розробки програм для Microsoft Office;

– вбудована підтримка рефакторингу коду, тобто покращення існуючої кодової бази, якщо така функція відсутня то перевірка коду стає рутинною роботою, що потребує багато часу;

– інструменти візуального конструювання класів;

– технології WCF, WPF, ASP.

Завдяки величезній кількості налаштувань, підтримуваних технологій, швидкодії та зручності Visual Studio вважається одним із найкращих середовищ розробки, що і зумовлює вибір саме даного середовища для подальшої роботи.

Для розробки графічного користувацького інтерфейсу типовими застосовуваними технологіями є Windows Form та WPF.

Windows Presentation Foundation – це графічна система у складі .Net Framework 3.0 та пізніших версій. Спроектована під впливом технологій HTML та Flash та використовує апаратне прискорення [31].

Технологія WPF була обрана для роботи в зв’язку з наявністю таких переваг:

– нова, а тому відповідає сучасним стандартам розробки;

– Microsoft використовує її у багатьох своїх програмах, наприклад Visual Studio;

– це більш гнучка система, що дозволяє зробити більше, без написання або покупки готових елементів керування;

– якщо є потреба скористатися готовими рішеннями із спільноти, швидше за все сторонні розробники будуть сфокусовані саме на WPF, оскільки це нова система;

– дозволяє створювати GUI як для додатків Windows, так і для додатків Web (Silverlight/XBAP), що доводить її універсальність.

Таким чином, на основі обраних засобів програмування далі в роботі буде здійснено практичну реалізацію вдосконаленого методу кластеризації даних для виявлення наявності загроз на веб-ресурсі.

2.5 Висновки до розділу

Отже, в даному розділі було описано розробку методу інтелектуального аналізу даних на основі нейронної мережі для виявлення наявності загроз веб-ресурсу. В ході роботи було здійснено розробку моделі аналізу даних на основі нейронної мережі, запропоновано та описано вдосконалення методу для аналізу даних, розробку алгоритму роботи програми на основі вдосконаленого методу, обґрунтовано вибір засобів розробки.

В якості модифікації запропоновано наступне рішення. Здійснивши навчання на основі карти Кохонена та провівши перший етап кластеризації, продовжимо дослідження за рахунок аналізу відстаней в кластері та міжкластерних відстаней, щоб виявити невраховані нейрони та підвищити достовірність формування кластеру.

Розроблена система на основі вдосконаленого алгоритму кластеризації в результаті аналізу дозволить здійснювати навчання нейромережі та опрацювання файлів, надавати рішення про наявність загроз для веб-ресурсу. Для реалізації обрано мову C# IDE Visual Studio та технологію розробки GUI WPF.

3 ПРАКТИЧНА РЕАЛІЗАЦІЯ ДОДАТКУ ДЛЯ ВИЗНАЧЕННЯ НАЯВНОСТІ ЗАГРОЗ ВЕБ-РЕСУРСУ НА ОСНОВІ ВДОСКОНАЛЕНОГО МЕТОДУ

Даний розділ описує етап роботи, що має на меті практичну реалізацію описаних алгоритмів для демонстрації практичних результатів на основі методу інтелектуального аналізу даних за рахунок використання нейронної мережі Кохонена та вдосконаленого алгоритму кластеризації даних для виявлення загроз веб-ресурсу.

В ході виконання розділу буде опрацьовано створення графічного користувацького інтерфейсу, описано особливості розробки, пояснено інструкцію користувача для роботи з додатком та здійснено тестування та аналіз отриманих результатів.

3.1 Проектування інтерфейсу програмного додатку

Розробка графічного користувацького інтерфейсу один із етапів практичної реалізації програмного додатку, що передбачений для використання штатним працівником.

Розробляючи інтерфейс додатку розробнику слід звернути увагу на основні вимоги, яких варто дотримуватись, серед них:

– простота та зрозумілість інтерфейсу (діалогові вікна повинні бути інтуїтивно зрозумілими, не викликати у користувача труднощів виконати певну дію);

– доступність (основні функції передбачені програмним додатком повинні бути розміщенні на головних діалогових вікнах, а не приховані за рядом кнопок);

– відповідність призначенню додатку (якщо додаток призначений для аналітики (як у межах даної роботи) використання яскравого фону та зайвих зображень є недоречним; розмір вікон повинен відповідати матеріалу, який вони відображають (наприклад, вікна з графіками не повинні бути надто малими);

– структурованість (перехід по діалоговим вікнам повинен бути

структурований, якщо користувач виконує певні передбачувані дії в додатку, то розміщення функціональних кнопок повинне бути відповідне, щоб не повертатись на кілька вікон назад для здійснення наступної команди).

Враховуючи наведені вимоги, спроектуємо діалогові вікна розроблюваного додатку, що призначений для інтелектуального аналізу даних на основі кластеризації та нейромережі для виявлення загроз веб-ресурсу.

Одним із основних діалогових вікон є вікно, що містить функціональні кнопки «Навчання», «Опрацювання» та «Історія». Дані кнопки користувач застосує для завантаження та вивантаження файлів в/із додатку (рис. 3.1).

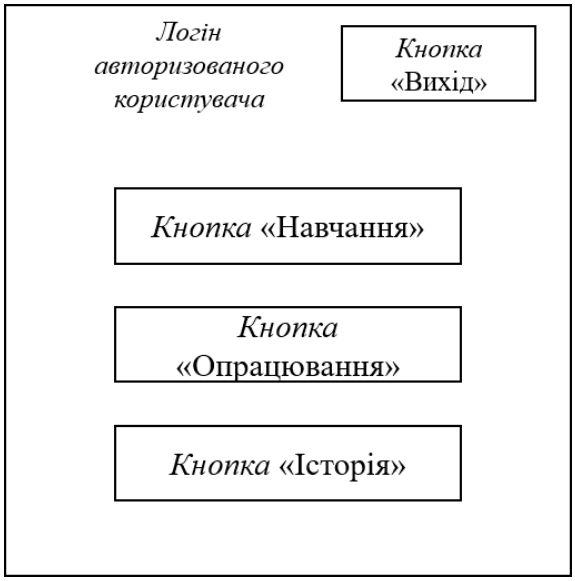


Рисунок 3.1 – Проектування діалогового вікна для завантаження / вивантаження файлів

До наступного діалогового вікна користувач потрапляє після натиснення функціональної кнопки «Навчання». Дане вікно має функціональне призначення, що полягає в керуванні навчанням нейронної мережі. У верхній частині вікна зазначимо назву процесу. Нижче розташована кнопка «завантажити файл», натискаючи яку користувачеві надається можливість відкрити директорію та обрати необхідний для навчання файл. Поряд із кнопкою розташоване поле, що відображатиме шлях до обраного файлу. Нижче розташоване поле, що передбачено для зазначення кількості кластерів, що будуть сформовані при навчанні мережі.

Для запуску процесу навчання користувачеві слід скористатись відповідною функціональною кнопкою «Опрацювати файл», що буде розміщена у нижній частині вікна (рис. 3.2).

Рисунок 3.2 – Проектування діалогового вікна для здійснення процесу навчання мережі

Наступне діалогове вікно призначене для опрацювання необхідного файлу на основі навченої мережі, перехід до нього відбувається через функціональну кнопку «Опрацювання» (рис. 3.3).

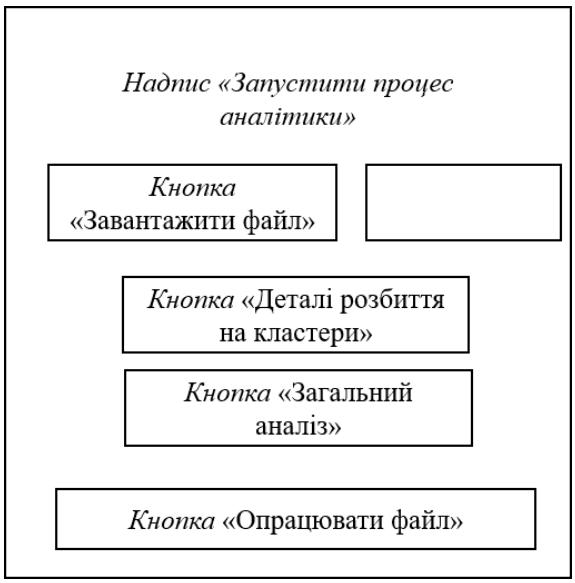


Рисунок 3.3 – Проектування діалогового вікна для здійснення процесу опрацювання файлів

У верхній частині даного вікна зазначено назву процесу, що дане вікно реалізує. Нижче розміщена кнопка «Завантажити файл», що аналогічно як у попередньому вікні, дозволяє користувачеві обрати файл із директорії, поряд знаходиться поле, яке відображатиме шлях до обраного файлу.

У середній частині вікна розмістимо функціональні кнопки «Деталі розбиття на кластери» та «Загальний аналіз», натиснувши на дані кнопки, користувач матиме змогу переглядати графічне представлення результатів інтелектуального аналізу даних для виявлення загроз веб-ресурсу. Дані кнопки на початку роботи у даному вікні залишатимуться у неактивному стані. Їх активація відбудеться лише після опрацювання файлу та наявності відповідних результатів.

В нижній частині вікна розташована функціональна кнопка «Опрацювати файл» для запуску відповідного процесу.

Таким чином, на основі поданих спроектованих діалогових вікон, далі в роботі буде реалізована програмна розробка додатку, що призначений для практичної демонстрації вдосконаленого методу на основі нейронної мережі та методу кластеризації.

3.2 Особливості програмної реалізації додатку

Виходячи з поставленої мети роботи було обрано відповідні засоби програмування (мова об’єктно – орієнтованого програмування C#; середовище програмування Visual Studio; технологія розробки інтерфейсу WPF) та практично реалізовано програмний додаток на основі вдосконаленого методу кластеризації даних з використанням математичного апарату нейронних мереж для виявлення кіберзагроз інформаційній системі (на прикладі веб-ресурсу).

В даному підрозділі опишемо основні фрагменти реалізованої кодової послідовності, що були написані при створенні програмного додатку.

Для програмної реалізації використовуються наступні бібліотеки:

```
using DataAnalyze.Context;
using DataAnalyze.Models;
```

```
using Newtonsoft.Json;
using System;
using System.Collections.Generic;
using System.ComponentModel;
using System.Data.Entity;
using System.IO;
using System.Linq;
using System.Text;
using System.Windows;
```

Реєстрація користувача реалізована на основі класу Registration:

```
public partial class Registration : Window, INotifyPropertyChanged
  {
    public Registration()
    {
      InitializeComponent();
    }
    public event PropertyChangedEventHandler PropertyChanged;
    private RecognizeContext _context = new RecognizeContext();
```

Для здійснення процесу реєстрації користувачеві необхідно вказати наступні дані: пошта, повне ім’я, логін та пароль:

```
User dbUser = new User()
            {
              Email = UserName.Text,
              FIO = UserName.Text,
              UserName = UserName.Text,
              Password = Password.Text
            };
```

У випадку, якщо пошта введена при реєстрації вже відома базі даних додатку, користувач отримує відповідне повідомлення:

```
if (user != null)
          {
            MessageBox.Show("Користувача вже зареєстровано");
          }
```

Авторизація користувача реалізована на основі класу Log:

```
public partial class Log : Window
  {  public Log()
    { InitializeComponent();
    }
    private void Window_Loaded(object sender, RoutedEventArgs e)
    {
      text1.Text = String.Join("\n", Helper.GeLogFile());   }  }
```

Якщо спроба авторизації користувача вдала, він отримує відповідне сповіщення:

```
MessageBox.Show("Спроба входу була вдалою!");
```

AddTraining – клас, що реалізовує процес навчання мережі описаний наступним чином:

```
public partial class AddTraining : Window
  {  public AddTraining()
     {   InitializeComponent();
     }
     private RecognizeContext _context = new RecognizeContext();
     private void Button_Click_2(object sender, RoutedEventArgs e)
     {   OpenFileDialog dlg = new OpenFileDialog();
        dlg.RestoreDirectory = true;
        if (dlg.ShowDialog() == true)
        {    string selectedFileName = dlg.FileName;
           FileName.Text = selectedFileName;
        }
     }
```

Первинна обробка даних в файлі:

```
private void ProcessCv(string path)
     {    using (TextFieldParser parser = new TextFieldParser(path))
        {    parser.TextFieldType = FieldType.Delimited;
           parser.SetDelimiters(",");
           while (!parser.EndOfData)
           {    string[] fields = parser.ReadFields();
              foreach (string field in fields)
              {
```

Після здійснення обробки даних. Користувач отримує відповідне сповіщення про успішне опрацювання файлу:

```
Helper.AddLog("Added file for train");
        MessageBox.Show("Файд успішно опрацьовано");
```

Клас, що реалізує аналіз завантаженого файлу на основі навченої нейронної мережі:

```
private void ProcessCv(string path)
     {
        var list = new string[7][];
        using (TextFieldParser parser = new TextFieldParser(path))
        {   int i = 0;
           parser.TextFieldType = FieldType.Delimited;
           parser.SetDelimiters(",");
```

```
while (!parser.EndOfData)
{ string[] fields = parser.ReadFields();
   list[i] = fields;
   i = i + 1;
}
csvObject = new CSVobject()
{   SessionAnalytic = ProcessArray(list[1]),
   CountryAnalytic = ProcessArray(list[2]),
   TraficAnalytic = ProcessArray(list[3]),
};
csvObjectBlok = ProcessArray(list[0]);
}
```

Після опрацювання завантаженого файлу відбувається виведення відповідного результату:

– виведення результату про відсутність загроз ресурсу:

```
if (csvObjectBlok.First().Key == "1")
        {
          MessageBox.Show("Файл було успішно опрацьовано. \r\nОтримані результати аналізу даних свідчать про відсутність загроз ресурсу.");
        }
```

– виведення результату про наявність загроз ресурсу:

```
if (csvObjectBlok.First().Key != "1")
        {
          MessageBox.Show("Файл було успішно опрацьовано. \r\nОтримані результати аналізу даних свідчать про наявність загроз ресурсу.");
        }
```

Для формування історії файлів, що доступні для перегляду та завантаження користувачем виконуємо наступну функцію:

```
{     string aText = null;
        System.Windows.Forms.FolderBrowserDialog openFileDlg = new System.Windows.Forms.FolderBrowserDialog();
        var result = openFileDlg.ShowDialog();
        if (result.ToString() != string.Empty)
        {
          aText = openFileDlg.SelectedPath;
        }
```

Підготовка файлів для завантаження з системи користувачем:

```
if (row != null)
            {  var uiUser = row.DataContext;
              var selectedId = Convert.ToInt32(uiUser.GetType().GetProperty("Id").GetValue(uiUser, null));
```

```
var file = await _context.TrainedFillets.FirstOrDefaultAsync(x => x.Id == selectedId);
File.WriteAllBytes(Path.Combine(aText, file.Name), Convert.FromBase64String(file.File));}
```

Повідомлення про успішне завантаження файлу:

```
MessageBox.Show("Файд успішно збережено");
```

Таким чином, на основі обраних засобів програмування була здійснена практична розробка програмного додатку, що реалізує навчання нейронної мережі, вдосконалений метод кластеризації, генерує відповідний результат та надає його графічне представлення.

3.3 Інструкція користувача для роботи з додатком

В даному розділі опишемо інструкцію користувача для роботи з розробленим додатком на основі вдосконаленого методу інтелектуального аналізу даних з підвищеною достовірністю з використанням математичного апарату нейронних мереж.

Враховуючи, що розроблюваний програмний додаток призначений для роботи з даними, що становлять певну цінність для їх власника, в додатку передбачений базовий модуль захисту – авторизація користувача для підтвердження особи.

Отже, після запуску виконуваного файлу, користувачеві відкривається форма входу, де йому необхідно ввести логін та пароль (рис. 3.4).

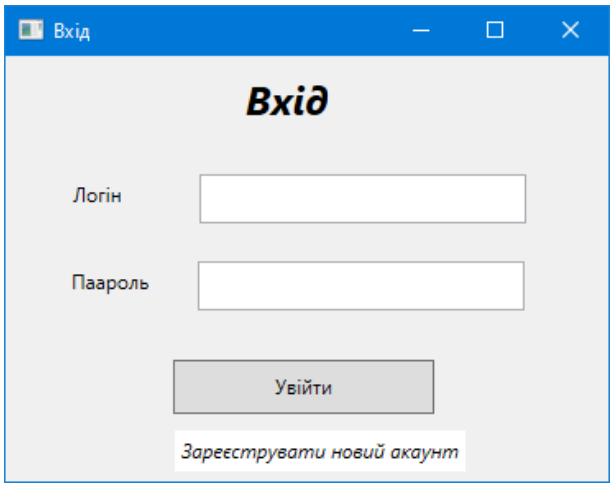


Рисунок 3.4 – Вигляд вікна форми входу для користувача

Після введення логіну та паролю, користувачеві слід натиснути на кнопку «Увійти» та дочекатись відповідного сповіщення. У випадку, якщо дані вказані вірно, відкривається сповіщення про успішну спробу входу (рис. 3.5).

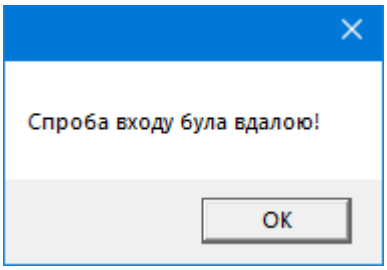


Рисунок 3.5 – Сповіщення про успішну спробу входу

У випадку якщо користувач вводить невірний або неіснуючий логін чи некоректний пароль, додаток відповідає відповідним сповіщенням (рис.3.6).

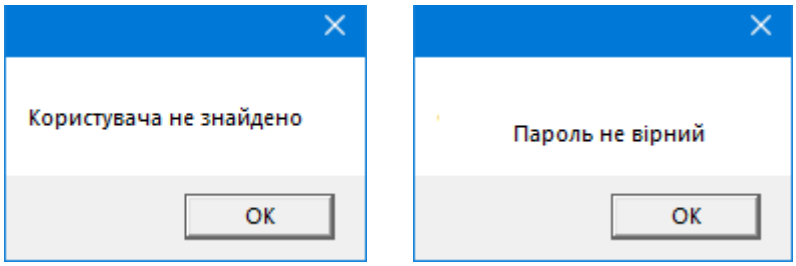


Рисунок 3.6 – Сповіщення про невдалу спробу входу

Якщо користувач ще немає облікового запису, йому слід скористатись кнопкою «Зареєструвати новий акаунт» та заповнити форму реєстрації. Дана форма містить базові поля «Логін», «Email» та «Пароль» (рис. 3.7)

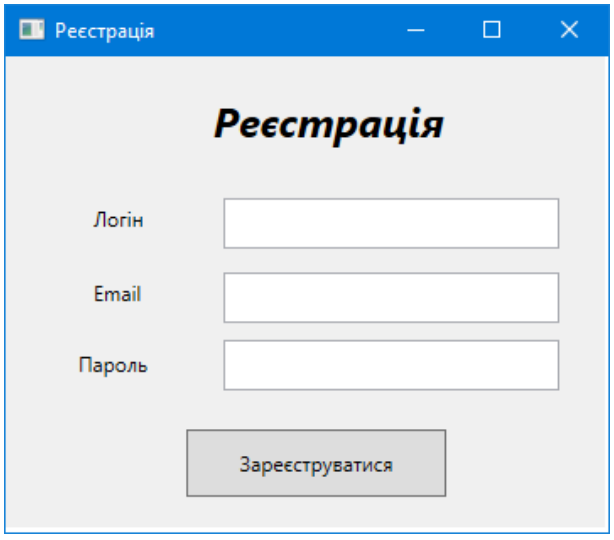


Рисунок 3.7 – Вигляд вікна форми авторизації

Після підтвердження реєстрації користувача адміністратором, він є повноцінним користувачем системи.

Наступний модуль додатку – опрацювання файлів, здійснення інтелектуального аналізу даних.

У вікні додатку розташовані функціональні кнопки:

– «Навчання» – для завантаження файлу на основі якого відбувається навчання нейронної мережі;

– «Опрацювання» – для завантаження файлів, що будуть піддаватись аналізу нейронною мережею;

– «Історія» – для перегляду та можливості завантаження аналізованих файлів для подальшого опрацювання.

Вигляд даного вікна представлено на рис. 3.8.

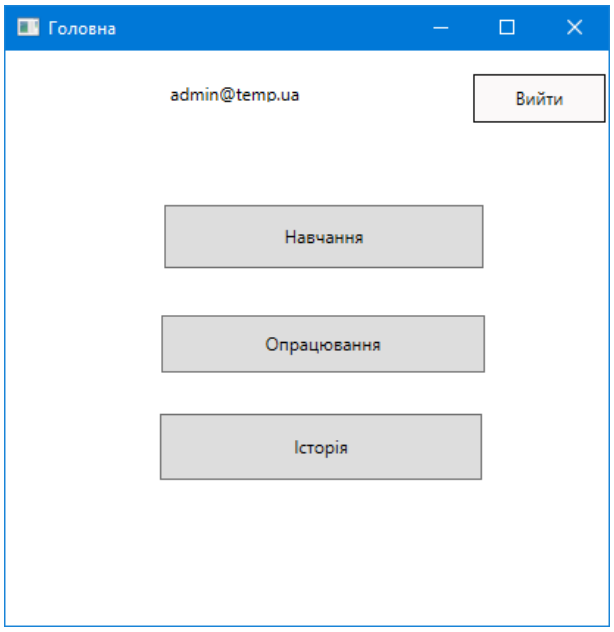


Рисунок 3.8 – Вигляд вікна з відображенням функціоналу додатку

Натиснувши кнопку «Навчання» відкривається наступне діалогове вікно. Для завантаження файлу користувачеві слід натиснути відповідну кнопку та обрати з директорії потрібний файл.

Нижче у відповідному полі вказується кількість кластерів, що задаються для нейронної мережі.

Для записку процесу навчання слід натиснути кнопку «Опрацювати файл».

Вигляд даного вікна представлено на рис. 3.9.

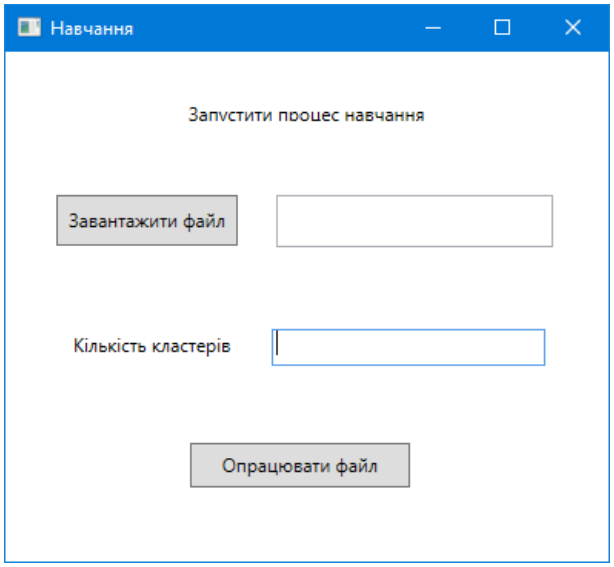


Рисунок 3.9 – Вигляд вікна з функціональною можливістю навчання нейромережі

Після здійсненого процесу навчання, користувач може скористатись кнопкою «Опрацювання» та здійснити аналіз необхідних даних на основі навченої нейронної мережі. Для цього через відповідну кнопку завантаження обирається потрібний файл із директорії та запускається процес опрацювання.

Розташовані у вікні кнопки «Деталі розбиття на кластери» та «Загальний аналіз» залишаються неактивним поки не буде завершено процес аналізу.

Вигляд даного вікна представлено на рис. 3.10.

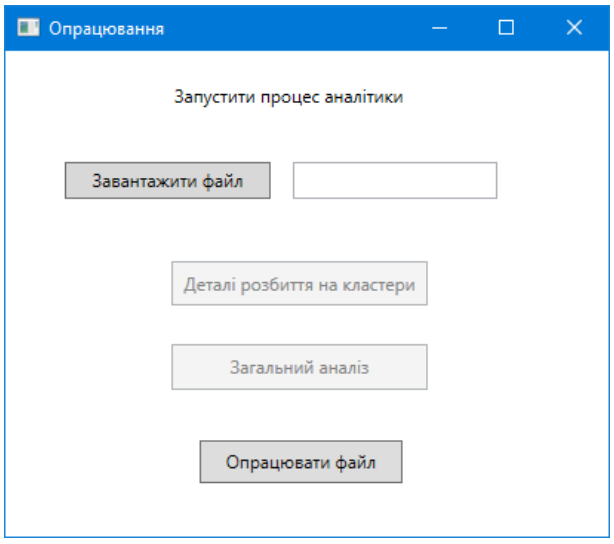


Рисунок 3.10 – Вигляд вікна з функціональною можливістю аналізу даних

Після завершення аналізу, користувач може скористатись кнопками «Деталі розбиття на кластери» та «Загальний аналіз», які функціонально передбачені для представлення в графічному форматі результатів аналізу. Виведені вікна з графіками міститимуть кнопку «Базове відображення», що відкриває результати навчання мережі та, відповідно, надає користувачеві можливість здійснити візуальний аналіз. Натиснувши кнопку «Історія» у початковому вікні (рис. 3.7) користувач зможе переглянути журнал опрацьованих файлів та завантажити необхідний файл для подальшої його обробки. Таким чином, на основі наведених діалогових вікон представлена практична розробка додатку, що призначений для демонстрації результатів інтелектуального аналізу даних для визначення наявності кіберзагроз на основі вдосконаленого методу.

3.4 Тестування та аналіз програмного додатку для виявлення загроз веб-ресурсу

Для перевірки практичних результатів вдосконаленого методу здійснимо аналіз даних статистики звернень до веб-ресурсу на основі розробленого програмного додатку. Після успішної авторизації в системі, натиснемо кнопку «Навчання»» та завантажимо файл для навчання нейронної мережі (рис. 3.11).

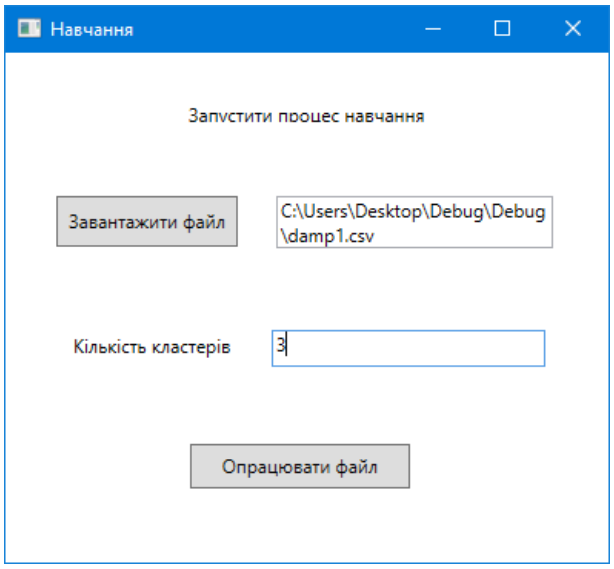


Рисунок 3.11 – Завантаження файлу для навчання системи

В даному вікні натиснемо «Завантажити файл» та оберемо необхідний файл із директорії, а також вкажемо кількість кластерів для першого етапу кластеризації на основі мережі Кохонена.

Далі опрацюємо файл. Отримуємо сповіщення про успішне опрацювання файлу, тобто підтвердження здійснення етапу навчання системи (рис.3.12).

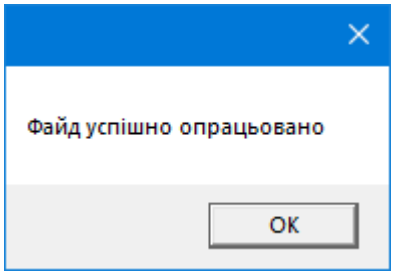


Рисунок 3.12 – Сповіщення про успішне опрацювання файлу

Далі натиснемо кнопку «Опрацювання» та у відкритому вікні завантажимо файл для аналізу вже на основі навченої мережі.

Обираємо файл із директорії та натискаємо «Опрацювати файл», після успішного опрацювання отримуємо відповідне сповіщення та повертаємось до вкладки опрацювання (3.13).

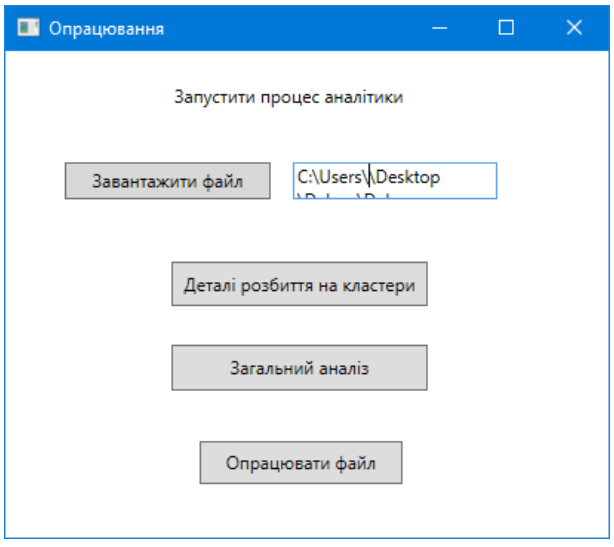


Рисунок 3.13 – Вікно додатку для завантаження файлу на основі навченої нейромережі

Звертаємо увагу, що функціональні кнопки для представлення аналітики стали активними. Натискаємо кнопку «Деталі розбиття на кластери» та

переглянемо отриманні результати кластеризації (рис. 3.14).

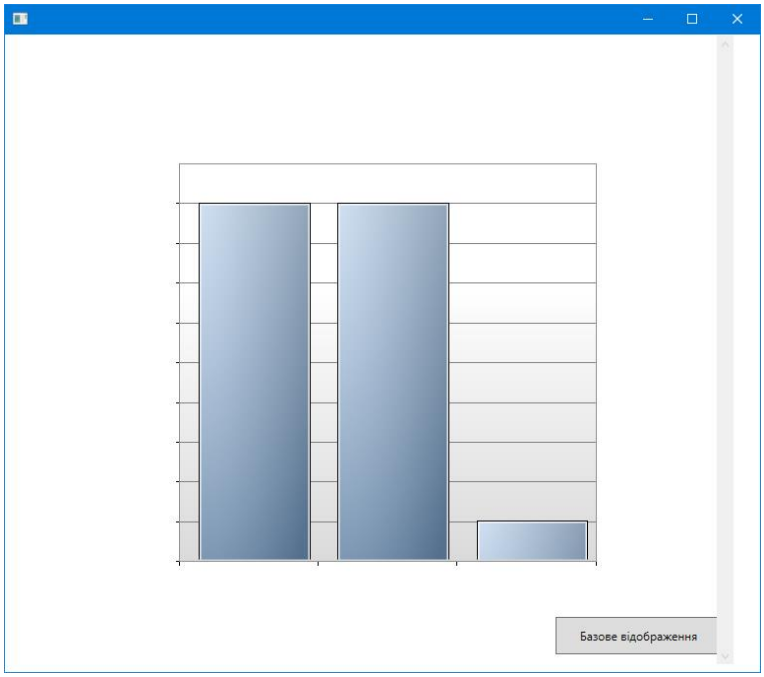


Рисунок 3.14 – Деталі розбиття на кластери аналізованого файлу (1 випадок)

Для порівняння отриманого результату кластеризації із кластеризацією файлу для навчання скористаємось кнопкою «Базове відображення» (рис.3.14).

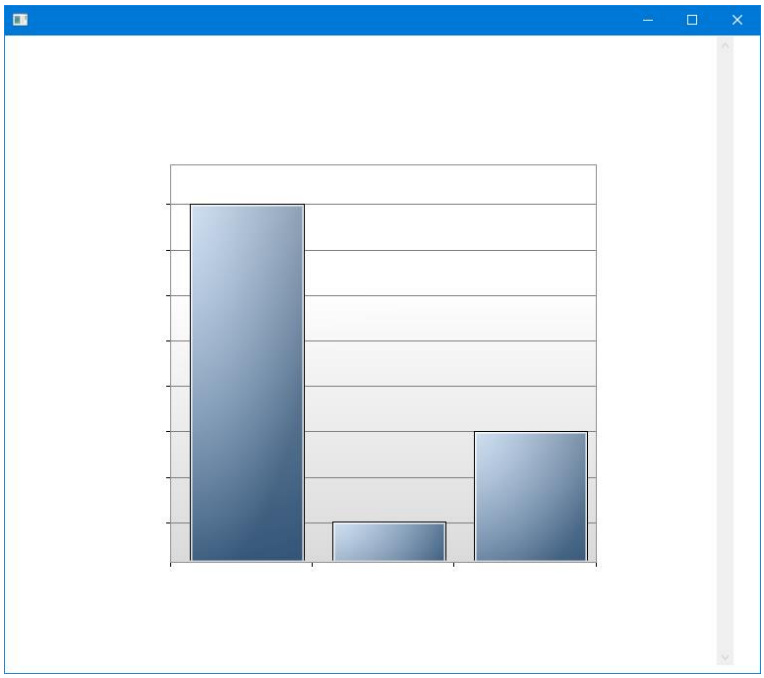


Рисунок 3.15 – Деталі розбиття на кластери файлу для навчання

Як можемо побачити, результати кластеризації аналізованого файлу суттєво відрізняються, від кластеризації файлу для навчання. Для більш

детальнішого аналізу переглянемо дані по кластерам. Для цього скористаємось кнопкою «Загальний аналіз» (рис. 3.16). Для перегляду даних файлу навчання нейронної мережі скористаємось кнопкою «Базове відображення» (рис.3.17).

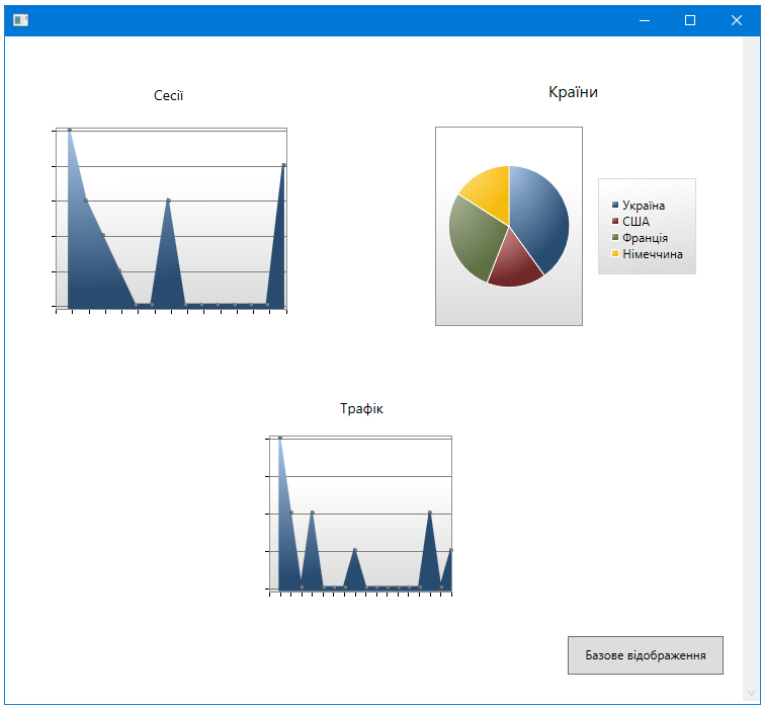


Рисунок 3.16 – Деталі аналізу даних опрацьованого файлу (1 випадок)

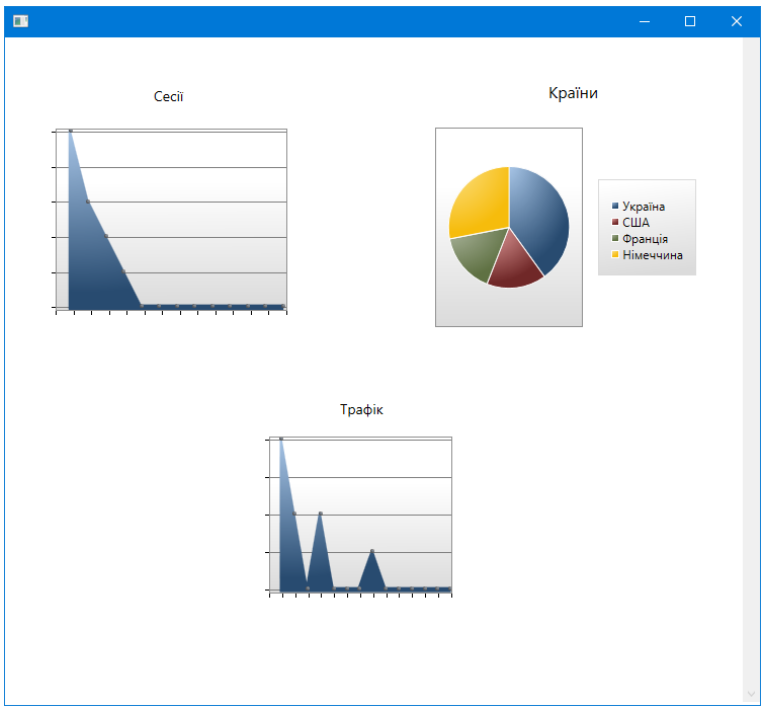


Рисунок 3.17 – Деталі аналізу даних файлу для навчання

Як можемо побачити, значні зміни спостерігаються при аналізі часу сесій,

часткові зміни діаграми звернень з країн та зміни показнику використовуваного обсягу трафіку.

Алгоритм аналізу даних на основі нейронної мережі, враховуючи інформацію отриману при навчанні та із врахуванням допустимих меж похибок, робить висновок про ймовірність загроз для аналізованого ресурсу (рис. 3.18).

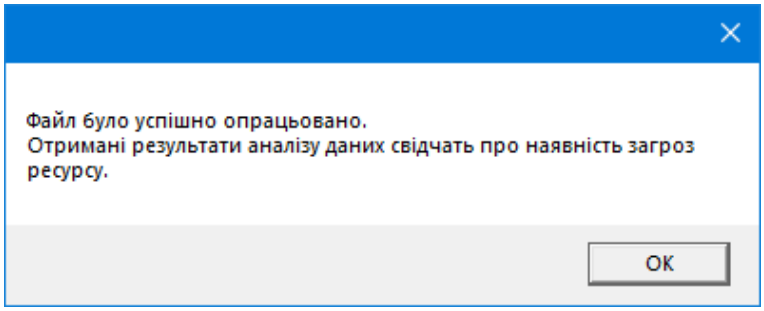


Рисунок 3.18 – Виведення сповіщення про наявність загроз ресурсу

Далі проаналізуємо наступний файл. Завантажимо файл та запустимо його опрацювання. Перейдемо у вкладку результатів кластеризації (рис.3.19).

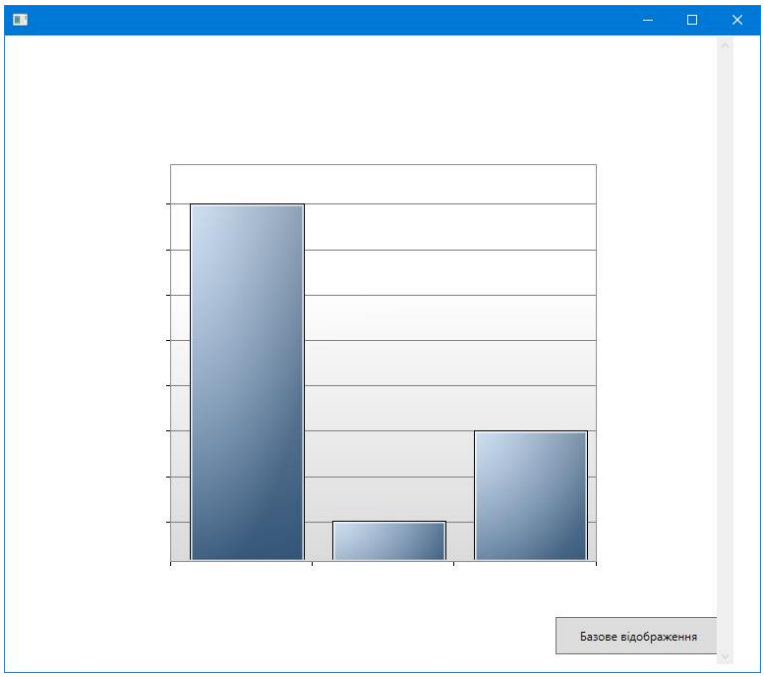


Рисунок 3.19 – Деталі розбиття на кластери аналізованого файлу (2 випадок)

Наступним кроком здійснимо перегляд детального аналізу даних (рис. 3.20).

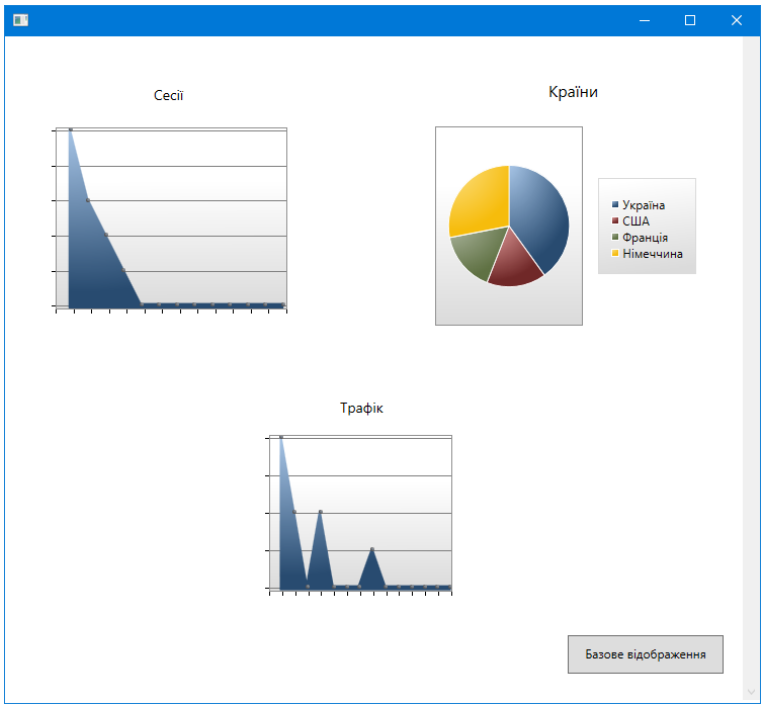


Рисунок 3.20 – Деталі аналізу даних опрацьованого файлу (2 випадок)

Як можемо побачити із наведених результатів та порівнявши отримані дані із даними файлу навчання, суттєвих відмінностей не помітно. На основі впровадженого алгоритму приймається відповідне рішення та подається у відповідному сповіщенні (рис. 3.20).

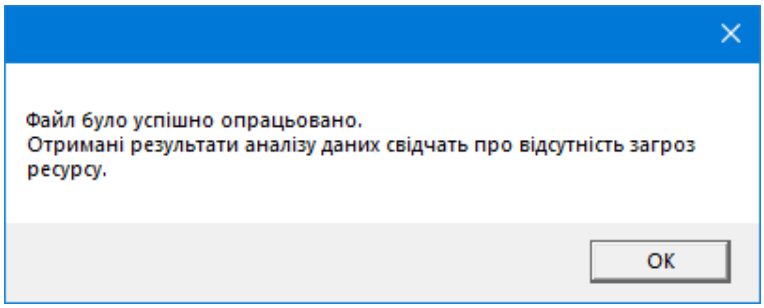


Рисунок 3.21 – Виведення сповіщення про відсутність наявності загроз ресурсу

Для подальшої обробки результати аналізу можна зберегти. Для цього користувачеві слід скористатись кнопкою «Історія», після чого відкриється відповідне діалогове вікно (рис. 3.22).

Обравши відповідний файл та шлях його завантаження через директорію системи, відбувається завантаження файлу із подальшим відповідним сповіщенням (рис. 3.23).

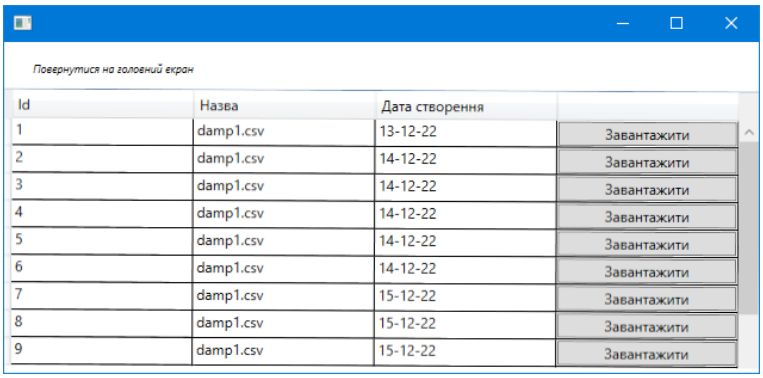


Рисунок 3.22 – Вигляд вікна історії опрацьованих файлів

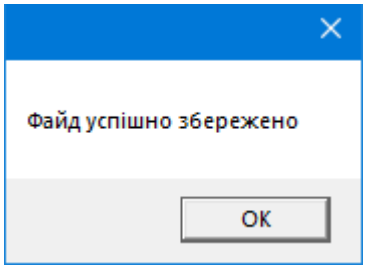


Рисунок 3.23 – Сповіщення про успішне завантаження файлу

Таким чином, на основі удосконаленого алгоритму кластеризації даних та подальшого їх аналізу було визначено наявність кіберзагроз для досліджуваного ресурсу.

Наступним етапом здійснимо оцінку якості кластеризації, для цього застосуємо одну із метрик оцінювання кластеризації – індекс силуетів [32].

Значення силуету вимірюється в межах від -1 до 1.

Якщо більшість об’єктів мають високе значення силуету, то кластеризація виконана вірно; якщо у багатьох об’єктів низьке або від’ємне значення силуету, це вказує на некоректну кількість кластерів.

Силует кожного кластеру можемо визначити наступним чином:

– нехай елемент x_j належить кластеру c_p;

– визначимо середню відстань від однієї точки до іншої того ж кластера c_p через значення a_{pj};

– визначимо середню відстань від x_j до точки із іншого кластера c_q, де $q \neq p$ через d_{qj};

– визначимо $b_{pj} = min_{q \neq p} d_{qj}$, де b_{pj} – міра відмінності окремого

елемента з елементами найближчого кластера.

– визначимо силует окремого кластеру (S_{xj}) поданою нижче формулою.

$$S_{xj} = \frac{b_{pj} - a_{pj}}{\max(a_{pj}, b_{pj})}$$

Отже, з поданої формули можна побачити, що високе значення S_{xj} характеризує «кращу» належність елемента x_j до кластеру p.

Для оцінки всієї кластерної структури знайдемо середнє значення показника по елементам за поданою формулою:

$$SWC = \frac{1}{N}\sum_{j=1}^{N} S_{xj}$$

В ході проведення аналізу на основі опрацьованих файлів було відповідно визначено індекс силуетів для отриманих векторів.

Пораховане програмно значення $SWC = 0,95$.

Даний показник свідчить про коректну кількість кластерів та розподілені дані у них.

Для визначення ефективності вдосконаленого методу, порівняємо його із стандартним методом кластеризації на основі нейромережі Кохонена.

Для цього скористаємось вибіркою файлів, в яких результати про наявність загроз заздалегідь відомі.

Вибірка становить 20 файлів із даними, в яких 15 файлів мають допустимі результати та не свідчать про наявність загроз в системі, 5 файлів тим чим іншим чином містять дані, що вказують на можливі наявні загрози.

Результати аналізу представимо у табл. 3.1 та на рис. 3.24 та 3.25.

Таблиця 3.1 – Результати порівняння методів кластеризації

Метод	Загроз не виявлено	Загрозу виявлено
Стандартна кластеризація	17	3
Вдосконалений метод	15	5

Рисунок 3.24 – Результати аналізу на основі стандартної кластеризації

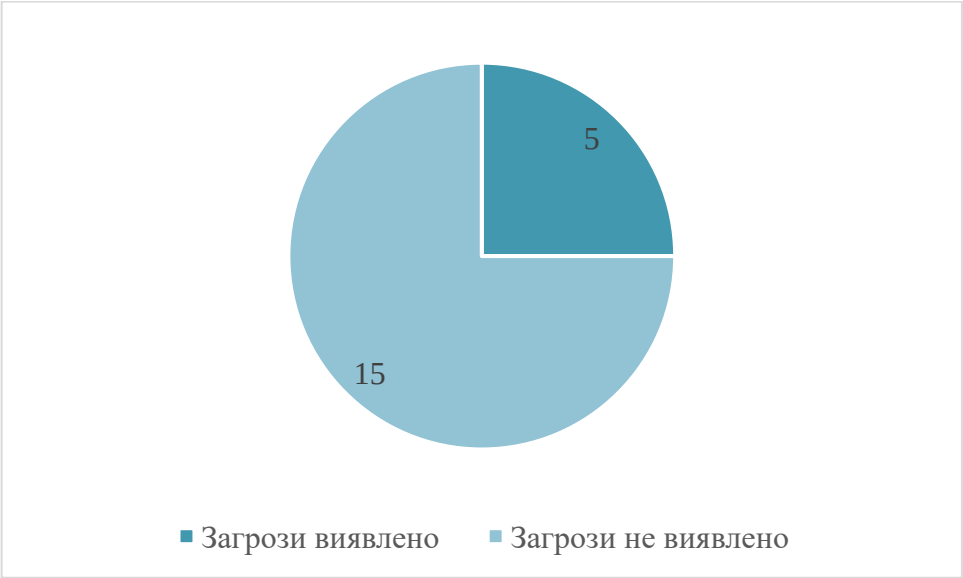


Рисунок 3.25 – Результати аналізу на основі вдосконаленого методу кластеризації

Як можемо побачити із наведених даних, використовуючи стандартний алгоритм кластеризації при аналізі даних для виявлення кіберзагроз, існує похибка. Це зумовлено недостовірним розподілом кластерів, що відповідно не дало змогу оцінити всіх необхідних даних. Врахування певного набору даних при аналізі наявності кіберзагроз інформаційній системі є важливим, оскільки найменші окремі відхилення можуть свідчити про можливу атаку чи злам.

Таким чином, застосування вдосконаленого алгоритму кластеризації даних дозволило підвищити достовірність аналізу даних орієнтовно на 10% та покращити загальний показник системи інтелектуального аналізу даних із високою оцінкою якості кластеризації.

3.5 Висновки до розділу

Отже, в даному розділі було описано розробку програмного додатку розроблюваного з метою реалізації вдосконаленого методу інтелектуального аналізу даних з підвищеною достовірністю на основі математичного апарату нейронних мереж призначеного для виявлення кіберзагроз інформаційній системі.

Для реалізації роботи було використано такі засоби програмування як мова об’єктно – орієнтованого програмування C#; середовище програмування Visual Studio; технологія розробки інтерфейсу WPF.

В ході виконання роботи було здійснено проектування користувацького інтерфейсу, описано особливості здійсненої програмної реалізації, описано інструкцію користувача та проведено аналізів результатів інтелектуального аналізу даних.

Аналіз результатів реалізованого додатку, що є демонстрацією вдосконаленого алгоритму для визначення наявності кіберзагроз інформаційної системи показав, що програма коректно виконує функції навчання нейронної мережі, інтелектуальний аналіз на основі навченої мережі, генерує відповідний результат та надає його графічне представлення.

Показник оцінки якості кластеризації – індекс силуетів $SWC = 0{,}95$ свідчить про коректну кількість кластерів та розподілені дані у них. Порівняння стандартної кластеризації нейронної мережі Кохонена та вдосконаленого методу показало, що достовірність вдосконаленого методу підвищена орієнтовано на 10%, що свідчить про виконання поставленої мети роботи.

4 ЕКОНОМІЧНА ЧАСТИНА

В дому розділі здійснимо економічний аналіз з метою визначення економічного потенціалу розробки. Для цього проведемо та опишемо в розділі оцінку комерційного потенціалу та ефекту впровадження розробки, спрогнозуємо витрати на виконання наукової роботи та впровадження її результатів, обчислимо ефективність вкладених інвестицій.

4.1 Оцінювання комерційного потенціалу розробки

Метою проведення технологічного аудиту є оцінювання комерційного потенціалу розробки, створеної в результаті науково-технічної діяльності [33].

Результатом магістерської кваліфікаційної роботи є розробка програмного засобу призначеного для аналізу загроз інформації з використанням вдосконаленого методу інтелектуального аналізу даних з підвищеною достовірністю на основі математичного апарату нейронних мереж.

Для проведення технологічного аудиту залучено трьох незалежних експертів. У межах даної роботи такими експертами є викладачі кафедри МБІС: Карпінець В. В. (к.т.н., доцент каф. МБІС ВНТУ), Шиян А.А. (к.ф.-м.н, проф. каф. МБІС ВНТУ) та Грицак А. В. (доц., викл. каф. МБІС ВНТУ). Оцінювання комерційного потенціалу здійснимо за критеріями, що наведені в табл. 4.1.

Таблиця 4.1 – Критерії оцінювання комерційного потенціалу розробки

Критерії оцінювання та бали (за 5-ти бальною шкалою)					
Кри-терій	0	1	2	3	4
Технічна здійсненність концепції:					
1	Достовірність концепції не підтверджена	Концепція підтверджена експертними висновками	Концепція підтверджена розрахунками	Концепція перевірена на практиці	Перевірено роботоздатність продукту в реальних умовах
Ринкові переваги (недоліки):					
2	Багато аналогів на малому ринку	Мало аналогів на малому ринку	Кілька аналогів на вел. ринку	Один аналог на великому ринку	Продукт не має аналогів на великому ринку

Продовження таблиці 4.1

3	Ціна продукту значно вища за ціни аналогів	Ціна продукту дещо вища за ціни аналогів	Ціна продукту приблизно дорівнює цінам аналогів	Ціна продукту дещо нижче за ціни аналогів	Ціна продукту значно нижче за ціни аналогів
Критерії оцінювання та бали (за 5-ти бальною шкалою)					
Кри-тер.	0	1	2	3	4
4	Технічні та споживчі властивості продукту значно гірші, ніж в аналогів	Технічні та споживчі властивості продукту трохи гірші, ніж в аналогів	Технічні та споживчі властивості продукту на рівні аналогів	Технічні та споживчі властивості продукту трохи кращі, ніж в аналогів	Технічні та споживчі властивості продукту значно кращі, ніж в аналогів
5	Експлуатацій ні витрати значно вищі, ніж в аналогів	Експлуатацій ні витрати дещо вищі, ніж в аналогів	Експлуатацій ні витрати на рівні експлуатаційн их витрат аналогів	Експлуатацій ні витрати трохи нижчі, ніж в аналогів	Експлуатаційні витрати значно нижчі, ніж в аналогів
Ринкові перспективи					
6	Ринок малий і не має позитивної динаміки	Ринок малий, але має позитивну динаміку	Середній ринок з позитивною динамікою	Великий стабільний ринок	Великий ринок з позитивною динамікою
7	Активна конкуренція великих компаній на ринку	Активна конкуренція	Помірна конкуренція	Незначна конкуренція	Конкурентів немає
Практична здійсненність					
8	Відсутні фахівці як з технічної, так і з комерційної реалізації ідеї	Необхідно наймати фахівців або витрачати значні кошти та час на навчання наявних фахівців	Необхідне незначне навчання фахівців та збільшення їх штату	Необхідне незначне навчання фахівців	Є фахівці з питань як з технічної, так і з комерційної реалізації ідеї
9	Потрібні значні фінансові ресурси, які відсутні. Джерела фінансування ідеї відсутні	Потрібні незначні фінансові ресурси. Джерела фінансування відсутні	Потрібні значні фінансові ресурси. Джерела фінансування є	Потрібні незначні фінансові ресурси. Джерела фінансування є	Не потребує додаткового фінансування

Продовження таблиці 4.1

10	Необхідна розробка нових матеріалів	Потрібні матеріали, що використовуються у військово-промисловому комплексі	Потрібні дорогі матеріали	Потрібні досяжні та дешеві матеріали	Всі матеріали для реалізації ідеї відомі та давно використовуються у виробництві
11	Термін реалізації ідеї більший за 10 років	Термін реалізації ідеї більший за 5 років. Термін окупності інвестицій більше 10-ти років	Термін реалізації ідеї від 3-х до 5-ти років. Термін окупності інвестицій більше 5-ти років	Термін реалізації ідеї менше 3-х років. Термін окупності інвестицій від 3-х до 5-ти років	Термін реалізації ідеї менше 3-х років. Термін окупності інвестицій менше 3-х років
12	Необхідна розробка регламентних документів та отримання великої кількості дозвільних документів на виробництво та реалізацію продукту	Необхідно отримання великої к-ті дозвільних документів на вир-во та реалізацію продукту, що вимагає значних коштів та часу	Процедура отримання дозвільних документів для виробництва та реалізації продукту вимагає незначних коштів та часу	Необхідно тільки повідомлення відповідним органам про виробництво та реалізацію продукту	Відсутні будь-які регламентні обмеження на виробництво та реалізацію продукту

Результати оцінювання комерційного потенціалу експертами розробки зведено в таблицю 4.2.

Таблиця 4.2 – Результати оцінювання комерційного потенціалу розробки

Критерії	Прізвище, ініціали, посада експерта		
	1 – Карпінець В.В.	2 – Шиян А.А.	3 – Грицак А.В.
1	4	4	4
Ринкові переваги (недоліки):			
2	3	3	4
3	3	3	3
4	3	3	4
5	4	4	4
Ринкові перспективи			
6	3	4	3
7	4	3	3
Практична здійсненність			
8	3	4	3

Продовження таблиці 4.2

9	3	4	4
10	3	4	3
11	3	3	3
12	4	4	4
Сума балів	$СБ_1 = 40$	$СБ_1 = 43$	$СБ_1 = 42$
Середньоарифметична сума балів $\overline{СБ}$	$\overline{СБ} = 41{,}6$		

За даними таблиці 4.2 можна зробити висновок, щодо рівня комерційного потенціалу розробки. Зважимо на результат й порівняємо його з рівнями комерційного потенціалу розробки, що представлено в таблиці 4.3.

Таблиця 4.3 – Рівні комерційного потенціалу розробки

Середньоарифметична сума балів $\overline{СБ}$, розрахована на основі висновків експертів	Рівень комерційного потенціалу розробки
0 – 10	Низький
11 – 20	Нижче середнього
21 – 30	Середній
31 – 40	Вище середнього
41 – 48	Високий

Рівень комерційного потенціалу розробки, становить $41{,}6$ балів, що відповідає рівню «високий».

Проаналізуємо суть технічної проблеми та розглянемо аналоги. Наукова новизна розробки полягає в реалізації програмного засобу на основі вдосконаленого методу інтелектуального аналізу даних з підвищеною достовірністю на основі математичного апарату нейронних мереж.

Завданням інтелектуального аналізу даних є виявлення правил та закономірностей у наборах даних. Тривалий час основним інструментом інтелектуального аналізу даних була традиційна математична статистика, проте часто даний підхід не дозволяв вирішити завдання з реального життя.

Спочатку застосування нейронних мереж в інтелектуальному аналізі даних викликало скептичне відношення, через недоліки, властиві нейронним мережам: складна структура, погана інтерпретованість і тривалий час навчання. Однак їх переваги, такі як висока допустимість до зашумлених даних і низький коефіцієнт помилок, безперервне вдосконалення та оптимізація різних алгоритмів навчання

мереж, алгоритму вилучення правил, алгоритму спрощення мереж роблять нейронні мережі все більш і більш перспективним напрямом у data mining.

Саме тому використання нейронних мереж у технології інтелектуального аналізу даних є актуальним напрямком, який було досліджено та реалізовано на практиці у програмному засобі.

Враховуючи такі переваги розробленого методу, можемо порівняти його з аналогами (табл. 1.2).

У таблиці 4.4 наведені основні технічні показники аналога і нового програмного продукту.

Таблиця 4.4 – Основні технічні показники аналога SAS Enterprise Miner і нового програмного продукту

Показники, %	Аналог	Нова розробка	Відношення параметрів нової розробки до параметрів аналога
Функціональність	80	100	1,25
Надійність	75	100	1,33
Сумісність	85	100	1,18
Супровід	70	100	1,42
Економія ресурсів і часу	85	100	1,18
Простота використання	80	100	1,25

Системи інтелектуального аналізу даних на основі нейронних мереж лише набирають своєї популярності та дозволяють знаходити рішення актуальних задач аналізу великих даних на сьогодні.

В якості запропонованого вдосконалення методу інтелектуального аналізу даних з підвищеною достовірністю на основі математичного апарату нейронних мереж було застосовано кластеризацію та передбачено, що розроблювана нейронна мережа навчається без вчителя та реалізує алгоритм кластеризації.

На сьогодні, одними з вимог, які пред'являються інтелектуальним системам здобуття знань, є ефективність і масштабованість. Робота з великими базами даних вимагає ефективності алгоритмів, а неточність і часто неповнота

даних породжують додаткові проблеми для отримання прихованих закономірностей. Нейронні мережі мають перевагу, оскільки вони є ефективним засобом роботи із зашумленими даними, що зумовлює їх значну перевагу порівняно з аналогами.

В представленому програмному засобі процес аналізу даних (data mining) представлений трьома основними фазами: підготовка даних; отримання правил; оцінки правил.

Розроблюваний додаток реалізує функції імпорту, обробки, візуалізації та експорту даних. Він може функціонувати і без сховища, отримуючи інформацію з будь-яких інших джерел, але найбільш оптимальним є їхнє спільне використання.

Головна перевага методів Data Mining полягає в тому, що вони поєднують у собі як математичний інструментарій (починаючи від статистичного аналізу та закінчуючи новітніми евристичними методами), так і останні розуміння в галузі IT. У методах інтелектуального аналізу даних комплексно використовуються формалізовані та неформальні способи та техніки аналітики, різні способи аналізу даних.

Робота користувача з додатком полягає у візуальній взаємодії та керуванні сценарієм. Сценарій – послідовність дій, дозволяють отримати з даних знання.

Таким чином, технічне рішення, що пропонується в роботі, матиме високі показники порівняно з аналогами та більшою мірою задовольнить потреби споживачів. Тому його розробка та впровадження є актуальним та доцільним.

Програмний засіб на сьогодні має перспективу та користь як для пересічних користувачів так і для комерційних. Для реалізованого проекту є характерним можливість надання якісних послуг з технічної підтримки на всіх етапах використання програмного продукту користувачами.

Засоби розповсюдження розробки: інтернет-магазини програмного забезпечення, відповідні сайти, поширення цільових рекламних матеріалів на них, встановлення середньої ціни задля збільшення попиту та залучення нових користувачів.

Оскільки, сьогодні питання якісного аналізу великих даних для забезпечення інформаційної безпека є актуальними – цілком ймовірний високий попит на запропонований програмний продукт.

4.2 Прогнозування витрат на виконання наукової роботи та впровадження її результатів

Прогнозування витрат на виконання науково-дослідної, дослідно-конструкторської та конструкторсько-технологічної роботи складається з таких етапів:

1-й етап: розрахунок витрат, які безпосередньо стосуються виконавців даного розділу роботи;

2-й етап: розрахунок загальних витрат на виконання даної роботи;

3-й етап: прогнозування загальних витрат на виконання та впровадження результатів даної роботи.

Виконаємо розрахунок витрат, які безпосередньо стосуються виконавців даного розділу роботи, за такими статтями та формулами, приймаючи до уваги те, що для розробки інформаційної технології було залучено одного розробника програмного забезпечення.

1. Основна заробітна $З_о$:

$$З_о = \frac{М}{Т_р} \cdot t, \text{грн.} \qquad (4.1)$$

де $М$ – місячний посадовий оклад – $20\,000$ грн.;

$Т_р$ – число робочих днів в місяці; приблизно $Т_р = 24$ дні;

t – число робочих днів роботи – 36 дні.

Таким чином:

$$З_о = \frac{20\,000}{24} \cdot 36 = 30\,000 \text{ (грн.)}$$

Таблиця 4.5 – Витрати по заробітній платі

Найменування посади	Місячний посадовий оклад, грн.	Оплата за робочий день, грн.	Число днів роботи	Витрати на заробітну плату
Розробник	20 000	833,3	36	30 000
Всього				30 000

2. Додаткова заробітна плата $З_д$ працівників розраховується як 12% від основної заробітної плати:

$$З_д = 0{,}12 \cdot 30\,000 \;\; = 3\,600\,(\text{грн.}) - \text{для розробника}$$

3. Нарахування на заробітну плату $Н_{зп}$ розробника становить:

$$Н_{зп} = (З_о + З_д) \cdot \frac{\beta}{100} \qquad (4.2)$$

де $З_о$ – основна заробітна плата розробника;

$З_д$ – додаткова заробітна плата розробника;

β – ставка єдиного внеску на загальнообов'язкове державне соціальне страхування – 22%.

$$Н_{зп} = (30\,000 \; + 3\,600) \cdot 0{,}22 = 7\,392\,(\text{грн.})$$

4. Амортизація обладнання, комп'ютерів та приміщень, які використовувались під час виконання даного етапу роботи розраховуємо за формулою:

$$А = \frac{Ц \cdot Т}{12 \cdot Т_В} \qquad (4.3)$$

де Ц – загальна балансова вартість обладнання, приміщення тощо, грн.;

Т – фактична тривалість використання, міс;

$Т_В$ – термін використання обладнання, приміщень тощо, роки.

Розробка програмного забезпечення ведеться орієнтовно 2 місяці.

Для офісного приміщення $А = \frac{80\,000 \cdot 1{,}5}{12 \cdot 5} = 2\,000\text{грн.}$; для комп'ютера $А = \frac{12\,000 \cdot 1{,}5}{12 \cdot 2} = 750\text{грн.}$; для монітора $А = \frac{5\,500 \cdot 1{,}5}{12 \cdot 2} = 343{,}8\text{грн.}$

Розрахунки зведено до таблиці 4.6:

Таблиця 4.6 – Амортизаційні відрахування

Наймену-вання	Балансова вартість (грн.)	Термін використання (років)	Фактична тривалість в-ня, (міс.)	Величина ам.. відрахувань, (грн..)
Офісне приміщення	80 000	5	1,5	2 000
Комп'ютер	12 000	2	1,5	750
Монітор	5 500	2	1,5	343,8
			Всього	3 093,8

5. Витрати на комплектуючі К, що були використані під час виконання даного етапу роботи, розраховуються за формулою:

$$К = \sum_{1}^{n} Н_i \cdot Ц_i \cdot К_i \text{ (грн.)} \quad (4.4)$$

де $Н_i$ – кількість комплектуючих *i*-го виду, шт.;

$Ц_i$– ціна комплектуючих *i*-го виду, грн.;

$К_i$ – коефіцієнт транспортних витрат, $К_i$ = (1,1…1,15);

n – кількість видів комплектуючих.

Таблиця 4.7 – Витрати на комплектуючі

Найменування комплектувальних	Кількість	Ціна за штуку, грн.	Сума, грн.	Примітка
Клавіатура	1	500 грн.	500 грн.	
Комп'ютерна мишка	1	350 грн.	350 грн.	
Всього:			$К_i = 1,2,$	850 грн.

6. Витрати на силову електроенергію $В_е$ розраховуються за формулою:

$$В_е = В \cdot П \cdot Ф \cdot К_п \text{ (грн.)} \quad (4.5)$$

де В – вартість 1 кВт-год. (на сьогодні для підприємців вартість 3,45грн./кВт-год);

П – установлена потужність обладнання – 0,8 кВт;

Ф – фактична кількість годин роботи обладнання – 360 годин,

$К_п$ – коефіцієнт використання потужності.

$$В_е = 6,4 \cdot 0,8 \cdot 360 \cdot 0,14 = 258 \text{ (грн).}$$

7. Інші витрати $В_{ін}$ охоплюють:

– витрати на управління організацією;

– оплату службових відряджень;

– витрати на утримання, ремонт та експлуатацію основних засобів;

– витрати на опалення, освітлення, водопостачання, охорону праці тощо.

Інші витрати $В_{ін}$ можна прийняти як 100% від суми основної заробітної плати розробника:

$$В_{ін} = 30\,000 \cdot 1 = 30\,000 \text{ (грн)}$$

Послуги Інтернету – 300 грн., канцтовари – 400 грн. Загальна вартість становить:

$$300 + 400 = 700 \text{ (грн.)}$$

8. Сума всіх попередніх статей витрат дає витрати на виконання даної частини роботи – В.

$$В = 30\,000 + 3\,600 + 7\,392 + 3\,093{,}8 + 850 + 258 + 30\,000 + 700$$
$$= 75\,893{,}8 \text{ (грн.)}$$

9. Проведемо прогнозування загальних витрат ЗВ на виконання та впровадження результатів виконаної наукової роботи. Прогнозування здійснюється за формулою:

$$ЗВ = \frac{В_{заг}}{\beta}, \text{грн.} \qquad (4.6)$$

де β – коефіцієнт, який характеризує етап (стадію) виконання даної роботи.

Так, якщо розробка знаходиться:

– на стадії науково-дослідних робіт, то $\beta \approx 0{,}1$;

– на стадії технічного проектування, то $\beta \approx 0{,}2$;

– на стадії розробки конструкторської документації, то $\beta \approx 0{,}3$;

– на стадії розробки технологій, то $\beta \approx 0{,}4$;

– на стадії розробки дослідного зразка, то $\beta \approx 0{,}5$;

– на стадії розробки промислового зразка, $\beta \approx 0{,}7$;

– на стадії впровадження, то $\beta \approx 0{,}9$.

$В_{заг}$ – загальна вартість всієї наукової роботи.

$$В = 75\,893{,}8\ (грн.)$$

$$ЗВ = \frac{75\,893{,}8}{0{,}7} = 108\,419{,}7\ (грн.)$$

Отже, прогноз загальних витрат ЗВ на виконання та впровадження результатів виконаної наукової роботи складає $108\,419{,}7$ (грн.)

4.3 Прогнозування комерційних ефектів від реалізації результатів розробки

У даному підрозділі проведемо кількісне прогнозування, яку вигоду, зиск можна отримати у майбутньому від впровадження результатів виконаної наукової роботи.

В умовах ринку узагальнюючим позитивним результатом, що його отримує підприємство від впровадження результатів тієї чи іншої розробки, є збільшення чистого прибутку підприємства. Зростання чистого прибутку можна оцінити у теперішній вартості грошей.

Зростання чистого прибутку забезпечить інвестору надходження додаткових коштів, які дозволять покращити фінансові результати діяльності.

Виконання даної наукової роботи та впровадження її результатів складає приблизно 1 рік. Позитивні результати від впровадження розробки очікуються вже в перші місяці після впровадження.

Проведемо детальне прогнозування позитивних результатів та кількісне їх оцінювання по роках.

Обчислимо збільшення чистого прибутку підприємства $\Delta П_i$ для кожного із років, протягом яких очікується отримання позитивних результатів від впровадження розробки, розраховується за формулою:

$$\Delta П_i = \sum_{1}^{n} (\Delta П_я \cdot N + П_я \cdot \Delta N)_i \qquad (4.7)$$

де $\Delta П_я$ – покращення основного якісного показника від впровадження результатів розробки у даному році;

N – основний кількісний показник, який визначає діяльність підприємства у даному році до впровадження результатів наукової розробки;

ΔN – покращення основного кількісного показника діяльності підприємства від впровадження результатів розробки;

$П_я$ – основний якісний показник, який визначає діяльність підприємства у даному році після впровадження результатів наукової розробки;

n – кількість років, протягом яких очікується отримання позитивних результатів від впровадження розробки.

Припустимо, що внаслідок впровадження результатів наукової розробки чистий прибуток підприємства збільшиться на 140 грн., а кількість одиниць реалізованої послуги збільшиться:

– протягом першого року – на 520 од.,

– протягом другого року – ще на 680 од.,

– протягом третього року – ще на 830 од.

Орієнтовно: реалізація продукції до впровадження результатів наукової розробки складала 1 шт., а прибуток, що його отримувало підприємство на одиницю продукції до впровадження результатів наукової розробки – 110 грн.

Потрібно спрогнозувати збільшення чистого прибутку підприємства від впровадження результатів наукової розробки у кожному році відносно базового.

Збільшення чистого прибутку підприємства $\Delta П_1$ протягом першого року складе:

$$\Delta П_1 = 140 \cdot 1 + (140 + 110) \cdot 520 = 118\,309 \text{ (грн.)}$$

Обчислимо збільшення чистого прибутку підприємства $\Delta П_2$ протягом другого року:

$$\Delta П_2 = 150 \cdot 1 + (140 + 110) \cdot (520 + 680) = 248\,049{,}6 \text{ (грн.)}$$

Збільшення чистого прибутку підприємства $\Delta П_3$ протягом третього року становитиме:

$$\Delta П_3 = 150 \cdot 1 + (140 + 110) \cdot (520 + 680 + 830) = 381\,397{,}4 \text{ (грн.)}$$

Отже, розрахунки показують, що відповідно прогнозуванню комерційний ефект від впровадження розробки виражається у значному збільшенні чистого прибутку підприємства.

4.4 Розрахунок ефективності вкладених інвестицій та періоду їх окупності

Основними показниками, які визначають доцільність фінансування наукової розробки певним інвестором, є абсолютна і відносна ефективність вкладених інвестицій та термін їх окупності.

Розрахунок ефективності вкладених інвестицій передбачає:

1-й крок. Розрахунок теперішньої вартості інвестицій PV, що вкладаються в наукову розробку.

Такою вартістю ми можемо вважати прогнозовану величину загальних витрат ЗВ на виконання та впровадження результатів НДДКР, тобто $ЗВ = PV = 108\,252{,}7 \text{ (грн.)}$

2-й крок. Розрахуємо очікуване збільшення прибутку $\Delta П_i$, що його отримає підприємство (організація) від впровадження результатів наукової розробки, для кожного із років, починаючи з першого року впровадження. Таке збільшення прибутку також було розраховане нами раніше та становить:

$$\Delta П_1 = 118\,309 \text{ (грн)}, \Delta П_2 = 248\,049{,}6 \text{ (грн)}, \Delta П_3 = 381\,397{,}4 \text{(грн)}.$$

3-й крок. Будуємо вісь часу, на якій відображаємо всі платежі (інвестиції та прибутки), що мають місце під час виконання науково-дослідної роботи та впровадження її результатів.

Рисунок 4.1 характеризує рух платежів (інвестицій та додаткових прибутків).

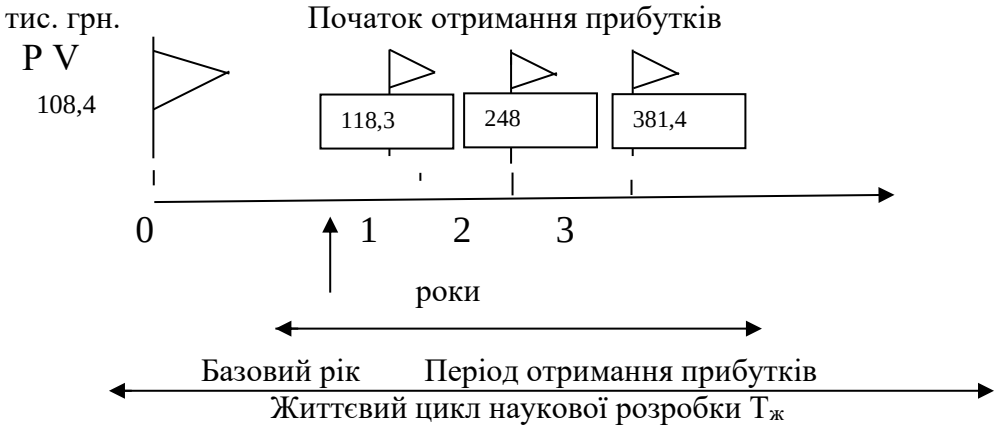


Рисунок 4.1 – Вісь часу з фіксацією платежів, що мають місце під час розробки та впровадження результатів НДДКР

4-й крок. Розрахуємо абсолютну ефективність вкладених інвестицій $Е_{абс}$ за формулою:

$$Е_{абс} = (ПП - PV), (грн.) \quad (4.8)$$

де ПП – приведена вартість всіх чистих прибутків, що їх отримає підприємство (організація) від реалізації результатів наукової розробки, грн.;

PV – теперішня вартість інвестицій $PV = ЗВ$, грн.

Приведена вартість всіх чистих прибутків ПП розраховується за формулою:

$$ПП = \sum_{1}^{Т} \frac{\Delta П_i}{(1+\tau)^t}, (грн) \quad (4.9)$$

де $\Delta П_i$ – збільшення чистого прибутку у кожному із років, протягом яких виявляються результати виконаної та впровадженої НДДКР, грн.;

Т – період часу, протягом якого виявляються результати впровадженої НДДКР, роки;

τ – ставка дисконтування, за яку можна взяти щорічний прогнозований рівень інфляції в країні – 0,1;

t – період часу (в роках) від моменту отримання чистого прибутку до точки «0»;

$$ПП = \frac{118\ 309}{(1+0{,}1)^1} + \frac{248\ 049{,}6}{(1+0{,}1)^2} + \frac{381\ 397{,}4}{(1+0{,}1)^3} = 747\ 756{,}1\ (\text{грн.})$$

$$Е_{абс} = 747\ 756{,}1 - 108\ 419{,}7 = 639\ 336{,}4\ (\text{грн.})$$

Оскільки $Е_{абс} > 0$, результат від проведення наукових досліджень щодо розробки програмного продукту та їх впровадження принесе прибуток, тобто є доцільним, проте це ще не свідчить про те, що інвестор буде зацікавлений у фінансуванні даної програми.

5-й крок. Розрахуємо відносну (щорічну) ефективність вкладених в наукову розробку інвестицій $Е_в$ за формулою:

$$Е_в = \sqrt[Т_ж]{1 + \frac{Е_{абс}}{PV} - 1} \qquad (4.10)$$

де $Е_{абс}$ – абсолютна ефективність вкладених інвестицій, грн.;

PV – теперішня вартість інвестицій $PV = ЗВ$, грн.

$Т_ж$ – життєвий цикл наукової розробки, роки.

$$E_в = \sqrt[3]{1 + \frac{639\ 336{,}4}{108\ 419{,}7}} - 1 = \sqrt[3]{6{,}9} - 1 = 0{,}82 \text{ або } 82\%$$

Порівняємо $E_в$ з мінімальною (бар'єрною) ставкою дисконтування τ_{min}, яка визначає ту мінімальну дохідність, нижче за яку інвестиції вкладатися не будуть.

Спрогнозуємо величину τ_{min}.

У загальному вигляді мінімальна (бар'єрна) ставка дисконтування τ_{min} визначається за формулою:

$$\tau_{min} = d + f \qquad (4.11)$$

де d – середньозважена ставка за депозитними операціями в комерційних банках; $d = 0{,}2$;

f – показник, що характеризує ризикованість вкладень; величина $f = 0{,}5$.

$$\tau_{min} = 0{,}2 + 0{,}5 = 0{,}7$$

Оскільки $E_в = 82\% > \tau_{min} = 70\%$, то у інвестора є потенційна зацікавленість у фінансуванні даної наукової розробки.

6-й крок. Розрахуємо термін окупності вкладених у реалізацію наукового проекту інвестицій $Т_{ок}$ за формулою:

$$Т_{ок} = \frac{1}{Е_в}, \text{рік} \qquad (4.12)$$

$$Т_{ок} = \frac{1}{0{,}82} = 1{,}2 \text{ (року)}$$

Оскільки термін окупності вкладених у реалізацію наукового проекту інвестицій менше трьох років ($Т_{ок} < 3$ років), то фінансування нової розробки є доцільним.

4.5 Висновки до розділу

В даному розділі було виконано оцінювання комерційного потенціалу розробки програмного засобу програмного засобу призначеного для аналізу загроз інформації з використанням вдосконаленого методу інтелектуального аналізу даних з підвищеною достовірністю на основі математичного апарату нейронних мереж.

Проведено технологічний аудит з залученням трьох незалежних експертів. Визначено, що рівень комерційного потенціалу розробки вище середнього. Проведено порівняння з аналогом.

Згідно з проведеним оцінюванням нова розробка є якісною та конкурентоспроможною.

Рівень комерційного потенціалу розробки, становить 41,6 бала, що відповідає рівню «високий».

Згідно із розрахунками всіх статей витрат на виконання науково-дослідної, дослідно-конструкторської та конструкторсько-технологічної роботи загальні витрати на розробку складають 108 419,7 (грн.). Розрахована абсолютна ефективність вкладених інвестицій в сумі 639 336,4 (грн.) свідчить про отримання прибутку інвестором від комерціалізації програмного продукту.

Щорічна ефективність вкладених в наукову розробку інвестицій складає 82%, що вище за мінімальну бар'єрну ставку дисконтування, яка складає 70%. Це означає потенційну зацікавленість інвесторів у фінансуванні розробки.

Термін окупності вкладених у реалізацію проекту інвестицій становить 1,2 (року), що також свідчить про доцільність фінансування нової розробки.

Отже, проаналізувавши отримані економічні показники, можна вважати, що запропонована розробка програмного засобу має високий комерційний потенціал, а тому є доцільною для подальшого впровадження.

ВИСНОВОК

В магістерській кваліфікаційній роботі описано розробку та реалізацію вдосконаленого методу виявлення загроз веб-ресурсу на основі інтелектуального аналізу даних із використанням нейронних мереж та кластеризації.

Після аналізу теоретичної бази та досліджень у даній галузі, для вирішення поставлених задач, в роботі було здійснено наступне.

В першому розділі роботи було проведено теоретичний огляд галузі та досліджено поняття інтелектуального аналізу як методу обробки даних; здійснено аналіз методів data mining та огляд існуючих аналогів їх реалізацій; проведено аналіз можливості застосування нейронних мереж для здійснення інтелектуального аналізу даних.

Другий розділ роботи описує вдосконалення методу інтелектуального аналізу даних на основі нейронної мережі та застосування кластеризації. В якості модифікації було запропоновано рішення, що передбачає навчання на основі карти Кохонена, проведення першого етапу кластеризації, а потім продовження дослідження за рахунок аналізу відстаней в кластері та міжкластерних відстаней із використанням метрики Евклідової відстані. Суть ідеї полягає в тому, що після певної кількості сформованих кластерів, продовжимо аналіз неврахованих кластерів, що не містять нейронів, а тому не були враховані до основної вибірки.

В ході написання розділу було здійснено розробку моделі аналізу даних на основі нейронної мережі, запропоновано та описано вдосконалення методу для аналізу даних, розроблено алгоритм роботи програми на основі вдосконаленого методу, визначено вхідні зміні та результат моделі, обґрунтовано вибір засобів розробки.

В третьому розділі роботи здійснено практичну розробку додатку: здійснено проектування користувацького інтерфейсу, описано особливості здійсненої програмної реалізації, описано інструкцію користувача та проведено результат аналізів інтелектуального аналізу даних.

Аналіз результатів реалізованого додатку, що є демонстрацією вдосконаленого алгоритму для визначення наявності кіберзагроз інформаційної системи показав, що програма коректно виконує функції навчання нейронної мережі, інтелектуальний аналіз на основі навченої мережі, генерує відповідний результат та надає його графічне представлення.

Показник оцінки якості кластеризації – індекс силуетів $SWC = 0{,}95$ свідчить про коректну кількість кластерів та розподілені дані у них. Порівняння стандартної кластеризації нейронної мережі Кохонена та вдосконаленого методу показало, що достовірність вдосконаленого методу підвищена орієнтовано на 10%, що свідчить про виконання поставленої мети роботи.

В четвертому розділі було виконано оцінювання комерційного потенціалу розробки програмного засобу. Визначено, що рівень комерційного потенціалу розробки, становить 41,6 бала, що відповідає рівню «високий». Термін окупності вкладених у реалізацію проекту інвестицій становить 1,2 (року), що також свідчить про доцільність фінансування нової розробки. Отже, запропонована розробка програмного засобу має високий комерційний потенціал, а тому є доцільною для подальшого впровадження.

Нейронна мережа, реалізована та навчена відповідно до вдосконаленого методу, дозволяє аналізувати поведінку користувачів веб-ресурсу на основі даних про їх дії, отриманих з таблиці журналу, що заповнюється підсистемою аудиту безпеки. Її застосування дозволить оперативно виявляти потенційних зловмисників, і навіть підвищить точність класифікації поведінки користувачів за рахунок інтелектуального аналізу даних.

Отже, результати виконання проведеної роботи свідчать про виконання поставленої мети, а саме – вдосконалення методу інтелектуального аналізу даних з підвищеною достовірністю на основі математичного апарату нейронних мереж для виявлення кіберзагроз для веб-ресурсу.

ПЕРЕЛІК ПОСИЛАНЬ

1. Securing web applications: top OWASP threats and what to do about them *Analytics:* веб-сайт. URL: https://www.ptsecurity.com/ww-en/analytics/knowledge-base/securing-webapplications-top-owasp-threats-and-what-to-do-about-them/ (дата звернення: 17.10.2022)

2. Захист веб-додатків, чому це важливо? *Itbiz:* веб-сайт. URL: https://itbiz.ua/statti-ta-obzori/zaxist-veb-dodatkiv-chomu-ce-vazhlivo/ (дата звернення: 17.10.2022)

3. Секрети кібербезпеки. *Sekrety-kiberbezpeky:* веб-сайт. URL: https://eba.com.ua/sekrety-kiberbezpeky-5-klyuchovyh-pravyl-bezpeky-v-interneti/ (дата звернення: 17.10.2022)

4. Загрози The Security Risk Management Guide. *Microsoft Corporation:* веб-сайт. URL: http://www.microsoft.com/technet/ security/topics/ (дата звернення: 17.10.2022)

5. Остапов С. Е. технологія захисту інформації : навчальний посібник / С. Е. Остапов, С. П. Євсеєв, О. Г. Король. – Х. : Вид. ХНЕУ, 2013. – 476 с.

6. Автоматизація виявлення та протидії інформаційним загрозам у мережі Інтернет / Р. В. Грищук [та інш.] // Наукоємні технології. – 2018. – № 2. – С. 158–165.

7. Черевко, О. В. Теоретичні засади поняття інформаційної безпеки та класифікація загроз системі інформаційного захисту / О. В. Черевко // Ефективна економіка. – 2014. – № 5.

8. Актуальні проблеми Data Mining : навчальний посібник для студентів факультету комп’ютерних наук та кібернетики / О. О. Марченко, Т. В. Россада. К. : КНУ ім. Т. Шевченка, 2017. – 150 с.

9. Інтелектуальний аналіз даних : підручник / О. І. Черняк, П. В. Захарченко. – К. : Знання, 2014. – 599 с.

10. Data Mining : пошук знань в даних / А. Я. Гладун, Ю. В. Рогушина. – К.: ТОВ ВД АДЕФ-Україна, 2016. – 452 с.

11. Бахрушин В. Є. Методи аналізу даних : навчальний посібник / В. Є. Бахрушин. – Запоріжжя : КПУ, 2011. – 268 с

12. Субботін, С. О. Нейронні мережі : навчальний посібник / С. О. Субботін, А. О. Олійник ; під заг. ред. проф. С. О. Субботіна. – Запоріжжя : ЗНТУ, 2014. – 132 с.

13. Інформаційні технології. Словник термінів. Частина 34. Штучний інтелект. Нейронні мережі (EN ISO/IEC 2382-34:1999, IDT) : ДСТУ ISO/IEC 2382-34-2003. – [Чинний від 2004-10-01]. – К. : Держспоживстандарт, 2005. – 20 с. – (Національний стандарт України).

14. Руденко, О. Г. Штучні нейронні мережі / О. Г. Руденко, Є. В. Бодянський. – Харків : Компанія СМІТ, 2006. – 404 с.

15. Нейронні мережі – шлях до глибинного навчання. *Codeguida:* веб-сайт. URL: https://codeguida.com/post/739 (дата звернення 20.10.2022)

16. Нейронні мережі: їх застосування, робота. *Журнал:* веб-сайт. URL: https://www.poznavayka.org/uk/nauka-i-tehnika-2/ (дата звернення 20.10.2022)

17. Що таке нейронні мережі та як вони працюють? Класифікація штучних нейромереж. *LivingFo:* веб-сайт. URL: https://livingfo.com/shcho-take-nejronni-merezhi-ta-iak-vony-pratsiuiut/ (дата звернення 20.10.2022).

18. SAS Enterprise Miner. *Офіційний сайт:* веб-сайт. URL: https://www.sas.com/ru_ua/software/enterprise-miner.html (дата звернення 20.10.2022)

19. STATISTICA Data Miner. *Стаття:* веб-сайт. URL: https://www.researchgate.net/publication/326265121_Bagatovimirnij_analiz_danih_u_sistemi_STATISTICA_Metodicni_vkazivki_dla_vikonanna_prakticnih_zavdan_z_kursu_Informacijno-analiticna_dialnist_u_miznarodnih_vidnosinah (дата звернення 20.10.2022)

20. Oracle Data Mining. *Офіційний сайт:* веб-сайт. URL: https://www.oracle.com/big-data/technologies/dataminer/ (дата звернення 20.10.2022)

21. KXEN Analytic Framework. *Software:* веб-сайт. URL: https://www.the-data-mine.com/Software/KXENAnalyticFramework (дата звернення 20.10.2022)

22. Microsoft SQL Server Analysis Services. *Microsoft:* веб-сайт. URL: https://learn.microsoft.com/en-us/analysis-services/ (дата звернення 20.10.2022)

23. Програмні продукти SPSS. *Стаття:* веб-сайт. URL: https://msn.khnu.km.ua/pluginfile.php/567764/ (дата звернення 20.10.2022)

24. Загальна характеристика задач кластерного аналізу. *Стаття:* веб-сайт. URL: https://sites.google.com/site/ne4itkalogika/necitka-klasterizacia/ (дата звернення 20.10.2022)

25. Методи штучного інтелекту. *Посібник:* веб-сайт. URL: https://library.krok.edu.ua/media/library/ (дата звернення 20.10.2022)

26. Godfried Toussaint, "The Euclidean algorithm generates traditional musical rhythms, "Proceedings of BRIDGES: Mathematical Connections in Art, Music, and Science, Banff, Alberta, Canada, July 31 to August 3, 2005, pp. 47 – 56.

27. Кластеризація за допомогою карт Кохонена. *Stud:* веб-сайт. URL: https://stud.com.ua/140001/informatika/klasterizatsiyi_dopomogoyu_kart_kohonena (дата звернення 20.10.2022)

28. Класифікація кластер-процедур. *Стаття:* веб-сайт. URL: http://ebooks.git-elt.hneu.edu.ua/babap/4-1-id4-1.html (дата звернення 20.10.2022)

29. C# programming language. *Microsoft:* веб-сайт. URL: https://learn.microsoft.com/en-us/dotnet/ (дата звернення 20.10.2022)

30. Visual Studio. *Microsoft:* веб-сайт. URL: https://visualstudio.microsoft.com (дата звернення: 22.10.2022).

31. WPF Designer. *Microsoft:* веб-сайт. URL: https://support.microsoft.com/uk-ua/ (дата звернення: 22.10.2022).

32. Оцінка якості кластеризації. *Stud:* веб-сайт. URL: https://stud.com.ua/139997/ (дата звернення: 10.11.2022).

33. Методичні вказівки до виконання економічної частини магістерських кваліфікаційних робіт / Уклад. : В. О. Козловський, О. Й. Лесько, В. В. Кавецький. – Вінниця : ВНТУ, 2021. – 42 с.

ДОДАТКИ

Додаток А. Технічне завдання

Вінницький національний технічний університет
Факультет менеджменту та інформаційної безпеки
Кафедра менеджменту та безпеки інформаційних систем

ЗАТВЕРДЖУЮ
Голова секції "Управління інформаційною безпекою" кафедри МБІС
д.т.н., професор
Юрій ЯРЕМЧУК
"24" вересня 2022 р.

ТЕХНІЧНЕ ЗАВДАННЯ

до магістерської кваліфікаційної роботи на тему:

Вдосконалений метод інтелектуального аналізу даних з підвищеною достовірністю на основі математичного апарату нейронних мереж

08-72.МКР.008.00.000.ТЗ

Керівник магістерської кваліфікаційної роботи
д.т.н., проф. Юрій Яремчук

Вінниця – 2022 р.

1. Найменування та область застосування

Вдосконалений метод інтелектуального аналізу даних з підвищеною достовірністю на основі математичного апарату нейронних мереж. Область застосування: Виявлення наявності загроз для веб-ресурсів.

2.Підстава для розробки

Розробка виконується на основі наказу ректора ВНТУ №203 від 14. 09. 2022 р.

3. Мета та призначення розробки

3.1 Мета розробки: розробка вдосконаленого методу інтелектуального аналізу даних з підвищеною достовірністю на основі математичного апарату нейронних мереж для виявлення загроз веб-ресурсу.

3.2 Призначення: розроблений програмний засіб виконує виявлення наявності загроз для веб-ресурсів.

4. Джерела розробки

4. 1. Securing web applications: top OWASP threats and what to do about them Analytics: веб-сайт. URL: https://www.ptsecurity.com/ww-en/analytics/knowledge-base/securing-webapplications-top-owasp-threats-and-what-to-do-about-them/ (дата звернення: 17.10.2022)

4.2. Актуальні проблеми Data Mining : навчальний посібник для студентів факультету комп’ютерних наук та кібернетики / О. О. Марченко, Т. В. Россада. К. : КНУ ім. Т. Шевченка, 2017. – 150 с.

4.3. Олійник, А. О. Інтелектуальний аналіз даних : навч. посіб. / А. О. Олійник, С. О. Субботін, О. О. Олійник. – Запоріжжя : ЗНТУ, 2011. – 271 с.

4.4. Субботін, С. О. Нейронні мережі : навчальний посібник / С. О. Субботін, А. О. Олійник ; під заг. ред. проф. С. О. Субботіна. – Запоріжжя : ЗНТУ, 2014. – 132 с.

5. Вимоги до програми

5.1 Вимоги до функціональних характеристик:

5.1.1 Програмний засіб повинен мати зручний, легкий у використанні інтерфейс користувача;

5.1.2 Реалізація методу не повинна вимагати спеціальних ліцензійних програмних додатків;

5.1.3 Програмний засіб повинен виконувати виявлення наявності загроз для веб-ресурсу на основі інтелектуального аналізу.

5.2 Вимоги до надійності:

5.2.1 Програмний засіб повинен працювати без помилок, у випадку виникнення критичних ситуацій необхідно передбачити виведення відповідних повідомлень;

5.2.2 Бази даних повинні бути налаштовані на автоматичне створення резервних копій;

5.2.3 Програмний засіб повинен виконувати свої функції.

5.3 Вимоги до складу і параметрів технічних засобів: процесор – Pentium 1500 МГц і подібні до них; оперативна пам’ять – не менше 512 Mb; середовище функціонування – операційна система сімейство Windows; вимоги до техніки безпеки при роботі з програмою повинні відповідати існуючим вимогам та стандартам з техніки безпеки при користуванні комп’ютерною технікою.

6. Вимоги до програмної документації

6.1 Обов’язкова поетапна інструкція для майбутніх користувачів, наведена у пункті 3.3

7. Вимоги до технічного захисту інформації

7.1 Необхідно забезпечити виявлення наявності загроз для веб-ресурсу на основі інтелектуального аналізу.

7.2 Неможливість отримання доступу незареєстрованих користувачів до інформаційних ресурсів.

8. Техніко-економічні показники

8.1 Цінність результатів використання даного проекту повинна перевищувати витрати на його реалізацію.

8.2 Має бути реалізований таким чином, щоб підходити для використання широкого загалу.

9. Стадії та етапи розробки

№ з/п	Назва етапів магістерської кваліфікаційної роботи	Початок	Закінчення
1	Визначення напрямку магістерської роботи, формулювання теми	15.09.2022	30.09.2022
2	Аналіз предметної області обраної теми	01.10.2022	10.10.2022
3	Апробація отриманих результатів	11.10.2022	15.10.2022
4	Розробка алгоритму роботи	16.10.2022	31.10.2022
5	Написання магістерської роботи на основі розробленої теми	01.11.2022	15.11.2022
6	Розробка економічної частини	15.11.2022	23.11.2022
7	Передзахист магістерської кваліфікаційної роботи	24.11.2022	25.11.2022
8	Виправлення, уточнення, корегування магістерської кваліфікаційної роботи	26.11.2022	15.11.2022
9	Захист магістерської кваліфікаційної роботи	19.12.2022	21.12.2022

10. Порядок контролю та прийому

10.1 До приймання магістерської кваліфікаційної роботи надається:

– ПЗ до магістерської кваліфікаційної роботи;

– програмний додаток;

– презентація;

– відзив керівника роботи;

– відзив опонента

Технічне завдання до виконання прийняв _____ Б.В. Панасюк

Додаток Б. Лістинг навчання нейромережі

```
using Microsoft.VisualBasic.FileIO;
using System;
using System.IO;
using System.Windows;
using Microsoft.Win32;
using DataAnalyze.Models;
using DataAnalyze.Context;
using Microsoft.VisualBasic.ApplicationServices;

namespace DataAnalyze
{
   public partial class AddTraining : Window
   {
      public AddTraining()
      {
         InitializeComponent();
      }
      private RecognizeContext _context = new RecognizeContext();

      private void Button_Click_2(object sender, RoutedEventArgs e)
      {
            OpenFileDialog dlg = new OpenFileDialog();
         dlg.RestoreDirectory = true;

         if (dlg.ShowDialog() == true)
         {
            string selectedFileName = dlg.FileName;
            FileName.Text = selectedFileName;
         }
      }

      private void Button_Click_1(object sender, RoutedEventArgs e)
      {
         ProcessCv(FileName.Text);

         Byte[] bytes = File.ReadAllBytes(FileName.Text);
         String file = Convert.ToBase64String(bytes);


         var trainedFillets = new TrainedFillet()
         {
            File = file,
            Name = Path.GetFileName(FileName.Text),
            CreateDate = DateTime.Now,
         };
```

```
            _context.TrainedFillets.Add(trainedFillets);
            _context.SaveChanges();

            Helper.AddLog("Added file for train");
            MessageBox.Show("Файд успішно опрацьовано");
        }

        private void ProcessCv(string path)
        {
            using (TextFieldParser parser = new TextFieldParser(path))
            {
                parser.TextFieldType = FieldType.Delimited;
                parser.SetDelimiters(",");
                while (!parser.EndOfData)
                {
                    //Process row
                    string[] fields = parser.ReadFields();
                    foreach (string field in fields)
                    {
                        var i = 8;
                        //TODO: Process field
                    }
                }
            }
        }

        private void FileName_Copy_TextChanged(object sender,
System.Windows.Controls.TextChangedEventArgs e)
        {

        }
    }
}
```

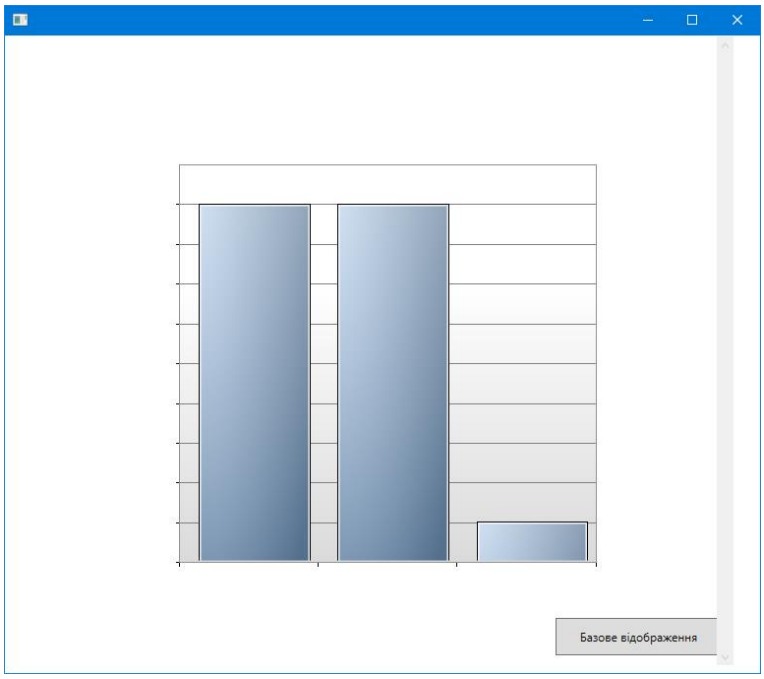
Додаток В. Лістинг опрацювання файлів

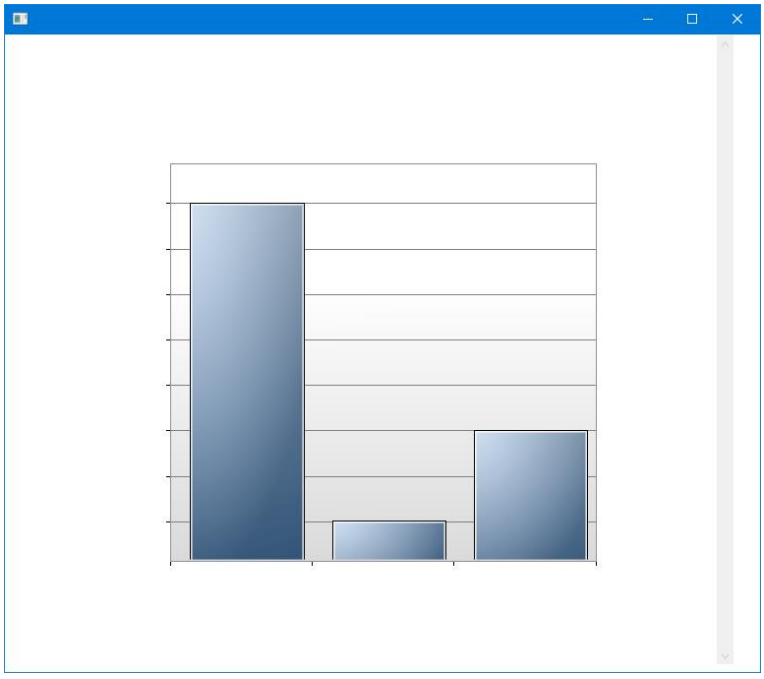
```
using System;
using System.Collections;
using System.Collections.Generic;
using System.Linq;
using System.Runtime.InteropServices.ComTypes;
using System.Windows;
using DataAnalyze.Models;
using Microsoft.VisualBasic.FileIO;
using Microsoft.Win32;
using WpfToolkitChart;

namespace DataAnalyze
{
   public partial class AddTrainingAnalytic : Window
   {
      public AddTrainingAnalytic()
      {
         InitializeComponent();
      }

      public CSVobject csvObject;
      public List<KeyValuePair<string, int>> csvObjectBlok;
      private void Button_Click(object sender, RoutedEventArgs e)
      {
         OpenFileDialog dlg = new OpenFileDialog();
         dlg.RestoreDirectory = true;

         if (dlg.ShowDialog() == true)
         {
            string selectedFileName = dlg.FileName;
            FileName.Text = selectedFileName;
         }
      }

      private void Button_Click_1(object sender, RoutedEventArgs e)
      {
         ProcessCv(FileName.Text);

         Helper.AddLog("Start analytic file");
         details.IsEnabled = true;
         detailsCommon.IsEnabled = true;
      }

      private void ProcessCv(string path)
      {
         //todo
```

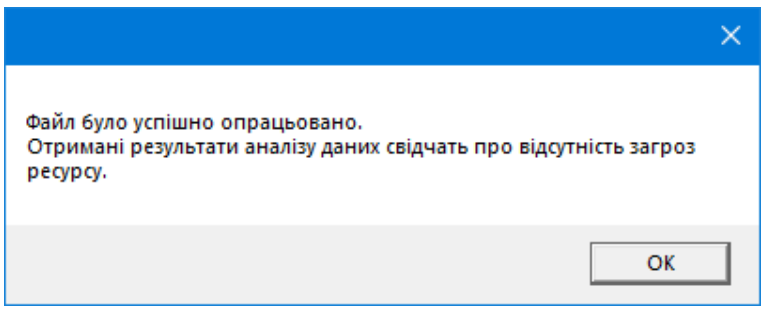
```
            var list = new string[7][];
            using (TextFieldParser parser = new TextFieldParser(path))
            {
               int i = 0;
               parser.TextFieldType = FieldType.Delimited;
               parser.SetDelimiters(",");
               while (!parser.EndOfData)
               {

                  //Process row
                  string[] fields = parser.ReadFields();

                  list[i] = fields;
                  //foreach (string field in fields)
                  //{
                  //    //TODO: Process field
                  //}
                  i = i + 1;
               }

               csvObject = new CSVobject()
               {
                  SessionAnalytic = ProcessArray(list[1]),
                  CountryAnalytic = ProcessArray(list[2]),
                  TraficAnalytic = ProcessArray(list[3]),
               };

               csvObjectBlok = ProcessArray(list[0]);
            }

            if (csvObjectBlok.First().Key == "1")
            {
               MessageBox.Show("Файл було успішно опрацьовано. \r\nОтримані результати
аналізу даних свідчать про відсутність загроз ресурсу.");
            }
            if (csvObjectBlok.First().Key != "1")
            {
               MessageBox.Show("Файл було успішно опрацьовано. \r\nОтримані результати
аналізу даних свідчать про наявність загроз ресурсу.");
            }
         }

         public List<KeyValuePair<string, int>> ProcessArray(string[] request)
         {
            var d = request.GroupBy(x => x);
            List<KeyValuePair<string, int>> valueList = new List<KeyValuePair<string, int>>();


            foreach (var item in d)
```

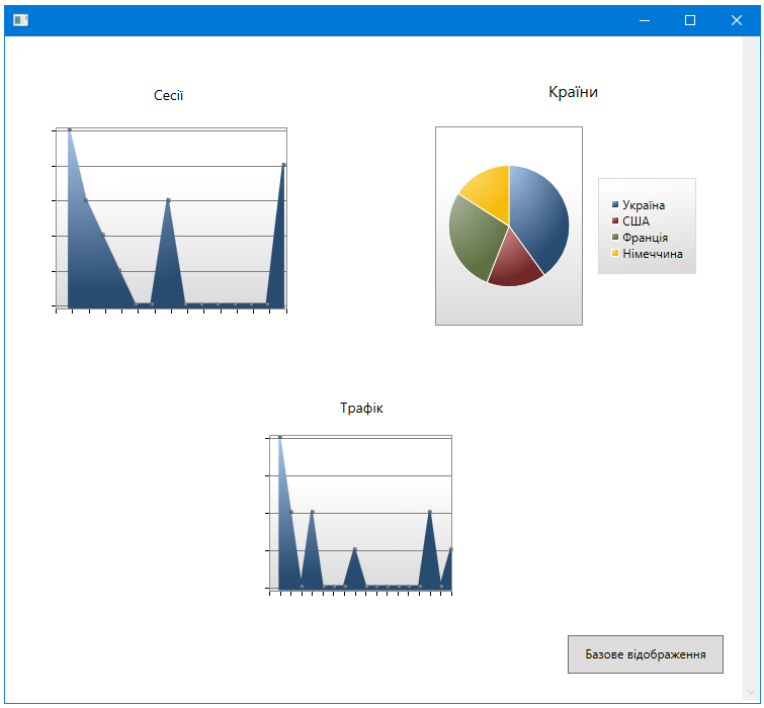
```
            {
                valueList.Add(new KeyValuePair<string, int>(item.Key, item.Count()));

            }

            return valueList;
        }

        private void Button_Click_3(object sender, RoutedEventArgs e)
        {

            //var w = Application.Current.Windows[0];
            //this.Hide();

            MainWindowClaster wfAbout = new MainWindowClaster();
            wfAbout.SetData(csvObjectBlok);
            wfAbout.ShowDialog();

            //this.Close();
        }

        private void Button_Click_2(object sender, RoutedEventArgs e)
        {

            //var w = Application.Current.Windows[0];
            //this.Hide();

            MainWindow wfAbout = new MainWindow();
            wfAbout.SetData(csvObject);
            wfAbout.ShowDialog();

            //this.Close();
        }
    }
}
```

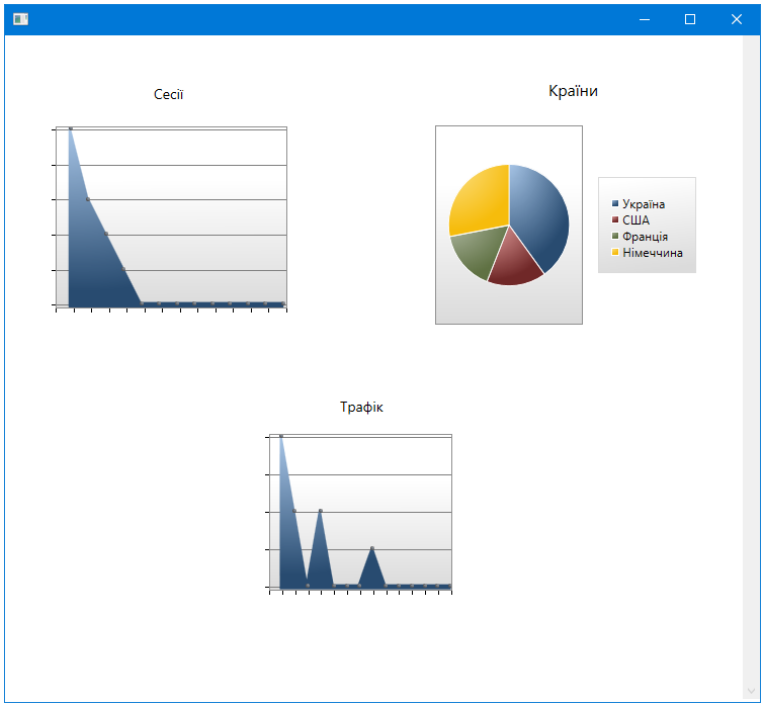
Додаток Г. Інтерфейс додатку

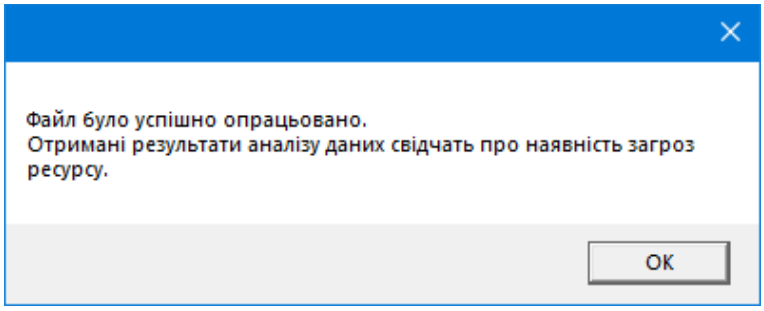







Сесії
Країни
Україна
США
Франція
Німеччина
Трафік
Базове відображення


Сесії
Країни
Україна
США
Франція
Німеччина
Трафік


Файл було успішно опрацьовано.
Отримані результати аналізу даних свідчать про наявність загроз ресурсу.
OK

Додаток Д. Ілюстративний матеріал

МАГІСТЕРСЬКА КВАЛІФІКАЦІЙНА РОБОТА

Вдосконалений метод інтелектуального аналізу даних з підвищеною достовірністю на основі математичного апарату нейронних мереж

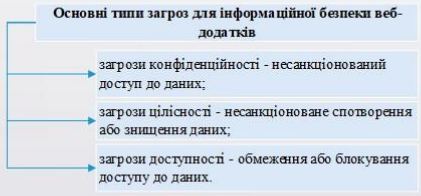
Виконав: ст. групи УБ-21м Панасюк Б.В.

Керівник: д.т.н., проф. Яремчук Ю.Є.

Вступ

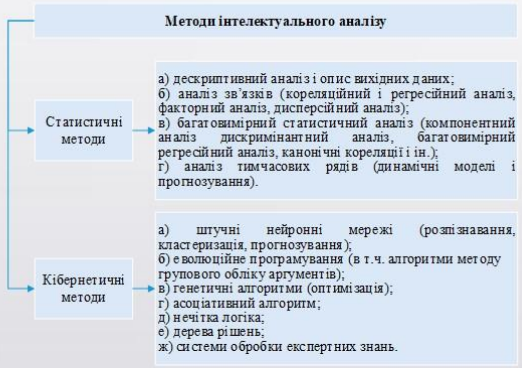
- Метою роботи є розробка вдосконаленого методу інтелектуального аналізу даних з підвищеною достовірністю на основі математичного апарату нейронних мереж для виявлення загроз веб-ресурсу.
- Практична цінність. Розроблено програмний продукт, який реалізує вдосконалений метод інтелектуального аналізу даних з підвищеною достовірністю на основі математичного апарату нейронних мереж.

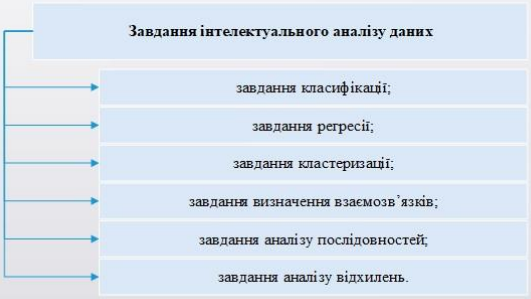
Аналіз загроз безпеки веб-ресурсу



- Аналіз загроз – це знання, отримані на основі фактичних даних про існуючу або виникаючу загрозу чи небезпеку для веб-ресурсів, які можуть використовуватися для прийняття рішень про реагування суб'єкта на цю загрозу чи небезпеку.

Методи та завдання інтелектуального аналізу даних





Аналіз програмних аналогів

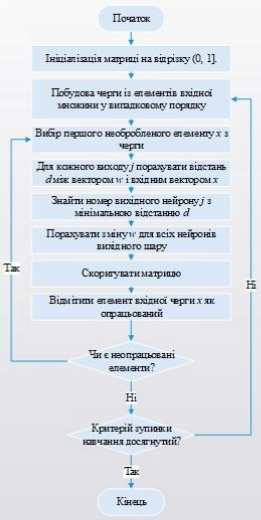
	SAS Enterprise Miner	STATISTICA Data Miner	Oracle Data Mining	KXEN Analytic Framework	Microsoft SQL Server Analysis Services	ПП SPSS	Як часто вик-тсья
Адаптивна Байєвська мережа			+				1
Аналіз тимчасових рядів		+		+	+	+	4
Граничні методи		+	+	+			3
Дерева рішень	+	+			+	+	4
Ієрархічна кластеризація	+	+	+	+			4
Лінійна регресія	+	+	+	+	+	+	6
Логістична регресія	+	+	+	+	+	+	6
Неієрархічна кластеризація	+	+	+	+	+		5
Нейронні мережі	+	+			+		3
Пошук асоціативних правил	+	+	+		+		4

Розробка вдосконаленого методу інтелектуального аналізу даних для виявлення загроз веб-ресурсу

1) Нейронна мережа на основі карт Кохонена

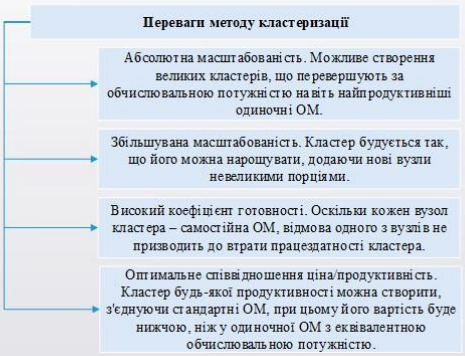
Ґрунтуючись на згаданих раніше принципах реалізації нейронних мереж, створимо нейронну мережу для визначення загроз веб-ресурсу.

Алгоритм процесу навчання нейронної мережі наведено на рисунку.



2) Вдосконалення на основі методу кластеризації

- В якості модифікації запропонуємо наступне рішення. Здійснивши навчання на основі карти Кохонена та провівши перший етап кластеризації, продовжимо дослідження за рахунок аналізу відстаней в кластері та міжкластерних відстаней.
- Суть ідеї полягає в тому, що після певної кількості сформованих кластерів, продовжимо аналіз неврахованих кластерів, що не містять нейронів, а тому не були враховані до основної вибірки. Аналіз проведемо за рахунок визначення відстаней за одною із метрик.



Алгоритм роботи програмного засобу для виявлення загроз веб-ресурсу

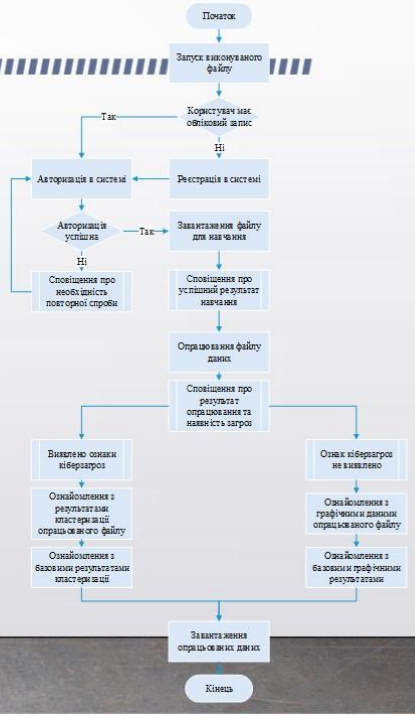
Вхідні параметри моделі для аналізу:
- країна підключення;
- час сесії;
- об`єм трафіку

Засоби обробки в моделі:
- вдосконалений метод на основі нейромережі Кохонена та кластеризації;
- виведення висновку;
- графічне представлення

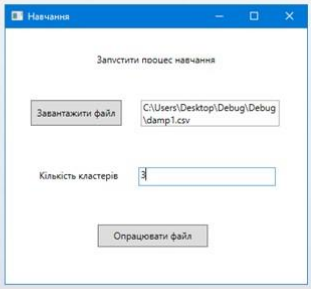
Результати моделі:
- ознак кіберзагроз ресурсу не виявлено;
- виявлено ознаки кіберзагроз ресурсу

- *Модель виявлення кіберзагроз на основі інтелектуального аналізу даних*

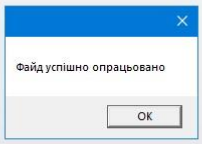
- *Алгоритм роботи програмного засобу*



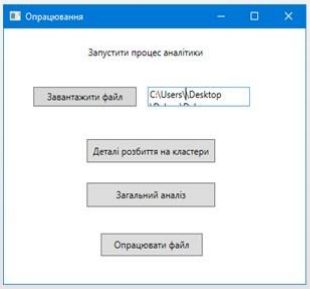
Тестування розробленого додатку для виявлення загроз веб-ресурсу на основі вдосконаленого методу



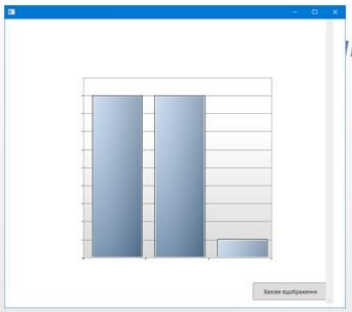
- *Вікно додатку для завантаження файлу для навчання системи*



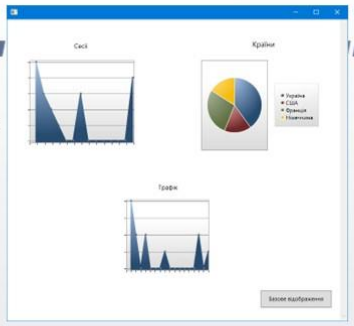
- *Сповіщення про успішне опрацювання файлу*



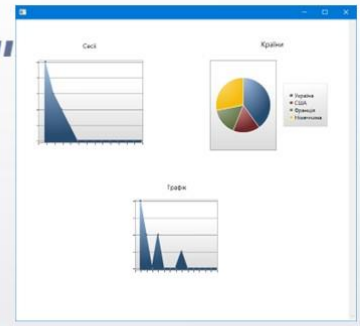
- *Вікно додатку для завантаження та файлу на основі навченої нейромережі*



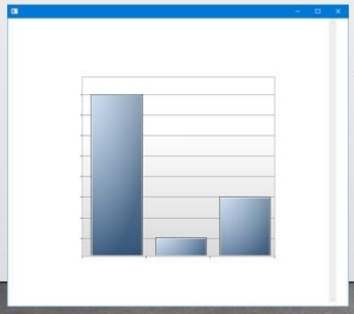
- *Деталі розбиття на кластери аналізованого файлу (1 випадок)*



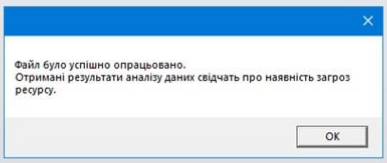
- *Деталі аналізу даних опрацьованого файлу (1 випадок)*



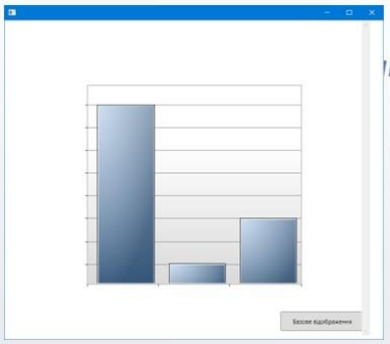
- *Деталі аналізу даних файлу для навчання*

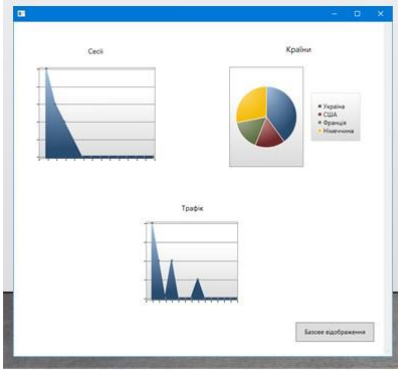


- *Деталі розбиття на кластери файлу для навчання*



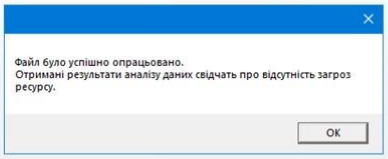
- *Виведення сповіщення про наявність загроз ресурсу в результаті здійсненого аналізу*



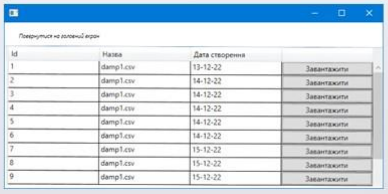


- *Деталі розбиття на кластери аналізованого файлу (2 випадок)*

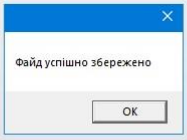
- *Деталі аналізу даних опрацьованого файлу (2 випадок)*



- *Виведення сповіщення про відсутність виявлених загроз ресурсу в результаті здійсненого аналізу*



- *Вигляд вікна історії опрацьованих файлів*



- *Сповіщення про успішне завантаження файлу*

Аналіз вдосконаленого методу

- Оцінка якості кластеризації здійснюється на основі метрики оцінювання кластеризації – індекс силуетів.
- $S_{xj} = \frac{b_{pj} - a_{pj}}{max(a_{pj}, b_{pj})}$
- $WC = \frac{1}{N}\sum_{j=1}^{N} S_{xj}$
- $WC = 0,95.$

- Для визначення ефективності вдосконаленого методу, порівняємо його із стандартним методом кластеризації на основі нейромережі Кохонена. Для цього скористаємось вибіркою файлів, в яких результати про наявність загроз заздалегідь відомі.

- *Результати аналізу на основі стандартної кластеризації*

- *Результати аналізу на основі вдосконаленого методу кластеризації*

Висновки

- В магістерській кваліфікаційній роботі описано розробку та реалізацію вдосконаленого методу виявлення загроз веб-ресурсу на основі інтелектуального аналізу даних із використанням нейронних мереж та кластеризації.
- В якості модифікації було запропоновано рішення, що передбачає навчання на основі карти Кохонена, проведення першого етапу кластеризації, а потім продовження дослідження за рахунок аналізу відстаней в кластері та міжкластерних відстаней із використанням метрики Евклідової відстані, що дозволить врахувати незалучені дані, що не містять нейронів, а тому не були враховані до основної вибірки.
- Показник оцінки якості кластеризації – індекс силуетів SWC=0,95 свідчить про коректну кількість кластерів та розподілені дані у них. Порівняння стандартної кластеризації нейронної мережі Кохонена та вдосконаленого методу показало, що достовірність вдосконаленого методу підвищена орієнтовано на 10%, що свідчить про виконання поставленої мети роботи.

Дякую за увагу!

ПРОТОКОЛ
ПЕРЕВІРКИ КВАЛІФІКАЦІЙНОЇ РОБОТИ
НА НАЯВНІСТЬ ТЕКСТОВИХ
ЗАПОЗИЧЕНЬ

Назва роботи: Вдосконалений метод інтелектуального аналізу даних з підвищеною достовірністю на основі математичного апарату нейронних мереж

Тип роботи: магістерська кваліфікаційна робота
(БДР, МКР)

Підрозділ: Кафедра менеджменту та безпеки інформаційних систем
Факультет менеджменту та інформаційної безпеки
(кафедра, факультет)

Показники звіту подібності Unicheck

Оригінальність 99% Схожість 1%

Аналіз звіту подібності (відмітити потрібне):

1. Запозичення, виявлені у роботі, оформлені коректно і не містять ознак плагіату.

2. Виявлені у роботі запозичення не мають ознак плагіату, але їх надмірна кількість викликає сумніви щодо цінності роботи і відсутності самостійності її виконання автором. Роботу направити на розгляд експертної комісії кафедри.

3. Виявлені у роботі запозичення є недобросовісними і мають ознаки плагіату та/або в ній містяться навмисні спотворення тексту, що вказують на спроби приховування недобросовісних запозичень.

Особа, відповідальна за перевірку _____ (підпис) Коваль Н.П. (прізвище, ініціали)

Ознайомлені з повним звітом подібності, який був згенерований системою Unicheck щодо роботи.

Автор роботи _____ (підпис) Панасюк Б.В. (прізвище, ініціали)

Керівник роботи _____ (підпис) Яремчук Ю.Є. (прізвище, ініціали)