1 Objetivos

- Implementar modelos de Data Science 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 ofuscado 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 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. Debe de analizar el artículo “Automated Behaviour’based Malware Detection Framework Based on NLP and Deep Learning Techniques,” en el artículo se explica cómo se construyó el dataset y el enfoque de NLP utilizado para la detección de malware.

3.1 Modelo 1

Se debe utilizar un modelo de representación numérica de las secuencias de las APIS que el grupo de trabajo definirá, por ejemplo, BoG o TF-IDF para obtener la lista de características que serán proporcionadas a un modelo de ML. Deben contemplar las fases típicas: 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).

3.2 Modelo 2

Se debe utilizar embeddings de clasificación generadas por alguno de los modelos enfocados a tareas de NLP de Gemini. Utilice los embeddings como información para una red neuronal.

Compare las métricas de los modelos en una tabla y discuta cual modelo detecto mejor el malware., justifique su análisis.

Rúbrica

Aspecto	Punteo (sobre 100 pts)
Modelo ML	45
Modelo DL	45
Comparación de modelos	10

Recursos

https://ai.google.dev/gemini-api/docs/embeddings

https://ai.google.dev/api/embeddings#v1beta.TaskType

https://github.com/google/generative-ai-docs/blob/main/site/en/gemini-api/tutorials/text_classifier_embeddings.ipynb

https://www.kaggle.com/code/markishere/day-2-embeddings-and-similarity-scores