

ДОДАТОК А

Перелік джерел посилання за науковими напрямками керівника та науковців кафедри програмної інженерії

15. Smelyakov K., Filipov O. Comparative Analysis of the Applicability of Five Clustering Algorithms for Market Segmentation. 2023 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream), м. Vilnius, Lithuania, 27 квіт. 2023 р. 2023. URL: <https://doi.org/10.1109/estream59056.2023.10134796> (дата звернення: 19.04.2024).

ДОДАТОК Б

Звіт результатів перевірки на унікальність тексту в базі ХНУРЕ



Ім'я користувача:
Кардаш Євген Вікторович каф.ПІ

ID перевірки:
1016354996

Дата перевірки:
13.06.2024 06:37:12 EEST

Тип перевірки:
Doc vs Internet + Library

Дата звіту:
13.06.2024 06:41:12 EEST

ID користувача:
100013622

Назва документа: 2024_М_ПІ_ІПЗздм-22-1_Приходько_Я_О_скорочений

Кількість сторінок: 50 Кількість слів: 9542 Кількість символів: 74735 Розмір файлу: 1.84 MB ID файлу: 1016156309

Виявлено модифікації тексту (можуть впливати на відсоток схожості)

1.78%
Схожість

Найбільша схожість: 0.42% з Інтернет-джерелом (<https://juandomingofarnos.wordpress.com/tag/ia>)

1.77% Джерела з Інтернету	232	Сторінка 52
0.71% Джерела з Бібліотеки	38	Сторінка 54

0% Цитат

Вилучення цитат вимкнене

Вилучення списку бібліографічних посилань вимкнене

0%
Вилучень

Немає вилучених джерел

Модифікації

Виявлено модифікації тексту. Детальна інформація доступна в онлайн-звіті.

Замінені символи 4

Підозріле форматування 9 сторінок

ДОДАТОК В

Слайди презентації



Дослідження методів впровадження та підтримки інформаційної безпеки при використанні DevSecOps практик в організації розробки ІТ-стартапів

Приходько Ян Олегович, ІПЗдм-22-1
Науковий керівник: доц.кафедри ПІ Лановий О.Ф.

17 червня 2024



Аналіз предметної області та актуальність дослідження

- зростання кількості ІТ-стартапів та їх важлива роллю у сучасному технологічному середовищі;
- забезпечення та підтримка інформаційної безпеки – критичний чинник для успішного розвитку та конкурентоспроможності;
- DevSecOps практики все частіше використовуються для ефективного управління ризиками та забезпечення відповідного рівня інформаційної безпеки;
- зростаюча кількість відомих вразливостей у базах «CVE», «VulnDB», «NVD» - ключовий фактор для ідентифікації та аналізу сучасних кіберзагроз;
- необхідність створення моделі загроз на основі DevSecOps є простим та ефективним рішенням для ІТ-стартапів – це дозволить системно підійти до ідентифікації та оцінки загроз, а також визначати найбільш ефективні заходи для їх нейтралізації.

Аналіз предметної області та актуальність дослідження

Метою цієї роботи є: дослідження методів та інструментів DevSecOps і їхнього зв'язку з конкретними сценаріями ризику для ефективного впровадження та підтримки інформаційної безпеки в організації розробки IT-стартапів.

Об'єктом дослідження є процес забезпечення інформаційної безпеки в організації розробки IT-стартапів, включаючи виклики та загрози, які виникають внаслідок швидкого розвитку і обмежених ресурсів.

Постановка задачі

- **дослідження існуючих методів DevSecOps:**
 - огляд існуючих методів та інструментів DevSecOps, прикладів успішного впровадження
- **практичне дослідження загроз інформаційної безпеки на основі моделі даних «Common Vulnerabilities and Exposures» та інструментів DevSecOps:**
 - модель машинного навчання для класифікації загроз з бази CVE на основі маркування даних за інструментами DevSecOps
 - аналіз результатів, висновки щодо актуальності використання різних та інструментів DevSecOps;
- **розробка рекомендацій щодо застосування методів та інструментів DevSecOps у IT-стартапах**
 - сформулювати модель загроз для IT-стартапу, що базується на отриманих результатах

Принципи та методи DevSecOps

З розвитком DevOps та швидким розгортанням програмного забезпечення, виникла необхідність інтегрувати безпеку безпосередньо в процес розробки.

Принципи DevSecOps

- Зміщення безпеки вліво (Shifting Security Left)
- Тренінги з безпеки
- Культура робочого місця
- Спостереження (observability) та моніторинг
 - Visibility
 - Traceability
 - Auditability

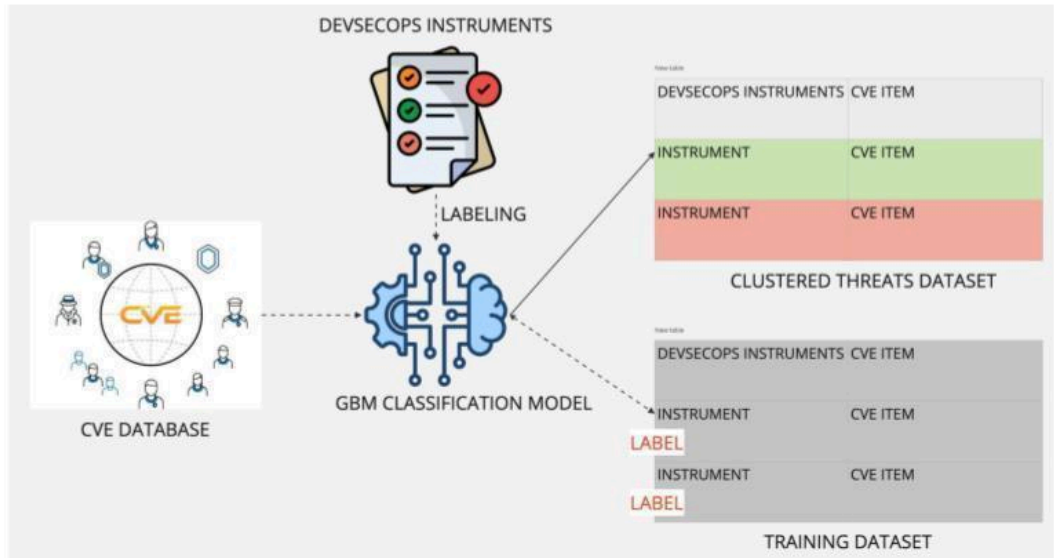
Методи DevSecOps

- моделювання загроз;
- тестування безпеки;
- аналіз і визначення пріоритетів;
- усунення наслідків;
- моніторинг.

Інструменти DevSecOps

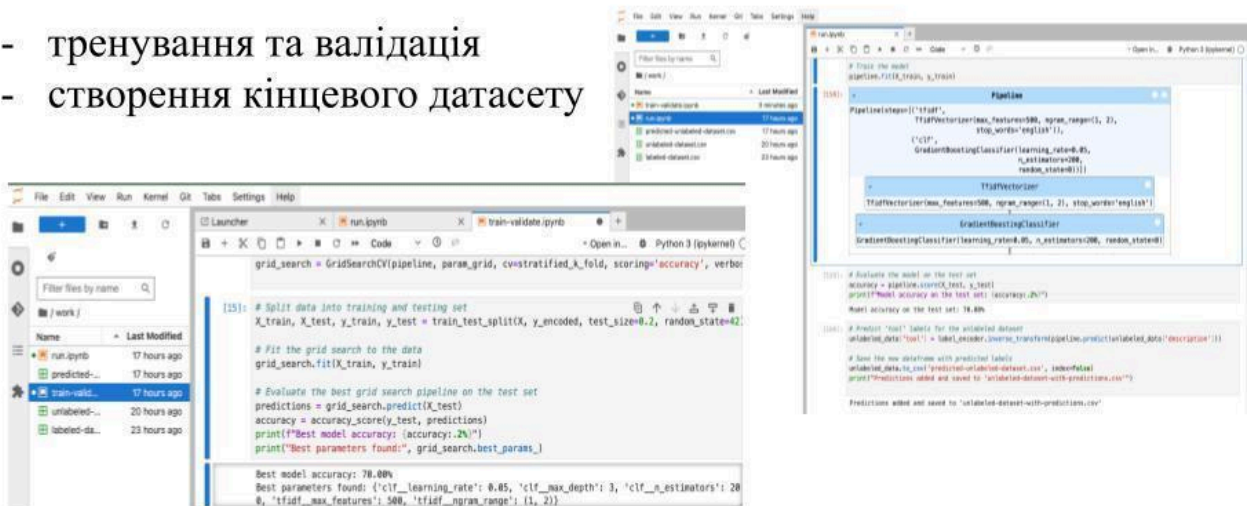
Метод	Інструмент	Приклад Інструменту
Моделювання Загроз	Інструменти для створення моделі загроз	IriusRisk, Microsoft Threat Modeling Tool, Pytm
Аналіз і визначення пріоритетів	Інструменти зберігання та моніторингу результатів тестування безпеки	SIEM
Тестування Безпеки	Інструменти сканування образів контейнерів	Snyk Container
Тестування Безпеки	Інструменти сканування операційних систем	OpenVAS
Тестування Безпеки	Інструменти сканування коду	Snyk Code
Тестування Безпеки	Інструменти сканування інфраструктур	ScoutSuite
Тестування Безпеки	Інструменти сканування коду інфраструктур	Chekhov
Тестування Безпеки	Інструменти тестування на проникнення	Pentester
Усунення наслідків	Інструменти захисту мережі	Cloudflare
Усунення наслідків	Інструменти виявлення аномалій та вторгнення	Snort
Усунення наслідків	Інструменти зберігання секретів	Vault
Усунення наслідків	Інструменти обмеження доступу	RBAC model
Моніторинг	Інструменти зберігання та перегляду логів	ELK Stack
Моніторинг	Інструменти зберігання та перегляду метрик	Grafana

Дослідження загроз інформаційної безпеки: CVE та інструменти DevSecOps



Дослідження загроз інформаційної безпеки: CVE та інструменти DevSecOps

- тренування та валідація
- створення кінцевого датасету



```

grid_search = GridSearchCV(pipeline, param_grid, cv=stratified_k_fold, scoring='accuracy', verbose=0)

[15]: # Split data into training and testing set
X_train, X_test, y_train, y_test = train_test_split(X, y_encoded, test_size=0.2, random_state=42)

# Fit the grid search to the data
grid_search.fit(X_train, y_train)

# Evaluate the best grid search pipeline on the test set
predictions = grid_search.predict(X_test)
accuracy = accuracy_score(y_test, predictions)
print(f"Best model accuracy: {accuracy:.2%}")
print(f"Best parameters found: {grid_search.best_params_}")

Best model accuracy: 78.88%
Best parameters found: {'clf__learning_rate': 0.05, 'clf__max_depth': 3, 'clf__n_estimators': 20, 'tfidf__max_features': 500, 'tfidf__ngram_range': (1, 2)}
  
```

```

# Train the model
pipeline.fit(X_train, y_train)

# Evaluate the model on the test set
accuracy = pipeline.score(X_test, y_test)
print(f"Model accuracy on the test set: {accuracy:.2%}")
Model accuracy on the test set: 78.88%

# Predict 'real' labels for the unlabeled dataset
unlabeled_data['real'] = label_encoder.inverse_transform(pipeline.predict(unlabeled_data['description']))

# Save the new dataset with predicted labels
unlabeled_data.to_csv('predicted-unlabeled-dataset.csv', index=False)
print("Predictions added and saved to 'predicted-unlabeled-dataset-with-predictions.csv'")

Predictions added and saved to 'predicted-unlabeled-dataset-with-predictions.csv'
  
```

Дослідження загроз інформаційної безпеки: CVE та інструменти DevSecOps

Структура даних кінцевого датасету (розподілена за роками)

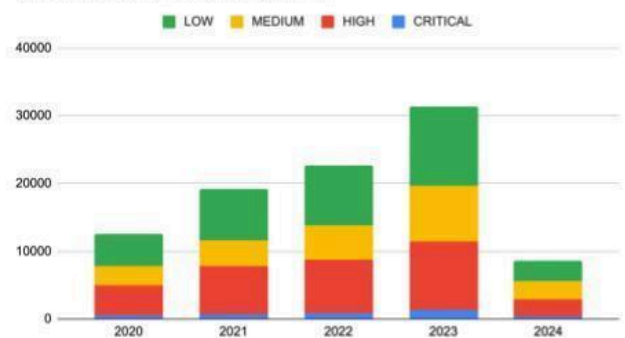
result-dataset-2024.csv

CVE	Description	Severity	Tool
CVE-2024-0007	A cross-site scripting (XSS) vulnerability in Palo Alto Networks	MEDIUM	Code scanning tools
CVE-2024-0008	Web sessions in the management interface ..	MEDIUM	Network protection tools
CVE-2024-0009	An improper verification vulnerability...	MEDIUM	Access restriction tools

Аналіз результатів дослідження

SEVERITY	YEAR				
	2020	2021	2022	2023	2024
CRITICAL	572	756	896	1323	318
HIGH	4422	7089	7896	10107	2658
MEDIUM	2843	3831	5039	8215	2672
LOW	4793	7535	8880	11728	2909

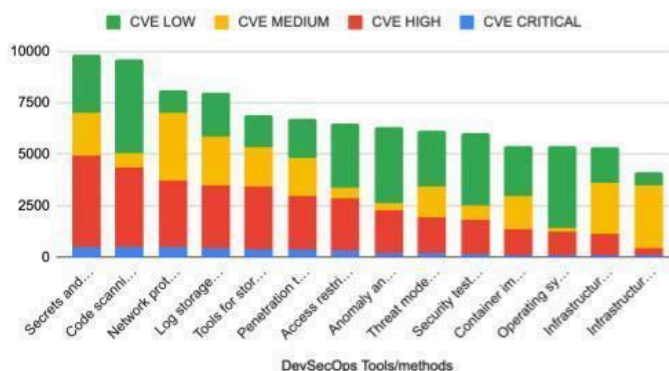
CVE SEVERITY over 2020-2024



Аналіз результатів дослідження

DevSecOps Tools/methods	CVE Critical	CVE High	CVE Medium	CVE Low	CVE TOTAL
Secrets and credentials storage tools	514	4406	2081	2819	9820
Code scanning tools	508	3849	699	4572	9628
Network protection tools	479	3272	3254	1127	8132
Log storage and viewing tools	439	3032	2395	2148	8014
Tools for storing and viewing metrics	395	3019	1939	1528	6881
Penetration testing tools	361	2606	1859	1912	6738
Access restriction tools	336	2544	479	3145	6504
....					

CVE SEVERITY BY DEVSECOPS TOOLS/METHODS



Застосування результатів дослідження - модель загроз (ТОП 5)

Загроза	Вразливість	Рейтинг та строки втілення	Інструменти та методи для усунення	Інтеграція
Витік секретів і облікових даних	Недостатній захист конфіденційної інформації	Високий, негайно	Secrets and credentials storage tools (Vault)	Централізоване сховище для секретів, автоматична ротація ключів
Вразливості в програмному коді	Недостатнє очищення вхідних даних, уразливості типу XSS, SQLi	Високий, протягом перших тижнів	Code scanning tools (Snyk Code)	Автоматичне сканування коду під час кожного коміту в репозиторій
Мережеві атаки (DDoS, Man-in-the-Middle)	Незахищені мережеві з'єднання	Високий, протягом перших місяців	Network protection tools (Cloudflare)	Використання мережевих фільтрів, VPN, системи запобігання DDoS
Недостатній моніторинг і аудит	Недостатнє логування подій	Високий, протягом перших тижнів	Log storage and viewing tools (ELK Stack)	Центральне сховище логів з можливістю їх перегляду і аналізу
Відсутність відстеження стану системи	Події, що загрожують функціям системи	Високий, протягом перших місяців	Tools for storing and viewing metrics (Grafana)	Налаштування дашбордів для моніторингу критичних метрик

Висновки

У результаті роботи, запропоновано шляхи підвищення рівня інформаційної безпеки IT-стартапів.

- виявлено ключові виклики інформаційної безпеки для IT-стартапів;
- проведено аналіз сучасних джерел загроз та вразливостей;
- виконано теоретичне дослідження методів DevSecOps на основі публікацій та статей;
- розглянуто приклади успішного впровадження методів DevSecOps;
- вивчено інструменти для впровадження методів DevSecOps;
- проведено практичне дослідження загроз інформаційної безпеки на основі моделі даних CVE та інструментів DevSecOps;
- проаналізовано застосування моделі для реального датасету для визначення актуальності методів DevSecOps;
- розроблено рекомендації щодо застосування методів DevSecOps у IT-стартапах на основі практичного та теоретичного дослідження;
- запропоновано новий підхід до інтеграції DevSecOps методів у процес розробки програмного забезпечення для IT-стартапі.

ДОДАТОК Г

Апробація результатів роботи

УДК 004.056.5:004.4

ІМПЛЕМЕНТАЦІЯ МЕТОДІВ DEVSECOPS В РОЗРОБКУ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ В СВІТЛІ СУЧАСНИХ ЗАГРОЗ

Приходько Я. О.

Науковий керівник – к.т.н., доц. Вечур О. В.

Харківський національний університет радіоелектроніки, каф. ПІ
м. Харків, Україна

e-mail: yan.prykhodko.cpe@nure.ua

This work is devoted to the analysis of the key information security challenges for IT startups and consideration on the implementation of the DevSecOps methods and instruments to ensure secure development and data security. It analyzes the modern threats database in order to find the correlation to the DevSecOps instruments that mitigate corresponding security risks which provides a possibility for IT-companies to concentrate on specific aspects of security measures implementation based on the limitations caused by modern company environments and teams.

У сучасному інформаційному середовищі, де IT-стартапи стають ключовим елементом інноваційного розвитку, питання інформаційної безпеки набувають особливої актуальності та значення. Інформаційна безпека не лише захищає конфіденційні дані та інтелектуальну власність, але й впливає на довіру клієнтів, інвесторів та партнерів. У світлі сучасних загроз, забезпечення належного рівня безпеки вимагає обізнаності, ефективного використання ресурсів і відповідних стратегій [1].

Відповідь на сучасні кіберзагрози стає надзвичайно важливою у світлі сучасних загроз інформаційній безпеці. Для впровадження та забезпечення відповідного рівня інформаційної безпеки в IT-стартапах, розробники все частіше звертаються до DevSecOps практик. Однією з ключових переваг DevSecOps є можливість виявлення вразливостей на ранніх стадіях розробки, що дозволяє запобігати серйозним проблемам безпеки в майбутньому. Основні принципи DevSecOps включають [2]: зміщення безпеки вліво (на початок життєвого циклу розробки); тренінги з безпеки та культура робочого місця; спостереження (observability) та моніторинг; моделювання загроз та тестування безпеки; аналіз і визначення пріоритетів та усунення наслідків.

З метою ідентифікації та кластеризації метою ідентифікації та кластеризації загроз можна використати базу загальновідомих вразливостей інформаційної безпеки (Common Vulnerabilities and Exposures) [4]. Це дозволить пов'язати загрози з інструментами DevSecOps [3], що мають найбільшу актуальність та зрозуміти і пріоритетувати заходи безпеки, необхідні для інформаційних систем.

Для аналізу використовуються датасет Kaggle “CVE (Common Vulnerabilities and Exposures)” [4] та основні особливості з датасету CVE:

числовий рейтинг вразливості (cvss) та текстовий опис вразливостей (summary). Виконується очищення та підготовка даних. Текстові дані перетворюються в числовий формат за допомогою TF-IDF векторизації, щоб вони могли бути проаналізовані за допомогою алгоритмів машинного навчання. Застосовується метод Principal Component Analysis (PCA) [5] для зменшення розмірності комбінованих даних (текст + cvss), щоб спростити візуалізацію та аналіз. Формула PCA:

$$X_{pca} = XW, \quad (1)$$

де X – вихідні дані, W – матриця ваг головних компонент.

Використовуючи метод кластеризації K-Means, дані розділяються на кластери, що представляють групи вразливостей з подібними характеристиками. Формула k-середніх:

$$\arg \min \sum_{i=1}^k \sum_{x \in S_i} \|x - \mu_i\|^2 \quad (2)$$

де S_i – i -тий кластер, x – точки даних, μ_i – центроїд i -того кластера.

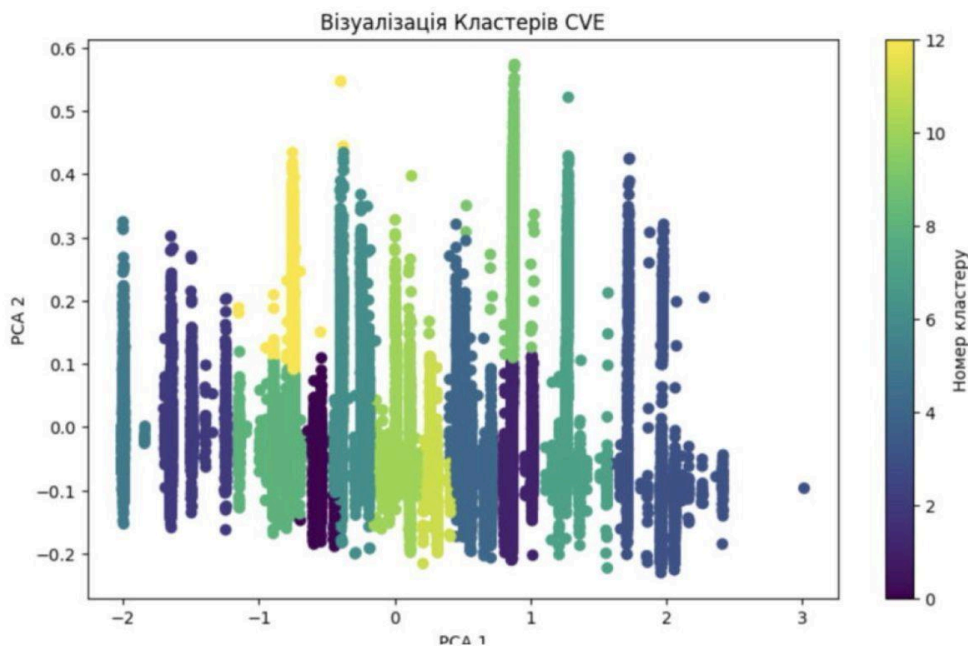


Рисунок 1 – Візуалізація результатів кластеризації

Дані, що були розподілені на кластери, марковані для відповідності інструментам DevSecOps[3] на основі ознак вразливостей (ключів) із бази загальновідомих вразливостей, відсортовані за рівнем актуальності відображено у таблиці 1:

Таблиця 1 – Результати кластеризації та відповідні інструменти

Кластер	Актуальність	Кількість вразливостей	Інструмент (назва)
5	10.000000	4504	Інструменти зберігання секретів
2	9.224815	7577	Інструменти сканування коду
8	7.582385	9479	Інструменти захисту мережі
12	7.501287	4847	Інструменти зберігання та перегляду логів
0	7.171997	3375	Інструменти зберігання та перегляду метрик
6	6.716870	12930	Інструменти тестування на проникнення

Таким чином, за результатами кластеризації, відповідно до актуальності інструментів, що покривають велику кількість DevSecOps, для усунення та запобігання загальновідомих загроз при впровадженні інформаційної безпеки проекту програмного забезпечення особливу увагу слід зосередити на інструментах для зберігання секретів, сканування коду, захисту мережі, моніторингу, а також інструментах тестування на проникнення. Це допомагає встановити пріоритети в інвестиціях у інструменти безпеки та стратегії DevSecOps.

Список використаних джерел:

1. Ayman Elsayah. “5 Problems With Startup Security” : вебсайт URL: <https://www.lastweekasavciso.com/p/5-problems-with-startup-security> (дата звернення 04.02.2024).
2. Microsoft. “What is DevSecOps?” : вебсайт. URL: <https://www.microsoft.com/en-us/security/business/security-101/what-is-devsecops> (дата звернення 04.02.2024).
3. Bright Security. “DevSecOps: Quick Guide to Process, Tools, and Best Practices” : вебсайт. URL: <https://www.hackerone.com/knowledge-center/devsecops-quick-guide-process-tools-and-best-practices> (дата звернення 04.02.2024).
4. Kaggle. “CVE (Common Vulnerabilities and Exposures)” : вебсайт URL: – <https://www.kaggle.com/datasets/andrewkronser/cve-common-vulnerabilities-and-exposures?select=cve.csv> (дата звернення 04.02.2024).
5. Christopher M. Bishop "Pattern Recognition and Machine Learning": 1st ed. 2006. Corr. 2nd printing 2011; Springer New York publisher. – 738 p.

ДОДАТОК Д.1

Програмний код для первинної обробки датасету «CVE»

```

const fs = require('fs');
const path = require('path');
const { parse } = require('json2csv');

async function generateCSV(inputDir, outputFile) {
  const files = fs.readdirSync(inputDir);
  let records = [];

  for (let file of files) {
    const filePath = path.join(inputDir, file);
    if (fs.statSync(filePath).isFile() && path.extname(filePath)
=== '.json') {
      let rawData = fs.readFileSync(filePath);
      let jsonData = JSON.parse(rawData);

      try {
        const cveId = jsonData.cveMetadata.cveId;
        const description =
jsonData.containers.cna.descriptions.find(desc => desc.lang ===
'en').value.replace(/\n/g, ' ').trim();
        const severity = jsonData.containers.cna.metrics.find(m
=> m.cvssV3_1.cvssV3_1.baseSeverity);
        if (cveId && description && severity) { // Check if
all required fields are available
          records.push({ CVE: cveId, description:
description, severity: severity });
        }
      } catch (error) {
        console.error(`Skipping file ${file}: Required data
is missing`);
      }
    }
  }
  if (records.length > 0) {
    const csv = parse(records, { fields: ["CVE", "description",
"severity"] });
    fs.appendFileSync(outputFile, csv + '\n', 'utf8');
    console.log(`CSV data has been appended to ${outputFile}`);
  } else {
    console.log("No valid records found to append.");
  }
}
// Example usage: node this_script.js
'./cvelistV5-main/cves/2024/0xxx' 'output.csv'
const inputDir = process.argv[2];
const outputFile = process.argv[3];
generateCSV(inputDir, outputFile);

```

ДОДАТОК Д.2

Програмний код для тренування та валідації моделі за алгоритмом Gradient Boosting Machines

```

import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.metrics import accuracy_score
from sklearn.model_selection import train_test_split, GridSearchCV,
StratifiedKFold

# Load the labeled data
labeled_data = pd.read_csv('labeled-dataset.csv')

# Check for missing values in text columns and fill with 'Missing'
text_columns = labeled_data.select_dtypes(include=['object']).columns
labeled_data[text_columns] =
labeled_data[text_columns].fillna('Missing')

# Extract features and labels
X = labeled_data['description']
y = labeled_data['tool']

# Encode the 'tool' labels
label_encoder = LabelEncoder()
y_encoded = label_encoder.fit_transform(y)

# Set up a pipeline with TF-IDF Vectorizer and Gradient Boosting
Classifier
pipeline = Pipeline([
    ('tfidf', TfidfVectorizer(stop_words='english')),
    ('clf', GradientBoostingClassifier(random_state=0))
])

# Define a grid of parameters to search (including both the vectorizer
and the classifier)
param_grid = {
    'tfidf__max_features': [500, 1000, None], # Example: number of
words to consider
    'tfidf__ngram_range': [(1,1), (1,2)], # Unigrams or bigrams
    'clf__n_estimators': [100, 200], # Number of boosting
stages to perform
    'clf__learning_rate': [0.05, 0.1, 0.2], # Learning rate
shrinks the contribution of each tree
    'clf__max_depth': [3, 5, 7] # Maximum depth of the
individual regression estimators
}

# Create a StratifiedKFold object to use for cross-validation

```

```
stratified_k_fold = StratifiedKFold(n_splits=5, shuffle=True,
random_state=42)

# Create a GridSearchCV object to evaluate the pipeline using the
parameter grid and StratifiedKFold
grid_search = GridSearchCV(pipeline, param_grid,
cv=stratified_k_fold, scoring='accuracy', verbose=1)

# Split data into training and testing set
X_train, X_test, y_train, y_test = train_test_split(X, y_encoded,
test_size=0.2, random_state=42)

# Fit the grid search to the data
grid_search.fit(X_train, y_train)

# Evaluate the best grid search pipeline on the test set
predictions = grid_search.predict(X_test)
accuracy = accuracy_score(y_test, predictions)
print(f"Best model accuracy: {accuracy:.2%}")
print("Best parameters found:", grid_search.best_params_)
```

ДОДАТОК Д.3

Повний програмний код моделі класифікації за алгоритмом Gradient Boosting
Machines

```

import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import LabelEncoder

# Load the labeled data
labeled_data = pd.read_csv('labeled-dataset.csv')

# Load the unlabeled data
unlabeled_data = pd.read_csv('unlabeled-dataset.csv')

# Check for missing values in text columns and fill with 'Missing'
text_columns = labeled_data.select_dtypes(include=['object']).columns
labeled_data[text_columns] = labeled_data[text_columns].fillna('Missing')
unlabeled_data[text_columns] = unlabeled_data[text_columns].fillna('Missing')

print(labeled_data.head())
print(unlabeled_data.head())

# Extract features and labels
X = labeled_data['description']
y = labeled_data['tool']

# Encode the 'tool' labels
label_encoder = LabelEncoder()
y_encoded = label_encoder.fit_transform(y)

# best_params = {
#     "clf__learning_rate": 0.05,
#     "clf__max_depth": 7,
#     "clf__n_estimators": 100,
#     "tfidf__max_features": 1000,
#     "tfidf__ngram_range": (1, 2),
# }

# Create a pipeline using the best parameters found from the grid
search
pipeline = Pipeline([
    ('tfidf', TfidfVectorizer(stop_words='english', max_features=500,
ngram_range=(1, 2))),
    ('clf', GradientBoostingClassifier(learning_rate=0.05,
max_depth=3, n_estimators=200, random_state=0))
])

```

```
# Split the data into training and test sets to evaluate the model
X_train, X_test, y_train, y_test = train_test_split(X, y_encoded,
test_size=0.2, random_state=42)

# Train the model
pipeline.fit(X_train, y_train)

# Evaluate the model on the test set
accuracy = pipeline.score(X_test, y_test)
print(f"Model accuracy on the test set: {accuracy:.2%}")

# Predict 'tool' labels for the unlabeled dataset
unlabeled_data['tool'] =
label_encoder.inverse_transform(pipeline.predict(unlabeled_data['desc
ription']))

# Save the new dataframe with predicted labels
unlabeled_data.to_csv('predicted-unlabeled-dataset.csv', index=False)
print("Predictions added and saved to
'unlabeled-dataset-with-predictions.csv")
```

