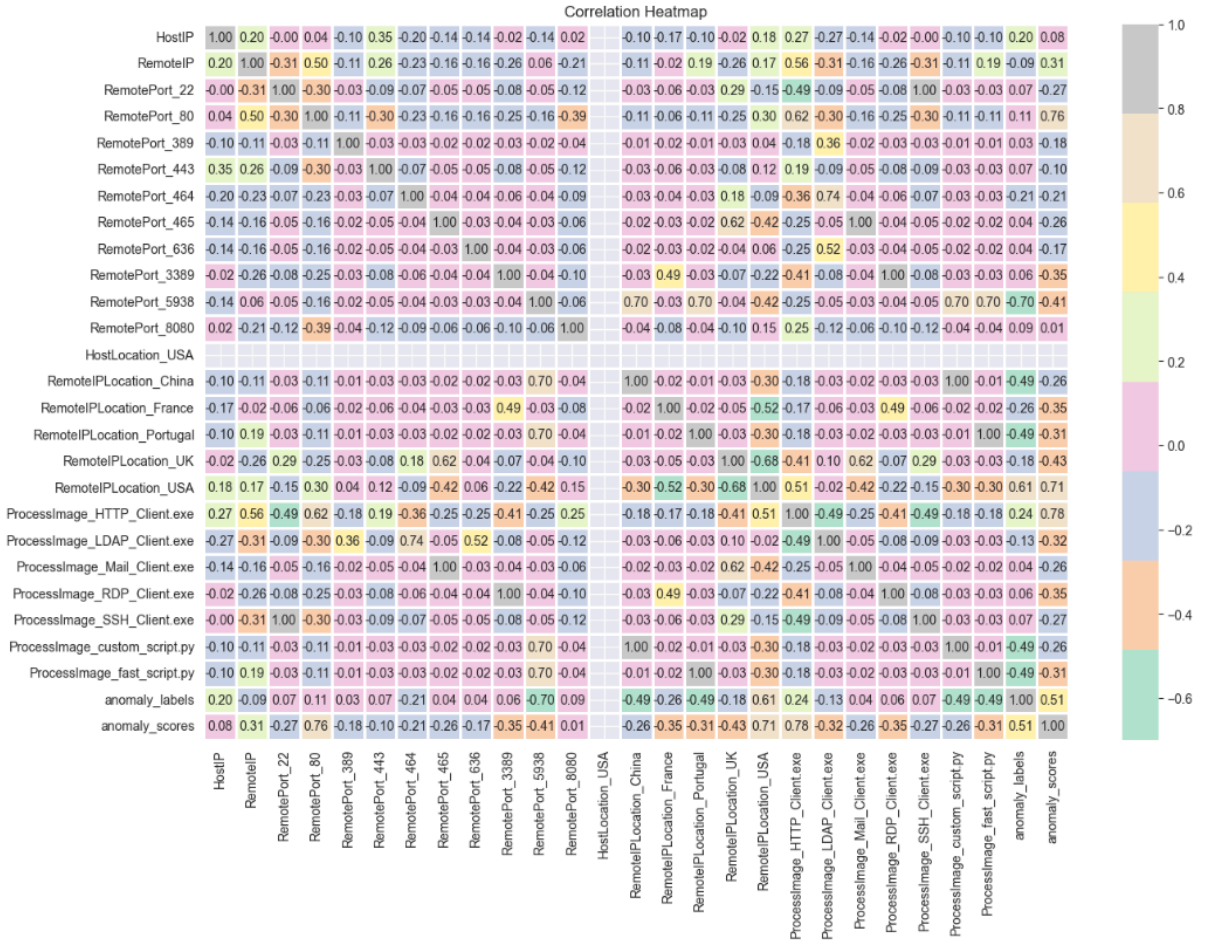
Datos, ciberseguridad e inteligencia artificial
Fecha: 25-06-2025.
Alumno: Taisen Romero Bañuelos.

Análisis del comportamiento de usuarios y entidades (UEBA)

A lo largo del reporte fui probando cosas extra para tener más opciones de trabajo para la actividad de mañana. Creo que lo más relevante de esta exploración más profunda de los datos se halla al final de mis notas. Al final retomé una matriz de correlación y unos gráficos de densidad que al inicio no me habían salido porque lo hice antes del one-hot encoding. Los resultados fueron interesantes, como se ve a continuación obtuve una matriz robusta con una lista amplia de variables con una correlación sumamente alta.



	Variable 1	Variable 2	Correlación
193	RemotePort_465	ProcessImage_Mail_Client.exe	1.000000
244	RemotePort_3389	ProcessImage_RDP_Client.exe	1.000000
70	RemotePort_22	ProcessImage_SSH_Client.exe	1.000000
321	RemoteIPLocation_China	ProcessImage_custom_script.py	1.000000
372	RemoteIPLocation_Portugal	ProcessImage_fast_script.py	1.000000
449	ProcessImage_HTTP_Client.exe	anomaly_scores	0.780591
99	RemotePort_80	anomaly_scores	0.757981
167	RemotePort_464	ProcessImage_LDAP_Client.exe	0.741057
424	RemoteIPLocation_USA	anomaly_scores	0.711019
261	RemotePort_5938	RemoteIPLocation_China	0.702673
271	RemotePort_5938	ProcessImage_custom_script.py	0.702673
263	RemotePort_5938	RemoteIPLocation_Portugal	0.702673
272	RemotePort_5938	ProcessImage_fast_script.py	0.702673
273	RemotePort_5938	anomaly_labels	0.698099
390	RemoteIPLocation_UK	RemoteIPLocation_USA	0.683451
91	RemotePort_80	ProcessImage_HTTP_Client.exe	0.621997
393	RemoteIPLocation_UK	ProcessImage_Mail_Client.exe	0.620331
189	RemotePort_465	RemoteIPLocation_UK	0.620331
423	RemoteIPLocation_USA	anomaly_labels	0.607315
41	RemoteIP	ProcessImage_HTTP_Client.exe	0.558009

Como vemos, la codificación One-hot fue fundamental para poder plotear el gráfico de densidad y la matriz de correlación.

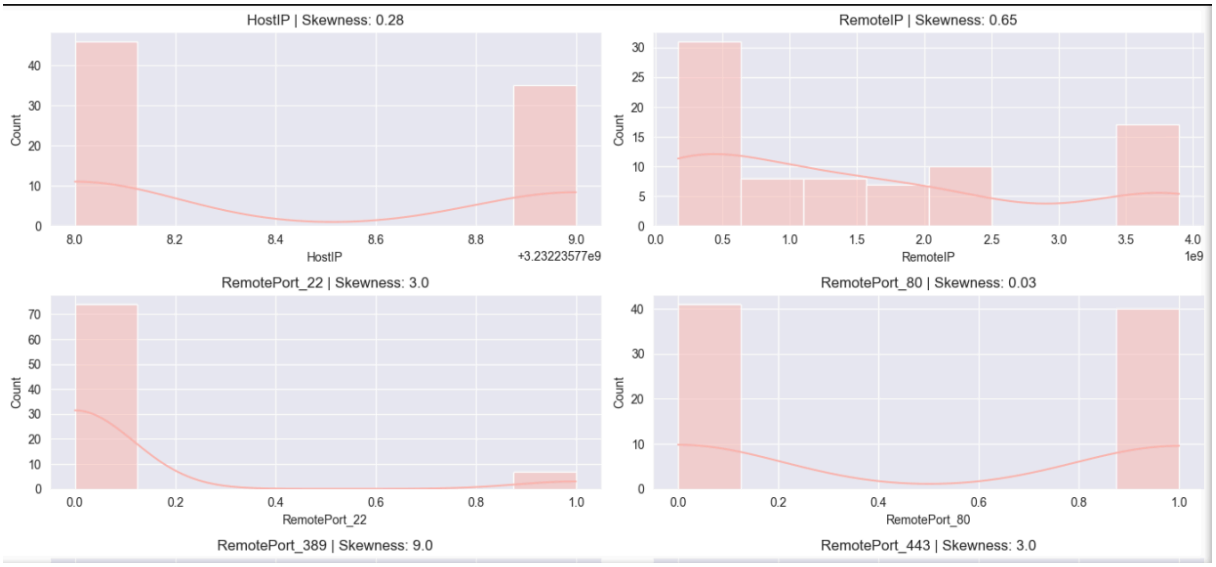
Podemos notar un efecto curioso en la matriz de correlación, y es que se hizo una especie de cruz al medio que separa en cuatro cuadrantes la matriz. Independientemente de eso, podemos observar correlaciones muy fuertes, por ejemplo, ProcessImage_custom_script.py y RemoteIPLocation_China tienen una correlación de 1.0, lo que significa que tienen una relación lineal perfecta positiva. En términos matemáticos, esto significa que no hay variación o desviación significativa en la relación entre ellas, por lo que cada vez que RemoteIPLocation_China es 1, también lo es ProcessImage_custom_script.py, y viceversa. Lo mismo sucede con otras variables:

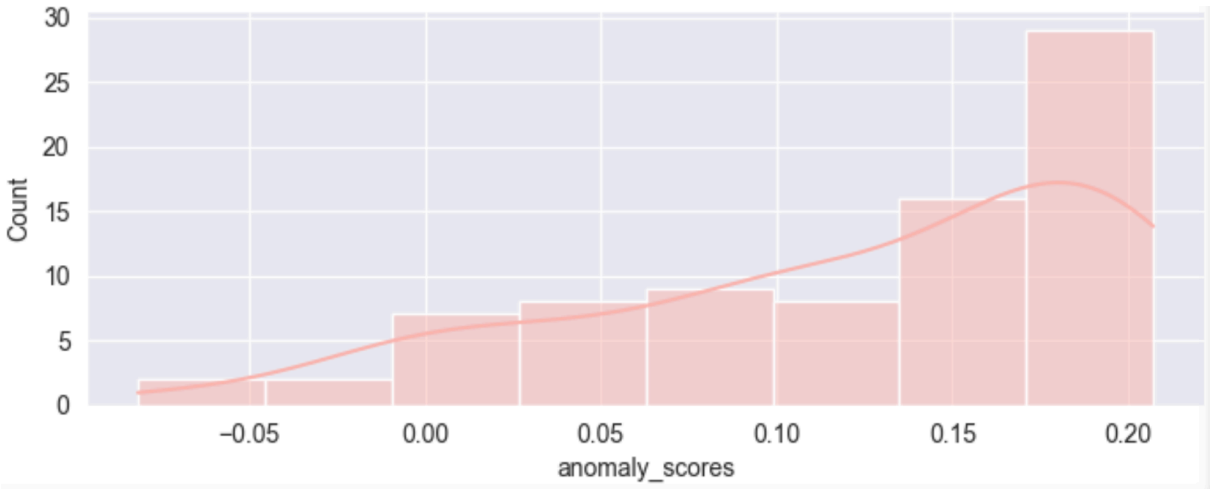
- ProcessImage_Mail_Client.exe y RemotePort_465
- ProcessImage_RDP_Client.exe y RemotePort_3389
- ProcessImage_SSH_Client.exe y RemotePort_22
- RemoteIPLocation_Portugal y ProcessImage_fast_script.py

Esto es interesante porque entonces una variable predice completamente la otra. Quizás a primera vista esto parezca ilusionante, pero antes debemos analizarlas con

detenimiento ya que alguna podría simplemente representar el comportamiento típico del tráfico de red. Por ejemplo, la correlación de RemotePort_465 y ProcessImage_Mail_Client.exe perfectamente puede significar que cada vez que se abre el puerto 465 (usado para SMTP seguro), se lanza el cliente de correo (comportamiento típico de aplicaciones de envío de correos). Algo similar ocurre con las otras variables de puertos. El uso de RDP (escritorio remoto) siempre está vinculado a la ejecución del cliente RDP. Y la conexión por SSH siempre implica ejecución del cliente SSH. El protocolo RDP puede ser interesante de cara a detectar sesiones remotas, sin embargo, hablemos de las correlaciones perfectas más interesantes. La correlación entre RemoteIPLocation_China y ProcessImage_custom_script.py puede implicar una ejecución automática de un script cuando la IP remota es de China, lo que puede ser señal de malware o comportamiento malicioso. Y lo mismo aplica para el caso de RemoteIPLocation_Portugal y ProcessImage_fast_script.py

Seguí con mi análisis de correlación entre variables en busca de posible uso de proxies o VPN (cosa que seguramente usaría un atacante), sin embargo, no supe encontrar algo que indicara el uso de estas herramientas. Aunque si noté que las variables relacionadas a HTTP (como la del puerto 80 y ProcessImage_HTTP_Client.exe) tienen una alta correlación con anomaly_scores, lo que sugiere que esta conexión tan fuerte con las anomalías y HTTP se deba a que exista tráfico web sospechoso.





Sobre el gráfico de densidad no hay mucho que decir, siento que como las variables categóricas se volvieron binarias (por el one-hot encoding) vemos distribuciones similares para casi todas las variables de su tipo. Tal vez podríamos mencionar que el gráfico de anomaly_score tiene una tendencia lineal ascendente, pero esto es normal porque según aumenta el conteo de observaciones naturalmente se irá acumulando la puntuación de anomalías.

P.D: En mis notas de notebook aparecen todos los gráficos de densidad que generé.