

## Clase 2: Ingreso y manejo de Kaggle

A continuación, se adjuntas capturas de imagen del Dataset escogido para el trabajo en clase:

The screenshot shows the Kaggle homepage with the sidebar open. The sidebar includes links for Create, Home, Competitions, Datasets (which is selected), Models, Code, Discussions, Learn, More, User Rankings, and View Active Events. The main content area is titled 'Datasets' and shows search results for 'cybersecurity'. The search bar contains 'cybersecurity'. Below it are filters: 'All datasets X' (selected), Computer Science, Education, Classification, Computer Vision, NLP, and Data Visualization. A 'Pre-Trained Model' button is also present. The results section shows '582 Datasets'. The first result is 'Cybersecurity: Suspicious Web Threat Interactions' by JanCSG, updated a year ago, with Usability 10.0, 1 File (CSV), and 4 kB. It has 64 votes and is a Bronze dataset. The second result is 'Cybersecurity Incidents Dataset'.

**kaggle**

- + Create
- Home
- Competitions
- Datasets
- Models
- <> Code
- Discussions
- Learn
- More
- Your Work
- View Active Events

HUZPSB · UPDATED 2 MONTHS AGO

# Cybersecurity Incidents Dataset

How many reports about cybersecurity incidents are filed to local authority



Data Card    Code (4)    Discussion (1)    Suggestions (0)

## About Dataset

The number of cybersecurity incident reports filed with local authorities and the estimated loss.

Loss is calculated in USD.

Some missing fields are imputed.

**Usability** 10.00

**License** CC0: Public Domain

**Expected update frequency** Annually

notebookbdcc53dec1 Draft saved

File Edit View Run Settings Add-ons Help

Share Save Version 0

[1]:

```
# This Python 3 environment comes with many helpful analytics libraries installed
# It is defined by the kaggle/python Docker image: https://github.com/kaggle/docker-python
# For example, here's several helpful packages to load

import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)

# Input data files are available in the read-only "../input/" directory
# For example, running this (by clicking run or pressing Shift+Enter) will list all files under the input directory

import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))

# You can write up to 20GB to the current directory (/kaggle/working/) that gets preserved as output when you create a version using "Save"
# You can also write temporary files to /kaggle/temp/, but they won't be saved outside of the current session
```

/kaggle/input/cybersecurity-incidents-dataset/LossFromNetCrime.csv

Nombre: María Fernanda Pacheco

```
# You can write up to 20GB to the current directory (/kaggle/working/) that gets preserved as output when you create a version using "Save Version".  
# You can also write temporary files to /kaggle/temp/, but they won't be saved outside of the current session
```

```
/kaggle/input/cybersecurity-incidents-dataset/LossFromNetCrime.csv
```

```
[1]: #Cargar el dataset (supongamos que esta en formato .csv)  
df_pandas= pd.read_csv("/kaggle/input/cybersecurity-incidents-dataset/LossFromNetCrime.csv")
```

+ Code + Markdown

```
[1]: #Cargar el dataset (supongamos que esta en formato .csv)  
df_pandas= pd.read_csv("/kaggle/input/cybersecurity-incidents-dataset/LossFromNetCrime.csv")
```

```
[2]: #mostrar el tamaño del dataset  
print(df_pandas.shape)  
print(f"Tamaño del dataset: {df_pandas.shape[0]} filas y {df_pandas.shape[1]} columnas")
```

+ Code + Markdown

## Paso 4: Discusión y Reflexión en grupos

Después de evaluar los conjuntos de datos, los estudiantes discuten:

- ¿Qué dataset es más útil para proyectos de ciberseguridad?
- ¿Cuáles son los desafíos al trabajar con estos datos?
- ¿Cómo podríamos mejorar la calidad de estos datasets?

### 0. Formato del Excel a subirse

TEMA	PREGUNTA	RESPUESTA	JUSTIFICACIÓN
Tamaño del dataset	¿Cuántas filas y columnas tiene?	Tamaño del dataset: 117 filas y 13 columnas	Ejecutando código: <pre>print(f"Tamaño del dataset: {df_pandas.shape[0]} filas y {df_pandas.shape[1]} columnas")</pre>
	¿Es suficiente para el análisis?	Sí.	Por tratarse de un Dataset pequeño, se recomienda analizar usando pandas; aunque también se podría usar polars, o dask, pero estos son para datasets de mayor tamaño que el elegido.
Formato de los datos	¿Está en CSV, JSON, SQL, u otro formato?	Está en formato .CSV	Al ejecutar el código para conocer el nombre del directorio se obtiene lo siguiente: /kaggle/input/cybersecurity-incidents-dataset/LossFromNetCrime.csv
	¿Es fácil de procesar?	Sí.	El dataset es relativamente pequeño y por lo tanto más fácil de procesar porque no requiere recursos computacionales extensos, lo cual facilita tareas como la limpieza, exploración y análisis de datos.
Etiquetado	¿Los datos tienen etiquetas claras?	Sí.	Los datos poseen columnas etiquetadas por país, año del dato desde 2019 hasta 2024, seguido de si se trata de una queja o una pérdida. Al ejecutar el siguiente código:  <pre># Mostrar las primeras filas print(df_pandas.head()) # Verificar nombres de columnas print(df_pandas.columns)</pre>

			<p>Se obtienen los datos de las etiquetas de cada columna:</p> <pre>Index(['Country', '2019_Complaints', '2019_Losses',        '2020_Complaints',        '2020_Losses', '2021_Complaints', '2021_Losses',        '2022_Complaints',        '2022_Losses', '2023_Complaints', '2023_Losses',        '2024_Complaints',        '2024_Losses'],       dtype='object')</pre>
	<p>¿Hay valores faltantes o datos inconsistentes?</p> <p>Sí, existe 1 null en la columna “Country”.</p>		<p>Al ejecutar el siguiente código:</p> <pre># Identificar valores faltantes print(df_pandas.isnull().sum())</pre> <p>Se obtiene:</p> <pre>Country      1 2019_Complaints    0 2019_Losses      0 2020_Complaints    0 2020_Losses      0 2021_Complaints    0 2021_Losses      0 2022_Complaints    0 2022_Losses      0 2023_Complaints    0 2023_Losses      0 2024_Complaints    0 2024_Losses      0 dtype: int64</pre>
Fuente y credibilidad	¿Quién creó el dataset?	Alvin Chou, Collaborator: huzpsb (Owner)	En Data Card del dataset se encontró esta información, al final de la página existe una sección “Metadata” en la cual se encuentran los nombres del autor, sus colaboradores, entre otra información.
	¿Está bien documentado?	Sí, lo suficiente al ser un dataset pequeño	Si bien no existe una gran sección explicativa, al menos existe en el encabezado algunos aspectos a considerar sobre los datos mostrados:

Nombre: María Fernanda Pacheco

		<p>About Dataset:</p> <p>Contiene el número de informes de incidentes de ciberseguridad presentados ante las autoridades locales y pérdida estimada.</p> <p>La pérdida se calcula en dólares estadounidenses.</p> <p>Se han imputado algunos campos faltantes.</p> <p>En general, coincide con el informe IC3.</p> <p><a href="#">El IC3 Report (Internet Crime Complaint Center) es el informe que las víctimas de delitos ciberneticos pueden presentar a la FBI a través del Internet Crime Complaint Center para alertar sobre un crimen en línea.</a></p>
--	--	--