1 Objetivos

- Implementar un modelo de ML que utilice la secuencia de llamadas a las APIs, para la detección 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 de NLP.

3 Desarrollo

A partir del dataset proporcionado se deberán implementar **dos** modelos de clasificación de malware. Se sugiere la lectura del artículo “Automated Behaviour’based Malware Detection Framework Based on NLP and Deep Learning Techniques” donde se explica cómo se construyó el dataset y el enfoque utilizado para la detección de malware.

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 , Recall y curva ROC para ambas clases (benigno, malware).

Los modelos pueden utilizar el mismo modelo de representación numérica de las secuencias de las APIS, pero deben utilizar un algoritmo diferente. Compare las métricas de los modelos en una tabla y discuta cual modelo detecto mejor el malware.

Rúbrica

Aspecto	**Punteo (sobre 100 pts)**
Pre-procesamiento Ingeniería de características	30
Implementación completa de los dos modelos (15 pts c/u)	30
Explicación de las métricas de rendimiento para cada modelo (15 pts c/u)	30
Comparación de modelos	10