

Entregable 3: Modelado Predictivo – Análisis Supervisado y/o No Supervisado

Deadline: 22/09

Objetivo: Aplicar técnicas de Machine Learning para desarrollar un MVP capaz de anticipar los movimientos de sitios impostores y/o predecir comportamientos potencialmente fraudulentos.

Proceso de Desarrollo de Modelos de Aprendizaje Automático



Made with Napkin

Vamos a plantear el modelado como un proceso de dos fases, esto es una práctica muy común y potente en la industria, ya que refleja cómo los proyectos de ciencia de datos a menudo evolucionan de una pregunta general a una más específica y accionable.

La Visión General: De la Detección a la Comprensión

Para este sprint, no solo vamos a construir un modelo que nos diga 'este sitio es fraudulento' o 'no lo es'. Vamos a ir un paso más allá, como se hace comúnmente en proyectos del mundo real, por eso nuestro objetivo es doble:

- 1. Detectar la Amenaza:** Primero, necesitamos una herramienta robusta que actúe como nuestro 'guardia de seguridad', identificando sitios con alta probabilidad de ser fraudulentos.
- 2. Entender al Sitio impostor/simulador:** Una vez que identificamos lasUrls sospechosas, necesitamos entender *cómo operan*. ¿Son todas iguales? ¿Hay diferentes 'tipos' de sitios fraudulentos? ¿Podemos encontrar patrones que nos ayuden a anticipar futuras amenazas?

Para lograr esto, vamos a dividir nuestro trabajo en dos fases claras y consecutivas.

Fase 1: El Modelo de Detección (Clasificación Binaria)

El "Quien":

- **Objetivo:** Construir un modelo de **clasificación binaria** que aprenda a distinguir entre un sitio web legítimo y uno potencialmente fraudulento.
- **Pregunta que respondemos:** *¿Es este sitio sospechoso de fraude? (Sí/No)*
- **Input:** El dataset maestro que preparamos en el Entregable 2, con todas nuestras features de ingeniería (longitud de dominio, uso de palabras clave, TLD, etc.).
- **Output:** Una probabilidad (de 0 a 1) de que un sitio sea fraudulento.
- **Técnica: Modelado Supervisado.** Usaremos algoritmos como Regresión Logística, Random Forest o Gradient Boosting.
- **Resultado clave:** Un modelo predictivo entrenado y validado, y una lista de sitios que nuestro modelo ha marcado como "alta probabilidad de fraude".

Fase 2: El Modelo de Perfilado (Clustering)

El "Cómo":

- **Objetivo:** Tomar *sólo* los sitios que nuestro modelo de Fase 1 clasificó como fraudulentos y agruparlos en clusters o "familias" de comportamiento similar.
- **Pregunta que respondemos:** *Dentro de los sitios fraudulentos, ¿existen grupos con características comunes? ¿Cuáles son los diferentes 'modus operandi' del fraude?*
- **Input:** El subconjunto de datos correspondiente a los sitios marcados como "alta probabilidad de fraude" por el modelo de clasificación.
- **Output:** Varios clusters (grupos) de sitios fraudulentos. Para cada cluster, un perfil que describa sus características distintivas.
- **Técnica: Modelado No Supervisado.** Usaremos algoritmos como K-Means, DBSCAN o Agrupamiento Jerárquico.
- **Resultado clave:** Perfiles detallados de los tipos de fraude. Por ejemplo:
 - **Cluster 1: "Phishing Bancario":** Dominios que imitan a bancos, usan palabras como "login", "cuenta", "seguridad".
 - **Cluster 2: "Estafas de E-commerce":** Sitios con ofertas increíbles, usan palabras como "gratis", "oferta limitada", "descuento".
 - **Cluster 3: "Malware Silencioso":** Sitios con poco contenido visible pero con scripts sospechosos o TLDs extraños.

Al combinar estas dos fases con un enfoque dual, no solo entregamos un 'detector de fraude', sino que generamos **inteligencia accionable**. Esto es lo que diferencia un proyecto académico de una solución de nivel industrial. Nuestros stakeholders no solo sabrán *qué* sitios bloquear, sino

que entenderán *por qué* y podrán anticipar nuevas tácticas de los adversarios. Este es el verdadero valor que como equipo de ciencia de datos podemos aportar.

Plan de Acción para el Equipo:

1. **Enfoque secuencial:** Primero, todos nos concentraremos en la **Fase 1**. Necesitamos el mejor modelo de clasificación posible.
2. **Hito intermedio:** Una vez que tengamos un modelo de clasificación con el que estemos satisfechos, definiremos un umbral de probabilidad para seleccionar a los "sospechosos".
3. **Comienzo de la Fase 2:** Con la lista de sospechosos, iniciaremos el análisis de clustering para descubrir los patrones ocultos.

Este enfoque les dará una estructura clara, donde aplicarán una metodología de trabajo avanzada que les permitirá ver cómo diferentes tipos de modelos (supervisados y no supervisados) pueden combinarse para resolver un problema complejo de manera integral.

Tareas:**1. Definición del Problema de Negocio y Traducción a Problema de ML:**

Clarificar el problema de negocio específico que el modelo intentará resolver (ej. detección de fraude, clasificación binaria: fraudulento / no fraudulento). Establecer los criterios de éxito del modelo desde una perspectiva de negocio.

2. Selección, Entrenamiento y Tuning de Hiperparámetros de Modelos Candidatos:

- **Selección de Modelos:** Investigar y seleccionar algoritmos de Machine Learning adecuados para el problema (ej. Regresión Logística, Árboles de Decisión, Random Forest, Gradient Boosting, SVM, Kmeans, entre otros).
- **División de Datos:** Dividir el dataset maestro (limpio y documentado del Entregable 2) en conjuntos de entrenamiento, validación y prueba.
- **Entrenamiento Inicial:** Entrenar los modelos seleccionados con los datos de entrenamiento.
- **Tuning de Hiperparámetros:** Utilizar técnicas como Grid Search o Random Search para optimizar los hiperparámetros de los modelos y mejorar su rendimiento.

3. Evaluación de Performance con Métricas Apropriadas:

- **Métricas de Evaluación:** Definir y calcular métricas relevantes para el problema (ej. Precisión, Recall, F1-Score, Curva ROC, AUC, Matriz de Confusión). Explicar por qué estas métricas son importantes en el contexto.
- **Análisis de Resultados:** Comparar el rendimiento de los diferentes modelos candidatos utilizando las métricas definidas.

4. Interpretación de Resultados y Validación del Modelo Final:

- **Interpretación del Modelo:** Analizar cómo el modelo toma sus decisiones (ej. Feature Importance para modelos basados en árboles, coeficientes para regresión).
- **Validación:** Realizar una validación cruzada para asegurar la robustez del modelo y su capacidad de generalización.
- **Análisis de Errores:** Identificar y analizar los tipos de errores que comete el modelo (falsos positivos, falsos negativos) y sus implicaciones de negocio.

Responsables:

- **Equipo de Ciencia de Datos Trainee:** Modelado, entrenamiento, evaluación inicial y documentación de los notebooks.
- **Científico de Datos Sr:** Guía en la definición del enfoque, revisión técnica de la selección de modelos y métricas, y apoyo en la interpretación de resultados.

Resultados Esperados:**• Notebooks de Entrenamiento y Validación de Modelos:**

Un notebook Jupyter o Google Colab conteniendo cada modelo significativo explorado:

- Código reproducible y comentado para la carga de datos, preprocesamiento final (si aplica), entrenamiento, tuning y evaluación de cada modelo.
- Visualizaciones claras de los resultados de la evaluación (ej. curvas ROC, matrices de confusión).
- Explicación de las decisiones tomadas en cada paso.
- **Documento Técnico-Funcional:**
 - Un informe conciso (1-2 páginas) que resuma los hallazgos clave del modelado.
 - Dashboard Interactivo (Opcional): Streamlit o Plotly Dash para crear una aplicación web simple que permita explorar los resultados del modelo o incluso realizar predicciones en tiempo real.

Stack Tecnológico (para este entregable):

- **Python:** pandas, numpy, scikit-learn (para modelado y métricas), matplotlib, seaborn (para visualizaciones de resultados).
- **Control de versiones:** Git/GitHub (para el repositorio de los notebooks y el documento técnico).

Criterios de Evaluación (relevantes para este entregable):

- **Aspectos Técnicos (50%):**
 - Calidad del código y documentación en los notebooks.
 - Rigor metodológico en la selección, entrenamiento y evaluación de los modelos.
 - Performance y robustez de los modelos.
 - Interpretabilidad de resultados.
- **Comunicación (50%):**
 - Claridad en el documento técnico-funcional.
 - Capacidad de síntesis y visualización de los resultados del modelo.

Consejos relevantes para este entregable:

- **Documentar todo:** Cada decisión técnica, desde la selección del modelo hasta la elección de las métricas, debe estar justificada y documentada.
- **Pensar en el usuario final:** ¿Cómo usaría este modelo un analista de ciberseguridad? Esto ayudará a enfocar la interpretación y las conclusiones.
- **Iteración frecuente:** Realizar pruebas y evaluaciones continuas para obtener feedback temprano sobre el rendimiento del modelo.
- **Mantener el enfoque:** Recordar siempre el objetivo de negocio de predecir comportamientos fraudulentos.