Actividad Evaluable 3

Descripción

MÓDULO	**CyberSecurity Management**
ASIGNATURA	**Data Driven Security**
Fecha Límite de Entrega	Domingo 4 de Febrero de 2024, a las 23:59
Puntos	25% de la Nota Total
Carácter	Grupo (max 2 personas)

Enunciado:

En esta actividad se planteará la mejora de la calidad de resultados de un modelo de Aprendizaje Automático. **El resultado final debe estar disponible en un repositorio de código creado en Github.**

El repositorio de código debe ser público, es decir, estar configurado de forma que sea accesible para otras personas más allá del estudiante. Adicionalmente, se requiere entregar un documento de texto que contendrá el enlace al repositorio de código a través del campus virtual hasta la fecha límite de entrega. Se considerará tanto la corrección de las soluciones como su presentación.

Parte de esta actividad implica programar código R. **Tal código debe ser entregado en un documento de código** (formato de fichero *.R o .RMD*), tal que el código debe poderse ejecutar directamente sobre un terminal nuevo en R. El código es imprescindible para la corrección del ejercicio y deberá estar incluido en el contenido del repositorio de código. **Adicionalmente, para esta entrega también deberá acompañarse el código de el documento RMarkdown renderizado en formato HTML o PDF.**

Las entregas tardías serán marcadas como “tarde”, y pueden NO ser evaluadas.

Optimización de un modelo de ML

Obtención y carga de los Datos:

Queremos mejorar la precisión con la que un modelo de aprendizaje automático es capaz de clasificar el tráfico de red. Para ello, disponemos de un modelo ya programado que nos ofrece una cierta garantía de precisión para determinar correctamente si un determinado flujo de datos se trata en realidad de un tráfico con características similares a las de un ciberataque y, por lo tanto, el sistema de gestión y protección de la red debería bloquearlo o, por el contrario, se corresponde a patrón de tráfico de red normal.

Nuestro programa es capaz ya de generar el modelo y verificar la precisión de los resultados utilizando el conjunto de datos de red.

Exploración de Datos

1. **Exploración de los datos de tráfico de red disponibles**
   Como en cualquier actividad de análisis de datos, empezaremos por la exploración de los datos disponibles. Para ello se deberá incluir en el documento a entregar datos como el volumen de la muestra, las categorías (columnas) así como la tipología de estas y cualquier observación destacable que podáis encontrar.

   Pistas:
   Explorad las frecuencias de los distintos valores que tienen las diferentes columnas. Revisad la codificación de cada columna y aseguraros que tiene los tipos correspondientes en relación con las características de los datos que representan.
   Comprobad el fichero features.csv incluido en la carpeta con los datos necesarios para esta practica.

Generación del modelo de ML

2. **Comprender el código utilizado para segmentar el conjunto de datos entre datos de entrenamiento y datos de validación.**
   Para responder a esta sección no se requiere de ningún código a programar ni ningún apartado en el documento RMarkdown. Sin embargo, será fundamental comprender el código incluido en la práctica para la correcta documentación en relación a la optimización del modelo.

   Pistas:

- *Aprovechando el uso de RStudio como entorno de desarrollo, utiliza la sección de consulta de la documentación para revisar el propósito de cada función desconocida. También puedes usar ´?´ seguido del nombre de la función y ejecutar la instrucción para abrir la documentación de cada método. Alternativamente, también podéis usar ChatGPT para mayor facilidad.*

Mejora de resultados del modelo

3. Mejora de los resultados

Una vez ejecutado el código, comprobad la accuracy del modelo generado cuando se verifica con los datos de validación.

Investigad el código y implementad al menos 2 mejoras que consigan aumentar la precisión del modelo.

Pistas:

- *Este apartado es totalmente abierto. Existen numerosas configuraciones diferentes a las incluidas en el código suministrado que permiten optimizar el modelo para conseguir una mayor precisión. Primeramente, revisad la documentación de los distintos métodos utilizados para* ***particionar los datos y entrenar el modelo*** *a fin de descubrir posibles variaciones en los parámetros de la función en la configuración utilizada.*
- *Alternativamente, implementad nuevas versiones en las secciones de* data *enrichment y feature engineering para también mejorar el rendimiento del modelo.*

Links de interés

- https://es.wikipedia.org/wiki/Creaci%C3%B3n_de_atributos
- https://es.wikipedia.org/wiki/Validaci%C3%B3n_cruzada
- https://rafalab.dfci.harvard.edu/dslibro/caret.html
- Curva ROC/AUC:
    - Formal: https://www.youtube.com/watch?v=fsgDD0pNkZ0
    - Informal: https://www.youtube.com/watch?v=TmhzUdPpVPQ
-