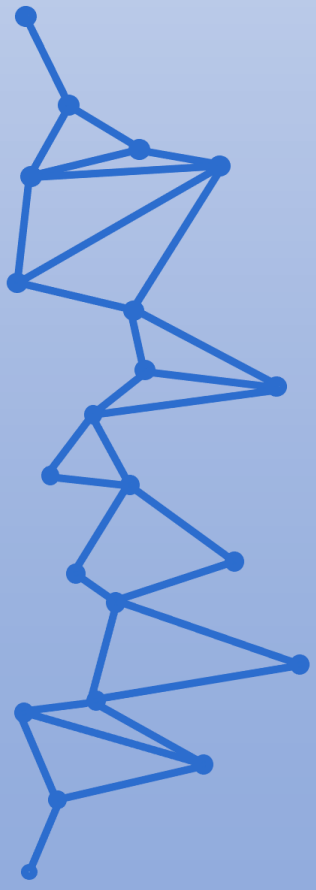
Curso de Especialización de
Inteligencia Artificial y Big Data (IABD)



Programación de Inteligencia Artificial

UD08. Planificación y estructura de un proyecto de Inteligencia Artificial. Resumen.

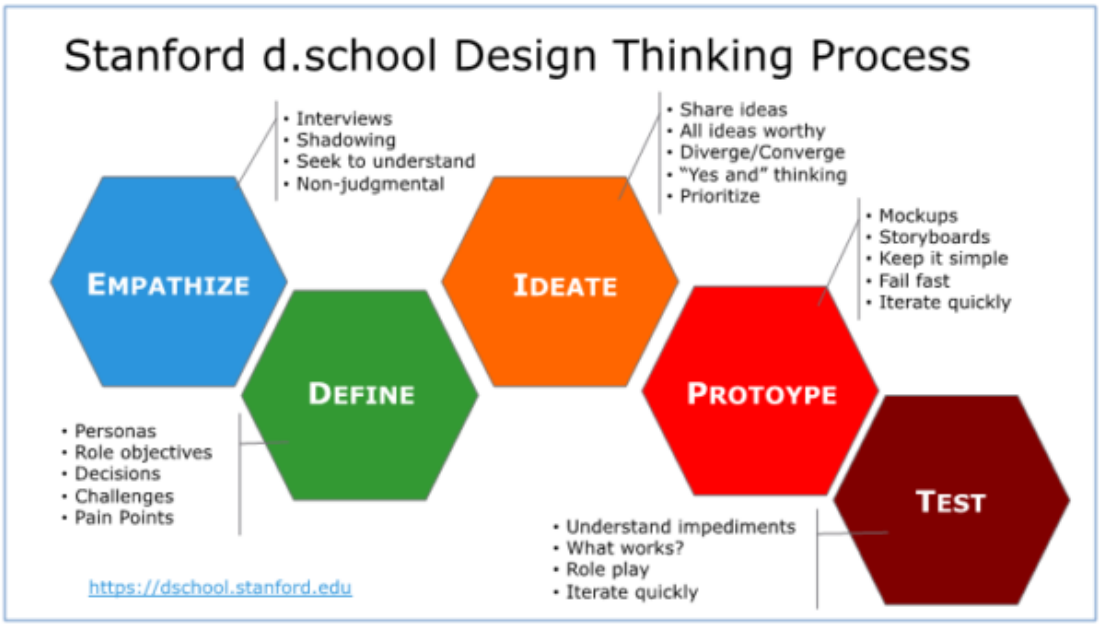
JUAN ANTONIO GARCIA MUELAS

Planteamiento y diseño de la solución.

Hay distintas formas de aproximarse a esta cuestión, dependiendo de la metodología.

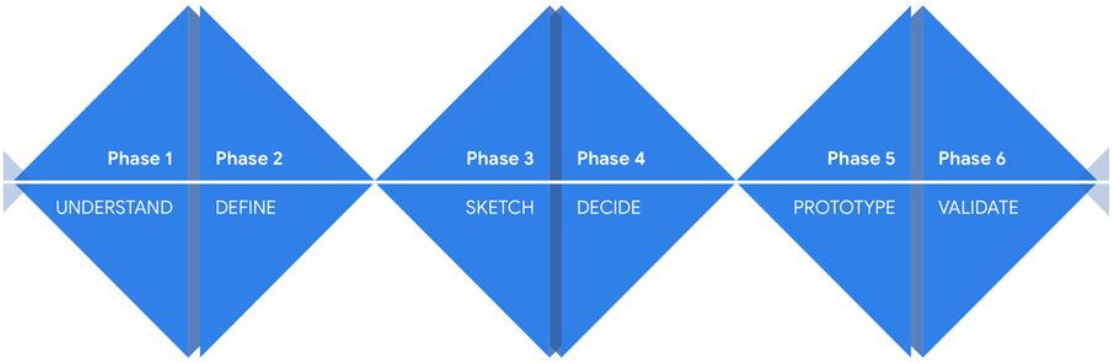
Una de esas metodologías es el **"Design Thinking", o Pensamiento de Diseño**, y traslada al **ámbito general** de proyectos el proceso que se sigue en el diseño de productos comerciales en la industria.

Emphatize, Define, Ideate, Prototype y test.



Google cuenta con "Design Sprint":

Understand, Define, Sketch, Decide, Prototype y Validate.



"Agile", como también se conoce este conjunto de metodologías, es un proceso mediante el cual un **equipo puede gestionar un proyecto dividiéndolo en varias etapas e involucrando** la **colaboración** constante de las **partes interesadas y una mejora e iteración continuas** en cada etapa.

En Agile, el producto se va adaptando a las necesidades del cliente en sucesivas iteraciones del proyecto.

En **Scrum** se realizan **entregas parciales y regulares del producto final**, priorizadas por el beneficio que aportan al receptor del proyecto.

Cada iteración en un proyecto gestionado con Scrum tiene que proporcionar un resultado completo.

La metodología Scrum utiliza Ciclos temporales cortos.

Kanban es un **método de gestión del flujo de trabajo (representado en tableros)** que ayuda a las organizaciones a gestionar y mejorar los sistemas de trabajo.

Todas ellas se caracterizan por identificar estos elementos:

- ✓ **Objetivos:**
  - Qué se **necesita o** qué **problemas** hay que resolver.
  - Qué **otras cosas** serían **deseables** y ayudan al objetivo principal.
- ✓ **Indicadores clave:**
  - Qué **variables** problema dan una **pista sobre si se está resolviendo** según su valor.
  - Qué **métricas** necesitamos **para calcular** cada indicador.
- ✓ **Tareas a realizar:**
  - Qué **trabajo** hay que hacer para **alcanzar los objetivos**.
  - Qué **división de trabajo** es la **que permite ejecutarlo** de forma **más eficaz**.

Fuentes habituales de datos

1.- Informes de plataformas, empresas y consultoras

En casos de **sectores muy concretos,** es **bastante común** que alguna empresa del sector o una consultora **haya realizado ya un estudio y publique datos** al respecto.

2.- Datos abiertos

Las **organizaciones y la administración elaboran sus propios estudios** de cara a desarrollar su actividad en base a la realidad. Podemos encontrar **portales de "open data".**

3.- Buscador Google de conjuntos de datos

Google, ha lanzado una interfaz orientada a buscar y encontrar datasets. https://datasetsearch.research.google.com/

4.- Datasets en Kaggle

El catálogo que más cantidad de conjuntos de datos interesantes concentra actualmente, es la plataforma de competiciones de machine learning Kaggle.com.

La plataforma ofrece un entorno con todo lo necesario para crear y ejecutar notebooks.

5.- Data.world

Y otro repositorio que contiene más de 130.000 datasets abiertos es data.world con datos que a veces es imposible encontrar en una búsqueda general y en esta web sí se encuentran.

6.- UCI

Una web que tradicionalmente contiene datasets bien preparados y con datos veraces, es la de la Universidad de California UCI, en su Machine Learning Repository.

Otras formas de obtener los datos

1.- Web Scrapping

Cuando los datos que necesitamos no están disponibles para descarga, pero son públicos en una página web. Web scraping o **raspado web**, es una técnica utilizada para **extraer información de sitios web, a través de un script** que simula la navegación de un humano.

2.- Encuestas y formularios

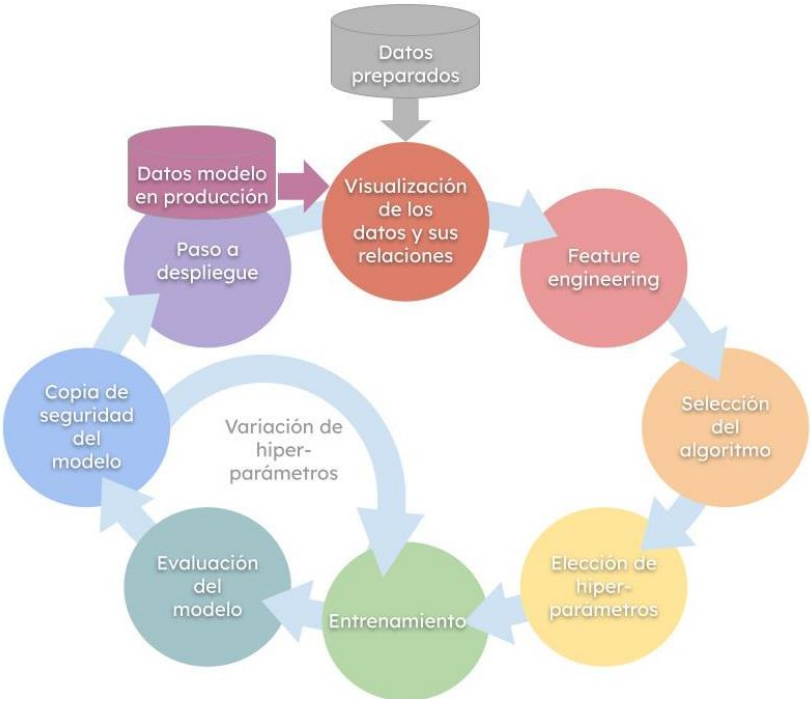
Es la **técnica más utilizada** para **datos** muy **concretos** de una **casuística particular**.

3.- Bases de datos propias

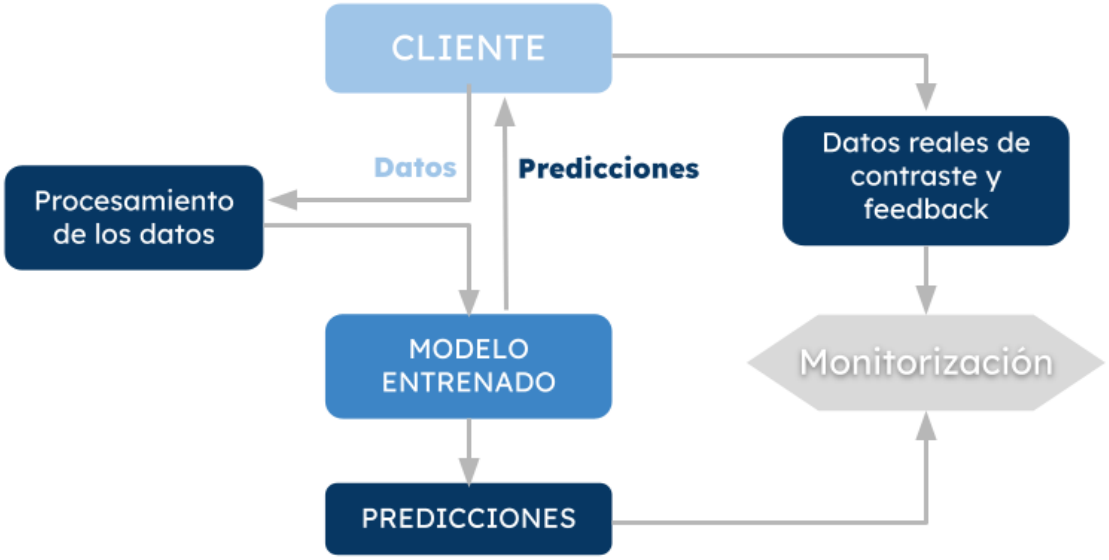
En las organizaciones que quieren implantar nuevos procesos de ciencia de datos, recurren a las propias bases de datos mediante estructuras “Data Lakes”

Diseño y preparación del modelo.

1. **Datos preparados**: partimos de la situación en la que ya contamos con un **dataset limpio y fiable**.
2. **Visualización de los datos y sus relaciones:** a través de representaciones estadísticas, gráficas de pares de variables, matrices de correlación y grafos, obtenemos una **primera impresión del comportamiento** de los datos, y empezamos a tomar decisiones.
3. **Feature engineering:** aplicamos **transformaciones** a las variables **o creamos variables** nuevas a partir de algunas **para obtener mejores resultados** en el entrenamiento.
4. **Selección del algoritmo:** dependiendo del tipo de problema, elegimos el algoritmo adecuado entre los vistos en anteriores unidades.
5. **Elección de hiperparámetros:** fijamos los parámetros del modelo y de su entrenamiento.
6. **Entrenamiento.**
7. **Evaluación del modelo**: lo ponemos a prueba con datos de entrada reservados (test). En el caso de deep learning, si hemos estado utilizando datos de validación, podemos saltarnos esta parte.
8. **Copia de seguridad del modelo**: es más que **conveniente ir guardando el archivo de los distintos modelos que se van obteniendo.** La aleatoriedad presente en cada parte del proceso puede generar, a veces, un modelo especialmente preciso que no se vuelve a repetir por más que repitamos el entrenamiento.
9. **Paso a despliegue:** implementación del modelo en una aplicación de negocio o para usuario final.
10. **Datos generados por el sistema en producción:** conviene repetir el proceso cada cierto tiempo, utilizando un dataset generado con los últimos datos que hayan pasado por el modelo y su correspondiente resultado.



Despliegue del modelo y pruebas.



Es recomendable repetir el proceso de entrenamiento con datos reales nuevos cada cierto tiempo para evitar que el modelo se desactualice y no haga buenas predicciones.

Cuando el modelo ya está entrenado, utilizarás una de estas dos opciones:

- ✓ Los **proveedores cloud ofrecen** entornos **"serverless"** o sin servidor, en los que se configura la aplicación o proyecto con una serie de recursos conectados (base de datos, almacenamiento, etc) y la infraestructura **escala según la demanda** de uso de forma automatizada.
- ✓ Se configura el proyecto en un **servidor privado o,** en el caso de proveedores cloud, un **servidor virtual privado.**