**Propuesta Proyecto Final**

**Samael de Jesus Salcedo Amortegui, Andres Fernando Roman, German David Garcia Nieves**

# 

1. ***Investigación y participación en el concurso de Quantopian con un algoritmo que utilice machine learning***

**Introducción:**

Quantopian es empresa de Boston que patrocina el desarrollo de algoritmos de trading los cuales permiten de manera automatizada la compra y venta de instrumentos financieros de tal forma que se obtenga una ganancia.

**Objetivos:**

Por medio del presente trabajo se pretende lograr los siguientes objetivos:

* Presentar un posible uso del machine learning como lo es el trading
* Desarrollo de una estrategia de trading por medio del uso de machine learning
* Participación en el concurso de Quantopian

1. ***Predicción de malware en máquinas de microsoft***

**Vista global del Reto**

El objetivo es **predecir la probabilidad de que una máquina Windows se infecte con varias familias de malware**, según las **diferentes propiedades de esa máquina**. Los datos de telemetría que contienen estas propiedades y las infecciones de la máquina se generaron combinando los informes de latidos y amenazas recopilados por la solución de protección de punto final de Microsoft, Windows Defender.

Cada fila en este conjunto de datos corresponde a una máquina, identificada únicamente por a **MachineIdentifier**. **HasDetections** Es un dato fundamental e indica que Malware fue detectado en la máquina. Utilizando la información y las etiquetas en train.csv, debe predecir el valor **HasDetections** para cada máquina en test.csv.

La metodología de muestreo es utilizada para crear este conjunto de datos, se diseñó para cumplir con ciertas restricciones comerciales, tanto en lo que respecta a la privacidad del usuario como al **período de tiempo durante el cual se ejecutó la máquina**. La detección de malware es intrínsecamente un problema de series de tiempo, pero se complica con la introducción de nuevas máquinas, máquinas que se conectan y desconectan, máquinas que reciben parches, máquinas que reciben nuevos sistemas operativos, etc. Mientras que el conjunto de datos que se proporciona aquí ha sido divididos por el tiempo, las complicaciones y los requisitos de muestreo mencionados anteriormente pueden significar que puede ver un acuerdo imperfecto entre sus validaciones cruzadas, las puntuaciones públicas y privadas. Además, este conjunto de datos no es representativo de las máquinas de los clientes de Microsoft en la naturaleza; se ha muestreado para incluir una proporción mucho mayor de máquinas de malware.

# 

# 

# 

# 

# 

# **2) Quora, clasificación de buenas preguntas**

**Vista global del Reto**

Un problema existencial para cualquier sitio web importante hoy en día es cómo manejar el contenido tóxico y divisivo. Quora quiere abordar este problema de frente para mantener su plataforma en un lugar donde los usuarios puedan sentirse seguros compartiendo sus conocimientos con el mundo.

[***Quora***](https://www.quora.com/) es una **plataforma que permite a las personas aprender unas de otras**. En Quora, las personas pueden hacer preguntas y conectarse con otras personas que contribuyen con ideas y respuestas de calidad unicas. Un desafío clave es eliminar las **preguntas no sinceras**, aquellas que se basan en premisas falsas o que pretenden hacer una declaración en lugar de buscar respuestas útiles.

En esta competencia, se desarrollarán modelos que identifiquen y marquen las preguntas que no sean buenas. Hasta la fecha, Quora ha empleado tanto el **aprendizaje automático** como la **revisión manual** para abordar este problema. Con su ayuda, pueden desarrollar métodos más escalables para **detectar contenido tóxico y engañoso**.

Aquí está tu oportunidad de combatir trolls en línea a escala. Ayude a Quora a mantener su política de "**Ser amable, ser respetuoso**" y continuar siendo un lugar para compartir y aumentar el conocimiento del mundo.

## **Descripción general**

En esta competencia, se predecirá si una pregunta hecha en Quora es sincera o no.

Una pregunta poco sincera se define como una pregunta que pretende contradecir una aclaración en lugar de buscar respuestas útiles. Algunas características que pueden significar que una pregunta no es sincera:

* No tiene un tono neutro
* Es despreciativo
* No está basado en la realidad
  + Basado en información falsa, o contiene suposiciones absurdas.
* Utiliza el contenido sexual

Los datos de capacitación incluyen la pregunta que se hizo y si se identificó como no sincera ( target = 1). Las etiquetas de Ground Ground contienen cierta cantidad de ruido: no se garantiza que sean perfectas.

Tenga en cuenta que la distribución de preguntas en el conjunto de datos no debe considerarse representativa de la distribución de las preguntas formuladas en Quora. Esto se debe, en parte, a la combinación de procedimientos de muestreo y medidas de sanitización que se han aplicado al conjunto de datos final.