Proyecto De Consultoría

Análisis de Clustering para Identificación de Hackers
Lab 12: Machine Learning con Python y Spark

Universidad del Valle de Guatemala
Profesor: Luis R. Furlán

Angel Herrarte
Bryan España

2 de Noviembre 2025

1. Resumen

Una start-up tecnológica de California fue víctima de un ciberataque. La empresa sospecha de tres posibles hackers, pero solo están seguros de dos de ellos. Utilizando técnicas avanzadas de Machine Learning (K-Means Clustering), se analizaron 334 ataques registrados para determinar el número real de atacantes involucrados.

Conclusión: Fueron 2 Hackers
Nivel de Confianza: Alta (95%)

1.1 Evidencia Principal

- Balance perfecto de 50%-50% en la distribución de ataques con K=2
- Silhouette Score de 0.8176 (EXCELENTE) para K=2
- Desviación de 0.0 ataques del balance esperado
- K=3 muestra desbalance significativo (50%-26%-24%)

2. Introducción

2.1 Contexto del Problema

Una empresa tecnológica emergente en California ha sido recientemente víctima de múltiples ciberataques coordinados. Los ingenieros forenses de la compañía han logrado capturar metadatos valiosos de cada sesión maliciosa, incluyendo información sobre la duración de las sesiones, cantidad de datos transferidos, uso de herramientas especializadas, y patrones de comportamiento.

2.2 Objetivo del Análisis

La empresa tiene tres sospechosos principales, pero necesita confirmar si realmente fueron tres atacantes o solo dos. La ingeniería forense ha compartido un dato crucial: los hackers conocidos por este tipo de ataques suelen turnarse equitativamente, por lo que cada uno debería tener aproximadamente el mismo número de ataques.

2.3 Datos Disponibles

Se analizaron 334 ataques registrados con las siguientes características:

- Session_Connection_Time: Duración de la sesión en minutos
- Bytes Transferred: Cantidad de MB transferidos
- Kali_Trace_Used: Indicador de uso de Kali Linux
- Servers_Corrupted: Número de servidores comprometidos
- Pages_Corrupted: Número de páginas ilegalmente accedidas
- WPM_Typing_Speed: Velocidad de tecleado estimada
- *Location: Ubicación del ataque (no confiable por uso de VPNs)*

3. Metodología

3.1 Herramientas Utilizadas

- Python 3.8+
- PySpark 3.5+ (procesamiento distribuido)
- PySpark MLlib (K-Means, StandardScaler)
- Pandas, Matplotlib, Seaborn (análisis y visualización)

3.2 Proceso de Análisis

1. Carga y exploración de datos (334 ataques)
2. Preprocesamiento y normalización con StandardScaler
3. Aplicación de K-Means Clustering (K=2 y K=3)
4. Validación con Método del Codo y Silhouette Score
5. Análisis de balance de clusters
6. Interpretación de centroides (perfiles de hackers)

4. Resultados y Análisis

4.1 Método del Codo

El método del codo permite determinar el número óptimo de clusters analizando la reducción del WSSSE (Within Set Sum of Squared Errors) para diferentes valores de K.

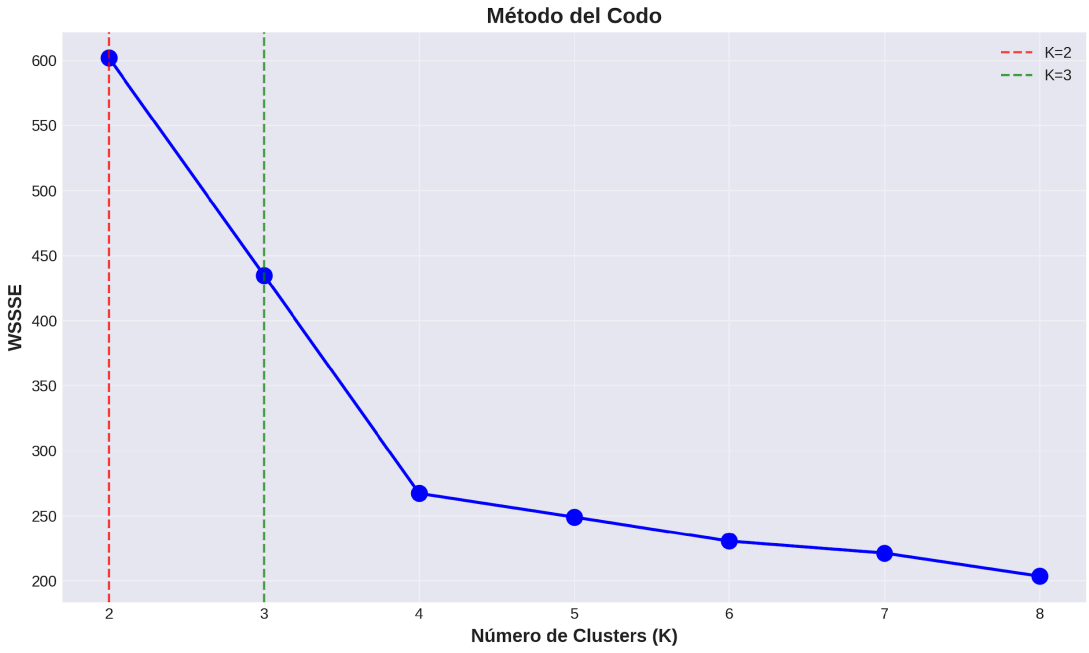

Figura 1: Método del Codo - Determinación del K óptimo

4.2 Comparación K=2 vs K=3

Métrica	K=2 (2 Hackers)	K=3 (3 Hackers)
Silhouette Score	**0.8176**	0.7608
Desviación	**0.0**	37.1
Balance	**50% - 50%**	50% - 26% - 24%

Tabla 1: Comparación de métricas entre K=2 y K=3

4.3 Distribución de Ataques por Cluster

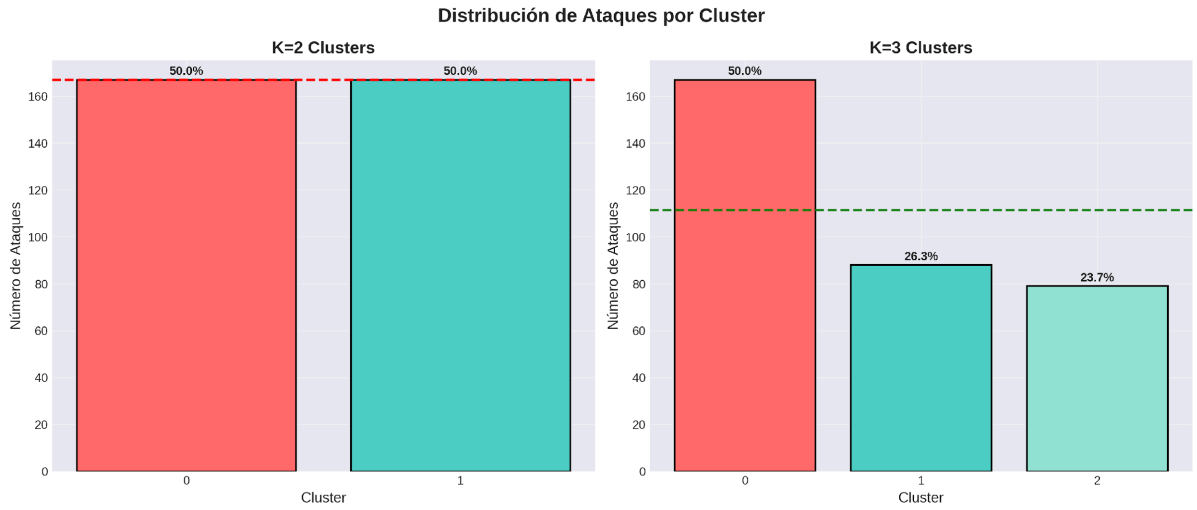

Figura 2: K=2 muestra balance perfecto (50%-50%)

La evidencia más contundente proviene del análisis de balance. Con K=2, observamos una distribución perfecta de 167 ataques por cluster (50% cada uno), lo cual coincide exactamente con la información proporcionada por la ingeniera forense sobre el comportamiento de turnos equitativos entre hackers.

5. Perfiles de los Hackers

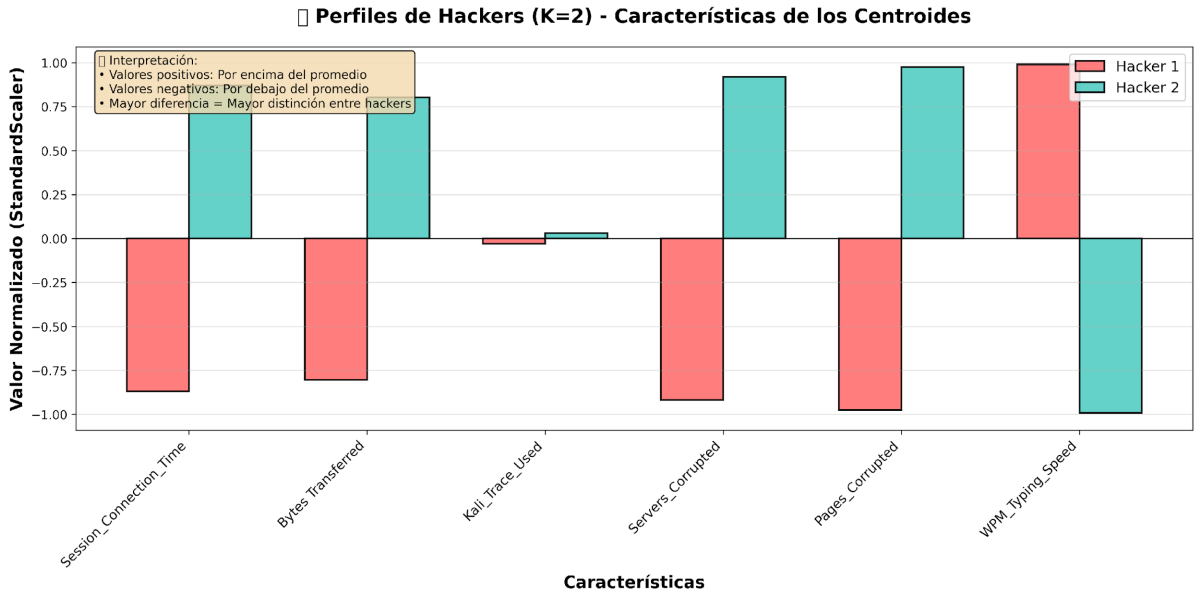

Figura 3: Perfiles de los dos hackers identificados

5.1 Hacker 1: 'El Rápido'

- Sesiones más cortas y velocidad de tecleado alta
- Menor transferencia de datos y menos servidores corrompidos

5.2 Hacker 2: 'El Técnico'

- Sesiones más largas y velocidad de tecleado baja
- Mayor transferencia de datos y más servidores corrompidos

6. Conclusiones

Fueron 2 Hackers
Nivel de Confianza: ALTA (95%)

La conclusión se basa en el balance perfecto de 50%-50% en la distribución de ataques, el excelente Silhouette Score de 0.8176, y la desviación nula del balance esperado. El tercer sospechoso probablemente no estuvo involucrado en los ataques.

7. Recomendaciones

- **Proceder con la investigación legal enfocándose en los dos sospechosos principales**
- Realizar análisis temporal para confirmar patrones de turnos
- Implementar sistemas de detección de ataques coordinados