BÁO CÁO THỰC HÀNH

**Môn học: Quản trị mạng và hệ thống**

**Lab 1: Setting Up Your Machine Learning for Cybersecurity Arsenal**

*GVHD: Nguyễn Hữu Quyền*

1. **THÔNG TIN CHUNG:**

*(Liệt kê tất cả các thành viên trong nhóm)*

Lớp: NT522.O21.ATCL.2- Nhóm 3

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Họ và tên** | **MSSV** | **Email** |
| 1 | Nguyễn Ngọc Trà My | 21520353 | [21520353@gm.uit.edu.vn](mailto:21520353@gm.uit.edu.vn) |
| 2 | Bùi Hoàng Trúc Anh | 21521817 | [21521817@gm.uit.edu.vn](mailto:21521817@gm.uit.edu.vn) |
| 3 | Lê Hoàng Oanh | 21521253 | [21521253@gm.uit.edu.vn](mailto:21521253@gm.uit.edu.vn) |
| 4 | Huỳnh Minh Tân Tiến | 21521520 | 21521520@gm.uit.edu.vn |

1. **NỘI DUNG THỰC HIỆN:**

|  |  |  |
| --- | --- | --- |
| **STT** | **Công việc** | **Kết quả tự đánh giá** |
| 1 | **Yêu cầu 1** | 100% |
| 2 | **Yêu cầu 2** | 100% |
| 3 | **Yêu cầu 3** | 100% |

**Phần bên dưới của báo cáo này là tài liệu báo cáo chi tiết của nhóm thực hiện.**

BÁO CÁO CHI TIẾT

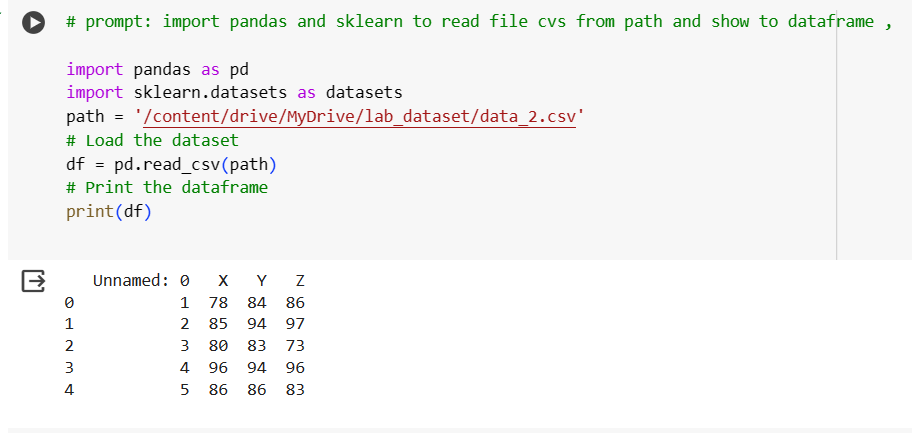
1. Sinh viên cho ví vụ về phép cộng, trừ hai ma trận numpy.

A screenshot of a computer program

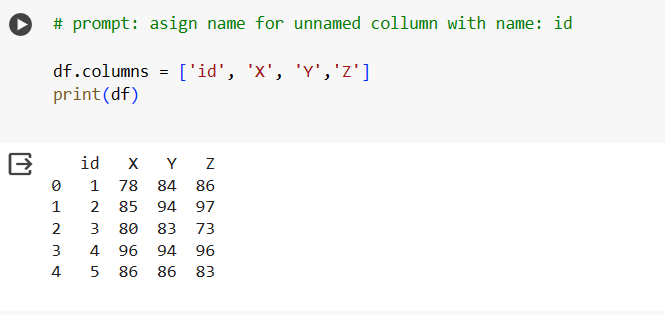
Description automatically generated

1. Sinh viên sử dụng pandas xử lý các yêu cầu sau:

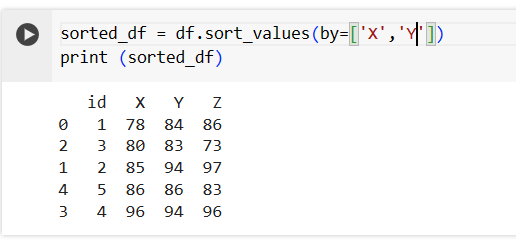
- Đọc CSV thành Dataframe và hiển thị



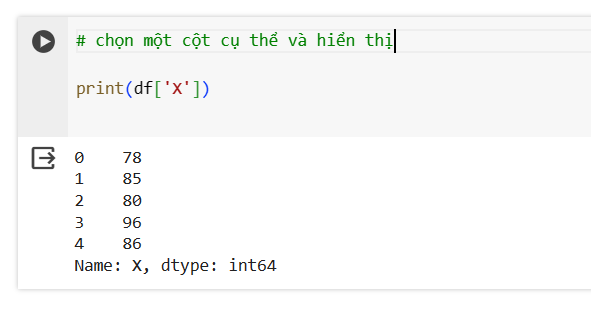
- Hãy chuyển index mặc định thành giá trị cột id



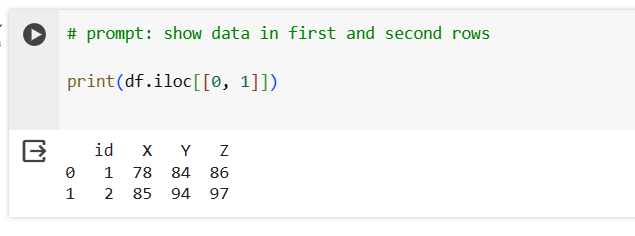
- Sắp xếp dữ liễu theo nhiều cột (sort)



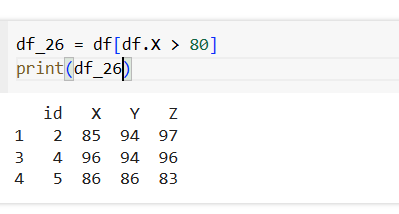
- Chọn một cột cụ thể và hiển thị nó



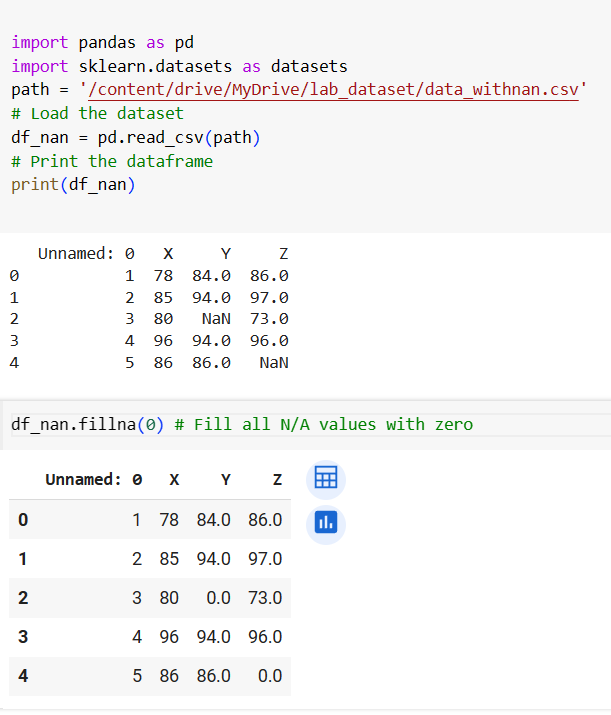
- Chọn 2 hàng đầu tiền và hiển thị chúng



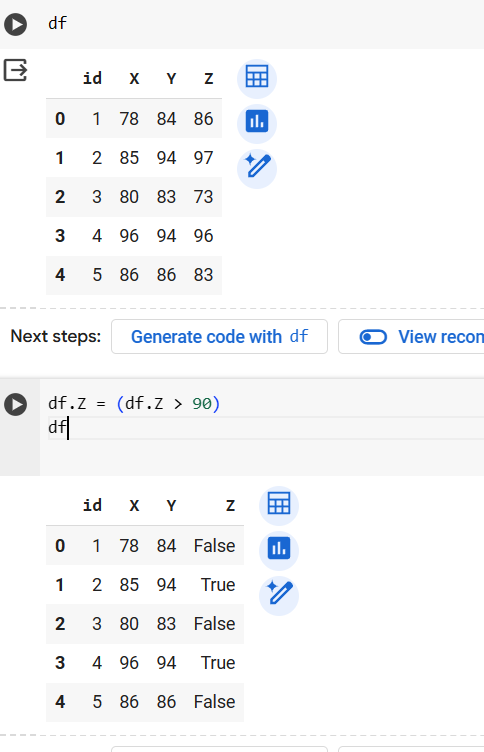
- Hãy chọn một hàng dựa trên một điều kiện giá trị của cột



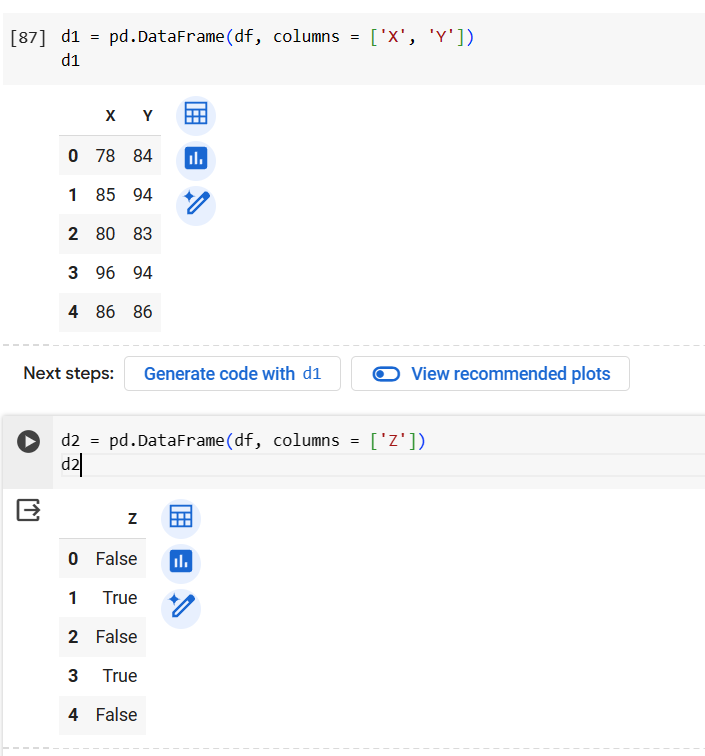
- Thay đổi một vài giá trị thành NaN ở CSV, sau đó đọc lên thành Dataframe và thay thế chúng bằng giá trị 0

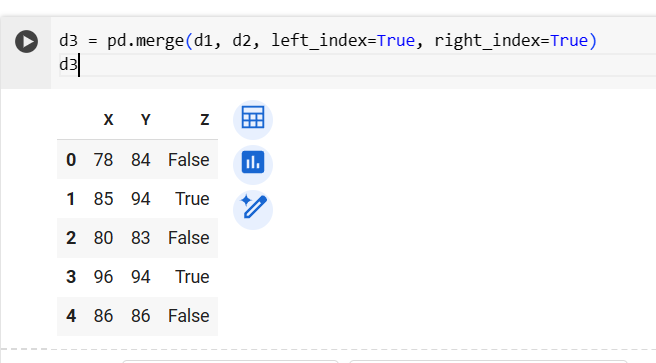


- Ở cột Z chuyển giá trị lớn hơn 90 là True và nhỏ hơn là False trong Dataframe

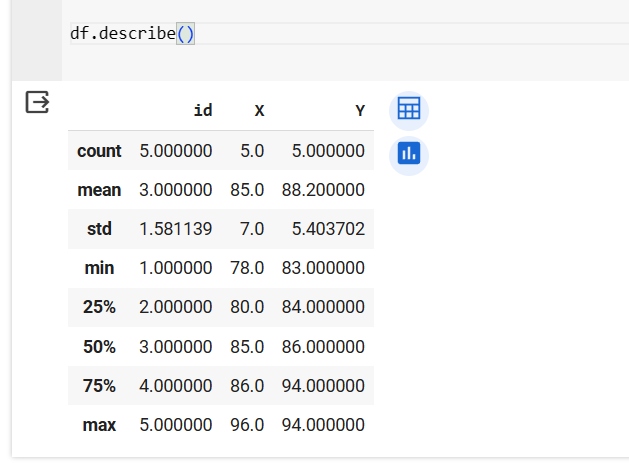


- Chuyển Dataframe trên thành 2 Dataframe d1 và d2; d1 chứ cột X và Y, d2 chứa cột Z; cuối cùng d3 là thành quả của nối 2 Dataframe d1 và d2





- Dùng tính năng thống kê hãy hiển thị kết quả thông kể các giá trị thuộc tính của Dataframe



1. Sinh viên tự tìm hiểu thực hiện lại ví dụ dùng mô hình Linear Regression trong thư viện scikit-learning bằng các thư viện sáu:

- TensorFlow

- Keras

- PyTorch

Cho biết cảm nghĩ về việc dùng 4 thư viện này

- TensorFlow

Ảnh có chứa văn bản, ảnh chụp màn hình, phần mềm, Trang web

Description automatically generated

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ

Description automatically generated

Ảnh có chứa văn bản, ảnh chụp màn hình

Description automatically generated

- Keras

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ

Description automatically generated

Ảnh có chứa văn bản, ảnh chụp màn hình, hàng, Sơ đồ

Description automatically generated

- PyTorch

Ảnh có chứa văn bản, ảnh chụp màn hình

Description automatically generated

Ảnh có chứa văn bản, ảnh chụp màn hình, tài liệu

Description automatically generated

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, thiết kế

Description automatically generated

Ảnh có chứa văn bản, ảnh chụp màn hình, hàng, Song song

Description automatically generated

* **scikit-learn:** Dễ sử dụng, thích hợp cho các công việc machine learning cơ bản và mô hình có cấu trúc đơn giản như Linear Regression.
* **TensorFlow và Keras:** Mạnh mẽ, linh hoạt và hỗ trợ mạnh mẽ từ cộng đồng. Đặc biệt, Keras cung cấp một giao diện cao cấp hơn cho việc xây dựng mô hình so với TensorFlow.
* **PyTorch:** Dễ dàng để xây dựng và debug mô hình nhờ vào cú pháp gần gũi với Python và cung cấp tính linh hoạt cao trong quá trình huấn luyện.

1. Sinh viên hoàn thành code phát hiện spam với SVMs và Linear regression

A screenshot of a computer

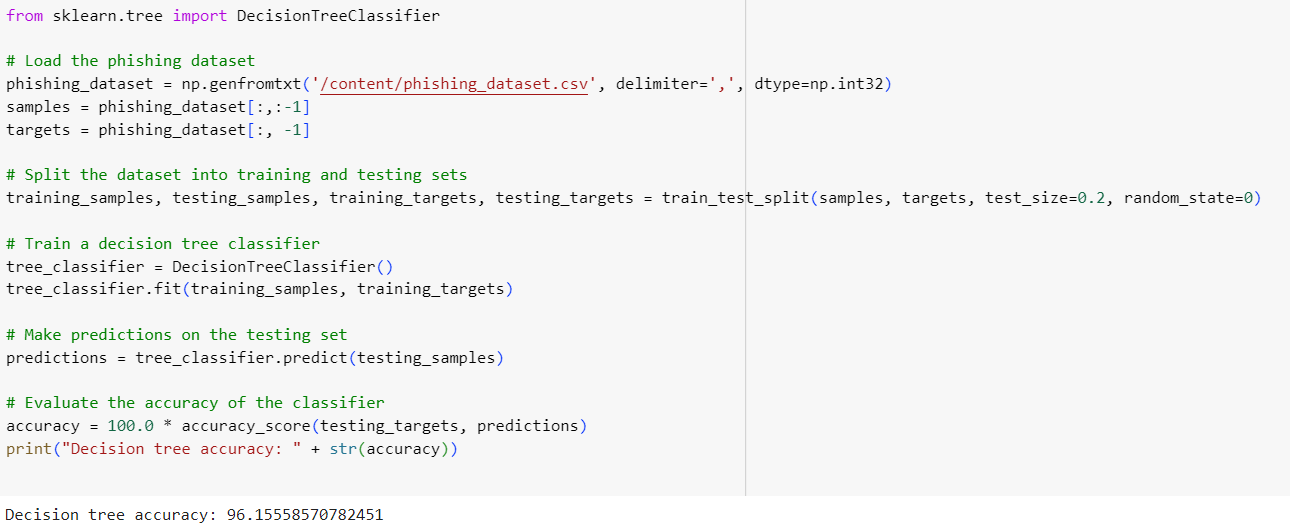
Description automatically generated

1. Sinh viên cho biết chức năng của phương thức genfromtxt() trong thư viện numpy.

genfromtxt() là một hàm trong thư viện NumPy của Python, được sử dụng để tải dữ liệu từ tệp văn bản (như CSV) vào một mảng NumPy. Nó cung cấp một cách linh hoạt để đọc dữ liệu từ các tệp có cấu trúc khác nhau và tự động xử lý các giá trị bị thiếu.

1. Sinh viên hoàn thiện code Decision trees trên và đánh giá kết quả nhận được so với phương pháp Logistic regression.

Decision trees



Logistic regression



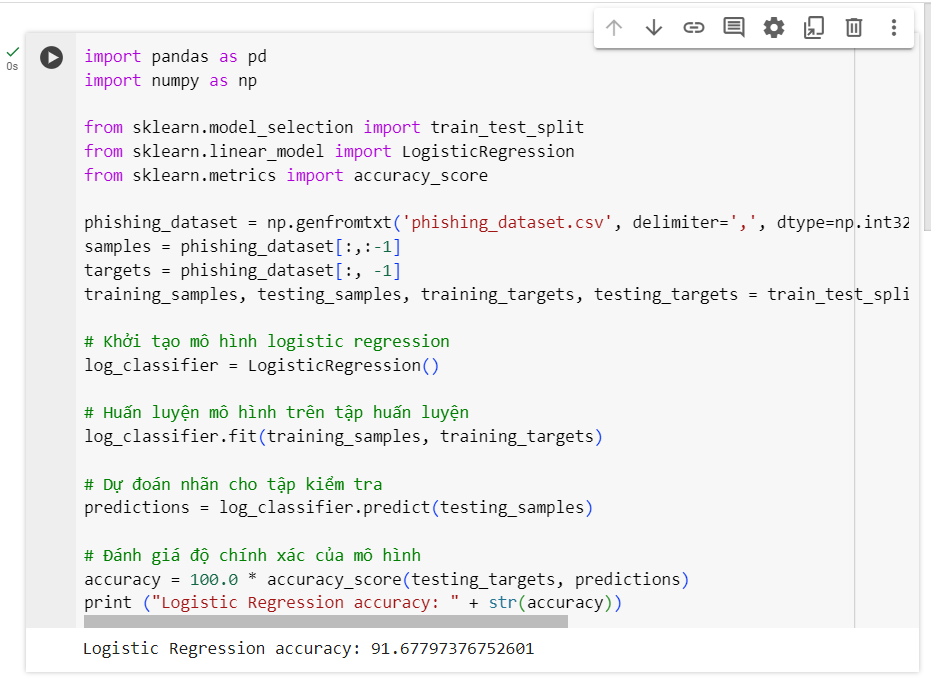
Kết quả của code Decision Trees cho ra accuracy cao hơn code Logistic Regression.

1. Sinh viên thực hiện code phát hiện phising website bằng mô hình học máy

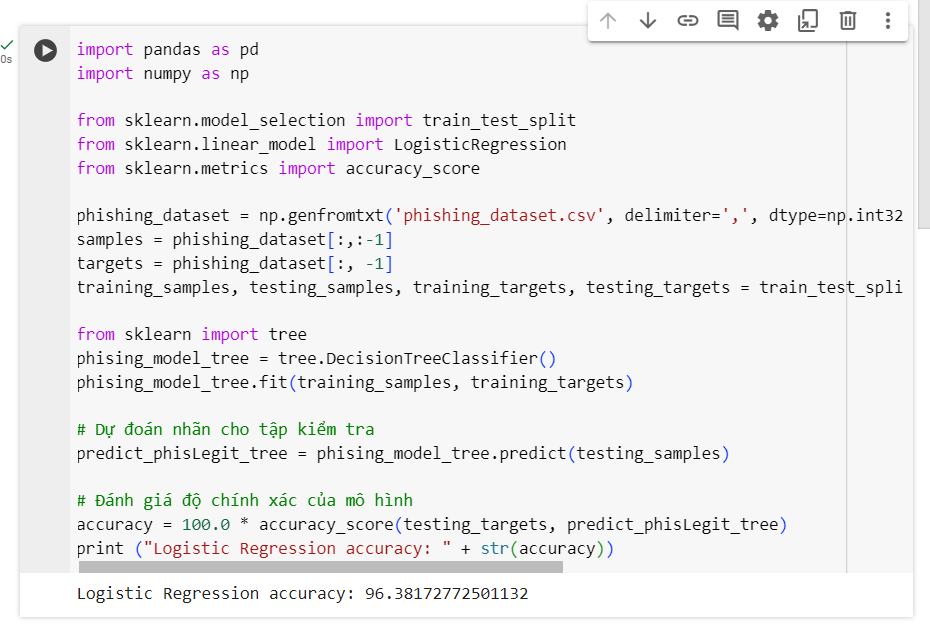
learning Logistic regression và Decision trees với train và test trên tập dữ liệu

<https://www.kaggle.com/shashwatwork/phishing-dataset-for-machine>

* Logistic Regression



* Decision Trees



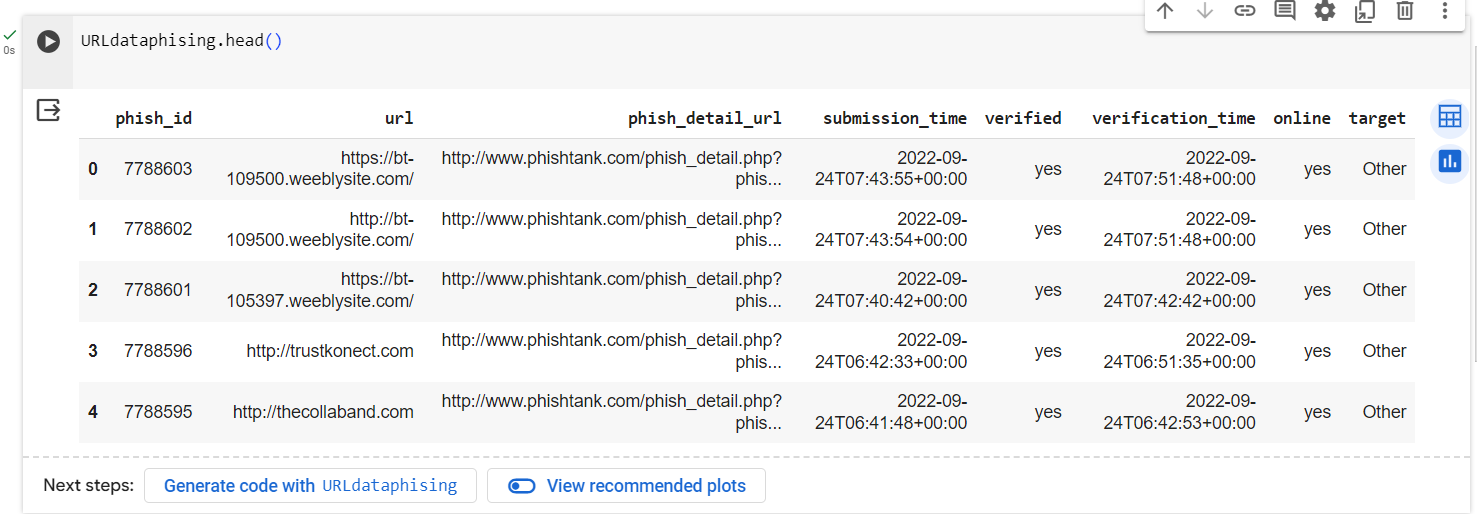
1. Sinh viên thực hiện code phát hiện phising website bằng mô hình học máy

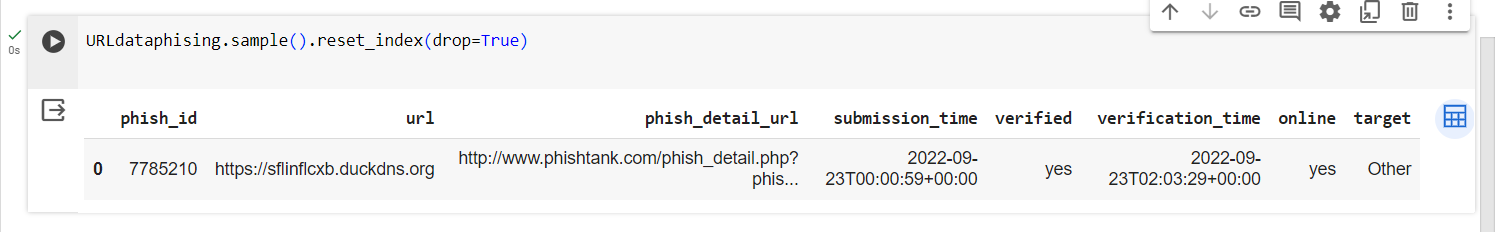
Logistic regression hoặc Decision trees với train và test trên tập dữ liệu phishtank. Tham khảo cách xử lý và trích xuất thuộc tích

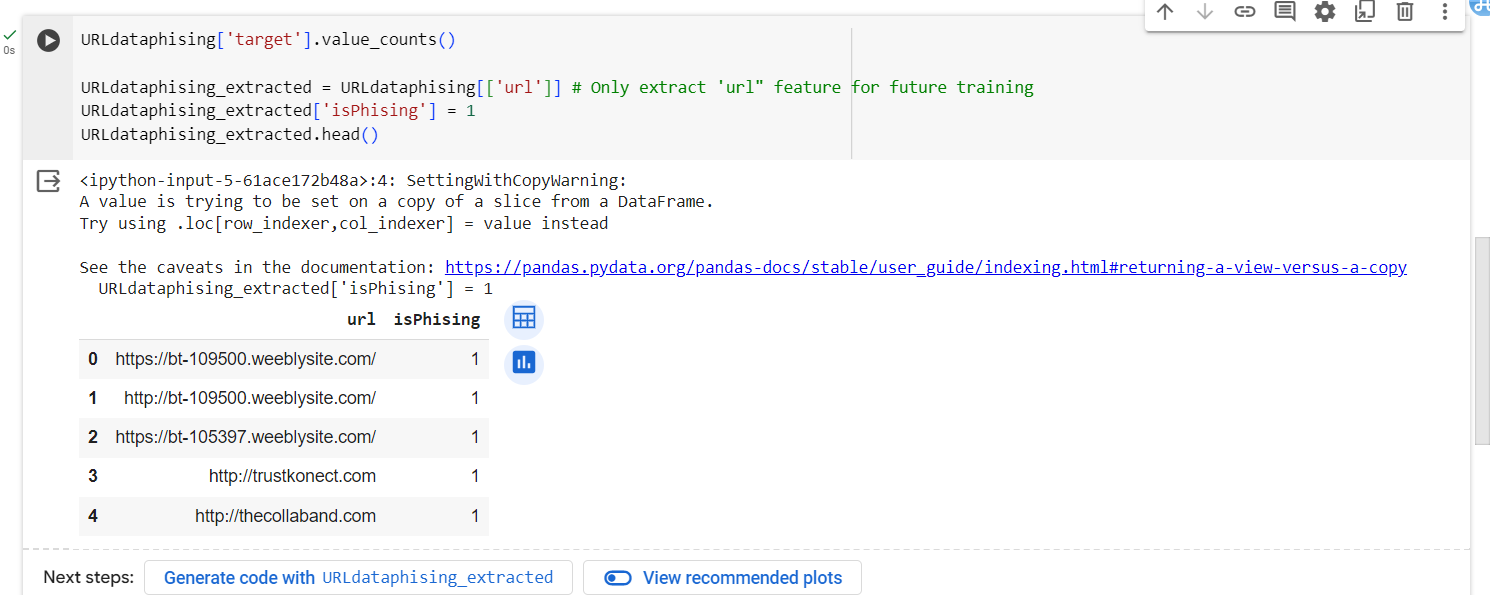
<https://github.com/surajr/URL-Classification>

Phising Data

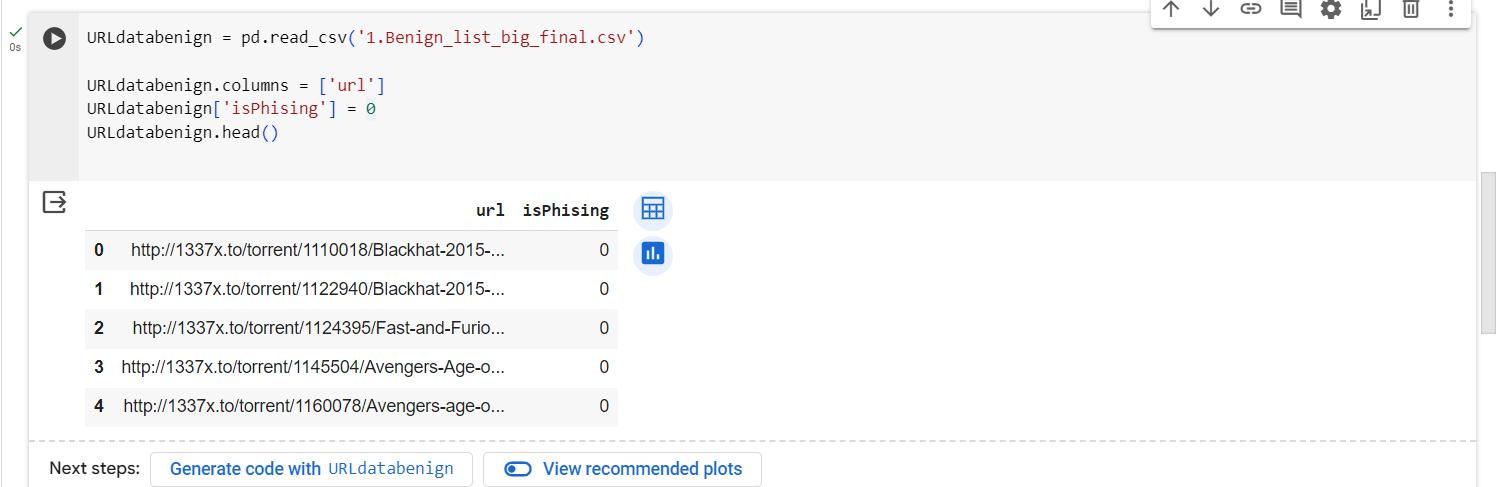




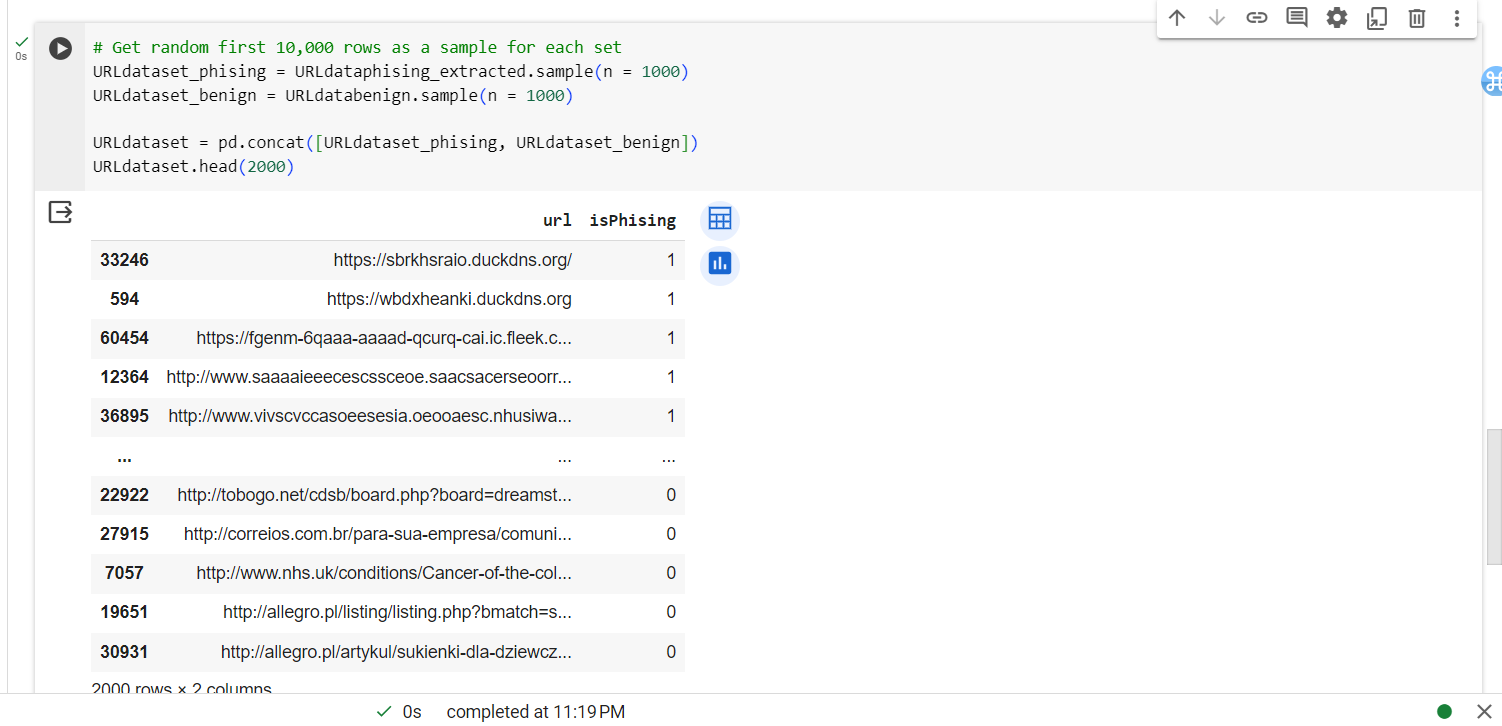




Benign Data

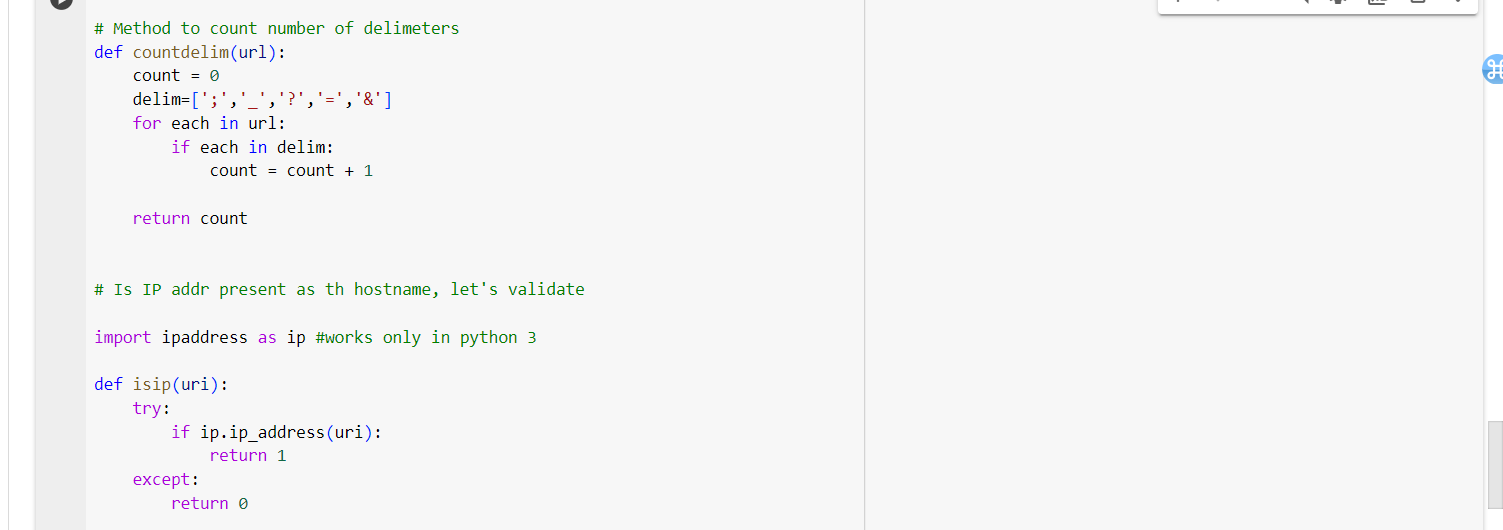


Create full dataset with phising and benign sample

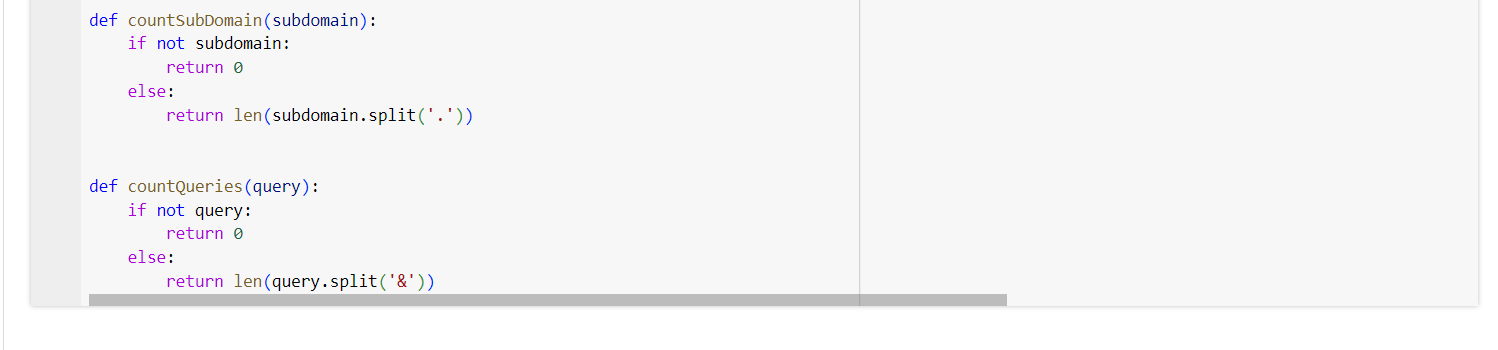


Define functions

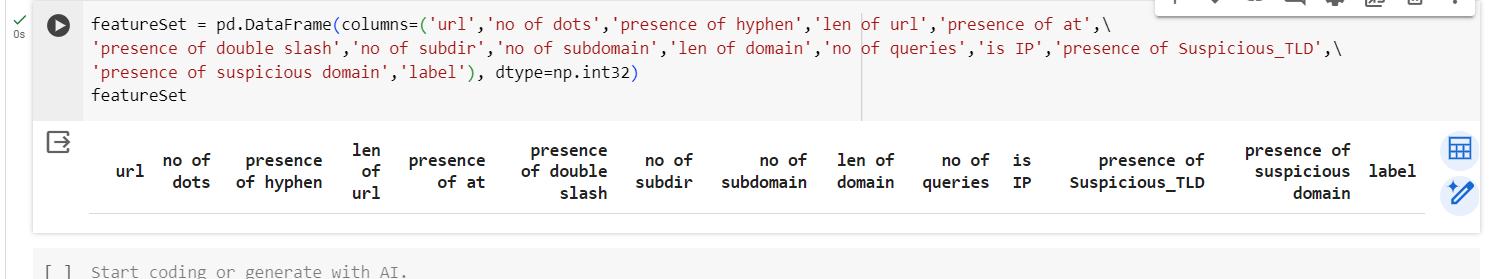


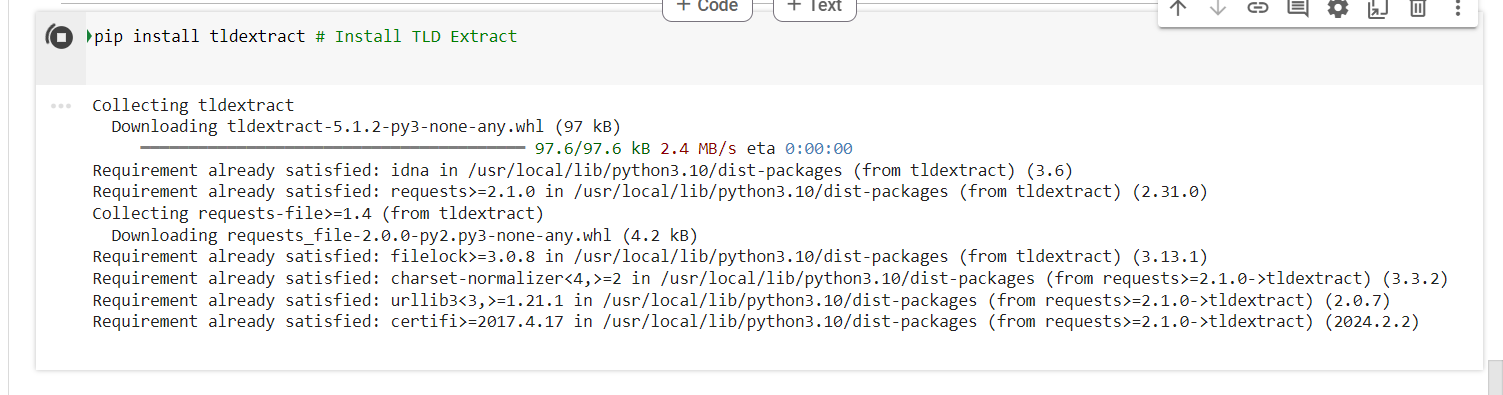




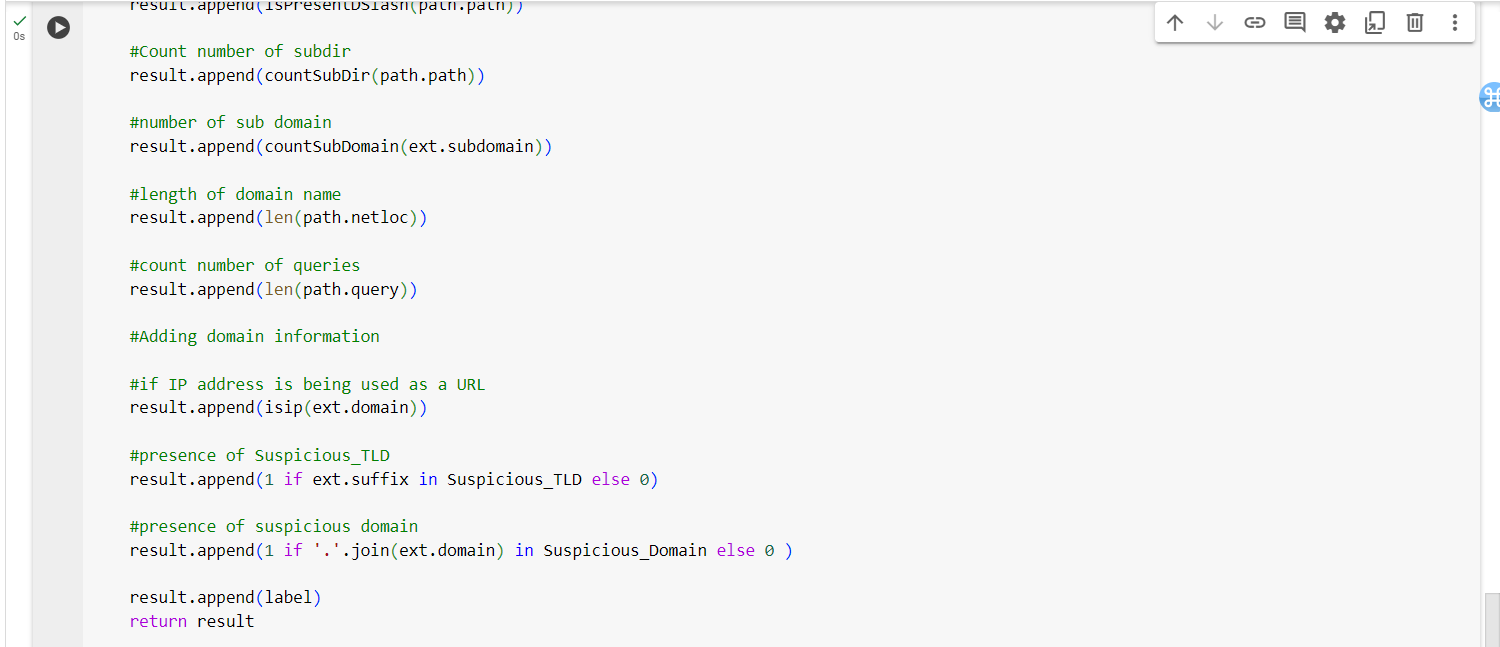


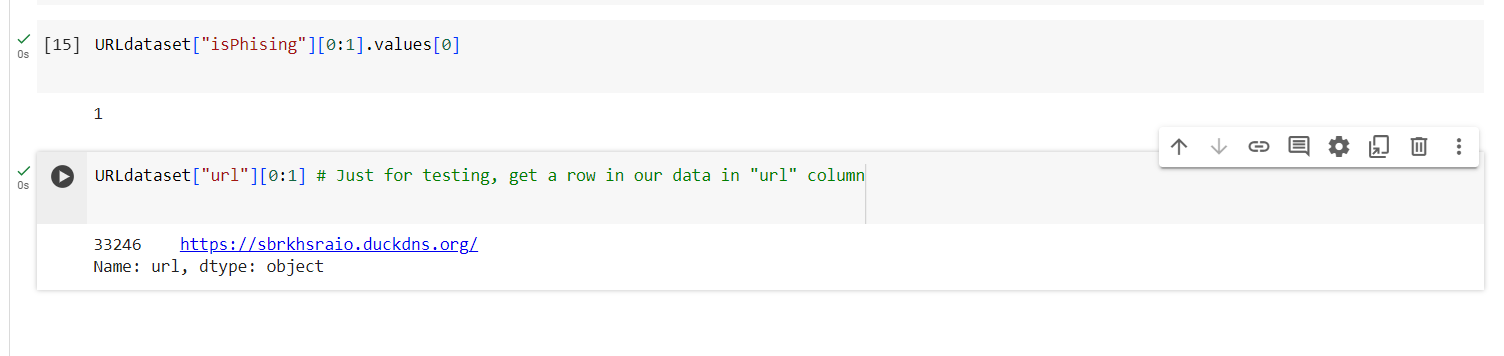
Create new dataset with number values

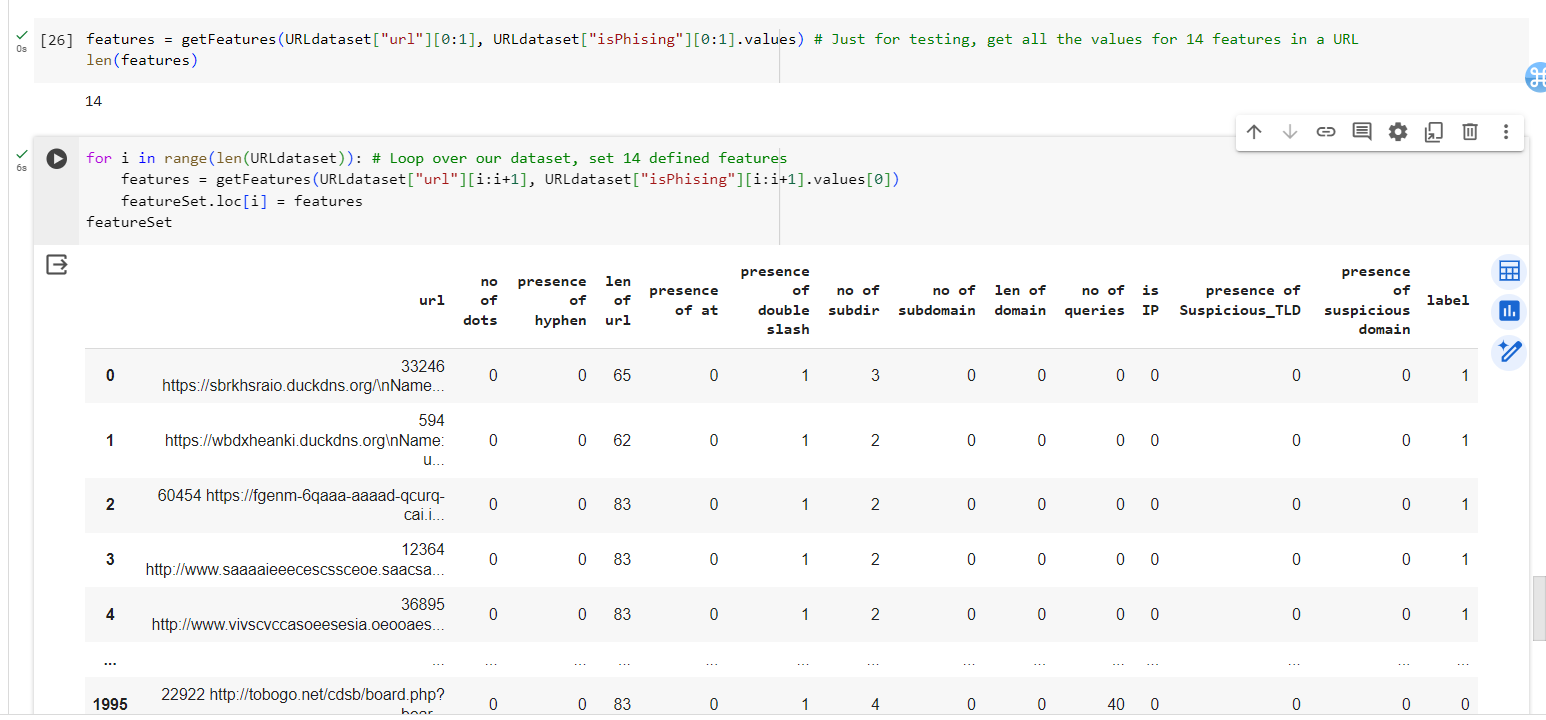












Using Decision Tree

