



fcfm

FACULTAD DE CIENCIAS
FÍSICAS Y MATEMÁTICAS
UNIVERSIDAD DE CHILE



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

Clase II: Definición de proyecto en Ciencia 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

¿Cómo definimos un proyecto?

- Necesitamos conocer de manera clara el problema a resolver para proponer una solución.
- La definición de un problema debe considerar al menos estas aristas:

Usuario + **Necesidad** + **Complicación** + **Hallazgo**

- ▶ Entonces, ¿quiénes son los usuarios? ¿Qué requieren o necesitan? ¿Qué es lo que no permite que esta necesidad sea resuelta? ¿Por qué?

Procesos



Observación de la realidad



Identificación de oportunidad



Definición del problema



Propuesta de solución

En general, podemos describir el flujo de definición de un proyecto en 4 etapas.

Ustedes implícitamente ya observaron la realidad e identificaron una oportunidad de mejora en el laboratorio pasado.

Vamos a ver un ejemplo...

Apliquemos lo anterior a un proyecto bajo el contexto de la Ciencia de Datos.

- Un cliente, *Borel*, dueño de una empresa de útiles escolares, ha recibido bastantes quejas por su página web.
- Los usuarios de la empresa de *Borel* se quejan porque las recomendaciones son malas. El buscador no funciona bien en la tienda en línea. Esto genera una mala experiencia de compra.
- *Borel* debe solucionar este problema pronto, porque si bien siempre estuvo a la vanguardia, la competencia se digitalizó y provee una mejor experiencia.

Borel nos pidió ayuda a nosotros. Primero debemos asegurarnos de entender bien el problema.

Definiendo el usuario

¿Quién es el usuario? *¿Borel?* *¿Los usuarios de la página web de Borel?* Lo cierto es que depende del enfoque que queramos darle al problema. Ambos en este caso tienen necesidades.

- Recuerden que su contraparte (para este curso, sus compañeros/as y equipo docente) no tiene por qué tener todo el contexto. Deben ser específicos.
 - ▶ ¿Qué significa esto? Hay que reunir más información si es necesario. La caracterización del usuario debe ser clara.

Definiendo la necesidad

En el ejemplo anterior, tomaremos como usuario a las personas que navegan en la página web de *Borel*. **¿Qué es lo que necesitan?**

En este caso está claro, la necesidad que tienen es buscar y comprar útiles escolares por internet. No tenemos más contexto, pero podría existir.

- La pregunta «¿por qué prefieren comprar por internet?» caracteriza a nuestra muestra.

Definiendo la complicación

Corresponde al «pero» que se desprende directamente de la necesidad. Es más, la palabra entre comillas es clave.

«El usuario X necesita (o requiere) Y **pero...**»

- Usualmente esta parte es la génesis del problema. **¿Qué es lo que está pasando?**

En nuestro ejemplo, la complicación es que no pueden realizar una búsqueda fructuosa en la página web de *Borel*.

Y por último (pero no menos importante)... el hallazgo

Es lo que se descubre en la investigación del estado del arte; la causal del problema. Responde a la pregunta: **¿por qué está pasando?**, y está conectado directamente con la situación.

- Un «¿por qué...?» se responde con «porque...». Esta es la palabra clave que incluiremos en la definición.

«El usuario X necesita Y pero [...] **porque** [...]»

En este caso, la causal del problema es que las tecnologías de recuperación de información involucradas para el sistema de búsqueda no están funcionando correctamente.

- Esto está bien, pero podría estar mejor. ¿Por qué no están funcionando correctamente? Hay que resumir esta formulación... aprovechen la atención de su contraparte.

Formulación formal del problema

Ya tenemos las 4 características esenciales para formular nuestro problema, que queda de la siguiente manera:

«Las personas que navegan en la página web de *Borel* **necesitan** buscar y comprar útiles escolares por internet, **pero** no pueden realizar una búsqueda fructuosa en el sitio mencionado **porque** las tecnologías de recuperación de información que implementa no están funcionando correctamente».

- ¡Esto no es una receta estricta! Pero fíjense que poniendo atención a los detalles que hemos conversado podemos describir nuestro problema considerando sus aristas principales.

¿Qué procede?

Ya tenemos la definición del problema. Con toda la búsqueda de información debiésemos estar claros de qué es lo que hay que resolver.

¿Qué hacemos ahora?



Lo que espero...

- En este curso, esperaré que ustedes apliquen técnicas tradicionales de aprendizaje automático (*Machine Learning*) en el contexto del área de Inteligencia Artificial. Las veremos en la cátedra del próximo sábado.
- En la cátedra pasada, vimos que principalmente nos enfocaremos en dos tareas: clasificación y agrupación. Deben ser capaces de decidir cuál usarán y plasmarlo en la propuesta de solución.

Definiendo la propuesta de solución

- Se espera que sean capaces de explicar con palabras **cuál** es la técnica que van a ocupar para resolver el problema y **por qué** la eligieron.
 - Revisitaremos cuándo hay que ocupar lenguaje técnico y cuándo no más adelante. Al menos en este curso, deben ocupar lenguaje técnico.
- Para el problema que vimos anteriormente, una propuesta válida podría ser:

«Se propone como solución usar técnicas de agrupación sobre el catálogo de productos. Esta técnica permite agrupar útiles escolares con características similares para mostrarlos en la búsqueda por lenguaje natural».

Por temas de tiempo, no nos preocuparemos de ahondar en cómo se procesa el lenguaje natural en esta definición (¡es un ejemplo!).

¿Es claro el trabajo a realizar? ¿O sigue siendo muy genérico?

Objetivos

Permiten trazar la línea de trabajo de manera clara. Se separa en dos definiciones:

- **Objetivo general**: se alinea con la propuesta de solución. Corresponde a una macrotarea.
- **Objetivos específicos**: cada una de las tareas que se realizarán para completar el objetivo general o macrotarea.

Todos los objetivos deben ser claros, empezar con un verbo en infinitivo (p. ej. «analizar», «crear», etc.), deben poder comprobarse (p. ej. «crear una página web» → «demostración de la página web»), y deben ser razonables (es decir, considerar el tiempo que tienen para completar cada objetivo específico).

Los objetivos específicos deben formar una partición del objetivo general.

Breve pincelada a la definición «SMART» de objetivos



FED Digital Business SL (2025). Foro de Economía Digital Business School. Recuperado de <https://foroeconomiadigital.com/courses/objetivos-s-m-a-r-t/>.

En la definición de algunos proyectos, se suele usar «SMART» para los objetivos. Mi recomendación: úsenlo con discreción.

- Permite estandarizar la manera en la cual se escriben los objetivos, considerando 5 aristas principales (funciona muy bien).
- Puede establecer metas inciertas (medible: ¿quiero reducir el 20% de un criterio? ¿El 50%?). Si las establecen, **deben cumplirlas al pie de la letra**.

Ejemplos de definición de objetivos

Caso I:

- **Objetivo general:** Desarrollar un modelo de *Machine Learning* que prediga el riesgo de abandono escolar en estudiantes de secundaria, utilizando datos académicos y socioeconómicos históricos.
- **Objetivos específicos:**
 1. Recopilar y limpiar los datos académicos y socioeconómicos de los últimos cinco años.
 2. Analizar la relación entre variables socioeconómicas y el rendimiento académico.
 3. Entrenar y validar modelos predictivos utilizando métricas que permitan evaluar el desempeño.
 4. Implementar un dashboard interactivo para visualizar las predicciones y riesgos.

¿Cuál de las dos definiciones es correcta...? ¿Por qué?

Caso II:

- **Objetivo general:** Mejorar la educación con tecnología y datos.
- **Objetivos específicos:**
 1. Hacer cosas con los datos de las escuelas.
 2. Ver si el modelo funciona bien.
 3. Usar IA para ayudar a los estudiantes.

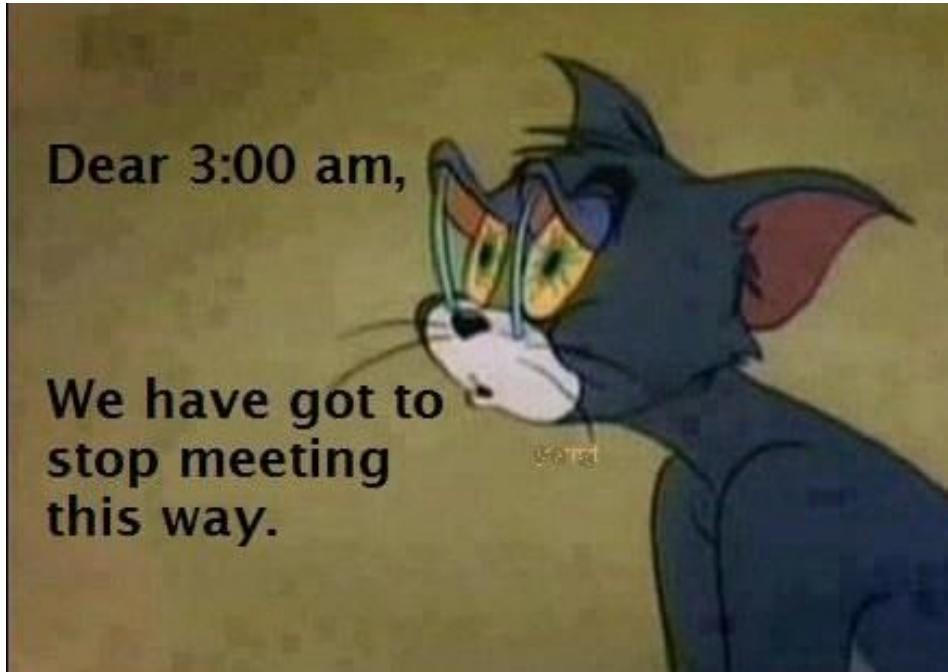
En su presentación final...

1. Deben definir formalmente su problema a resolver.
2. Deben definir su propuesta de solución.
3. Deben dejar claro cuáles fueron los objetivos de su proyecto, si se cumplieron o no, y la evidencia de su cumplimiento.

No deben seguir necesariamente las «recetas» que hemos estado viendo, pero sí es altamente sugerido. Si realizan cambios en sus definiciones, no se les pueden escapar detalles que introduzcan ambigüedades.

Spoiler: La evaluación no se limita sólo a esto. Más adelante seguiremos mencionando aspectos que evaluaremos.

Sí, sí, todo muy bien, pero...



¿Cómo se supone que debemos encontrar la información?

En este curso, también veremos *tips* para buscar información en internet de manera eficiente. Deben administrar bien su tiempo.

Páginas recomendadas

! ¡Las búsquedas en inglés dan mejores resultados!

- Para temas académicos, recomiendo mucho las siguientes páginas (no se limiten sólo a ellas):
 - Google Académico: <https://scholar.google.com/>.
 - arXiv: <https://arxiv.org/>.
 - ScienceDirect: <https://www.sciencedirect.com/>.

Sirve tanto para buscar información del estado del arte en diversos tópicos, como para buscar términos con los cuales no estén familiarizados. En general, son investigaciones científicas, así que el público objetivo es esa comunidad.

Páginas recomendadas

- Para temas más generales, generalmente Medium (<https://medium.com/>) cumple muy bien la tarea.
- Y obviamente... las páginas que implementan modelos grandes de lenguaje (LLMs), como ChatGPT, Claude, Gemini, etc.
 - ▶ Sí o sí les subiré un video asíncrono hablándoles sobre el tópico de *prompt engineering*, para que puedan sacarle el máximo provecho a estos modelos. Por ahora, les recomiendo buscarlo en las páginas que les recomendé :-).

Técnicas de búsqueda en Google: *Dorking*

Una de las técnicas que merece ser mencionada es el *dorking* en Google ([guía](#)). Corresponde a una serie de instrucciones que uno puede insertar en la búsqueda para hacerla mucho más específica. Algunos de los ejemplos:

- ¿Quieres buscar un texto **exacto**? Enciérralo entre comillas: "este es un ejemplo".
- ¿Quieres buscar en una página **exacta**? Usa `site:<página web>` al final de tu búsqueda, donde `<página web>` debe ser reemplazado por la URL.
- ¿Quieres buscar un tipo de archivo **exacto**? Usa `filetype:<extensión>` al final de tu búsqueda. Por ejemplo, `filetype:pdf` si les interesa extraer documentos en ese formato.

En internet les va a aparecer como *Google Hacking*, lo que puede sonar imponente, pero esto es sólo si le dan un mal uso... (no lo hagan).

¿Dudas?