Universidad Rey Juan Carlos

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA INFORMÁTICA

GRADO EN INGENIERÍA DE LA CIBERSEGURIDAD

PRÁCTICA 2

SISTEMAS DE INFORMACIÓN

CURSO 2021 - 2022

Redactado por: Óscar del Río Jiménez

Daniel Paciencia Miguel

Rodrigo Regaliza Alonso

Índice

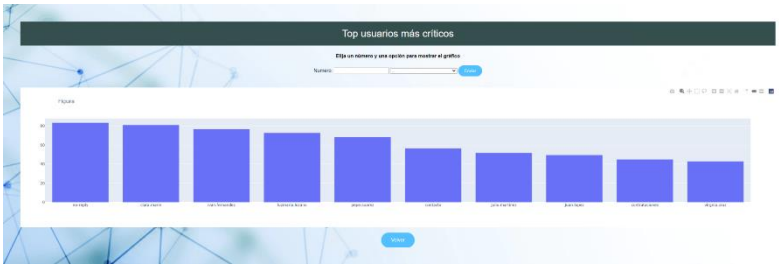
ENLACE AL REPOSITORIO DE GITHUB: 3
EJERCICIO 2: 3
EJERCICIO 3: 3
EJERCICIO 4: 4
EJERCICIO 5: 4
EJERCICIO 6: 5

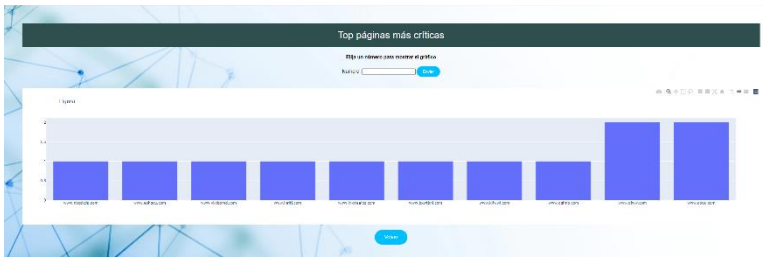
ENLACE AL REPOSITORIO DE GITHUB:

https://github.com/Oscarmanz/SSII-GB

EJERCICIO 2:

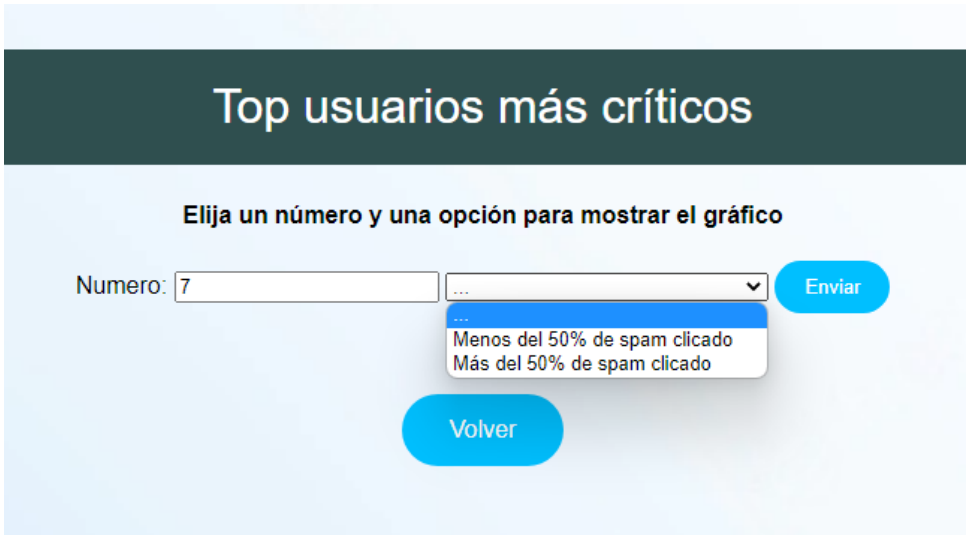
Hemos utilizado los datos de la práctica anterior para recoger los usuarios más críticos en una base de datos actualizable y posteriormente hemos representado esos datos en la web utilizando Flask. Para las webs más inseguras el procedimiento ha sido el mismo, utilizando los datos previos y Flask lo hemos representado en la web.





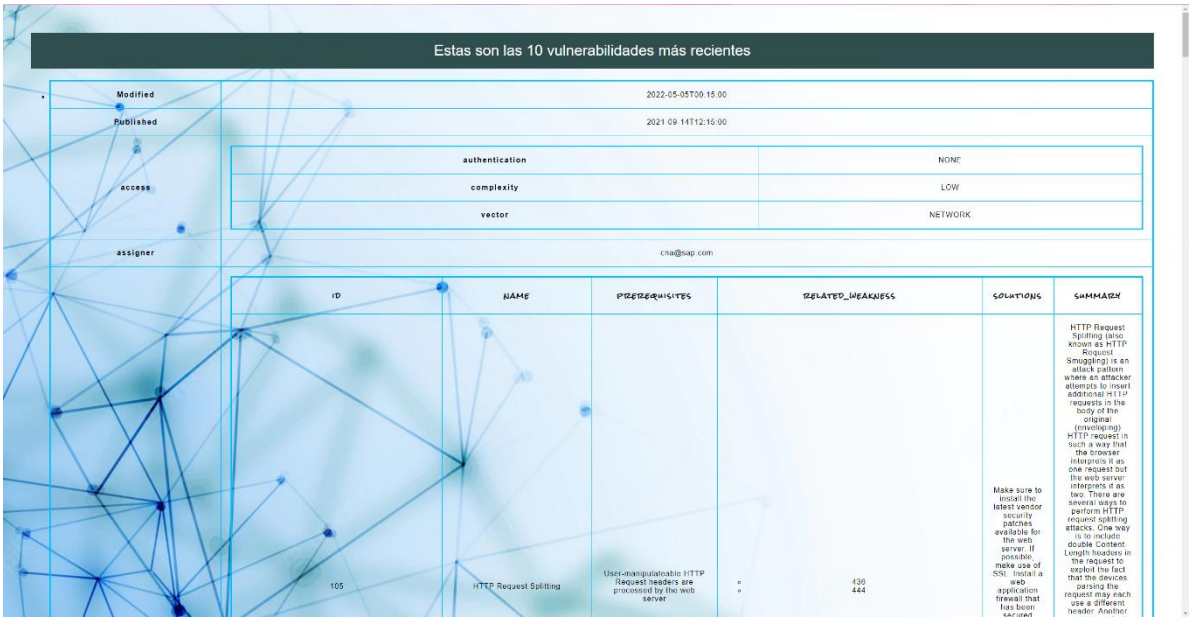
EJERCICIO 3:

Para este apartado hemos realizado una ampliación de la funcionalidad de la parte de usuarios críticos del ejercicio 2, añadiendo una opción en la misma página para visualizar el gráfico normal, los que han clicado más del 50% de emails y los que han clicado menos del 50%.



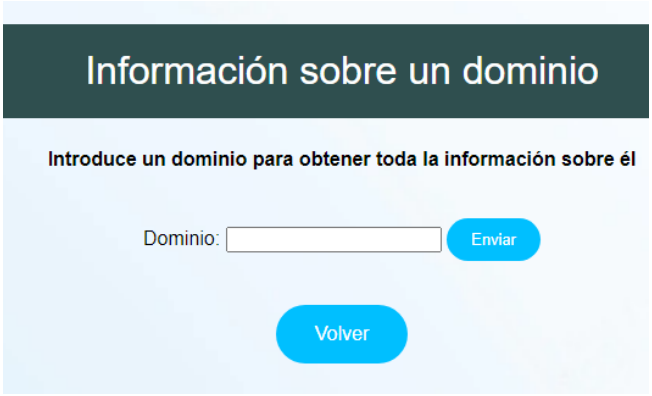
EJERCICIO 4:

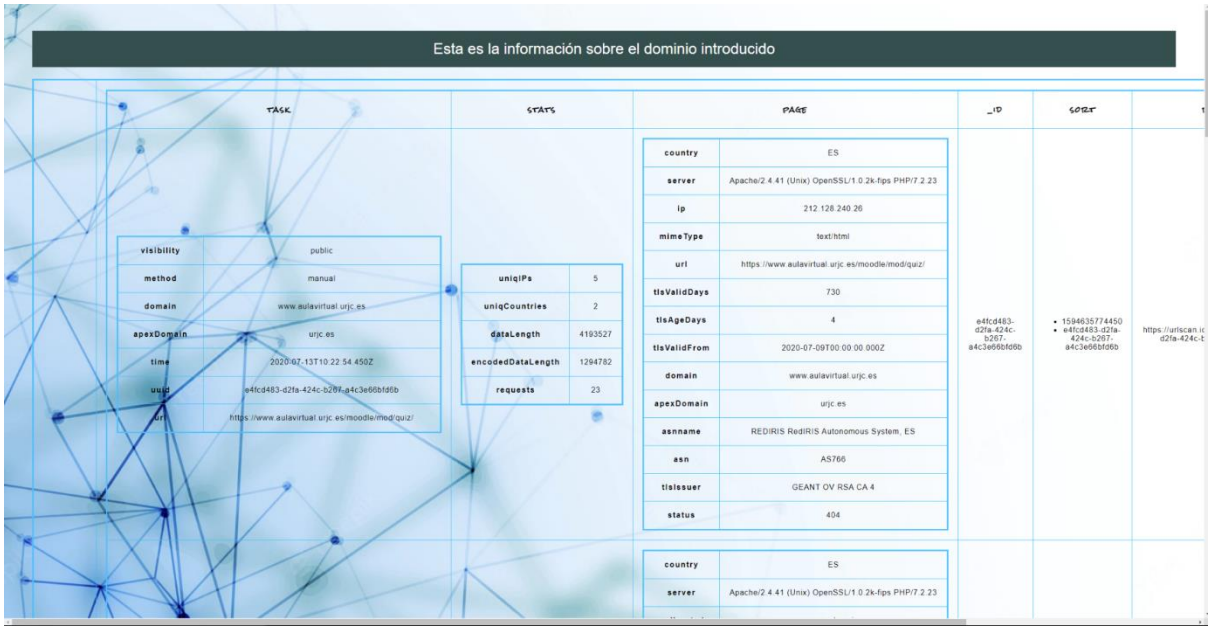
Para este apartado primero hacemos una petición a la API web y obtenemos sus resultados en JSON, después transformamos los resultados a HTML con la función json2html y luego lo incorporamos a nuestro template html utilizando Flask para que se muestren los datos en una tabla de la siguiente forma.



EJERCICIO 5:

Para este ejercicio hemos decidido utilizar otra API, en este caso hemos creado un formulario en el que introduces un dominio y a través de la API urlscan se obtiene información sobre él, después repitiendo la mecánica del ejercicio 4 lo mostramos en nuestra plantilla HTML con un formato más legible.





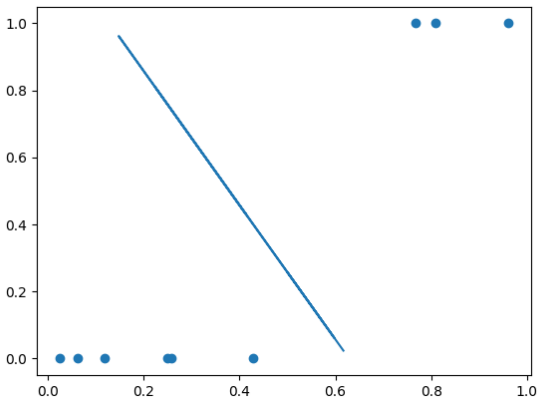
EJERCICIO 6:

Para este apartado hemos utilizado 3 modelos. Antes de utilizar cada uno hemos cargado los datos del JSON a nuestro código en Python, hemos eliminado los nombres ya que daban problemas y hemos separado los datos en entrenamiento y test. En nuestro caso hemos decidido que el 70% serían para entrenar y el 30% restante para el test.

El primer modelo es el de regresión lineal, lo que hace este modelo es representar en el eje X la probabilidad de pinchar en spam y en él Y si es crítico o no en función de esto. El resultado de la regresión es una recta que separa claramente 2 regiones, si está por debajo será no vulnerable y si está por encima lo será. Aplicado a los datos de test el accuracy es de 66,66% y en el modelo real los resultados obtenidos son los siguientes:

```
El modelo de regresion lineal ha predicho:
3  vulnerables y  6  no vulnerables
El accuracy es de:  66.66666666666666 %

El modelo de regresion lineal sobre los datos reales ha predicho:
7  vulnerables y  23  no vulnerables
--------------------------------------------------------------------
```

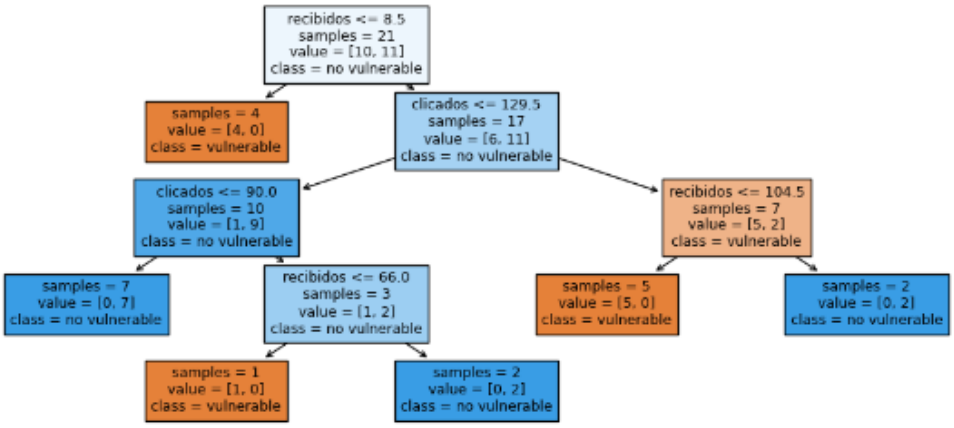


El segundo método es el del árbol de decisión, construimos el modelo y le pasamos los mismos datos que al anterior obteniendo lo siguientes resultados y la representación del árbol creado:

```
El modelo de arbol de decisión ha predicho:
5  vulnerables y  4  no vulnerables
El accuracy es de:  88.88888888888889 %

Profundidad del árbol: 4
Número de nodos terminales: 6

El modelo de arbol de decisión sobre los datos reales ha predicho:
10  vulnerables y  20  no vulnerables
```



Por último, el Random Forest que es similar al árbol ya que lo único que hace es crear varios árboles más sencillos con conjuntos de datos diferentes, luego para clasificar hace un consenso entre todos y deciden cual es el mejor resultado. Se obtiene lo siguiente y los gráficos de los árboles se crean en una carpeta llamada randomForestTrees.

```
El modelo de random forest ha predicho:
4  vulnerables y  5  no vulnerables
El accuracy es de:  77.77777777777779 %

El modelo de random forest sobre los datos reales ha predicho:
12  vulnerables y  18  no vulnerables
```

Vemos que el árbol de decisión es el mejor modelo con un acuraccy del 88% seguido del random forest y por último el de regresión. Sorprende que el random forest no sea el mejor pero probablemente se deba a la escasez de datos para el análisis ya que los árboles salen muy sencillos y es razonable que con un conjunto de datos mayor los resultados se ajustasen más a la realidad y el random obtuviese los mejores resultados.