

Portafolio de Ciencia de Datos

Oscar Castro Rosso

Curso: Fundamentos de Ciencia de Datos

Módulo 10 – Desarrollo de Portafolio para Especialidades

Fecha: Septiembre 2025

Resumen Ejecutivo

Este portafolio se desarrolla en el marco del **Módulo 10 – Desarrollo de Portafolio para Especialidades**, con el propósito de consolidar y presentar los aprendizajes obtenidos a lo largo del programa de Ciencia de Datos. El portafolio busca mostrar la capacidad de integrar proyectos aplicados, comunicar resultados de manera profesional y reflejar competencias técnicas y analíticas.

Se incluyen proyectos de análisis exploratorio de datos, manejo de grandes volúmenes de información, automatización de reportes y selección asistida de algoritmos de machine learning. Además, se presenta un caso práctico aplicado a la detección de fraude con tarjetas, ejemplo de cómo trasladar el conocimiento adquirido a problemas reales. El portafolio sintetiza el progreso en herramientas, metodologías y reflexión crítica sobre el rol del científico de datos.

Proyecto 1: EDA Automático y Reportes

Herramienta en Python para generar informes automáticos de exploración de datos (EDA), con diagnóstico de calidad, outliers y gráficos básicos. Permite estandarizar y agilizar análisis iniciales.

Proyecto 2: EDA Robusto para Big Data

Script que procesa datasets grandes por chunks, limpia duplicados, detecta nulos y outliers, y exporta un dataset preparado para modelado. Incluye generación de gráficos estadísticos.

Proyecto 3: Asistente SelectML (AutoDataX)

Asistente formativo que guía la selección de modelos de machine learning según el tipo de problema y condiciones del dataset. Sugiere algoritmos y genera un script base listo para entrenar.

Proyecto 4 (Bonus): Detección de Fraude con Tarjetas

Caso aplicado con modelos balanceados, validación estratificada y métricas apropiadas (ROC-AUC, PR-AUC). Permite comprender cómo enfrentar problemas de desbalance extremo en escenarios reales.

Habilidades Técnicas

- Python (Pandas, NumPy, Scikit-Learn, Matplotlib, Seaborn, Statsmodels)
- Automatización de reportes con ReportLab y python-docx
- Procesamiento de datos masivos con chunks
- Análisis exploratorio de datos (EDA) automatizado
- Modelado supervisado: regresión y clasificación
- Selección asistida de algoritmos ML
- Control de versiones con Git/GitHub

Reflexión Personal Final

El proceso de construcción de este portafolio me permitió integrar proyectos técnicos con un enfoque profesional, mostrando mi capacidad de aplicar ciencia de datos a distintos contextos y comunicar resultados de forma clara. La automatización, la escalabilidad y el diseño de asistentes son aprendizajes clave que refuerzan mi preparación para desafíos futuros en el ámbito laboral.

Este portafolio es una carta de presentación profesional, evidencia de mis competencias adquiridas y una base sólida para seguir desarrollando especialidades en Ciencia de Datos.