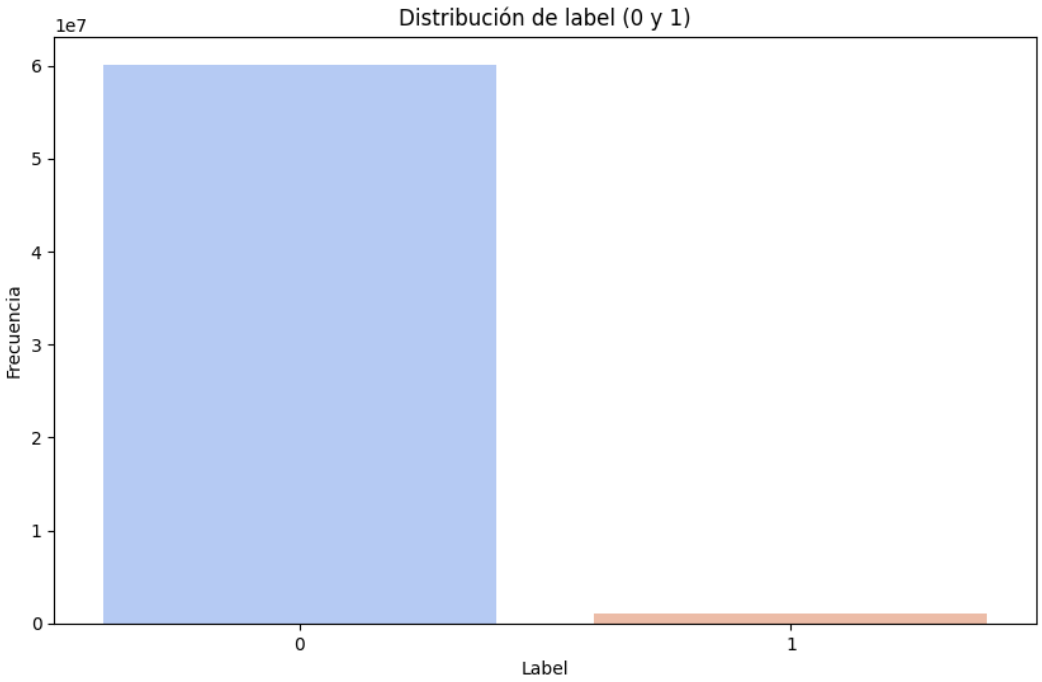
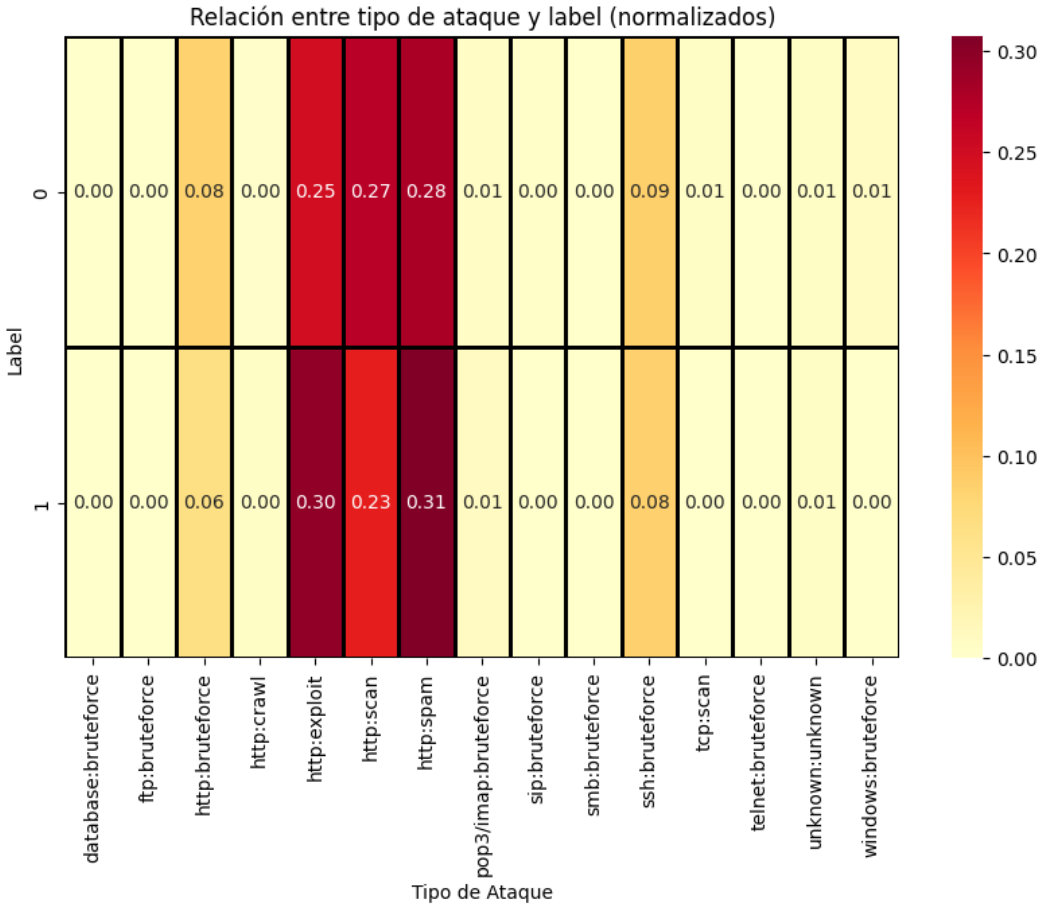


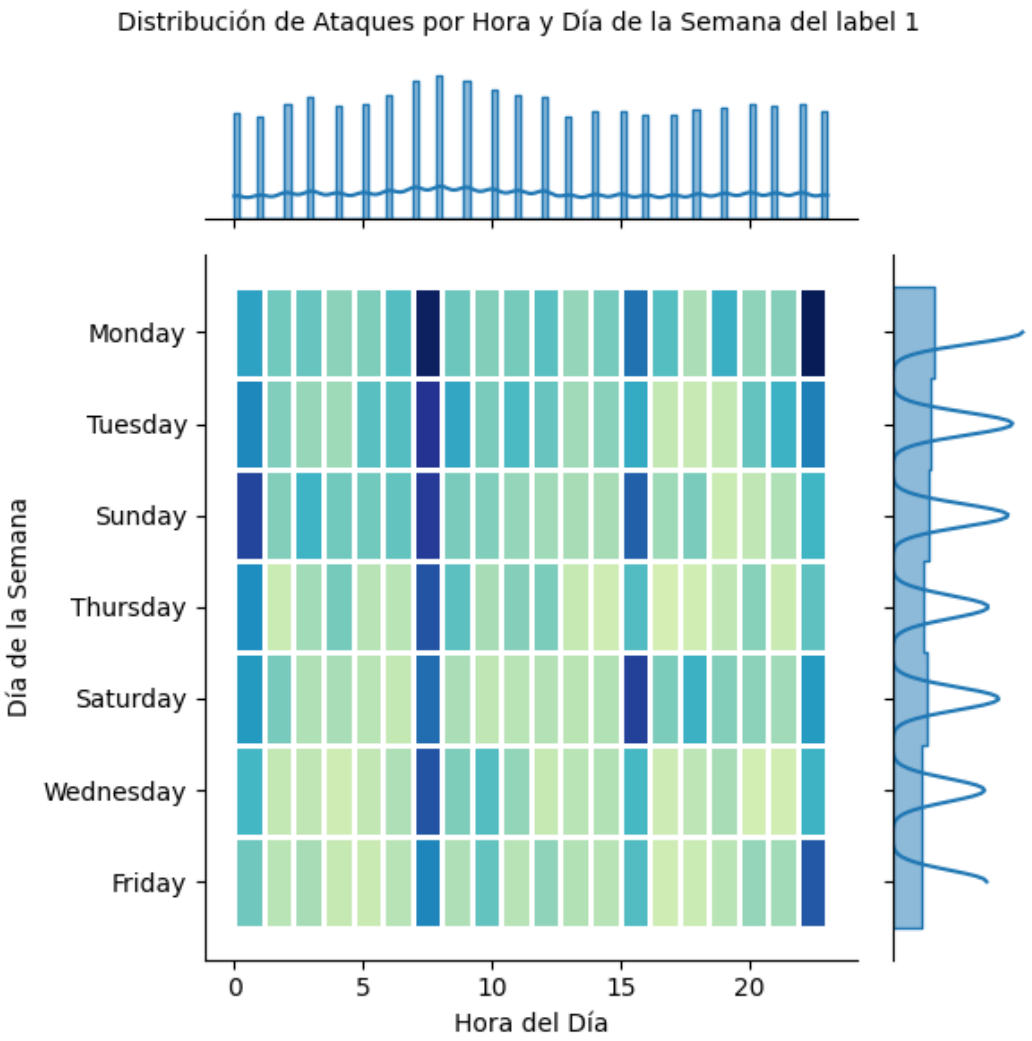
Link a colab/referencias

Este plot muestra el gran desbalanceo que contienen las clases del label. Puede dar un indicio de como maneja los datos debido al tarjet de la clase minoritaria.



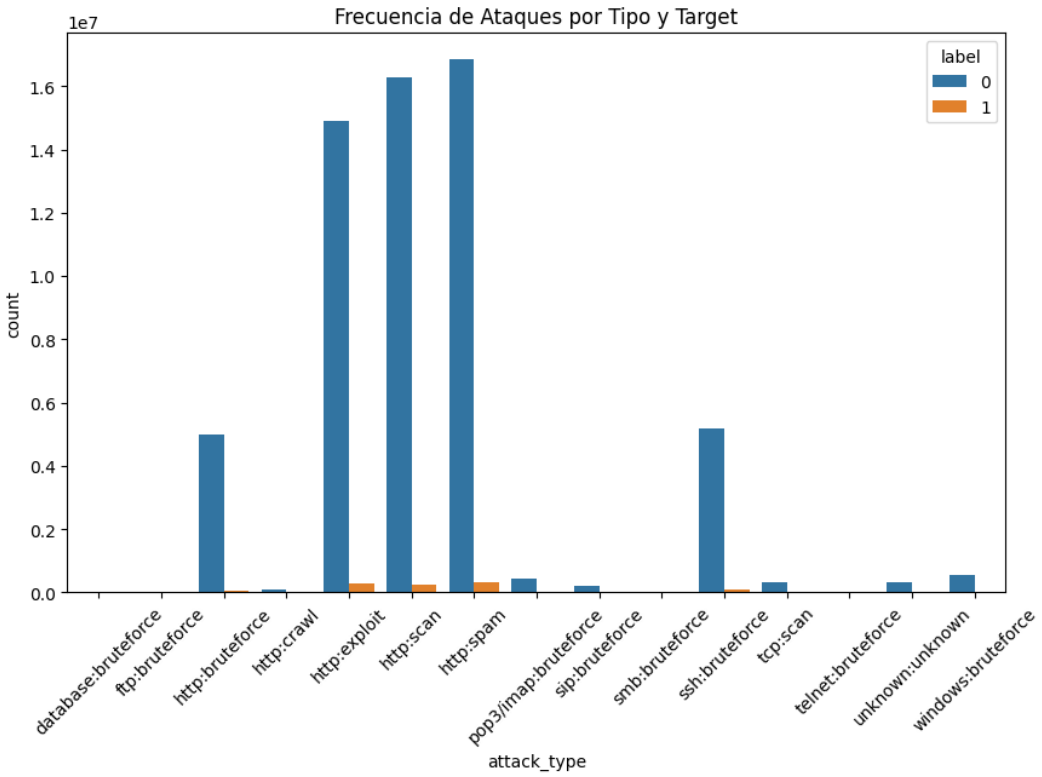
Link a colab/referencias

Es este plot se quizo buscar si hay algun diferencia porcentual en los ataques realizados por cada label (por eso son valores normalizados). Si bien la difrencia es minima, puede verse que existe.



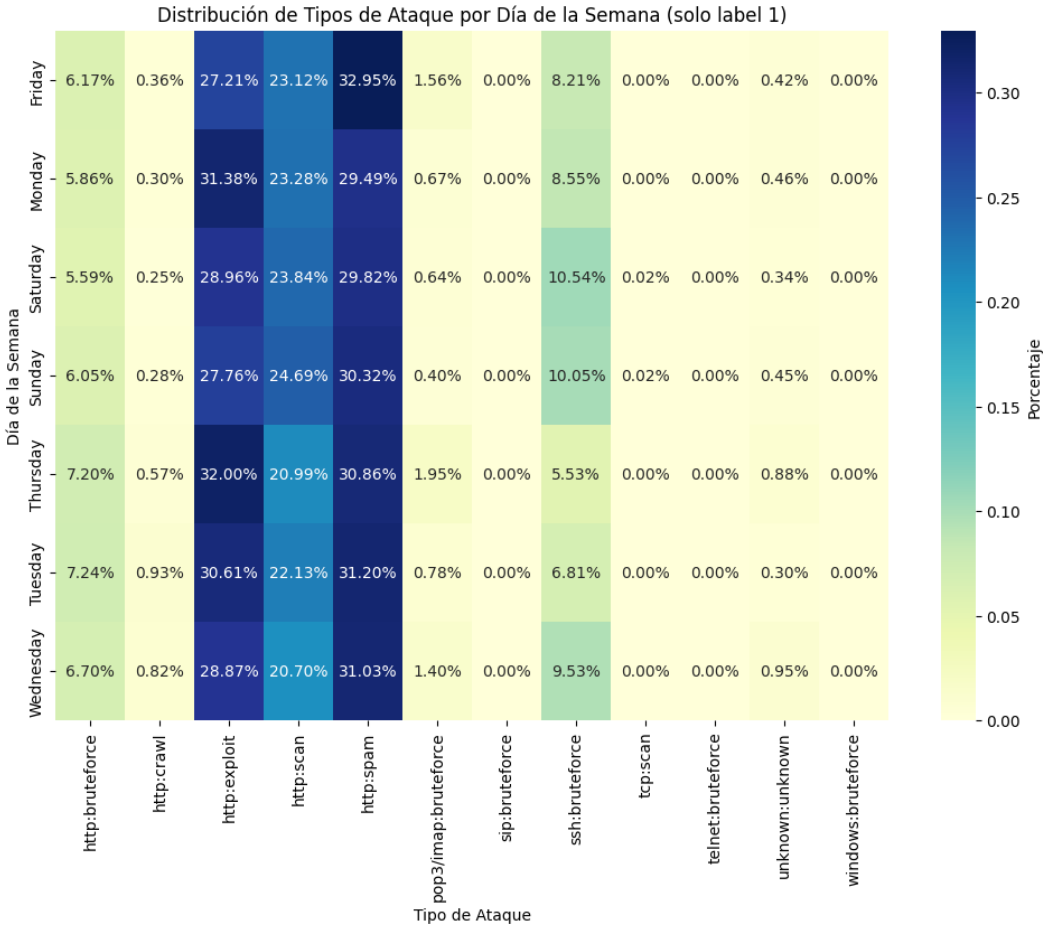
Link a colab/referencias

Se quizo buscar un relacion entre los horarios y dias de ataques (label 1). Puede verse que la mayor cantida de ataques son realizados en ciertos rangos horarios.



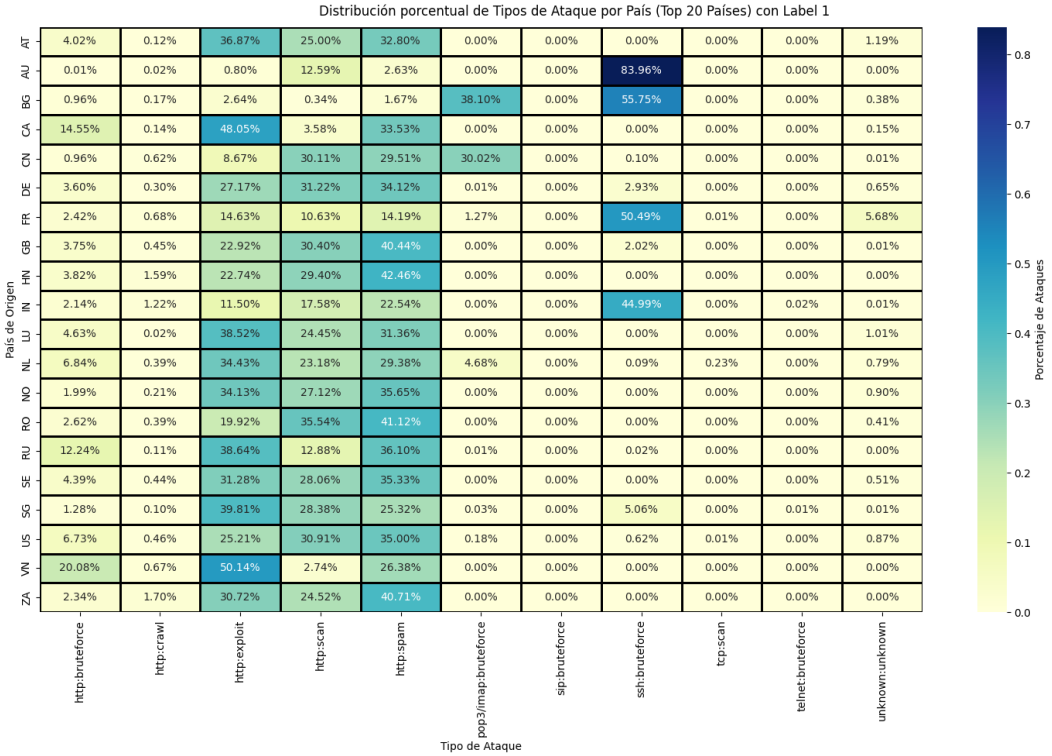
Link a colab/referencias

Se quizo mostrar si existe alguna entre los tipos de ataque y labels. Si bien este grafico sigue mostrando el gran desbalanceo que hay entre los labels, dentro del collab se encuentran ambos graficos por separado.



Link a colab/referencias

Esta plot muestra la diferencia minima que hay entre los porcentajes de cada ataque ejecutado diariamente.

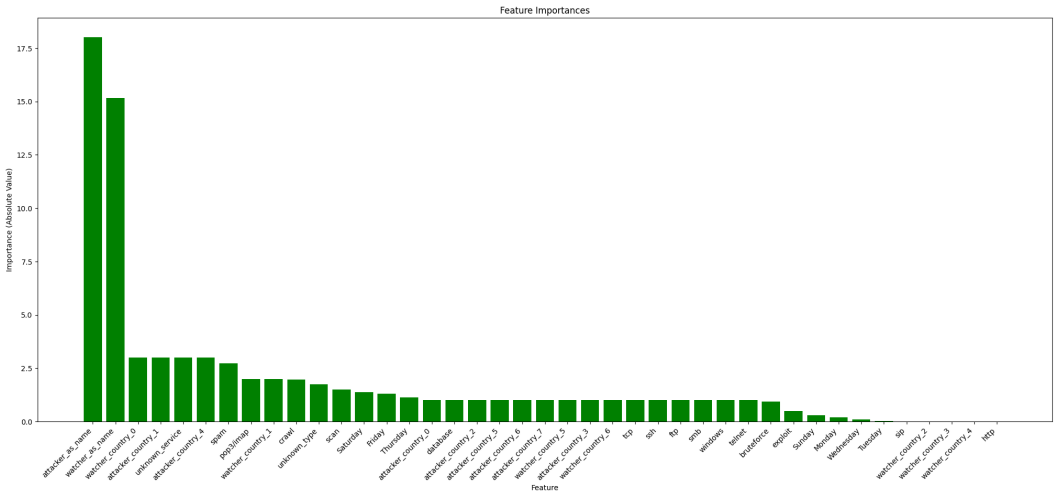


Link a colab/referencias

Este plot muestra en un top 20 de paises, el porcentaje de ataques que realizo cada uno (siempre referidos al label = 1).

Baseline

Link al colab o carpeta del Baseline



Link a colab/referencias

Se evidencia de manera notable el impacto significativo de las características attacker as name y watcher as name en este modelo. Este fenómeno podría explicarse por la tendencia de las empresas que ofrecen servicios de vigilancia (watcher as name) a tener áreas de vigilancia asignadas similares (attacker as name). Siguiendo estas dos características, se observan los valores más prominentes para watcher country y attacker country. Dado que estos están codificados con Binary Encoding, se puede inferir que representan los valores menos frecuentes para esos países. A continuación, encontramos la presencia de unknown service, que, según el diagrama de barras anterior, parece existir únicamente para la etiqueta 0. Finalmente, la característica spam destaca como aquella con mayor repetición dentro del conjunto de datos.

- F1 - Val : 0.5641434262948207
- F1 - Test : 0.53418
- Features:
  - Attack Service — OHE — Se realiza OHE y se agrupa la suma de los servicios de ataques por IP.
  - Attack type — OHE — Se realiza OHE ,se suma la cantidad de tipos de ataque y se normalizan para representar las proporciones de ataques por IP.
  - Watcher as name — Mean encoding — Como es unico por IP, no fue necesario hacer ninguna operacion extra
  - Attacker as name — Mean encoding — Como es unico por IP, no fue necesario hacer ninguna operacion extra
  - Days — OHE — Se realiza OHE ,se suman las cantidades de las columnas generadas y se normalizan para representar las proporciones de ataques por IP.
  - Attacker country — Binary encodingg — Como es unico por IP, no fue necesario hacer ninguna operacion extra
  - Watcher country — Binary encoding — Como es unico por IP, no fue necesario hacer ninguna operacion extra

Mejor Modelo

Link al colab o carpeta del Modelo

- Elegi hacer este modelo porque Random Forest puede manejar conjuntos de datos desequilibrados, lo que significa que puede trabajar bien cuando las clases objetivo no tienen el mismo número de muestras.
- Este modelo(RandomForest) es el mejor porque el F1 Val dio mas alto que el F1 val del modelo 2 (XGB)
- F1 - Val : 0.6044444444444445
- F1 - Test : 0.52986
- Features:
  - Attack Service(nueva feature) — OHE — Se realiza OHE y se agrupa la suma de los servicios de ataques por IP.
  - Attack type(nueva feature) — OHE — Se realiza OHE ,se suma la cantidad de tipos de ataque y se normalizan para representar las proporciones de ataques por IP.
  - Watcher as num — Mean encoding — Como es unico por IP, no fue necesario hacer ninguna operacion extra
  - Attacker as num — Mean encoding — Como es unico por IP, no fue necesario hacer ninguna operacion extra
  - Days(nueva feature) — OHE — Se realiza OHE ,se suman las cantidades de las columnas generadas y se normalizan para representar las proporciones de ataques por IP.
  - Attacker country — Binary encodingg — Como es unico por IP, no fue necesario hacer ninguna operacion extra
  - Watcher country — Binary encoding — Como es unico por IP, no fue necesario hacer ninguna operacion extra
  - combined countries(nueva feature) — Binary encoding — Se agrupa Attacker con Watcher country y Como es unico por IP, no fue necesario hacer ninguna operacion extra
  - hours(nueva feature) — OHE — Se realiza OHE en rangos de 2hs ,se suman las cantidades de las columnas generadas y se obtiene la cantidad de ataques por rangos de 2hs por IP.
  - ports features — OHE — Se realiza OHE creando nuevas features con los puertos más comunes VPN y no VPN y el resto, se los coloca en una nueva columna llamada open ports count, que contiene la cantidad de puertos abiertos que no estan en las nuevas features.
  - protocols features — OHE — Se realiza OHE y se obtienen las columnas protocol udp y tcp, y una extra que contiene la cantida de protocolos que utilizo una misma IP.

Segundo Modelo

Link al colab o carpeta del Modelo

- Elegi hacer este modelo porque XGBoost está optimizado para un rendimiento rápido y eficiente. Puede entrenar modelos de manera más rápida en comparación con otros algoritmos, especialmente en conjuntos de datos grandes.
- F1 - Val : 0.6026392961876833
- F1 - Test : 0.53584
- Features:
  - Attack Service(nueva feature) — OHE — Se realiza OHE y se agrupa la suma de los servicios de ataques por IP.
  - Attack type(nueva feature) — OHE — Se realiza OHE ,se suma la cantidad de tipos de ataque y se normalizan para representar las proporciones de ataques por IP.
  - Watcher as num — Mean encoding — Como es unico por IP, no fue necesario hacer ninguna operacion extra
  - Attacker as num — Mean encoding — Como es unico por IP, no fue necesario hacer ninguna operacion extra
  - Days(nueva feature) — OHE — Se realiza OHE ,se suman las cantidades de las columnas generadas y se normalizan para representar las proporciones de ataques por IP.
  - Attacker country — Binary encodingg — Como es unico por IP, no fue necesario hacer ninguna operacion extra
  - Watcher country — Binary encoding — Como es unico por IP, no fue necesario hacer ninguna operacion extra
  - combined countries(nueva feature) — Binary encoding — Se agrupa Attacker con Watcher country y Como es unico por IP, no fue necesario hacer ninguna operacion extra
  - ports features — OHE — Se realiza OHE creando nuevas features con los puertos más comunes VPN y no VPN y el resto, se los coloca en una nueva columna llamada open ports count, que contiene la cantidad de puertos abiertos que no estan en las nuevas features.
  - protocols features — OHE — Se realiza OHE y se obtienen las columnas protocol udp y tcp, y una extra que contiene la cantida de protocolos que utilizo una misma IP.