1 Objetivos

- Investigar sobre la defensa contra ataques de evasión, inferencia, extracción y envenenamiento
- Utilizar el framework Adversarial Robustness ToolBox para atacar y defender modelos de ML y DL

2 Preámbulo

Seguridad en modelos de data science

La defensa de modelos de ML usa los mismos conceptos utilizados para los ataques. Por ejemplo, los ataques de envenenamiento usan perturbaciones sobre ciertas observaciones. Una técnica de defensa consiste en añadir perturbaciones a un dataset, y comparar los resultados. Si la predicción de ciertas observaciones no cambia a pesar de las perturbaciones añadidas, es probable que dichas observaciones fueron envenenadas con un patrón específico. En el caso de ataques adversariales, una técnica consiste en entrenar a un modelo con observaciones falsas para que aprenda a detectarlas.

3 Desarrollo

El laboratorio consiste en el desarrollo de dos ataques y su defensa. Un ataque es obligatorio (adversarial) y el otro ataque es libre.

Para la defensa del ataque adversarial, se debe entrenar al modelo con observaciones falsas generadas para su detección. Para el ataque libre, se puede utilizar el ART como defensa o utilizar otros frameworks de seguridad. Se puede utilizar el mismo ataque del laboratorio #7.

Sugerencia: instalar el ART framework y probar los ejemplos vistos en clase para asegurar que la herramienta fue instalada correctamente y que funciona sin problemas.

4 Calificación

- El grupo de trabajo será el mismo grupo que trabajó el laboratorio #6 – Clasificación de malware con DL.
    - Se debe entregar el link al repositorio en Github del laboratorio que debe incluir:
        - Jupyter Notebook: explicación de los ataques, evidencia de los pasos realizados, prueba del ataque en el modelo del laboratorio #6, explicación de las técnicas de defensa elegidas, evidencia de los pasos realizados, y evidencia de la efectividad de la defensa.
- La fecha de entrega será el martes **23 de mayo a las 17:20 horas.**
- Plagio parcial o total anula el proyecto, y se elevará el caso a la Dirección para las sanciones administrativas.