

Міністерство освіти і науки України
Державний університет «Житомирська політехніка»
Факультет інформаційно-комп'ютерних технологій
Кафедра комп'ютерної інженерії та кібербезпеки

ПОЯСНЮВАЛЬНА ЗАПИСКА

до випускної кваліфікаційної роботи бакалавра
на тему:

«Детектор аномалій на основі служб машинного навчання
Azure для рішення завдань кібербезпеки»

Виконав студент 4-го курсу, групи КБ-1
спеціальності 125 «Кібербезпека»

 Олексюк Б. Ю.

Керівник, доцент кафедри КІ та КБ, к.т.н.,

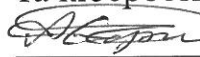
 Пулеко І. В.

Рецензент, доцент кафедри КІ та КБ, к.т.н.

 Россінський Ю.М.

Житомир 2021

ДЕРЖАВНИЙ УНІВЕРСИТЕТ «ЖИТОМИРСЬКА ПОЛІТЕХНІКА»
ФАКУЛЬТЕТ ІНФОРМАЦІЙНО-КОМП'ЮТЕРНИХ ТЕХНОЛОГІЙ
КАФЕДРА КОМП'ЮТЕРНОЇ ІНЖЕНЕРІЇ ТА КІБЕРБЕЗПЕКИ
СПЕЦІАЛЬНІСТЬ 125 «КІБЕРБЕЗПЕКА»

ЗАТВЕРДЖУЮ
Завідувач кафедри
комп'ютерної інженерії
та кібербезпеки
 Сфіменко А. А.
«12» 06 2021 р.

ЗАВДАННЯ
на випускню кваліфікаційну роботу

Студента Олексюк Богдана Юрійовича

Тема роботи: Детектор аномалій на основі служб машинного навчання Azure для
рішення завдань кібербезпеки

затверджена Наказом по університету від «12» червня 2021 р. № 134/С

Термін здачі студентом закінченої роботи – 20. 06 2021 р.

Вихідні дані для роботи:

Консультанти з випускної кваліфікаційної роботи із зазначенням розділів, що їх
стосуються

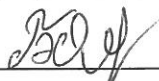
Розділ	Консультант	Підпис, дата	
		Завдання видав	Завдання прийняв
1	Пулеко І. В.		
2	Пулеко І. В.		
3	Пулеко І. В.		

Керівник



Календарний план

№ з/п	Назва етапів випускної кваліфікаційної роботи	Термін виконання етапів роботи	Примітка
1	Постановка завдання	6 квітня 2020 – 7 квітня 2020	Виконано
2	Пошук, огляд та аналіз літературних джерел	8 квітня 2020 – 11 квітня 2020	Виконано
3	Пректування структури системи	12 квітня 2020 – 16 квітня 2020	Виконано
4	Розробка конвеєра та налаштування конструктура	17 квітня 2020 – 19 квітня 2020	Виконано
5	Тестування	27 квітня 2020 – 4 травня 2020	Виконано
6	Захист	4 червня 2021	Виконано

Студент  Олексюк Б. Ю.

Керівник  Пулеко І. В.

РЕФЕРАТ

Випускна кваліфікаційна робота бакалавра направлена на розробку детектора аномалій на основі служб машинного навчання Azure.

Пояснювальна записка до роботи містить 68 сторінок, 21 ілюстрацій та 5 таблиць.

Метою дослідження є виявлення аномалій за рахунок розробки та застосування детектора на основі служб машинного навчання Azure. Об'єктом дослідження є процес пошуку аномалій у мережевому трафіку на основі служб машинного навчання Azure з використанням KDD Cup. Предметом дослідження алгоритми класифікації Конструктора машинного навчання студії Microsoft Azure для вирішення завдання виявлення аномалій.

В роботі визначено основні завдання на розробку детектора, який попередить та захистить від зловмисних вторгнень в системи. Також в роботі проаналізовані методи машинного навчання, обґрунтовано актуальність даного детектора аномалій та описана сфера його використання, обґрунтований вибір виробника обладнання – Microsoft. Розроблені, сформовані та налагоджені загальний проект детектора. Наведено дослідження переваг та недоліків технологій машинного навчання Azure, алгоритм побудови детектора аномалій, дослідження та порівняння алгоритмів прогнозування студії машинного навчання Azure, рекомендації щодо застосування технологій виявлення аномалій та вторгнень в майбутньому. Також наводяться результати навчання моделі детектора аномалій.

КЛЮЧОВІ СЛОВА: ЗАХИСТ, КЛАСИФІКАЦІЯ, МАШИННЕ НАВЧАННЯ, AZURE, ДЕТЕКТОР АНОМАЛІЙ.

					КБ.КР.Б – 125 – 21 – ПЗ			
Зм.	Арк.	№ докум.	Підпис	Дата	Детектор аномалій на основі служб машинного навчання Azure для рішення завдань кібербезпеки Пояснювальна записка	Літ.	Арк.	Аркушів
Розроб.		Олексюк Б. Ю.		23.06.21			4	
Керівник		Пулеко І. В.		23.06.21				
Рецензент.		Росіньський Ю.М.		29.06.21				
Зав. каф.		Єфіменко А. А.		29.06.21		Державний університет «Житомирська політехніка», група КБ-1		

ABSTRACT

The qualification work consists of an anomaly detector project based on Azure machine learning services and an explanatory note to it. The explanatory note to the work contains 68 pages, 21 illustrations and 5 tables.

The aim of the study is to identify anomalies through the development and application of a detector based on Azure machine learning services. The object of the study is the process of finding anomalies in network traffic based on Azure machine learning services using the KDD Cup. The subject of the study is the classification algorithms of the Machine Learning Designer of the Microsoft Azure studio to solve the problem of detecting anomalies.

The main tasks for the development of a detector that will prevent and protect against malicious intrusions into the system are identified. Also in the work the methods of machine learning are analyzed, the relevance of this anomaly detector is substantiated and the sphere of its use is described, the choice of the equipment manufacturer - Microsoft is substantiated. Developed, formed and adjusted the general project of the detector. A study of the advantages and disadvantages of Azure machine learning technologies, an algorithm for constructing an anomaly detector, research and comparison of forecasting algorithms for the Azure machine learning studio, recommendations for the use of technologies to detect anomalies and intrusions in the future. The results of learning the model of the anomaly detector are also given.

KEYWORD: SECURITY, NETWORK CLASSIFICATION, MACHINE LEARNING, AZURE, ANOMALY DETECTOR.

					КБ.КР.Б – 125 – 21 – ПЗ	
						5

ЗМІСТ

РЕФЕРАТ	4
ABSTRACT	5
ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ.....	7
ВСТУП.....	8
РОЗДІЛ 1. АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ ТА ОБҐРУНТУВАННЯ АКТУАЛЬНОСТІ ТЕМИ РОБОТИ	10
1.1. ОБҐРУНТУВАННЯ АКТУАЛЬНОСТІ ТЕМИ РОБОТИ.....	10
1.2. АНАЛІЗ МЕТОДІВ МАШИННОГО НАВЧАННЯ.....	18
1.3. АНАЛІЗ ІСНУЮЧИХ МЕТОДІВ КЛАСИФІКАЦІЇ	22
1.4. АНАЛІЗ ІСНУЮЧИХ МЕТОДИК ВИЯВЛЕННЯ АНОМАЛІЙ І ВТОРГНЕНЬ З ВИКОРИСТАННЯМ МЕТОДІВ МАШИННОГО НАВЧАННЯ.....	25
1.5. АНАЛІЗ АНАЛІТИЧНИХ МОЖЛИВОСТЕЙ ХМАРНОГО СЕРЕДОВИЩА AZURE З ВИКОРИСТАННЯМ МАШИННОГО НАВЧАННЯ	27
1.6. ПОСТАНОВКА ЗАВДАННЯ	30
РОЗДІЛ 2. РОЗРОБКА ХМАРНОГО РІШЕННЯ ЗАДАЧІ КЛАСИФІКАЦІЇ НА ОСНОВІ КОНСТРУКТОРА МАШИННОГО НАВЧАННЯ AZURE	31
2.1. РОЗРОБКА АРХІТЕКТУРИ РІШЕННЯ З ВИКОРИСТАННЯМ СТУДІЇ МАШИННОГО НАВЧАННЯ AZURE.....	31
2.2. АНАЛІЗ ВХІДНИХ ДАНИХ ТА ЇХ ПОПЕРЕДНЯ ОБРОБКА	36
2.3. РОЗРОБКА МОДЕЛІ ДЛЯ КЛАСИФІКАЦІЇ	44
РОЗДІЛ 3. ПЕРЕВІРКА ПРАЦЕЗДАТНОСТІ ТА ЕФЕКТИВНОСТІ РІШЕННЯ ЗАДАЧІ КЛАСИФІКАЦІЇ.....	48
3.1. ПОКАЗНИКИ ЯКОСТІ КЛАСИФІКАЦІЇ ПРИ ВИЯВЛЕННІ АНОМАЛІЙ.....	48
3.2. ДОСЛІДЖЕННЯ ТА ПОРІВНЯННЯ АЛГОРИТМІВ ПРОГНОЗУВАННЯ СТУДІЇ МАШИННОГО НАВЧАННЯ AZURE	52
3.3. РЕКОМЕНДАЦІЇ ПО ЗАСТОСУВАННЮ РОЗРОБЛЕНОЇ ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ	56
ВИСНОВКИ.....	60
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ.....	62

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ

KDD – Knowledge Discovery in Databases

GDPR – General Data Protection Regulation

CCPA – California Consumer Privacy Act

МН – машинне навчання

SVM – Support Vector Mashine

ELM – Extreme Learning Mashine

SLFN – Single hidden Learning Feedforward Network

IDS – Intrusion Detection System

DOS – Denial-of-service

R2L – remote to local

U2R – user to root

KNN – k-nearest neighbor

ВСТУП

Актуальність теми. В наш час все більше інформації переходить на електронні носії, зокрема Інтернет. Інтернет дуже корисна технологія, яка при розповсюдженні змінила напрямок досліджень інформаційних технологій. Його використовують для навчання, пошуку інформації, спілкування. Цей список можна вести до нескінченності. В Інтернета величезна сфера застосування. Тому не дивно що його стали використовувати для не дуже законних діяч. Так як, ми зберігаємо нашу персональну інформацію на електронних носіях, які часто підключені до Інтернету, зловмисники, які хочуть заволодіти нашими даними використовують різні способи взлому. Зараз технології різних вірусів і атак настільки розвинулись та пришвидшились, що людина ніяк не може поспівати при захисті систем. На сьогоднішній день, для вирішення проблем безпеки часто передбачається задіяння методів штучного інтелекту, зокрема методів машинного навчання. В інформаційній безпеці штучний інтелект - це програмне забезпечення, здатне інтерпретувати стан середовища, розпізнавати відбуваються в ній події і самостійно вживати необхідних заходів. Штучний інтелект особливо добре справляється з розпізнаванням закономірностей і аномалій, тому може бути прекрасним інструментом виявлення загроз.

У нашому випадку системи машинного навчання - це програмне забезпечення, здатне самостійно навчатися на введених людиною даних і результатах виконаних дій. Засоби машинного навчання здатні будувати прогнози, спираючись на відомості про розвиток подій в минулому.

Хорошу базу для створення рішень штучного інтелекту надають сучасні хмарні сервіси, наприклад, такі як Microsoft Azure. Тому тема роботи яка направлена на розробку детектора аномалій на основі служб машинного навчання Azure для рішення завдань кібербезпеки є **актуальною**.

Метою дослідження є виявлення аномалій за рахунок розробки та застосування детектора на основі служб машинного навчання Azure.

Перед нами стоятиме декілька завдань:

- Ознайомлення з технологіями служб машинного навчання Azure;

					КБ.КР.Б – 125 – 21 – ПЗ	8

- Дослідження переваг та недоліків технологій машинного навчання Azure;
- Визначення алгоритму побудови детектора аномалій;
- Вивчення та налаштування технологій служб машинного навчання Azure;
- Розроблення технологія виявлення аномалій та вторгнень в KDD Cup з використанням методів машинного навчання Azure.
- Дослідження та порівняння алгоритмів прогнозування студії машинного навчання Azure;
- Розробка рекомендацій щодо застосування технологій виявлення аномалій та вторгнень в майбутньому.

Об’єктом дослідження є процес пошуку аномалій у мережевому трафіку на основі служб машинного навчання Azure з використанням KDD Cup.

Предметом дослідження алгоритми класифікації Конструктора машинного навчання студії Microsoft Azure для вирішення завдання виявлення аномалій.

Наукова новизна дослідження полягає в використанні можливостей більш сучасної студії машинного навчання середовища Microsoft Azure для розробки моделі виявлення аномалій і вторгнень в KDD Cup.

Методи досліджень. В даній роботі були використані можливості моделювання хмарного середовища Microsoft Azure, зокрема «машинне навчання Azure», методи класифікації машинного навчання, набір даних KDD Cup Data set та стандартний набір даних для аудиту KDD.

РОЗДІЛ 1. АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ ТА ОБҐРУНТУВАННЯ АКТУАЛЬНОСТІ ТЕМИ РОБОТИ

1.1. Обґрунтування актуальності теми роботи

Штучний інтелект у задачах кібербезпеки за підтримки машинного навчання має бути потужним інструментом у близькому майбутньому. Як і у інших галузях, людське втручання довгий час було необхідно та незамінно в безпеці. Хоча в даний час кібербезпека в значній мірі покладається на людський вклад, ми поступово бачимо, що технології стають кращими у виконанні конкретних завдань, ніж ми.

Кожне вдосконалення технології дещо наближає нас до більш ефективного доповнення людських ролей. Серед цих розробок в основі всього лежить декілька напрямків досліджень:

1) Штучний інтелект (ШІ) розроблений, щоб надати комп'ютерам повну здатність реагувати як людський розум. Це основна дисципліна, під яку потрапляють багато інших, включаючи машинне навчання та глибоке навчання.

2) Машинне навчання (МН) використовує існуючі моделі поведінки, формуючи процес прийняття рішень на основі минулих даних та висновків. Для деяких змін все ще потрібне втручання людини. Машинне навчання є, мабуть, найактуальнішою дисципліною з питань кібербезпеки на сьогодні [1].

3) Глибоке навчання (ГН) працює подібно до машинного навчання, приймаючи рішення за попередніми моделями, але вносить корективи самостійно. Глибоке навчання в галузі кібербезпеки в даний час підпадає під сферу машинного навчання, тому ми зосередимось тут переважно на МН.

Штучний інтелект та кібербезпека були визнані революційними та набагато ближчими, ніж ми могли б подумати. Однак це лише часткова істина, до якої потрібно підходити із стриманими очікуваннями. Реальність така, що ми можемо зіткнутися з відносно поступовими вдосконаленнями в майбутньому. У перспективі, що може здатися поступовим, в порівнянні з

повністю автономним майбутнє насправді досі стрибає за те, що ми були здатні в минулому.

Коли ми досліджує можливі наслідки для безпеки в машинному навчанні та штучному інтелекту, важливо сформулювати поточні проблеми в кібербезпеці. Є багато процесів та аспектів, які ми давно прийняли як нормальні, і їх можна розглядати під егідою технологій штучного інтелекту.

Людські помилки - це значна частина слабких сторін кібербезпеки. Наприклад, правильною конфігурацією системи може бути надзвичайно складно управляти, навіть якщо великі ІТ-команди беруть участь у налаштуванні. У процесі постійних інновацій комп'ютерна безпека стала як ніколи багат шаровою. Адаптивні інструменти можуть допомогти командам знайти та пом'якшити проблеми, що виникають при заміні, модифікації та оновлення мережевих систем [1].

Розглянемо, як новішу Інтернет-інфраструктуру, таку як хмарні обчислення, можна розмістити поверх старих локальних систем. У корпоративних системах ІТ-команда повинна забезпечити сумісність для захисту цих систем. Процеси, що проводяться вручну для оцінки безпеки конфігурації, змушують команди відчувати втомленість, оскільки вони врівноважують нескінченні оновлення та звичайні щоденні завдання підтримки. Завдяки розумній адаптивній автоматизації команди можуть отримувати своєчасні поради щодо нещодавно виявлених проблем. Вони можуть отримати поради щодо подальших дій або навіть мати в наявності системи для автоматичного регулювання налаштувань за потреби.

Ефективність людини - ще одна проблема, яка виникає в галузі кібербезпеки. Жоден ручний процес не може бути ідеально повтореним щоразу, особливо в динамічному середовищі, подібному до нашого. Індивідуальне налаштування багатьох машин кінцевих точок в організації є одним із найбільш трудомістких завдань. Навіть після початкового налаштування ІТ-команди згодом переглядають машини знову для виправлення помилкових конфігурацій або застарілих налаштувань, які неможливо виправити у віддалених оновленнях.

Крім того, коли співробітникам дається завдання зреагувати на загрозу, масштаб знайденої загрози може швидко змінюватися. Там, де увага людини може сповільнюватися несподіваними проблемами, система, заснована на штучному інтелекті та машинному навчанні, може рухатися з мінімальними затримками.

Втомлюваність від загроз надає організаціям ще одну слабкість, якщо з ними не дбайливо поводитися. Поверхні атаки збільшуються, оскільки вище зазначені шари безпеки стають все більш досконалішими та розлогими. Багато систем безпеки налаштовані реагувати на багато відомих проблем набором суто рефлекторних попереджень. Як результат, ці індивідуальні підказки залишають людські команди для аналізу потенційних рішень та вжиття заходів.

Високий приплив попереджень робить цей рівень прийняття рішень особливо вимогливим процесом. Зрештою, втома від прийняття рішень стає щоденним досвідом для персоналу з кібербезпеки. Активні дії щодо цих виявлених загроз та вразливостей є ідеальними, але багатьом командам не вистачає часу та персоналу, щоб покрити всі виявлені проблеми.

Іноді командам доводиться вирішувати спочатку найбільші проблеми, а вторинні цілі відходять на другий план. Використання штучного інтелекту в рамках заходів з кібербезпеки може дозволити ІТ-командам ефективно та практично управляти більшістю з цих загроз. Протистояння кожній із цих загроз можна зробити набагато простішим, якщо використовувати автоматичне маркування. Окрім цього, деякі проблеми дійсно можуть бути розглянуті самим алгоритмом машинного навчання [3].

Час реакції на загрозу є абсолютно одним з найважливіших показників ефективності команд з кібербезпеки. Як відомо, зловмисні атаки відбуваються дуже швидко. Раніше діючим загрозам доводилося перебирати дозволи мережі та “обеззброювати” безпеку, іноді тижнями поспіль, перш ніж розпочати свою атаку.

На жаль, експерти в галузі кіберзахисту не єдині, хто виграє від технологічних інновацій. Останнім часом автоматизація стає більш звичним

явищем у кібератаках. Такі загрози, як нещодавні атаки-вимагателі LockBit, значно прискорили час атаки. В даний час деякі атаки можуть відбуватися навіть швидше, ніж півгодини. Реакція людини може відставати від початкової атаки, навіть із відомими типами атак. З цієї причини багато команд частіше беруть участь у реакціях на успішні атаки, а не в запобіганні спробам атаки. З іншого боку, невиявлені атаки самі по собі є небезпекою. Система захисту за допомогою машинного навчання може витягувати дані з атаки, щоб негайно їх згрупувати та підготувати до аналізу. Він може надати командам з кібербезпеки спрощені звіти, щоб зробити обробку та прийняття рішень більш чистою роботою. Виходячи за рамки простого звітування, цей тип захисту може також запропонувати рекомендовані дії для обмеження подальшої шкоди та запобігання майбутнім атакам.

Ідентифікація та прогнозування нових загроз є ще одним фактором, який впливає на терміни реагування на кібератаки. Як зазначалося раніше, час затримки вже має місце при наявних загрозах. Невідомі типи атак, поведінка та інструменти можуть додатково звести команду до повільних реакцій на загрози. Ще гірше, загрози, такі як крадіжка даних, іноді можуть залишатися абсолютно невиявленими. Опитування Fugue, проведене в квітні 2020 року, показало, що приблизно 84% ІТ-команд були стурбовані зломом їх хмарних систем без їх усвідомлення.

Постійна еволюція атак, що веде до скорочення часу та зусиль що витрачаються на атаки до мінімуму, завжди є основною проблемою в рамках зусиль захисту мереж. Але хороша новина, кібератаки зазвичай не будуються з нуля. Оскільки вони часто будуються поверх поведінки, фреймворків та вихідних кодів минулих атак, машинне навчання має вже існуючий шлях для роботи.

Програмування, засноване на машинному навчанні, може допомогти виділити спільність між новою загрозою та раніше виявленими, щоб допомогти розпізнати напад. Це те, що люди не можуть ефективно зробити вчасно, і надалі підкреслює необхідність адаптивних моделей безпеки. З цієї точки зору,

машинне навчання потенційно може спростити командам також прогнозування нових загроз та зменшення часу затримки завдяки підвищенню обізнаності про загрози [4].

Кадровий потенціал підпадає під сферу постійних проблем, які турбують багато команд ІТ та кібербезпеки у всьому світі. Залежно від потреб організації кількість кваліфікованих фахівців може бути обмежена. Однак найпоширеніша ситуація полягає в тому, що наймання нових людей може також коштувати організаціям великої частини бюджету. Підтримка людського персоналу вимагає не лише в оплаті щоденної праці, але надання допомоги у постійній потребі в освіті та сертифікації. Бути в професіоналом з питань кібербезпеки є вимогливим, особливо стосовно постійних інновацій.

Засоби безпеки, засновані на штучному інтелекті, можуть взяти на себе керівництво з менш щільною командою, яка забезпечує персонал і підтримує його. Хоча цьому персоналу потрібно буде йти в ногу з найсучаснішими напрямками штучного інтелекту та машинного навчання, економія коштів та часу буде супроводжуватися меншими вимогами до персоналу.

Адаптивність не є такою очевидною проблемою, як інші згадані моменти, але може суттєво змінити можливості безпеки організації. Людським командам може бути недостатньо можливостей адаптувати свої навички відповідно до заданих спеціалізованих потреб.

Якщо персонал не навчений певним методам, інструментам та системам, ви можете виявити, що в результаті ефективність вашої команди загальмована.

Навіть такі, здавалося б, прості потреби, як прийняття нової політики безпеки, можуть повільно рухатися в командах, де базою є людина. Це просто природа бути людиною, оскільки ми не можемо навчитися нових способів робити речі миттєво, і для цього потрібно мати час. За допомогою правильних наборів даних, навчені алгоритми можуть бути перетворені на індивідуальне рішення спеціально для вас.

Машинне навчання відмінно справляється з нудними завданнями, такі як ідентифікація та адаптація шаблону даних. Люди погано підходять для таких

					КБ.КР.Б – 125 – 21 – ПЗ	
						14

типів завдань через втому від завдань і, як правило, низьку толерантність до монотонності. Отже, хоча інтерпретація аналізу даних все ще перебуває в людських руках, машинне навчання може допомогти у розділенні даних та формування у зручній для розуміння формі.

Кібербезпека машинного навчання має кілька різних форм, кожна з яких має свої унікальні переваги:

1) Класифікація даних працює за допомогою попередньо встановлених правил для присвоєння категорій точкам даних. Позначення цих пунктів є важливою частиною побудови профілю атак, вразливостей та інших аспектів активного захисту. Це є фундаментальним для перетину машинного навчання та кібербезпеки.

2) Кластеризація даних бере результати що входять до заздалегідь встановлених правил класифікації, розміщуючи їх у “кластерних” колекціях даних із загальними рисами чи непарними ознаками. Наприклад, це можна використовувати при аналізі даних атак, для яких система ще не навчена. Ці кластери можуть допомогти визначити, як сталася атака, а також те, що було використано та показано.

3) Рекомендовані напрямки дій піднімають активні заходи системи безпеки машинного навчання. Це рекомендації, засновані на моделях поведінки та колишніх рішеннях, що пропонують звичайно запропоновані напрями дій. Тут важливо повторити, що це не розумне прийняття рішень за допомогою справжнього автономного ШІ. Швидше, це адаптивна система висновків, яка може за допомогою вже існуючих точок даних, укласти логічні зв'язки. Відповідь на загрози та пом'якшення ризиків може надзвичайно допомогти за допомогою цього типу інструментів [4].

4) Синтез можливостей дозволяє синтезувати нові можливості на основі уроків з попередніх даних та нових незнайомих наборів даних. Це дещо відрізняється від рекомендацій, оскільки більше зосереджується на шансах того, що дія чи стан системи відповідають подібним минулим ситуаціям.

Наприклад, цей синтез може бути використаний для попереджувального зондування слабких місць у системах організації.

5) Прогнозування є найбільш перспективним з процесів складової МН. Ця перевага досягається передбаченням потенційних результатів шляхом оцінки існуючих наборів даних. Це може бути використано в першу чергу для побудови моделей загроз, окреслення запобігання шахрайству, захисту від порушення даних і це є основним елементом багатьох прогнозних рішень кінцевих точок.

Всі ці форми машинного навчання використовуються в різних процесах кібербезпеки. Також існують частини кібербезпеки які важко уявити без засобів машинного навчання. Такі як:

1. Класифікація та відповідність конфіденційності даних [6].

Захист організацій від порушень законодавства про конфіденційність, мабуть, став головним пріоритетом за останні кілька років. Зі загальним положенням про захист даних (GDPR) з'явилися інші законодавчі заходи, такі як Каліфорнійський закон про захист споживачів (CCPA).

Управління зібраними даними ваших клієнтів та користувачів повинно здійснюватися відповідно до цих актів, що, як правило, означає, що ці дані повинні бути доступними для видалення за запитом. Наслідки недотримання цих законодавчих норм включають значні штрафи, а також шкоду репутації організації.

Класифікація даних може допомогти вам відокремити ідентифікаційні дані користувачів від тих, які є анонімними або без ідентифікації. Це позбавляє вас від фізичної праці при спробах проаналізувати величезні колекції старих і нових даних, особливо у великих або старих організаціях.

2. Профілі безпеки поведінки користувачів

Сформувавши власні профілі для співробітників мережі, що базуються на поведінці користувачів, безпеку можна адаптувати відповідно до вашої організації. Потім ця модель може встановити, як може виглядати неавторизований користувач, виходячи з випусків поведінки користувачів.

Тонкі риси, такі як натискання клавіш, можуть сформувати модель прогнозування загрози. Окресливши можливі результати потенційної несанкціонованої поведінки користувачів, служба безпеки машинного навчання може запропонувати рекомендований спосіб зменшення потенційно небезпечних поверхонь атак.

3. Профілі безпеки продуктивності системи

Подібно до концепції профілю поведінки користувача, спеціальний діагностичний профіль роботи всього вашого комп'ютера може бути скомпільований у звичайному стані. Моніторинг використання процесора та пам'яті поряд з такими рисами, як високий рівень використання даних в Інтернеті, може свідчити про шкідливу діяльність. Тим не менш, деякі користувачі можуть регулярно використовувати великі обсяги даних за допомогою відеоконференцій або частого завантаження великих медіафайлів. Дізнавшись, як зазвичай виглядає базовий показник роботи системи, можна встановити, як він не повинен виглядати, подібно до правил поведінки користувачів, про які ми згадали в попередньому прикладі машинного навчання.

4. Блокування ботів на основі поведінки

Діяльність ботів може спричинити спад вхідної пропускної здатності веб-сайтів. Це особливо актуально для тих, хто залежить від трафіку, наприклад, у тих, що мають спеціальні вітрини для електронної комерції та не мають постійного місця перебування.

Класифікуючи цю діяльність, засоби захисту МН можуть блокувати мережу ботів, незалежно від таких інструментів, як віртуальні приватні мережі, які можуть їх анонімізувати. Точки даних про поведінку на зловмисних сторонах можуть допомогти засобу захисту машинного навчання сформувати прогностичні моделі навколо цієї поведінки та превентивно блокувати нові веб-адреси для відображення цієї самої діяльності [2].

1.2. Аналіз методів машинного навчання

Машинне навчання використовує 2 основні методи навчання:

- індуктивне;
- дедуктивне.

Дедуктивне навчання відносять до області експертних систем, то його не використовують з причини не спроможності узгоджувати бази даних, що лежать в їх основі, з реляційною моделлю даних, що є дуже незручно при спробах наповнення баз даних експертних систем, промисловими базами даних.

Індуктивне навчання, в свою чергу поділяють на три основних типи: навчання з учителем або контрольоване навчання, навчання без учителя або неконтрольоване навчання, і навчання з підкріпленням (рис.1.1).

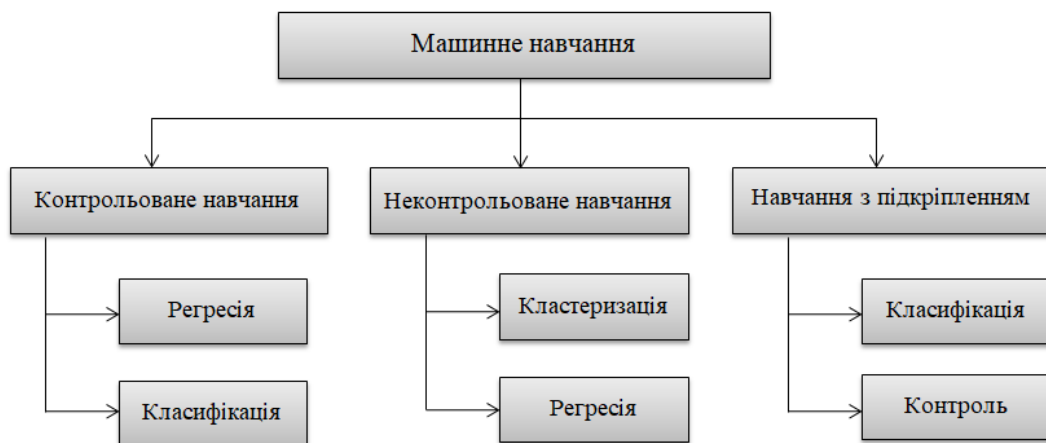


Рисунок 1.1 – Огляд типів та завдань машинного навчання

Навчання з учителем, також відоме як машинне навчання з учителем, є підкатегорією машинного навчання та штучного інтелекту. Це визначається використанням маркованих наборів даних для навчання алгоритмів, які дозволяють класифікувати дані або точно прогнозувати результати. Оскільки вхідні дані подаються в модель, вона регулює важливість, поки модель не буде встановлена належним чином, що відбувається як частина процесу перехресної перевірки. Навчання під контролем допомагає організаціям вирішувати різноманітні масштабні проблеми, наприклад класифікувати спам в окремій папці з папки "Вхідні".

Навчання з учителем використовує тренувальний перелік для навчання моделей, щоб отримати бажаний результат. Цей навчальний набір даних включає вхідні дані та правильні вихідні дані, які дозволяють моделі навчатися з часом. Алгоритм вимірює свою точність за допомогою функції втрат, регулюючи, поки помилка не буде мінімізована в достатній мірі.

Навчання з вчителем можна розділити на два типи проблем під час аналізу даних:

- класифікація;
- регресія.

Класифікація використовує алгоритм для точного віднесення тестових даних до конкретних категорій. Він розпізнає конкретні сутності в наборі даних і намагається зробити деякі висновки щодо того, як ці сутності слід позначати або визначати. Загальними алгоритмами класифікації є лінійні класифікатори, векторні машини підтримки (SVM), дерева рішень, k-найближчий сусід та випадковий ліс, які більш докладно описані нижче;

Регресія використовується для розуміння зв'язку між залежними та незалежними змінними. Зазвичай його використовують для прогнозування, наприклад, для отримання виручки від продажу для певного бізнесу. Лінійна регресія, логістична регресія та поліноміальна регресія - популярні алгоритми регресії.

Навчання без вчителя, також відоме як машинне навчання без вчителя, використовує алгоритми машинного навчання для аналізу та кластеризації немічених даних наборів. Ці алгоритми виявляють приховані шаблони або групування даних без потреби втручання людини. Він спроможний виявити схожість та різницю в інформації, роблячи ідеальні рішення для дослідницького аналізу даних, стратегію перехідних продажів, сегментації клієнтів та розпізнавання зображень[3].

- кластеризація - це метод інтелектуального аналізу даних для групування немаркованих даних на основі їх подібності чи відмінності. Наприклад, алгоритми кластеризації K-засобів призначають подібні точки

даних у групи, де значення K відображає розмір групування та деталізацію. Цей прийом корисний для сегментації ринку, стиснення зображень тощо;

- асоціація - це ще один тип методу навчання без нагляду, який використовує різні правила для пошуку зв'язків між змінними в даному наборі даних. Ці методи часто використовуються для аналізу ринкових кошиків та механізмів рекомендацій, відповідно до рекомендацій "Клієнти, які придбали цей товар".

Основна відмінність між двома підходами полягає у використанні маркованих наборів даних. Простіше кажучи, контрольоване навчання використовує позначені вхідні та вихідні дані, тоді як алгоритм некерованого навчання – ні [4].

При контрольованому навчанні алгоритм «навчається» із набору навчальних даних шляхом ітераційного прогнозування даних та коригування правильної відповіді. Хоча моделі контрольованого навчання, як правило, є більш точними, ніж моделі навчання без нагляду, вони потребують втручання людини заздалегідь, щоб належним чином позначити дані. Наприклад, контрольована модель навчання може передбачити, як довго буде їхати ваша поїздка на дорогу, залежно від часу доби, погодних умов тощо. Але по-перше, ви повинні навчити його, щоб знати, що дощова погода збільшує час водіння.

Моделі навчання без нагляду, навпаки, працюють самостійно, щоб виявити властивості структури немічених даних. Зверніть увагу, що для перевірки вихідних змінних вони все ще потребують втручання людини. Наприклад, модель без нагляду може визначити, що користувачі в Інтернеті часто купують групи товарів одночасно. Однак аналітику даних доведеться підтвердити,.

Моделі також відрізняються цілями застосування. Під час контрольованого навчання метою є прогнозування результатів для нових даних. Ви заздалегідь знаєте, яких результатів очікувати. Завдяки неконтрольованому алгоритму навчання метою є отримання знань із великих обсягів нових даних. Машинне навчання саме визначає, що відрізняється чи цікаве в наборі даних.

Моделі навчання, що контролюються, ідеально підходять для виявлення спаму, аналізу настроїв, прогнозування погоди та прогнозування цін, серед іншого. Навпаки, неконтрольоване навчання чудово підходить для виявлення аномалій, механізмів рекомендацій, персональних даних клієнтів та медичної візуалізації.

Також є різниця в складності використання. Контрольоване навчання - це простий метод машинного навчання, який, як правило, обчислюється за допомогою таких програм, як R або Python. При навчанні без нагляду вам потрібні потужні інструменти для роботи з великими обсягами некласифікованих даних. Моделі навчання без нагляду є складними в обчислювальному плані, оскільки їм потрібен великий навчальний набір для отримання запланованих результатів.

В кожного з методів є свої недоліки. Моделі навчання під наглядом можуть зайняти багато часу, а мітки для змінних введення та виведення вимагають знань. Тим часом, некеровані методи навчання можуть мати надзвичайно неточні результати, якщо у вас не буде людського втручання для перевірки вихідних змінних.

Вибір правильного підходу для потрібної ситуації залежить від того, як науковці оцінюють структуру та обсяг заданих даних, а також варіант використання.

Класифікація великих даних може бути справжньою проблемою у навчанні під наглядом, але результати є дуже точними та надійними. Навпаки, неконтрольоване навчання може обробляти великі обсяги даних у режимі реального часу. Але існує недостатня прозорість щодо кластеризації даних та більший ризик отримання неточних результатів. Тут виникає напівконтрольоване навчання[5].

Напівконтрольоване навчання пропонує компроміс між навчанням під наглядом та без нагляду. Під час навчання він використовує менший набір маркованих даних для керівництва класифікацією та вилучення ознак із більшого, немаркованого набору даних. Навчання під наглядом може вирішити

проблему недостатньої кількості маркованих даних (або неможливості дозволити собі маркування достатньої кількості даних) для підготовки керованого алгоритму навчання.

Машинне навчання з підкріпленням - це поведінкова модель машинного навчання, поділене на контрольне навчання та класифікацію, але алгоритм не навчений при використанні зразкових даних. Ця модель навчається, використовуючи програму та помилки. Послідовність успішних результатів буде посилено для вдосконалення найкращих рекомендацій або політики щодо даної проблеми.

1.3. Аналіз існуючих методів класифікації

1) Нейронні мережі

В основному, використовуючи алгоритми глибокого навчання, нейронні мережі обробляють навчальні дані, імітуючи взаємозв'язок людського мозку через шари вузлів. Кожен вузол складається з входів, вагових коефіцієнтів, змінних (або порогового значення) та вихідних даних. Якщо вихідне значення перевищує заданий порядок, воно «спрацьовує» або активує вузол, передаючи дані наступного шару в мережі. Нейронні мережі вивчають цю функцію картографування через контрольоване навчання, регулюючи втрати основної функції в процесі градієнтного спуску. Коли функція витрат дорівнює нулю або близько нього, ми можемо бути впевнені в точності моделі, та дати правильну відповідь.

2) Наївні Байєси

Наївний Байєс - це класифікаційний підхід, який приймає принцип класової умовної незалежності від теореми Байєса. Це означає, що наявність однієї ознаки не впливає на присутність іншої в імовірності даного результату, і кожен провісник має рівний вплив на цей результат. Існує три типи класифікаторів наївського Байєса: багаточленні наївські Байєси, Бернуллі Наївні Байєси та Гауссові наївні Байєси. Цей прийом в основному

використовується в системах класифікації тексту, ідентифікації спаму та рекомендацій[6].

3) Підтримка векторної машини (SVM)

SVM були спочатку запропоновані Вапником (1995) для вирішення проблем класифікації та регресійного аналізу.

SVM - це контрольована техніка навчання, яка навчена класифікувати різні категорії даних з різних дисциплін. Вона використовувалась для класифікації двох класів і застосовуються як для лінійних, так і для нелінійних завдань класифікації даних. SVM створює гіперплощину або декілька гіперплощин у багатовимірному просторі, і найкращою гіперплощиною є та, яка оптимально ділить дані на різні класи з найбільшим розділенням між класами. Нелінійний класифікатор використовує різні функції ядра для оцінки полів. Основною метою цих функцій ядра (тобто лінійної, поліноміальної, радіальної базисної та сигмовидної) є максимізація полів між гіперплощинами. Нещодавно дослідниками було розроблено багато дуже перспективних додатків через зростаючий інтерес до SVM. SVM широко використовується в програмах обробки зображень та розпізнавання образів.

4) Випадковий ліс

Випадковий ліс - ще один гнучкий керований алгоритм машинного навчання, що використовується як для класифікації, так і для регресії. "Ліс" посиляється на колекцію некорельованих дерев рішень, які потім об'єднуються разом, щоб зменшити дисперсію та створити більш точні прогнози даних. ВЛ працює, створюючи різні дерева рішень на етапі навчання та мітки класів результатів, які мають більшість голосів. ВЛ досягає високої точності класифікації та може обробляти викиди та шум у даних. У цій роботі використовується ВЛ, оскільки він менш схильний до надмірного припасування, і раніше він показав хороші результати класифікації[7].

5) Екстремальна навчальна машина(ELM)

ELM - це інша назва для одно- або багаторазових прихованих шарів прямого зв'язку нейронних мереж [13]. ELM може бути використаний для

					КБ.КР.Б – 125 – 21 – ПЗ	23

вирішення різних проблем класифікації, кластеризації, регресії та інженерних характеристик. Цей алгоритм навчання включає вхідний рівень, один або кілька прихованих шарів та вихідний рівень.

У традиційних нейронних мережах завдання регулювання ваги вхідного та прихованого шарів є дуже обчислювально дорогими і трудомісткими, тому що для зближення потрібно кілька ітерацій. Щоб подолати цю проблему, Huang et al. запропонував SLFN шляхом довільного вибору вагових коефіцієнтів та прихованих ухилів шару, щоб мінімізувати час навчання. Вичерпні деталі ELM доступні у Huang et al. та Qayyum та ін. Автори стверджують, що ці моделі швидше навчаються та досягають більш високих можливостей узагальнення порівняно з іншими моделями мереж прямого зв'язку. Ефективність ELM порівнянна з SVM або іншими сучасними класифікаторами машинного навчання. ELM має найбільшу здатність ефективніше працювати у дуже складних наборах даних.

6) Нечітка логіка

НЛ вона походить від теорії нечітких множин, яка займається міркуваннями, які є приблизними, а не точно похідними від класичної логіки предикатів. Застосована сторона теорії нечітких множин має справу з продуманими реальними експертними цінностями для складної проблеми. При такому підході дані класифікуються на основі різних статистичних показників.

Фрагменти даних застосовуються з використанням правил НЛ, щоб класифікувати їх як нормальні чи шкідливі. Існують різні інші нечіткі методи видобування даних для вилучення шаблонів, що представляють нормальну поведінку для виявлення вторгнень, які описують різні модифікації існуючих алгоритмів видобутку даних для підвищення ефективності та точності.

1.4. Аналіз існуючих методик виявлення аномалій і вторгнень з використанням методів машинного навчання

Система виявлення вторгнень (IDS) - це активний процес або пристрій, який аналізує діяльність системи та мережі на предмет несанкціонованої діяльності. Ідентифікатор - це апаратне чи програмне забезпечення або їх поєднання, яке використовується для моніторингу системи чи мережі систем від будь-якої зловмисної чи несанкціонованої діяльності. Системи виявлення вторгнень (IDS) використовуються для підвищення безпеки мережі. Ідентифікатор покращує безпеку мережі, виявляючи, оцінюючи та повідомляючи про несанкціоновані дії мережі. IDS поділяються на два класи: на основі мережі та на основі хосту. Мережеві системи виявлення вторгнень аналізують мережеві пакети, отримані з мережі. Система виявлення вторгнень на основі хоста аналізує системні виклики, генеровані окремими хостами. Потік даних через мережу дуже великий, і важко аналізувати та виявляти атаки за допомогою традиційних методів [8].

Аномалії - це шаблон у даних, який не відповідає чітко визначеній нормальній поведінці. Причиною аномалії може бути зловмисна діяльність або якість вторгнення. Ця аномальна поведінка, виявлена в наборі даних, цікава аналітику, і це є найважливішою характеристикою для виявлення аномалій.

Виявлення аномалій - це процес пошуку шаблонів у наборі даних, поведінка яких не є очікуваною. Ці несподівані способи поведінки також називаються аномаліями або відхиленнями. Аномалії не завжди можна класифікувати як атаку, це може бути виражено несподіваною поведінкою, якої раніше не було помічено. Це може мати шкідливий вплив, а може і не мати. Виявлення аномалій забезпечує безпеку для дуже важливої та критичної інформації в різних додатках, наприклад, при крадіжках кредитних карток або крадіжки особистих даних. Коли дані потрібно проаналізувати, щоб знайти взаємозв'язок або передбачити поведінку, зазвичай використовуються відомі або невідомі методи видобутку даних. Сюди входять класифікація,

кластеризація та методи на основі машинного навчання. Для досягнення вищого рівня точності виявлення аномалій також створюються гібридні підходи. Поєднання існуючих алгоритмів аналізу та класифікації для отримання кращих результатів - основна ідея цього підходу. Таким чином, виявлення несподіваної або нестандартної поведінки тобто аномалію дає змогу дослідити та класифікувати її як новий тип атак або унікальний тип вторгнень.

Існує безліч підходів до розробки алгоритму пошуку аномалій. Але в кожному підході є основні незмінні етапи (рис.1.2).

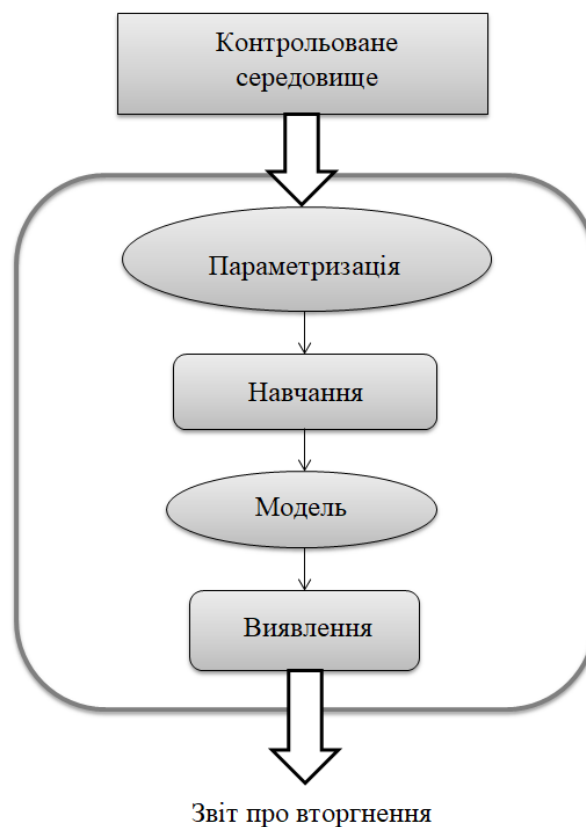


Рисунок 1.2 – Методологія виявлення аномалій

Етап параметризації - це попередня обробка даних у заздалегідь встановлені формати, які будуть прийняті та відповідатимуть поведінці цільової системи.

Етап навчання - модель побудована на основі нормальної (або ненормальної) поведінки системи. Способи навчання різняться в залежності від типу розглянутого методу виявлення аномалії. Може налаштовуватися вручну або автоматичним методом.

Етап виявлення - якщо модель для системи доступна, її порівнюють із (параметризованим або заздалегідь визначеним) спостережуваним трафіком. Якщо знайдена аномалія, перевіряється чи перевищує (або є меншою, від зазначених моделей відхилень) заздалегідь визначений поріг значень. При виявленні не очікуваного результату спрацюватиме сигнал тривоги.

1.5. Аналіз аналітичних можливостей хмарного середовища Azure з використанням машинного навчання

Azure ML - це наскрізна, хмарна, розширена / прогнозована платформа аналітики. Вона обслуговує організації, користувачів та дослідників даних усіх рівнів кваліфікації та досвіду. Azure ML дозволяє користувачам імпортувати навчальні дані, створювати, навчати та розгортати моделі машинного навчання та навіть прогнозувати результати та кластерні дані із простого веб-браузера.

Крім того, користувачі можуть легко імпортувати навчальні дані з ряду джерел, включаючи плоскі файли, сховище Azure BLOB, запити Hive, бази даних Azure SQL, списки веб-сайтів тощо. Потім вони можуть використовувати простий, звичний інтерфейс перетягування для створення загальних завдань робочого циклу для підготовки даних, вибору функцій та навчання, оцінки та порівняння моделей[9].

Azure ML пропонує велику колекцію вбудованих завдань перетворення даних, надійні та перевірені алгоритми машинного навчання Microsoft, а також підтримку часто використовуваних мов програмування науки про дані, таких як R. Після того, як модель розроблена, її можна легко розгорнути та опублікувати в форма веб-служби, розміщеної на Azure. Після розгортання модель може бути доступна майже з будь-якого місця, включаючи власні програми, веб-сайти, Azure Data Factory, Excel та Power BI та багато іншого для використання при прогнозуванні та кластеризації нових даних за потреби. Користувачі та компанії також можуть монетизувати створені ними моделі, публікуючи їх на Azure Market Learning Marketplace.

Azure ML також має декілька особливих функцій, які виділяють її на фоні інших аналогічних платформ:

1. Більш інтуїтивне створення веб-сервісу - одним клацанням миші користувачі можуть взяти “навчальну модель” і перетворити її на “бальну модель”. Azure ML також автоматично запропонує / створить точки введення та виведення моделі веб-служби. Нарешті, після створення Azure ML представляє файл Excel, який можна завантажити та використовувати для взаємодії з веб-службою для введення функцій та виведення оцінок / прогнозів.

2. Можливість навчання / перекваліфікації моделей за допомогою API - розробок, користувачі та клієнти тепер можуть періодично перекваліфікувати розгорнуту модель із власними, свіжими даними програмно через API. Це означає, що шаблони даних змінюються з часом, або якщо нові клієнти мають власні «різні» дані, вони можуть легко підключитися до існуючої моделі, перекваліфікувати її, використовуючи нові дані, і розпочати прогнозування з оновленою моделлю.

3. Підтримка Python - загальноживана мова програмування Python тепер повністю підтримується в Azure ML. Додати власний код Python дуже просто, потрібно лише перетягнути завдання “Виконати сценарій Python” у модель та вводити код безпосередньо у діалогове вікно. Azure ML навіть надає підказки щодо вхідних та вихідних даних сценарію. Завдяки підтримці Python, користувач тепер може інтегрувати алгоритми Python, R та Microsoft ML в один робочий процес.

Компанія Forrester Research випустила свій звіт за 3 квартал 2020 року для прогнозованої аналітики та машинного навчання на базі ноутбуків. Microsoft Azure визнано лідером у цьому напрямку, отримавши найвищий бал у здатності виконувати критерії та найвищий рейтинг у категорії стратегій.

Forrester поставив Azure Machine Learning найвищий з можливих балів у 13 оціночних показниках (рис.1.3)

		Forrester's weighting	Amazon Web Services*	Anaconda*	Civis Analytics	Cloudera	Databricks*	Domino Data Lab	Google	MathWorks	Microsoft	OpenText	Oracle	RStudio
Current offering		50%	2.43	2.67	2.05	3.38	2.48	3.19	3.12	3.55	3.36	2.57	2.71	2.93
Data	14%		2.00	1.00	2.00	2.00	1.00	1.00	2.50	3.50	3.00	2.00	1.50	
Modeling	14%		2.33	2.33	1.67	3.00	2.33	2.33	3.00	3.00	3.67	2.33	3.00	1.67
Collaboration	14%		3.00	4.33	3.67	5.00	4.33	5.00	3.00	3.67	4.33	3.00	3.00	4.33
Model evaluation	14%		2.33	1.67	1.00	1.00	1.67	2.33	4.33	4.33	3.00	1.00	4.33	1.67
Model operations (ModelOps)	14%		2.33	2.33	1.00	3.67	1.00	3.67	3.00	2.33	3.00	1.67	1.67	2.33
Methods and algorithms	14%		4.00	4.00	4.00	4.00	4.00	3.00	5.00	3.00	5.00	4.00	4.00	4.00
Platform infrastructure	14%		1.00	3.00	1.00	5.00	3.00	5.00	1.00	5.00	1.00	3.00	1.00	5.00
Strategy		50%	4.50	2.50	1.50	4.50	2.50	4.50	4.50	3.50	5.00	2.50	3.00	3.00
Ability to execute	25%		5.00	1.00	1.00	3.00	3.00	5.00	5.00	5.00	5.00	1.00	3.00	3.00
Solution roadmap	25%		5.00	1.00	1.00	5.00	1.00	5.00	3.00	3.00	5.00	3.00	3.00	3.00
Enablement	25%		3.00	3.00	3.00	5.00	3.00	5.00	5.00	5.00	5.00	3.00	3.00	3.00
Partners	25%		5.00	5.00	1.00	5.00	3.00	3.00	5.00	1.00	5.00	3.00	3.00	3.00
Market presence		0%	4.33	2.33	1.00	2.33	4.33	1.67	3.67	3.67	4.33	1.00	1.67	3.67
Customer adoption	33%		5.00	1.00	1.00	3.00	5.00	1.00	3.00	5.00	5.00	1.00	3.00	5.00
Evaluated product revenue	33%		3.00	1.00	1.00	1.00	3.00	1.00	3.00	3.00	3.00	1.00	1.00	1.00
Market awareness	33%		5.00	5.00	1.00	3.00	5.00	3.00	5.00	3.00	5.00	1.00	1.00	5.00

All scores are based on a scale of 0 (weak) to 5 (strong).
*Indicates a nonparticipating vendor.

157464

Source: Forrester Research, Inc. Unauthorized reproduction, citation, or distribution prohibited.

Рисунок 1.3 - Результат дослідження Forrester Research аналогів Azure Machine Learning

Також вони описали можливості та переваги Azure Machine Learning так:

Корпорація Майкрософт надає вченим, що займаються кодуванням даних, усі навороти. З колекції різномірних пропозицій PAML - Azure ML Workbench, Azure ML Studio та Azure Batch AI - Microsoft створила нову уніфіковану пропозицію - Microsoft Azure Machine Learning. Результат трансформаційний. Машинне навчання Microsoft Azure пропонує повний набір можливостей корпоративного PAML, від централізованих реєстрів моделей до налаштування гіперпараметрів та модульних конвеєрів навчання та розгортання моделей[10].

Корпорація Майкрософт приділяла особливу увагу співпраці - наприклад, дозволяючи користувачам одночасно працювати в одному ноутбукі та інтеграції з взаємодіями Jenkins та GitHub, щоб забезпечити можливості MLOps, а також справедливості та відповідальне машинне навчання - наприклад, створення можливостей для тестування моделі на чутливих змінних,

таких як вік і стать, що рекомендують моделі пом'якшення наслідків та захист даних шляхом додавання шуму або навчання без нагляду. У машинному навчанні Microsoft Azure також є майстр AutoML, інструменти перетягування для створення конвеєрів ML та інтеграції для побудови моделей у редакторах SQL для підтримки розробників, аналітиків даних та інших науковців, що не належать до даних, які хочуть створювати та розгортати моделі.

1.6. Постановка завдання

Проведений аналіз показав що для вирішення проблем безпеки необхідне використання методів штучного інтелекту, зокрема методів машинного навчання.

В інформаційній безпеці штучний інтелект- це програмне забезпечення, здатне розпізнавати чи відбуваються в досліджуваному середовищі аномальні події і самостійно вживати необхідних заходів. Штучний інтелект особливо добре справляється з розпізнаванням закономірностей і незвичайної поведінки, тому ШІ бути прекрасним інструментом побудови програмного забезпечення для виявлення аномалій та загроз.

У якості методу виявлення аномалій будуть використовуватись методи класифікації (із застосуванням різних алгоритмів машинного навчання).

Класифікація - це метод який має завдання взяти кожен екземпляр набору даних, що розглядається, і віднести його до певного класу. Він може бути ефективним як для виявлення зловживань, так і для виявлення аномалій, але частіше використовується для виявлення зловживань.

Класифікацію будемо проводити з використанням конструктора Машинного навчання Azure.

Необхідно побудувати загальну модель з методами класифікації в Конструкторі конвеєра Машинного навчання Azure, навчити її на вхідних даних, оцінити якість виявлення аномалій і порівняти різні методи класифікації між собою.

РОЗДІЛ 2. РОЗРОБКА ХМАРНОГО РІШЕННЯ ЗАДАЧІ КЛАСИФІКАЦІЇ НА ОСНОВІ КОНСТРУКТОРА МАШИННОГО НАВЧАННЯ AZURE

2.1. Розробка архітектури рішення з використанням студії машинного навчання Azure

Машинне навчання Azure - це окрема та модернізована послуга, яка забезпечує повну платформу з обробки даних. Він підтримує як досвід повністю налаштування кодом, так і досвід низької кількості коду. Для роботи буде використовуватися досвід низької кількості коду.

МН Azure має перелік компонентів які взаємодіють між собою в процесі компіляції, розгортання та підтримки моделей машинного навчання (рис.2.1)

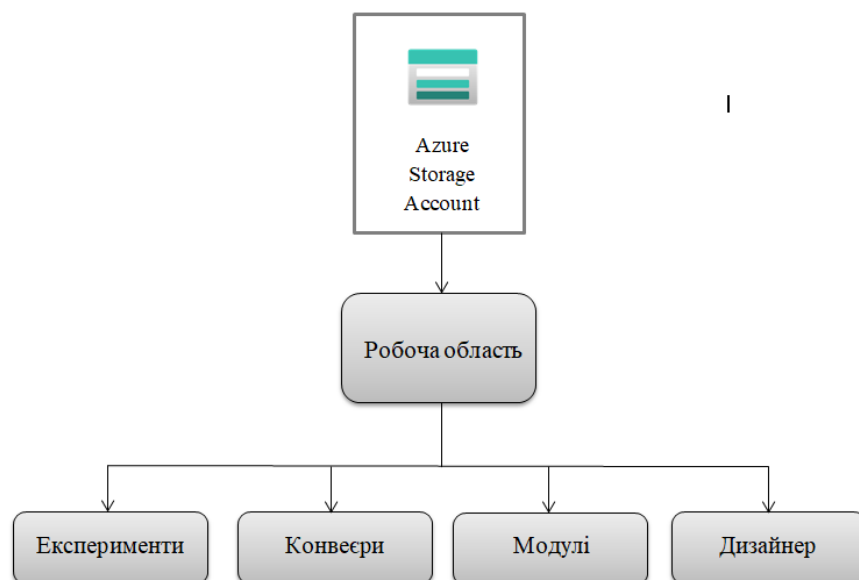


Рисунок 2.1 – Основні компоненти для роботи Azure Machine Learning

Робоча область - це ресурс верхнього рівня для вашої діяльності в машинному навчанні, що забезпечує централізоване місце для перегляду та управління модулів, які ви створюєте під час використання машинного навчання Azure. Обчислювальні ресурси забезпечують заздалегідь налаштоване хмарне середовище, яке можна використовувати для навчання, розгортання, автоматизації, управління та відстеження моделей машинного навчання [11].

Експеримент складається з наборів даних, що надають дані для модулів аналітики, які з'єднуються один з одним для створення моделі прогнозу аналітики.

Коли модель буде готова, навчальний експеримент можна перетворити в прогнозний, а потім опублікувати його як веб-службу, щоб модель стала доступна іншим користувачам. При цьому не потрібно бути фахівцем з програмування. Виконується візуальне з'єднання наборів даних і модулів, щоб створити модель прогнозу аналітики в робочій області експеримента.

Дизайнер конвеєру надає цю робочу область для побудови, тестування та розгортання моделей машинного навчання. З дизайнером ви можете:

1. Перетягувати набори даних та модулі на робочу область.
2. Підключати модулі, для створення конвеєрного проекту.
3. Виконувати запуск конвеєра, використовуючи обчислювальні ресурси у робочій області машинного навчання Azure.
4. Перетворювати свої навчальні конвеєри на конвеєри виводу.
5. Публікувати свої конвеєри в кінцевій точці конвеєру REST, щоб мати можливість опубліковувати новий конвеєр, який працює з різними параметрами та наборами даних.

5.1 Опубліковувати навчальний конвеєр для повторного використання при навчанні кількох моделей при зміні параметрів та наборів даних.

5.2 Опубліковувати конвеєр пакетного висновку, щоб робити прогнози щодо нових даних за допомогою попередньо навченої моделі.

6. Розгортувати конвеєр виводу до кінцевої точки в реальному часі, щоб робити прогнози щодо нових даних у реальному часі.

Конвеєри машинного навчання Azure організовують кілька етапів машинного навчання та обробки даних в один ресурс. Конвеєри дозволяють організовувати, керувати та повторно використовувати складні робочі процеси машинного навчання між проектами та користувачами [12].

Конвеєр складається з наборів даних та аналітичних модулів, які ви підключаєте. Він має багато застосувань: ви можете зробити конвеєр, який готує одну модель, або той, який навчає декілька моделей. Ви можете створити конвеєр, який робить прогнози в режимі реального часу або пакетно, або зробити конвеєр, який очищає лише дані. Конвеєр дозволяють вам повторно використовувати свою роботу та організовувати свої проекти.

Коли ви редагуєте конвеєр у конструкторі, ваш прогрес зберігається як проект конвеєра. Ви можете редагувати проект конвеєра в будь-який момент, додаючи або видаляючи модулі, конфігуруючи обчислювальні цілі, створюючи параметри тощо.

Конвеєр має такі характеристики:

- набори даних можуть підключатися лише до модулів;
- модулі можуть підключатися лише до наборів даних або інших модулів;
- всі вхідні порти для модулів повинні мати певне з'єднання з потоком даних;
- для кожного модуля повинні бути встановлені всі необхідні параметри.

Кожного разу, коли ви запускаєте конвеєр, конфігурація на момент запуску та його результати зберігаються у вашій робочій області як запуск конвеєра. Ви можете повернутися до будь-якого запуску конвеєра, щоб перевірити його для аудиту або усунення несправностей. Кожен новий запуск конвеєра створює клон існуючого конвеєрного проекту, який можна редагувати.

Запуски конвеєра згруповані в експерименти для упорядкування історії запусків. За потреби можливе створення нового експеременту для кожного запуску конвеєра.

Модуль - це алгоритм, який ви можете використовувати для своїх даних. Конструктор має кілька видів модулів, починаючи від функцій введення даних і закінчуючи процесами навчання, оцінки та перевірки.

Приклади модулів:

- обчислення елементарної статистики - виконує елементарні статистичні обчислення, такі як отримання математичного очікування, середньоквадратичного відхилення і ін;

- імпортування даних - завантажує дані з веб-URL-адрес або з різних хмарних сховищ в Azure;

- класифікація - підпункт розділу алгоритми машинного навчання. Містить набір модулів з методами алгоритмів класифікації.

Для нашого випадку потрібно створити модель класифікації за допомогою конструктора.

Конструктор - це засіб який дозволяє створювати моделі машинного навчання за допомогою візального інтерфейсу не використовуючи при цьому рядки коду. Основними етапами створення моделі в конструкторі є :

1. Створення робочої області.
2. Створення конвеєра.
3. Імпортування даних
4. Розробка моделі класифікації
5. Навчання моделей машинного навчання
6. Оцінка якості моделі
7. Розгортання моделі у вигляді кінцевого продукту

Для створення робочої області потрібно увійти в портал Azure за допомогою облікових даних підписки Azure та створити ресурс. Після цього, використовуючи сервіс Azure машинне навчання, відкриваємо вікно налаштувань нової робочої області. Як тільки закінчуємо налаштування параметрів робочої області натискаємо Перевірити і створити (рис.2.2).

Главная > Создание ресурса > Машинное обучение >

Машинное обучение

Создание рабочей области машинного обучения

Основные сведения Сеть Дополнительно Теги Проверка и создание

Сведения проекта

Выберите подписку для управления развернутыми ресурсами и затратами. Используйте группы ресурсов, например папки, для упорядочения и контроля всех ваших ресурсов.

Подписка * ⓘ Azure for Students

Группа ресурсов * ⓘ kb1_obyu
Создать

Сведения рабочей области

Укажите имя и регион для рабочей области.

Имя рабочей области * ⓘ KDDExpirement

Регион * ⓘ Восточный регион США

Учетная запись хранения * ⓘ kddetection1840015119
Создать

Хранилище ключей * ⓘ kddetection3321544104
Создать

Application Insights * ⓘ kddetection4769008581
Создать

Регистр контейнеров * ⓘ Нет
Создать

Рисунок 2.1 - Створення робочої області

Для створення конвеєра потрібно увійти на сайт ml.azure.com та вибрати робочу область, з якою ви хочете працювати. Після цього потрібно назвати конвеєр використовуючи значок “Налаштування” і рядок “Назва проекту”. Також конвеєр працює на цілі обчислення, яка є обчислювальним ресурсом, приєднаним до вашої робочої області. Для її створення потрібно в зоні “Налаштування”, яка відкриється справа від робочої області, вибрати створити ціль обчислення.

Після створення цілі обчислення ви можете використовувати її повторно для подальших запусків. Створення обчислювального ресурсу займає приблизно п'ять хвилин. Після створення ресурсу ви можете використовувати його повторно та пропустити цей час очікування для подальших запусків.

Обчислюваний ресурс автоматично масштабується до нульових вузлів, коли він не працює, щоб заощадити витрати. При повторному використанні з затримкою може виникнути приблизно п'ять хвилин очікування, поки він масштабується.

Після цього робоча область конвеєра готова до роботи з конструктором.

2.2. Аналіз вхідних даних та їх попередня обробка

Для навчання моделей будемо використовувати набір даних KDD Cup.

Детектор аномалій(вторгнень) KDD - програмне забезпечення для виявлення вторгнень у мережу захищає комп'ютерну мережу від несанкціонованих користувачів, включаючи, можливо, інсайдерів. Навчальним завданням детектора вторгнень є побудова прогнозової моделі (тобто класифікатора), здатної розрізняти ``погані" з'єднання, які називаються вторгненнями або атаками, і ``хороші" звичайні з'єднання.

Вихідні навчальні дані становили близько чотирьох гігабайт стиснених двійкових даних дампа TCP за сім тижнів мережевого трафіку. Це було оброблено приблизно в п'ять мільйонів записів про з'єднання. Подібним чином, два тижні тестових даних дали близько двох мільйонів записів про підключення.

З'єднання - це послідовність TCP-пакетів, що починаються і закінчуються в певний час, між якими дані надходять до вихідної IP-адреси та до цільової IP-адреси за деяким чітко визначеним протоколом. Кожне з'єднання позначене як звичайне або як атака з точно одним конкретним типом атаки. Кожен запис з'єднання складається приблизно із 100 байт.

Напади діляться на чотири основні категорії:

1. DOS: відмова в обслуговуванні, напр. синхронний флуд;
2. R2L: несанкціонований доступ з віддаленої машини, наприклад вгадування пароля;
3. U2R: несанкціонований доступ до локальних привілеїв суперкористувача (кореня), наприклад, різні атаки ``переповнення буфера";
4. Probing: спостереження та інші зондування, наприклад, сканування портів.

Важливо зазначити, що дані тесту не походять з того самого розподілу ймовірностей, що й дані навчання, і вони включають конкретні типи атак, що не містяться в даних тренувань. Це робить завдання більш реалістичним. Деякі

експерти про втручання вважають, що більшість нових атак є варіантами відомих атак, і "підпису" відомих атак може бути достатнім для виявлення нових варіантів. Набори даних містять загалом 24 типи навчальних атак, з додатковими 14 типами лише в тестових даних.

Stolfo та ін. визначені функції вищого рівня, які допомагають відрізнити звичайні зв'язки від атак. Існує кілька категорій похідних ознак.

Функції «того самого хоста» перевіряють лише з'єднання за останні дві секунди, які мають той самий хост призначення, що і поточне з'єднання, і обчислюють статистику, пов'язану з поведінкою протоколу, службою тощо.

Подібні функції «тієї ж служби» перевіряють лише ті з'єднання за останні дві секунди, які мають ту саму послугу, що і поточне з'єднання.

Функції «того самого хоста» і «тієї ж служби» разом називаються функціями трафіку, що базуються на часі, у записах з'єднання.

Деякі зондуючі атаки сканують хости (або порти) із значно більшим інтервалом часу, ніж дві секунди, наприклад раз на хвилину. Тому записи з'єднання також були відсортовані за хостом призначення, а функції побудовані за допомогою вікна із 100 підключень до того самого хоста замість часового вікна. Це дає набір так званих особливостей трафіку на основі хоста.

На відміну від більшості атак DOS та зондування, схоже, немає послідовних шаблонів, які часто зустрічаються у записах атак R2L та U2R. Це пов'язано з тим, що атаки DOS та зондування передбачають багато з'єднань з деякими хостами за дуже короткий проміжок часу, але атаки R2L та U2R вбудовані в частини даних пакетів і зазвичай включають лише одне з'єднання.

Корисні алгоритми автоматичного видобування неструктурованих частин даних пакетів - це відкрите питання дослідження. Stolfo та ін. використовували знання домену для додавання функцій, які шукають підозрілу поведінку в частинах даних, наприклад кількість невдалих спроб входу. Ці функції називаються функціями «вмісту»[13].

Для кращого розуміння, повний перелік набору функцій приведений в табл. 2.1 - 2.3.

					КБ.КР.Б – 125 – 21 – ПЗ	
						37

Таблиця 2.1

Основні особливості окремих з'єднань TCP

Назва функції	Опис	Тип
duration	тривалість (кількість секунд) з'єднання	безперервна
protocol_type	тип протоколу, напр. tcp, udp тощо	дискретна
service	послуга мережі за призначенням, наприклад, http, telnet тощо.	дискретна
src_bytes	кількість байтів даних від джерела до пункту призначення	безперервна
dst_bytes	кількість байтів даних від пункту призначення до джерела	безперервна
flag	нормальний або стан помилки з'єднання	дискретна
land	1 якщо з'єднання здійснюється з / до одного хосту / порту; 0 інакше	дискретна
wrong_fragment	кількість `` неправильних " фрагментів	безперервна
urgent	number of urgent packets	безперервна

Таблиця 2.2

Функції вмісту в межах з'єднання, запропонованого доменом

Назва функції	Опис	Тип
1.	2.	3.
hot	кількість «гарячих» показників	безперервна
num_failed_logins	кількість невдалих спроб входу	безперервна
logged_in	1 при успішному вході в систему; 0 інакше	дискретна
num_compromised	кількість «скомпрометованих» умов	безперервна
root_shell	1 якщо отримана коренева оболонка; 0 інакше	дискретна
su_attempted	1 при спробі команди «su root»; 0 інакше	дискретна
num_root	кількість «корневих» доступів	безперервна

Продовження табл. 2.2

1.	2.	3.
num_file_creations	кількість операцій створення файлів	безперервна
num_shells	кількість підказок оболонки	безперервна
num_access_files	кількість операцій з файлами контролю доступу	безперервна
num_outbound_cmds	кількість вихідних команд у сеансі ftp	безперервна
is_hot_login	1, якщо логін належить до «гарячого» списку; 0 інакше	дискретна
is_guest_login	1, якщо логін є «гостьовим» логіном; 0 інакше	дискретна

Таблица 2.3

Характеристики трафіку, обчислені за допомогою двосекундного часового вікна

Назва функції	Опис	Тип
count	кількість підключень до того самого хосту, що і поточне підключення за останні дві секунди	безперервна
serror_rate	% з'єднань, які мають помилки «SYN»	безперервна
rerror_rate	% з'єднань, які мають помилки «REJ»	безперервна
same_srv_rate	% підключень до тієї самої послуги	безперервна
diff_srv_rate	% підключень до різних служб	безперервна
srv_count	кількість підключень до тієї самої служби, що і поточне з'єднання за останні дві секунди	безперервна
srv_diff_host_rate	% of connections to different hosts	безперервна

Для навчання моделі буде використовуватися набір даних - network intrusion detection.

З цього набору даних буде використовуватися 43 стовпця і 125973 рядки.

Для завантаження та попередньої обробки даних використовуємо модулі Конструктора:

1. *Import Data* - цей модуль використовується для завантаження даних у конвеєр машинного навчання із існуючих хмарних служб даних.

Модуль *Import Data* підтримує зчитування даних із таких джерел:

- URL via HTTP;
- Azure cloud storages (через Datastores);
- Azure Blob Container;
- Azure File Share;
- Azure Data Lake;
- Azure Data Lake Gen2;
- Azure SQL Database;
- Azure PostgreSQL.

Перш ніж використовувати хмарне сховище, спочатку потрібно зареєструвати сховище даних у робочій області Azure Machine Learning.

Після того, як ви визначите потрібні дані та підключитесь до джерела, Імпорт даних визначає тип даних кожного стовпця на основі значень, які він містить, і завантажує дані у ваш конструктор. Результатом імпорту даних є набір даних, який можна використовувати з будь-яким конструкторським конвеєром(рис.2.2).

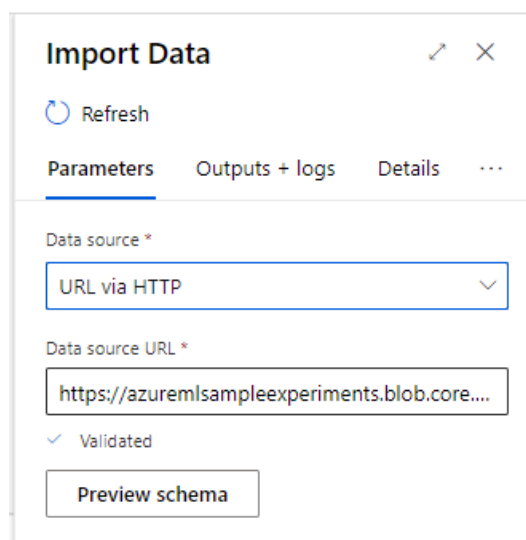


Рисунок 2.2 - Налаштування модуля для роботи з нашим набором даних

Edit Metadata - цей модуль використовується для редагування метаданих, пов'язаних зі стовпцями набору даних. Значення та тип набору даних змінюватимуться після використання модуля *Edit Metadata*.

Типові зміни метаданих можуть включати:

- перерахування булевих або числових стовпців як категоріальних значень;
- вказівка, який стовпець містить мітку класу або містить значення, які ви хочете класифікувати або передбачити;
- позначення стовпців як ознак;
- зміна значень дати / часу на числові значення або навпаки;
- перейменування стовпців.

Edit Metadata використовується в будь-який час, коли потрібно змінити визначення стовпця, як правило, щоб задовольнити вимоги до подальшого модуля. Наприклад, деякі модулі працюють лише з певними типами даних або вимагають прапорців у стовпцях, такі як *IsFeature* або *IsCategorical*.

Якщо потрібно після виконання необхідної операції можна скинути метадані до початкового стану.(рис. 2.3)

The image shows a 'Edit Metadata' window with the following fields:

- Column**: class
- Data type**: Unchanged
- Categorical**: Categorical
- Fields**: Unchanged
- New column names**: (empty field)

Рисунок 2.3 - Налаштування модуля для роботи з нашим набором даних

Convert to Indicator Values - цей модуль використовується, щоб «Перетворити у значення показників» у конструкторі машинного навчання Azure, щоб перетворити стовпці, що містять категоріальні значення, у ряд двійкових стовпців-показників.

Модуль також видає визначення трансформації, що використовується для перетворення значень показників. Ви можете повторно використовувати це перетворення в інших наборах даних, що мають ту саму схему, за допомогою модуля «Застосувати перетворення»(рис. 2.4).

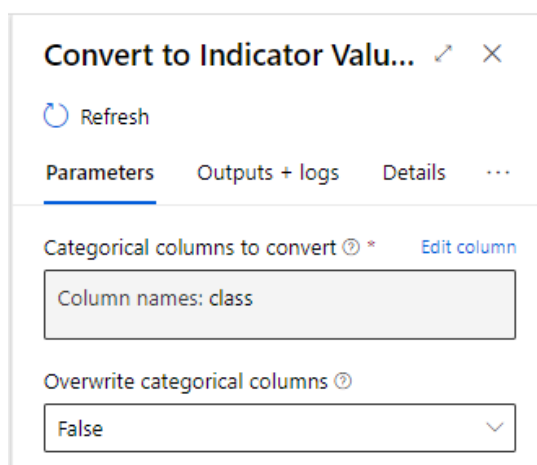


Рисунок 2.4 - Налаштування модуля для роботи з нашим набором даних

Select Columns in Dataset- цей модуль використовується, щоб вибрати підмножину стовпців для використання в подальших операціях. Модуль фізично не видаляє стовпці з вихідного набору даних; натомість він створює підмножину стовпців, подібно до бази даних або проекції.

Модуль корисний тим що, коли потрібно обмежити стовпці, доступні для подальшої операції, або якщо ви хочете зменшити розмір набору даних, видаливши непотрібні стовпці.

Стовпці в наборі даних виводяться в тому ж порядку, що й у вихідних даних, навіть якщо ви вказуєте їх в іншому порядку(рис. 2.5).

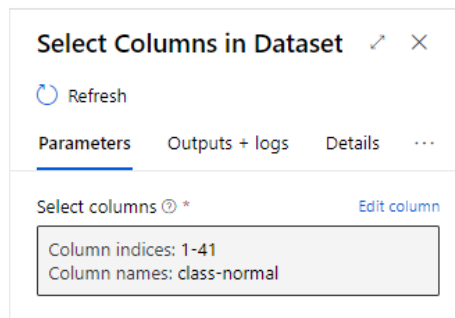


Рисунок 2.5 - Налаштування модуля для роботи з нашим набором даних

В результаті попередніх налаштувань та підвантаження даних модель в робочому полі буде мати вигляд дерева(рис. 2.6).

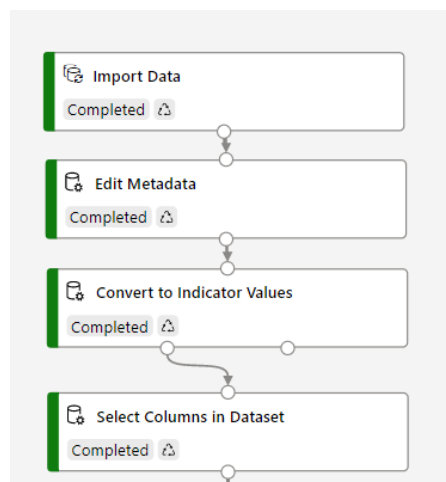


Рисунок 2.6 - Модель завантаження та попередньої обробки даних

Після проведенних налаштувань можна віалізувати загальний вид завантажених даних(рис. 2.7)



Рисунок 2.7 - Візуалізація завантажених та оброблених даних

2.3. . Розробка моделі для класифікації

Конструктор Машинного навчання Azure має цілий ряд модулів двокласової класифікації, які можуть бути використані для визначення аномалій. Їх основні характеристики:

1. Підсилена модель дерева прийняття рішень (Two-Class Boosted Decision Tree module) - це метод навчання «ансамблю», при якому друге дерево виправляє помилки першого дерева, третє дерево виправляє помилки першого та другого дерев тощо. Прогнози базуються на цілому «ансамблі» дерев, що робить прогноз точніше.

Як правило, при правильній конфігурації підсилені дерева рішень є найпростішим методом, за допомогою якого можна досягти максимальної продуктивності в широкому спектрі завдань машинного навчання. Однак вони також є одними з найбільш інтенсивних в навчанні, і поточна реалізація зберігає все в пам'яті. Тому підсилена модель дерева прийняття рішень може бути не в змозі обробити великі набори даних, які можуть обробляти деякі лінійні способи навчання.

2. Машини опорних векторів (Two-Class Support Vector Machine module)- Це модель SVM є контрольованою моделлю навчання, яка вимагає маркованих даних. У процесі навчання алгоритм аналізує вхідні дані та розпізнає закономірності у багатовимірному просторі об'єктів, який називається гіперплощиною. Усі приклади введення представлені у вигляді точок у цьому просторі та зіставляються з вихідними категоріями таким чином, що категорії поділяються на якомога більш широкий та чистий пробіл.

Для прогнозування алгоритм SVM призначає нові приклади до тієї чи іншої категорії, відображаючи їх у тому самому просторі.

3. Усереднений перцептрон (Two-Class Averaged Perceptron module) - цей алгоритм класифікації є контрольованим методом навчання і вимагає додавання тегу до набору даних, який включає стовпець міток. Ви можете навчити модель, надавши модель та позначений набір даних як вхідні дані для

навчання моделі. Навчена модель потім може бути використана для прогнозування значень для нових входних прикладів.

4. Модель лісу прийняття рішень (Two-Class Decision Forest module) - це швидкі, контрольовані моделі ансамблів. Цей модуль є хорошим вибором, якщо ви хочете передбачити ціль із максимум двома результатами.

5. Нейронна мережа (Two-Class Neural Network module) - класифікація за допомогою нейронних мереж є контрольованим методом навчання, і тому потрібен тегований набір даних, який включає стовпець міток. Наприклад, ви можете використовувати цю модель нейронної мережі для прогнозування бінарних результатів, таких як наявність у пацієнта певної хвороби, чи може машина вийти з ладу протягом певного періоду часу.

6. Логістична регресія (Two-Class Logistic Regression module) - це добре відома статистична техніка, яка використовується для моделювання багатьох видів проблем. Цей алгоритм є контрольованим методом навчання. Отже, наданий набір даних уже повинен містити результати для навчання моделі.

Кожен з цих модулів повинен бути навчений. Навчання моделі здійснюється за допомогою модуля «Навчання моделі» (Train Model). Переміщаємо його на полотно експерименту. З'єднуємо виходи модулів класифікації з лівим входом модуля навчання моделі, а потім приєднуємо вихід (лівий порт) модуля «Поділ даних» (Split Data) до правого входу модуля навчання моделі.

Модуль «Оцінити модель» (Evaluate Model) використовується для вимірювання точності навченої моделі. Ви надаєте набір даних, що містить оцінки, згенеровані з моделі, а модуль «Оцінити модель» обчислює набір стандартних показників оцінки. Вибираємо модуль «Аналіз моделі» (Evaluate Model) і перетягуємо його на полотно експерименту, а потім з'єднуємо вихід модуля Оцінка моделі з лівим входом модуля Аналіз моделі (рис. 2.8).

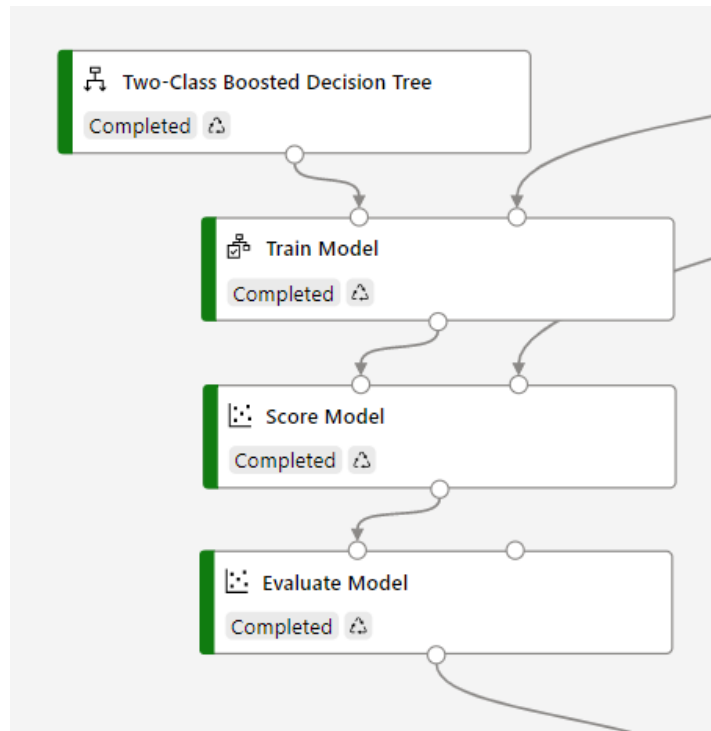


Рисунок 2.8 - Приклад підключення модуля Навчання моделі і Оцінки моделі

Для згрупування звітності досліджень всіх методів було б розумно використати модулі для об'єднання даних(рис.2.9).

Execute R Script - Виконує сценарій R з експерименту машинного навчання Azure. За допомогою R можна виконувати завдання, які не підтримуються існуючими модулями, наприклад:

- створювати власні перетворення даних;
- використовувати власні показники для оцінки прогнозів;
- створювати моделі, використовуючи алгоритми, які не реалізовані як окремі модулі в конструкторі.

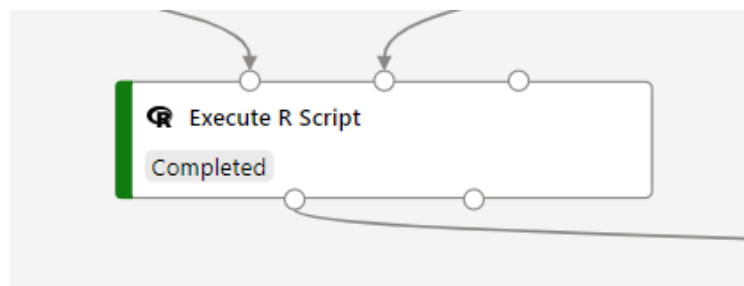


Рисунок 2.9 – Загальний вигляд модуля Execute R Script

Для групування даних використовується простий код для додавання стовпця назви алгоритма та зведення даних навчання двох алгоритмів(рис.2.10).

```

1
2 azureml_main <- function(dataframe1, dataframe2){
3     datainfo1 <- data.frame(Algorithm='Two-Class Boosted Decision Tree',dataframe1)
4     datainfo2 <- data.frame(Algorithm='Two-Class Support Vector Machine',dataframe2)
5
6     data.set = rbind(datainfo1, datainfo2);
7     plot(data.set);
8
9     return(list(dataset1=data.set, dataset2=dataframe2))
10 }

```

Рисунок 2.10 - Лістинг коду для модуля Execute R Script

Add Rows - цей модуль використовується для об'єднання двох наборів даних. Під час об'єднання рядки другого набору даних додаються в кінець першого набору даних.

Об'єднання рядків корисно в таких сценаріях:

- створили низку статистичних оцінок, і хочете об'єднати їх в одну таблицю для зручності звітування;
- працювали з різними наборами даних, і хочете їх, щоб створити кінцевий набір даних.

Так як ці сценарії збігаються з нашим завданням додаємо декілька модулів в робочий простір проекту та з'єднуємо їх між собою та з модулем Execute R Script.(рис.2.11).

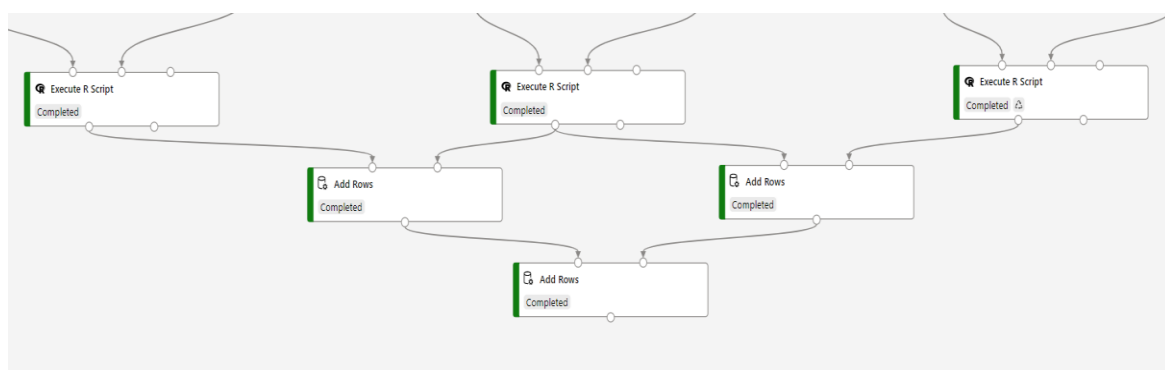


Рисунок 2.11 – Загальний вигляд з'єднання модулів Add Rows

РОЗДІЛ 3. ПЕРЕВІРКА ПРАЦЕЗДАТНОСТІ ТА ЕФЕКТИВНОСТІ РІШЕННЯ ЗАДАЧІ КЛАСИФІКАЦІЇ

3.1. Показники якості класифікації при виявленні аномалій

Коли говориться про оцінку якості моделі, ми говоримо або про регресійну модель (безперервний вихід), або про модель класифікації (номінальний або двійковий вихід). Показники оцінки, що використовуються в кожній із цих моделей, різні.

У задачах класифікації використовуються два типи алгоритмів (залежно від виду результату, який він створює):

1. Результат класом. Алгоритми, такі як SVM та KNN, створюють вихід класом. Наприклад, у двійковій задачі класифікації виходи будуть або 0, або 1. Однак є алгоритми, які можуть перетворити ці результати класу на ймовірність. Але ці алгоритми погано сприймаються статистичним співтовариством.

2. Результат вірогідністю. Алгоритми, такі як логістична регресія, випадковий ліс, посилення градієнта, Adaboost тощо, дають результати вірогідності. Перетворення вихідних даних імовірності у вихідні дані класу - це лише питання створення порогової ймовірності [14].

Проблеми моделі регресії не мають таких невідповідностей у результатах. Вихід завжди має безперервний характер і не потребує подальшої обробки.

В даній роботі використовувалися алгоритми двокласової класифікації тому результати будуть розглядатися за допомогою моделі класифікації.

Перший спосіб оцінки це матриця плутанини. Матриця плутанини - це матриця $N \times N$, де N - кількість прогнозованих класів. Для розглянутої задачі маємо $N = 2$, а отже, отримуємо матрицю 2×2 . Після проведення класифікації можливе отримання чотирьох видів результатів:

TP (True Positive – істинно позитивний),

TN (True Negative – істинно негативний),

FP (False Positive – псевдо позитивний),

FN (False Negative – псевдо негативний).

Вигляд матриці плутанини

	Прогнозовано НІ	Прогнозовано ТАК
Насправді НІ	істинно негативний(TN)	Псевдо позитивний(FP)
Насправді ТАК	Псевдо негативний(FN)	Істинно позитивний(TP)

Ось декілька визначень, які потрібно пам'ятати для матриці плутанини:

- акуратність (Accuracy) - частка загальної кількості правильних прогнозів;

$$\text{Accuracy} = \frac{TP+TN}{Totalinstances} \quad (3.1)$$

- точність (Precision) - показує, як часто насправді передбачено правильно результат;

$$\text{Precision} = \frac{TP}{TotalPredictedYES} \quad (3.2)$$

- відкликання (повнота) або чутливість (Recall or Sensitivity) - частка фактично позитивних випадків, які були правильно визначені;

$$\text{Recall} = \frac{TP}{TotalActuallyYES} \quad (3.3)$$

- специфічність (Specificity) - частка фактично негативних випадків, які були правильно визначені.

$$\text{Specificity} = \frac{TN}{TotalActuallyNO} \quad (3.3)$$

Загалом, ми маємо справу з однією із визначених вище метрик. Наприклад, у фармацевтичній компанії їх більше турбує мінімальний неправильний позитивний діагноз. Отже, їх більше турбуватиме висока специфічність. З іншого боку, модель стирання більше стосуватиметься чутливості. Матриці плутанини зазвичай використовуються лише з моделями виводу класу.

F1-оцінка є середнім гармонічним значенням точності та відкликання для класифікаційної задачі.

Обчислюється як середньозважене значення точності та відкликання від 0 до 1, де ідеальне значення F-оцінки дорівнює 1.

$$F_1 = \left(\frac{\text{recall}^{-1} + \text{precision}^{-1}}{2} \right)^{-1} = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}, \quad (3.4)$$

де β - приймає значення в діапазоні $0 < \beta < 1$, якщо пріоритет віддається точності, і $\beta > 1$, якщо пріоритет віддається відкликанню.

F-оцінка досягає максимуму при повноті і точності, що дорівнює одиниці, і близька до нуля, якщо один з аргументів близький до нуля.

Площа під кривою ROC (AUC - ROC) - одна з найпопулярніших метрик, що використовуються в галузі. Найбільша перевага використання кривої ROC полягає в тому, що вона не залежить від зміни частки респондентів.

Для розрахунку кривої буде використовуватися 2 значення з матриці плутанини:

- чутливість;
- специфічність.

Отже, для кожної чутливості ми отримуємо різну специфічність (рис.3.1)

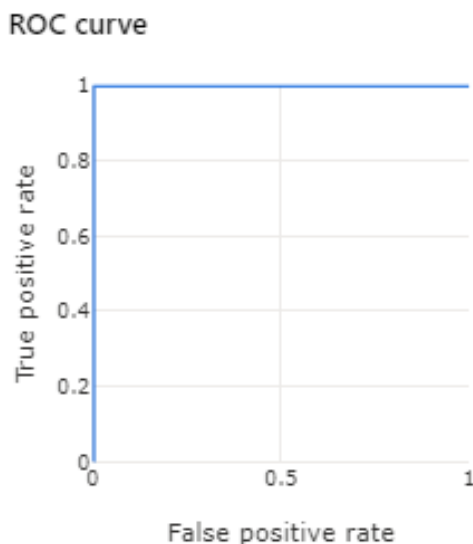


Рисунок 3.1 – Приклад Приклад вигляду ROC -кривої

Критерій AUC-ROC стійкий до незбалансованих класів і може бути інтерпретований як ймовірність того, що випадково обраний позитивний об'єкт

буде ранжований класифікатором вище (матиме більш високу ймовірність бути позитивним), ніж випадково обраний негативний об'єкт [14].

Крім ROC-кривої існує PR-крива (від англ. Precision-Recall Curve), що показує відношення точності до повноти [15].

Крива точність-повнота відображає зв'язок між точністю і повнотою при зміні порога прийняття рішення. Для вирішення деяких задач може знадобитися більш висока повнота, а для деяких - більш висока точність, в залежності від відносної важливості запобігання помилково негативні і хибнопозитивних результатів (рис.3.2).

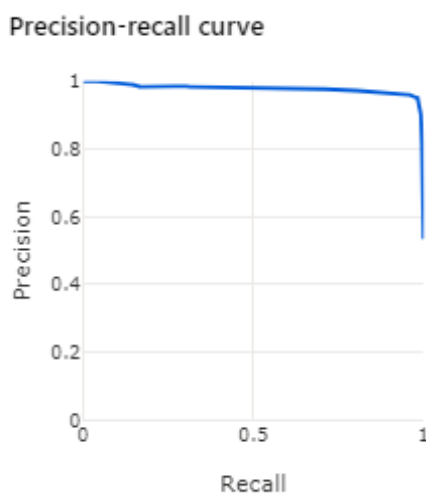


Рисунок 3.2 – Приклад вигляду PR-кривої

Крім вище згаданих показників наступні показники також подаються при оцінці моделей класифікації.

Середня втрата журналу - це одна оцінка, яка використовується для вираження покарання за неправильні результати. Вона обчислюється як різниця між двома розподілами ймовірностей - істинним і тим, що є в моделі.

Втрата журналу тренувань - це одна оцінка, яка представляє перевагу класифікатора перед випадковим передбаченням. Втрата журналу вимірює невизначеність вашої моделі шляхом порівняння ймовірностей, які вона видає, із відомими значеннями (основна істина) в ярликах. Необхідно мінімізувати втрати журналу для моделі в цілому.

3.2. Дослідження та порівняння алгоритмів прогнозування студії машинного навчання Azure

Основна мета цього дослідження полягає в тому, щоб порівняти між собою показники якості алгоритмів двокласової класифікації за допомогою конструктора машинного навчання при вирішенні завдань кібербезпеки.

Загальний вигляд конвеєра експерименту в конструкторі машинного навчання Azure показаний в додатку 1.

Щоб оцінити якість створеної моделі потрібно запустити виконання експерименту. Після виконання всіх модулів перевіряємо вихідні дані аналізу модулів.

Це можна зробити двома способами:

1. Зберегти результати в базі даних набору даних для аналізу за допомогою інструментів вихідних даних.
2. Створити візуалізацію даних в інтерфейсі машинного навчання Azure.

Щоб перевірити вихідні дані модуля “Аналіз моделей” (Evaluate Model) в інтерфейсі, необхідно вказати порт виводу та вибрати елемент “Візуалізація” (Visualize).

Якщо ви підключите набори даних до обох входів Оцінка моделі, результати будуть містити метрики для обох наборів даних або обох моделей.

Розглянемо проміжні результати отримані для двійкового класифікатора з використанням алгоритму підтримка векторної машини (Two-Class Support Vector Machine)(рис.3.3 – 3.4).

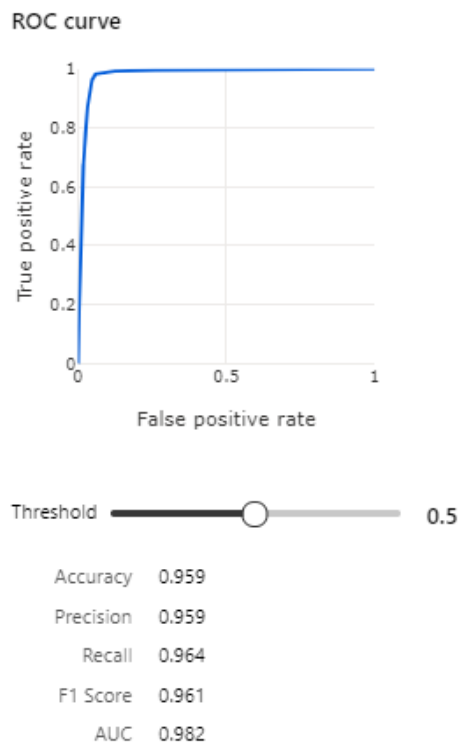


Рисунок 3.3 – Показана крива ROC для моделі двійкового класифікатора з використанням алгоритму підтримка векторної машини.

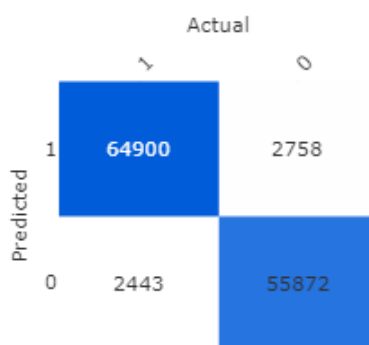


Рисунок 3.4 – Показана матриця плутанини для моделі двійкового класифікатора з використанням алгоритму підтримка векторної машини.

Точність розглянутої проблеми виявляється 96%. Як видно з двох наведених вище таблиць, позитивне передбачувальне значення є дуже високим, негативне прогнозне значення є досить низьким. Те саме стосується чутливості та специфічності. Це в першу чергу визначається вибраним нами пороговим

значенням. Якщо ми зменшимо наше порогове значення, дві пари різних чисел різко зблизяться.

Змінюючи значення порога можна підрегулювати і криву ROC. В ідеальній моделі крива йде вгору зліва по верхньому краю, охоплюючи всю область діаграми. Чим більше область під кривою (може бути будь-яким значенням від 0 до 1), тим вища продуктивність моделі.

Це показує метрика AUC, яка приведена разом з іншими метриками.

Якщо метрика AUC для моделі має більше значення, ніж значення для моделі бінарної класифікації (0,5), то в цьому випадку модель більш ефективна, ніж вгадування. Чим ближче AUC до 1 тим краще модель розпізнає атаки і вторгнення.

Як бачимо за прикладом поріг = 0,5.

Чутливість при цьому порозі становить 96,4%, а (1-специфічність) ~ 95%. Ця координата стає точкою на нашій кривій ROC. Щоб звести цю криву до одного числа, ми знаходимо площу під цією кривою (AUC).

Зверніть увагу, що площа всього квадрата дорівнює $1 * 1 = 1$. Отже, AUC сам по собі є відношенням під кривою та загальною площею. Для даного випадку ми отримуємо AUC ROC як 98,2%. Нижче наведено кілька оцінок відповідних значень:

- .90-1 = excellent (A);
- .80-.90 = good (B);
- .70-.80 = fair (C);
- .60-.70 = poor (D);
- .50-.60 = fail (F).

Бачимо, що ми підпадаємо під excellent діапазон для поточної моделі. Тобто навчання моделі можна вважати успішним.

Також на PR-кривій можна бачити аналогічний результат(рис.3.5)

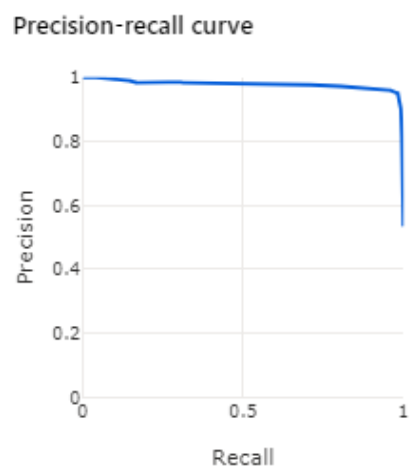


Рисунок 3.5 - Показана крива PR для моделі двійкового класифікатора з використанням алгоритму підтримка векторної машини.

Для зручності результати з всіх навчених моделей були зведені в єдину таблицю(рис 3.6).

Algorithm	AUC	Accuracy	Precision	Recall	F1	True.Negative	False.Positive	False.Negative	True.P
Two-Class Boosted Decision Tree	1	0.999889	0.999881	0.999911	0.999896	58622	8	6	67337
Two-Class Support Vector Machine	0.995213	0.953292	0.923756	0.994728	0.957929	53101	5529	355	66988
Two-Class Decision Forest	0.999999	0.99977	0.999777	0.999792	0.999785	58615	15	14	67329
Two-Class Support Vector Machine	0.995213	0.953292	0.923756	0.994728	0.957929	53101	5529	355	66988
Two-Class Decision Forest	0.999999	0.99977	0.999777	0.999792	0.999785	58615	15	14	67329
Two-Class Logistic Regression	0.996207	0.97278	0.966921	0.982701	0.974747	56366	2264	1165	66178
Two-Class Neural Network	0.999636	0.99276	0.992527	0.993941	0.993234	58126	504	408	66935

Рисунок 3.6 - Показана крива PR для моделі двійкового класифікатора з використанням алгоритму підтримка векторної машини.

Аналіз зведених результатів показує що:

Всі розглянуті алгоритми духклассової класифікації виявляють аномалії і вторгнення в більшій мірі, є методи, які повністю впорались із завданням. Це методи, що використовували для навчання алгоритми посиленого дерева

рішень (Two-Class Boosted Decision Tree) та лісу прийняття рішень (Two-Class Decision Forest). Ці методи показали ідеальні властивості по виявленню аномалій та вторгнень(рис.3.7).

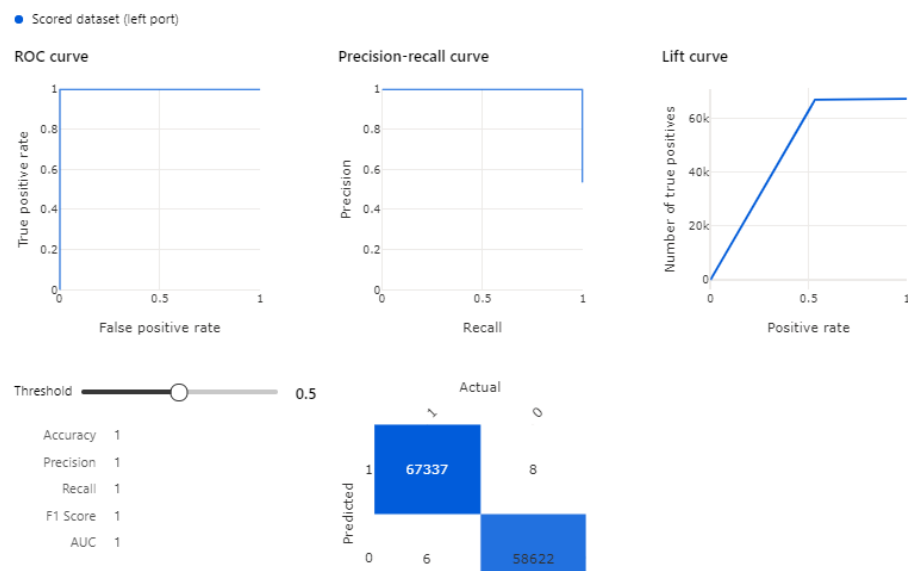


Рисунок 3.7 - Показані результати аналізу моделі двійкового класифікатора з використанням алгоритму посиленого дерева рішень.

Найгірші властивості по виявленню показав метод Двоклассовой підтримки векторної машини (Two-Class Support Vector Machine). Хоча результат AUC високий, але в порівнянні з іншими методами акуратність та точність мають достатньо низькі показники.

3.3. Рекомендації по застосуванню розробленої інформаційної технології

Незважаючи на те, що велика робота була пророблена з використанням незалежних алгоритмів, на сьогоднішній день широко використовуються гібридні підходи, оскільки вони забезпечують найкращі результати і долають недолік одного підходу в порівнянні з іншим. Кожен день спостерігаються нові невідомі атаки, і тому існує потреба в тих підходах, які можуть виявити невідому поведінку в наборі даних, що зберігаються, передаються або

змінюються. Використовується злиття або комбінація вже існуючих алгоритмів, які були запропоновані раніше.

Наприклад, існують різні нові підходи в модифікації дерев рішень (такі як ID3, C4.5), GA, SVM (включаючи оптимізовані і засновані на декількох ядрах підходи). Це може дати більш точні результати.

Використання будь-якого конкретного алгоритму саме по собі не дає належних результатів. Раз у раз реєструються нові атаки, при цьому використання якогось одного алгоритму буде недостатньо. В останні кілька років підходи були зроблені або шляхом об'єднання, або шляхом злиття різних алгоритмів разом.

Декілька прикладів методів створення нових алгоритмів.

Каскадні контрольовані методи: різні алгоритми класифікації об'єднуються разом, щоб отримати більш високу точність. Була запропонована комбінація наявного Байеса і алгоритму дерева рішень. Цей гібридний алгоритм був протестований в наборі даних Knowledge Data Discovery (KDD) cup, і досягнута точність склала 99 відсотків.

Також можна об'єднувати продуктивності наявного байєсівського класифікатора (NB) і алгоритму ID3. Також було запропоновано гібридний підхід злиття дерева рішень (DT) та машини опорних векторів (SVM). У ньому описувався ансамблевий підхід, в якому використовувалися дерево рішень (DT), машина опорних векторів (SVM) і гібридний класифікатор DT-SVM з очікуваннями. Ансамблевий підхід привів до 100-відсоткової точності тестованого набору даних. Можуть бути різні типи комбінацій, тому можна запропонувати безліч підходів, а найкращі результуючі підходи можуть бути реалізовані практично[16].

Комбінування контрольованих і неконтрольованих методів: існує ряд неконтрольованих і контрольованих алгоритмів навчання, комбінації яких можуть бути зроблені. В останні роки з'явилося багато подібних гібридних методів. Таким чином, ефективність контрольованого алгоритму значно підвищується, оскільки точність виявлення аномалій може бути значно

підвищена при використанні неконтрольованих алгоритмів. Для класифікації аномальних і нормальних дій в трафіку протоколу дозволу комп'ютерних адрес (ARP) була запропонована комбінація k-means та ID3, де була досягнута точність 98 відсотків.

Запропоновано новий підхід до виявлення мережових атак, метою якого є вивчення ефективності методу, заснованого на машинному навчанні при виявленні вторгнень, що включає штучні нейронні мережі і машини опорних векторів. Експериментальні результати, отримані при застосуванні цього підходу до набору даних KDD CUP '99, показують, що запропонований підхід забезпечує високу продуктивність, особливо для атак типу U2R і U2L. Було запропоновано гібридний підхід для об'єднання ентропії мережових ознак і SVM, який перевершив індивідуальні методи ентропії і SVM. Таким чином, гібридні підходи дають найкращі результати, оскільки об'єднують різні методи, долаючи недоліки один одного і приводячи до більш високої точності виявлення аномалій. У таблиці 3.2 представлені кілька гібридних підходів, запропонованих для виявлення аномалій[17].

Таблиця 3.2

Вигляд матриці плутанини

1.	2.	3.	4.
Ім'я автора	Використані методи	Методологія	Плюси і мінуси
Chitrakar, Roshan, and Chuanhe (2012)	Класифікація SVM і k-кластеризація медоїдов	Подібні екземпляри даних згруповані методом k-медоїдов і отримані кластери класифікуються у використанні класифікаторів SVM	Більш висока точність. Тимчасова складність більше, коли набір даних дуже великий.

Продовження табл. 3.2

1.	2.	3.	4.
Chitrakar, Roshan, and Chuanhe (2012)	Кластеризація k-Медоїдов і наївна байєсівську класифікація	Подібні екземпляри даних групуються за допомогою методу кластеризації k - Medoids. Отримані кластери класифікуються за допомогою наївного байєсівського класифікатора	Збільшення частоти виявлення та зменшення середнього часу помилкової тривоги. Важко передбачити, коли наївний байєсовський класифікатор знаходиться в різних середовищах.
Fu, Liu and Pannu(2012)	Один клас і два класи Підтримка Векторні Машин (В хмарні обчислення)	Перший клас SVM використовується для визначення оцінки аномалій. По-друге, детектор переучується, коли певні нові записи даних включаються в існуючий набір даних.	Він не вимагає попередньої історії збоїв і самоадаптується шляхом навчання на основі спостережуваних подій збою. Точність виявлення відмов не може досягати 100%.
Farid, Harbi, and Rahman (2010)	Наївний Байєс і рішення дерево для адаптивного вторгнення виявлення	Він виконує виявлення балансу і зберігає помилкові спрацьовування на прийнятному рівні для різних типів мережних атак.	Зведення до мінімуму помилкових спрацьовувань і максимізація швидкості виявлення балансу. Потрібно поліпшення хибнопозитивної швидкості для віддалених атак користувачів.

Таким чином, перспективним в подальших дослідженнях та розробках детектора аномалій є використання комбінованих (ансамблевих) алгоритмів машинного навчання для створення методів виявлення.

ВИСНОВКИ

В даній кваліфікаційній роботі проведено аналітичну роботу для доведення актуальності теми роботи, а саме розробки детектора аномалій на основі служб машинного навчання Azure для рішення завдань кібербезпеки. Також було проведено дослідження методів класифікації машинного навчання при використанні їх в детекторі аномалій для рішення задач кібербезпеки.

В результаті дослідження було виявлено, що данна тема має велику актуальність на сьогоднішній день. Дослідження теми машинного навчання показало що нові підходи та методи не перестають створюватися.

В роботі було розглянуто основні методи навчання машинного навчання, які використовуються в створенні нових більш ефективніших та зручніших методів.

Було досліджено існуючі методики виявлення аномалій та вторгнень з використанням машинного навчання. Була розглянута система виявлення вторгнень та досліджено природу аномалій. Також було досліджено процес виявлення аномалій та методики виявлення відхилень в досліджуваних даних. При використанні класифікації, кластеризації та методів на основі машинного навчання були виявленні основні етапи процесу виявлення аномалій.

Розроблена технологія виявлення аномалій і вторгнень, яка використовує методи машинного навчання конструктора машинного навчання Azure.

Модель конвеєра машинного навчання побудована на основі методів класифікації Конструктора машинного навчання Azure.

Оцінка якості навчання і дослідження ефективності класифікації різними методами проведена за допомогою показників: акуратність, точність (precision), повнота(чутливість), чутливість, специфічність.

Також була побудована ROC-крива, яка значно упростила процес аналізу продуктивності алгоритмів навчання та методів класифікації машинного навчання

У роботі використані методи візуального математичного моделювання хмарної середовища Microsoft Azure, методи класифікації машинного навчання, методи програмування.

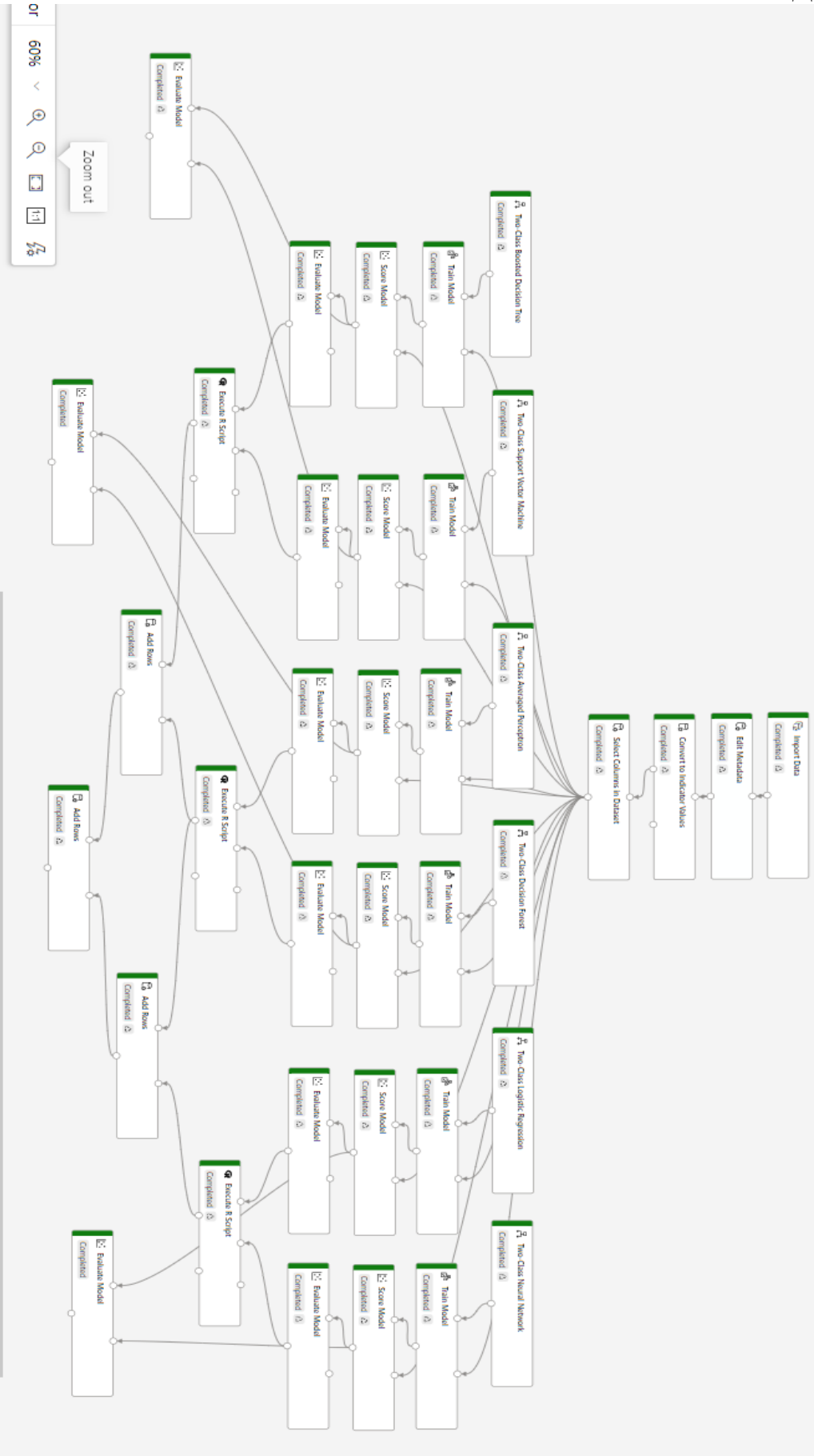
Таким чином, завдання на кваліфікаційну бакалаврську роботу виконано в повному обсязі.

					КБ.КР.Б – 125 – 21 – ПЗ	
						61

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Phua C., Lee V., Smith K., Gayler R., A comprehensive survey of data mining-based fraud detection ; research; 2010; p. 1-14.
2. Syarif I., Prugel-Bennett A., Wills G., Data mining approaches for network intrusion detection from dimensionality reduction to misuse and anomaly detection; Journal of Information Technology Review ; 3(2); 2012; p. 70-83.
3. Wu S. Y., Yen E., Data mining-based intrusion detectors; Expert Systems with Applications; 36(3); 2009; p. 5605-5612.
4. Amor N. B., Benferhat S., Elouedi Z., Naive Bayes vs decision trees in intrusion detection systems, In Proceedings of the ACM symposium on Applied computing; 2014; p. 420-424
5. Chitrakar R., . Chuanhe,H., Anomaly detection using Support Vector Machine classification with k-Medoids clustering; In Proceedings of IEEE Third Asian Himalayas International Conference on Internet (AH-ICI); 2012; p. 1-5.
6. Peddabachigari S., Abraham A., Grosan C., Thomas J., Modeling intrusion detection system using hybrid intelligent systems; Journal of network and computer applications; 30(1); 2007; p. 114-132.
7. Padhy N., Mishra P. , Panigrahi R., The Survey of Data Mining Applications and Feature Scope; International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), 2(3) ;2012; p. 43-58.
8. Ambusaidi, M., He, X., Tan, Z., Nanda, P., Lu, L. F., and Nagar, U. T., “A novel feature selection approach for intrusion detection data classification”,2014 IEEE 13th International Conference on Trust,Security and Privacy in Computing and Communications, 2014, pp.82-89.
9. Sharma, S. and Gupta, R. K.,“Intrusion Detection System: A Review”,International Journal of Security and Its Applications, vol.9, no.5,2015, pp.69-76.

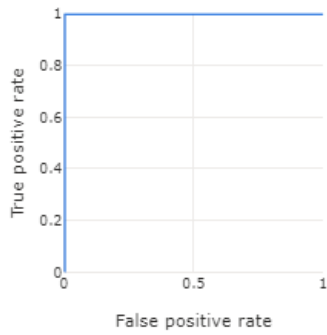
10. H. Kaur, G. Singh and J. Minhas, "A Review of Machine Learning Based Anomaly Detection Techniques" *International Journal of Computer Applications Technology and Research* volume 2-issue 2, 185-187, 2013
11. Top 10 BEST Intrusion Detection Systems (IDS) [2020 Rankings]. Available online: <https://www.softwaretestinghelp.com/intrusion-detection-systems/> (accessed on june 24, 2021)
12. S. Agrawal and J. Agrawal, "Survey on anomaly detection using data mining techniques", *Computer Science*, vol. 60, pp. 708-713, 2015.
13. Purohit and H. Gupta, "Hybrid intrusion detection system model using clustering classification and decision table", *IOSR Journal of Computer Engineering*, vol. 9, no. 4, pp. 103-107, 2013.
14. Meena, G., Choudhary, R.R., 2017. A review paper on ids classification using kdd 99 and nsl kdd dataset in weka, in: 2017 International Conference on Computer, Communications and Electronics (Comptelix), IEEE. pp. 553–558.
15. M. Fernández-Delgado, E. Cernadas, S. Barro and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?", *Journal of Machine Learning Research*, vol. 15, pp. 3133-3181, 2014.
16. Choudhary S., Kesswani N. A survey: Intrusion detection techniques for internet of things *International Journal of Information Security and Privacy(IJISP)*, 13 (2019), pp. 86-105
17. Tavallaee M., Bagheri E, Lu W, Ghorbani A.A. A detailed analysis of the kdd cup 99 data set 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications, IEEE (2009), pp. 1-6.



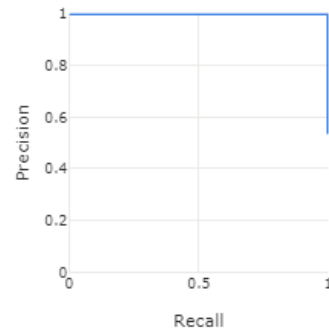
Two-class Boosted Decision Tree

● Scored dataset (left port)

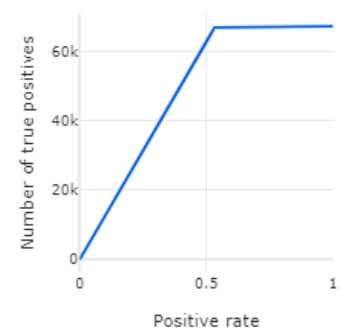
ROC curve



Precision-recall curve

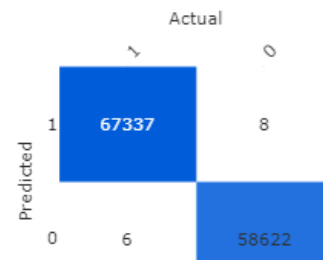


Lift curve



Threshold 0.5

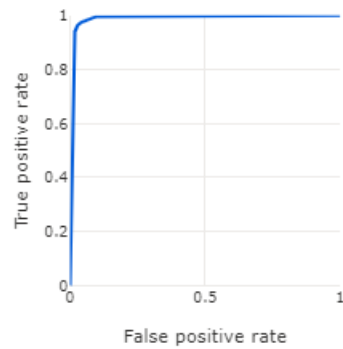
Accuracy 1
Precision 1
Recall 1
F1 Score 1
AUC 1



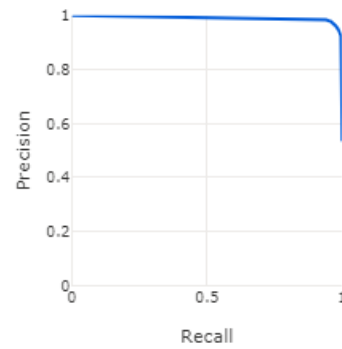
Two-class Averaged Perceptron

● Scored dataset (left port)

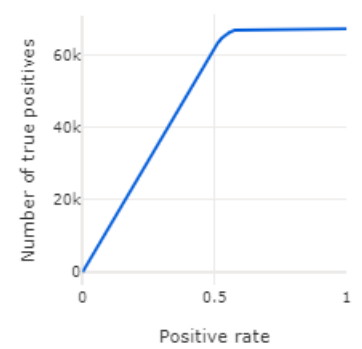
ROC curve



Precision-recall curve

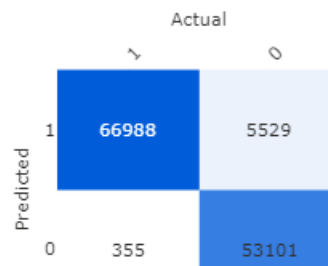


Lift curve



Threshold 0.5

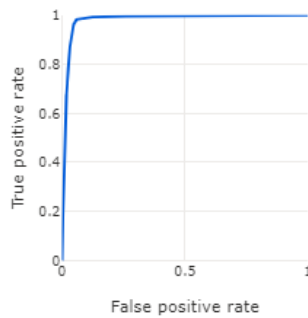
Accuracy 0.953
Precision 0.924
Recall 0.995
F1 Score 0.958
AUC 0.995



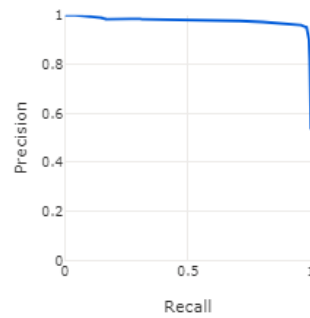
Two-class Support Vector Machine

● Scored dataset (left port)

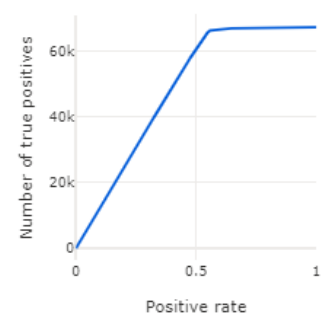
ROC curve



Precision-recall curve

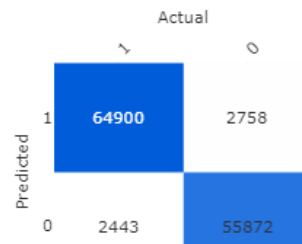


Lift curve



Threshold 0.5

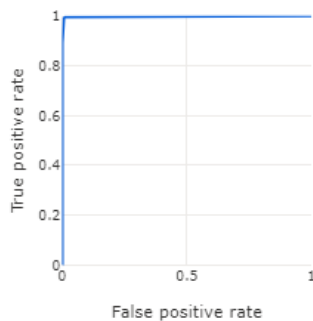
Accuracy 0.959
Precision 0.959
Recall 0.964
F1 Score 0.961
AUC 0.982



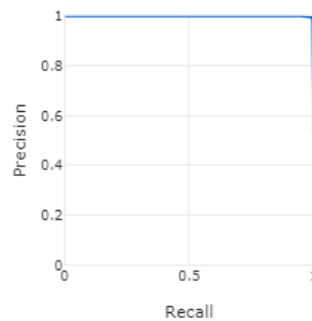
Two-class Decision Forest

● Scored dataset (left port)

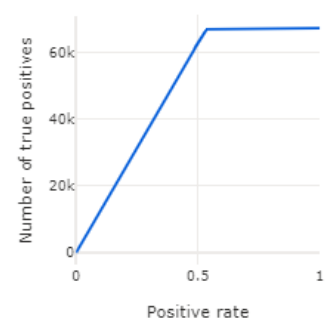
ROC curve



Precision-recall curve

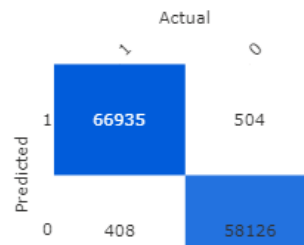


Lift curve



Threshold 0.5

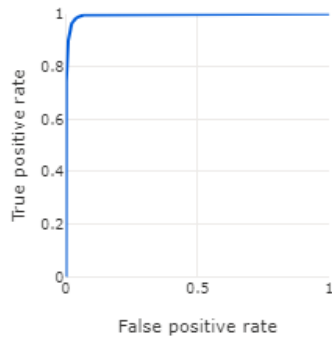
Accuracy 0.993
Precision 0.993
Recall 0.994
F1 Score 0.993
AUC 1



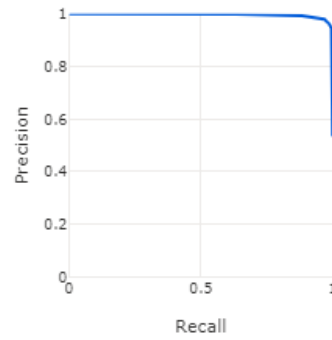
Two-class Logistic Regression

● Scored dataset (left port)

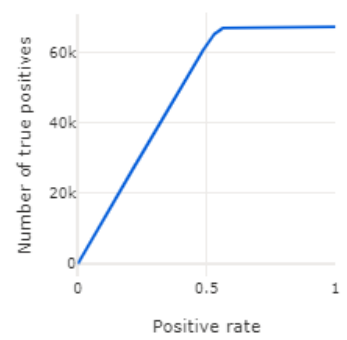
ROC curve



Precision-recall curve



Lift curve



Threshold 0.5

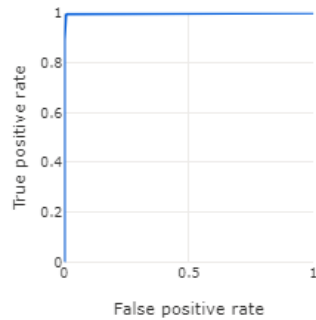
Accuracy 0.973
Precision 0.967
Recall 0.983
F1 Score 0.975
AUC 0.996

		Actual	
Predicted	1	66178	2264
	0	1165	56366

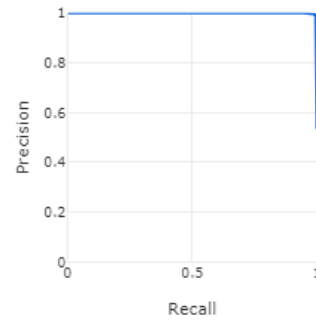
Two-class Neural Network

● Scored dataset (left port)

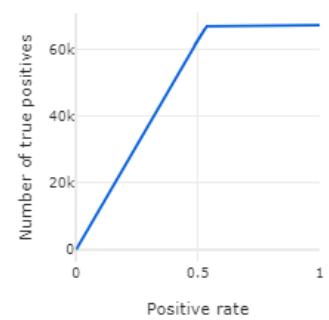
ROC curve



Precision-recall curve



Lift curve



Threshold 0.5

Accuracy 0.993
Precision 0.993
Recall 0.994
F1 Score 0.993
AUC 1

		Actual	
Predicted	1	66935	504
	0	408	58126