Instituto Tecnológico de Hermosillo

Minería de datos

Proyecto Final: Aumento de incidentes cibernéticos

Maestro: Eduardo Antonio Hinojosa Palafox

Alumna: Janeth Torres Cruz

Número de control: 19330670

Hermosillo, Sonora 13 de dic. de 2022

Introducción

Gracias a los conocimientos adquiridos dentro de Minería de Datos se trabaja con un conjunto de datos, se leen, se normalizan y estandarizan para poder trabajar con datos limpios y evitar trabajar con datos erróneos y nulos. Gracias a la Minería de datos nos permite extraer información de un conjunto de datos y transformarla en una estructura comprensible para el uso deseado.

Descripción del problema a desarrollar

El aumento de los incidentes cibernéticos ha ido aumento año con año, se debe de identificar qué tipo de ataque ha tenido un mayor aumento, a quien se está realizando el ataque, conocer las víctimas, el responsable, se debe de conocer las diferentes categorías y tipos de los ataques para llegar a una solución e ir disminuyendo el número de los ataques.

Descripción del conjunto de datos

Los datos han sido reportados por parte del Consejo de Relaciones Exteriores entre los años 2005 y 2020: “Growth of Cyber Incidents from 2005-2020”. El conjunto de datos cuenta con 481 filas y 12 columnas. Algunos datos son los siguientes: título del ataque, fecha, afiliaciones, descripción, víctimas, responsable, categoría y tipo de ataque.

Descripción de la solución propuesta.

Agrupar los datos para tener un mejor entendimiento y visualización de la importancia que se le debe de dar a la ciberseguridad y evitar el aumento continuo de los ciberataques. Mostrar los datos por medio de diferentes gráficas que permitan ver con mayor claridad los datos, haciendo un énfasis en la fecha, en el año, en el tipo y categoría de los ataques cibernéticos. Por último, predecir cual es la categoría que seguirá recibiendo ataques y de qué tipo de ataque cibernético en los próximos años.

Descripción del código utilizado

Se utilizaron las siguientes librerías de pandas, matplot y seaborn para graficar, y las diferentes funciones de sklearn para crear el modelo.

```
[ ]  import pandas as pd
     import matplotlib.pyplot as plt
     from sklearn.model_selection import train_test_split
     from sklearn.linear_model import LogisticRegression
     from sklearn.metrics import accuracy_score
     import seaborn as sns
     import seaborn as sn
     from sklearn.preprocessing import StandardScaler
```

Primero se carga el dataset extraído de Kaggle

```
df = pd.read_csv('/content/cyber-operations-incidents 2.csv')
df.head()
```

	Title	Date	Affiliations	Description	Response	Victims	Sponsor	Type	Category	
0	Attack on Austrian foreign ministry	2/13/2020	Turla	The suspected Russian hackers conducted a week...	Confirmation https://www.theregister.co.uk/2...	Austrian Foreign Ministry	Russian Federation	Espionage	Government	https://www.theregister.co.uk/202
1	Spear-phishing campaign against unnamed U.S. g...	1/23/2020	Konni Group	The suspected North Korean threat actor Konni ...	NaN	Employees of the U.S. government	Korea (Democratic People's Republic of)	Espionage	Government	https://unit42.paloaltonetworks.c
2	Australian Signals Directorate	4/6/2020	NaN	Responsible for attacking infrastructure that ...	NaN	NaN	Australia	Data destruction	Private sector	https://www.minister.defence.gov
3	Catfishing of Israeli soldiers	2/16/2020	APT-C-23	The Hamas-associated threat actor APT-C-23	Hack Back https://www.bleepingcomputer.com/n...	Israeli Defense Forces (IDF)	Palestine, State of	Espionage	Military	https://www.bleepingcomputer.com/

Se eliminan las columnas que no se utilizarán, dejando solo las columnas con las que se van a trabajar.

```
[ ] columns_to_drop = ['Title', 'Affiliations', 'Description', 'Response', 'Victims', 'Sponsor', 'Sources_1', 'Sources_2', 'Sources_3']

[ ] df.drop(columns_to_drop, axis='columns', inplace=True)
    df.head(10)
```

	Date	Type	Category
0	2/13/2020	Espionage	Government
1	1/23/2020	Espionage	Government
2	4/6/2020	Data destruction	Private sector
3	2/16/2020	Espionage	Military
4	8/10/2020	Espionage	Government, Private sector
5	3/29/2020	Espionage	Private sector
6	5/12/2020	Financial Theft	Private sector
7	8/31/2020	Espionage	Civil society, Private sector, Government
8	1/13/2020	Espionage	Private sector
9	1/28/2020	Espionage	Civil society

Se eliminan los datos nulos y se cuentan para verificar que no hay datos nulos en el nuevo dataframe, en este caso df2.

```
[ ] df2= df.dropna()

[ ] df2.shape

    (433, 3)

[ ] df2.isnull().sum()

    Date        0
    Type        0
    Category    0
    dtype: int64
```

Se hacen diferentes agrupaciones, en este caso el número de ataques agrupados por el tipo de ataque.

```
[ ] cyber_incidents = df2['Type'].value_counts()
    cyber_incidents.head(15)

    Espionage            361
    Sabotage              22
    Denial of service     18
    Data destruction      14
    Financial Theft        7
    Doxing                 6
    Defacement             5
    Name: Type, dtype: int64
```

Se observa como el ataque de tipo espionaje se obtuvieron 361 siendo el más realizado.

```
cyber_incidents2 = df['Category'].value_counts()
cyber_incidents2
```

```
Private sector                                          128
Government                                              104
Government, Private sector                               55
Civil society                                            52
Military                                                 21
Government, Military                                     17
Government, Civil society                                15
Private sector, Government                               14
Government, Private sector, Civil society                12
Military, Government                                      8
Civil society, Government                                 6
Civil society, Private sector                             6
Private sector, Military                                  4
Military, Private sector                                  3
Military, Civil society                                   2
Private sector, Civil society                             2
Civil society, Private sector, Government                 2
Government, Military, Private sector                      1
Private sector, Government, Military, Civil society       1
Government, Civil society, Private sector                 1
Military, Government, Civil society                       1
Private sector, Government, Civil society, Military       1
Private sector, Government, Civil society                 1
Civil society, Private sector, Military, Government       1
```

Se logra observar cómo el Sector privado es el que más recibió ataques con un número de 128 y seguido del gobierno con 104 ataques.

Se dividen los datos de entrenamiento y prueba, imprimiendo ‘x’ y ‘y’ de entrenamiento
Siendo x_test con los datos de fecha y categoría

```
X_Train,X_Test,Y_Train,Y_Test = train_test_split(X,Y,test_size = 0.2, random_state = 10)
```

```
X_Test
```

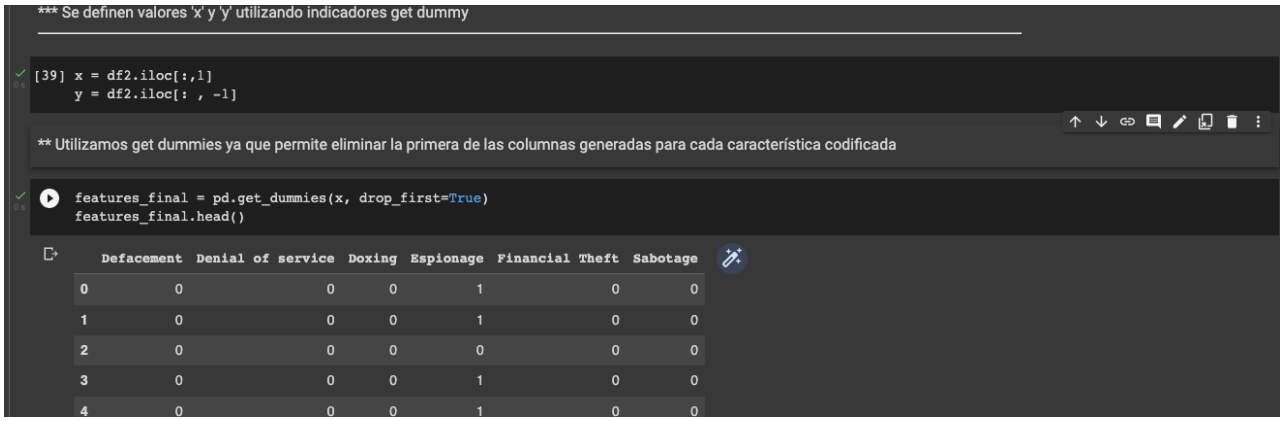
	Date	Category
92	4/4/2019	Private sector
410	5/20/2013	Private sector
94	10/21/2019	Military, Government, Civil society
117	9/5/2019	Civil society
298	12/13/2016	Private sector
...	...	...
449	3/20/2009	Government
133	8/7/2019	Government, Private sector
421	9/18/2012	Private sector

Y ‘Y_train’ con los tipos de ciberataques

```
Y_Train
```

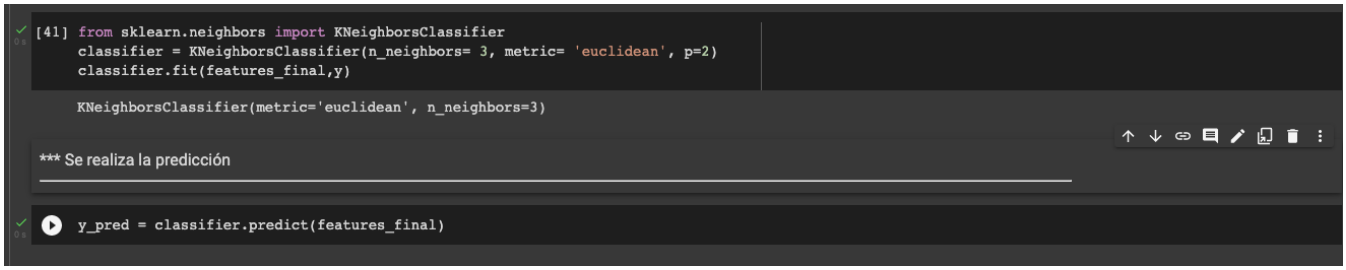
	Type
472	Espionage
346	Espionage
105	Espionage
357	Espionage
98	Espionage
...	...
409	Espionage
359	Espionage
15	Espionage
151	Espionage
303	Sabotage

Se definen los valores ‘x’ y ‘y’ y se utilizan los indicadores get dummy, ya que permite eliminar la primera de las columnas generadas para cada característica codificada.



	Defacement	Denial of service	Doxing	Espionage	Financial Theft	Sabotage
0	0	0	0	1	0	0
1	0	0	0	1	0	0
2	0	0	0	0	0	0
3	0	0	0	1	0	0
4	0	0	0	1	0	0

Se importa la función KNeighborsClassifier de Sklearn para crear el modelo y classifier nos ayuda a ajustar los parámetros del modelo y se indican 3 vecinos



Se crea dataframe a partir de diccionario, los valores originales y los valores que el modelo predijo

	y	y pred
0	Government	Government
1	Government	Government
2	Private sector	Private sector
3	Military	Government
4	Government, Private sector	Government
...	...	...
469	Government	Government
470	Government	Government
471	Military	Government
472	Military	Government
473	Military, Government	Government

Observando cómo se predice que el gobierno y el sector privado seguirán recibiendo un alto número de ciberataques.

Se obtienen las métricas, precisión, sensibilidad, puntuación F1 y matriz de confusión

```
confusion_matrix(y,y_pred)
```

```
array([[ 0,  0,  0,  0, 46,  3,  0,  0,  3,  0,  0,  0,  0,  0,  0,  0,
         0,  0,  0,  0,  0,  0],
       [ 0,  0,  0,  0,  5,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,
         0,  0,  0,  0,  0,  0],
       [ 0,  0,  1,  0,  5,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,
         0,  0,  0,  0,  0,  0],
       [ 0,  0,  0,  0,  2,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,
         0,  0,  0,  0,  0,  0],
       [ 0,  0,  5,  0, 89,  0,  0,  0,  4,  0,  0,  0,  0,  0,  0,  4,
         0,  0,  0,  0,  0,  0],
       [ 0,  0,  0,  0, 12,  2,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,
         0,  0,  0,  0,  0,  0],
       [ 0,  0,  0,  0,  1,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,
         0,  0,  0,  0,  0,  0],
       [ 0,  0,  2,  0, 14,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  1,
         0,  0,  0,  0,  0,  0],
       [ 0,  0,  1,  0, 42,  0,  0,  0,  1,  0,  0,  0,  0,  0,  0,  1,
         0,  0,  0,  0,  0,  0],
       [ 0,  0,  0,  0, 11,  0,  0,  0,  1,  0,  0,  0,  0,  0,  0,  0,
         0,  0,  0,  0,  0,  0],
       [ 0,  0,  1,  0, 17,  0,  0,  0,  3,  0,  0,  0,  0,  0,  0,  0,
         0,  0,  0,  0,  0,  0],
```

Resultados

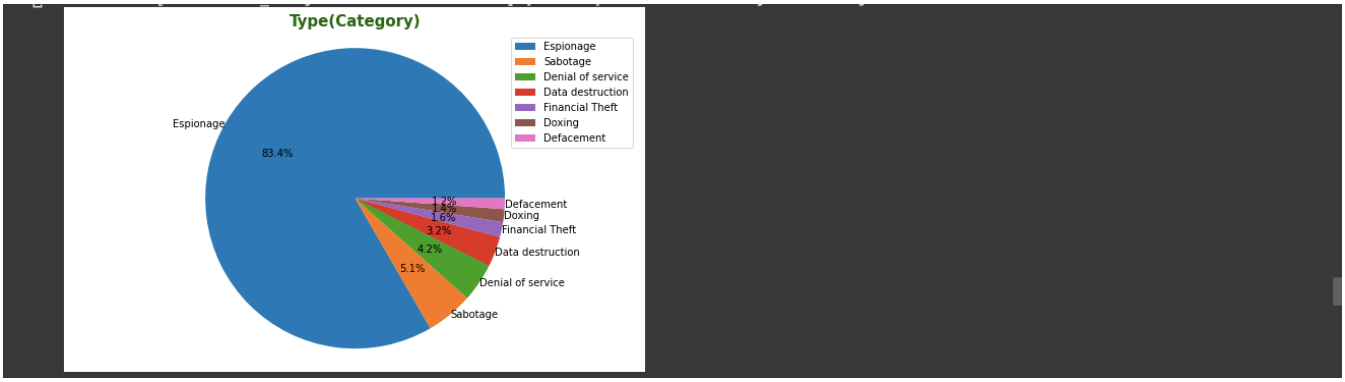
Por último, se grafican el tipo de ataque junto con la categoría observando el tipo de ataque que se realiza con más frecuencia.

```
[51] title_font = {"family": "arial", "color":"darkgreen","weight":"bold","size":15}
     axis_font = {"family": "arial", "color":"darkred","weight":"bold","size":10}
     for i,z in list(zip(categorical,categorical_axis_name)):
         fig, axis = plt.subplots(figsize=(10,6))

         observational_values = list(df2[i].value_counts().index)
         total_observation_values = list(df2[i].value_counts())

         axis.pie(total_observation_values,labels= observational_values, autopct = "%1.1f%%",startangle = 0 , labeldistance = 1.0)
         axis.axis("equal")

         plt.title((i+ "(" + z + ")"), fontdict = title_font)
         plt.legend()
         plt.show()
```

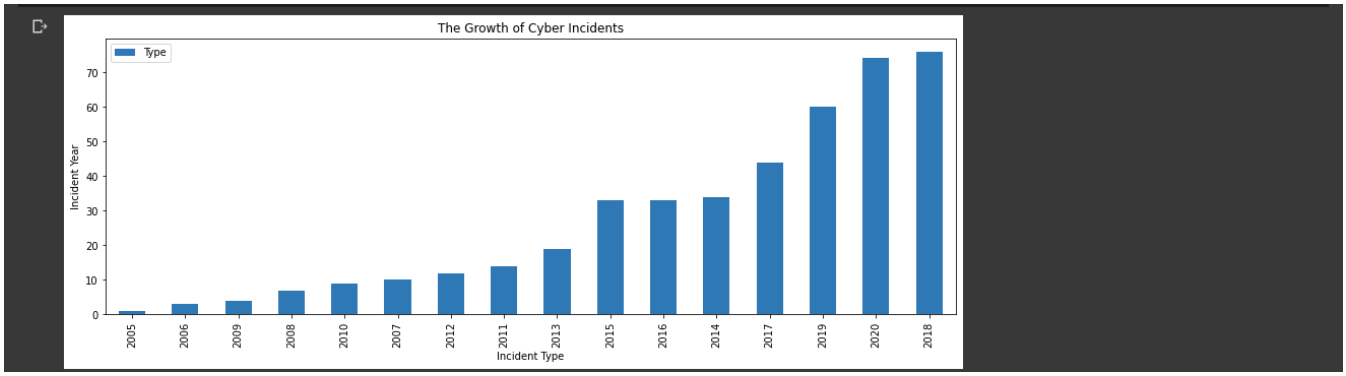


Se observa como el espionaje es el tipo de ataque que más se realizó con un 83.4%, seguido de sabotaje con un 5.1%, servicio denegado 4.2%, destrucción de datos 3.2%, robo financiero 1.6%, doxing 1.4% y desfiguración 1.2%.

Se grafica por año el aumento de los incidentes cibernéticos entre 2005 y 2020

```
[56] df2[['Year', 'Type']] \
                .groupby('Year') \
                .count() \
                .sort_values('Type', ascending=True)\
                .plot(kind='bar', figsize=(15,5));

    plt.title("The Growth of Cyber Incidents")
    plt.xlabel("Incident Type")
    plt.ylabel("Incident Year");
```



Se observa como el aumentó ha sido mayor en los últimos años, comparando como fue el aumento entre 2005 y 2010. Los mayores ataques comenzaron en el 2014, después solo ha ido aumentando el número.

Conclusiones

En conclusión, gracias a las diferentes técnicas, funciones y algoritmos de Minería de Datos se logra tener información valiosa de una gran cantidad de datos. Sin el conocimiento de esos diferentes algoritmos no se puede llegar a extraer la información necesaria para poder tener resultados, soluciones y observaciones a cerca de un problema, solo sería una inmensa cantidad de datos sin valor. En este caso, pudimos observar la importancia de agrupar la información y poder conocer cuál es el tipo de ataque que se realiza con más frecuencia y hacía que categoría va dirigido el ataque.

Se concluye que los aumentos de ciberataques seguirán siendo hacia el gobierno y sector privado de tipo espionaje. Gracias a la visualización de datos se puede concluir en prestar atención a estas categorías y que herramientas se deben de adquirir y prestar atención para evitar el mayor tipo de ataque realizado, siendo espionaje.

Referencias

https://www.kaggle.com/code/christinamq/growth-of-cyber-incidents-from-2005-2020#%F0%9F%92%AA%F0%9F%8F%BD-Conclusion

Liga Github

https://github.com/janethtrs/IncidentesCiberneticos