Diseño y Evaluación de un Filtro de Dominios DGA

Los dominios generados automáticamente (DGA) son frecuentemente utilizados en ataques de ciberseguridad para evadir sistemas de defensa y organizar comunicaciones maliciosas. En esta actividad, los alumnos diseñarán un filtro para identificar dominios DGA a partir de datos recopilados sobre longitud de nombre de dominio, conteo de n-gramas, y entropía. El objetivo es elegir el modelo de clasificación que mejor discrimine entre dominios legítimos y dominios DGA, minimizando el riesgo para los usuarios.

Objetivos de la Actividad

1. Aplicar técnicas de análisis exploratorio de datos (EDA) para comprender las características y patrones del dataset.
2. Entrenar y evaluar modelos de aprendizaje supervisado para clasificar dominios como legítimos o DGA.
3. Seleccionar el mejor modelo en función de las métricas de evaluación.
4. Proponer mejoras o recomendaciones basadas en los resultados obtenidos.

Instrucciones y Pasos a Seguir

La actividad se desarrollará en varias fases que deberán ser documentadas y justificadas. Cada grupo debe seguir los siguientes pasos:

Fase 1: Análisis Exploratorio de Datos (EDA)

- **Comprensión del Dataset**: Describan las características del conjunto de datos proporcionado, que contiene información sobre longitud de los dominios, n-gramas y entropía.
- **Visualización de Datos**:
    - Utilicen diferentes gráficos como diagramas de caja (box plots), diagramas de violín (violin plots) y diagramas de dispersión para explorar la distribución de las características.
    - **Objetivo**: Identificar si existe una relación clara entre las variables que permita distinguir entre dominios DGA y legítimos.
- **Preguntas de Orientación**:
    - ¿Qué variables parecen tener más relevancia para la clasificación?
    - ¿Existen patrones claros que diferencien los dominios legítimos de los DGA?

Fase 2: Preprocesamiento de Datos

- **Tratamiento de Datos**:
    - Manejo de valores faltantes y outliers (si los hubiera).
    - Normalización o estandarización de las características, considerando las necesidades de cada modelo.
- **Selección de Características**:
    - Evaluar la relevancia de cada característica para el problema y justifiquen si alguna debe ser eliminada o modificada.

Fase 3: Entrenamiento de Modelos

- **Modelos a Entrenar**:
  - **k-Nearest Neighbors (kNN)**
  - **Regresión Logística**
  - **Support Vector Machines (SVM)**
  - **Red Neuronal**
  - **Árbol de Decisión (Decision Tree)**
  - **Naive Bayes**
- **Configuración de Entrenamiento**:
  - Dividir el conjunto de datos en entrenamiento y prueba.
  - Utilizar validación cruzada para garantizar la robustez de los modelos.
- **Preguntas de Orientación**:
  - ¿Qué hiperparámetros muestra cada modelo? ¿Cómo decidisteis ajustarlos?

Fase 4: Evaluación y Comparación de Modelos

- **Métricas de Evaluación**:
  - Calcular y comparar las siguientes métricas: **AUC (Área Bajo la Curva ROC)**, **Exactitud (Accuracy)**, **F1 Score**, **Precisión (Precision)**, y **Sensibilidad (Recall)**.
- **Análisis de Resultados**:
  - Comparar las métricas obtenidas por cada modelo.
  - Analizar las **matrices de confusión** para identificar los errores más comunes de cada modelo, especialmente falsos positivos y falsos negativos.
  - Visualizar los resultados con curvas ROC y lift curve para analizar el rendimiento de los modelos en distintas condiciones.
- **Preguntas de Orientación**:
  - ¿Qué modelo tiene mejor rendimiento general? ¿Cuál tiene menos errores críticos (falsos negativos)?
  - ¿Cuál de los modelos sería el más adecuado para implementar el filtro, considerando los riesgos de los errores de clasificación?

Fase 5: Recomendación y Mejoras

- **Selección del Modelo Final**:
  - Proponer cuál de los modelos sería el mejor para el filtro de dominios DGA, considerando el objetivo de minimizar falsos negativos (es decir, dominios DGA no identificados).
- **Mejoras Propuestas**:
  - Sugerir posibles mejoras al filtro, como la combinación de modelos (**Ensamblaje de Modelos**), ajustes adicionales en las características, o técnicas de regularización.
  - Evaluar el impacto del tamaño del conjunto de datos: ¿Cómo podría mejorar el rendimiento si se tuviera acceso a más datos?

Entrega

Cada grupo deberá entregar un informe que incluya:

1. **Resumen del análisis exploratorio de datos**, con gráficos relevantes y discusiones.
2. **Descripción del preprocesamiento realizado** y la justificación detrás de las decisiones tomadas.
3. **Detalles del entrenamiento de cada modelo**, incluyendo hiperparámetros, justificación y rendimiento.
4. **Análisis de resultados**, con tablas y gráficas de métricas, y discusión sobre cuál es el modelo más adecuado.
5. **Conclusiones y recomendaciones**, detallando la elección del modelo final y mejoras propuestas para el filtro.

Formato del Informe: 8-12 páginas, con una estructura clara (introducción, análisis, metodología, resultados, conclusiones).

Fecha de Entrega: 10 de noviembre a las 23:59

Criterios de Evaluación:

- **Calidad del Análisis Exploratorio**: Claridad y profundidad del análisis, incluyendo las visualizaciones y conclusiones.
- **Justificación del Preprocesamiento**: Razonamiento adecuado para las decisiones de limpieza y selección de características.
- **Entrenamiento y Evaluación de Modelos**: Correcta implementación y comparación de los modelos propuestos.
- **Conclusión y Selección del Modelo**: Capacidad para argumentar de manera clara y lógica cuál es el mejor modelo para el problema.
- **Originalidad y Mejora**: Creatividad en las propuestas de mejora para el filtro y el proceso.