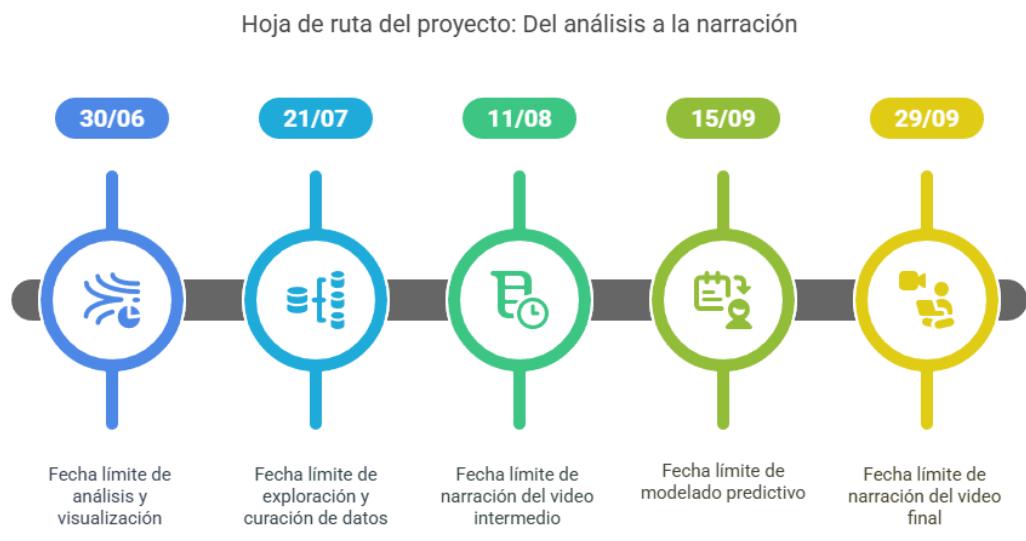


Roadmap del Proyecto: "El Robo del Siglo → Digital"



Contenido

Roadmap del Proyecto: "El Robo del Siglo → Digital"	1
Entregable 1: Análisis y Visualización	2
Entregable 2: Exploración y Curación de Datos	2
Video Intermedio: Storytelling de Entregables 1 y 2	3
Entregable 3: Modelado Predictivo (Supervisado y/o No Supervisado)	4
Video Final: Storytelling del Proyecto Completo	4
Stack Tecnológico	5
Metodología	5
Criterios de Evaluación	5
Consejos para el Éxito	6

Entregable 1: Análisis y Visualización

- **Fecha límite:** 30/06
- **Objetivo:** Nuestro primer objetivo es "**reconocer el terreno**". Debemos entender la anatomía (estructura) de los datos, evaluar su calidad inicial y descubrir los primeros hallazgos visuales que nos guíen en la investigación.
- **Tareas:**
 - Revisión general de datasets disponibles + estrategia dual de enriquecimiento: data enrichment para completar datos existentes y data augmentation para incorporar nuevos datasets
 - Ingesta y perfilamiento inicial de datos (data profiling): carga, inspección de tipos de datos, valores nulos y estadísticas descriptivas.
 - Desarrollo de un Análisis Exploratorio de Datos (EDA) inicial, generando las primeras visualizaciones clave.
 - Identificación de patrones, outliers y formulación de las hipótesis iniciales a partir de los insights.
- **Responsables:**
 - *Equipo de Ciencia de Datos Trainee* (web scraping, análisis, visualización)
 - *Científico de Datos Sr* (supervisión, validación de enfoque, feedback constructivo)
- **Resultados esperados:**
 - Notebook de Análisis Exploratorio (EDA) documentado, explicando el código y los hallazgos.
 - Informe de hallazgos iniciales (1-2 páginas) con las visualizaciones más relevantes y un resumen de las primeras conclusiones.

Entregable 2: Exploración y Curación de Datos

- **Fecha límite:** 21/07
- **Objetivo:** Ahora que conocemos el terreno, es hora de "**preparar las herramientas**". El objetivo es profundizar el análisis, mejorar la calidad de los datos y preparar un dataset robusto, limpio y confiable, que será la base para construir nuestro modelo predictivo.
- **Tareas:**
 - Tratamiento de valores faltantes, duplicados e inconsistencias.
 - Ingeniería de características (Feature Engineering) básica: curación, normalización y transformación de variables para prepararlas para el modelado.

- Análisis más profundo de correlaciones y relaciones entre variables para seleccionar las más prometedoras.
- Generación de un dataset maestro (master dataset) limpio y bien documentado.
- **Responsables:**
 - *Equipo de Ciencia de Datos Trainee* (data wrangling, EDA avanzado)
 - *Científico de Datos Sr* (guía de calidad de datos, apoyo en feature engineering y evaluación)
- **Resultados esperados:**
 - Dataset maestro curado (en formato CSV o similar), listo para el modelado.
 - Notebook de curación con código y justificación de las decisiones tomadas.
 - Documento técnico-funcional con hallazgos clave post-curación.

Video Intermedio: Storytelling de Entregables 1 y 2

- **Fecha límite:** 11/08
- **Objetivo:** Presentar nuestro primer "**informe de inteligencia**". Debemos comunicar de forma clara y narrativa los descubrimientos y aprendizajes de los dos primeros entregable a los stakeholders y/o público objetivo.
- **Tareas:**
 - Síntesis de los resultados clave del análisis y la curación.
 - Construcción de una narrativa de datos (Data Storytelling) coherente: problema → datos → descubrimientos.
 - Preparación del guión y grabación del video.
- **Estructura Narrativa:**
 1. Introducción: El problema del fraude web en la era digital
 2. Exploración: ¿Qué nos revelan los datos?
 3. Patrones identificados: Características distintivas de sitios fraudulentos
 4. Preparación: Cómo transformamos datos brutos en información valiosa
 5. Próximos pasos: Hacia la construcción de modelos predictivos
- **Responsables:**
 - *Equipo de Ciencia de Datos Trainee* (guión, visuales, exposición)
 - *Científico de Datos Sr* (revisión narrativa, apoyo técnico)
- **Resultados esperados:**
 - Video grabado de máximo 10 minutos.
 - Slides o material de apoyo con gráficos e insights clave.

Entregable 3: Modelado Predictivo (Supervisado y/o No Supervisado)

- **Fecha límite:** 15/09
- **Objetivo:** Llegó el momento de "**construir una herramienta predictiva**". Nuestro objetivo es aplicar técnicas de Machine Learning para desarrollar un modelo predictivo capaz de anticipar los movimientos del adversario (predecir comportamientos potencialmente fraudulentos).
- **Tareas:**
 - Definición clara del problema de negocio y su traducción a un problema de ML (ej. clasificación binaria: fraudulento / no fraudulento).
 - Selección, entrenamiento y tuning de hiperparámetros de modelos candidatos (regresión, árboles, k-means, etc.).
 - Evaluación de performance con métricas apropiadas para un problema de fraude (ej. Precisión, Recall, F1-Score, Matriz de Confusión).
 - Interpretación de resultados y validación del modelo final.
- **Responsables:**
 - *Equipo de Ciencia de Datos Trainee* (modelado y evaluación)
 - *Científico de Datos Sr* (guía en la definición del enfoque y revisión técnica)
- **Resultados esperados:**
 - Notebooks de entrenamiento y validación de modelos, reproducibles y comentados.
 - Documento técnico-funcional con interpretación del modelo (ej. feature importance), resultados de la evaluación y conclusiones.

Video Final: Storytelling del Proyecto Completo

- **Fecha límite:** 29/09
- **Objetivo:** El objetivo es comunicar de forma integral la "**historia del proyecto**", desde la hipótesis inicial hasta la solución final, demostrando el valor generado.
- **Tareas:**
 - Revisión y consolidación de todos los entregables.
 - Guión completo de storytelling: problema → datos → análisis → modelo → resultados → impacto.
 - Producción audiovisual con enfoque narrativo.
- **Estructura Narrativa Completa:**

- Planteamiento: El desafío del fraude web contemporáneo
- Exploración: Insights del análisis de datos
- Metodología: Estrategias de preprocessing y feature engineering
- Modelado: Enfoques de machine learning implementados
- Resultados: Performance del modelo y casos de uso
- Impacto: Aplicación práctica en ciberseguridad
- **Responsables:**
 - *Equipo de Ciencia de Datos Trainee* (contenido y presentación)
 - *Científico de Datos Sr* (curaduría final y orientación estratégica en narrativa técnica)
- **Entregables esperados:**
 - Video final de presentación (máximo 8-10 minutos), orientado a stakeholders.
 - Resumen Ejecutivo (Executive Summary) del proyecto a modo de presentación, con resultados clave y recomendaciones (1-2 páginas)
 - Repositorio de código final, limpio, documentado y reproducible.

Stack Tecnológico

- **Python:** pandas, numpy, scikit-learn, matplotlib, seaborn
- **Visualización:** Plotly, Seaborn, Bokeh, etc
- **Control de versiones:** Git/GitHub (opcional)
- Dashboard Final: Plotly Dash ó Streamlit para crear una aplicación web (opcional)
- **Documentación:** Jupyter Notebooks, Markdown, Visual Studio Code

Metodología

- **CRISP-DM:** Cross Industry Standard Process for Data Mining
- **Agile Data Science:** Iteraciones cortas con feedback continuo

Criterios de Evaluación

Aspectos Técnicos (50%):

- Calidad del código y documentación
- Rigor metodológico en el análisis

- Performance y robustez de los modelos
- Interpretabilidad de resultados

Comunicación (50%):

- Claridad en reportes escritos
- Efectividad del storytelling en videos
- Capacidad de síntesis y visualización
- Presentación de insights accionables

Consejos para el Éxito

Documentar todo: Cada decisión técnica debe estar justificada

Pensar en el usuario final (potenciales interesados en el proyecto, stakeholders): Ejemplo ¿Cómo usaría este modelo un analista de ciberseguridad?

Iteración frecuente: Buscar feedback temprano y continuo

Mantener el enfoque: Recordar siempre el/los objetivo/s de negocio

Colaboración efectiva: Aprovechar la diversidad de perspectivas del equipo