| МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ |
| --- |
| ДЕРЖАВНИЙ УНІВЕРСИТЕТ ІНФОРМАЦІЙНО-ТЕЛЕКОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ |

Навчально-науковий інститут захисту інформації

Кафедра інформаційної та кібернетичної безпеки

**Організація проведення наукових досліджень**

**ЗВІТ**

за результатами виконання лабораторної роботи на тему:

**“Проведення наукових досліджень в кібербезпеці”**

**Варіант 31**

Виконав: Терно Ярослав Анатолійович,

студент групи БСДМ-52

**Київ, 2024**

**Метою** необхідно виконати комплексну лабораторну роботу за варіантом у вигляді файлу (розширення .ipynb) та розмістити у системі Moodle. Для цього необхідно провести наступні етапи машинного навчання:

проаналізувати вхідний набір даних (Exploratory data analysis, EDA);

підготувати дані до обробки (Data Preprocessing);

застосувати алгоритм машинного навчання (Modelling);

оцінити оцінювання продуктивності моделей машинного навчання (Evaluation);

зробити висновки про доцільність застосування алгоритму машинного навчання для виявлення та реагування на шкідливі процеси в інформаційній системі.

**Завдання 1.**

Комплексну лабораторну роботу рекомендується виконати в середовищі Jupyter Notebook (<https://jupyter.org/>) або в середовищі Google Colaboratory (<https://colab.google/>).

**Завдання 2.**

Пропонуються наступні набори даних для застосування алгоритмів машинного навчання (див. варіант завдання нижче):

1. Wednesday-28-02-2018\_TrafficForML\_CICFlowMeter.csv

Детальна інформація про набори даних розміщено за наступним посиланням: <https://www.unb.ca/cic/datasets/ids-2018.html>

Також набори даних є доступними за наступним посиланням: <https://www.kaggle.com/datasets/solarmainframe/ids-intrusion-csv>

**Завдання 3.**

Пропонуються наступні алгоритми машинного навчання (див. варіант завдання нижче):

для розв’язання задачі класифікації:

7. метод k найближчих сусідів (k-Nearest Neighbours, kNN);

Детальна інформація про відповідні інструменти машинного навчання розміщено в бібліотеці scikit-learn (<https://scikit-learn.org/stable/>).

**Завдання 4.**

 Пропонуються наступні методи оцінювання продуктивності моделей машинного навчання (sklearn.metrics) (див. варіант завдання нижче):

3. обчислення показника F1 (f1\_score);

Детальна інформація про відповідні інструменти машинного навчання розміщено в бібліотеці scikit-learn (<https://scikit-learn.org/stable/>).

Виконання роботи:

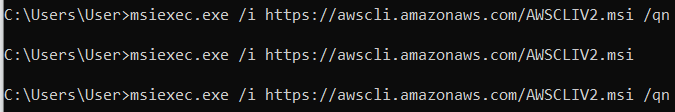
У ході виконання лабораторної роботи були отримані наступні результати:

1. Ознайомитися з інформацією про набори даних для дослідження шкідливих процесів в інформаційній системі організації з використання методів машинного навчання за посиланням:

CSE-CIC-IDS2018 on AWS: <https://www.unb.ca/cic/datasets/ids-2018.html>.

2. Для завантаження набору даних відповідно до вашого варіанту необхідно встановити AWS CLI. Інсталятор Ви знайдете за посиланням:

AWS Command Line Interface: <https://aws.amazon.com/cli/>.



Процес інсталяції aws на Windows 10

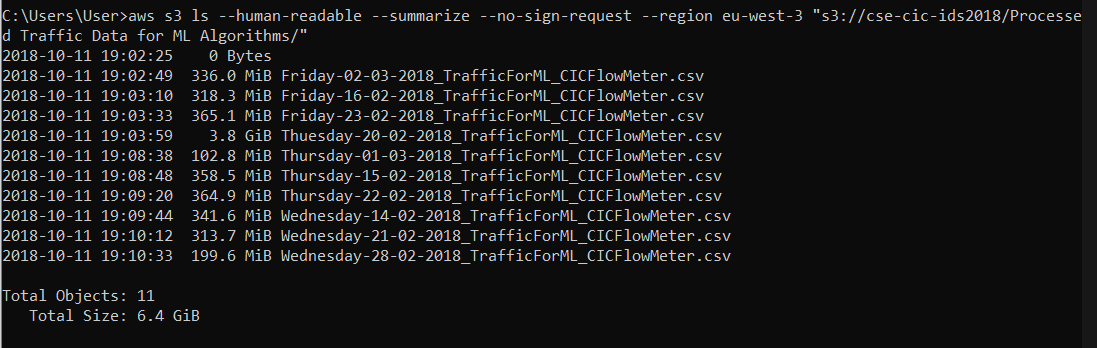
3. Щоб перевірити поточну встановлену версію, скористайтеся такою командою:





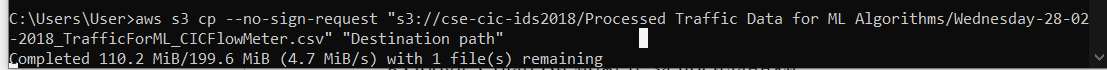
4. Для перегляду переліку наборів даних у форматі .csv скористайтеся такою командою:

aws s3 ls --human-readable --summarize --no-sign-request --region eu-west-3 "s3://cse-cic-ids2018/Processed Traffic Data for ML Algorithms/"

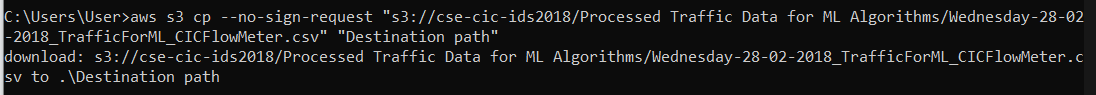


5. Для завантаження набору даних відповідно до вашого варіанту (наприклад, Wednesday-28-02-2018\_TrafficForML\_CICFlowMeter.csv) необхідно виконати таку команду:

aws s3 cp --no-sign-request "s3://cse-cic-ids2018/Processed Traffic Data for ML Algorithms/Wednesday-28-02-2018\_TrafficForML\_CICFlowMeter.csv" "Destination path"



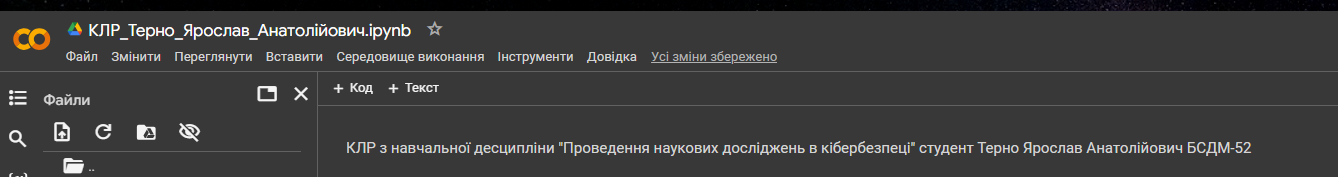
Процес завантаження потрібного нам файлу



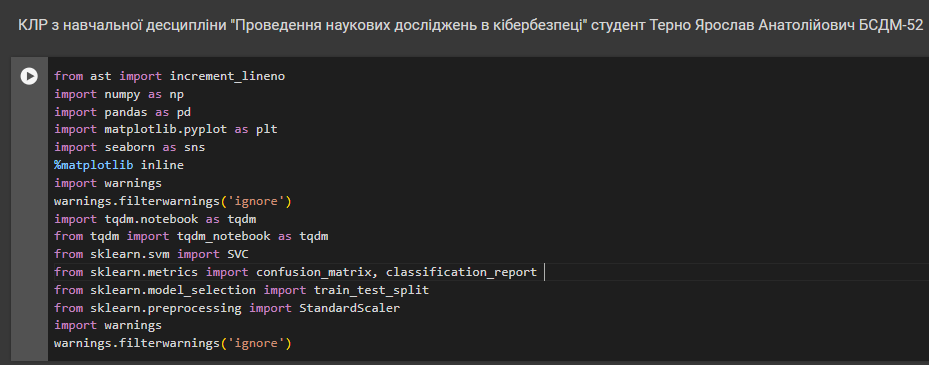
6. Для проведення дослідження із застосуванням методів машинного навчання необхідно застосовувати середовище Google Colab або інші. Для використання Google Colab необхідно мати обліковий запис Google. Перейти в Google Colab Ви можете за посиланням:

<https://colab.research.google.com/#scrollTo=5fCEDCU_qrC0>

7. У середовищі Google Colab необхідно створити новий записник (блокнот), який буде вашим звітом з комплексної лабораторної роботи.



8. Здійснюється імпорт потрібних бібліотек.



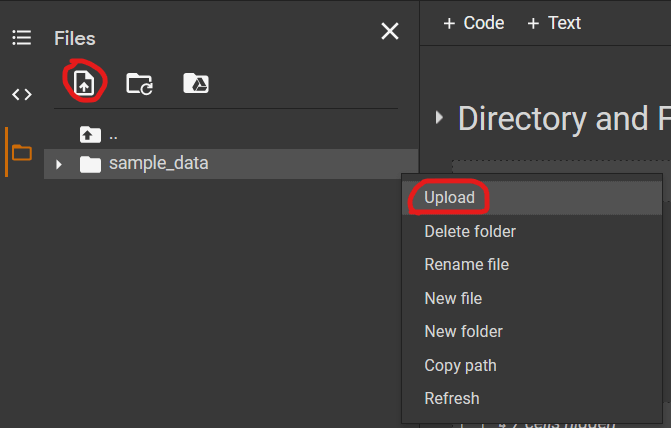
9. Здійснюється імпорт набору даних (попередньо скачаний відповідно варіанту завдання).

Для завантаження будь-яких файлів з локальної файлової системи до поточної робочої директорії Colab можна скористатися опцією Upload у верхній частині панелі менеджера файлів.

Для завантаження файлів безпосередньо в піддиректорію потрібно:

натисніть на три точки, що з'являються при наведенні курсору на каталог;

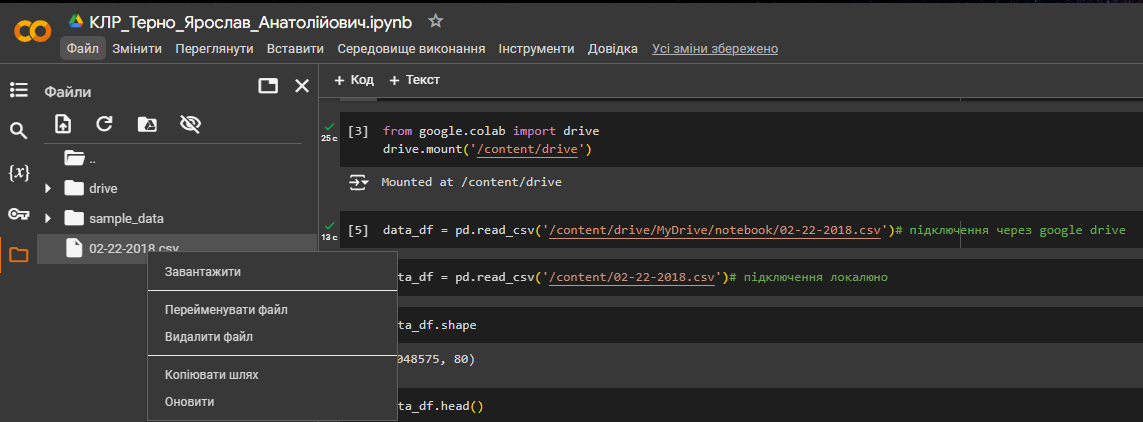
вибрати опцію Upload;



вибрати файли для завантаження із діалогового вікна File Upload;

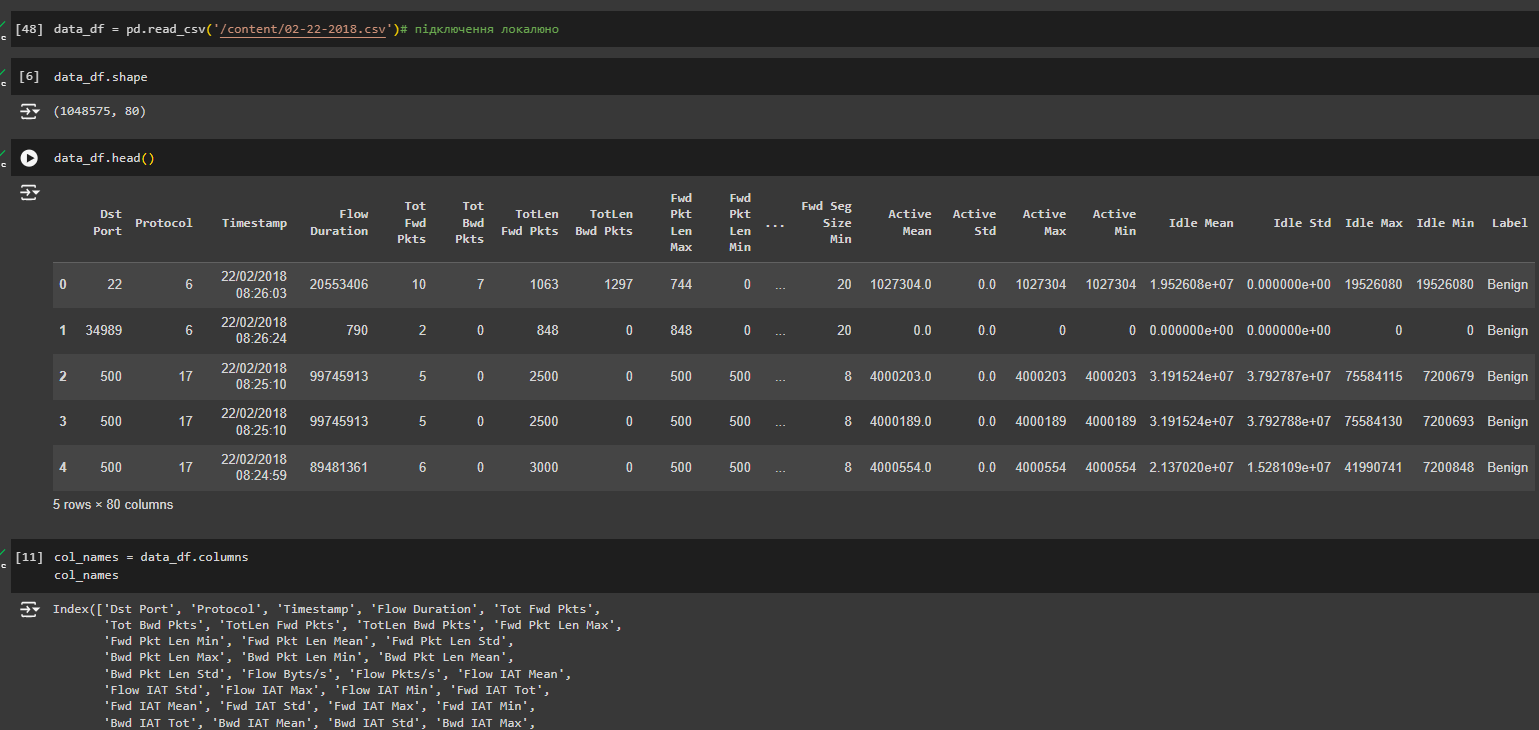
зачекати на завершення завантаження (звертаю увагу!), процес виконання якого відображається в нижній частині панелі менеджера файлів.

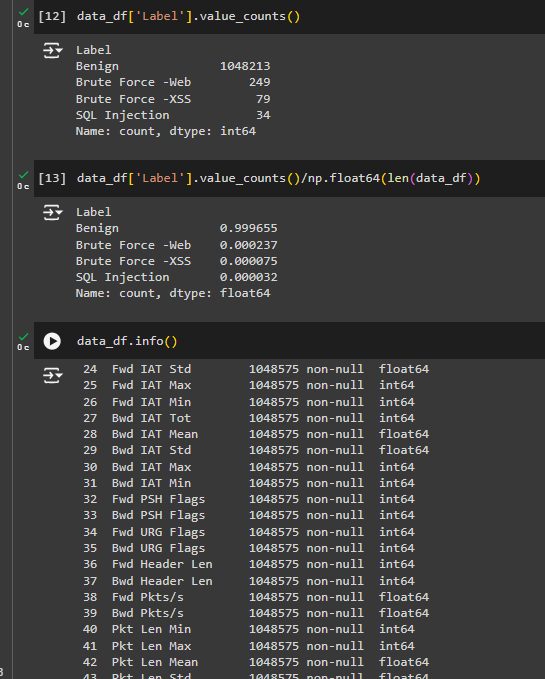
Після закінчення процесу завантаження Ви можете скопіювати шлях до датасету та відкрити в Colab. Завантажувати датасет необхідно кожного разу при входження в середовище Colab.



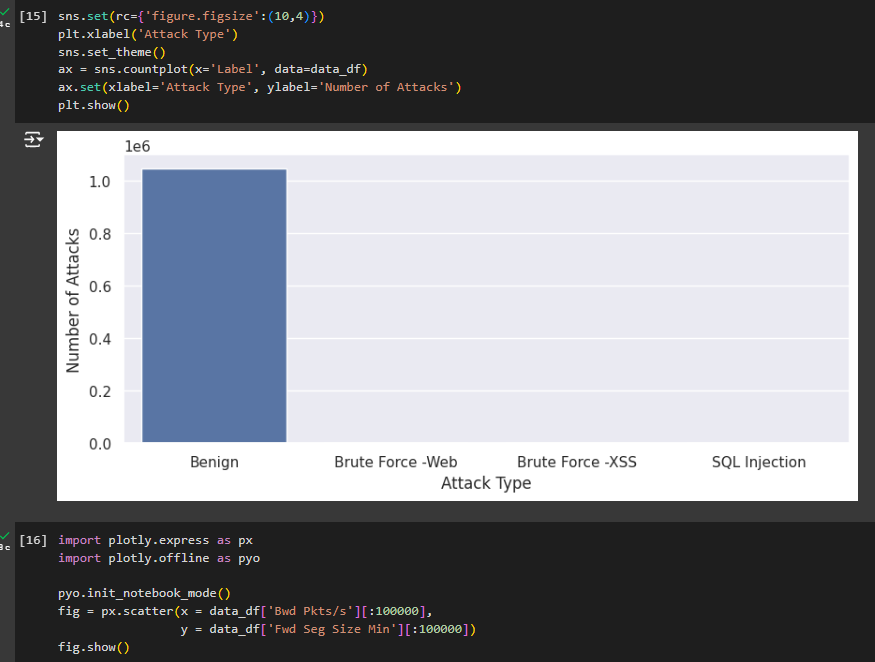
10. Проводиться дослідницький аналіз даних (Exploratory data analysis).

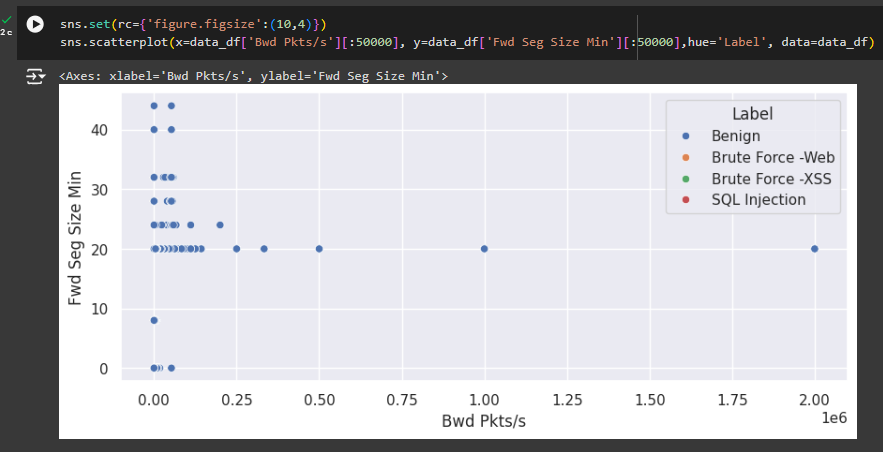
Дослідницький аналіз даних проводиться для правильного розуміння набору даних, який застосовується. Для цього здійснюється візуалізація даних, розуміння та їх аналіз для постановки прикладних задач (висунення гіпотез).



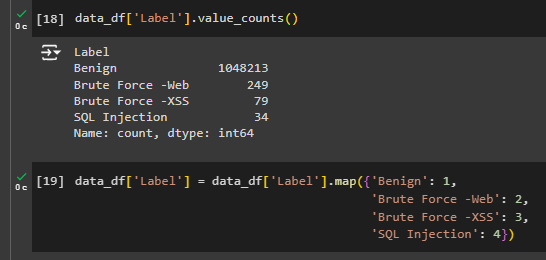


11. Здійснюємо візуалізацію даних. Отримавши деяку корисну інформацію про дані, необхідно створити візуальні зображення набору даних, щоб побачити, як розвивається тенденція в даних. Візуальні елементи включають стовпчасті діаграми, діаграми розподілу, точкові діаграми тощо.

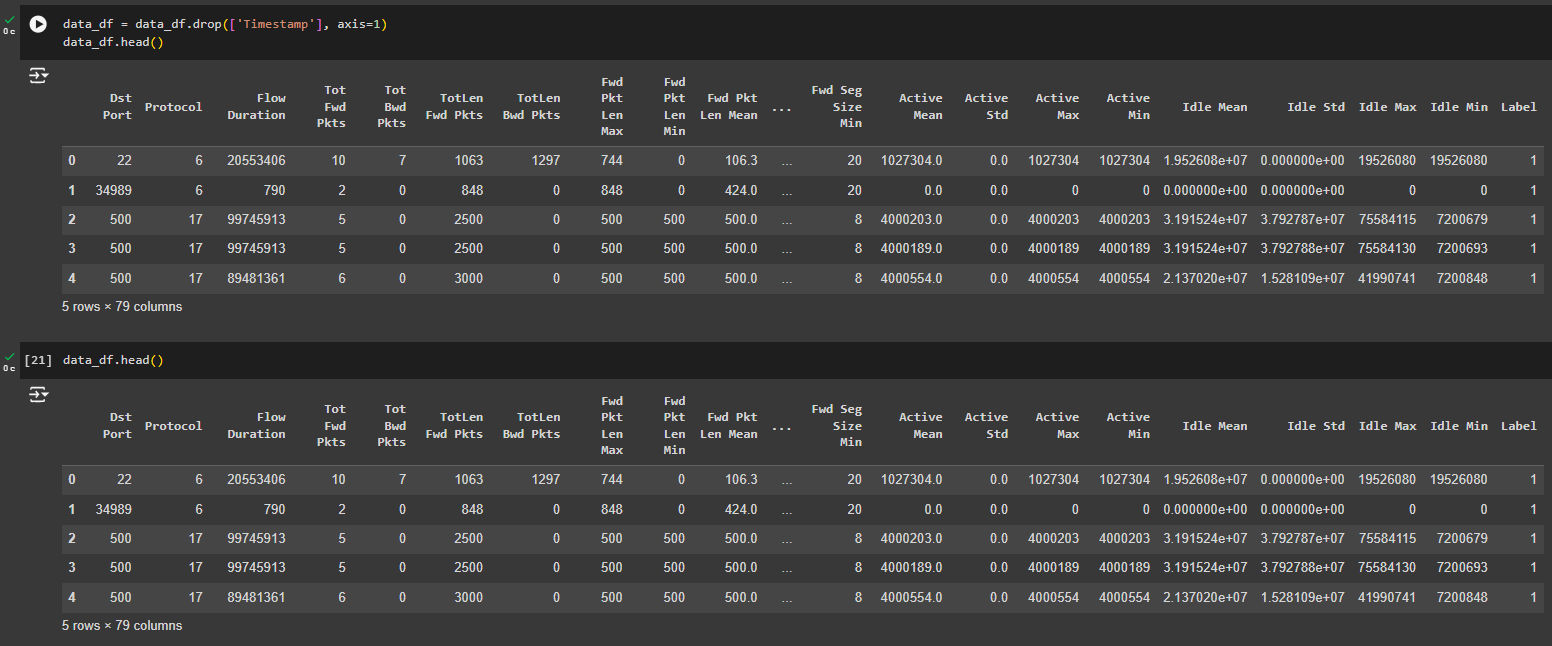


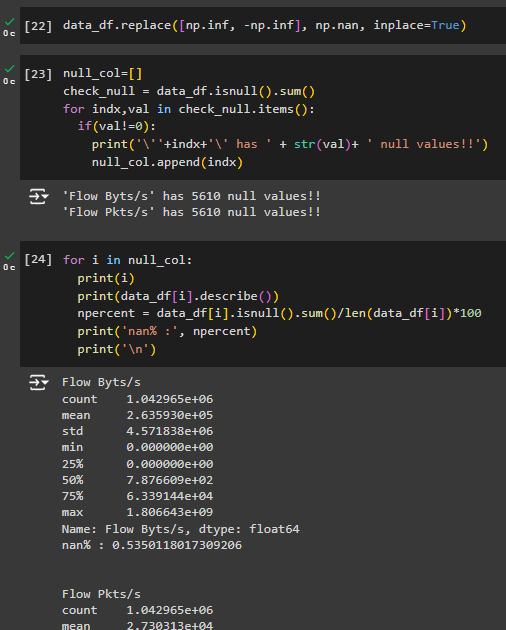


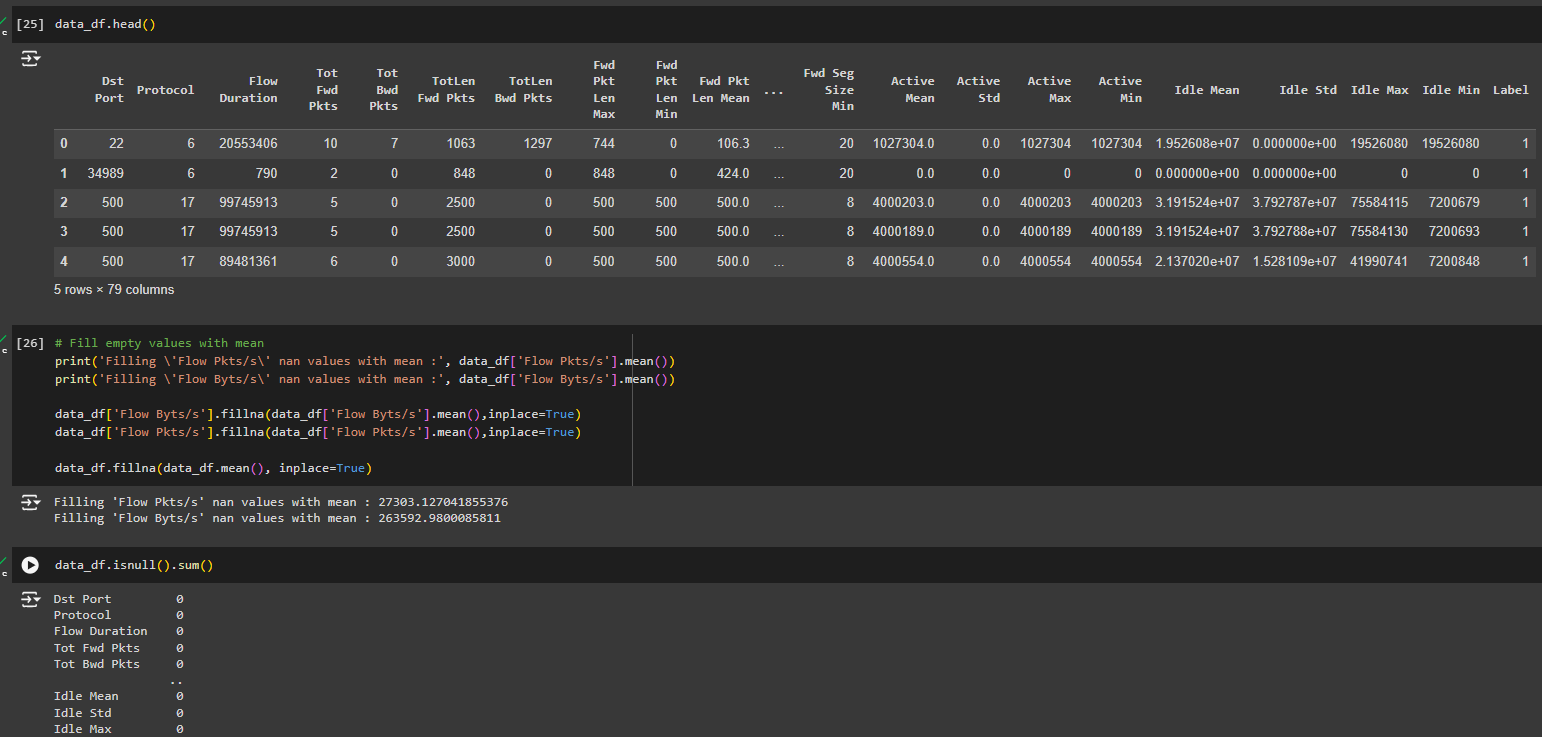
12. Здійснюємо попередню обробку даних. Попередня обробка даних відіграє важливу роль, оскільки дані можуть містити відсутні або нульові значення, а також викиди тощо.

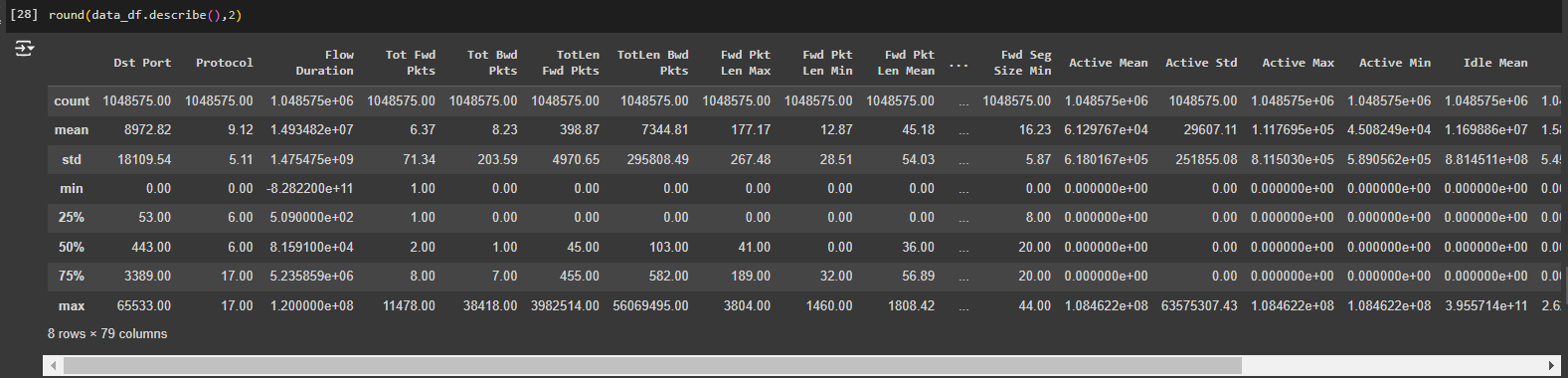


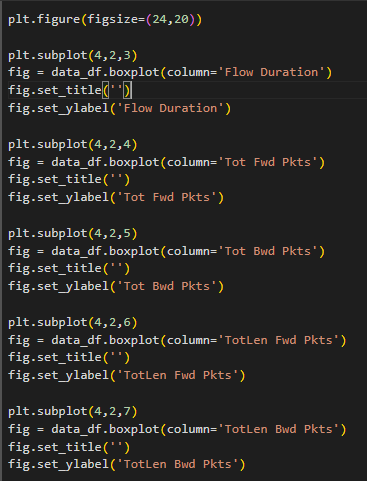
або так:

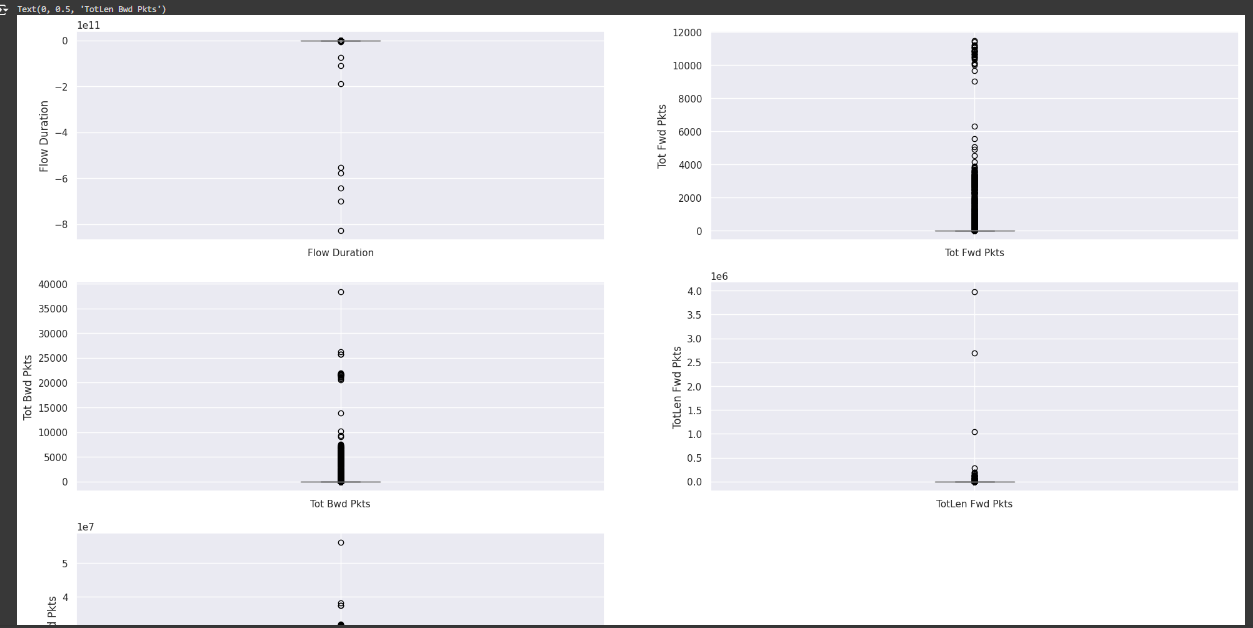


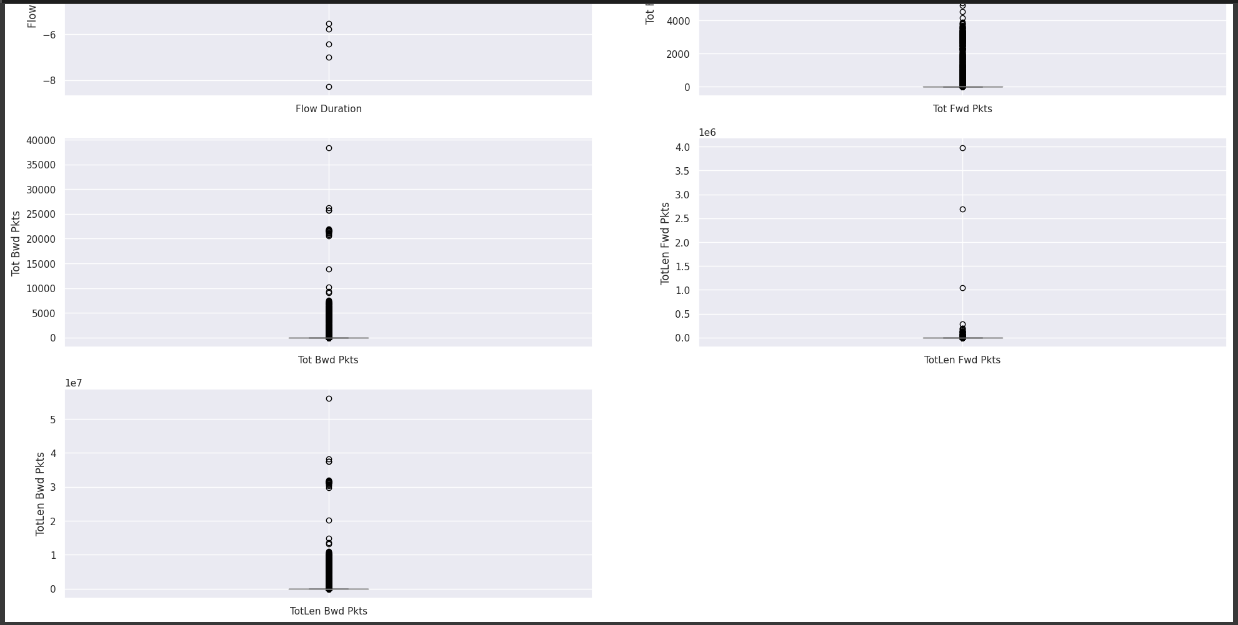


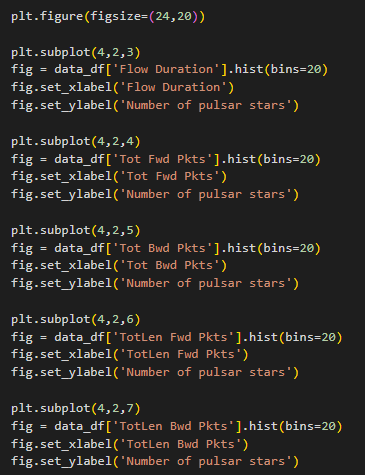


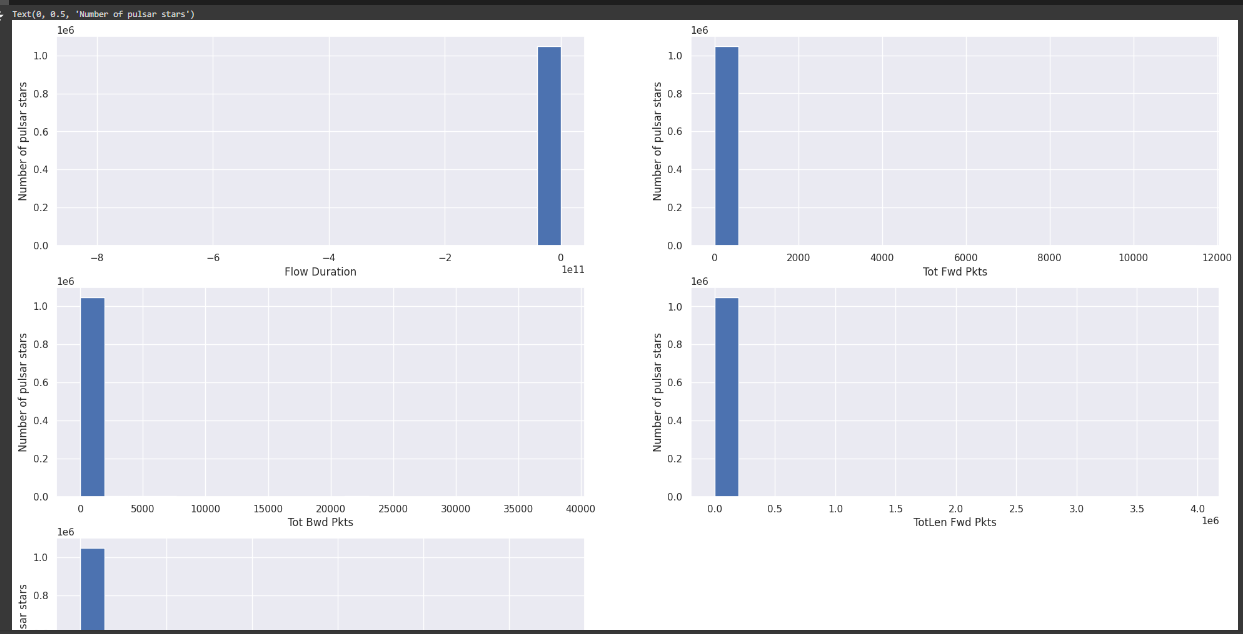


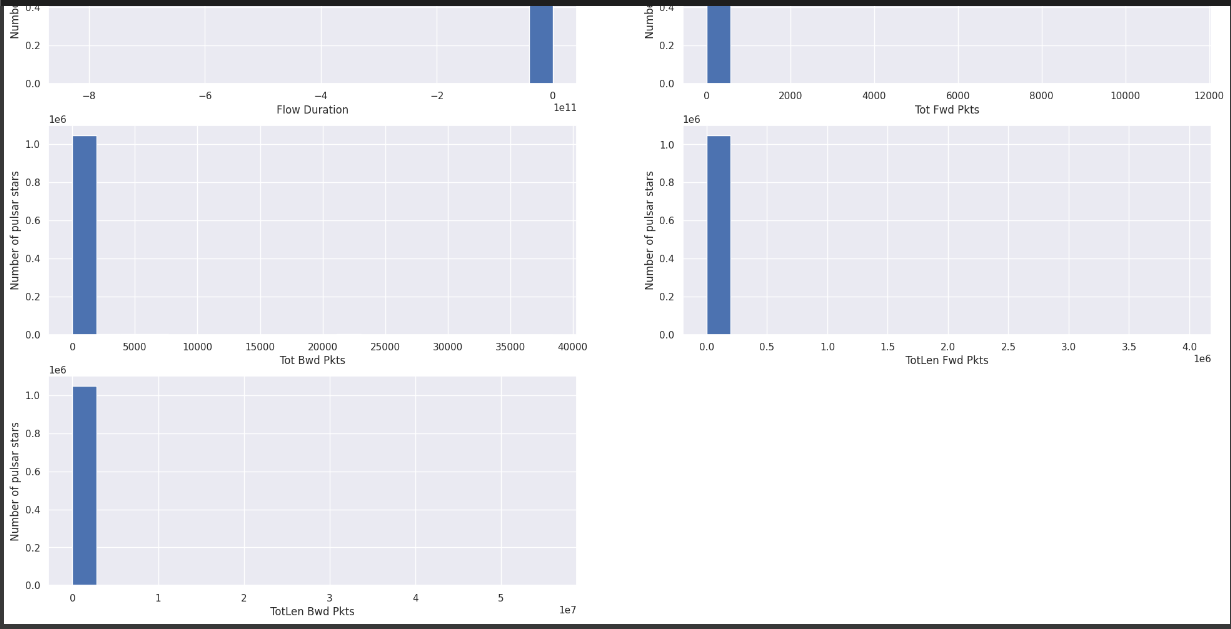




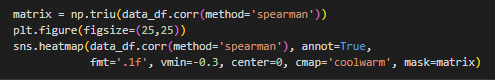


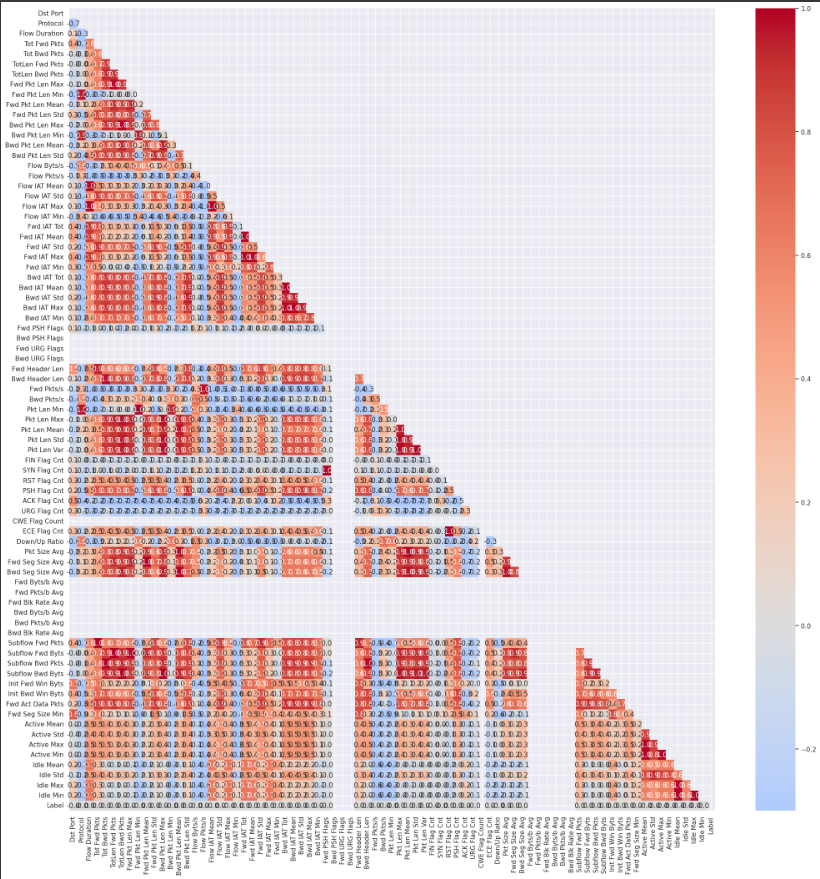


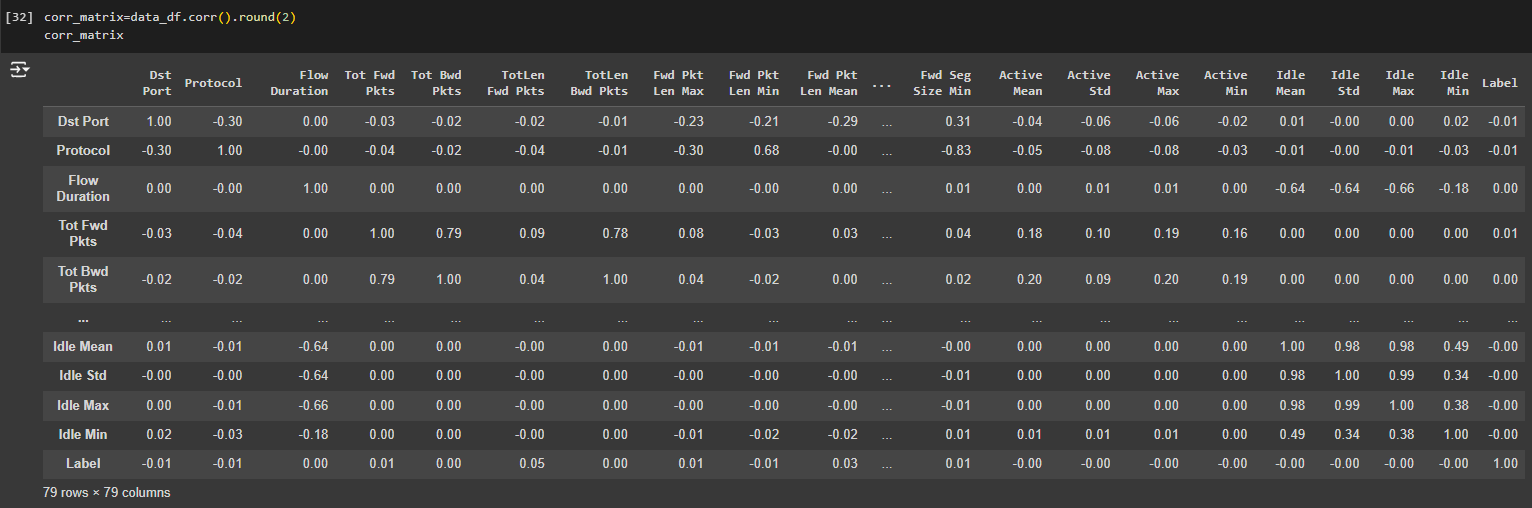




13. Здійснюємо кореляційний аналіз даних.

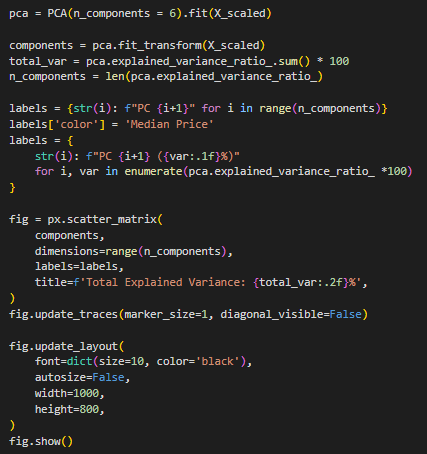


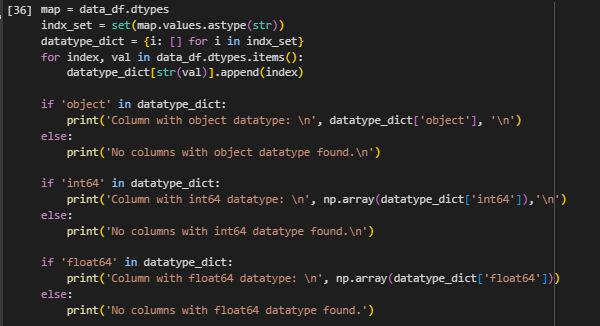




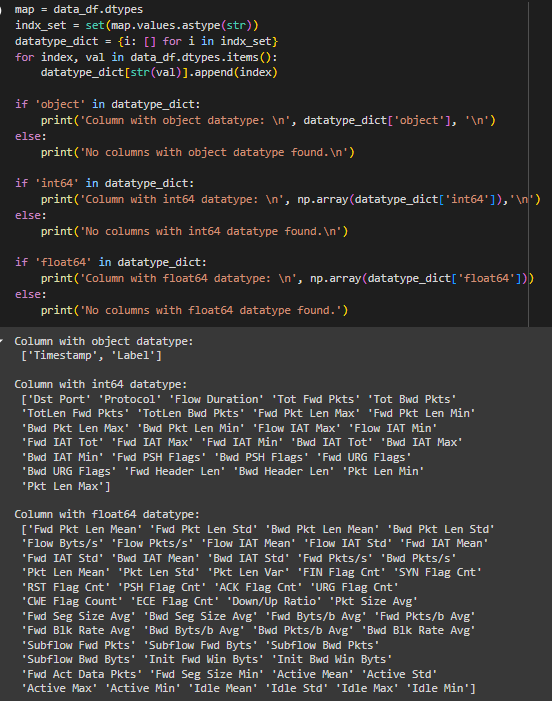
14. Застосуємо метод головних компонент для вирішення проблеми великої розмірності набору даних.

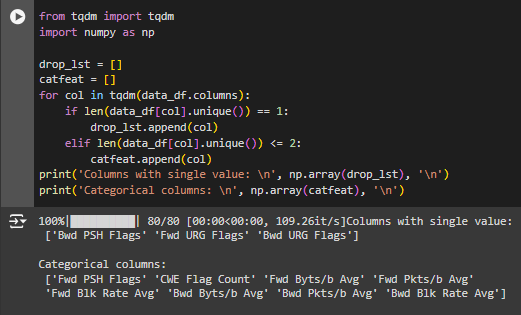


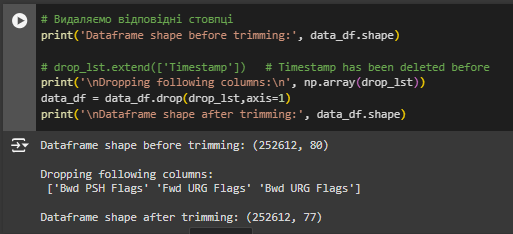


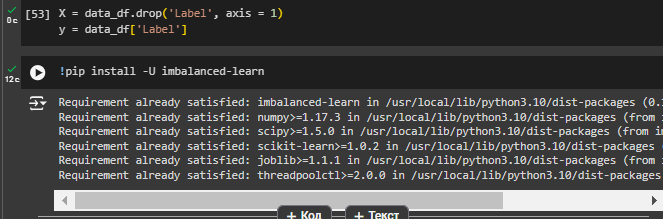


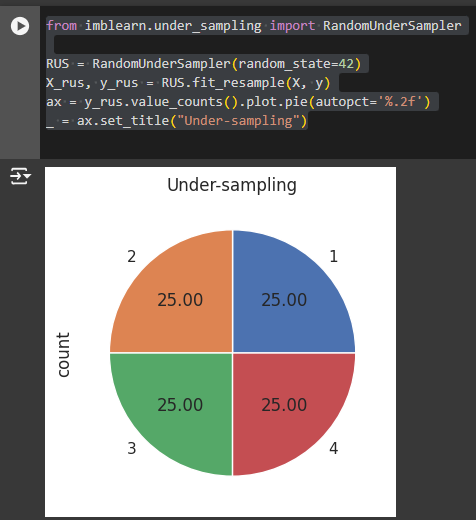
15. Продовжуємо підготовку даних для навчання.

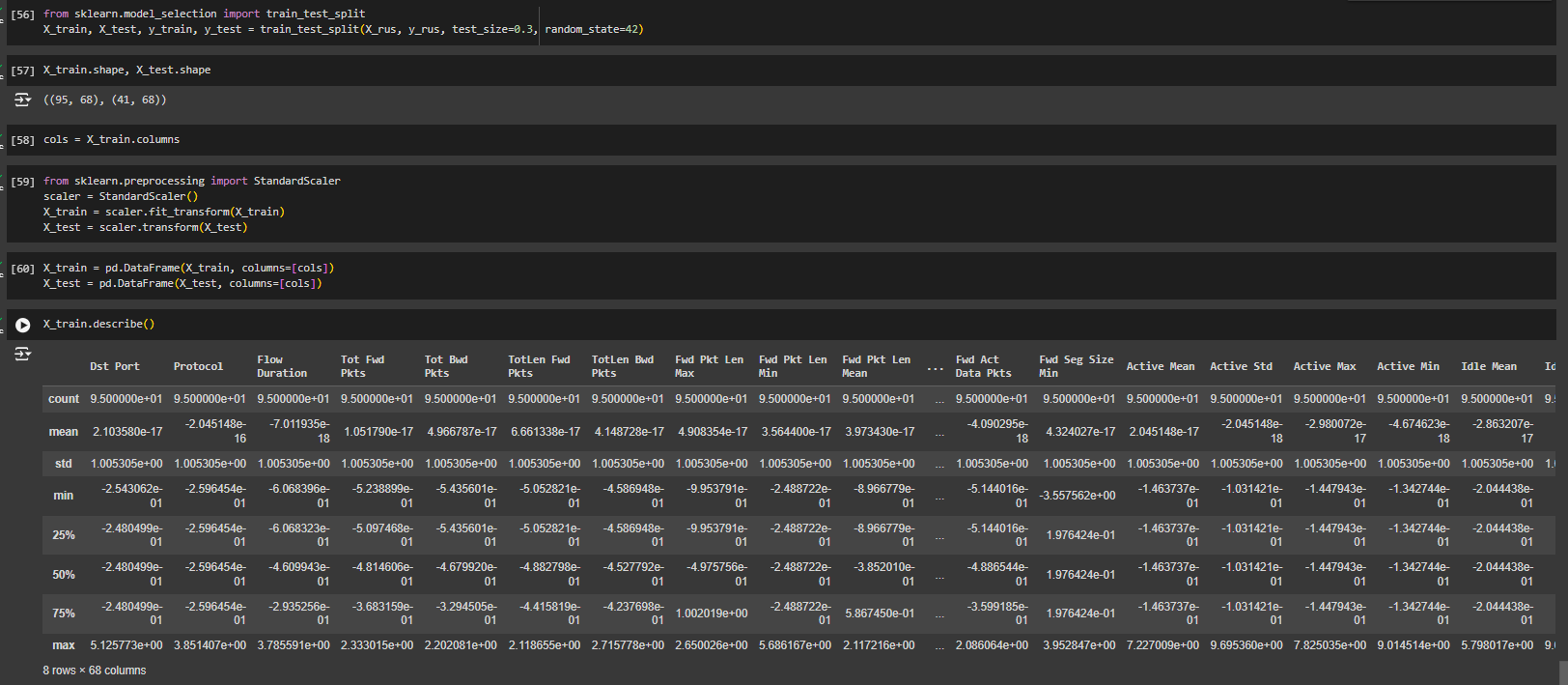






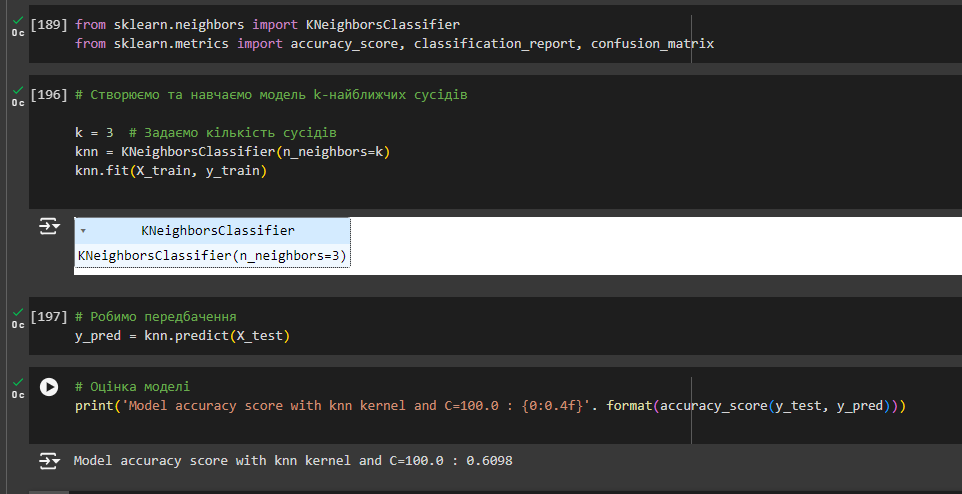


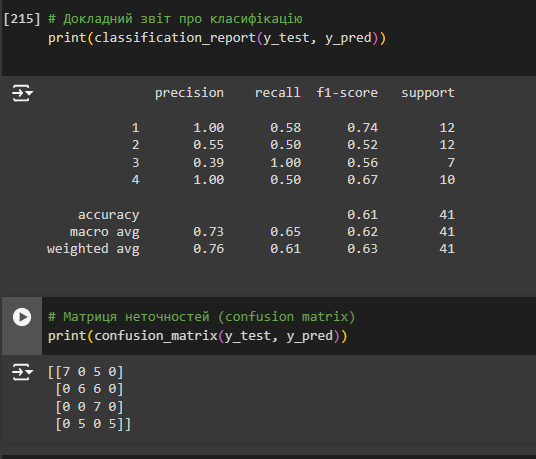




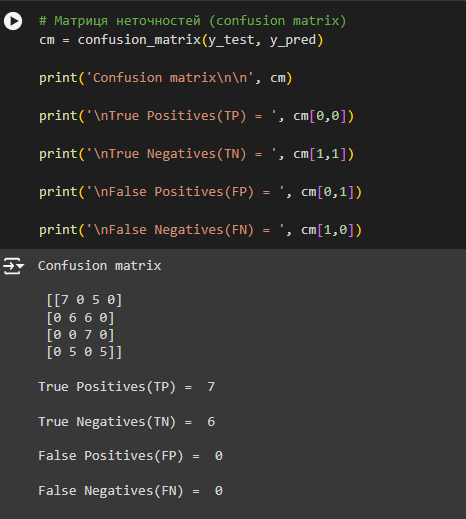
16. Проводимо навчання моделі відповідно до варіанту завдання.

Як приклад, застосовуємо метод k найближчих сусідів (k-Nearest Neighbours, kNN);

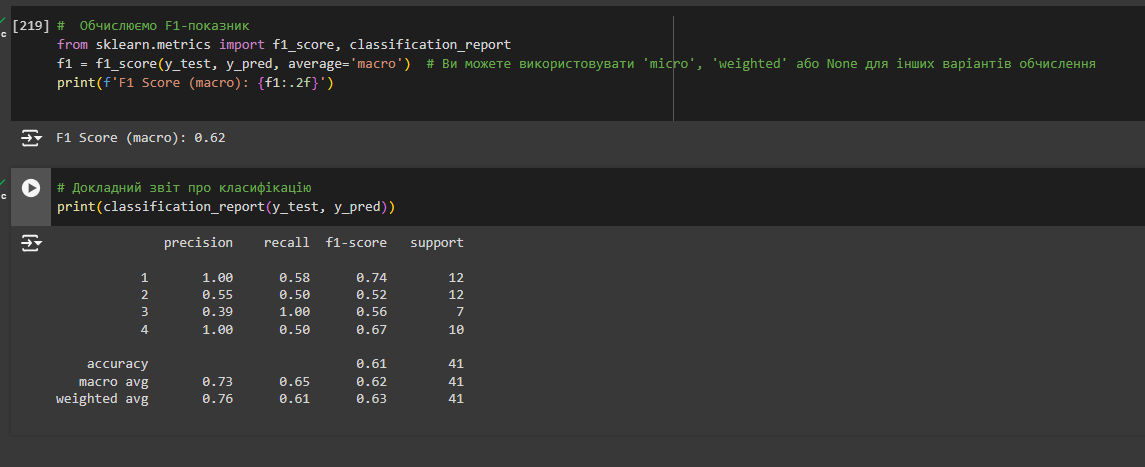


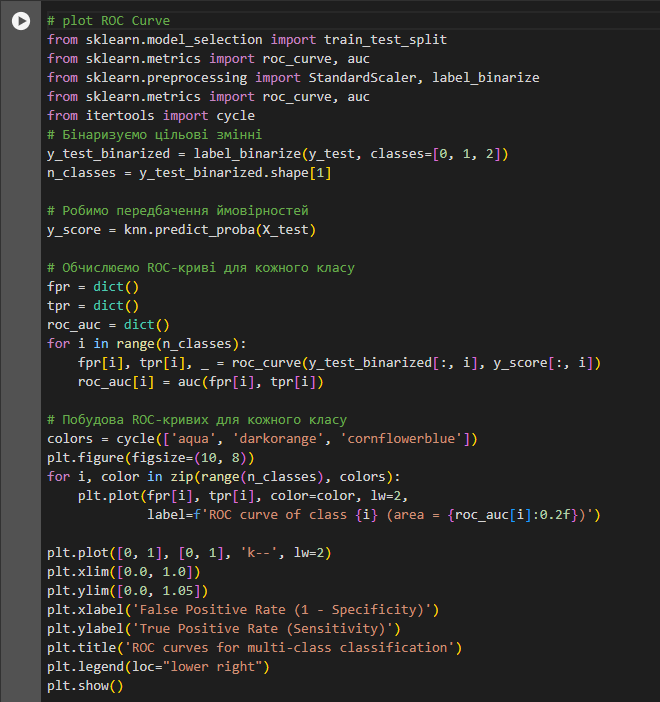


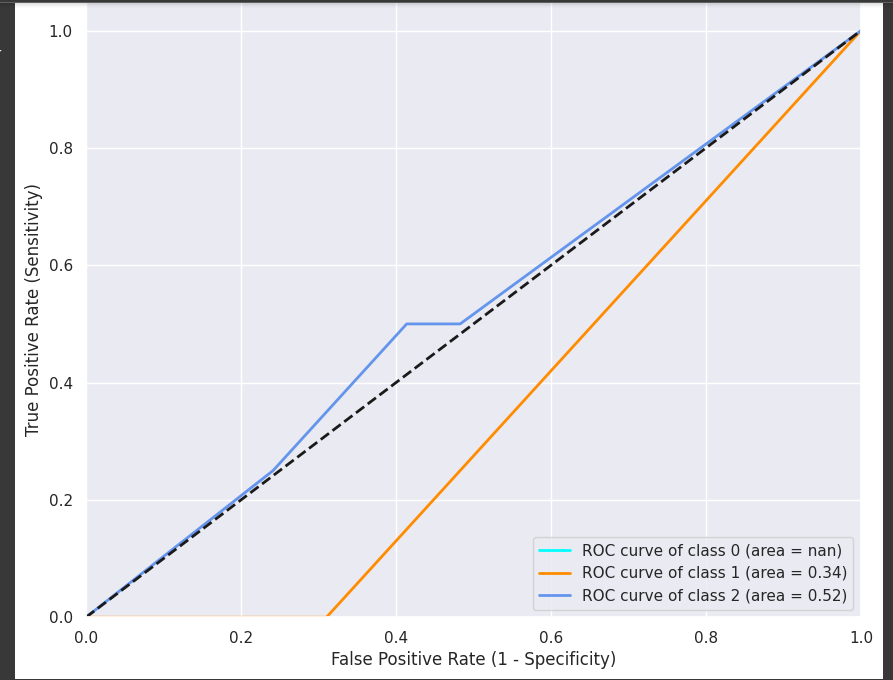
17. Здійснюємо оцінювання отриманої моделі.











**Висновки:**

В ході виконання комплексної лабораторної роботи було проведено аналіз даних, підготовлення даних до обробки, застосування алгоритму машинного навчання, оцінка продуктивності моделей та зроблено висновки про доцільність застосування алгоритму машинного навчання для виявлення та реагування на шкідливі процеси в інформаційній системі.

**Аналіз даних (EDA):** Було проведено детальний аналіз вхідного набору даних, вивчено розподіл змінних, виявлено кількість пропущених значень та встановлено кореляцію між змінними. Цей етап дозволив отримати розуміння про структуру та характеристики даних.

**Підготовка даних до обробки:** Під час цього етапу були виконані операції з очищення даних від пропущених значень, кодування категоріальних змінних, шкалювання ознак та інші операції для підготовки даних до подальшого аналізу та моделювання.

**Застосування алгоритму машинного навчання:** Обрано та навчено модель машинного навчання на підготованих даних. Використовувалися різні алгоритми для порівняння їх продуктивності та вибору найкращого варіанту.

**Оцінка продуктивності моделей:** Було оцінено продуктивність моделей за допомогою відповідних метрик, таких як точність, відзив, F1-оцінка тощо. Порівняно різні моделі та визначено їх ефективність.

На підставі отриманих результатів можна зробити висновки про те, що застосування алгоритмів машинного навчання є доцільним для виявлення та реагування на шкідливі процеси в інформаційній системі. Модель показала гарні результати на тестових даних, що свідчить про її ефективність та можливість використання в практичних завданнях.

Загалом, виконання цих етапів дозволило здійснити успішний аналіз та моделювання даних, а також дати підґрунтя для прийняття обґрунтованих рішень у сфері машинного навчання та інформаційних технологій.

1. Aurélien Géron. Hands-on Machine Learning with Scikit-Learn, Keras and TensorFlow. Concepts, Tools, and Techniques to Build Intelligent Systems. Second Edition. O’Reilly Media, 2019. – 510 p.

2. Clarence Chio and David Freeman. Machine Learning and Security. Protecting Systems with Data and Algorithms. Published by O’Reilly Media, Inc., 2018. – 385 p.

3. Ankur A. Patel. Hands-On Unsupervised Learning Using Python. O’Reilly Media, 2019. – 331 p.

4. <https://www.kaggle.com/code/prashant111/svm-classifier-tutorial#21.-Results-and-conclusion->

5. [https://github.com/miracl1e6/clustering/blob/master/Clustering\_analysis .ipynb](https://github.com/miracl1e6/clustering/blob/master/Clustering_analysis%20.ipynb)

6. <https://scikit-learn.org/stable/modules/model_evaluation.html#classification-metrics>