Proyecto "Detección de Sitios de Phishing en Argentina"

Diplomatura en Ciencia de Datos – FAMAF | Mentoría 2025
Sitio web funcional: Sitio detector phishing
Repositorio: detección-phishing-argentina-Diplo2025

---

1. Descripción general del proyecto

Este proyecto lo desarrollé dentro de un grupo de cuatro integrantes pertenecientes a la Diplomatura y fui acompañada por la Mentora Noelia Ferrero

Inspirado conceptualmente en el histórico robo al Banco Río de Acassuso (2006). Mientras aquel hecho representó un robo físico meticulosamente planificado, este proyecto abordo un tipo de "robo moderno": el robo de datos y credenciales a través de **sitios web fraudulentos** en Argentina.

El objetivo central fue **analizar, modelar y detectar sitios de phishing y fraudes digitales**, aplicando herramientas de Ciencia de Datos y técnicas de Machine Learning. El proyecto combino:

- **Web scraping** para estudiar el ecosistema web argentino.
- **Análisis exploratorio y curación de datos** para construir un dataset limpio y consistente.
- **Ingeniería de características** orientada a URLs, dominios, estructura HTML y respuesta del servidor.
- **Modelos supervisados** para clasificar sitios legítimos vs. fraudulentos.
- **Construcción de un sitio web funcional** para pruebas de detección en tiempo real.

---

2. Objetivos del proyecto

Crear un dataset propio de sitios web legítimos y fraudulentos, identificar señales estructurales que caracterizan a los sitios de phishing, entrenar modelos de Machine Learning y desarrollar un producto final utilizable por usuarios o equipos de ciberseguridad: un sitio donde se puede ingresar una URL y obtener una predicción.

---

3. Metodología de trabajo:

El proyecto se desarrolló en las siguientes etapas: Construcción del dataset, a través de web scraping. Curación y limpieza de datos. Entrenamiento de modelos supervisados, entremos varios modelos como: Árboles de decisión, Random Forest, Gradient Boosting. Y analizando diferentes métricas concluimos el mejor modelo, lo entrenamos. Finalmente llegamos a un producto final.

---

4. Herramientas principales utilizadas:

Python, Pandas, NumPy - **BeautifulSoup**, Requests - **Matplotlib, Seaborn - scikit-learn - Streamlit** (producto final) - **Git + GitHub** para control de versiones - **Google Colab** como entorno principal de desarrollo

---

5. Producto final:

- **Plataforma de detección:**
  El proyecto culminó con la creación de un sitio web interactivo desarrollado en Streamlit, donde cualquier usuario puede ingresar una URL y obtener una predicción inmediata sobre si el sitio es legítimo o fraudulento.

  Sitio web: https://diplophishing.streamlit.app/

  Este producto permite demostrar la aplicabilidad práctica del proyecto y su potencial de uso para ciberseguridad, trazabilidad y educación digital.

- **Disponibilidad: Repositorio del proyecto**

  GitHub: detección-phishing-argentina-Diplo2025

  Incluye: Notebooks del proceso completo - Dataset procesado - Código del modelo - Documentación - Versión del sitio en Streamlit

---

7. Difusión y presentación:

El proyecto fue seleccionado para presentarse en la **Córdoba Tech Week 2025**.La presentación contribuyó a visibilizar la problemática del phishing en Argentina y las oportunidades del uso de Ciencia de Datos para fortalecer la ciberseguridad.