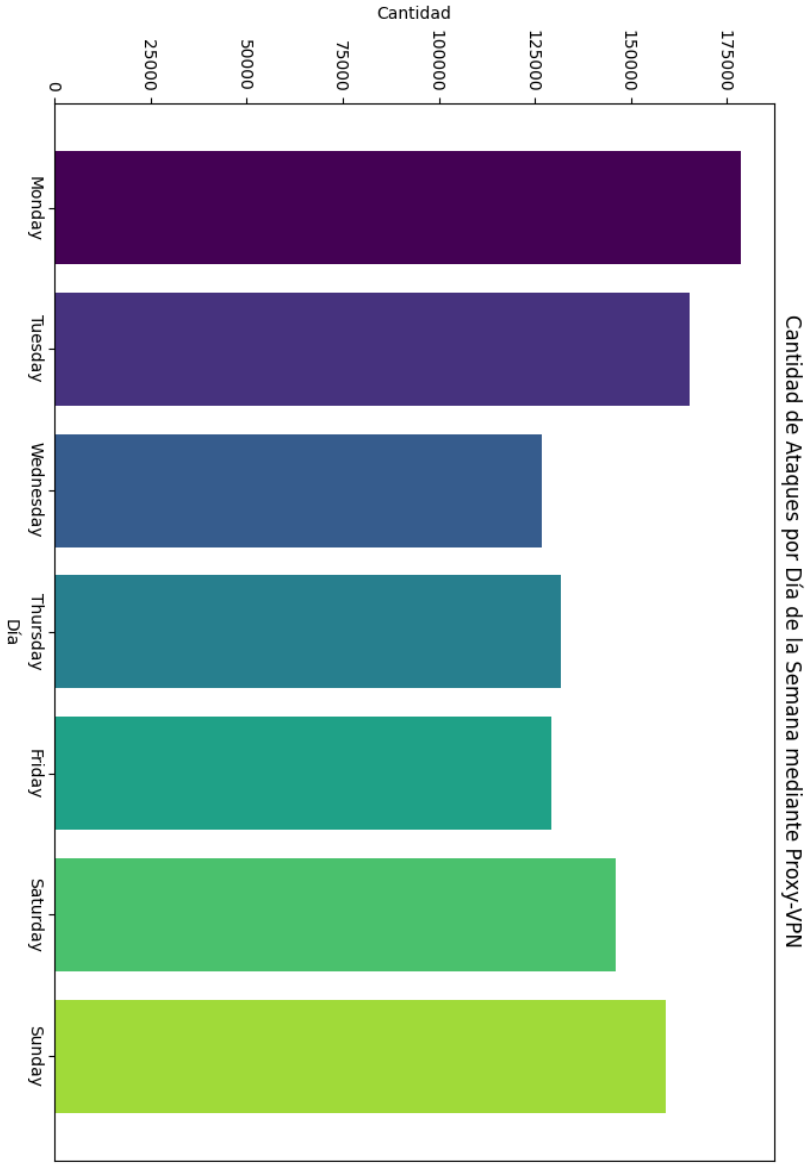
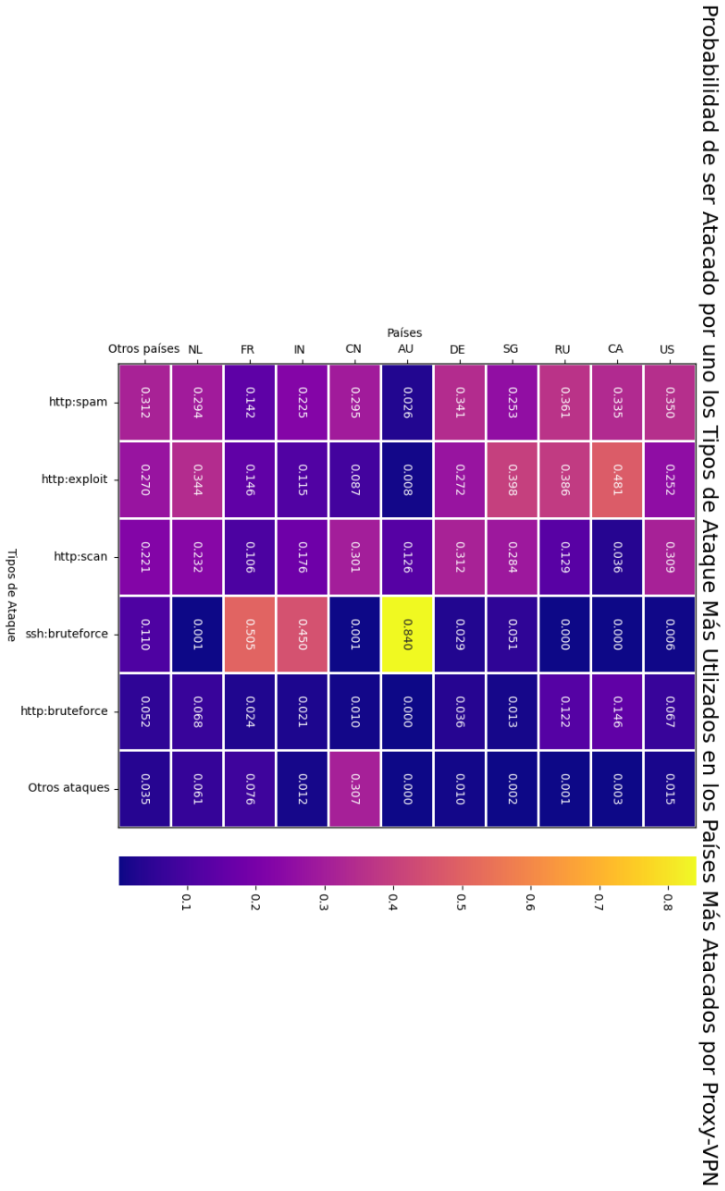


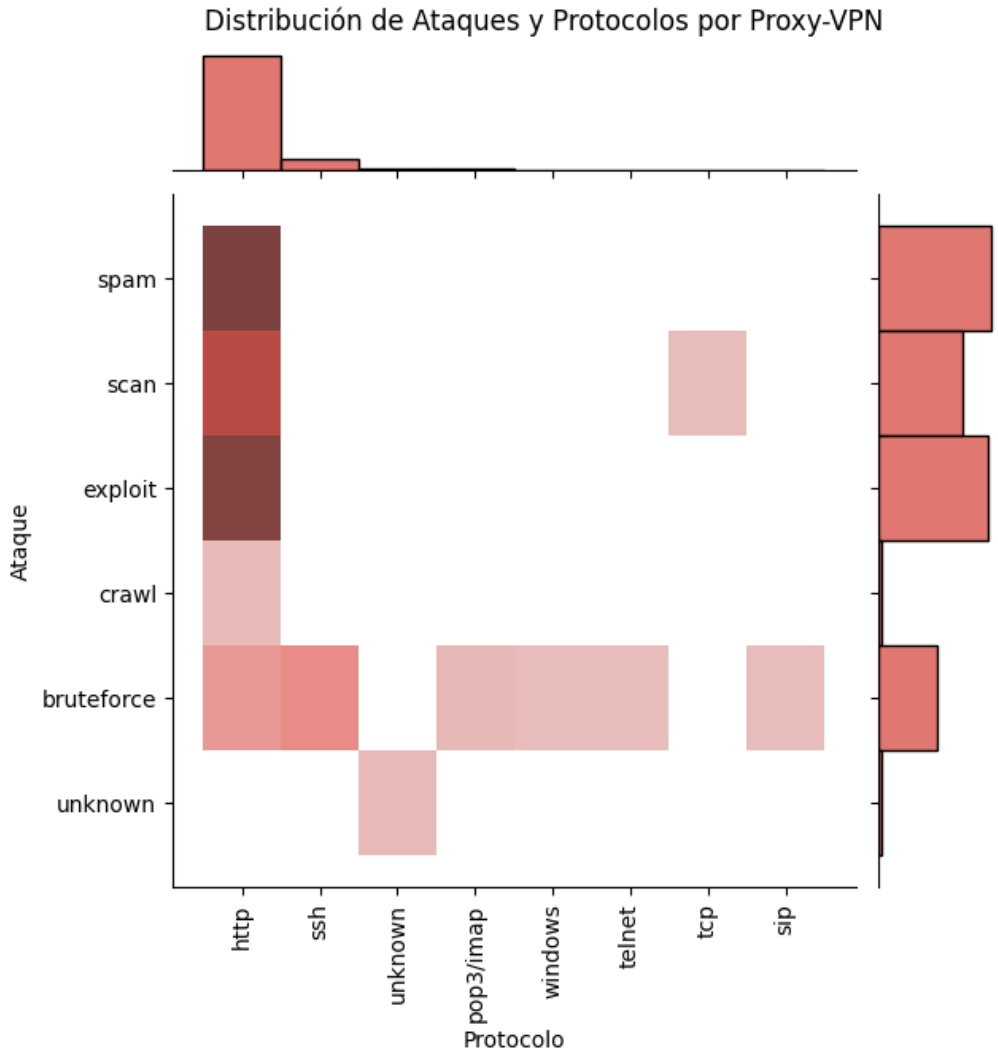

Link a colab/referencias

Los ataques de Proxy-VPN suceden más a menudo los Lunes, Martes y Domingos, y que el día de la semana es importante.



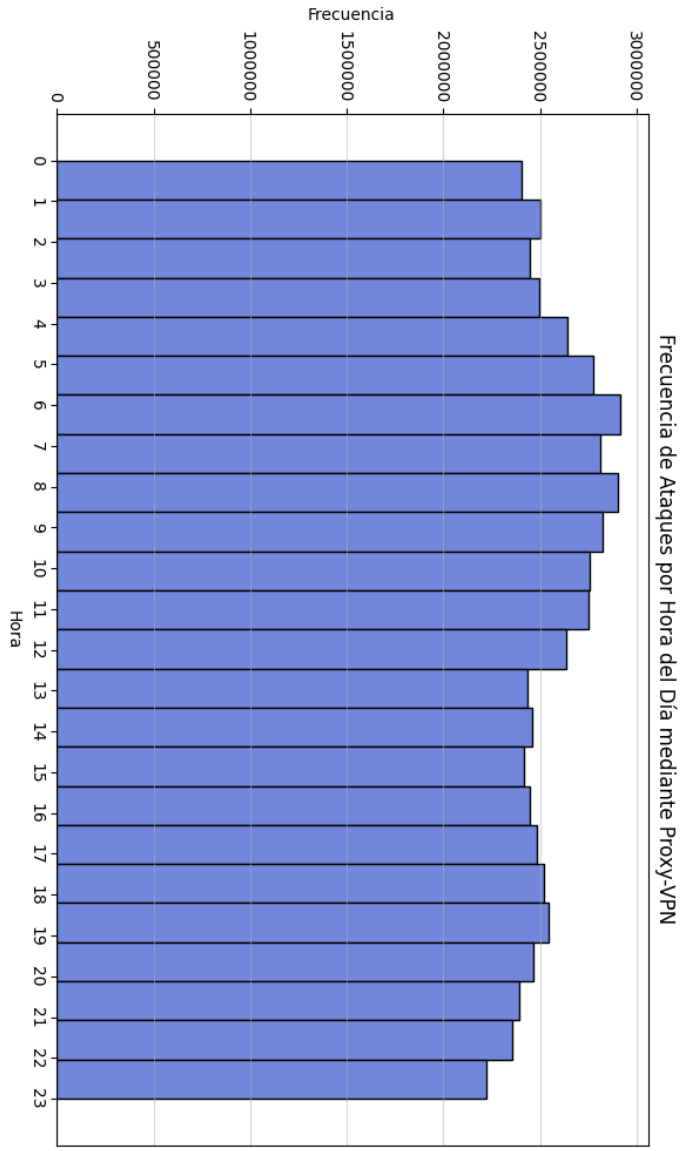

Link a colab/referencias

Determinados países son más probables a ser atacados por determinados ataques, usando un Proxy-VPN.



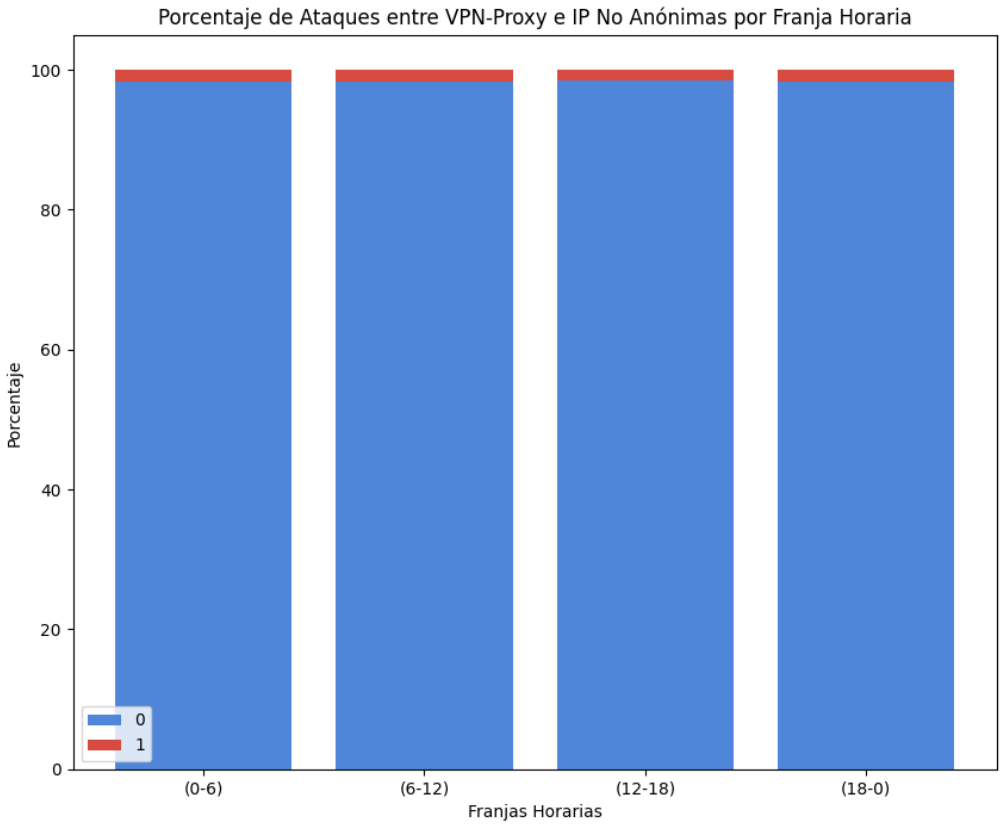

Link a colab/referencias

Hay más cantidad de ataques por spam, scan, exploit y bruteforce, junto a los protocolos http y ssh, mientras que los demás son despreciables.



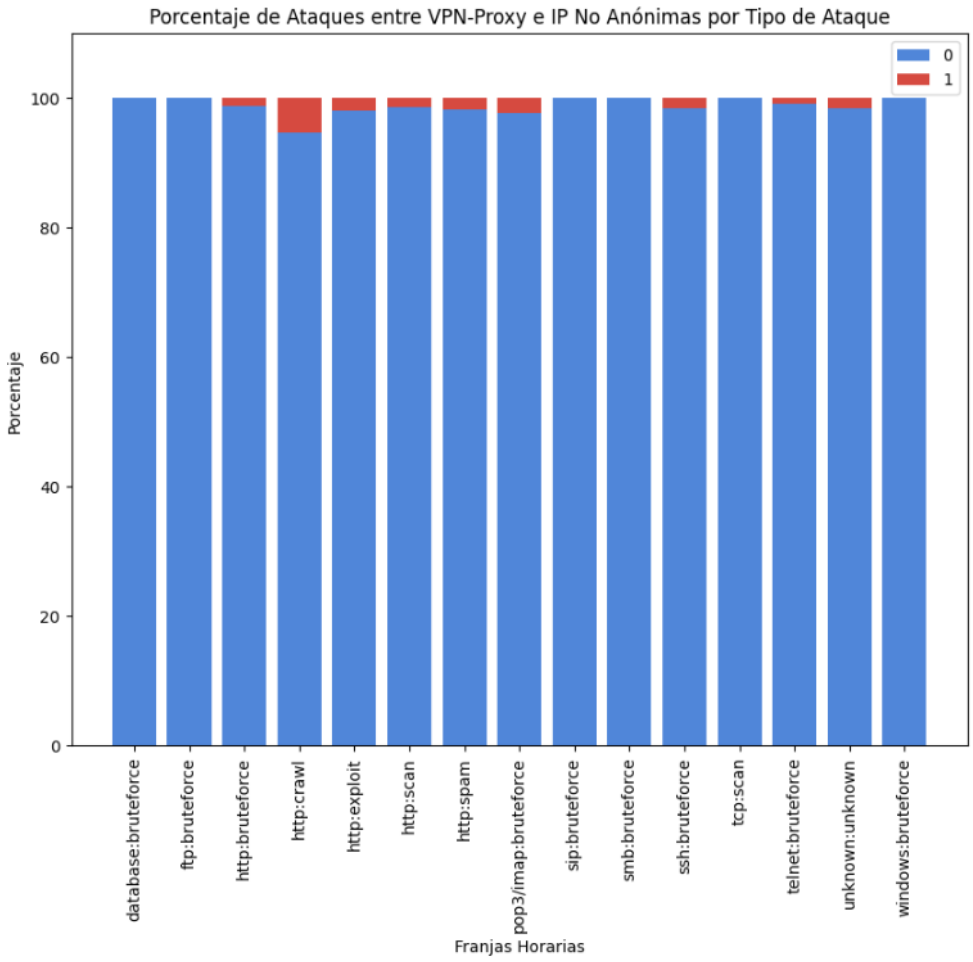

Link a colab/referencias

Hay una frecuencia elevada de ataques a ciertas horas, por lo que me sirve para crear dos nuevas features relacionadas al horario de ataque.




Link a colab/referencias

Los ataques son parejos en todos los rangos horarios y son despreciables en proporción.




Link a colab/referencias

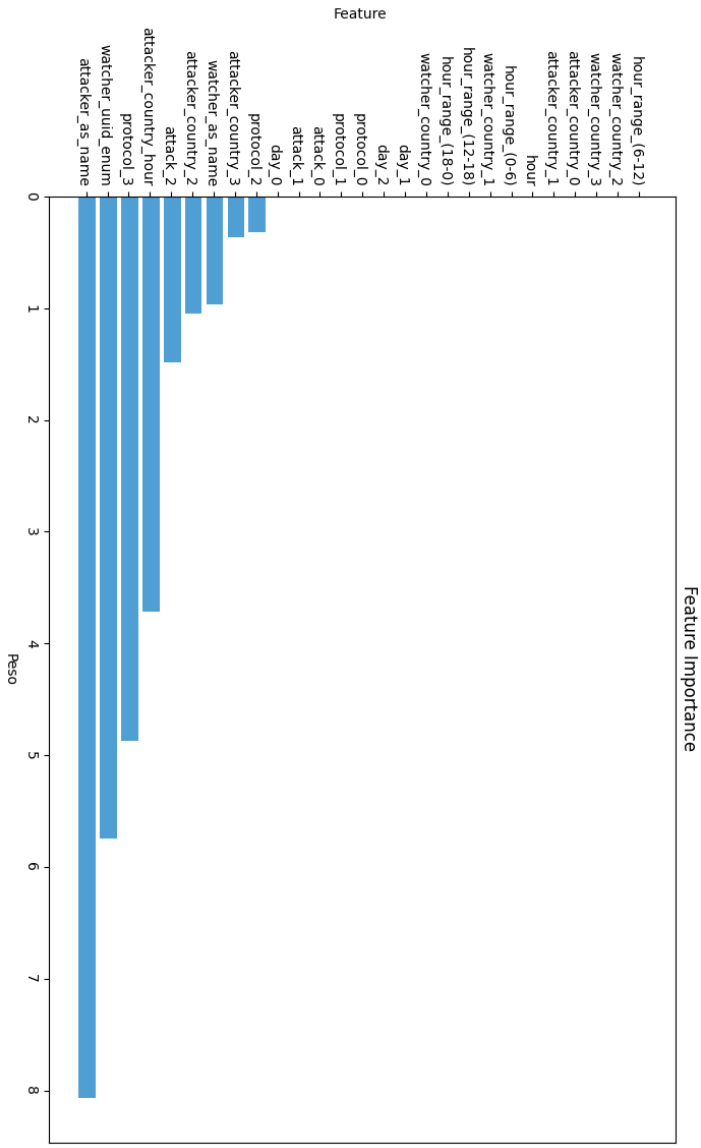
Los ataques son parejos en casi cada tipo de ataque y son despreciables en proporción.

# Visus Extra

Link a la carpeta con el resto de las visus

# Baseline

Link al colab o carpeta del Baseline




Link a colab/referencias

El Attacker AS Name y el Watcher UUID Enum son más importantes que las otras features, para este modelo al menos.

- F1 - Val : 0.30581533877194983
- F1 - Test : 0.4304
- Features:
  - Hour Range (Attack Time) — One Hot Encoding — Agrupé en intervalos de 6 horas
  - Day (Attack Time) — Binary Encoding — Nombre del día de la semana
  - Protocol (Attack Type) — Binary Encoding — Separé protocol:attack en protocolo
  - Attack (Attack Type) — Binary Encoding — Separé protocol:attack en ataque
  - Attacker Country Hour (Attacker Country, Hour) — One Hot Encoding — Agrupé el país del atacante junto a la hora del ataque

# Mejor Modelo (Random Forest)

Link al colab o carpeta del Modelo

- Elegí hacer este modelo porque: dada la información que obtuve del análisis exploratorio y baseline, donde pude concluir que hay datos más importante que otros y se tiene que cumplir ciertas condiciones para predecir si una IP es Proxy-VPN o no, lo que mejor se ajusta es un Decision Tree, entonces probé con eso (está en Modelos extra) y me dió buenos resultados. Como Random Forest usa múltiples Decision Tree para entrenar y promediar las predicciones obtenidas, decidí utilizarlo y me dio aún mejores resultados, por lo que me parece un mejor modelo.
- Este es el mejor modelo porque: además de que me dio mejor score en validación y test, y de lo explicado anteriormente, el Random Forest utiliza bagging, construyendo varios conjuntos de entrenamiento en un subconjunto de datos reducido utilizando los Decision Tree, lo que me garantiza para el dataset que tengo obtener un buen modelo para las predicciones. Además, las decisiones son colectivas entre los múltiples árboles para decidir si una IP es Proxy-VPN o no, ya que estos se promedian, y me evita el overfitting.
- F1 - Val : 0.7466193448518398
- F1 - Test : 0.55656
- Features:
  - Hour Range (Attack Time) — One Hot Encoding — Agrupé en intervalos de 6 horas
  - Day (Attack Time) — Binary Encoding — Nombre del día de la semana
  - Protocol (Attack Type) — Binary Encoding — Separé protocol:attack en protocolo
  - Attack (Attack Type) — Binary Encoding — Separé protocol:attack en ataque
  - Attacker Country Hour (Attacker Country, Hour) — One Hot Encoding — Agrupé el país del atacante junto a la hora del ataque

# Segundo Modelo (LightGBM)

Link al colab o carpeta del Modelo

- Elegí hacer este modelo porque: es más rápido y más eficiente manejando la memoria que otros modelos. Primero probé con XGBoost, ya que tiene parecido con LightGBM en cuanto a que ambos utilizan técnicas de Gradient Boosting, pero este último resultó ser superior para el conjunto de datos que tengo porque puede manejar conjuntos de datos más grandes y tiene mas precisión para las predicciones, debido a que ambos utilizan árboles para el entremiento de forma diferente.
- F1 - Val : 0.6046510563928006
- F1 - Test : 0.50857
- Features:
  - Hour Range (Attack Time) — One Hot Encoding — Agrupé en intervalos de 6 horas
  - Day (Attack Time) — Binary Encoding — Nombre del día de la semana
  - Protocol (Attack Type) — One Hot Encoding — Separé protocol:attack en protocolo
  - Attack (Attack Type) — One Hot Encoding — Separé protocol:attack en ataque
  - Attacker Country Hour (Attacker Country, Hour) — Mean Encoding — Agrupé el país del atacante junto a la hora del ataque

# Modelos extra

Perceptron con Mean Encoding
- Perceptron con Mean Encoding y Mejor Score
- Random Forest con Random Search a Batches
- Decision Tree con Random Search a Batches
- Decision Tree
- XGB