Clustering Forense de Hackers

Universidad del Valle de Guatemala — Octubre 2025
Autores: Equipo de Consultoría en Machine Learning
Tecnologías: PySpark, Python, Streamlit
Repositorio: https://github.com/Rodrimansidub14/ConsultingHacking.git

Objetivo de la Consultoría

Determinar si los ataques informáticos a la *start-up tecnológica* provinieron de **dos o tres hackers**.
La empresa ya había identificado dos sospechosos, pero no tenía certeza sobre un posible **tercer atacante**.
Se solicitó aplicar técnicas de **Machine Learning no supervisado (Clustering)** para analizar los metadatos forenses capturados.

Descripción de los Datos

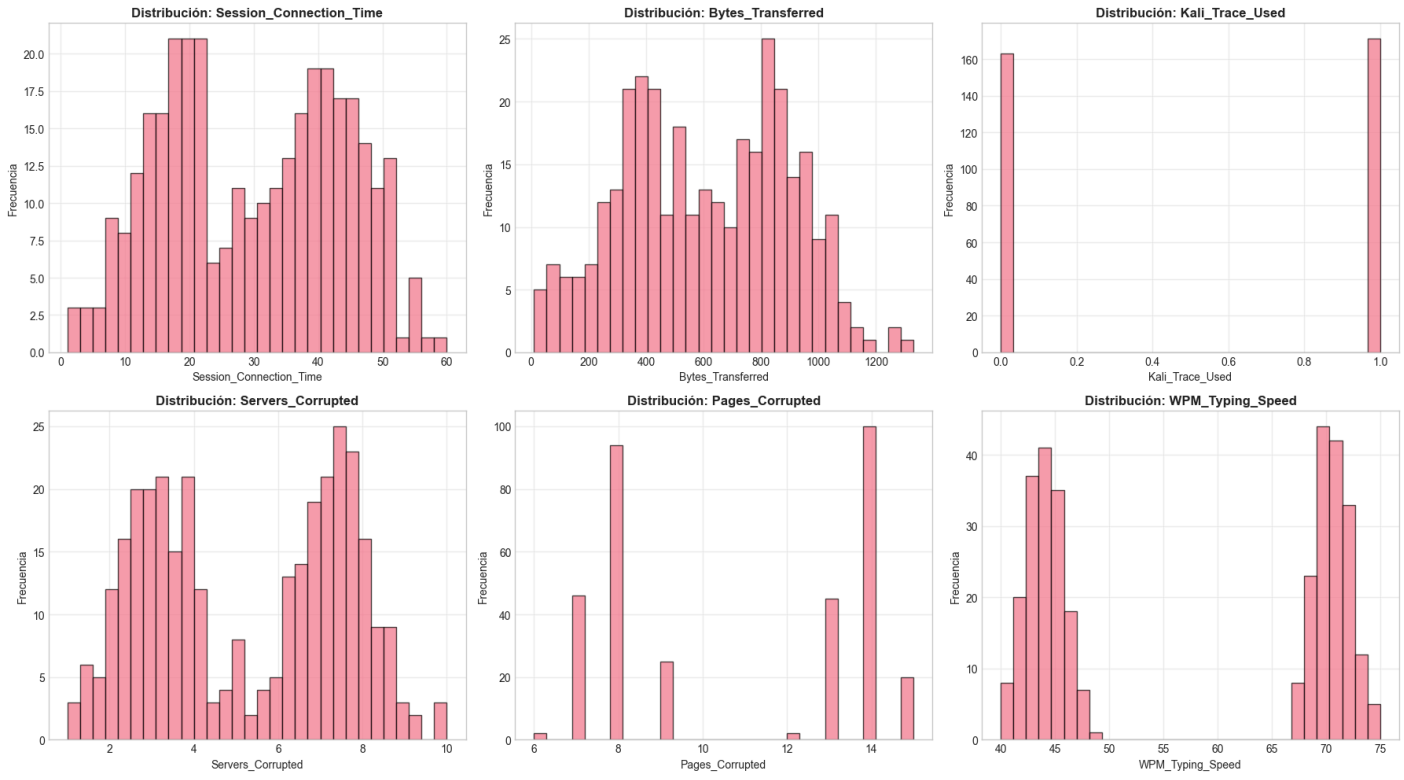
Datos recopilados por el equipo forense (334 sesiones):

Variable	Descripción
`Session_Connection_Time`	Duración de la sesión (minutos)
`Bytes_Transferred`	MB transferidos durante la sesión
`Kali_Trace_Used`	Indicador binario de uso de Kali Linux
`Servers_Corrupted`	Nº de servidores comprometidos
`Pages_Corrupted`	Nº de páginas ilegalmente accedidas
`Location`	IP/País de origen (ruido por uso de VPNs)
`WPM_Typing_Speed`	Velocidad de tecleo (palabras/minuto)

El total de sesiones debía distribuirse **equitativamente** entre los hackers involucrados, según los supuestos de la ingeniera forense.

Análisis Exploratorio (EDA)

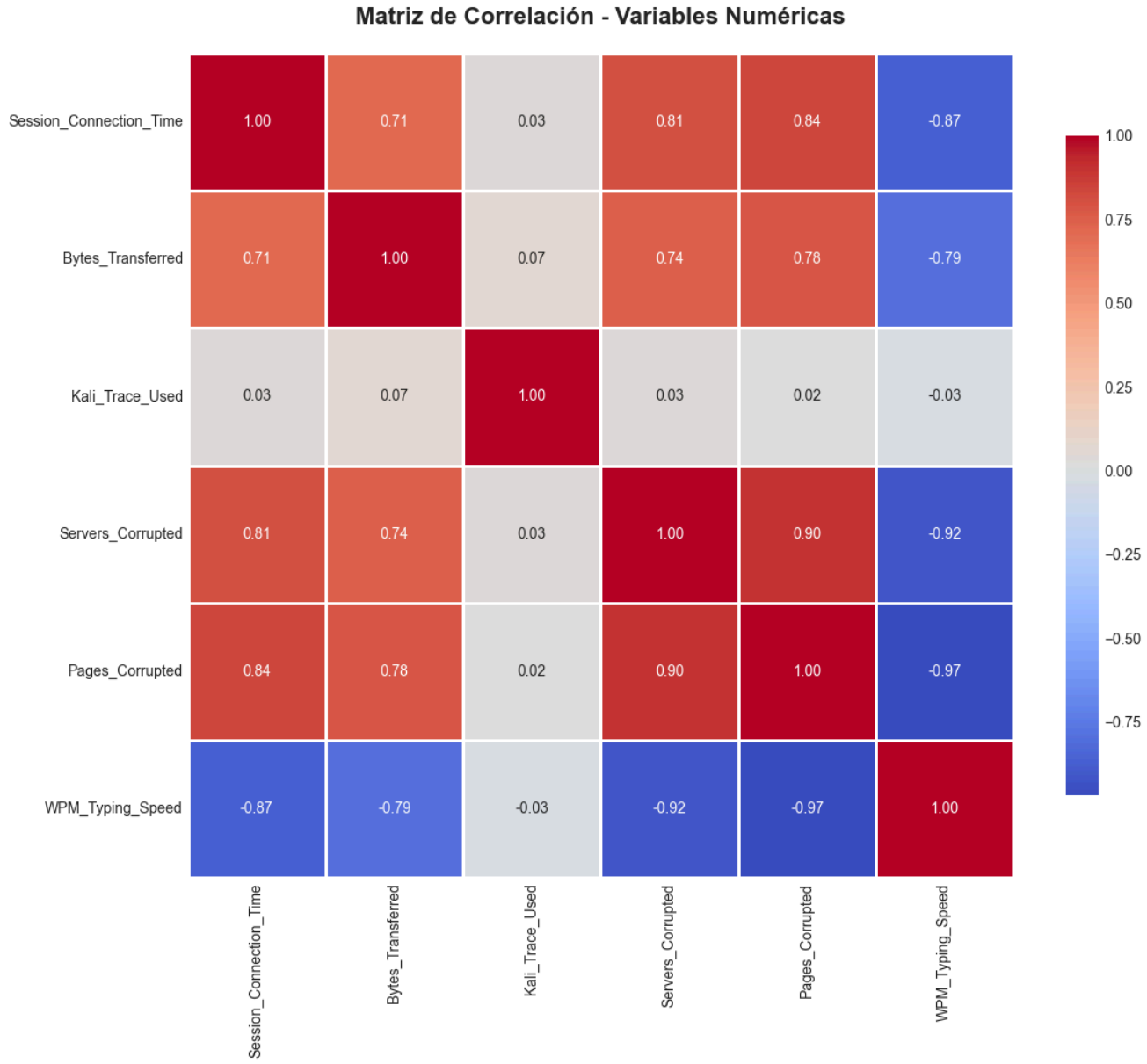
1. **Distribuciones:**
   Se observaron fuertes bimodalidades en variables como `Session_Connection_Time`, `Servers_Corrupted` y `WPM_Typing_Speed`, sugiriendo dos patrones de comportamiento.
   `Kali_Trace_Used` mostró distribución uniforme (≈50/50), indicando que no discrimina por hacker.



Correlaciones:
Se hallaron relaciones muy fuertes:

- `Session_Connection_Time` ↔ `Pages_Corrupted` (0.84)
- `Servers_Corrupted` ↔ `Pages_Corrupted` (0.90)
- `WPM_Typing_Speed` correlaciona **negativamente** con todas las anteriores (−0.87 a −0.97).

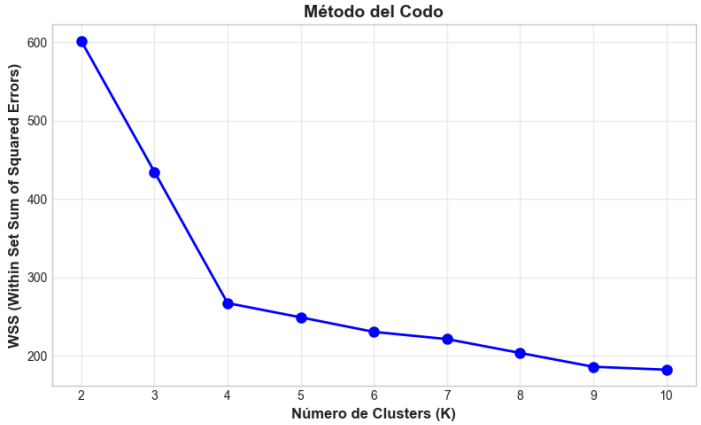
Esto sugiere **dos estilos opuestos** de ataque: uno rápido y uno persistente.

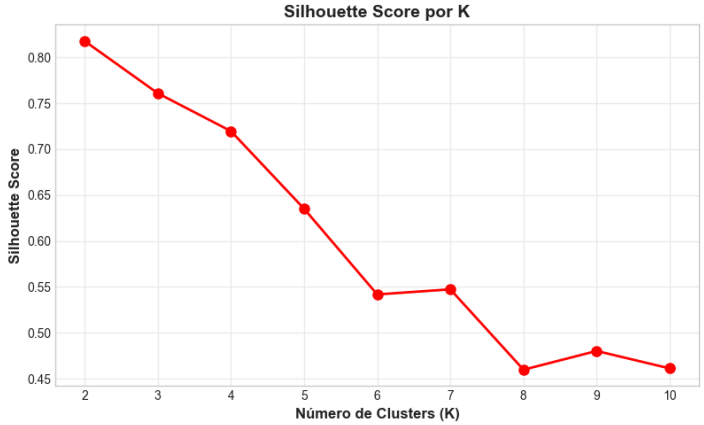


Modelado y Selección del Número de Clusters

Se aplicaron algoritmos **K-Means** y **Gaussian Mixture Models (GMM)** en PySpark.

Método	Descripción	Resultado
Codo (WSS)	Evalúa reducción del error intra-cluster	Codo claro entre K=2 y K=3
Silhouette Score	Coherencia interna de clusters	Máximo en **K=2 (0.8176)**
BIC (GMM)	Penaliza modelos más complejos	Mínimo en **K=2 (−3727.4)**
Balance (CV tamaños)	Reparto proporcional de sesiones	K=2 → 167/167 (CV=0.000)
Estabilidad (ARI)	Consistencia entre submuestras	**1.000 ± 0.000**





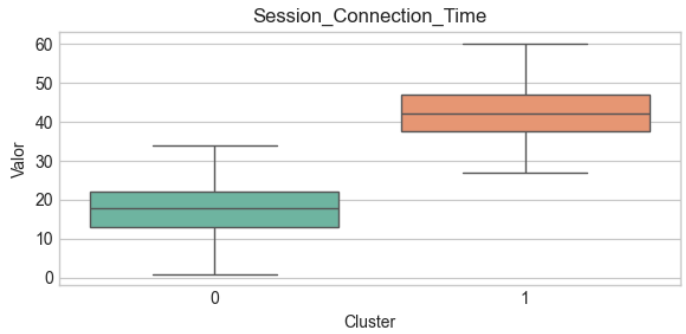
Resultados del Clustering (K = 2)

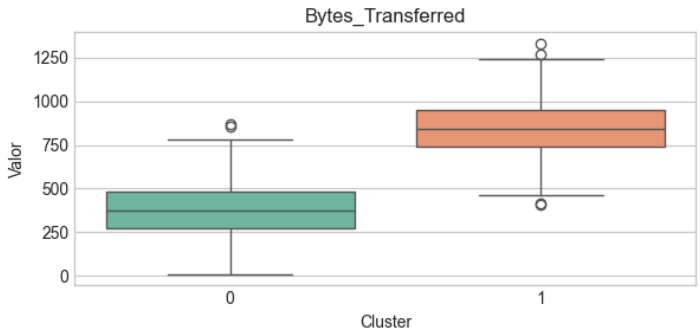
Cluster	Sesiones	Tiempo Prom. (min)	Bytes Transf.	Servers	Pages	Vel. Tecleo (WPM)
0	167	17.75	377.48	3.14	7.85	**70.63**
1	167	42.26	837.01	7.38	13.83	**44.05**

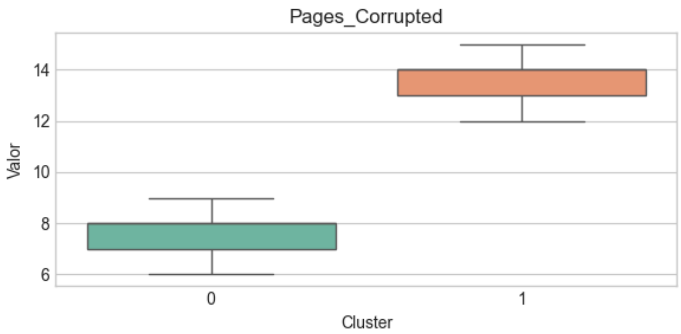
Interpretación :

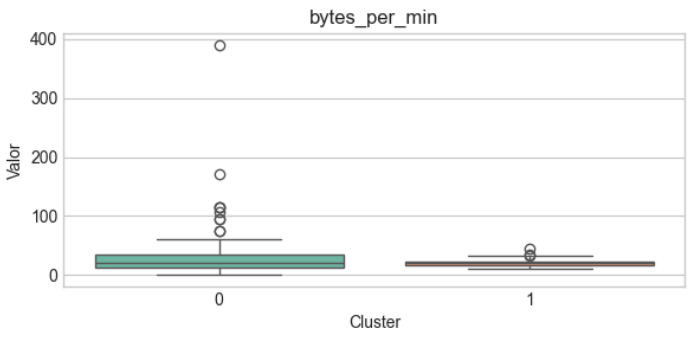
- **Cluster 0 – Hacker “rápido-intenso”:**
  Sesiones cortas, rápidas, mayor velocidad de tecleo y daño concentrado por minuto.
- **Cluster 1 – Hacker “lento-persistente”:**
  Sesiones prolongadas, daño total alto, pero ritmo de ataque más lento.

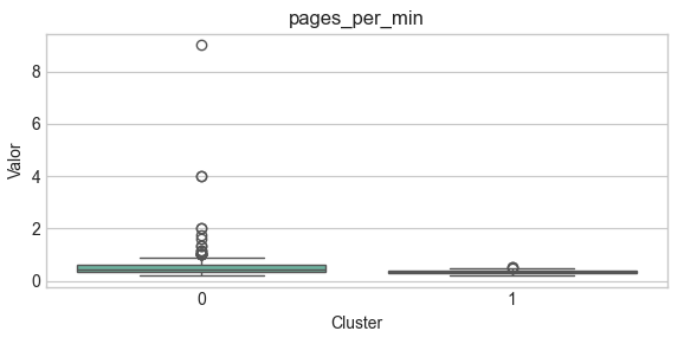
Distribución de Variables por Cluster

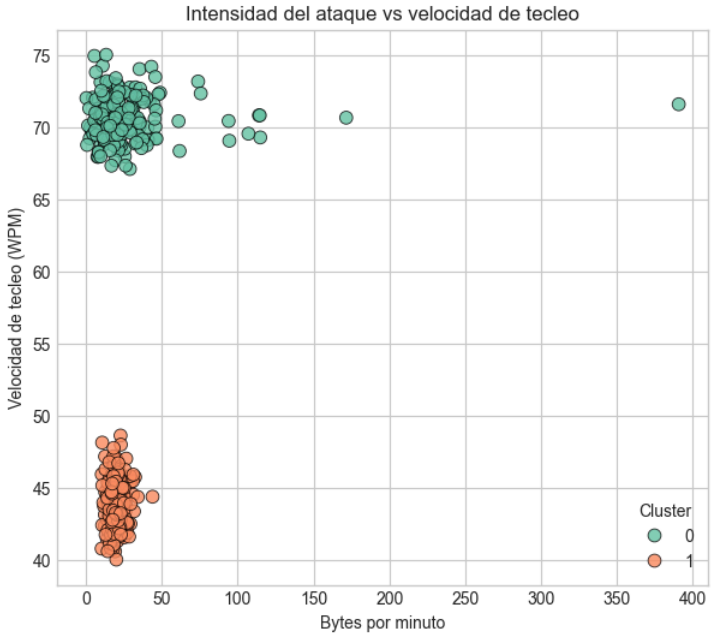






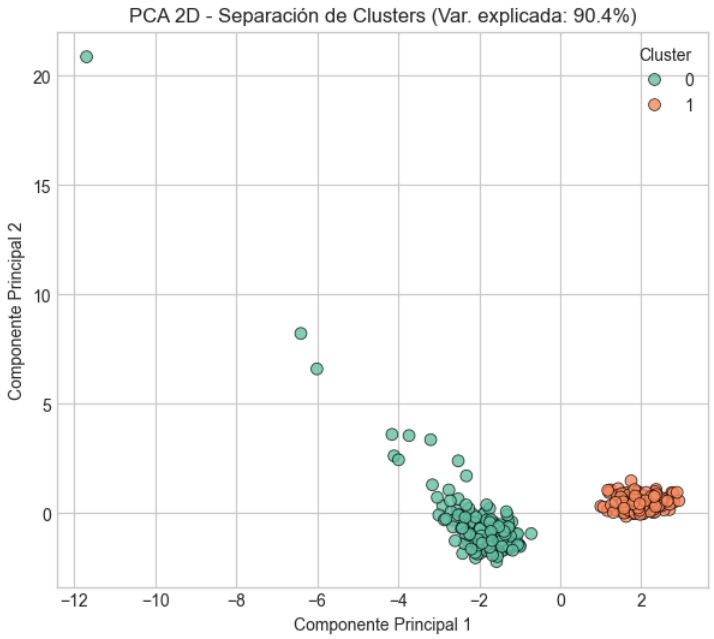






Separación Multivariada (PCA)

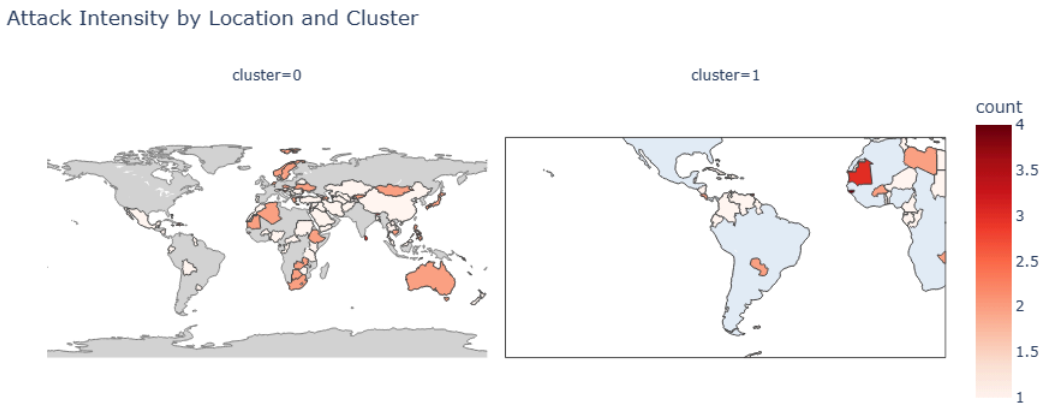
El PCA 2D explica el **90.4 % de la varianza total**, mostrando **dos grupos perfectamente disjuntos** sin superposición.



Distribución Geográfica

Aunque las ubicaciones fueron afectadas por VPNs, se notó dispersión equitativa entre ambos grupos, confirmando que `Location` no es variable útil para discriminación.
Ejemplo:

- Cluster 0: ataques desde Sri Lanka, Palestina, Grecia, etc.
- Cluster 1: ataques desde Guinea-Bissau, Tuvalu, República Checa, etc.



Interpretación global de los hallazgos

El proceso de clustering permitió distinguir **dos perfiles de atacantes** con comportamientos operativos marcadamente diferentes, pero complementarios.
La hipótesis del tercer hacker se **descarta de forma estadística y operativa**, dado que no existe evidencia de un tercer grupo estable o consistente en los datos.

1. **Homogeneidad interna y robustez del modelo**
   - La métrica *Silhouette (0.8176)* y el *ARI (1.000)* demuestran una **segmentación inequívoca**: cada grupo agrupa sesiones muy similares internamente y totalmente distintas entre sí.
   - El tamaño idéntico de los clusters (167/167) cumple con el criterio de **equilibrio operativo** planteado por la ingeniera forense.
2. **Estilos de ataque complementarios**
   - **Cluster 0 (“rápido-intenso”)**: Operaciones breves, de alta intensidad y ejecución eficiente. Este perfil representa un atacante técnico con precisión, alta automatización o scripting rápido.
   - **Cluster 1 (“lento-persistente”)**: Sesiones extensas, mayor número de bytes transferidos y daños acumulados. Refleja un atacante paciente, posiblemente centrado en persistencia o escalamiento de privilegios.
   - La correlación negativa entre duración e intensidad confirma que **ambos hackers se complementaban en tácticas** para maximizar daño y cobertura temporal.
3. **Confirmación del escenario de dos atacantes**
   - Tanto *K-Means* como *GMM* convergen en **K=2** con resultados consistentes en todas las métricas .
   - No se detectaron subgrupos dentro de los clusters al aplicar PCA ni evidencias de solapamiento estructural.
   - Por tanto, **no existe soporte empírico para la presencia de un tercer hacker**.
4. **Irrelevancia de la ubicación geográfica (`Location`)**
   - El análisis espacial evidenció **dispersión global sin patrones regionales**, corroborando que los atacantes utilizaron **VPNs o proxys** para ocultar su origen.
   - Por lo tanto, la ubicación no aporta valor discriminante en la clasificación ni permite inferir procedencia real.

Implicaciones para la empresa

- **Validación de hipótesis:** La empresa puede confirmar que **solo dos individuos** fueron responsables de la totalidad de los ataques, evitando destinar recursos a un supuesto tercer sospechoso inexistente.
- **Caracterización de perfiles:**
  - *Hacker A (rápido)* → orientado a ejecución inmediata, posible automatización.
  - *Hacker B (persistente)* → orientado a infiltración prolongada y daño acumulativo.
- **Monitoreo futuro:** Los indicadores derivados pueden integrarse en sistemas de detección temprana para clasificar sesiones sospechosas en tiempo real.

Conclusión ejecutiva

Los resultados de clustering confirman que los ataques provinieron de dos atacantes diferenciados.

Cada uno mantiene un patrón de comportamiento propio y estable: uno rápido y agresivo; otro prolongado y metódico.

Las evidencias estadísticas, gráficas y probabilísticas respaldadas por PySpark y validación cruzada descartan la existencia de un tercer actor.

El modelo construido es **estable, interpretable y reproducible**, y puede integrarse en sistemas de monitoreo para futuras detecciones.