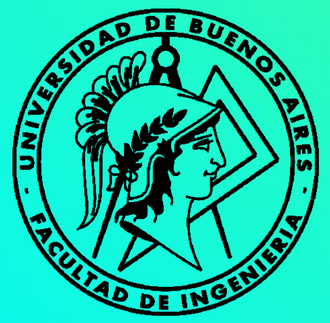


INTRODUCCIÓN

APRENDIZAJE DE MAQUINA I - CEIA - FIUBA

Antonio Zarauz Moreno

Datos elaborados por BCP para uso Interno

INTRODUCCIÓN

- Materia de 8 clases teórico-prácticas
- Clases con diapositivas y desarrollo en notebooks
- Estructuras de las clases:
  - 10 minutos repaso de clase anterior
  - 3 bloques de 50 minutos de clases teórico-practicas
  - 2 recreos de 10 minutos
  - Ejercicio de practica entre las clases. Sin entrega y evaluación.

Datos elaborados por IACP para uso interno

INTRODUCCIÓN

Aula virtual

- https://campusposgrado.fi.uba.ar/course/view.php?id=251

Repositorio de la materia:

- https://github.com/FIUBA-Posgrado-Inteligencia-Artificial/aprMaqI_CEIA

Consultas

- Foro de consulta en el aula virtual

Correo

Antonio Zarauz Moreno: hedrergudene@gmail.com

Datos elaborados por IACP para uso interno

EVALUACIÓN

- Trabajo práctico final que implementar y entregar 7 días posteriores del final de la cursada.
- Obligación de trabajar en grupo mínimo de 2 y máximo de 6. Excepciones se pueden hacer mediante un correcto justificativo.
- Implementación de algún modelo de Machine Learning en un set de datos a elección.
- Puede ser entregado en cualquier formato, preferentemente en notebook de ipython (formato ipynb). También puede entregarse mediante un documento en Google Colab.
- El trabajo no solo consiste en aplicar el modelo, sino el correcto manejo de datos, feature engineering y evaluación. Visualización es también una de las partes más importantes. Justificar elección de las herramientas y desarrollo teórico correcto. No olvidar usar referencias.

Datos elaborados por IACP para uso interno

DATASETS

¿No tiene algún set de datos?

- Kaggle: https://www.kaggle.com/datasets
- Awesome public datasets: https://github.com/awesomedata/awesome-public-datasets
- Real world fake data: https://sonsofhierarchies.com/real-world-fake-data/
- Maven analytics: https://www.mavenanalytics.io/data-playground

Sigues si encontrar un dataset, les dejamos algunos en el aula virtual.

Datos elaborados por IACP para uso interno

HERRAMIENTAS

Lenguaje de Programación

- Python >=3.10
- Anaconda/Pip/Poetry

Librerías

- Numpy, Pandas, SciPy, Statsmodels
- Matplotlib, Seaborn
- Scikit-Learn, PyTorch, XGBoost

Consola Interactiva de Python

- IPython
- Jupiter Notebook

Herramientas

- GitHub para repositorios

IDE Recomendados

- Visual Studio Code
- PyCharm Community Edition

PROGRAMA

Clase a clase

01 Introducción a Machine Learning. Definiciones. Ciclo de vida de un proyecto de Aprendizaje Automático. Metodología para construir modelos.

02 Aprendizaje supervisado. Clasificador KNN. Medidas de distancia. Encontrando a los vecinos. Métodos de Ajuste de los hiper-parámetros. Frameworks de búsqueda

03 Máquinas de vectores de soporte. Maximal Margin Classifier. Clasificador de Vector de soportes. Máquinas de vectores de soporte en regresión.

04 Arboles de decisión. Arboles de clasificación y arboles regresión.

Datos elaborados por IACP para uso interno

PROGRAMA

Clase a clase

05	Redes neuronales simples. Red feed-foward y Backpropagation.
06	Métodos de ensamble. Modelos que votan. Bagging. Bosques aleatorios. Boosting. XGBoost
07	Aprendizaje No Supervisado. Clustering. K-Means. Suma de Cuadrados Intracluster. Índice de la silueta. Fuerza de Predicción. Modelo de Mixtura Gaussiana. Reducción de dimensionalidad. Análisis de componentes principales.

Datos elaborados por IACP para uso interno

BIBLIOGRAFIA

- Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python - Peter Bruce (Ed. O'Reilly)
- The Elements of Statistical Learning - Trevor Hastie (Ed. Springer)
- An Introduction to Statistical Learning - Gareth James (Ed. Springer)
- Pattern Recognition And Machine Learning - Christopher Bishop (Ed. Springer)
- Deep Learning - Ian Goodfellow https://www.deeplearningbook.org/

Datos elaborados por IACP para uso técnico

APRENDIZAJE AUTOMÁTICO

Datos elaborados por BCP para uso Interno

APRENDIZAJE AUTOMÁTICO

Tal como se aprendió en Inteligencia Artificial, **Aprendizaje Automático** se entiende a:

Una computadora observa algunos datos, construye un modelo basado en estos datos, y usa el modelo como una hipótesis sobre el mundo y como una pieza de software que puede resolver problemas.

- Recibe un conjunto de datos y aprende por sí mismo.
- Reconoce patrones entre los datos y hace una predicción
- No requiere que una persona programe instrucciones

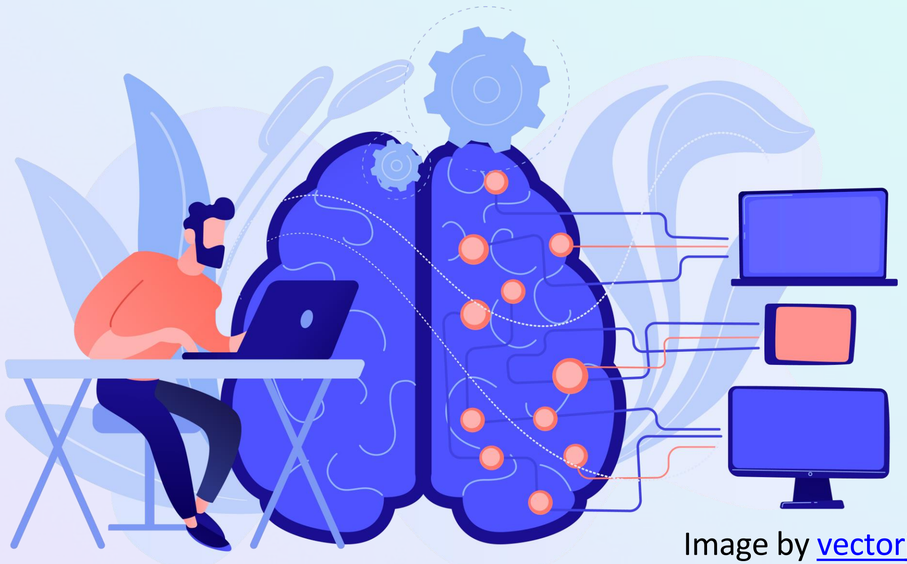


Image by vectorjuice

Datos elaborados por IICP para uso interno

¿CUÁNDO USARLO?

Datos elaborados por BCP para uso Interno

¿CUÁNDO USARLO?

Aprendizaje automático no es una herramienta mágica que pueda resolver todos los problemas. Incluso para los problemas que Aprendizaje Automático puede resolver, las soluciones pueden no ser las soluciones óptimas.

Antes de iniciar un proyecto de Aprendizaje Automático, debemos ver si es necesario o rentable:

- Para que un sistema de ML aprenda, debe haber algo de qué aprender. Debemos tener una buena cantidad de **Datos.** y de calidad (garbage in, garbage out)

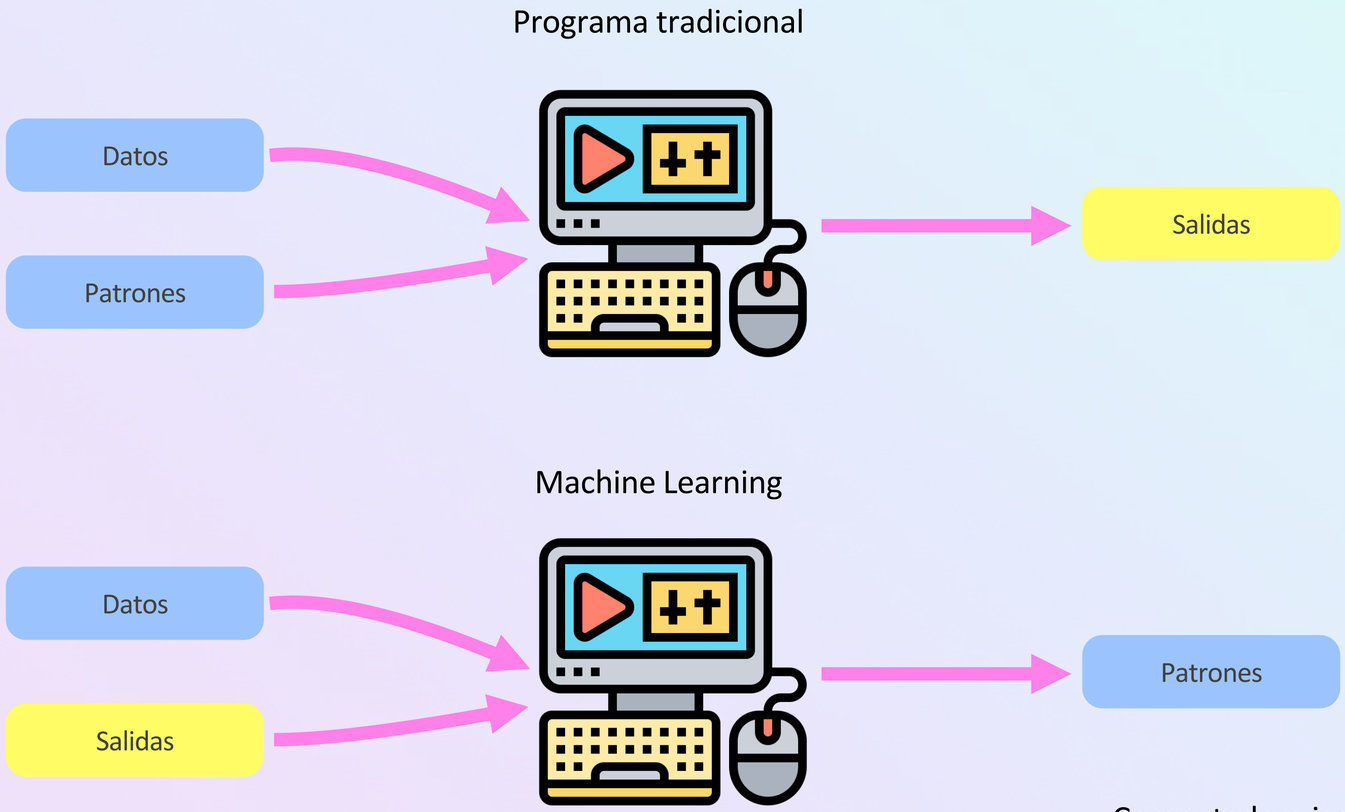
- Las soluciones de Aprendizaje Automático solo son útiles cuando hay **patrones que aprender**. Por ejemplo, no tiene sentido usar modelos para hacer un controlador PID.

  Es posible que no sea obvio si existe un patrón, o si existen patrones, el conjunto de datos o el algoritmo podrían no ser suficientes para capturarlos. *Por ejemplo, podría haber un patrón en cómo los tweets de Elon Musk afectan los precios de las criptomonedas*.

Datos elaborados por BCP para uso interno

¿CUÁNDO USARLO?

VS. PROGRAMACIÓN TRADICIONAL



Computadora iconos creados por Eucalyp - Flaticon

¿CUÁNDO USARLO?

- No solo se necesitan Datos, sino que estos deben **estar disponibles o que sea posible recopilarlos**. Sin datos y sin aprendizaje continuo, muchas empresas siguen un enfoque de *“fake-it- til-you make it”*, es decir, lanzar un producto que **sirva predicciones hechas por humanos**, en lugar de modelos de ML, con la esperanza de utilizar los datos generados para entrenar ML modelos.

Es importante que los datos esten disponibles porque se puede introducir sesgo

- Los modelos de ML hacen predicciones, por lo que solo pueden resolver **problemas que requieren respuestas predictivas**. El aprendizaje automático puede resultar especialmente atractivo cuando puede beneficiarse de una gran cantidad de predicciones baratas pero aproximadas. Por ejemplo, ¿cómo será el tiempo mañana? ¿Quién ganará el Super Bowl este año? ¿Qué película querrá ver un usuario a continuación?

- Los datos para lo que se usará en el futuro y los datos de entrenamiento **deben provenir de distribuciones similares**. ¿Cómo sabemos de qué distribución provienen? No lo sabemos, pero podemos hacer suposiciones y esperar que nuestras suposiciones se cumplan. Si no lo hacen, tendremos un modelo que funciona mal.

Por ejemplo, si se graba audio en una provincia al probar con otra region el modelo no va a funcionar bien

Datos elaborados por BCP para uso interno

¿CUÁNDO USARLO?

- **Cuando una tarea es repetitiva**. Cada patrón se repite varias veces, lo que facilita que las máquinas lo aprendan.
- Si los patrones cambian constantemente. Las culturas cambian. Los gustos cambian. Las tecnologías cambian. Lo que está de moda hoy puede ser una vieja noticia mañana. Consideremos la tarea de clasificar el correo no deseado. Hoy un indicio de un correo electrónico no deseado es un príncipe nigeriano, pero mañana podría ser un escritor vietnamita angustiado. Debido a que Aprendizaje Automático aprende de los datos, **se puede actualizar los modelos con nuevos datos sin tener que descubrir cómo han cambiado los datos**.

Datos elaborados por IICP para uso interno

¿CUÁNDO USARLO?

- Los modelos se equivocan. Por lo que hay que dar más atención en problemas que pueden tener **consecuencias catastróficas.** Desarrollar vehículos autónomos es un desafío porque un error puede provocar la muerte. Muchas empresas todavía quieren desarrollar vehículos autónomos porque tienen el potencial de salvar muchas vidas una vez que los vehículos autónomos sean estadísticamente más seguros que los conductores humanos.

- Las soluciones de aprendizaje automático generalmente requieren una inversión inicial no trivial en datos, computación, infraestructura y talento, por lo que se justifica **si el uso de esta solución es frecuente**.

Datos elaborados por IACP para uso interno

¿CUÁNDO USARLO?

La mayoría de los algoritmos de Aprendizaje Automático **no deberían usarse bajo ninguna de las siguientes condiciones**:

- No es ético
- Soluciones más simples funcionan bien
- No es rentable

Datos elaborados por IICP para uso interno

DEFINICIONES

Datos elaborados por BCP para uso Interno

TERMINOLOGIA BÁSICA

En Machine Learning generalmente se utilizan arrays 2D y notaciones vectoriales para referirnos a los datos, de la siguiente forma:

datos tabulares

- Cada fila del array es una muestra, observación o dato puntual
- Cada columna es una característica (feature o atributo), de la observación mencionada en el punto anterior.
- En el caso más general habrá una columna, que llamaremos objetivo, label, etiqueta o respuesta, y que será el valor que se pretende predecir.

Datos elaborados por IACP para uso interno

TERMINOLOGIA BÁSICA

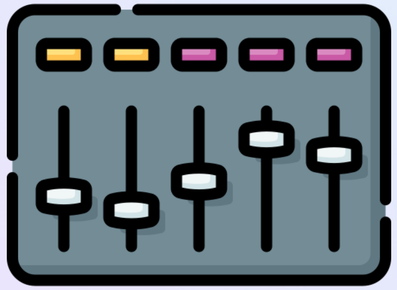
	Features					Label
	Position	Experience	Skill	Country	City	Salary ($)
Observation →	Developer	0	1	USA	New York	103100
	Developer	1	1	USA	New York	104900
	Developer	2	1	USA	New York	106800
	Developer	3	1	USA	New York	108700
	Developer	4	1	USA	New York	110400
	Developer	5	1	USA	New York	112300
	Developer	6	1	USA	New York	116100
	Developer	7	1	USA	New York	117800

Datos elaborados por BCP para uso interno

TERMINOLOGIA BÁSICA

Los algoritmos de Machine Learning tienen parámetros “internos” que no dependen de los datos. Estos parámetros se llaman hiperparámetros. Por ejemplo, una red neuronal tiene como hiperparámetros la función de activación o la constante de entrenamiento.

Por ejemplo, en el algoritmo de los K-vecinos mas proximos, el valor K es un hiperparametro



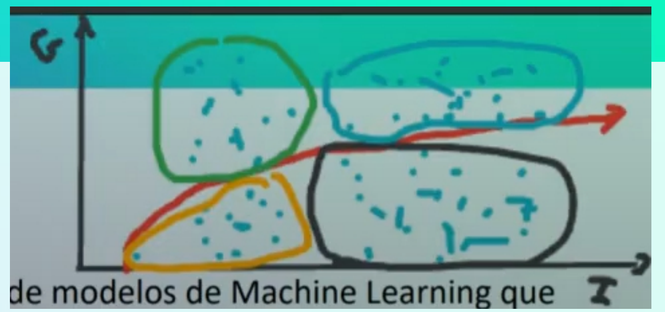
Llamamos generalización a la capacidad del modelo de hacer predicciones nuevas utilizando datos nuevos.

Robustez frente a datos que no a visto durante el entrenamiento

Consola iconos creados por Freepik - Flaticon

Datos elaborados por IICP para uso interno

TIPOS DE APRENDIZAJE



Aprendizaje supervisado: Se refiere a un tipo de modelos de Machine Learning que se entrenan con un conjunto de ejemplos en los que los resultados de salida son conocidos.

Aprendizaje no supervisado: El objetivo será la extracción de información significativa, sin la referencia de variables de salida conocidas, y mediante la exploración de la estructura de dichos datos sin etiquetar.

Una aplicacion de Aprendizaje no supervisado, muy comun, es la de construir variables.
En base al analisis se determina que significa cada grupo

Aprendizaje profundo: Es un subcampo de Machine Learning, que usa una estructura jerárquica de redes neuronales artificiales, que se construyen de una forma similar a la estructura neuronal del cerebro humano, con los nodos de neuronas conectadas.

En este campo, hay ademas auto supervisado, fuerte, debil, aprendizaje por refuerzo

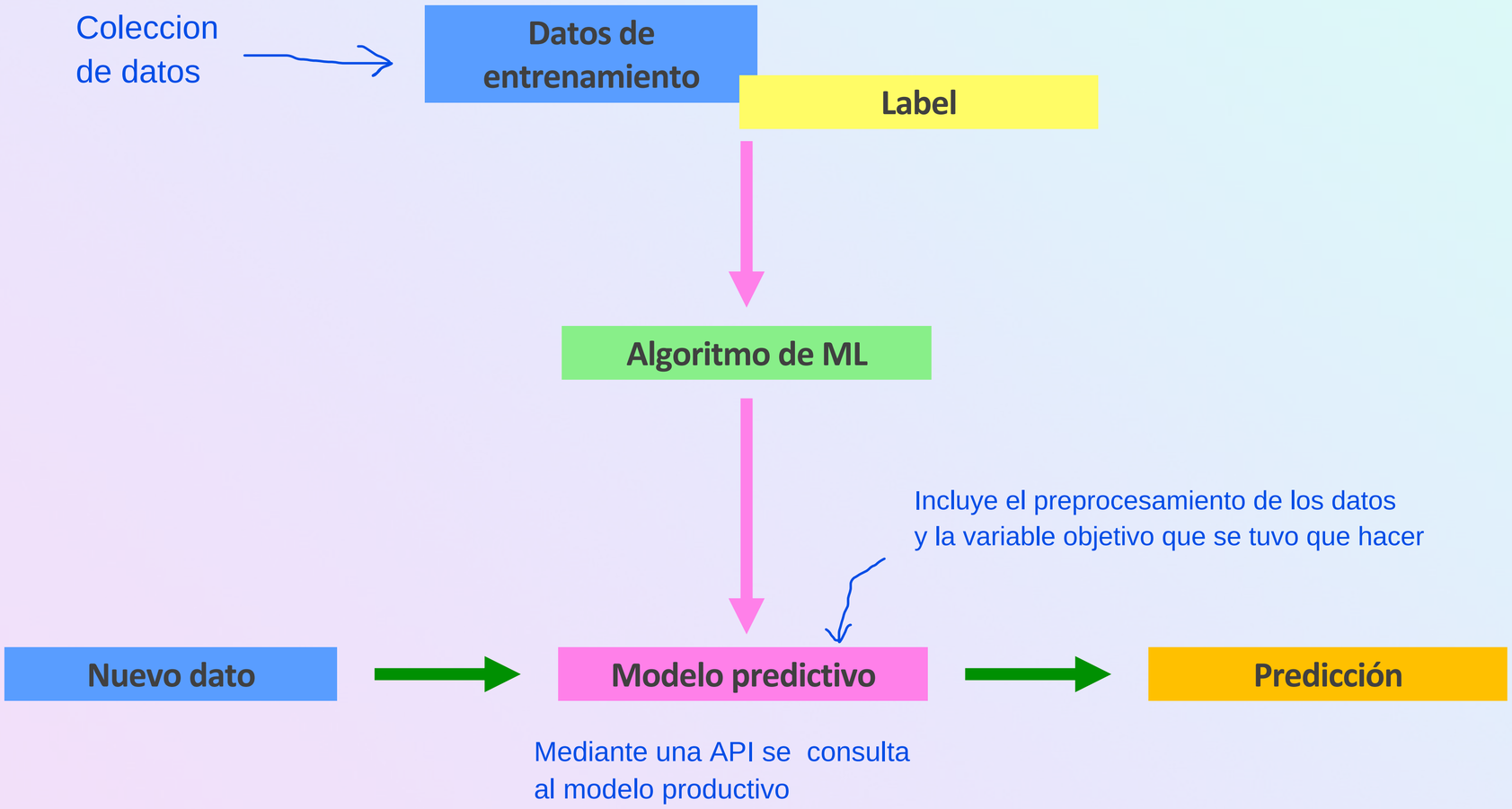
Datos elaborados por IACP para uso interno

APRENDIZAJE SUPERVISADO

- Los modelos aprenden de los resultados conocidos y realizan ajustes en sus parámetros interiores para adaptarse a los datos de entrada.
- Una vez que el modelo es entrenado adecuadamente, y los parámetros internos son coherentes con los datos de entrada y los resultados de los datos de entrenamiento, el modelo podrá realizar predicciones adecuadas ante nuevos datos

Image by vectorjuice

APRENDIZAJE SUPERVISADO



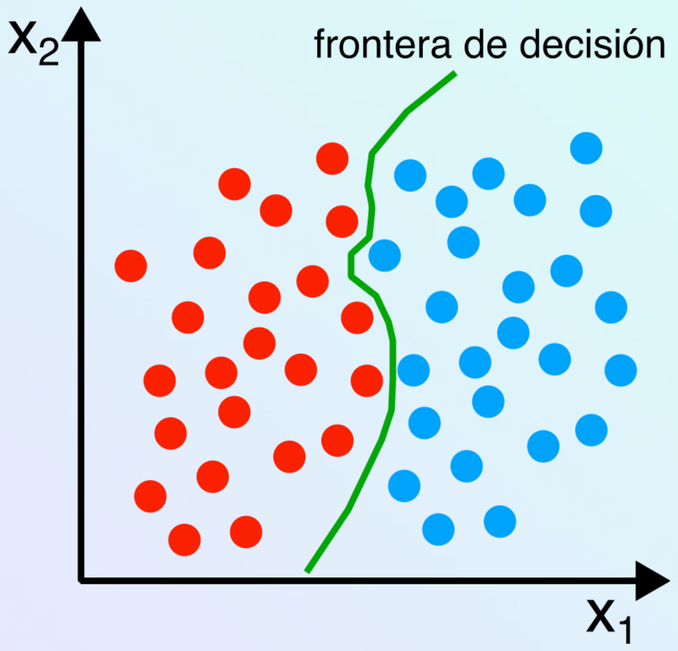
Datos elaborados por BCP para uso interno

APRENDIZAJE SUPERVISADO

→ CLASIFICACIÓN

Clasificación es una subcategoría de aprendizaje supervisado en la que el objetivo es predecir clases categóricas (valores discretos, no ordenados, pertenencia a grupos).

- Detección de SPAM… clasificación binaria (es spam o no es spam).
- Clasificación multi-clase: Múltiple clases, como por ejemplo clasificación del nivel socioeconómico de una persona (alta, media y baja).



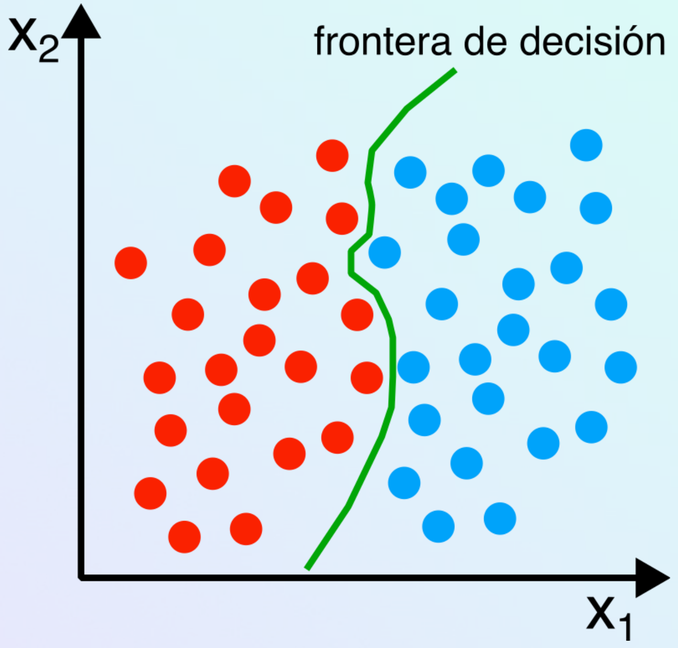
Datos elaborados por IACP para uso interno

APRENDIZAJE SUPERVISADO

CLASIFICACIÓN

Un ejemplo de clasificación binaria: hay dos clases de objetos, círculos y cruces, y dos características de los objetos, X_1 y X_2

- Se entrena un modelo que dado las dos características puede predecir la clase (circulo o cruz).
- Lo que hace el modelo es crear una frontera que le permite separar a las dos clases. Si tenemos un nuevo dato, en función de qué lado de la frontera queda, se clasificará de una forma u otra.

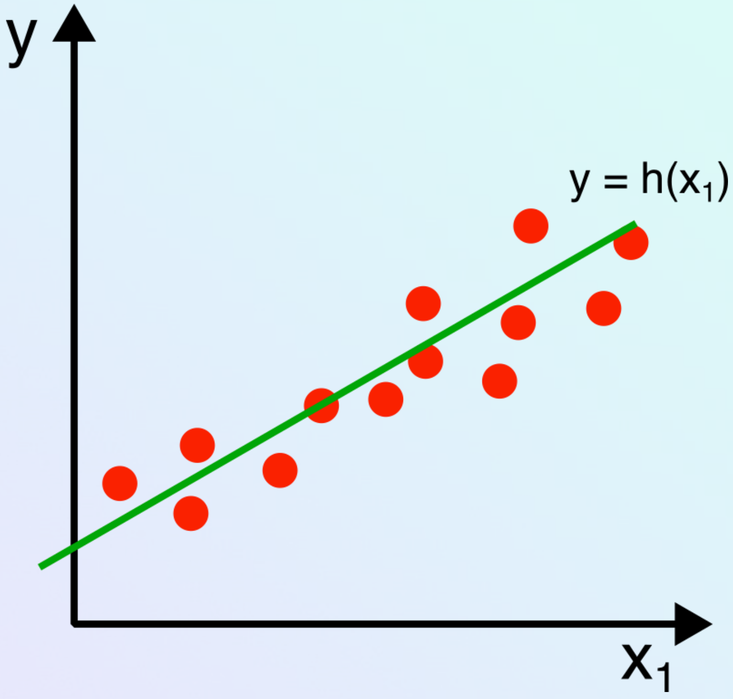


APRENDIZAJE SUPERVISADO

→ REGRESIÓN

En este tipo de aprendizaje tenemos un número de variables predictoras (explicativas) y una variable de respuesta continua (resultado), y se tratará de encontrar una relación entre dichas variables que nos proporciones un resultado continuo.

- La regresión lineal: dados X e Y, establecemos una línea recta que minimice la distancia entre los puntos de muestra y la línea ajustada. Después, utilizaremos la fórmula obtenida de la regresión para predecir nuevos datos de salida.
- Un ejemplo típico es la regresión del precio de casas en ventas en una ciudad dado barrio, cantidad de habitaciones, etc.



Datos elaborados por IICP para uso interno

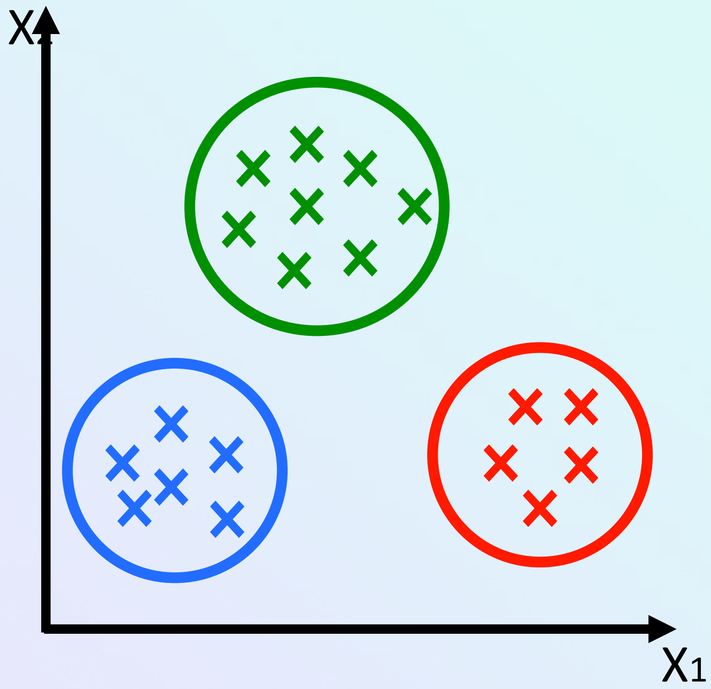
APRENDIZAJE NO SUPERVISADO

→ AGRUPAMIENTO O CLUSTERING

Hay dos tipos de algoritmos: basadas en distancia y densidad

El agrupamiento es una técnica exploratoria de análisis de datos, que se usa para organizar información en grupos con significado sin tener conocimiento previo de su estructura.

- Cada grupo es un conjunto de objetos similares que se diferencia de los objetos de otros grupos.
- El objetivo es obtener un número de grupos de características similares.
- Un ejemplo de aplicación es establecer dado el comportamiento de muchas cadenas de comida, como se agrupan en el mercado en diferentes grupos para evaluar verdaderos competidores.



Datos elaborados por IACP para uso interno

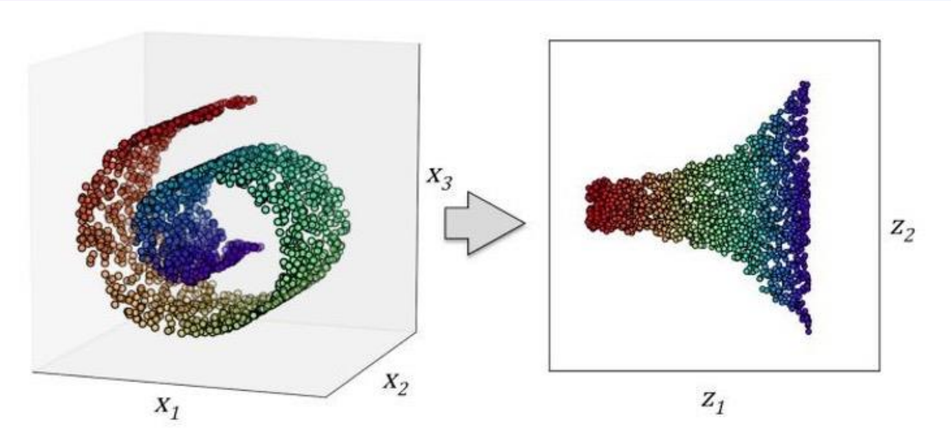
APRENDIZAJE NO SUPERVISADO

- REDUCCIÓN DIMENSIONAL

Hay escenarios donde tener demasiadas variables es contraproducente porque el modelo no puede centrarse en determinado aspecto de la informacion

La reducción dimensional funciona encontrando correlaciones entre las características, lo que implica que existe información redundante, ya que alguna característica puede explicarse parcialmente con otras (por ejemplo, puede existir dependencia lineal).

Estas técnicas eliminan *ruido* de los datos (que puede también empeorar el comportamiento del modelo), y comprimen los datos en un sub-espacio más reducido, al tiempo que retienen la mayoría de la información relevante



APRENDIZAJE PROFUNDO

Esta arquitectura permite abordar el análisis de datos de forma no lineal.

La primera capa de la red neuronal toma datos en bruto como entrada, los procesa, extrae información y la transfiere a la siguiente capa como salida.

Este proceso se repite en las siguientes capas, cada capa procesa la información proporcionada por la capa anterior, y así sucesivamente hasta que los datos llegan a la capa final, que es donde se obtiene la predicción.

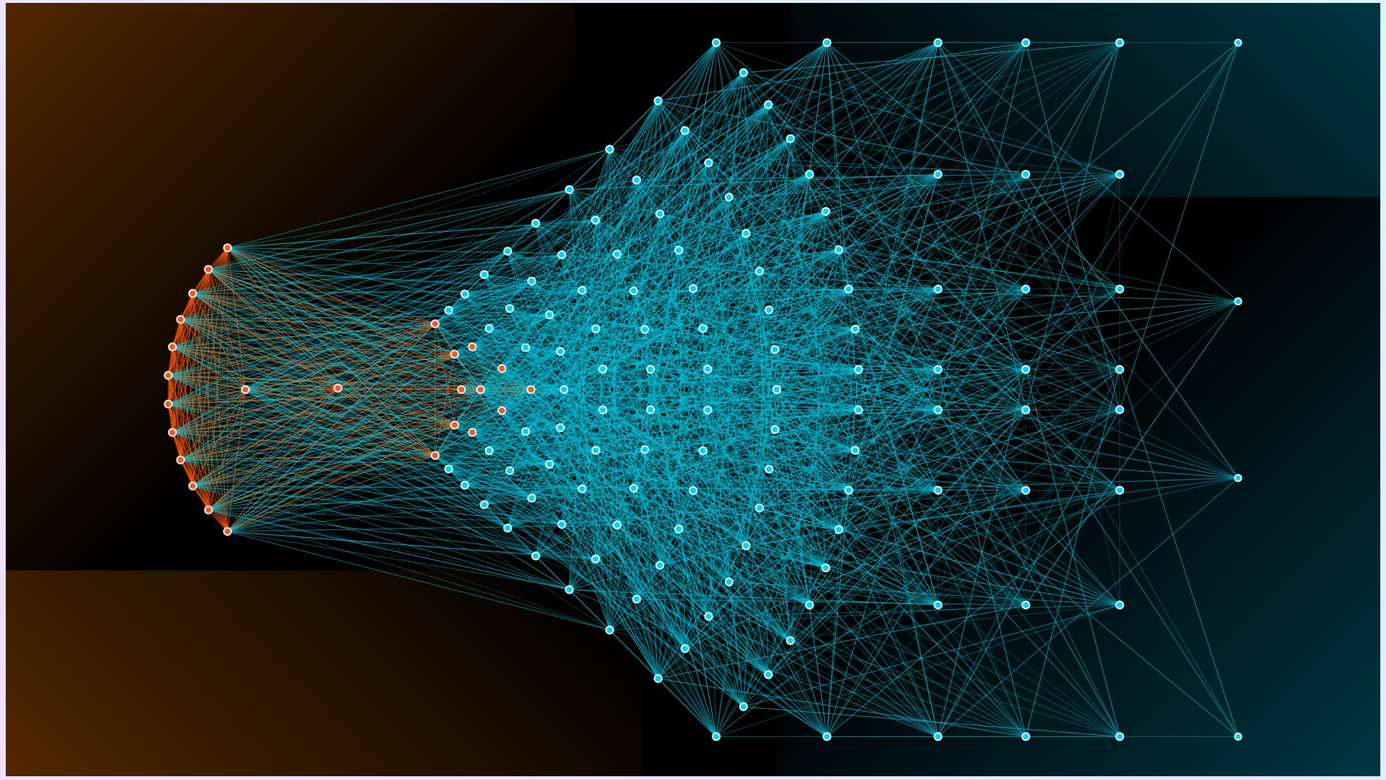
Esta predicción se compara con el resultado conocido, y así por análisis inverso el modelo es capaz de aprender los factores que conducen a salidas adecuadas.

Es uno de los principales algoritmos utilizados en la creación de aplicaciones y programas para reconocimiento de imágenes.

Datos elaborados por IACP para uso interno

APRENDIZAJE PROFUNDO

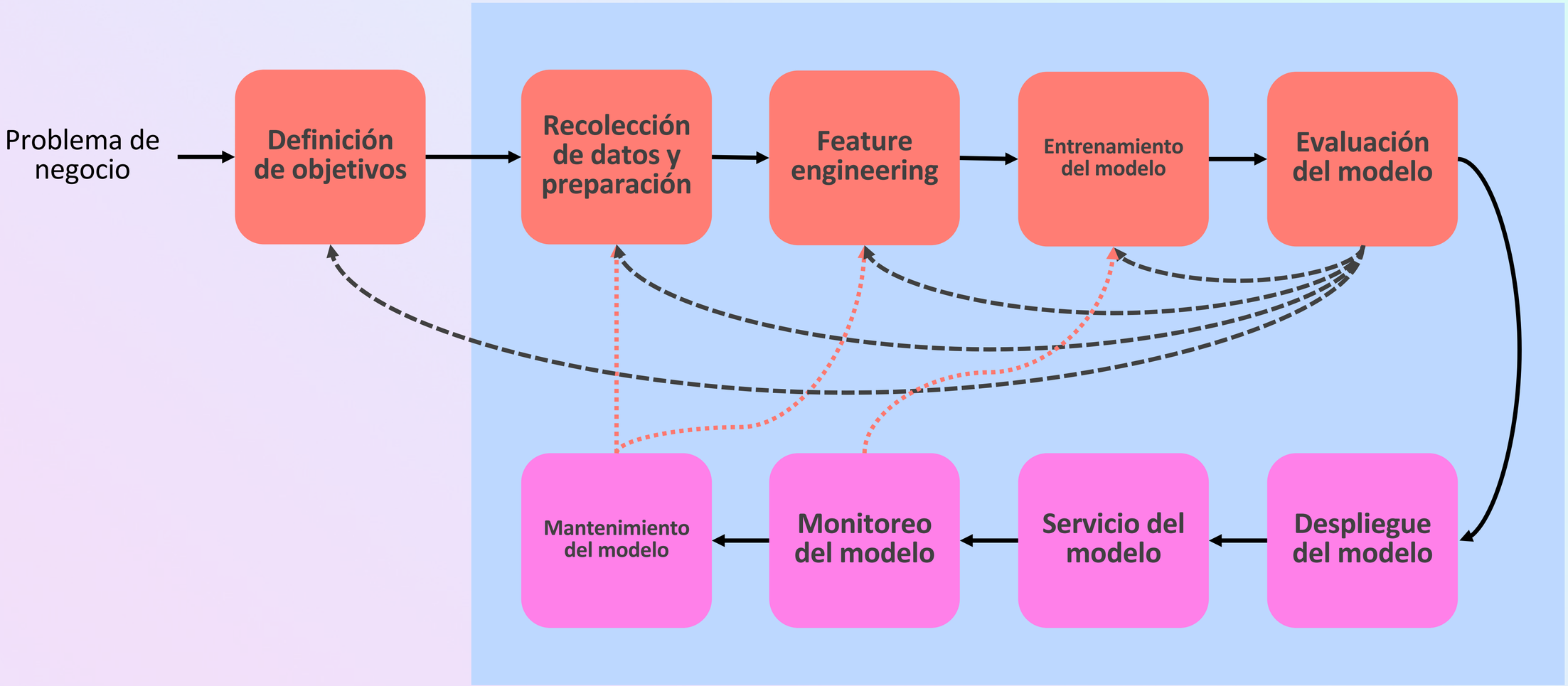
Para tener buenos resultados, necesitamos muchos datos y en general son “caras" de entrenar.



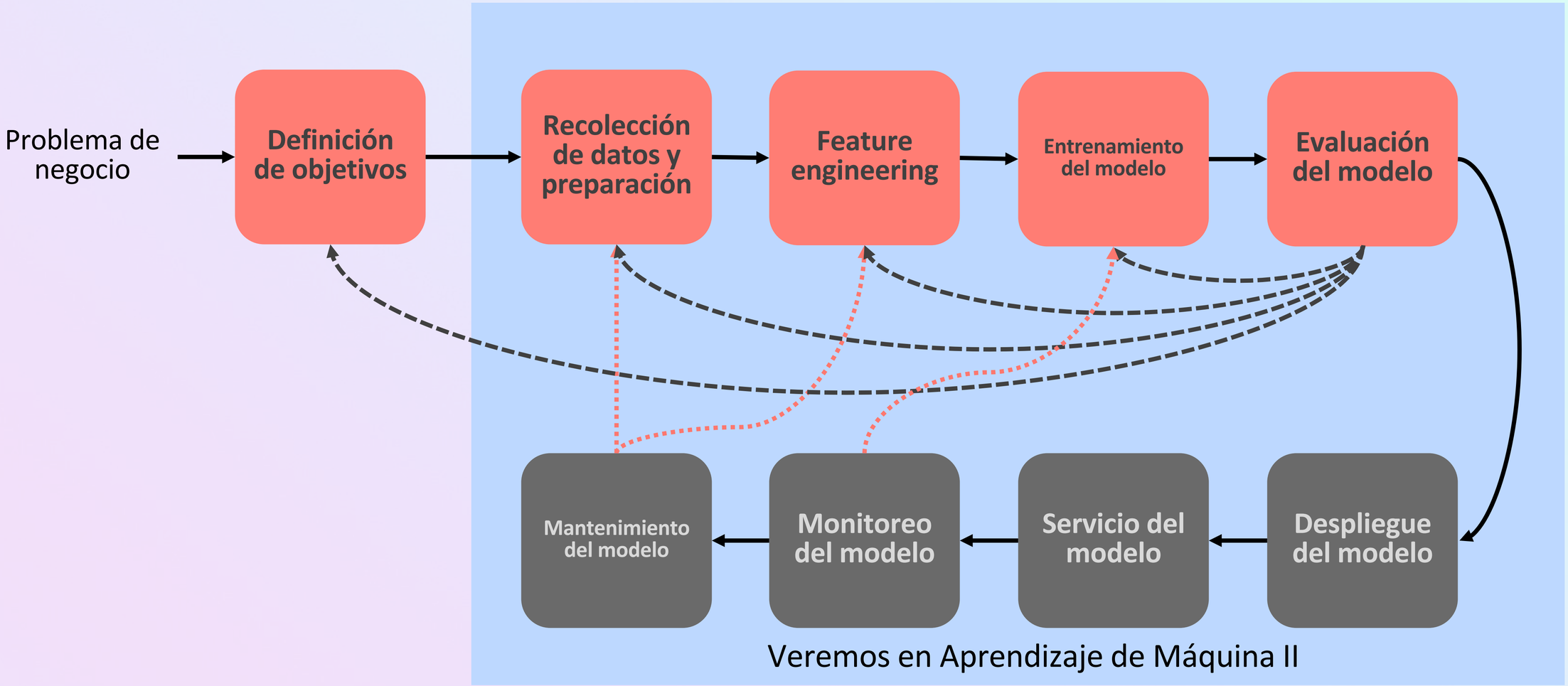
CICLO DE VIDA

Datos elaborados por BCP para uso Interno

CICLO DE VIDA DE UN PROYECTO DE APRENDIZAJE AUTOMÁTICO



CICLO DE VIDA DE UN PROYECTO DE APRENDIZAJE AUTOMÁTICO



ENTENDIENDO EL NEGOCIO

Datos elaborados por BCP para uso Interno

ENTENDIENDO EL NEGOCIO

Cuando se trabajan en un proyecto de Aprendizaje Automático, los científicos de datos tienden a preocuparse por los objetivos de Aprendizaje Automático: Las métricas que pueden medir sobre el rendimiento, como la precisión, F1, la latencia de inferencia, etc.

Los equipos dan mucha significancia a mejorar la exactitud de 94% a 94.5%, inclusive gastando muchos recursos. La verdad es que es que a la mayoría de las empresas no les importan las sofisticadas métricas de ML. No les importa aumentar la precisión de un modelo del 94% al 94,2% a menos que cambie algunas métricas comerciales.

Solo tiene sentido en empresas que buscan mas exactitud, como OpenAI

Entonces, ¿qué métricas les interesan a las empresas? Para que un proyecto de ML tenga éxito dentro de una organización empresarial, es fundamental **vincular el rendimiento de un sistema de ML con el rendimiento empresarial general**.

¿En qué métricas de rendimiento de negocio se supone que influirá el nuevo sistema de aprendizaje automático, por ejemplo, la cantidad de ingresos por publicidad y el número de usuarios activos mensuales?

Hay una metrica popular llamada Time to First Token: el tiempo que tarda el primer token en ser visualizado por el usuario

Datos elaborados por BCP para uso interno

ENTENDIENDO EL NEGOCIO

Es importante lograr vincular métricas del modelo con las métricas del negocio. Una de las razones por las que la predicción de las tasas de clics en anuncios y la detección de fraude se encuentran entre los casos de uso más populares de ML en la actualidad es que es fácil asignar el rendimiento de los modelos a las métricas de negocio: cada aumento en la tasa de clics da como resultado ingresos publicitarios reales , y cada transacción y cada transacción fraudulenta detenida da como resultado un ahorro de dinero real.

Para obtener una respuesta definitiva a la pregunta de cómo las métricas de ML influyen en las métricas de negocio, a menudo es necesario realizar experimentos.

Muchas veces se hacen con experimentos como las pruebas A/B y eligen el modelo que conduce a mejores métricas comerciales, independientemente de si este modelo tiene mejores métricas de ML.

Datos elaborados por IACP para uso interno

ENTENDIENDO EL NEGOCIO

A veces, es difícil manejar expectativas de negocios. Debido al hype que genera Machine Learning, se debe ser realista acerca de los beneficios que pueden dar un modelo.

Algunos negocios pueden tener la noción de que el Aprendizaje Automático puede transformar una solución mágicamente de la noche a la mañana.

- **Mágicamente: Posible**
- **De la noche a la mañana: No**

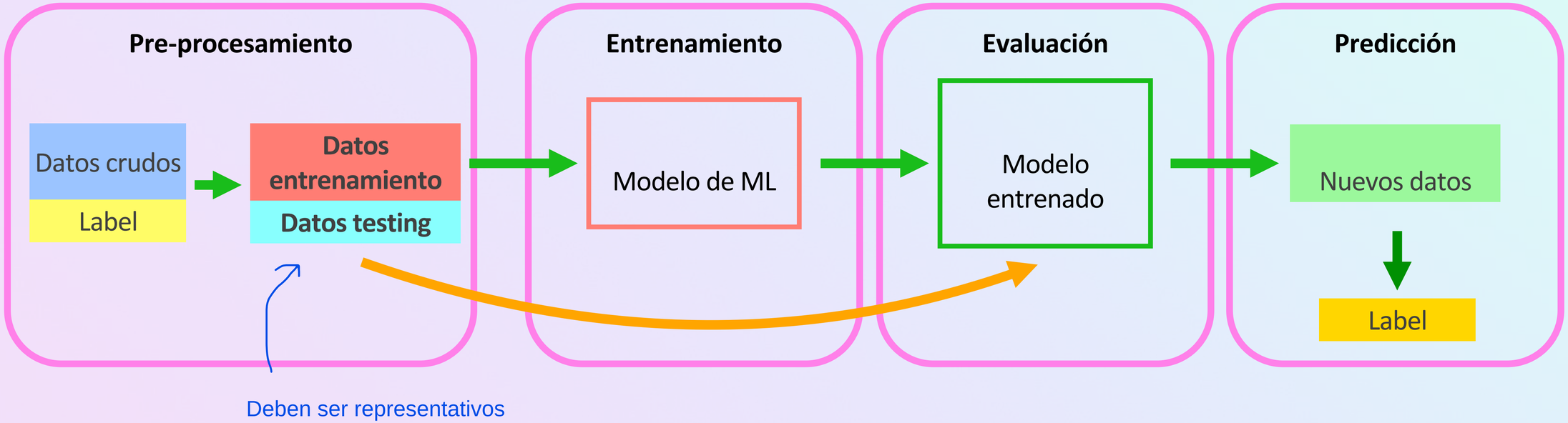
Hay que dar concientizacion de lo que ML puede ofrecer.
Tambien hay que pensar en soluciones escalables.

Datos elaborados por BCP para uso interno

METODOLOGÍA PARA CONSTRUIR MODELOS

Datos elaborados por BCP para uso Interno

METODOLOGÍA PARA CONSTRUIR ALGORITMOS DE ML



METODOLOGÍA PARA CONSTRUIR ALGORITMOS DE ML

→ PRE-PROCESAMIENTO

- Es el paso más importante. Recordar: Garbage in, garbage out.
- Usualmente los datos se presentan en formatos no óptimos (o incluso inadecuados) para ser procesados por el modelo.
- Muchos algoritmos requieren que las características estén en la misma escala para optimizar su rendimiento, lo que se realiza frecuentemente aplicando técnicas de normalización o estandarización en los datos.
- Podemos también encontrar en algunos casos que las características seleccionadas están correlacionadas, y por tanto son redundantes para extraer información con significado correcto de ellas. En este caso tendremos que usar técnicas de reducción dimensional para comprimir las características en subespacios con menores dimensiones.
- Hacer un correcto análisis estadístico es vital para lograr tener los datos para el algoritmo que queremos usar.

Datos elaborados por IACP para uso interno

METODOLOGÍA PARA CONSTRUIR ALGORITMOS DE ML

SELECCIÓN Y ENTRENAMIENTO DE MODELO

- Es esencial comparar los diferentes algoritmos de un grupo para entrenar y seleccionar el de mejor rendimiento. Para realizar esto, es necesario seleccionar una métrica para medir el rendimiento del modelo.
- Cuanto no tenemos nada para comparar inicialmente, elegimos un modelo sencillo que llamamos baseline, que nos servirá para comparar el desarrollo de nuevos modelos. No siempre debe ser un modelo.
- Para asegurarnos de que nuestro modelo funcionará adecuadamente con datos reales, podemos usar validación cruzada (Cross Validation) antes de utilizar el conjunto de datos de prueba para la evaluación final del modelo.

Validacion cruzada + busqueda bayesiana de hiperparametros: es algo que funciona muy bien para elegir la mejor version del algoritmo

- En general, los parámetros por defecto de los algoritmos de Machine Learning proporcionados por las librerías no son los mejores para utilizar con nuestros datos, por lo que usaremos técnicas de optimización de hiperparámetros

Datos elaborados por IACP para uso interno

METODOLOGÍA PARA CONSTRUIR ALGORITMOS DE ML

EVALUANDO Y PREDICIENDO CON DATOS NUEVOS

- Una vez que hemos seleccionado y ajustado un modelo a nuestro conjunto de datos de entrenamiento, podemos usar los datos de prueba para estimar el rendimiento del modelo en los datos nuevos, por lo que podemos hacer una estimación del error de generalización del modelo, o evaluarlo utilizando alguna otra métrica.
- Si el modelo cumple nuestros objetivos, luego es ponerlo productivo en el contexto que se quiere utilizar. En esta materia no veremos este paso.

Hay que considerar que los datos de evaluacion sean representativos y tambien las metricas.
Ejemplo, si los datos estan desbalanceados, la metrica de accuracy no es muy conveniente

Datos elaborados por IACP para uso interno