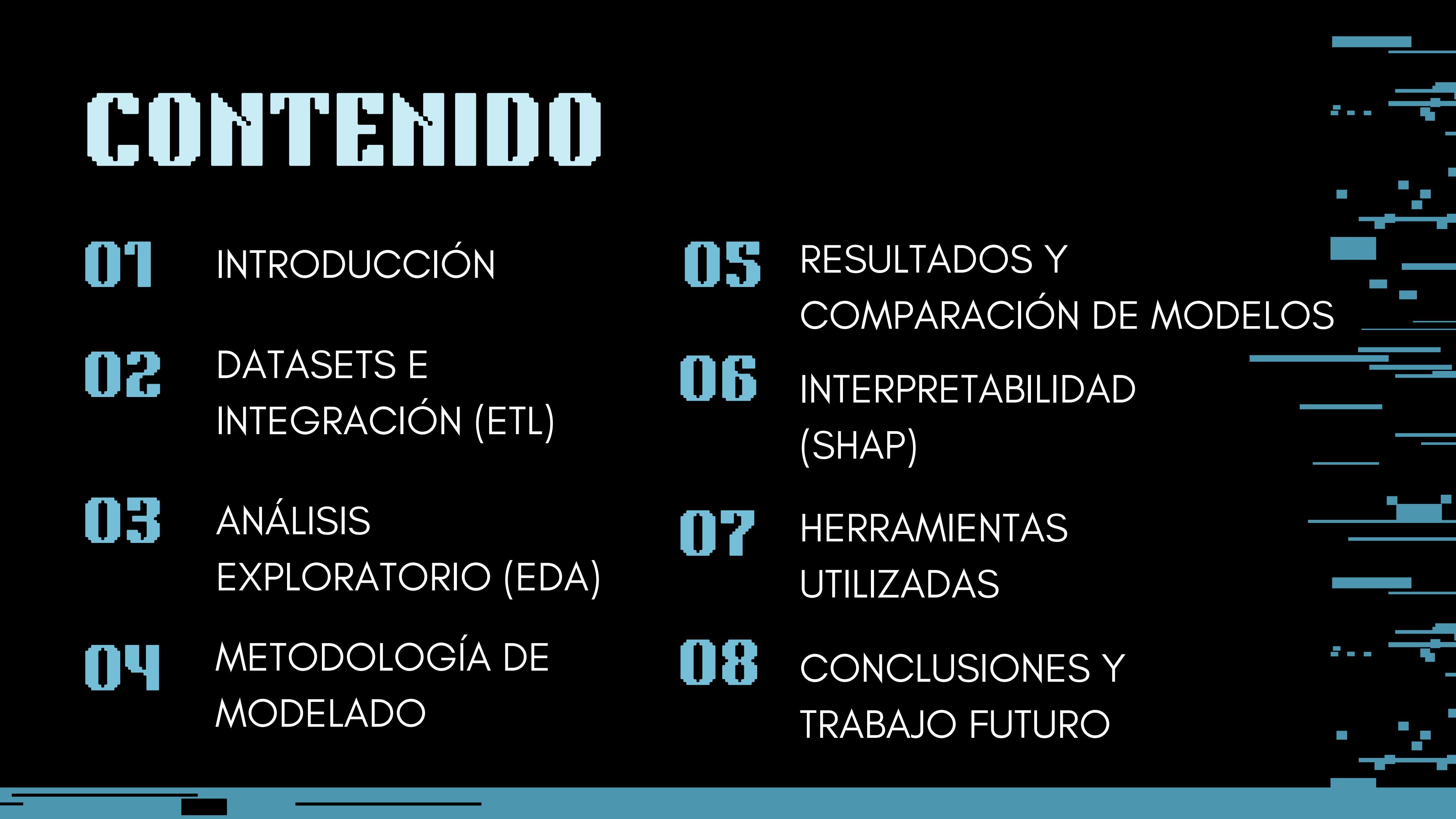


ANALITICA DE DATOS (PHISHING)



ELIAS DEL HOYO CESAR EDUARDO
DIEGO EMANUEL SAUCEDO ORTEGA
CARLOS DANIEL TORRES MACIAS

CONTENIDO

- 
- 01** INTRODUCCIÓN
 - 02** DATASETS E INTEGRACIÓN (ETL)
 - 03** ANÁLISIS EXPLORATORIO (EDA)
 - 04** METODOLOGÍA DE MODELADO
 - 05** RESULTADOS Y COMPARACIÓN DE MODELOS
 - 06** INTERPRETABILIDAD (SHAP)
 - 07** HERRAMIENTAS UTILIZADAS
 - 08** CONCLUSIONES Y TRABAJO FUTURO

INTRODUCCIÓN

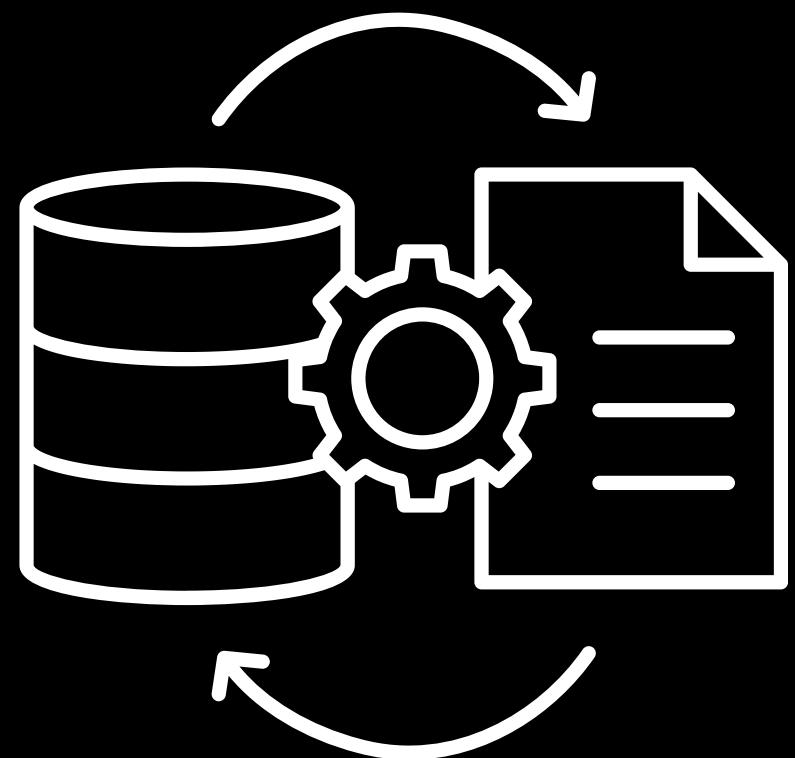
DETECCIÓN DE PHISHING MEDIANTE ANALÍTICA DE DATOS

- Proyecto enfocado en identificar correos electrónicos maliciosos utilizando técnicas avanzadas de analítica de datos.
- Integración de datos heterogéneos, modelado, calibración e interpretabilidad.
- Objetivo: desarrollar un modelo robusto que mejore la detección temprana de ataques de phishing.

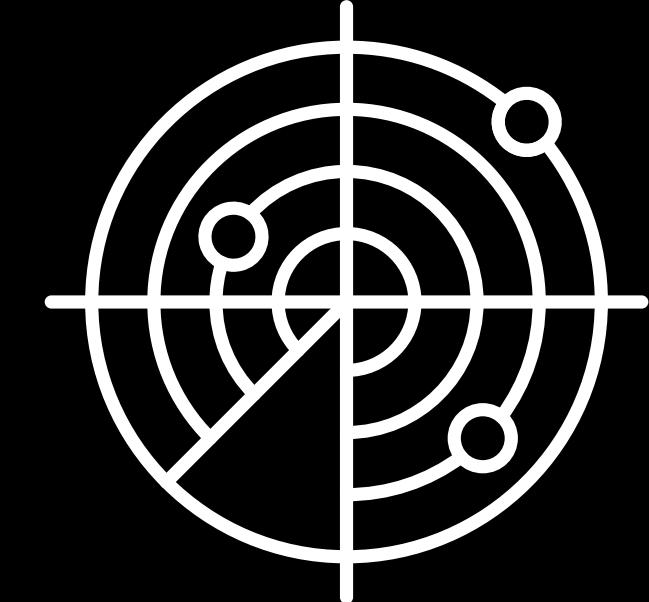


INTEGRACIÓN DE DATOS HETEROGÉNEOS

- Se integraron 3 datasets diferentes, con columnas no coincidentes.
- Tras limpieza, normalización y unión, el dataset final contiene 10,706 registros.
- Justificación del tamaño: La naturaleza heterogénea de los datos limita su volumen pero maximiza la calidad analítica.
- Proceso ETL documentado: estandarización, eliminación de duplicados, unificación de claves y estructura final.



EXPLORACIÓN DE DATOS EN EXCEL Y POWER BI



► EXCEL PERMITIÓ REVISIÓN INICIAL, DETECCIÓN DE VALORES EXTREMOS Y DISTRIBUCIÓN DE VARIABLES.

► POWER BI REVELÓ PATRONES CLAVE: FRECUENCIA DE ENVÍO POR DOMINIO, URGENCIA DE LOS MENSAJES Y CARACTERÍSTICAS DEL CUERPO DEL CORREO.

► HALLAZGOS PRINCIPALES

- EL URGENCY_SCORE MOSTRÓ CORRELACIÓN POSITIVA CON LA ETIQUETA PHISHING.
- ALGUNOS DOMINIOS EMISORES FUERON SIGNIFICATIVAMENTE MÁS FRECUENTES.

MODELOS IMPLEMENTADOS Y VALIDACIÓN

MODELOS UTILIZADOS

- ▶ REGRESIÓN LOGÍSTICA:
INTERPRETABILIDAD Y BASE COMPARATIVA.
- ▶ RANDOM FOREST: MEJOR DESEMPEÑO GLOBAL.
- ▶ GRADIENT BOOSTING:
- ▶ REFINAMIENTO DEL ERROR SECUENCIAL.

TÉCNICAS APLICADAS

- ▶ VALIDACIÓN CRUZADA K-FOLD ($K=5$).
- ▶ GRID SEARCH PARA HIPERPARÁMETROS.
- ▶ CALIBRACIÓN CON PLATT E ISOTÓNICA.

MÉTRICAS DE DESEMPEÑO



**RANDOM FOREST CALIBRADO OBTUVO
EL MEJOR RENDIMIENTO.**

INDICADORES CLAVE

- AUC elevado.
- Mejor F1-Score.
- Menor Brier Score tras calibración.
- La calibración mejoró la confiabilidad de las probabilidades generadas por el modelo.



INTERPRETABILIDAD (SHAP)

EXPLICABILIDAD DEL MODELO

Se aplicaron valores SHAP para conocer el impacto de cada variable.

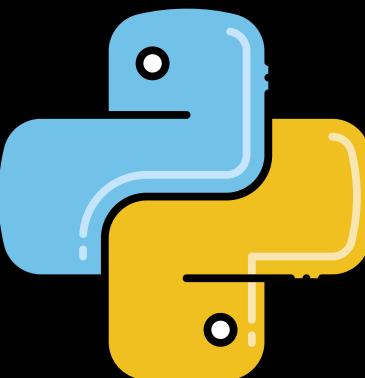
Atributos más influyentes:

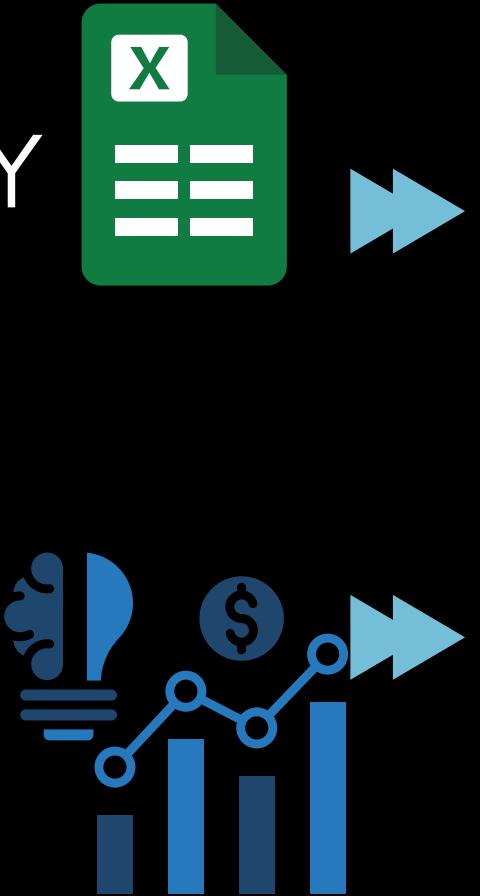
- url_count
- urgency_score
- body_upper
- subject_upper

A binary matrix visualization where each row represents a data point and each column represents a feature. The matrix is composed of blue '1's and white '0's. The columns are labeled with feature names: url_count, urgency_score, body_upper, and subject_upper. The matrix shows that the 'url_count' feature has a significant impact across most rows, while 'urgency_score' and 'body_upper' also show patterns of influence. The 'subject_upper' feature appears to have minimal influence based on this visualization.

Se generaron explicaciones locales y globales para casos específicos.

HERRAMIENTAS UTILIZADAS ECOSISTEMA TECNOLÓGICO



- ▶ EXCEL: LIMPIEZA INICIAL Y MUESTREO.
 - ▶ POWER BI: DASHBOARDS INTERACTIVOS
 - ▶ ORANGE DATA MINING: CLASIFICACIÓN Y CLUSTERING RÁPIDO.
- 
- ▶ PYTHON: MODELADO, CALIBRACIÓN Y ANÁLISIS INTERPRETATIVO.
 - ▶ HERRAMIENTA DE AUTOESTUDIO: (ESPECIFICAR) Y SU APORTE AL PIPELINE.
- 

CONCLUSIONES Y TRABAJO FUTURO

- SE LOGRÓ INTEGRAR CON ÉXITO UN CONJUNTO DE DATOS HETEROGÉNEO PARA LA DETECCIÓN DE PHISHING.
 - RANDOM FOREST CALIBRADO DEMOSTRÓ EL MEJOR BALANCE ENTRE DESEMPEÑO Y ESTABILIDAD.
 - LA INTERPRETABILIDAD MEDIANTE SHAP PERMITIÓ COMPRENDER DECISIONES DEL MODELO.
- * FUTURAS MEJORAS:
- INTEGRACIÓN DE MODELOS NLP MODERNOS (BERT/DISTILBERT).
 - INCORPORAR ANÁLISIS MULTIMODAL.
 - EXPANDIR EL DATASET MEDIANTE APIs O SCRAPING CONTROLADO



THANK YOU