1 Objetivos

- Implementar un modelo de ML que utilice la secuencia de llamadas a las APIs, para la clasificación de distintas categorías de malware.

2 Preámbulo

El análisis dinámico ofrece información sobre el comportamiento de un malware y cómo interactúa con el sistema que infecta. Al registrar, observar y analizar este comportamiento es posible evadir las técnicas de ofuscamiento que dificultan el análisis estático, pues el malware ejecuta las funciones cuyo código intenta ocultar.

Entre la información relevante que ofrece el análisis dinámico se encuentra la secuencia de llamadas a las APIs. A diferencia de un análisis estático donde podemos obtener el conjunto de APIs que un malware utiliza, la secuencia de llamadas muestra el orden en el tiempo en el que estas APIs son ejecutadas, información que se puede utilizar para derivar nuevas características en un modelo de aprendizaje de máquina, como los n-gramas.

3 Desarrollo

Obtenga del repositorio https://github.com/khas-ccip/api_sequences_malware_datasets el archivo VirusSample.csv

A partir de este dataset se deberán implementar **dos** modelos de clasificación de malware. Se sugiere la lectura del artículo “New Datasets for Dinamyc Malware Classification” donde se explica cómo se obtuvo la información que conforma el dataset.

Los modelos deben contemplar todas las fases de machine learning: exploración de datos, pre – procesamiento, ingeniería de características, implementación y validación (70% entrenamiento y 30 pruebas), validación cruzada con K folds para k = 10, y cálculo y explicación de las métricas de Accuracy, Precision y Recall para cada categoría de malware.

El artículo “New Datasets for Dinamyc Malware Classification” sirve como un benchmark para comparar modelos de clasificación, ¿se lograron obtener mejores métricas que las obtenidas en el artículo para la clasificación de malware?