

Planificación y estructura de un proyecto de inteligencia artificial.



Caso práctico



[LookStudio \(CC BY-SA\)](#)

La empresa Pick&Deliver se dedica a almacenaje, logística y entregas de pedidos de empresas de comercio electrónico. En solo un año, han pasado de ser una empresa muy analógica a ser pioneros en Transformación Digital y uso de Inteligencia Artificial para sus procesos internos.

El equipo de innovación tecnológica de la empresa es bastante joven y con ganas de probar, equivocarse todo lo que haga falta y aprender todo lo posible en el proceso. Gracias a esto, han logrado implementar un gran número de mejoras y la empresa está entre las 10 primeras del sector en reputación y facturación.

Lorena es de las más entusiastas del equipo, y siempre está proponiendo ideas. Esta vez, llega con una idea inesperada: "¿Qué os parece si nos apuntamos a esta iniciativa?" y les muestra una web de una ONG que organiza un hackathon para equipos profesionales con el objetivo de ayudar a una región que ha sufrido una fuerte crisis económica a recuperar su tejido comercial y de negocios. "Con todo lo que hemos aprendido y probado nosotros aquí, seguro que les podemos ayudar"

Todos se contagian de la motivación de Lorena. Ayudar y crear impacto real en la sociedad es un desafío importante, pero es de las cosas que más sentido dan al desarrollo de la tecnología.

"¿Cuándo empezamos?" Dice Miguel sonriendo.

En esta unidad vamos a seguir el proceso que se recorre cuando se lanza un nuevo proyecto de inteligencia artificial en una organización. Trabajaremos sobre los elementos implicados en la planificación de un proyecto y en las características de cada fase de su desarrollo. En concreto, revisaremos:

- ✓ Identificación de necesidades y diseño de la solución.
- ✓ Búsqueda de los recursos necesarios.
- ✓ Proceso de diseño, entrenamiento y evaluación del modelo de aprendizaje automático.
- ✓ Despliegue y puesta en marcha de la solución.
- ✓ Seguimiento del proyecto y propuesta de nuevas actualizaciones.

Para ello, vamos a apoyarnos en conceptos generales que afectan a todo tipo de proyectos, pero también particularizaremos muchos condicionantes al ámbito concreto del desarrollo e implantación de soluciones de inteligencia artificial.



[Ministerio de Educación y Formación Profesional](#) (Dominio público)

Materiales formativos de FP Online propiedad del Ministerio de Educación y Formación Profesional.

[Aviso Legal](#)

1.- Planteamiento y diseño de la solución.



Caso práctico



[LookStudio \(CC BY-SA\)](#)

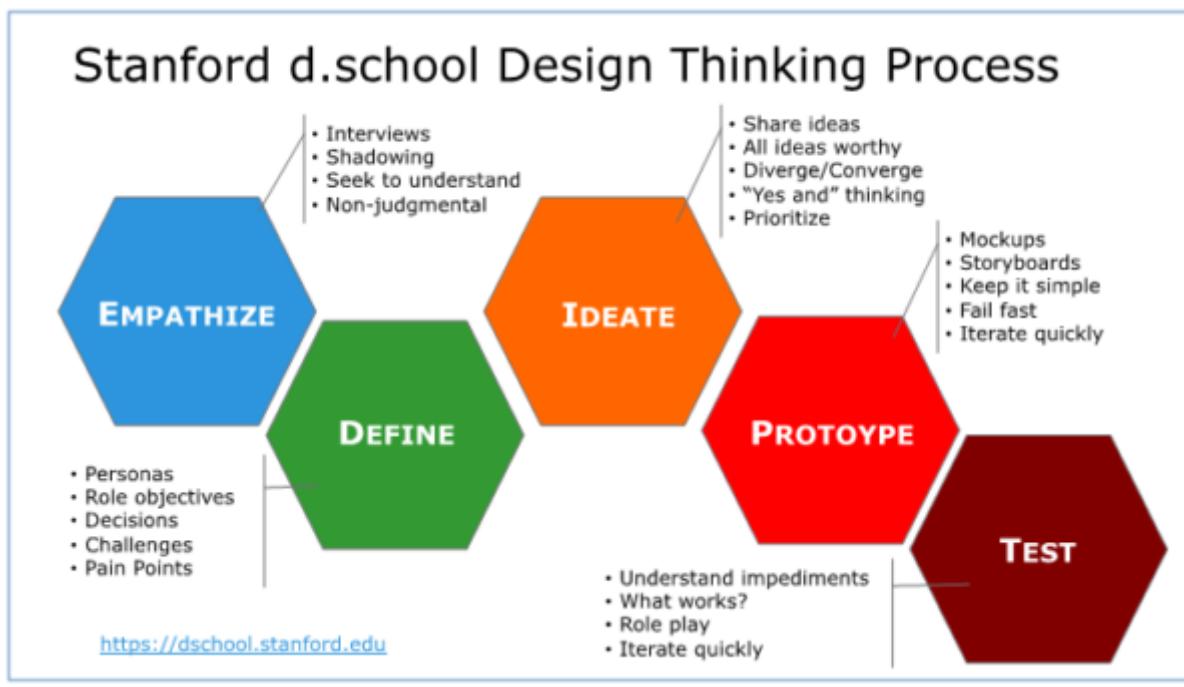
Lorena y Miguel revisan a fondo la información sobre los retos planteados para el hackathon solidario al que se han apuntado. Todos parecen muy interesantes, y cuesta decidirse por uno de ellos. "¿Qué te parece éste de ayudar a automatizar una planta de tratamiento y envasado de frutas y hortalizas? Parece que tienen varios problemas con los que podemos ayudar a la propia cooperativa y a la comunidad que vive de ello" Dice Miguel fijándose en un proyecto que ve bastante inspirador.

"Sí, yo también me estaba fijando en ese. ¿Seremos capaces de ayudarles desde aquí?" Contesta Lorena.

En los días siguientes, van cumpliendo con los requisitos de inscripción y entran en contacto con varios voluntarios de la ONG que les van dando indicaciones. Finalmente, se ponen en contacto con el representante de la cooperativa y revisan con él los problemas que tienen que resolver.

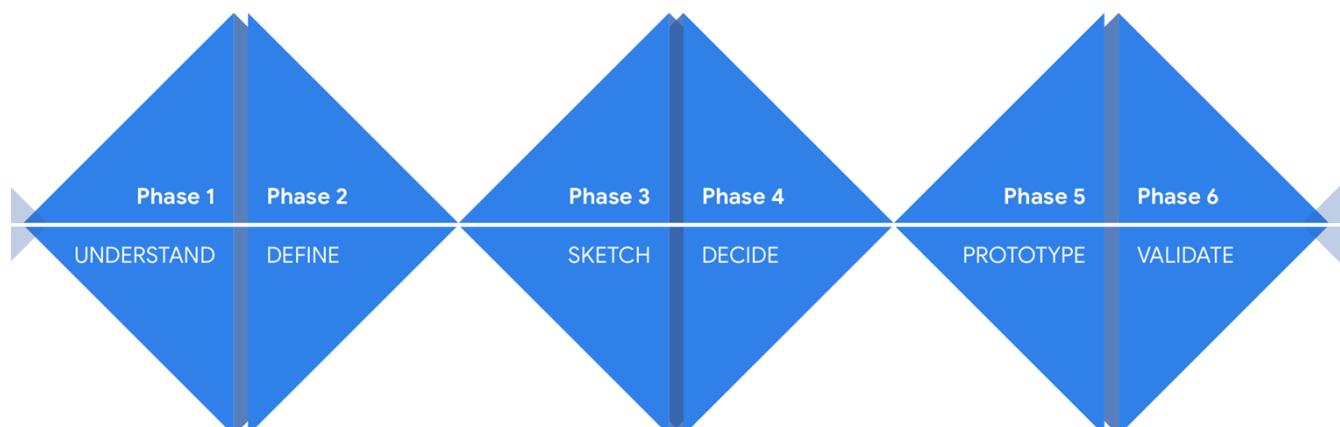
Existen muchas metodologías que ayudan a plantear la planificación de un proyecto de forma efectiva. En todos ellos, se parte de un análisis previo de la situación y de la identificación de una serie de necesidades o problemas. Hay distintas formas de aproximarse a esta cuestión, dependiendo de la metodología.

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.



[dschool.stanford.edu \(CC BY\)](#)

Google también cuenta con varias metodologías entre las que, para proyectos, destaca la de Design Sprint:



[Google \(CC BY-SA\)](#)

Estas dos metodologías ponen especial foco en la detección del problema y en identificar muy bien su anatomía para aportar ideas de solución que realmente se ajusten a las

necesidades reales. La ideación de la solución a partir de lluvia de ideas, procesos divergentes-convergentes y consolidación de ideas, nos conduce a la creación de un prototipo. En este punto, podemos aplicar metodologías más enfocadas en la creación de ese prototipo.

En el ámbito del desarrollo de software, las denominadas metodologías ágiles, constituyen el marco de trabajo elegido por las empresas más innovadoras y con mejor desempeño. Desde que se popularizó el "Manifesto for Agile Software Development" en 2001, las dos metodologías con más éxito han sido Scrum y Kanban.

"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. La metodología Agile comienza con la descripción de los clientes sobre cómo se utilizará el producto final y qué problema resolverá. Esto aclara las expectativas del cliente al equipo del proyecto. Una vez que comienza el trabajo, los equipos realizan un ciclo de un proceso de planificación, ejecución y evaluación, que podría cambiar la entrega final para adaptarse mejor a las necesidades del cliente. La colaboración continua es clave, tanto entre los miembros del equipo como con las partes interesadas del proyecto, para tomar decisiones plenamente informadas.



[katemangostar \(CC BY-SA\)](#)

En **Scrum** se realizan entregas parciales y regulares del producto final, priorizadas por el beneficio que aportan al receptor del proyecto. Por ello, Scrum está especialmente indicado para proyectos en entornos complejos, donde se necesita obtener resultados pronto, donde los requisitos son cambiantes o poco definidos, donde la innovación, la competitividad, la flexibilidad y la productividad son fundamentales. En Scrum un proyecto se ejecuta en ciclos temporales cortos y de duración fija (iteraciones que normalmente son de 2 semanas, aunque en algunos equipos son de 3 y hasta 4 semanas, límite máximo de feedback de producto real y reflexión). Cada iteración tiene que proporcionar un resultado completo, un incremento de producto final que sea susceptible de ser entregado con el mínimo esfuerzo al cliente cuando lo solicite.

Kanban es un método de gestión del flujo de trabajo que ayuda a las organizaciones a gestionar y mejorar los sistemas de trabajo. El trabajo se representa en tableros Kanban, lo que te permite optimizar la entrega de trabajo a través de múltiples equipos y manejar, incluso los proyectos más complejos en un solo entorno. Fue desarrollado y aplicado por primera vez por Toyota como sistema de programación para la fabricación JIT ("Just In Time": "justo a tiempo"). Este enfoque representa un Sistema con un comportamiento "pull"; Esto significa que la producción se basa en la demanda de los clientes, en lugar de la práctica habitual de, producir bienes y llevarlos al mercado.

En casi todas las metodologías de proyectos necesitan identificar estos elementos para un correcto diseño de la solución:

- ✓ Objetivos:
 - ⇒ Qué se necesita o qué problemas hay que resolver.
 - ⇒ Qué otras cosas serían deseables y ayudan al objetivo principal.
- ✓ Indicadores clave:
 - ⇒ Qué variables del problema nos 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.

De esta manera, al inicio de un proyecto, es conveniente tener una o más sesiones en las que pensar, idear y decidir. Si no, se puede dar la situación de trabajar duro desarrollando un sistema o modelo para darse cuenta demasiado tarde de que no sirve realmente para lo que se necesita.



Para saber más

Las metodologías vistas en esta sección suelen requerir formación y práctica para poder ser aplicadas con éxito en las empresas y equipos. Si quieres trabajar en empresas de desarrollo de software o vas a ser técnico en una startup, es muy recomendable hacer un curso de Scrum y documentarte todo lo que puedas de las metodologías que sepas que se utilizan en la empresa.

En todo caso, puedes saber más sobre Scrum en [su propia web](#).

2.- Adquisición y tratamiento de los datos.



Caso práctico



[LookStudio \(CC BY-SA\)](#)

Una vez el equipo se ha puesto de acuerdo con cuál va a ser la solución en la que van a trabajar, y empiezan a organizarse, se dan cuenta de que la parte fundamental del proyecto son los datos, pues, sin ellos, no van a ir a ninguna parte.

"Tenemos varias opciones a la hora de obtener suficientes datos, pero empecemos por la más obvia: pedírselos a los propios miembros de la cooperativa" Dice Lorena, que disfruta bastante con la parte de gestión de los datos.



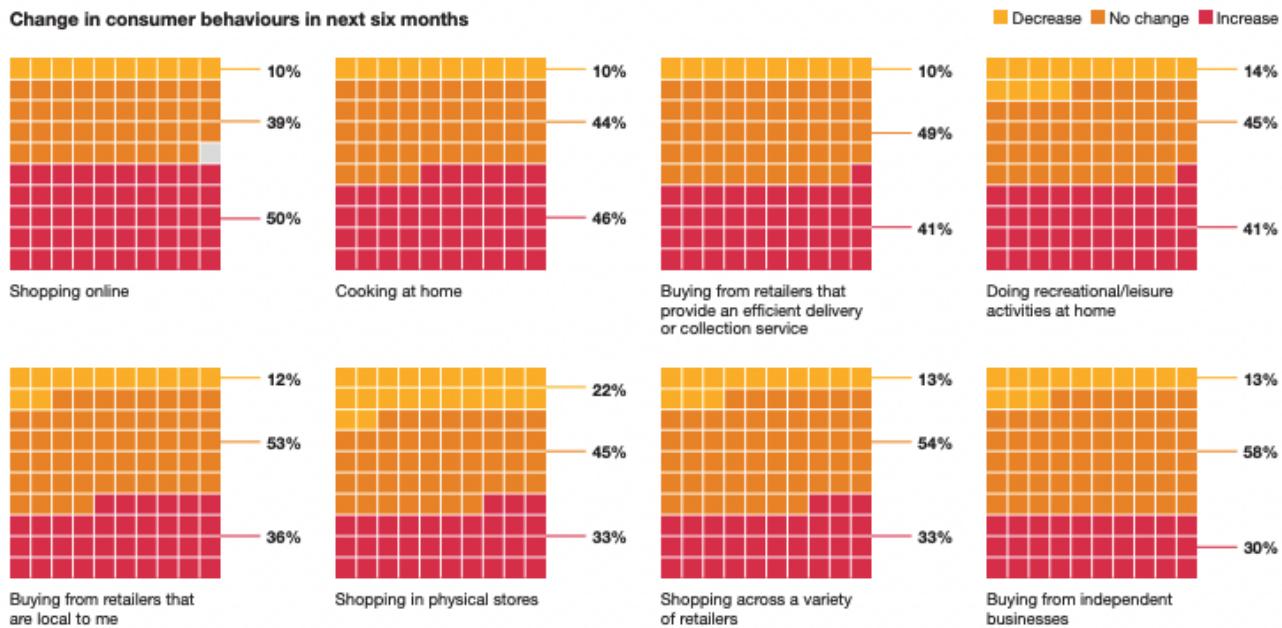
[vectorsmarket15 \(CC BY-SA\)](#)

En todo proyecto de inteligencia artificial o ciencia de datos, la parte de los datos suele representar el 80% de la carga de trabajo del proyecto. Y, análogamente, representa un ratio similar en cuanto a la calidad de la solución final. Es de vital importancia trabajar con datos completos, veraces y sin sesgos. Ya hemos visto en la unidad 6 que si queremos una solución suficientemente general y que tenga un buen comportamiento con nuevos datos de entrada, necesitaremos un conjunto de datos de partida lo suficientemente grande.

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. Estos estudios suelen presentar los datos ya procesados en estadísticas e indicadores finales, pero pueden utilizarse para completar un dataset en el que nos falte alguna de las variables o también se pueden solicitar.



[PWC \(CC BY-SA\)](#)

También es posible encargarle a una consultora la confección de un dataset con los campos de interés, pero este tipo de proyectos son caros y requieren de un presupuesto concreto antes de lanzar el proyecto.

2.- Datos abiertos

Las organizaciones y la administración elaboran sus propios estudios de cara a desarrollar su actividad en base a la realidad. En los últimos años se ha popularizado la opción de compartir esos datos de forma abierta y gratuita para que puedan ser aprovechados por investigadores y quién los quiera consultar. Podemos encontrar portales de "open data"

Catálogo de datos.



[datos.gob.es \(Dominio público\)](#)

3.- Buscador Google de conjuntos de datos

Google, que también se beneficia de que la comunidad en torno a la ciencia de datos pueda crecer y descubrir nuevas formas de predecir y dar soluciones utilizando la gran cantidad de datos que se generan diariamente, ha lanzado una interfaz especialmente orientada a buscar y encontrar datasets disponibles en internet. La herramienta solo rastrea la red y muestra dónde se encuentran esos datasets. Hay que pulsar en el enlace y ya en la correspondiente web nos encontraremos el conjunto de datos en cuestión, que no siempre es gratuito.

The screenshot shows a Google search results page with the query "retail". At the top, there are several filter buttons: "Última actualización", "Formato de descarga", "Derechos de uso", "Tema", and "Gratis". Below the search bar, it says "Se han encontrado más de 100 conjuntos de datos". The first result is a link to "List of 38M Retail companies worldwide" from "kaggle". The result includes details like the source ("Online Retail II Data Set from ML Repository"), URL ("www.kaggle.com"), file type ("zip"), and last update ("Jun 14, 2021"). The second result is a link to "China Retail Sales YoY" from "tradingeconomics.com". It includes details like the source ("tradingeconomics.com"), URL ("tr.tradingeconomics.com"), file type ("+12más"), and last update ("Aug 15, 2022"). The third result is a link to "List of 38M Retail companies worldwide" from "datarade.ai". It includes details like the source ("datarade.ai"), URL ("Última actualización: Apr 28, 2021"), file type (".json, .csv, .xls, .txt"), and last update ("Apr 28, 2021"). Below the results, there is a note about the data being held in-house by BoldData, mentioning 300 million companies in 150+ countries, and a link to "Google Datasearch (Dominio público)".

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.

Tanto las empresas que lanzaban sus competiciones como otros usuarios que han ido donando datasets, han hecho crecer este banco de datos a nivel internacional, que proporciona materia prima para aprender, practicar y también entrenar nuevos modelos. No siempre se encuentra exactamente lo que se está buscando, pero puede servir para hacer un modelo de aproximación o prototipo para luego ir afinando las variables que realmente hacen falta para conseguir un buen modelo.

Datasets

Explore, analyze, and share quality data. [Learn more](#) about data types, creating, and collaborating.

+ New Dataset

Your Work



Search datasets

Filters

All datasets

Computer Science

Education

Classification

Computer Vision

NLP

Data Visualization

Pre-Trained Model

Trending Datasets

See All

Yoga Posture Dataset
Mrinal Tyagi · Updated 11 hours ago
Usability 8.8 · 468 MB
2759 Files (other, JSON)

14

iZMİR-MANİSA İlceleri Saatlik Hava Durumu
huseyincotel · Updated a day ago
Usability 9.7 · 9 MB
1 File (CSV)

17

Top 50 Bestselling Novels 2009-2021 of Amazon
Zhi Weng Lim · Updated 7 days ago
Usability 10.0 · 37 kB
2 Files (CSV)

15

Agricultural crops image classification
Md Waquar Azam · Updated 11 days ago
Usability 8.8 · 83 MB
829 Files (other)

20

[Kaggle.com](#) (Dominio público)

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.

The screenshot shows the data.world interface for the 'Untitled project'. On the left, there's a sidebar with a file tree: 'Project directory' (with '+ Add'), 'Home', 'Project summary', 'Data dictionary', 'PROJECT FILES' (empty), 'CONNECTED DATASETS' (with 'Coffee Chain (1)'), and 'QUERIES' (empty). The main area shows a table titled 'Coffee Chain.txt' with 22 rows and columns like '# area_code', '# cogs', '# difference_between_actual_and_target_profit', 'date', '# inventory', '# margin', and 'market'. To the right of the table are sections for 'ABOUT THIS FILE' (Last Updated 2 years ago, Owner Haiyun YU, Created 2 years ago, Size 151.6 KB) and 'TABLE COLUMNS' (descriptions for each column). At the bottom, there's a footer with 'data.world' and '(Dominio público)'.

[data.world](#) (Dominio público)

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](#).



Check out the beta version of the new UCI Machine Learning Repository we are currently testing! Contact us if you have any issues, questions, or concerns. Click here to try out the new site. X

Browse Through: 622 Data Sets

	Name	Data Types	Default Task	Attribute Types	# Instances	# Attributes	Year
 Abalone	Multivariate	Classification	Categorical, Integer, Real	4177	8		1995
 Adult	Multivariate	Classification	Categorical, Integer	48842	14		1996
 Annealing	Multivariate	Classification	Categorical, Integer, Real	798	38		
 Anonymous Microsoft Web Data		Recommender-Systems	Categorical	37711	294		1998
 Arrhythmia	Multivariate	Classification	Categorical, Integer, Real	452	279		1998
 Artificial Characters	Multivariate	Classification	Categorical, Integer, Real	6000	7		1992
 Audiology.(Original)	Multivariate	Classification	Categorical	226			1987
 Audiology.(Standardized)	Multivariate	Classification	Categorical	226	69		1992
 Auto MPG	Multivariate	Regression	Categorical, Real	398	8		1993

[UCI](#) (Dominio público)

Otras formas de obtener los datos

1.- Web Scrapping

En algunas ocasiones, 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. Es una técnica un poco polémica porque a veces podría ir en contra de los términos de uso de algunos sitios web. En nuestro caso, no suele haber problema porque no vamos a hacer un aprovechamiento comercial de los datos tal cual se extraen, sino que los vamos a utilizar para otro propósito.

2.- Encuestas y formularios

Es la técnica más utilizada cuando requerimos datos muy concretos de una casuística particular. Se suelen lanzar a través de la propia internet, utilizando sorteos o promociones como estímulo para la participación.

3.- Bases de datos propias

En las organizaciones donde se quieren implantar nuevos procesos basados en ciencia de datos, solo hay que recurrir a las propias bases de datos y trabajar sobre éstos. En este

ámbito, se suele recurrir a infraestructuras denominadas "Data lakes" que permiten gestionar los datos de forma rápida y mejorar la accesibilidad a los datos desde otros nodos del sistema. En los siguientes módulos del curso vas a ver todo esto con más detalle.

Sobre el tratamiento de los datos

Una vez que tenemos el conjunto de datos con el que queremos trabajar, es necesario explorarlo y prepararlo para su paso por el modelo a la hora de entrenarlo. En los **siguientes módulos** se va a ver en más detalle las técnicas para la gestión de los datos, y en la unidad 3, cuando viste las librerías de Python implicadas en procesos de machine learning, viste como trabajar los datos de forma más fina justo antes de utilizarlos en el entrenamiento del modelo utilizando funciones del paquete **Pandas**. Repasa ese tema siempre que tengas que explorar un dataset nuevo.



Autoevaluación

El buscador de datasets que ha lanzado Google se llama:

- Dataset Search
- Databuscador
- Google datasets flow

Opción correcta

Google lanzó el buscado especializado Dataset Search para el ámbito de la ciencia de datos

Google lanzó el buscado especializado Dataset Search para el ámbito de la ciencia de datos

Solución

1. Opción correcta
2. Incorrecto
3. Incorrecto



3.- Diseño y preparación del modelo.



Caso práctico



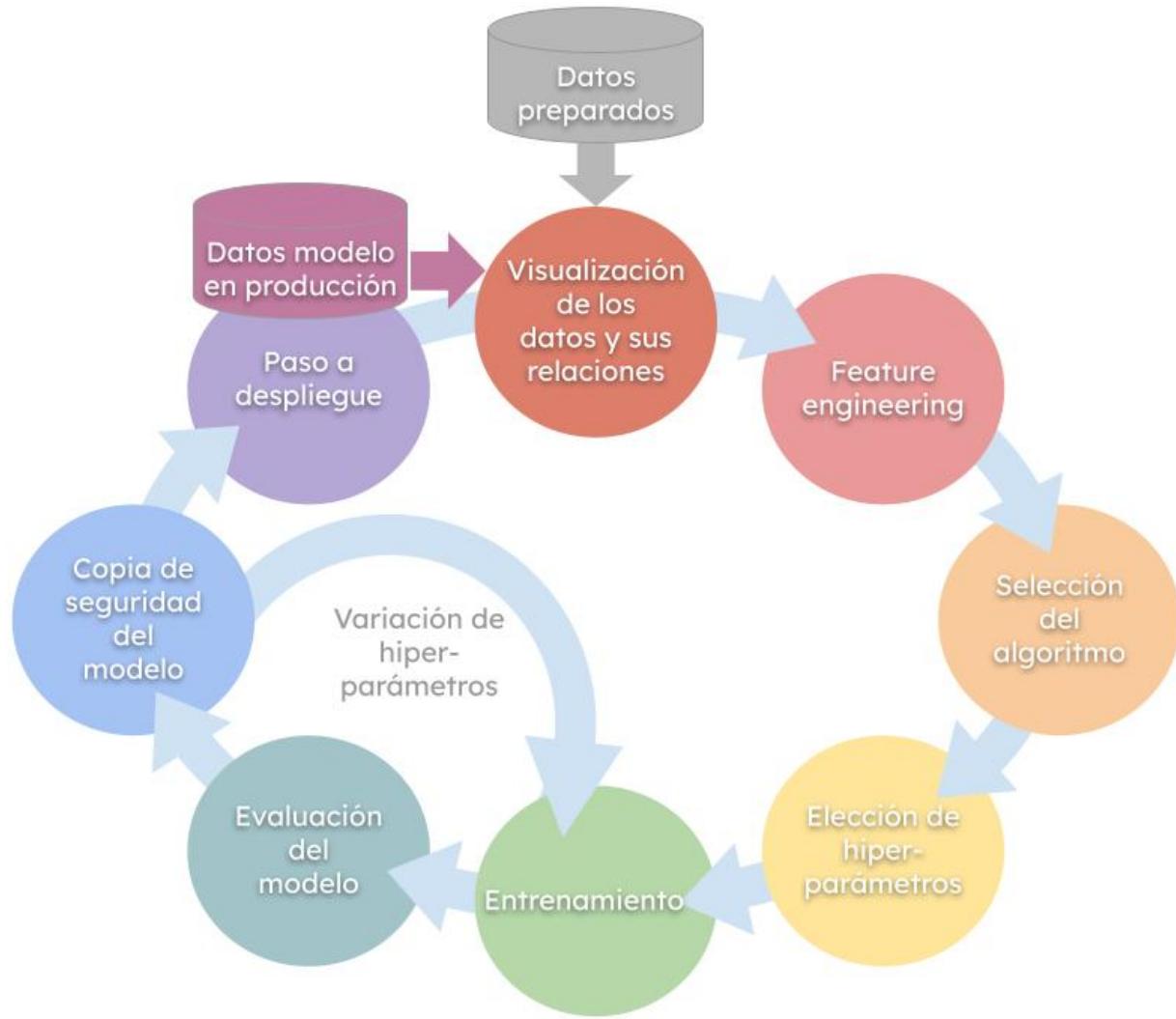
[LookStudio \(CC BY-SA\)](#)

La cooperativa a la que están ayudando, está facilitando al equipo de Miguel y Lorena los datos que van necesitando según el planteamiento inicial que han hecho del proyecto. Ahora que ya está definida la arquitectura esquemática de la solución, toca ponerse con la arquitectura interna, con las tripas del proyecto. Deben seleccionar el tipo de algoritmo más adecuado, programarlo con los parámetros que inicialmente estiman más convenientes y entrenarlo.

"¡Esto va tomando forma!" Dice Miguel suspirando "Aunque todavía queda mucho trabajo por hacer"

"Estoy en contacto con un par de amigos que se han ofrecido para echarnos una mano en la optimización de hiperparámetros" Afirma Lorena contenta "Voy a crear un documento colaborativo con una tabla en la que ir registrando cada equipo los resultados con distintas configuraciones de red"

Esta etapa viene conformada por las tareas que hemos estado viendo durante las unidades 4, 5, 6 y 7. Si te has estado fijando en el flujo de trabajo de los ejemplos, habrás detectado ya las fases que se suelen dar en casi todos los procesos de Machine Learning. Mira la siguiente infografía y vuelve a los ejemplos de las unidades anteriores para identificar cada parte.



Carmen Bartolomé ([CC0](#))

- 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.



Autoevaluación

Como los procesos de machine learning utilizan configuraciones de partida aleatorias, puede darse el caso de alcanzar modelos con una leve diferencia de precisión, aun habiendo configurado los mismos hiperparámetros y usado los mismos datos.

- Verdadero Falso

Verdadero

Ciertamente, el grado de aleatoriedad de los procesos actuales de aprendizaje automático favorecen pequeñas diferencias entre modelos en igualdad de condiciones de programación.

4.- Despliegue del modelo y pruebas.



Caso práctico

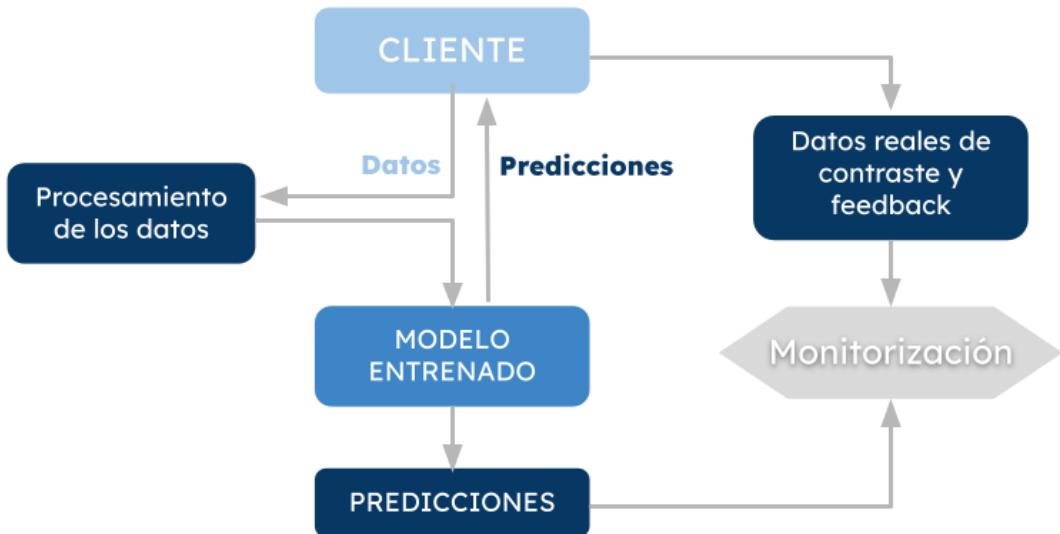


[LookStudio \(CC BY-SA\)](#)

El equipo ha estado empleándose a fondo durante el fin de semana, en la fase de desarrollo del hackathon solidario, para tener listo el proyecto para su despliegue y pasar a la fase de pruebas.

"¿Dónde lo desplegamos?" le pregunta Miguel a Andrew, que se conoce bastante bien las opciones de los diferentes proveedores de infraestructura en la nube.

"En la empresa utilizamos servicios de varios proveedores cloud, y cada uno tiene sus ventajas. Voy a pedirle a cada uno que nos de infraestructura y servicios sin coste para este proyecto que, al fin y al cabo, es altruista y con fines sociales. El que quiera unirse a la causa y darnos el servicio en condiciones especiales, será el más indicado, ¿no os parece?" responde Andrew.



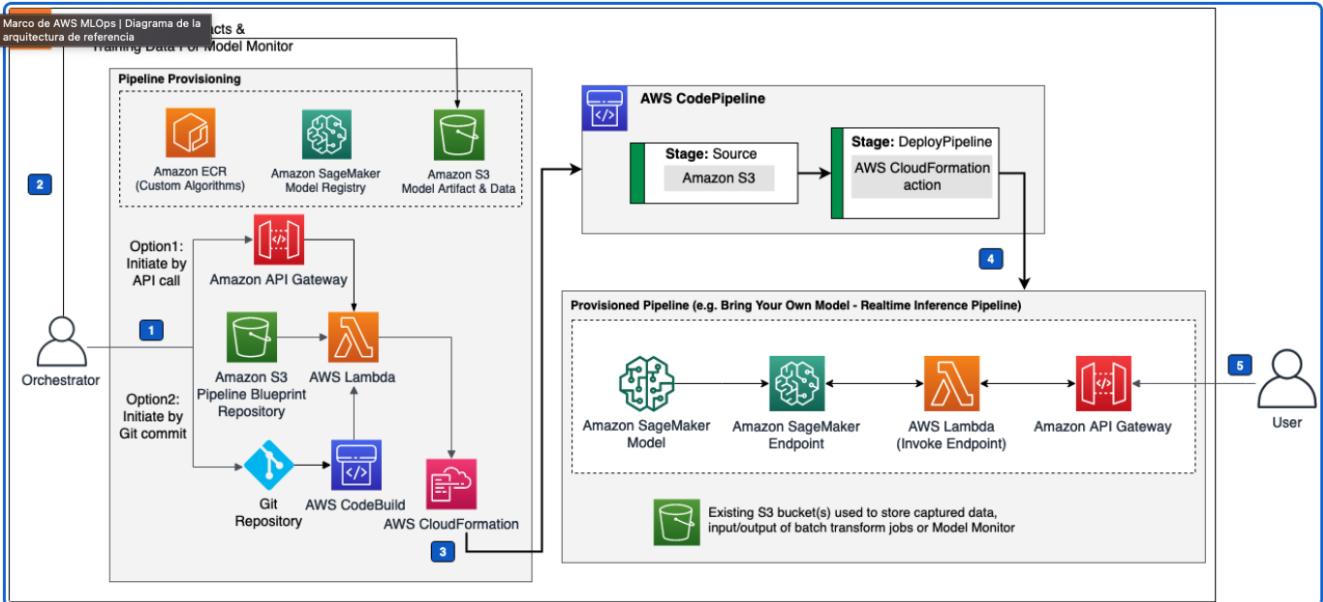
Carmen Bartolomé ([CC0](#))

Cuando el modelo ya está entrenado y hemos repetido el proceso suficientes veces hasta tener el mejor modelo posible, pasamos a su implementación en la aplicación final. En general, 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.

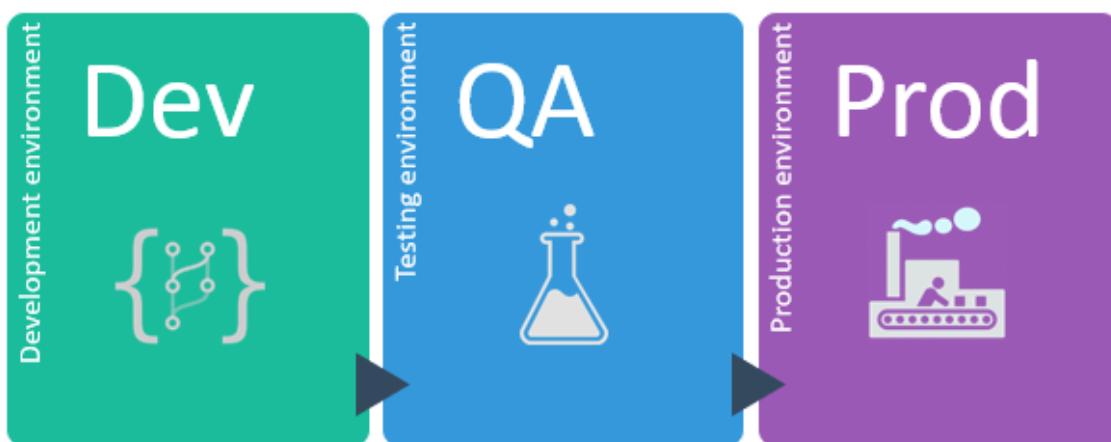
No es recomendable recurrir a otro tipo de hosting de menor capacidad en el caso de modelos de machine learning.

Como vimos en las unidades 1 y 2, los proveedores cloud actuales ofrecen también la posibilidad de conectar el modelo que hemos estado trabajando en un entorno tipo "notebook" directamente a la infraestructura de explotación.



[AWS](#) (Dominio público)

En todo caso, nuestra recomendación es que tengas dos entornos o servidores, uno de **desarrollo** y otro de **producción**. En el entorno de desarrollo harás las primeras pruebas en despliegue, sin que trasciendan al cliente. Después, se monta el proyecto en el entorno de producción preparado con más recursos y mantenimiento más cuidadoso. Pero el entorno de desarrollo nos seguirá sirviendo para probar cambios, actualizaciones y buscar "bugs" sin interferir con la aplicación en producción. También se suele clonar el proyecto en un entorno de "Quality Assurance" o QA para pruebas más extremas, e incluso un entorno tipo Research, para experimentos y pruebas de concepto.



[hypertchnologyweb.com \(CC BY\)](#)