Machine Learning

Primeros pasos

Herramientas:

Python:

- Entorno de desarrollo
- Variables, tipos de datos, operadores
- Estructuras de datos
- Estructuras selectivas y repetitivas
- Funciones
- Clases y objetos

PANDAS

Numpy -> arrays, acceso a elementos, redimensionamiento, operaciones matemáticas.

Creación de gráficos (Matplotlib, plotly, Seaborn ...)

Scikit-learn

Introducción: contexto

Análisis descriptivo

Análisis predictivo

Análisis prospectivo

Análisis generativo

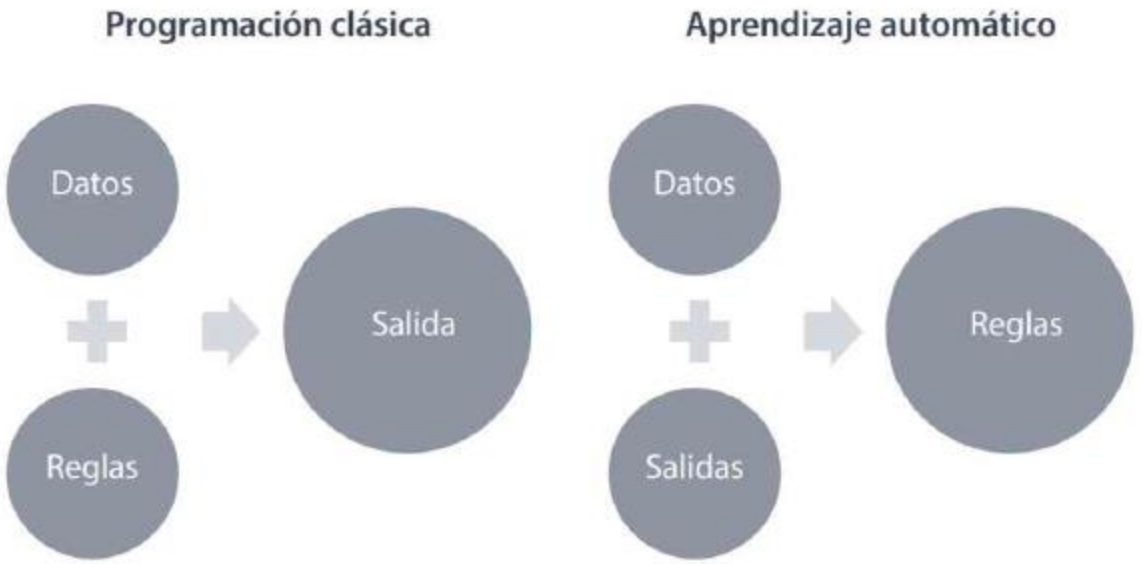
Introducción: Concepto

"Campo de estudio que da al computador la habilidad de aprender sin haber sido explícitamente programado para ello."
Arthur Samuel, 1959

"Se dice que un programa aprende de la experiencia E con respecto a alguna tarea T y alguna medida de rendimiento P, si su rendimiento en T, medido por P, mejora con la experiencia E"
Tom Mitchell

Introducción: Concepto

Programación tradicional: Se procesan datos de entrada y por medio de una serie de reglas se genera una salida



Machine Learning: Datos y salidas son los datos iniciales, y mediante proceso de entrenamiento generan reglas (modelo)

Etapas


Preprocesamiento de datos
Separación en conjunto de entrenamiento y prueba
Configuración del algoritmo
Entrenamiento del modelo
Predicción
Evaluación

Preprocesamiento de datos

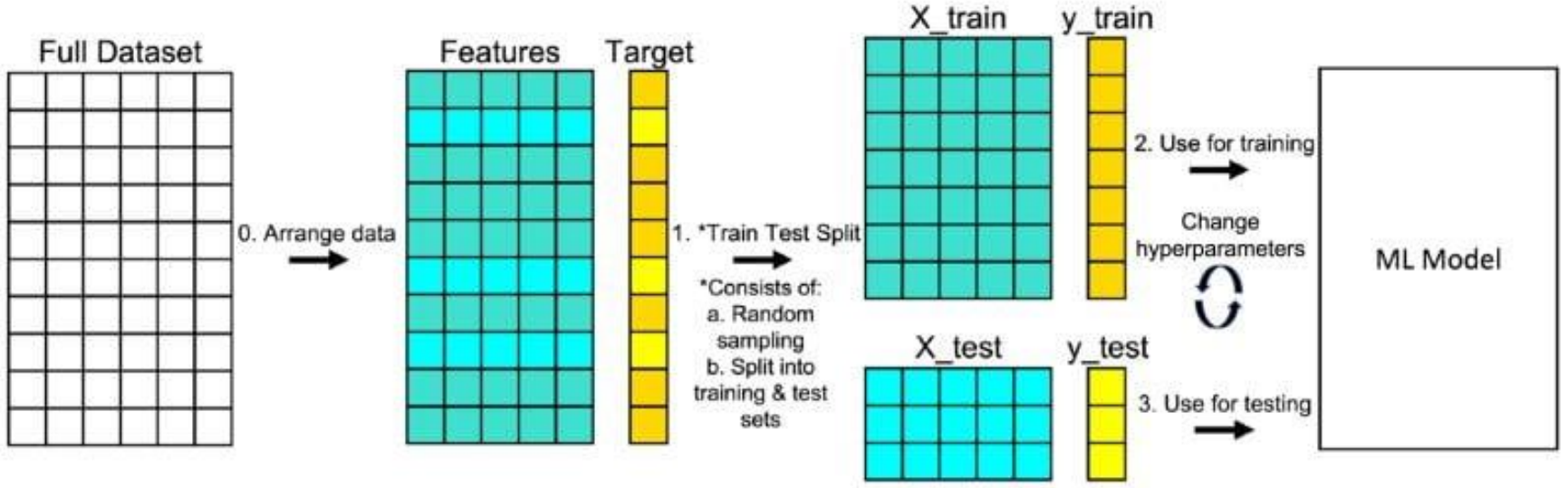
Etapa crucial, si los datos no son adecuados el modelo de echa a perder.

Los algoritmos de ML trabajan con datos numéricos y en una misma escala mejor.

Limpieza y transformación de datos, gestión de datos nulos, etc.

Conjunto de entrenamiento y prueba

Dividimos el dataset en dos subconjuntos, entrenamiento y prueba. Uno para entrenas y estimar los **parámetros** del modelo y otro para hacer predicciones y probar el modelo con datos diferentes a los de entrenamiento (70-30% u 80-20%)

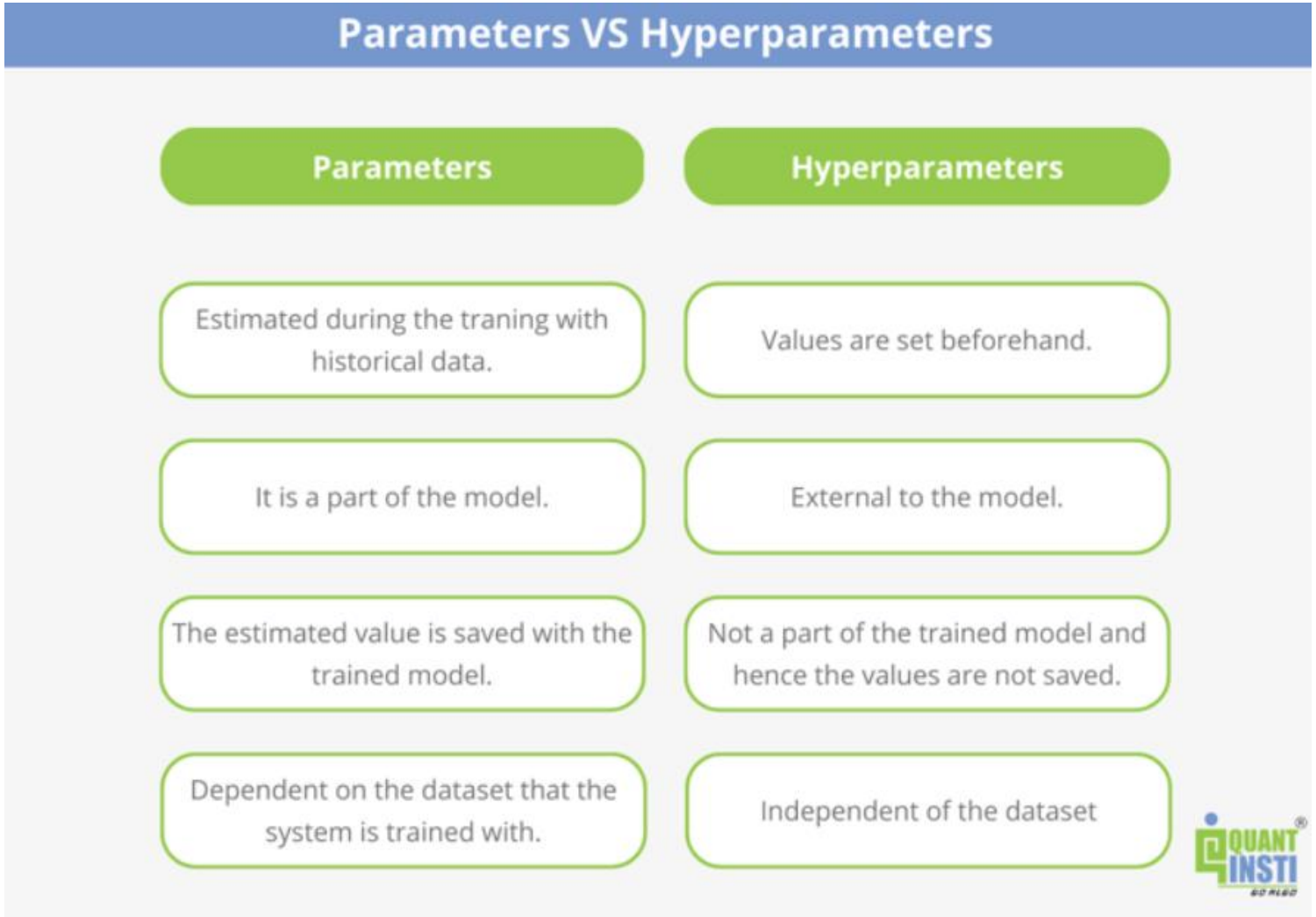


Configuración del algoritmo

Se crea una instancia del algoritmo a utilizar y se definen los **hiperparámetros.**

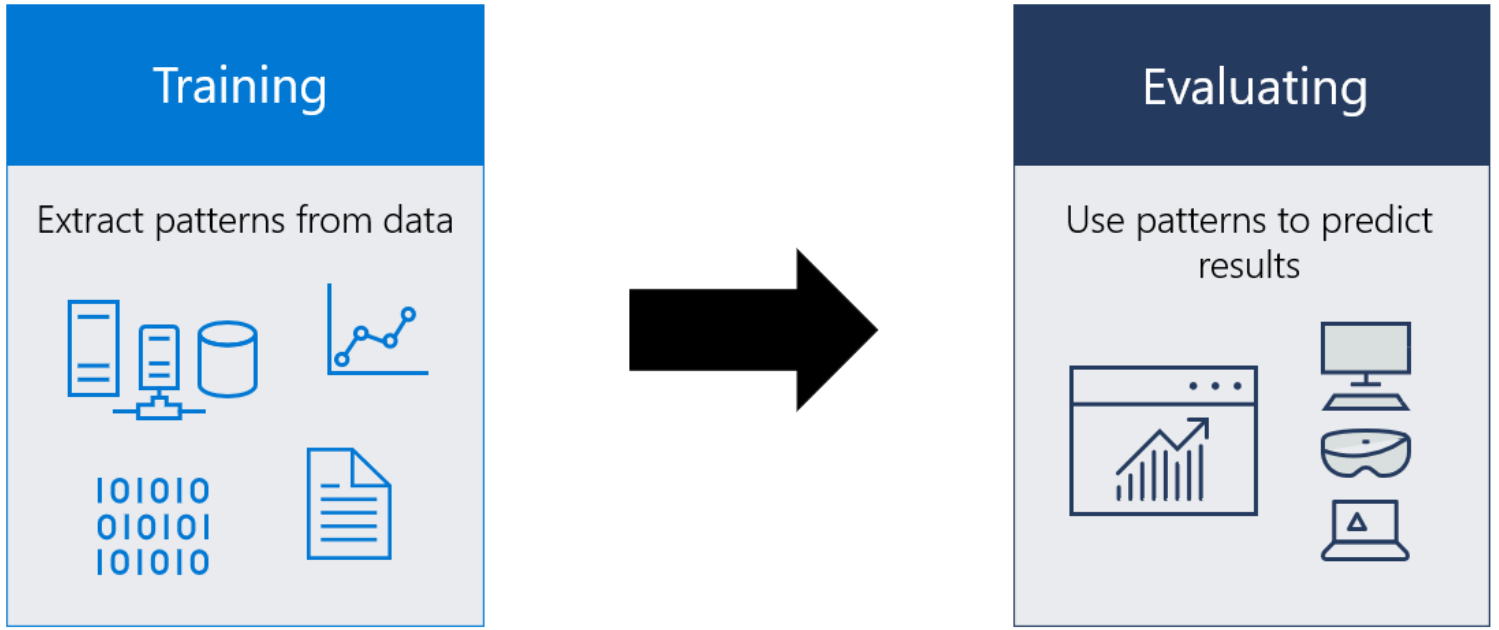
Hiperparámetros: Los parámetros los define el modelo internamente durante el entrenamiento, nosotros no participamos directamente en su generación.

EJ: Número de épocas, tasa de aprendizaje.



Entrenamiento del modelo

Le proporcionamos al modelo o instancia del modelo un dataset para que pueda aprender, logrando estimar los parámetros del modelo.



Predicción

Una vez generado el modelo se puede probar su nivel de presicción pasándole muestras del conjunto de pruebaas.

El resultado sobre la muestra es un valor contínuo o discreto lo más cercano al valor esperado.

Evaluación

Determinar numéricamente la efectividad de nuestro modelo. Esta efectividad parte del supuesto de que a menos diferencia entre salida esperada y salida predicha mejor es la evaluación.

Existen diferentes métricas, que se usan dependiendo del tipo de problema de ML.

Glosario

Dataset: Datos recibidos por el algoritmo de ML para ser entrenado y probado. Normalmente de tipo estructurado y no estructurado (tabla vs cuerpo de correo electrónico)

Variables descriptivas: Describen las instancias(filas) almacenadas en el conjunto de datos. Tienen un tipo asociado (int, object…): num –habitaciones, area –m2…

Glosario

Variable destino o etiqueta: Característica o conjunto de ellas usadas para etiquetas instancias(necesario para un tipo de entrenamiento del modelo) en el caso de las casas, sería el precio

Tipos de datos: Numéricos, categóricos, ordinales

Modelo: Resultado del proceso de aprendizaje automático

	Atributos descriptivos				*Etiqueta*
Id	**Salario**	**Historia crediticia**	**Ocupación**	**Edad**	**Préstamo**
1	1.800.000	Bien	Profesional	24	SI
2	700.000	-	Técnico	18	NO
3	2.000.000	Bien	Profesional	20	SI
4	1.000.000	Mal	Tecnólogo	55	NO

Variable destino o etiqueta: Característica o conjunto de ellas usadas para etiquetas instancias(necesario para un tipo de entrenamiento del modelo) en el caso de las casas, sería el precio

Tipos de datos: Numéricos, categóricos, ordinales

Modelo: Resultado del proceso de aprendizaje automático

	Atributos descriptivos				*Etiqueta*
Id	**Salario**	**Historia crediticia**	**Ocupación**	**Edad**	**Préstamo**
1	1.800.000	Bien	Profesional	24	SI
2	700.000	-	Técnico	18	NO
3	2.000.000	Bien	Profesional	20	SI
4	1.000.000	Mal	Tecnólogo	55	NO

Tipos de Aprendizaje Automático

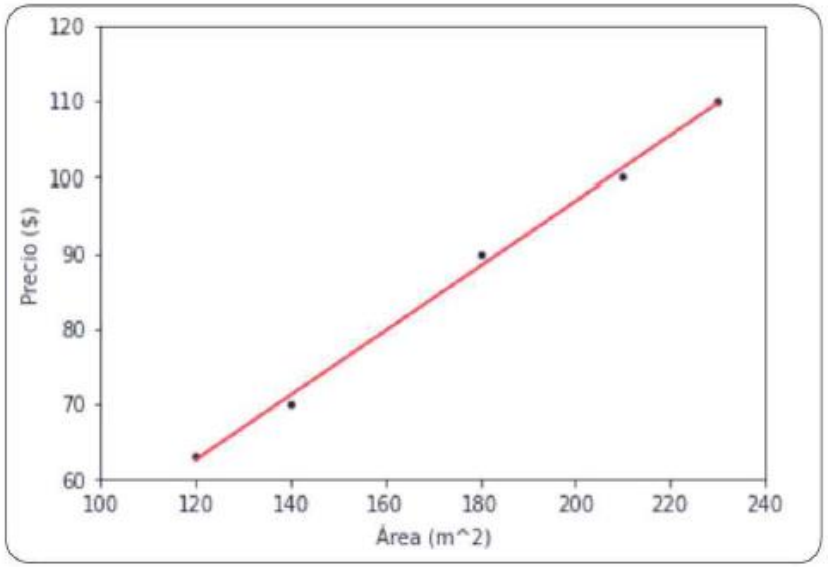
Tipos de Aprendizaje Automático: Supervisado

Todas las muestras son etiquetadas con un resultado o categoría utilizando una variable llamada destino u objetivo. (préstamo SI o NO)

Regresión: Predice un valor continuo, tomando variables de entrada. Busca encontrar la relación entre una variable dependiente con algunas independientes.

Predecir el precio de una casa a partir del área en m2, temperatura el próximo mes, tiempo de vida de un dispositivo electrónico.

- Predecir el tiempo de vida de un dispositivo electrónico.



Tipos de Aprendizaje Automático: Supervisado

Clasificación: El valor a predecir es discreto, asocia una instancia a una categoría o clase a la que pertenece. (modelo detector de SPAM, clasificar flores, determinar si un tumor es maligno o benigno, predecir el ganador de las proximas elecciones entre dos candidatos)

Tipos de Aprendizaje Automático: No supervisado

No contamos con etiquetas para las muestras de entrenamiento del modelo, el modelo actúa directamente sobre los datos de entrada y busca relaciones entre ellos basándose en características en común.

- Clustering
- Reducción de dimensionalidad.

Muy usado en medicina para el diagnóstico de enfermedades, en marketing para segmentar clientes según preferencias.

Otros algoritmos de agrupamiento: **Semisupervisados**

Usa datos de entremaniento tanto etiquetados como no etiquetados. En ocasiones se consiguen modelos más exactos.

Tipos de Aprendizaje Automático: Por refuerzo

Modelos que se consiguen a pertir de ensayo y error, teniendo en cuenta que cada vez que se da un acierto se establece una recompensa y en caso contrario una penalización.

Estos algoritmos, llamados agentes, buscan aprender estrategias para minimizar las penalizaciones (como los niños)

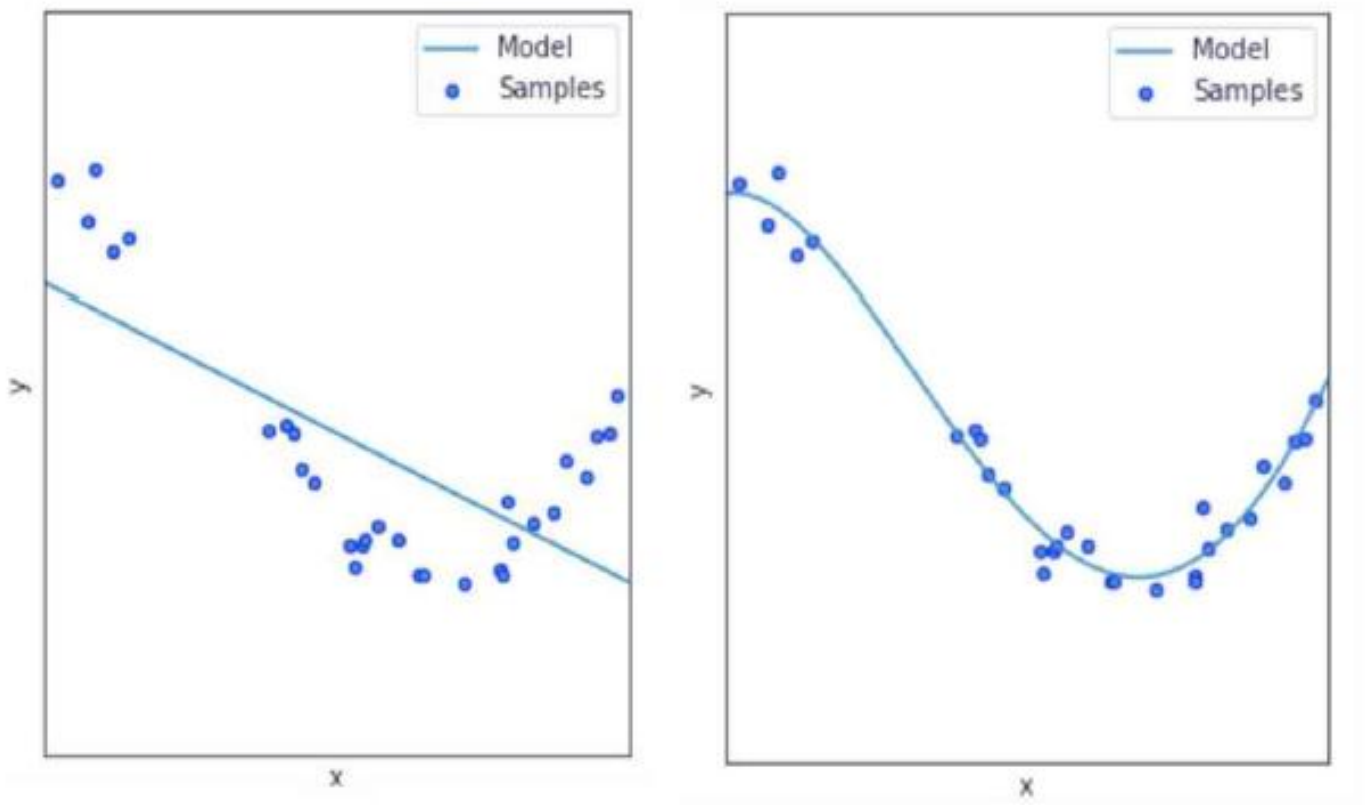
Problemas típicos de ML

Calidad de los datos

Maldición de la dimensionalidad (columnas con alta correlación)

Subajuste (underfitting) -> Sesgo elevado

Sobreajuste (overfitting)-> Información no deseada, valores atípicos, ruido y alta varianza.



CRISP(Cross Industry Standard for Data Mining)

Modelo genérico, flexible y cíclico diseñado para ser adaptado para cubir necesidades en procesos de minería de datos.

- Conocimiento de negocio
- Conocimiento de los datos
- Preparación de los datos
- Modelado
- Evaluación
- Despliegue