Cuadro de textoInterfaz de usuario gráfica, Texto

Descripción generada automáticamente

**Proyecto de Prácticas**

**CREDENTIAL RISK ASSESSMENT TOOL**

**María Díaz Alba**

[**1.** **Introducción** 4](#_Toc172032308)

[**1.1.** **Contexto** 4](#_Toc172032309)

[**1.2.** **Especificaciones y Objetivos** 4](#_Toc172032310)

[**2.** **Fases del proyecto** 5](#_Toc172032311)

[**2.1.** **Investigación** 5](#_Toc172032312)

[**2.2.** **Desarrollo** 5](#_Toc172032313)

[**2.3.** **Documentación** 5](#_Toc172032314)

[**3.** **Herramientas utilizadas** 6](#_Toc172032315)

[**4.** **Desarrollo del procedimiento analítico** 7](#_Toc172032316)

[**4.1.** **Selección y definición de variables** 7](#_Toc172032317)

[**4.2.** **Desarrollo del Dataset** 8](#_Toc172032318)

[4.2.1. Recolección y preparación de datos 8](#_Toc172032319)

[4.2.1.1. Simulación de muestras 8](#_Toc172032320)

[4.2.1.2. Muestreo 9](#_Toc172032321)

[4.2.2. Sistema de asignación de puntuaciones y pesos 9](#_Toc172032322)

[4.2.2.1. Asignación de puntos 10](#_Toc172032323)

[4.2.2.2. Asignación de pesos 10](#_Toc172032324)

[**5.** **Análisis exploratorio de datos** 12](#_Toc172032325)

[**5.1.** **Objetivos del análisis exploratorio de datos** 12](#_Toc172032326)

[**5.2.** **Limpieza de datos** 12](#_Toc172032327)

[5.2.1. Manejo de datos faltantes 12](#_Toc172032328)

[5.2.2. Comprobación de columnas irrelevantes 12](#_Toc172032329)

[5.2.3. Eliminación de filas repetidas 13](#_Toc172032330)

[5.2.4. Estudio de valores extremos o atípicos (outliers) 13](#_Toc172032331)

[5.2.5. Corrección de errores tipográficos 13](#_Toc172032332)

[**5.3.** **Análisis individual de variables (Análisis univariado)** 13](#_Toc172032333)

[5.3.1. Comprensión y visualización de las variables numéricas 14](#_Toc172032334)

[5.3.2. Comprensión y visualización de las variables categóricas 16](#_Toc172032335)

[5.3.3. Conclusiones del análisis univariado 19](#_Toc172032336)

[**5.4.** **Análisis bivariado** 20](#_Toc172032337)

[5.4.1. Análisis y visualización de las relaciones entre pares de variables 20](#_Toc172032338)

[5.4.1.1. Diagramas de barras de tablas de contingencia (o cruzadas) 20](#_Toc172032339)

[5.4.1.2. Análisis de regresión mediante diagramas de dispersión 21](#_Toc172032340)

[5.4.1.3. Matriz de correlación 22](#_Toc172032341)

[5.4.1.4. Análisis de distribución condicional o de frecuencia por grupo 23](#_Toc172032342)

[5.4.2. Conclusiones del análisis bivariado 27](#_Toc172032343)

[**5.5.** **Análisis multivariado** 27](#_Toc172032344)

[5.5.1. Análisis y visualización de las relaciones entre múltiples variables 27](#_Toc172032345)

[5.5.1.1. Selección de características con el método SelectKBest 27](#_Toc172032346)

[5.5.1.2. Análisis discriminante de las características del bloque “Users” 28](#_Toc172032347)

[5.5.1.3. Multicolinealidad entre variables independientes 30](#_Toc172032348)

[5.5.1.4. Análisis multidimensional de las variables del bloque “Password” 33](#_Toc172032349)

[5.5.2. Conclusiones del análisis multivariado 34](#_Toc172032350)

[**5.6.** **Conclusiones generales del análisis exploratorio de datos** 35](#_Toc172032351)

[**6.** **Valor añadido del procedimiento a la operativa diaria** 36](#_Toc172032352)

[**ANEXO I** 37](#_Toc172032353)

[**ANEXO II** 39](#_Toc172032354)

## **Introducción**

### **Contexto**

El servicio de ciber inteligencia del departamento de ciberseguridad de TRC utiliza la plataforma de mensajería instantánea de Telegram en su operativa diaria para obtener información relevante para las medidas de ciberseguridad que este servicio ofrece de manera interna y externa.

En lo que respecta al contexto de este proyecto, cabe mencionar que uno de los intereses en esta plataforma se centra en la difusión de credenciales exfiltradas que tiene lugar por parte de una gran cantidad de usuarios a través de determinados canales y grupos que existen en esta plataforma. Se trata de una actividad ilícita común entre otras muchas que tienen lugar a través de esta aplicación, ya que debido a sus atractivas características que protegen la privacidad y actividad de sus usuarios Telegram se ha convertido en un terreno fértil para la ciberdelincuencia.

Ante este interés surge la idea de desarrollar un Bot en Telegram que de forma automatizada busque en estos grupos y canales las credenciales de los clientes que se benefician de este servicio.

El Bot, por lo tanto, tiene como objetivo fundamental la búsqueda filtrada de credenciales coincidentes con los dominios de nuestros clientes, y enfocada exclusivamente a archivos en formato .txt. Es decir, el procesamiento de la información por parte del Bot consiste en buscar, descargar y analizar a tiempo real todos los archivos en este formato que se envíen por cualquier grupo o canal al que pertenezca la cuenta de Telegram a través de la cual se realiza la monitorización. Si encuentra coincidencias, el Bot genera una alerta enviando un archivo en formato Excel con la información de las credenciales detectadas en un mensaje a un chat personal de la cuenta de Telegram.

### **Especificaciones y Objetivos**

A medida que la base de grupos y canales supervisados se expande, también lo hace el número de alertas generadas por el Bot. Este incremento continuo en las alertas impone una carga significativa sobre los analistas, quienes deben revisar y evaluar cada alerta manualmente. Este proceso es no solo laborioso, sino también propenso a errores humanos y falta de uniformidad en la evaluación del riesgo asociado a cada alerta.

Para optimizar el proceso de revisión de alertas y asegurar que se da prioridad a aquellas con mayor criticidad, se propone el desarrollo de un procedimiento analítico. Este procedimiento tiene como objetivo asignar automáticamente un grado de criticidad a cada alerta, utilizando una escala de 0 a 100. La criticidad se basará en una serie de características predefinidas que influyen en el riesgo potencial de las credenciales detectadas. Así, se busca mejorar la eficiencia y efectividad del proceso de revisión, permitiendo a los analistas enfocarse en las alertas más críticas primero.

## **Fases del proyecto**

En la gestión de cualquier proyecto, seguir un enfoque estructurado es clave para alcanzar el éxito. Este enfoque puede dividirse en varias fases, cada una con un propósito específico y vital para el avance del proyecto. En nuestro caso, hemos diseñado nuestro proyecto en torno a tres fases fundamentales: investigación, desarrollo y documentación.

### **Investigación**

Esta fase ha consistido en la familiarización con la aplicación de Telegram, especialmente en el contexto de filtraciones de datos a través de grupos y canales. También, se ha tratado de observar con precisión las alertas generadas por el Bot en concreto para conocer las características más relevantes a la hora de determinar el grado de riesgo que pueda presentar una credencial filtrada en Telegram.

### **Desarrollo**

En esta fase se transforman las ideas y conclusiones obtenidas en la investigación en soluciones concretas y funcionales. Durante esta etapa se ha planteado, diseñado y desarrollado la herramienta con la que se predice el valor de riesgo que presenta una credencial, basada en las características obtenidas de la fase anterior.

### **Documentación**

Finalmente, llegamos a la fase de documentación. Aunque a menudo subestimada, esta fase es esencial para el éxito a largo plazo de cualquier proyecto. Aquí, recopilamos y organizamos toda la información relevante sobre el desarrollo y la implementación del proyecto, creando una guía detallada que servirá tanto para la referencia futura como para la optimización continua.

## **Herramientas utilizadas**

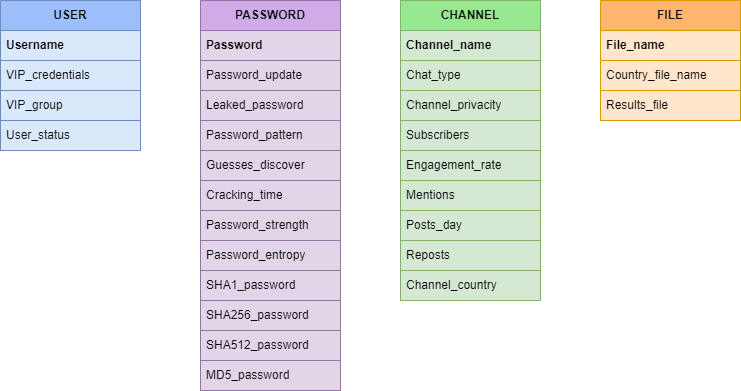
## **Desarrollo del procedimiento analítico**

### **Selección y definición de variables**

El primer paso en la creación del procedimiento analítico es identificar y seleccionar las variables que influirán en la evaluación del riesgo de una credencial filtrada. Tras un análisis exhaustivo del funcionamiento de Telegram y una revisión de las alertas generadas por el Bot, se determinan las características más relevantes. Estas variables se dividen en dos categorías principales: variables independientes (que determinan el riesgo) y una variable dependiente (que representa el riesgo o criticidad de la credencial). Véase la “Documentación de Variables del Dataset” anexa.

Para facilitar la evaluación, las variables independientes seleccionadas se organizan en cuatro bloques según el tipo de información que aportan sobre el riesgo global de la credencial, teniendo:

* Bloque “USER”: Las variables de este bloque hacen referencia al usuario de la credencial y a la información que este aporta al riesgo de la credencial completa.
* Bloque “PASSWORD”: Las variables de este bloque hacen referencia a la contraseña de la credencial y a la información que esta aporta al riesgo de la credencial completa.
* Bloque “CHANNEL”: Las variables de este bloque hacen referencia al canal o grupo por el que se comparte el archivo en el que se detecta la credencial y a la información que este aporta al riesgo de la credencial completa. Se le da nombre de “channel” porque para el caso de difusión de credenciales y data leaks en Telegram, es más habitual que ocurra a través de canales que de grupos.
* Bloque “FILE”: Las variables de este bloque hacen referencia al fichero en formato .txt que contiene las credenciales detectadas por el Bot y a la información que este aporta al riesgo de la credencial completa.



Se presenta además una tabla que resume las variables seleccionadas, su bloque correspondiente, el tipo de datos que representan, y una breve descripción en el libro de Excel “Proyecto Bot de Telegram.xlsx” en la hoja “Dataset”.

### **Desarrollo del Dataset**

Una vez definidas todas las variables, se procede a la creación del dataset, que es el conjunto de datos estructurados que se utilizará para extraer y analizar patrones, investigar y diseñar el procedimiento analítico.

Para diseñar la estructura del dataset, partimos de una serie de variables independientes y una variable dependiente, que se organizarán en una tabla. Esta tabla se completará posteriormente con los datos correspondientes a cada variable (columnas) para cada muestra estudiada (filas).

La variable dependiente se completará con un valor numérico basado en todas las variables independientes del dataset. La influencia de cada variable en el grado de criticidad de una credencial determinará el valor de la variable dependiente.

##### Recolección y preparación de datos

Una vez creada la estructura del dataset, se deben recolectar y preparar los datos necesarios de manera ética y responsable. Es fundamental cumplir con todas las regulaciones éticas y legales en la recolección y uso de datos de credenciales, incluyendo el consentimiento informado, la anonimización de datos y la protección de la privacidad de los usuarios.

Las supuestas credenciales detectadas por el Bot y almacenadas en una base de datos, debido a la información que alberga estarían reservadas a los analistas con la Habilitación Personal de Seguridad[[1]](#footnote-2) (HPS) aprobada que da acceso a documentos clasificados.

Al ser estudiante en prácticas y no poseer esta habilitación no se han podido utilizar credenciales reales para el pool de muestras del proyecto. Por lo tanto, es necesario generar un pool de muestras inventadas y aleatorias a través de una simulación.

###### Simulación de muestras

Para completar los datos del dataset, se simularán muestras representativas de credenciales. Estas simulaciones deben reflejar las condiciones y características reales de las credenciales filtradas en Telegram, componiéndose de cuatro variables fundamentales: el nombre de usuario, la contraseña, el nombre del canal en el que se ha detectado la credencial y el nombre del archivo que contiene la credencial.

La simulación de nombres de usuario se realiza creando un dominio que simule la entidad o cliente que obtiene el servicio de monitorización del Bot. En este caso, se creó el dominio inventado @telebot.com. Luego, se crea una lista de usuarios inventados con este dominio y se diseña un diccionario en el que se determina el puesto de trabajo de cada usuario, así como su estatus (activo o inactivo).

Las contraseñas se generan automáticamente con un script que crea contraseñas de longitudes y mezcla de caracteres aleatorias, intercalando con contraseñas de un diccionario de contraseñas.

Para los canales, se crea un diccionario con los nombres de los canales y grupos de Telegram que se quieren utilizar el análisis. En este caso, los nombres pueden ser inventados o reales, pero es necesario determinar el tipo de chat (canal o grupo) y su tipo de privacidad (público o privado).

De manera similar, se crea un diccionario con los nombres de los archivos que contienen las credenciales. Se incluyen nombres de archivos que contengan palabras como “Spain” y “Europe” en cualquier idioma o términos similares, como “SP” o “EU”.

Para crear una gran cantidad de muestras aleatorias que combinen estas cuatro variables, se puede hacer manualmente usando herramientas en línea para aleatorizar listas, o se puede usar el script ***randomizer.py*** diseñado específicamente para ello, siendo esta última una opción más eficiente.

###### Muestreo

Este paso es crucial para asegurar que el dataset utilizado para desarrollar el procedimiento sea representativo y de alta calidad. Una vez generadas las muestras aleatorias, se introducen en el dataset y se completan todas las variables independientes del dataset basadas en la información obtenida de las muestras.

Por ejemplo, para una credencial cuyo nombre de usuario sea [xxxx@telebot.com](mailto:xxxx@telebot.com), se completarán las características del bloque de usuarios según la información correspondiente, como su estatus (activo o inactivo) y clasificación (VIP o no VIP), y a qué grupo de usuarios VIP corresponde en tal caso.

El muestreo se realiza con el script ***muestreo.py*** con el que de forma automatizada podemos completar todas las variables independientes para la muestra analizada formada por el nombre de usuario, contraseña, nombre de canal y nombre de archivo.

##### Sistema de asignación de puntuaciones y pesos

Para completar los datos de la variable dependiente, primero se debe establecer un sistema interno de asignación de puntuaciones y pesos para cada variable independiente que influye sobre el resultado final. De esta manera, una vez se tengan los datos sobre una muestra, se obtendrá además el valor del grado de criticidad representado por la variable dependiente.

Se utiliza Excel como herramienta para completar los datos, con la que se diseñan las fórmulas necesarias para que, a medida que se rellenan los datos de todas las variables independientes del dataset, se obtenga automáticamente una puntuación basada en la asignación de puntos que se define a continuación. Los datos de la variable dependiente se completarán a partir de una fórmula que suma las puntuaciones de las variables de cada bloque multiplicadas por el peso de cada bloque, obteniendo así el valor final sobre la criticidad de la credencial. Una vez recolectados todos los datos de las 200 muestras, se creará el dataset final con los datos de todas las variables independientes junto con los datos de la variable dependiente.

###### Asignación de puntos

Por un lado, cada una de las variantes o posibilidades de las variables independientes recibirá una puntuación específica en una escala de 0 a 10 basada en su contribución al riesgo global de la credencial, donde 0 corresponde a la posibilidad menos crítica y 10 a la más crítica.

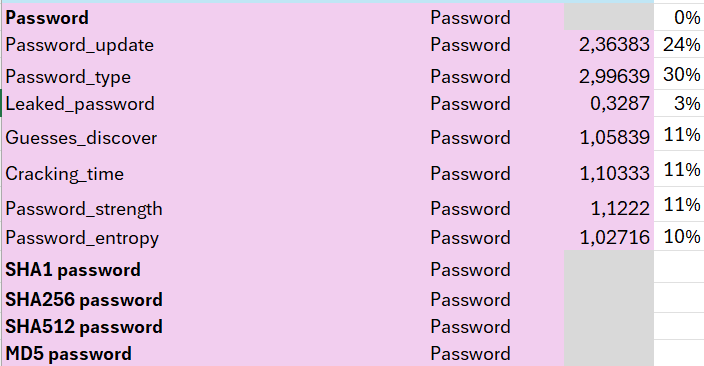
Según el tipo de dato de las variables, las variantes se definen por intervalos para aquellas de tipo numérico o por categorías para aquellas de tipo categórico. Las diferentes variantes de las variables y su correspondiente puntuación, así como el criterio aplicado para establecerla se reflejan en el documento ubicado en la documentación interna del sharepoint en la ruta: Documentos\DocumentaciónInterna\ÁreadeCiberseguridad\Bot\_Telegram\DocumentaciónFinal\Definición Variables del Dataset.docx

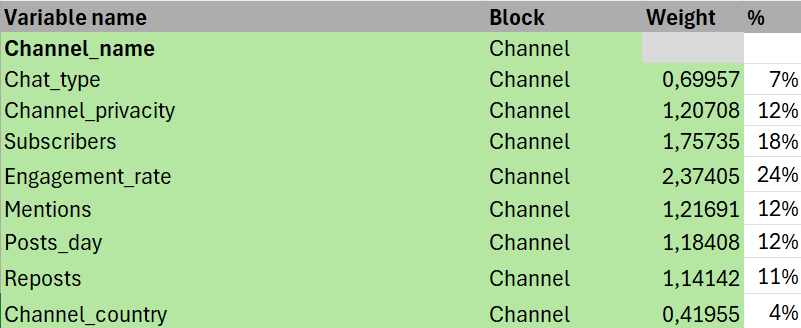
###### Asignación de pesos

Además de las puntuaciones individuales, se asignarán pesos a cada variable y a cada bloque de variables (USER, PASSWORD, CHANNEL, FILE) según su relevancia en la evaluación del riesgo.

En primer lugar, se determina individualmente el peso que representa cada variable independiente dentro de su correspondiente bloque. A continuación, se puede ver la asignación de pesos para cada variable por bloque, donde la suma de pesos de cada bloque equivale a 10 (o un 100%).







Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

El siguiente paso es definir el peso de cada bloque sobre el valor de la variable objetivo en relación con el grado de criticidad de una credencial. Se establecen los pesos por bloque de la siguiente manera:



El criterio para asignar los pesos se basa en la investigación y observación inicial, donde en un primer momento se determinaron unos pesos provisionales y a medida que se obtuvieron resultados estos se fueron ajustando hasta alcanzar los pesos que daban resultados mejor ajustados a una valoración humana por parte de un analista especializado en esta actividad.

Respecto a los bloques, se considera que las características del bloque de usuarios son más críticas sobre el riesgo de una credencial, con una notable diferencia respecto al resto de bloques, debido a que el usuario es la parte más crítica en una filtración. Al fin y al cabo, no tenemos forma de demostrar si la contraseña asociada a ese usuario es real o inventada por el ciberdelincuente, e independientemente de sus características, si esta fuese real el riesgo ya estaría presente, en cualquier caso. Por este motivo el peso del bloque de las contraseñas es relativamente bajo, aunque se sigue teniendo en cuenta para evaluar y considerar las contraseñas igualmente. En cambio, el usuario es relativamente sencillo de obtener, por ejemplo, investigando el patrón que emplea una empresa para el username de sus empleados, de manera que es común que los usuarios sí sean reales, aunque las contraseñas no lo sean, y esto llevaría ya implícito un riesgo de amenazas como el phishing, por ejemplo. Por su parte, el bloque del canal y el archivo tienen un peso bastante bajo, pero presente aún así para que de esta forma marque una diferenciación especialmente para aquellos casos en los que los usuarios de varias credenciales sean de similares características.

En resumen, con esta asignación de pesos, las características del bloque de usuarios serán las que determinen con mayor influencia el riesgo de la credencial, correspondiéndose con un 75% del total, pero sin embargo el resto de las características sobre la contraseña, el canal y el archivo, aunque en menor medida, siguen formando parte de ese 25% que en muchos casos será determinante para poder diferenciar el riesgo de las credenciales.

## **Análisis exploratorio de datos**

### **Objetivos del análisis exploratorio de datos**

El Análisis Exploratorio de Datos (EDA) desempeña un papel crucial en la validación y comprensión del dataset diseñado específicamente para este proyecto. El objetivo principal de este análisis es asegurar la coherencia y la integridad de la variable a predecir, así como estudiar el comportamiento de todas las variables predictoras incluidas en el dataset.

El EDA proporciona una base sólida para preparar el dataset para el modelado predictivo. Esto incluye decisiones sobre transformaciones de variables, manejo de datos faltantes y selección de variables relevantes. Además, podremos con esta fase asegurar que los datos son confiables y útiles, el dataset será eficaz y que la variable objetivo, derivada del sistema de puntuaciones y pesos, es válida y refleja adecuadamente el fenómeno de interés.

Cada una de las fases del análisis exploratorio de datos se puede consultar en los cuadernos de *Jupyter Lab* almacenados en la carpeta “Análisis EDA”, en los cuales se incluyen los cuerpos de código necesarios para realizar todos los apartados de cada una de las fases de este análisis, así como sus correspondientes explicaciones.

### **Limpieza de datos**

En el ámbito del análisis de datos, la limpieza de datos es una fase fundamental y crítica que precede cualquier tipo de análisis exploratorio o modelado predictivo. Los datos de baja calidad pueden llevar a conclusiones incorrectas, modelos ineficaces y decisiones empresariales equivocadas. Para que las conclusiones sean precisas y confiables, se realizan las siguientes acciones:

##### Manejo de datos faltantes

Los datos faltantes son valores ausentes en una o más columnas de un dataset. Pueden surgir debido a errores de recolección, problemas de almacenamiento o simplemente porque la información no estaba disponible. Para identificar si en nuestro dataset hay datos faltantes podemos utilizar la función .info() para comprobar si hay el mismo número de registros en todas las columnas.

La gestión de datos faltantes incluye varias estrategias, como eliminar la fila entera, aplicar la media aritmética, la moda, la mediana, completar con el dato más repetido de esa columna o el menos repetido. En nuestro caso, aplicaremos la estrategia de completar con el dato más frecuente.

##### Comprobación de columnas irrelevantes

No todas las columnas en un dataset son útiles para el análisis. Algunas pueden no aportar información valiosa o pueden estar fuera del alcance del análisis en cuestión. La identificación y eliminación de columnas irrelevantes es crucial para simplificar el análisis y mejorar la eficiencia de los modelos.

Para ello se comprueban los subniveles de las variables categóricas, para detectar aquellas variables que tengan un único subnivel y por tanto no sean significativas, con la función .*nunique().* En este caso, no se encontró ninguna de estas características.

Para las variables numéricas se comprueba que la desviación estándar de sus valores no sea igual a 0, lo cual indicaría que no aporta valor al dataset, con la función *.describe()*. Tampoco se encontró en este caso ninguna columna numérica con estas características.

##### Eliminación de filas repetidas

Las filas duplicadas son copias exactas de otras filas y pueden inflar artificialmente los resultados del análisis, llevando a conclusiones erróneas. Identificar y eliminar estas filas es esencial para mantener la integridad de los datos.

Para identificarlas en nuestro dataset se visualizan las filas repetidas con la función .duplicated(). Luego se eliminan haciendo uso de la función drop\_duplicates(inplace=True) las filas repetidas identificadas**.**

##### Estudio de valores extremos o atípicos (outliers)

Los valores atípicos son observaciones que se desvían significativamente del resto de los datos. Estos pueden ser resultado de errores de entrada de datos, variaciones naturales o eventos excepcionales. Los valores extremos pueden distorsionar los resultados del análisis.

Estos valores siempre se refieren a datos numéricos, y no siempre deben ser eliminados, aunque si deben estudiarse todos para comprobar su relevancia en el dataset, con ayuda de representaciones gráficas de tipo boxplot.

##### Corrección de errores tipográficos

Este tipo de errores son inconsistencias en la entrada de datos, como errores de ortografía, uso inconsistente de mayúsculas, formatos incorrectos, etc. Estos errores pueden causar problemas en el análisis y deben ser corregidos.

Para ello comprobaremos las categorías únicas existentes para cada una de las variables categóricas del dataset, con ayuda de un diagrama de barras que muestre todas las categorías únicas. Una vez detectados los errores tipográficos, podremos corregirlos renombrando las categorías.

### **Análisis individual de variables (Análisis univariado)**

El análisis individual de las variables tiene como objetivo entender cada variable del dataset de manera independiente antes de explorar las relaciones entre ellas. Al realizar un análisis individual de las variables, se pueden identificar patrones, distribuciones y posibles anomalías que son esenciales para cualquier análisis posterior y ayudan a entender la naturaleza de los datos

##### Comprensión y visualización de las variables numéricas

Para comprender la distribución de las variables numéricas, se pueden calcular medidas de tendencia central (media, mediana, moda) y de dispersión (rango, desviación estándar, varianza).

Con la función .*describe()* podemos obtener esta información de todas las variables numéricas del dataset.

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Podemos visualizar la distribución de las variables mediante histogramas, gráficos de cajas y bigotes para proporcionar una visión clara de la forma y la dispersión de los datos, facilitando la comprensión de estas distribuciones individuales y el comportamiento de variable.

Gráfico, Gráfico de barras, Histograma

Descripción generada automáticamente Gráfico, Histograma

Descripción generada automáticamente

Gráfico, Histograma

Descripción generada automáticamente Gráfico, Histograma

Descripción generada automáticamente

Gráfico, Histograma

Descripción generada automáticamente Gráfico, Histograma

Descripción generada automáticamente

Para aquellas variables con datos atípicos cuyas representaciones son más complejas de comprender visualmente, podremos realizar un ajuste eliminando aquellos datos atípicos para su representación. Esto no quiere decir que no se tengan en cuenta, ya que estos datos son significativos, pero dificultan la visualización gráfica del resto de valores.

Podemos emplear para ello una representación distinta, por ejemplo un diagrama de violín, que representa la distribución de los datos y su densidad de probabilidad.

Eliminamos los outliers de “Leaked\_password” y “Craking\_time” para una visualización más precisa del resto de valores. De esta manera para estas dos variables ya podemos entender que en Leaked password la mayoría de datos equivalen a 0 o son valores próximos a este número y a medida que aumentan va observándose una menor densidad. En cambio en Cracking time aunque también se acumula la mayoría de datos en el valor de 0, al realizar el ajuste descubrimos que hay otra notable acumulación en el valor de 10^11.

Gráfico

Descripción generada automáticamente con confianza media

Gráfico, Gráfico de dispersión, Gráfico de cajas y bigotes

Descripción generada automáticamente

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Gráfico

Descripción generada automáticamente

Observaciones:

* Algunas variables presentan picos notables en su distribución de valores bajos cercanos a 0, como “Leaked\_password” o “Cracking\_time”, “Posts\_day” o “Reposts”.
* Algunas variables presentan una distribución más homogénea como “Guesses\_discover” o “Results\_file”.
* La variable dependiente “TOTAL” presenta una distribución aproximadamente simétrica y normal, con un pico referente a la media en valores cercanos a 50.

##### Comprensión y visualización de las variables categóricas

Para el estudio de las variables categóricas, se analizará la frecuencia de aparición de cada categoría en el conjunto de datos. Con este análisis podemos visualizar la distribución de los datos, concluyendo qué características de las credenciales son las que se repiten con mayor frecuencia en las alertas generadas por el Bot.

Escala de tiempo

Descripción generada automáticamente con confianza media

Gráfico

Descripción generada automáticamente

Gráfico, Gráfico de barras

Descripción generada automáticamente

Gráfico

Descripción generada automáticamente

Observaciones:

* Podemos diferenciar visualmente las variables binarias de las no binarias, así como identificar todas las categorías presentes en cada una de estas variables categóricas, incluyendo las 5 categorías de la variable dependiente a predecir.
* La mayoría de las variables categóricas de este dataset presentan una diferencia notable en la distribución de datos entre sus categorías. En el caso de las variables binarias siempre hay una categoría que presenta una mayor frecuencia de datos. En el caso de las no binarias la distribución tampoco resulta homogénea, por ejemplo en la variable “Password\_type” hay una clara mayoría de datos clasificados como “Personal\_password”.

En el caso de la variable categórica nominal “*CHANNEL\_NAME*” resulta también de interés visualizar la distribución de sus datos. En este caso cada categoría corresponde a un canal monitorizado por el Bot de Telegram en el que se han detectado credenciales filtradas, por lo tanto, con esta representación podremos observar cuáles son los canales que difunden una mayor cantidad de credenciales.

En el caso de la variable categórica nominal “*CHANNEL\_NAME*” resulta también de interés visualizar la distribución de sus datos. En este caso cada categoría corresponde a un canal monitorizado por el Bot de Telegram en el que se han detectado credenciales filtradas, por lo tanto, con esta representación podremos observar cuáles son los canales que difunden una mayor cantidad de credenciales.

Gráfico

Descripción generada automáticamenteObservaciones:

* En el diagrama se observa que los grupos son los chats en los que se han detectado más credenciales, seguidos del canal “NoName\_Cloud” situado el segundo en el ranking. Por el contrario, “HQ COMBO | FRESH BASES | VALID COMBOLIST” es el canal en el que menos actividad de este tipo se ha detectado por el Bot según el dataset estudiado.

##### Conclusiones del análisis univariado

Un análisis detallado de cada variable proporciona información valiosa que puede influir en la selección y preparación de datos para el modelado predictivo. Entender la distribución y las características de cada variable permite elegir las transformaciones adecuadas y los algoritmos de modelado más efectivos.

Al realizar el análisis individual de las distribuciones de las variables podemos conocer a qué tipo de usuario corresponden la mayoría de las credenciales que aparecen filtradas en Telegram, de qué tipo son las contraseñas que suelen filtrarse con más frecuencia y sus características. También se podría estudiar de qué países provienen la mayor parte de filtraciones mirando el origen de los canales o incluso el nombre de los archivos, así como el tipo de chat por el cual se difunde una mayor cantidad de credenciales detectadas por el Bot.

Una vez tenemos una idea clara de la naturaleza de nuestros datos y de sus distribuciones podemos empezar a plantear hipótesis sobre su comportamiento en el dataset y sus posibles relaciones o influencia sobre otras variables.

### **Análisis bivariado**

El análisis bivariado examina dos variables para identificar relaciones y correlaciones entre ellas, pudiendo revelar cómo los cambios en una variable afectan a otra. Este tipo de análisis es crucial para comprender cómo las variables interactúan y se influyen mutuamente, proporcionando una visión más profunda y rica de los datos en comparación con el análisis univariado. Este análisis presenta los siguientes objetivos:

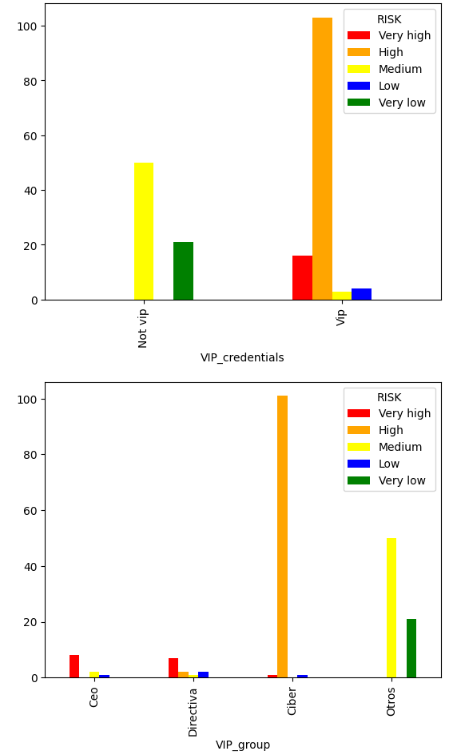
* **Identificación de Relaciones:** El análisis bivariado permite detectar la presencia y la naturaleza de las relaciones entre dos variables, ya sean lineales, no lineales, positivas, negativas o nulas. Esto es vital para comprender las dinámicas subyacentes en los datos y formular hipótesis sobre posibles causalidades.
* **Evaluación de la Fuerza y la Dirección de la Relación:** Mediante el uso de coeficientes de correlación (como el coeficiente de Pearson), el análisis bivariado cuantifica la fuerza y la dirección de la relación entre dos variables. Esto ayuda a identificar qué variables están más estrechamente relacionadas y podrían ser más relevantes en un procedimiento analítico o modelo predictivo.
* **Detección de Patrones y Tendencias:** Visualizar las relaciones entre dos variables a través de gráficos de dispersión, diagramas de caja y bigote, o gráficos de barras facilita la identificación de patrones y tendencias.
* **Información para el Modelado Predictivo:** Identificar variables que tienen una fuerte relación con la variable objetivo puede guiar la selección de características y la construcción del modelo, mejorando su precisión y efectividad.

##### Análisis y visualización de las relaciones entre pares de variables

###### Diagramas de barras de tablas de contingencia (o cruzadas)

Haciendo uso de las tablas de contingencia podemos representar datos categóricos en términos de conteos de frecuencia, es decir, analizar la cantidad de credenciales del dataset que poseen un subnivel de una de las características analizadas y otro subnivel de otra característica de forma simultánea. Podremos visualizar la relación de cada una de las variables con la variable “RISK” a través de diagramas de barras con leyenda que creamos con *pandas* y *matplotlib.pyplot*, accesibles en el [Anexo I](#_ANEXO_I).

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente 

Observaciones:

* Las características asociadas en mayor grado a las categorías que denotan mayor criticidad de las credenciales (“very high” y “high”) son de forma destacable según estas visualizaciones
  + Los usuarios VIP y de los grupos Ceo y Directiva, así como los usuarios activos.
  + Aquellas contraseñas de tipo personal, con puntuaciones de fortaleza más bajas (Password\_strength) y no actuales.
  + Las credenciales provenientes de canales públicos provenientes de grupos no españoles.

###### Análisis de regresión mediante diagramas de dispersión

Existen métodos que modelan la relación entre una variable dependiente y una o más variables independientes, proporcionando una comprensión más detallada de cómo las variables están relacionadas. Para las variables numéricas, podemos hacer un análisis de regresión para comprobar si existe una regresión lineal visualizando el análisis mediante gráficos de dispersión, que permiten hacerse una idea de la fuerza de correlación que estas presentan.

Gráfico

Descripción generada automáticamente

Gráfico, Histograma

Descripción generada automáticamente

Observaciones:

* En relación con la variable dependiente “TOTAL” no se observa a penas relación lineal entre las variables independientes y esta.
* Entre las variables independientes sí se observa relación lineal con alta fuerza de correlación en el caso de “Password\_entropy” y “Guesses\_discover”, y con algo menos fuerza de correlación en el caso de “Resposts” y “Subscribers”.

###### Matriz de correlación

Existen medidas estadísticas que cuantifican la dirección y la fuerza de la relación entre dos variables continuas; la correlación y la covarianza. Podemos crear una matriz de correlación para visualizar de manera conjunta las relaciones entre todos los posibles pares de variables y seleccionar los que resulten más relevantes para reducir el campo de visión a determinados grupos de variables.

Para realizar la matriz incluyendo todas las variables se realiza un preprocesamiento de los datos, transformando las variables categóricas en variables binarias con un codificador *OneHot*, y además escalando todas las variables numéricas con el escalador *MinMaxScaler*, quedando así todos los datos del dataset con valores entre 0 y 1.

Con los índices de correlación además podemos determinar qué variables independientes tienen una relación inversa con respecto a la variable dependiente (índice negativo) y cuáles tienen una relación directa (índice positivo).

Así, además, podremos visualizar cuáles son las variables independientes que tienen una mayor relación lineal con la variable a predecir “TOTAL”, siendo aquellas con los índices más cercanos a 1 o -1.

Imagen que contiene Gráfico de dispersión

Descripción generada automáticamente

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente Imagen que contiene Escala de tiempo

Descripción generada automáticamente

###### Análisis de distribución condicional o de frecuencia por grupo

Este análisis se enfoca en contar la frecuencia de aparición de un canal dentro de cada nivel de otra variable categórica, es decir, se examina cómo la distribución de los canales varía entre las diferentes categorías de otra variable. Este análisis es útil para identificar patrones y relaciones entre las dos variables, determinar si ciertos canales son más comunes en algunas categorías que en otras.

Para su representación visual se realiza un diagrama de barras interactivo con Dash en el que se representa la frecuencia de los canales para cada una de las variables categóricas, pudiendo seleccionar para cada una de ellas la categoría de la cual se quiere visualizar el recuento de aparición de los canales.

Podríamos comparar qué canales son los que distribuyen más cantidad de credenciales de usuarios de cada departamento.

Gráfico, Gráfico de barras

Descripción generada automáticamente

Imagen que contiene Gráfico

Descripción generada automáticamente

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente con confianza media

Gráfico, Gráfico de barras

Descripción generada automáticamente

Observaciones:

* “*FREE RED/VPS & BINNERS CC*” es el canal de Telegram en el que más credenciales del grupo “Ceo” se han detectado con el Bot y en “*NoName\_Cloud*” credenciales del grupo “Directiva”.
* Las credenciales de usuarios pertenecientes a otros departamentos se detectan con más frecuencia en grupos públicos.

También podríamos estudiar la distribución de los canales según la categoría de la variable dependiente, es decir, comprobar qué canales suelen tener con más frecuencia un determinado grado de criticidad en la valoración final de la credencial detectada en ese canal.

Se muestran en las siguientes gráficas los canales en los que se detectan más credenciales de riesgo “very high”, y los canales en los que se detectan más credenciales de riesgo “very low”.

Gráfico, Gráfico de barras

Descripción generada automáticamente

Gráfico

Descripción generada automáticamente con confianza baja

Gráfico

Descripción generada automáticamente con confianza bajaObservaciones:

* El canal “*FREE RED/VPS & BINNERS CC*” es el canal en el que se han detectado más credenciales reconocidas de riesgo “very high”.
* El canal “Daily Combolist [Backup]” es el canal en el que se han detectado más credenciales reconocidas de riesgo “very low”.

Por otro lado, podemos aprovechar esta técnica para visualizar la clasificación de los canales, por ejemplo, según la privacidad del canal, recogida en los datos que se almacenan en la variable “*Channel\_privacity*”. De esta forma podemos reconocer de forma clara cuáles son los canales y grupos privados y cuáles son públicos, así como ver la cantidad de veces que aparece cada uno de estos dentro de su categoría.

Gráfico, Gráfico de barras

Descripción generada automáticamente

Gráfico

Descripción generada automáticamente con confianza media

Observaciones:

* Hay una mayor cantidad de grupos / canales de tipo público que privados.
* En la categoría de chat público, la mayoría de las credenciales analizadas se han detectado en un grupo.
* En la categoría de chat privado, la mayoría de las credenciales analizadas se han detectado en el canal “NoName\_Cloud”.

##### Conclusiones del análisis bivariado

Estos análisis de pares de variables que nos permiten comprender las relaciones entre ellas hacen que podamos optimizar la selección de variables identificando aquellas que son más relevantes y tienen un impacto significativo en la predicción de la variable objetivo. Además, nos permiten formular hipótesis informadas, basando el modelado en una comprensión profunda de cómo las variables interactúan y se influyen mutuamente.

### **Análisis multivariado**

El análisis multivariado es una etapa avanzada dentro del Análisis Exploratorio de Datos (EDA) que se enfoca en el examen simultáneo de más de dos variables. Este tipo de análisis permite entender las complejas relaciones e interacciones entre múltiples variables, proporcionando una visión integral y holística de los datos. Presenta los siguientes objetivos:

* **Identificación de Patrones y Grupos:** Podremos identificar patrones y subgrupos dentro de los datos, agrupando observaciones similares. Esto es útil para segmentar los datos y descubrir estructuras ocultas.
* **Modelado Predictivo:** Técnicas de regresión múltiple, análisis discriminante y modelos de Machine Learning permiten predecir la variable objetivo basándose en múltiples variables predictoras simultáneamente.
* **Detección y Manejo de Multicolinealidad:** La multicolinealidad ocurre cuando dos o más variables predictoras están altamente correlacionadas entre sí, lo que puede distorsionar los resultados de los modelos. El análisis multivariado ayuda a identificar y mitigar estos problemas, asegurando la fiabilidad de los modelos.

##### Análisis y visualización de las relaciones entre múltiples variables

###### Selección de características con el método SelectKBest

El análisis de selección de características evalúa la importancia de las variables independientes sobre la variable dependiente. Con el método SelectKBest se seleccionan las K características más importantes según un puntuaje estadístico, en este caso “f\_regression”, que basa la selección de características en su relación con la variable dependiente usando correlaciones.

Se representa la selección de las 12 mejores características a través de un diagrama de barras interactivo:

Gráfico

Descripción generada automáticamente



Pantalla de computadora con letras

Descripción generada automáticamente con confianza media

Observaciones:

* Las variables con mayor peso sobre la predicción son “User\_status\_Inactive”, seguida de “VIP\_group\_Otros”, “VIP\_credentials\_Vip” y “VIP\_group\_Ciber”.
* Estas características que presentan de forma notable un mayor peso que el resto se corresponden al bloque de “USERS”. Esto indica que **el bloque que recoge las características relacionadas con el usuario de la credencial es condicionante sobre el resto de cara a la predicción final sobre la criticidad de la credencial.**

###### Análisis discriminante de las características del bloque “Users”

Al conocer que las características del bloque de usuarios son las que presentan mayor influencia sobre la variable dependiente, realizaremos un análisis discriminante sobre ellas. Esta técnica es utilizada para clasificar observaciones en diferentes grupos basándose en múltiples variables predictoras.

A partir de los diagramas de barras de las tablas de contingencia generados durante el análisis bivariado, podemos observar en la de “*VIP\_group”* vs “*RISK”* que los usuarios con posiciones más altas obtienen un mayor grado de criticidad y los usuarios menos importantes obtienen un grado de criticidad medio-bajo.

A su vez, en la gráfica de “*User\_status”* vs “*RISK”* se identifica una clara influencia del estatus del usuario sobre el riesgo de la credencial, siendo todas las credenciales de alto o muy alto riesgo de usuarios activos, y todas las de riesgo bajo o muy bajo de usuarios inactivos.

Gráfico

Descripción generada automáticamenteGráfico

Descripción generada automáticamente

Para explicar por qué en los usuarios VIP de posiciones elevadas también aparecen credenciales de riesgo bajo estudiamos la relación entre “*VIP\_group”* y “*User\_status”.* De esta manera comprobamos que las credenciales de bajo riesgo que aparecían en los grupos VIP de alta importancia se corresponden con usuarios inactivos.

Gráfico, Gráfico de barras

Descripción generada automáticamente

Para visualizar el comportamiento de estas dos variables con respecto al grado de criticidad de la credencial se crea un diagrama de barras interactivo a través de *Dash*.

Gráfico

Descripción generada automáticamente



Observaciones:

* La variable “*User\_status”* tiene una mayor influencia sobre el riesgo final asignado a la credencial respecto a la variable “*VIP\_group”*, aunque esta tenga también una influencia significativa.
* Todas las muestras que aparezcan con medios y bajos grados de criticidad en los altos grupos de usuarios VIP se corresponden con usuarios inactivos.

###### Multicolinealidad entre variables independientes

Partiendo de la matriz de correlación que engloba todas las variables del dataset realizada en el análisis bivariado, creamos una matriz filtrada que muestre únicamente los índices menores que -0.5 o mayores que 0.5, para identificar más fácilmente las correlaciones elevadas entre variables.

Gráfico

Descripción generada automáticamente

Observaciones:

* Todos los pares de variables con mayor índice de relación se corresponden a variables del mismo bloque, no hay ningún par de distinto bloque que tengan un índice de correlación mayor a 0.5.
* **No existe correlación significativa entre pares de variables de distinto bloque**, es decir que cada bloque se comporta de forma independiente con respecto a la variable objetivo.
* Se aprecia **multicolinealidad entre variables del mismo bloque**, es decir, que las variables dentro de un mismo bloque miden aspectos diferentes del mismo concepto o fenómeno, lo que lleva a una alta correlación entre ellas.

Por otro lado, se realiza un análisis discriminante de cada bloque sobre la multicolinealidad de las variables independientes correlacionadas entre sí en cada uno de ellos, a partir de las matrices de correlación adjuntas en el [Anexo II](#_ANEXO_II). De estos análisis se obtienen las siguientes conclusiones:

**Bloque USERS**:

La variable binaria “*VIP\_credentials”* y la variable categórica “*VIP\_group\_Otros”* representan una **redundancia de variables** con una correlación negativa perfecta de -1. Esto se debe a que todos aquellos usuarios no VIP corresponderán a la categoría “*Otros*” de “*VIP\_group*”, es decir, vendrían a representar la misma característica. Sin embargo, la variable VIP\_credentials no se eliminará por esta redundancia ya que para otros estudios puede resultar relevante su distinción binaria entre usuarios VIP y no VIP.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

**Bloque PASSWORD**:

Presentan una alta correlación negativa las variables correspondientes a aquellos datos transformados con One-Hot Encoding, reflejando la estructura subyacente de la variable categórica original. Esto ocurre por ejemplo con las variables de “Password\_type”, de tal manera que cuando “Password\_type\_Personal password” tenga valor de 1, las otras dos alternativas siempre tendrán un valor de 0.

Presentan una alta correlación positiva las variables de *Password\_strength, Guesses\_discover* y *Password\_entropy,* ya que estos grupos representan características similares relacionadas con la complejidad y aleatoriedad de la contraseña.

Además, esta relación entre características se puede estudiar con un **análisis de regresión para visualizar la** **linealidad** que existe entre ellas. Lo haremos a través de un gráfico de dispersión de puntos, en el que además se representa la distinción de la variable dependiente sobre el grado de criticidad de cada muestra.

Gráfico, Gráfico en cascada

Descripción generada automáticamente

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

**Bloque CHANNEL:**

Presentan una alta correlación positiva las variables de “Reposts” vs “Subscribers” y las variables de “Channel\_privacity\_Public” vs “Engagement\_rate”. Esto sugiere que los canales con más suscriptores tienden a tener más reposts, lo cual puede ser indicativo de una audiencia más activa o comprometida, y por otro lado que hay una tendencia positiva moderada de que los canales públicos tienen mayor ratio de compromiso.

**Gráfico

Descripción generada automáticamente**

**Bloque FILE:**

Presentan una alta correlación negativa las variables correspondientes a aquellos datos transformados con One-Hot Encoding, al igual que vimos anteriormente. Ene ste caso, podemos observar este efecto en las variables de “Country\_file\_name”, de tal manera que cuando “Country\_file\_name\_Spain” tenga valor de 1, las otras dos alternativas siempre tendrán un valor de 0.

Gráfico

Descripción generada automáticamente

###### Análisis multidimensional de las variables del bloque “Password”

Este análisis representa la visualización de cómo algunas de las variables del bloque PASSWORD interaccionan entre sí y por lo tanto se refleja el funcionamiento interno de este bloque del dataset. Dentro de este estudio distinguimos de forma más precisa los siguientes análisis:

* **Análisis de distribución**: se muestra cómo se distribuyen los datos (contraseñas) de la variable categórica nominal “PASSWORD” según la puntuación de fortaleza asignada a cada una de ellas y recogida en la variable “*Password\_strength*”.
* **Análisis multivariado**: Se establece la relación simultánea entre las variables “Password\_strength”, “Password\_type” y “Password\_entropy”, pudiendo observar, para cada tipo de contraseña (personal, por defecto o de cambio) la distribución de contraseñas para cada puntuación de fortaleza, así como el valor de entropía que presenta cada una de ellas.
* **Análisis de Segmentación o Agrupamiento**: Se dividen los datos en grupos en función de las categorías de la variable “Password\_type”.

La representación se realiza mediante un **gráfico Sunburst**, una técnica de visualización que muestra la jerarquía y relaciones de datos categóricos en una forma radial, así como la **visualización de color-codificación**, usando distintos colores para representar la variable adicional en el gráfico, en este caso haciendo referencia al grado de criticidad, añadiendo una capa adicional de información a la visualización.

Gráfico, Gráfico de proyección solar

Descripción generada automáticamente

##### Conclusiones del análisis multivariado

Con este análisis hemos podido estudiar y observar el comportamiento de múltiples variables, tanto independientes como dependientes, de forma simultánea, permitiendo conocer cuáles tienen mayor correlación entre ellas, cuáles de las independientes afectan en mayor grado a la variable dependiente final y cómo se comportan las variables dentro de los bloques.

* **el bloque que recoge las características relacionadas con el usuario de la credencial es condicionante sobre el resto de cara a la predicción final sobre la criticidad de la credencial.**

### **Conclusiones generales del análisis exploratorio de datos**

A continuación, se presentan las conclusiones obtenidas tras un exhaustivo análisis exploratorio de datos realizado sobre el conjunto de datos utilizado para diseñar el procedimiento analítico cuyo objetivo es evaluar el grado de criticidad de las credenciales detectadas por el Bot de Telegram. El análisis exploratorio ha permitido identificar patrones significativos, relaciones clave y tendencias que son fundamentales para el desarrollo y la optimización del procedimiento analítico.

1. El dataset presenta una división de variables por bloques en el análisis de las muestras, lo cual se ve reflejado en la correlación entre variables independientes. Los bloques están claramente definidos:
   * El bloque USERS contiene aquellas características referentes a los usuarios como “VIP\_group” o “User\_status”.
   * El bloque PASSWORD contiene aquellas características referentes a las contraseñas como "Password\_strength” o “Password\_type”.
   * El bloque CHANNEL contiene aquellas características referentes a los canales por los que se difunden las credenciales en Telegram como “Channel\_name” o “Subscribers”.
   * El bloque FILE contiene aquellas características referentes a los archivos que contienen las credenciales detectadas por el Bot, como “File\_name” o “Country\_file\_name”.
2. Algunas de las variables independientes presentan multicolinealidad entre las variables dentro de su mismo bloque, pero nunca con las variables de otros bloques, lo cual reafirma la independencia de las variables de cada bloque.

Esta multicolinealidad permite comprender el comportamiento y la interacción entre las variables independientes dentro de cada bloque de características.

1. El bloque que contiene las características referentes a los usuarios presenta una mayor influencia sobre la variable objetivo que determina el grado de criticidad de la credencial en comparación con el resto de las variables independientes del dataset. Esto quiere decir que el usuario es determinante en el riesgo que presenta una credencial.

## **Valor añadido del procedimiento a la operativa diaria**

Este procedimiento analítico añade un valor al trabajo de un analista del servicio de Ciber Inteligencia que haga uso del bot de Telegram para analizar las credenciales detectadas, que se basa en los siguientes puntos fundamentales:

* **Identificación Eficiente de Alertas Críticas (sistema de triaje)**: Permite al analista detectar rápidamente las alertas más urgentes y establecer un orden de prioridad para su análisis, optimizando el uso de su tiempo y reduciendo el riesgo de pasar por alto credenciales altamente sensibles.
* **Objetividad y Consistencia**: Este procedimiento garantiza que todas las alertas se evalúen de manera uniforme, utilizando criterios estandarizados. Así se elimina la subjetividad y se asegura un análisis riguroso y objetivo. Además, se tiene una documentación detallada y trazable del análisis realizado, valiosa para auditorías y mejoras continuas.
* **Mejora de la Información de los Informes**: Proporciona alertas más concretas y priorizadas, mejorando la precisión y la calidad de los informes y recomendaciones. Esto fomenta una comunicación más completa hacia el cliente al recibir informes más relevantes y oportunos.

Este enfoque no solo optimiza el trabajo del analista de inteligencia, sino que también fortalece la capacidad de la organización para responder eficazmente a incidentes de seguridad.

## **ANEXO I**

**Gráfico, Gráfico de barras

Descripción generada automáticamente Gráfico, Gráfico en cascada

Descripción generada automáticamente**

Gráfico, Gráfico en cascada

Descripción generada automáticamente Gráfico, Gráfico de barras

Descripción generada automáticamente

Gráfico, Gráfico de barras

Descripción generada automáticamenteGráfico, Gráfico de barras

Descripción generada automáticamente

**Gráfico, Gráfico de barras

Descripción generada automáticamente**

## **ANEXO II**

Gráfico

Descripción generada automáticamente

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

**Gráfico

Descripción generada automáticamente**

**Gráfico, Gráfico en cascada

Descripción generada automáticamente**

**Gráfico

Descripción generada automáticamente**

**Gráfico

Descripción generada automáticamente**

**Gráfico, Gráfico de rectángulos

Descripción generada automáticamente**

**Gráfico

Descripción generada automáticamente**

1. HABILITACIÓN PERSONAL DE SEGURIDAD (HPS): <https://www.defensa.gob.es/Galerias/portalservicios/seginfoemp/Habilitacion_Personal_de_Seguridad.pdf> [↑](#footnote-ref-2)