

Proyecto Final

[Erick Naunay / Fundamentos de Data Science]

10 de abril de 2025

Introducción

Este documento describe las **instrucciones** y el **formato** que deberán seguir los estudiantes para la **entrega final** del curso de *Fundamentos de Data Science*. El objetivo es integrar todo el trabajo previo (proyectos parciales de EDA, Data Wrangling, Feature Engineering, Modelado, etc.) en un **informe final** y una **presentación ejecutiva** que sigan las fases de la metodología **CRISP-DM** y que se presenten tanto en *un repositorio de GitHub* como en un documento y exposición formal.

1. Entregables Principales

1. **Informe Final:** Documento escrito que incluya tanto la parte gerencial como la parte técnica, utilizando el **outline** descrito en las siguientes secciones.
2. **Presentación Ejecutiva:** Diapositivas o formato similar, para exponer el proyecto en una junta o ante directivos.
3. **Repositorio en GitHub:** Estructurado y completo, con todos los notebooks, scripts, datos y reportes relevantes.

2. Estructura Requerida

Para la **estructura temática** del Informe Final y la Presentación Ejecutiva, se utilizará el siguiente **outline** de referencia (adaptado de CRISP-DM). El mismo se compone de:

2.1. 1. Título y Objetivo del Proyecto

- **Título del Proyecto:** Nombre conciso que describa la esencia del proyecto.
- **Objetivo General y Visión (Business Understanding – CRISP-DM):**
 - *Resumen Ejecutivo:* ¿Qué problema de negocio o investigación se pretende resolver?

- *Valor e Impacto*: ¿Por qué es relevante para la empresa u organización?

Para la Presentación Ejecutiva:

- Explique en términos de *retorno de inversión*, *reducción de costos* u *oportunidad de crecimiento*.
- Mencione brevemente la relevancia estratégica.

Para el Informe Final (Detalle Técnico):

- Incluya la *pregunta central*, *hipótesis inicial* y la alineación con objetivos corporativos o de investigación.

2.2. 2. Contexto y Alcance

■ **Antecedentes (Business Understanding – CRISP-DM):**

- Situación Actual: contexto de la industria o departamento.
- Dolor del Negocio: problema que experimentan los usuarios/clientes/área.

Presentación Ejecutiva:

- Use ejemplos de negocio concretos.
- Destaque el *tamaño de la oportunidad* y posibles impactos de resolver el problema.

Informe Final (Técnico):

- Profundice en *cómo se diagnosticó* la necesidad (entrevistas, hipótesis, datos preliminares).
- Explique el *propósito* y *requisitos* que el proyecto cubrirá.

■ **Definición de Alcance:**

- Scope: qué sí y qué no incluirá el proyecto.
- Limitaciones: restricciones de tiempo, recursos o tecnología.

2.3. 3. Entendimiento de los Datos (Data Understanding – CRISP-DM)

- **Fuentes de Datos:** origen (bases internas, APIs, sensores, datasets externos), tipo (estructurados, no estructurados), volumen, frecuencia.
- **Descripción y Calidad de los Datos:** perfil de variables, outliers, faltantes, consistencia y confiabilidad de las fuentes.

Presentación Ejecutiva:

- Resaltar aspectos críticos: calidad general y viabilidad del proyecto.

- Mencionar requerimientos de sistemas o licenciamiento, si aplica.

Informe Final (Técnico):

- Incluir estadísticas descriptivas detalladas.
- Análisis de correlaciones y visualizaciones iniciales (charts, histograms, etc.).

2.4. 4. Preparación de los Datos (Data Preparation – CRISP-DM)

■ Limpieza y Transformaciones:

- Procesos de limpieza (tratamiento de faltantes, outliers).
- Feature Engineering (normalizaciones, escalados, variables derivadas).

■ Integración de Datos:

- Fusión de múltiples fuentes, validación de consistencia, ajustes de formatos.

Presentación Ejecutiva:

- Resaltar *costos* y *complejidad* de la preparación de datos.
- Cómo esto mejora la calidad y confiabilidad del análisis.

Informe Final (Técnico):

- Detallar metodologías y scripts utilizados.
- Explicar cada decisión de diseño (por qué imputar, escalar, etc.).

2.5. 5. Modelado (Modeling – CRISP-DM)

- **Selección de Modelos:** razonamiento sobre la técnica (regresión, clasificación, clustering, series de tiempo, etc.) y algoritmos escogidos.
- **Entrenamiento y Validación:** metodología (train/test split, validación cruzada), hiperparámetros y justificación.
- **Métricas de Evaluación:** RMSE, Accuracy, F1, AUC, etc., comparación de modelos y justificación de la elección final.

Presentación Ejecutiva:

- Conectar con KPIs de negocio (costo de error, precisión esperada, impacto financiero).
- Usar *gráficos claros* y *tablas de alto nivel* para comparativas.

Informe Final (Técnico):

- Detallar los pasos de validación (n-fold CV, configuraciones, frameworks).
- Mostrar resultados numéricos y visualizaciones (ROC, tablas detalladas).

2.6. 6. Evaluación e Interpretación de Resultados (Evaluation – CRISP-DM)

- **Análisis de Desempeño:** interpretar resultados en términos de negocio o hipótesis, destacar fortalezas y limitaciones.
- **Factores de Éxito y Riesgos:** sesgos potenciales, riesgos en adopción o implementación.

Presentación Ejecutiva:

- Énfasis en la *eficacia* y *viabilidad* para el negocio.
- Posibles impactos financieros o en la toma de decisiones.

Informe Final (Técnico):

- Explicar errores más comunes y ejemplos de casos (falsos positivos/negativos).
- Proponer mejoras (más datos, features adicionales, nuevas arquitecturas).

2.7. 7. Plan de Implementación (Deployment – CRISP-DM)

- **Propuesta de Despliegue:** arquitectura de la solución, entorno de producción, API, dashboard, recursos necesarios.
- **Estrategia de Monitoreo y Mantenimiento:** mecanismos de seguimiento de desempeño, plan de reentrenamiento.

Presentación Ejecutiva:

- *Roadmap* de implementación y cronograma.
- Requerimientos de inversión y retorno esperado.

Informe Final (Técnico):

- Diagramas de arquitectura, entornos de prueba/producción, herramientas DevOps.
- Plan de retrain o actualización de parámetros.

2.8. 8. Conclusiones, Próximos Pasos y Recomendaciones

- **Resumen de Hallazgos Clave:** logros principales del proyecto y su relevancia.
- **Próximos Pasos:** extensiones futuras, nuevas líneas de investigación o mejoras.
- **Recomendaciones:** acciones concretas para la dirección, priorización de mejoras de datos, integración con otros sistemas.

Presentación Ejecutiva:

- Destacar *impacto estratégico* y recomendaciones inmediatas.
- Oportunidades de escalabilidad.

Informe Final (Técnico):

- Propuestas de investigación adicional y experimentación.
- Documentación consolidada para desarrollo continuo.

2.9. 9. Apéndices (Opcional)

- **Documentación Técnica Detallada:** pipeline completo, scripts principales.
- **Referencias y Fuentes:** bibliografía, papers, librerías y frameworks utilizados.

3. Repositorio de GitHub

Todos los materiales (código, datos, reportes) deben ser versionados y **organizados** en un repositorio de GitHub, manteniendo la estructura propia de proyectos de ciencia de datos:

- **Carpeta data/:** datos en crudo (**raw/**), datos intermedios (**interim/**), datos finales (**processed/**).
- **Carpeta notebooks/:** notebooks ordenados de EDA, Data Wrangling, Feature Engineering, Modelado, etc.
- **Carpeta src/:** scripts modulares para limpieza, transformación, modelado, utilidades.
- **Carpeta reports/:** informes finales, presentaciones, figuras.
- **README.md:** descripción del proyecto, uso, requerimientos, etc.
- **.gitignore:** indicar archivos o carpetas que no serán versionados.

4. Formato de Presentación

- **Informe Final:** puede entregarse en formato PDF siguiendo el outline. Deberá contener tanto el enfoque ejecutivo como el técnico (en secciones separadas o integradas con claridad).
- **Presentación Ejecutiva:** debe ser concisa (ejemplo: 10-15 diapositivas) y enfocarse en resultados y conclusiones de negocio, alineados con la información técnica mínima necesaria.
- **Exposición Oral (si aplica):** preparar una defensa de 15-20 minutos, con espacio para preguntas y demostraciones.

5. Evaluación

Se tendrán en cuenta los siguientes criterios:

1. **Claridad y coherencia** en la aplicación de CRISP-DM.
2. **Calidad del Análisis** (EDA, Modelado, Evaluación de Resultados).
3. **Integración de todos los avances** previos (proyectos parciales) en un solo documento y repositorio unificado.
4. **Presentación profesional** del Informe y la Exposición (estructuración, nivel de detalle adecuado a las audiencias).
5. **Organización y limpieza** del repositorio en GitHub (legibilidad del código, documentación, versiones).

Conclusión

Siguiendo esta guía, el estudiante podrá entregar un proyecto final **completo** que aborde tanto las necesidades ejecutivas como los requerimientos técnicos. La combinación de la **Presentación Ejecutiva** y el **Informe Técnico** permite comunicar el valor del proyecto a todo nivel de la organización, mientras que el **Repositorio en GitHub** aporta la trazabilidad y reproducibilidad necesarias en Data Science.

¡Éxito en su Proyecto Final!