



fcfm

FACULTAD DE CIENCIAS
FÍSICAS Y MATEMÁTICAS
UNIVERSIDAD DE CHILE



EdV
Para estudiantes de Educación Básica y Media.
UNIVERSIDAD DE CHILE

Clase III: Análisis exploratorio de datos

FM849 Proyecto de Ciencia de Datos: Inteligencia Artificial (IA) y sus aplicaciones

Máximo Flores Valenzuela (mflores@dcc.uchile.cl)

Universidad de Chile • 9 de agosto de 2025

Contexto

- En la clase anterior, quedó claro que hay que conocer bien el problema antes de idear una solución.
- Una vez ideada la solución, hay que plantear los objetivos de trabajo para poder fragmentar los esfuerzos y tener claro el «fin» del proyecto.
- Ahora estamos en la fase de **implementación de la solución**.

Antes de cualquier cosa: metodología de trabajo

La metodología que usaremos en este curso es una adaptación de CRISP-DM (1999). Es una metodología bastante buena para personas que están recién partiendo en el desarrollo de proyectos de *Machine Learning*.

1. Entendimiento del problema: ¿qué es lo que busco resolver?
2. Entendimiento de los datos: ¿qué datos tenemos o necesitamos para resolver el problema? ¿Están «limpios»?
3. Preparación de los datos: ¿cómo puedo organizar los datos para modelar el problema?
4. Modelamiento: ¿qué técnicas de *Machine Learning* debo aplicar para resolver este problema?
5. Evaluación: ¿qué forma de evaluación se ajusta al problema que quiero resolver?
6. Despliegue: ¿cómo despliego esto en producción para su uso? ← no lo veremos.

Entendimiento de los datos

Es una fase muy importante para partir trabajando con un conjunto de datos. Permite responder preguntas como: **¿Qué tienen mis datos? ¿Qué problemas existen? ¿Hay información que pueda eliminar? ¿En qué afecta que yo elimine esta información?** En cuestión:

- Permite complementar el conocimiento adquirido en la definición del problema y la solución.
- Permite saber si es factible con los datos que tengo solucionar el problema.
- Permite cuestionarse qué porción de los datos es realmente útil, y cuál no.
- Permite encontrar errores de formato, datos faltantes, redundantes, etc.
- Permite encontrar comportamientos o patrones previos en los datos.

Análisis exploratorio de datos (EDA)

El análisis exploratorio de datos, o en inglés, EDA, permite comprender las características esenciales de un conjunto de datos.

Se basa en dos criterios principalmente:

1. Estadísticas de resumen.
2. Visualización de datos.

Al final de esta clase veremos un ejemplo de aplicación.



Principales herramientas a usar (por ahora)



Google Colab (motor)



NumPy (operaciones)



Matplotlib (gráficos)



pandas

Pandas (carga y trabajo de datos)

¡A la sección de código!