ntic master School

UNIVERSIDAD COMPLUTENSE DE MADRID

Introducción

Minería de Datos y Modelización Predictiva

Máster en Big Data y Business Analytics

Universidad Complutense de Madrid

Curso 2020-2021

- Clases teóricas y prácticas utilizando R.
- Contenidos:
  - Introducción a la minería de datos
  - Análisis exploratorio y depuración de datos
  - Regresión lineal
  - Regresión logística
  - Selección de variables en modelos de regresión.

- Según IBM:

  "*La minería de datos es una forma* **innovadora** *de obtener información comercial valiosa mediante el análisis de los datos contenidos en la base de datos de la empresa.* **Revela información** *comercial exhaustiva utilizando técnicas avanzadas de análisis y creación de modelos.*"

- Según Microsoft:

  "*La minería de datos es el proceso de* **detectar información** *procesable de grandes conjuntos de datos para deducir los* **patrones y tendencias** *que existen. Normalmente, estos patrones no se pueden detectar mediante la exploración tradicional de los datos porque las relaciones son demasiado complejas o hay* **demasiado datos**."

- Según Wikipedia:

  "*La minería de datos es un campo de la* **estadística** *y las ciencias de la* **computación** *referido al proceso que intenta descubrir patrones en* **grandes volúmenes** *de conjuntos de datos. El objetivo general del proceso de minería de datos consiste en* **extraer información** *de un conjunto de datos y transformarla en una estructura comprensible para su uso posterior.*"

Diferencias entre Minería de datos y Estadística clásica

- La minería de datos trabaja con datos **heterogéneos**, **abundantes** (muchas observaciones y/o variables) y, en ocasiones, de **mala calidad** (con muchos datos atípicos y/o faltantes).
- Se pone más enfasis en la **predicción** que en la explicación.
- El gran tamaño de muestra hace inutil la aplicación de inferencia en la predicción, debido a la **potencia del test**.
- Se realiza frecuentemente una **partición** de los datos en varias muestras: entrenamiento, validación y prueba.
  1. **Entrenamiento** (*training*): Sirve para **generar** y definir el modelo.
  2. **Validación**: Se utiliza para **afinar** el modelo obtenido en la fase de entrenamiento (dependiendo del modelo, esta muestra puede no ser necesaria).
  3. **Prueba** (*test*): Permite **evaluar** la bondad del modelo obtenido.

  Los valores recomendados para llevar a cabo esta particion son: $70\,\%, 15\,\%, 15\,\%$.

Una compañía de seguros quiere adquirir clientes a través de envíos de correos electrónicos. No obstante, en lugar de enviar correos electrónicos a toda su base de datos, sería preferible hacer una selección de aquellos posibles clientes con mayor propensión.

Para ello, dispone de una base de datos con información de campañas similares previas de la que forman parte individuos que pudieron o no convertirse en clientes.

Además, dispone de información adicional de dichos individuos como son: sexo, nº de hijos, nº de tarjetas de crédito, nivel de ingresos,..

De cara a ofrecer una atención más personalizada, un banco quiere predecir, basándose en los productos que tienen sus clientes, qué productos podrían comprar próximamente.

En este caso, la información de la que dispone el banco consiste en los productos que ya tienen contratados sus clientes y las características socio-demográficas de los mismos.

En una compañía telefónica, se desea predecir la fuga de clientes. En este caso, se pueden plantear dos objetivos diferentes: cuánto tiempo va a permanecer el cliente en la compañía o, si es un plazo determinado, va a permanecer o no.

La información disponible será el histótico de clientes, junto con otras variables como son: edad, ingreso, gasto en los meses precedentes, llamadas a atención al cliente, etc.

Según su función

- Identitificativas: Sirven para **identificar observaciones** y no son útiles.
- Input o de entrada: Son las variables **predictoras** (también llamadas independientes).
- Objetivo: Es la variable que se pretende **predecir**.
- rechazadas: Son variables **eliminadas** antes de la fase de modelización.

Según su tipología

- Contínuas o cuantitativas: Toman **cualquier valor** en un intervalo (que puede estar limitado o no).
- Nominales, cualitativas o categóricas: Toman un número **finito** de valores.
- Dicotómicas: Variables nominales que toman sólo **dos** valores.
- Fecha/Hora: Son variables que representan una **fecha y/o hora**. Para poder aprovechar su potencial en la fase de modelización, hay que **obtener otras variables** de ellas.

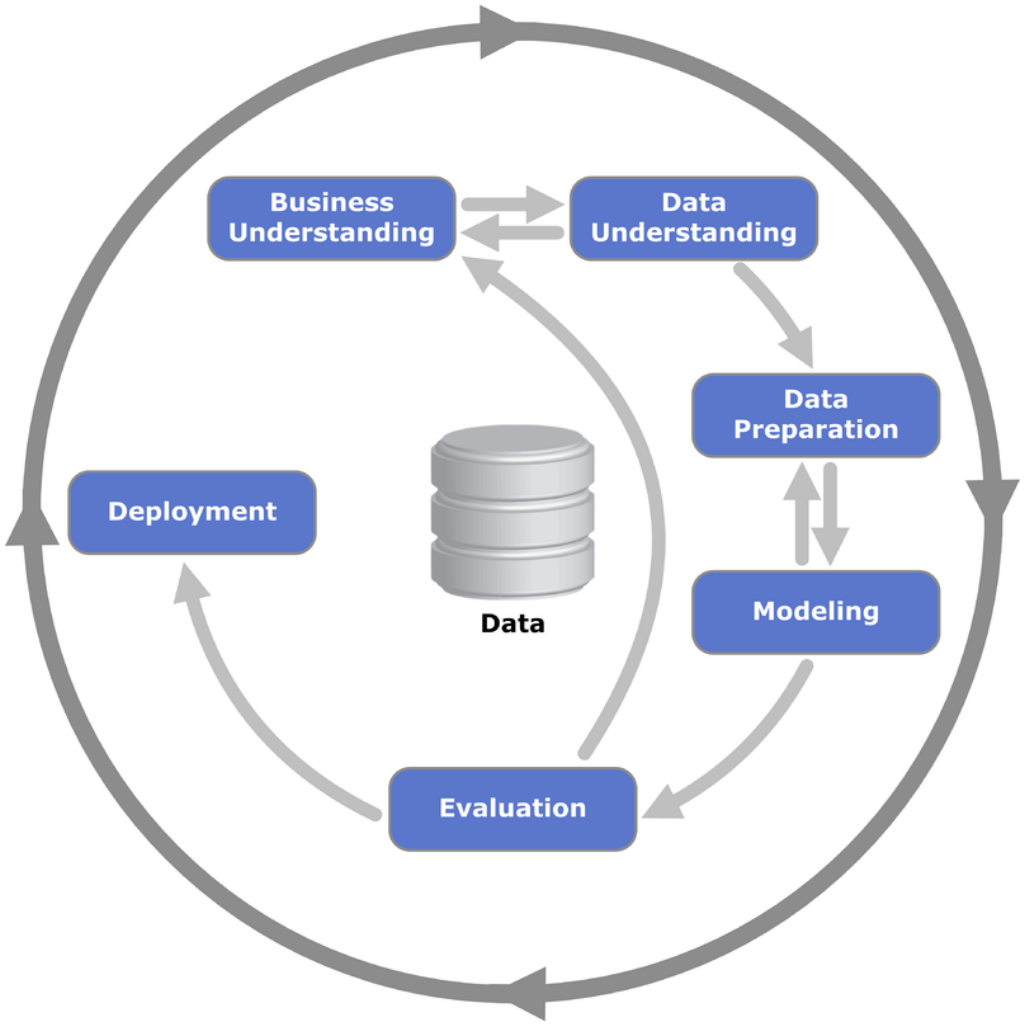
Existen distintas metodologías/esquemas dentro de la minería de datos. Destacan la metodología **SEMMA** (Sample-Explore-Modify-Model-Assess), desarrollada principalmente por la empresa de software SAS, y **CRISP-DM** (Cross Industry Standard Process for Data Mining), desarrollada, entre otros, por la empresa IBM.

Aunque se puede establecer un **paralelismo** claro entre ellas, CRISP-DM es más completo porque tiene en cuenta la aplicación al **entorno de negocio** de los resultados, y por ello, se trata de la metodología **más utilizada**.

En este curso nos centraremos en la metodología CRISP-DM.

- BUSINESS UNDERSTANDING (comprensión del negocio): Esta fase inicial se enfoca en la comprensión de los objetivos de proyecto.
- DATA UNDERSTANDING (estudio y comprensión de los datos): La fase de entendimiento de datos comienza con la colección de datos inicial y continúa con las actividades que permiten familiarizarse con los datos para detectar relaciones, anomalias y tendencias.
- DATA PREPARATION (análisis de los datos y selección de características): La fase de preparación de datos cubre todas las actividades necesarias para construir el conjunto final de datos a partir de los datos en bruto iniciales. Las tareas incluyen la selección de registros y atributos, así como la transformación y la limpieza de datos de cara a facilitar la modelización.
- MODELING (modelado): En esta fase, se seleccionan y aplican las técnicas de modelado que sean pertinentes al problema (cuantas más mejor), y se calibran sus parámetros.
- EVALUATION (evaluación): En esta etapa en el proyecto, es importante evaluar a fondo los modelos construidos y revisar los pasos ejecutados para crearlos, comparándolos entre ellos y evaluando si se han cumplido los objetivos de negocio.
- DEPLOYMENT (despliegue): Dependiendo de los requisitos, la fase de desarrollo puede ser tan simple como la generación de un informe o tan compleja como la realización periódica (y quizás automatizada) de un proceso de análisis de datos en la organización.

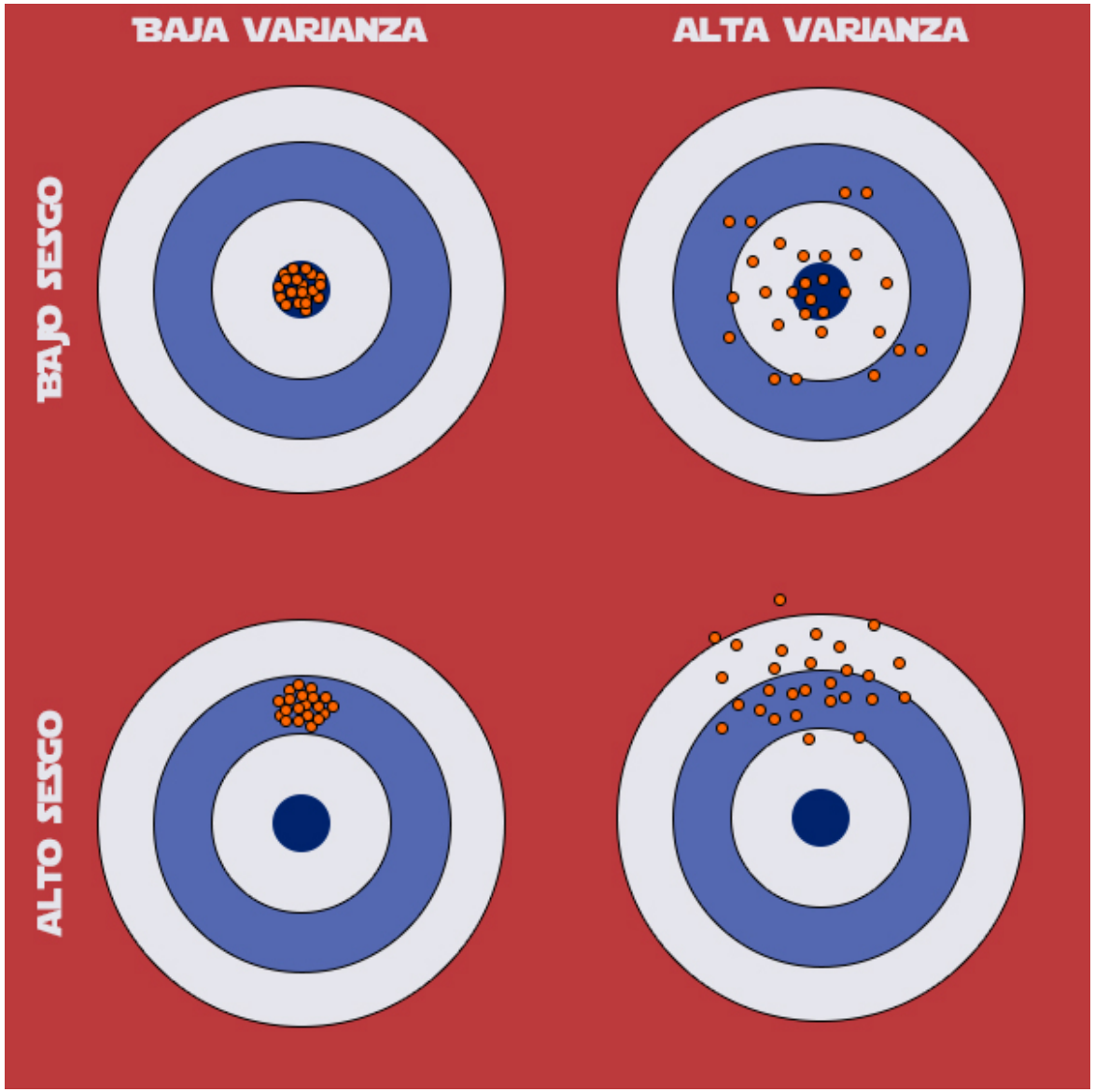
Las fases pueden repetirse y el orden de las mismas puede modificarse.


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

- Los modelos de predicción tienen como objetivo **predecir una variable objetivo** a partir de un conjunto de variables input.
- A falta de modelo de predicción, la mejor estimación que se puede realizar para variables objetivo de intervalo será la **media** de dicha variable y, para variables objetivo de clase, la **clase mayoritaria**.
- Los errores cometidos cuando no se aplica ningún modelo de predicción sirven como **referencia** para evaluar la bondad del modelo construido.

SESGO Y VARIANZA

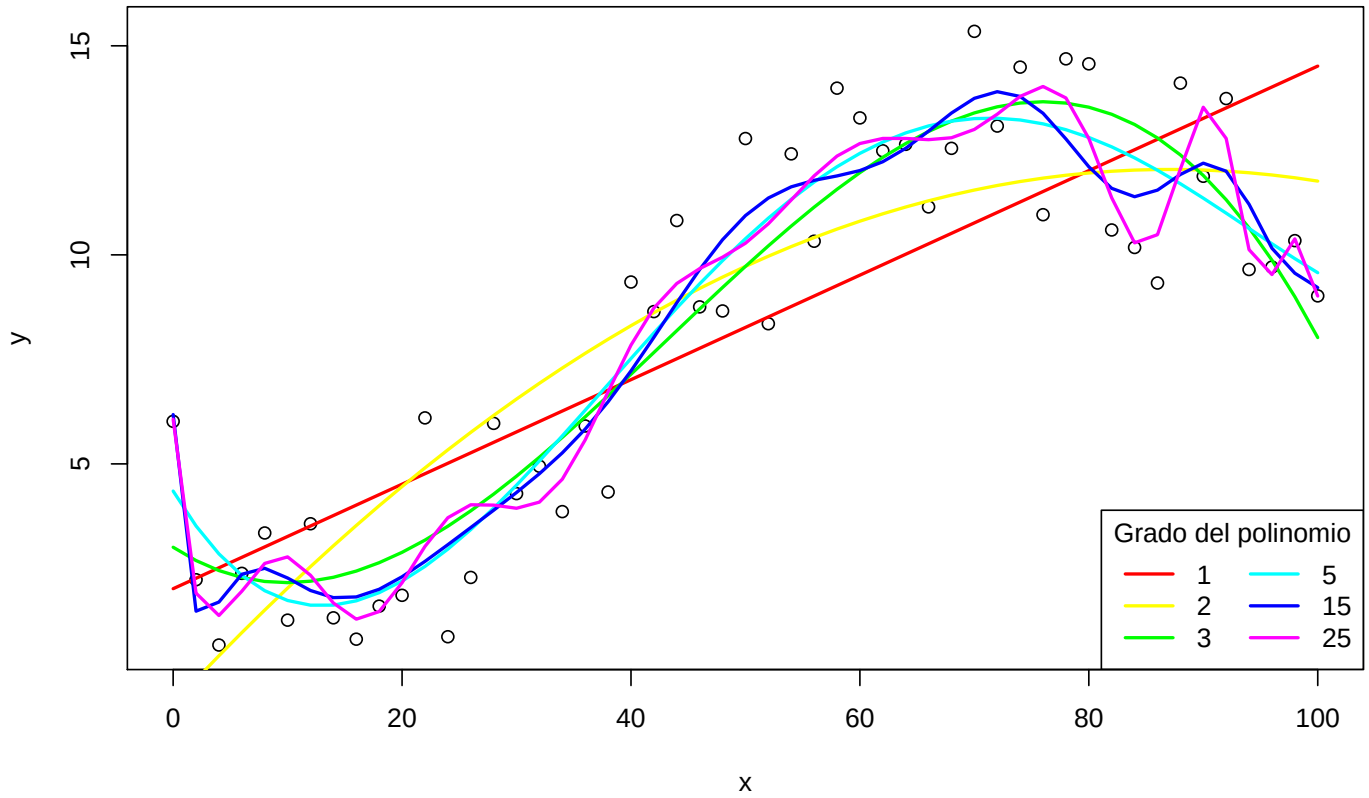
El **sesgo** de un modelo de predicción mide el error cometido al predecir la variable objetivo para un conjunto de datos determinado (conjunto de **entrenamiento**), mientras que la **varianza** es la cantidad en la que cambiaría mi predicción si la estimáramos con un conjunto de datos diferente al utilizado para la construcción del modelo (conjunto de **prueba/test**).



https://koldopina.com/equilibrio-varianza-sesgo/

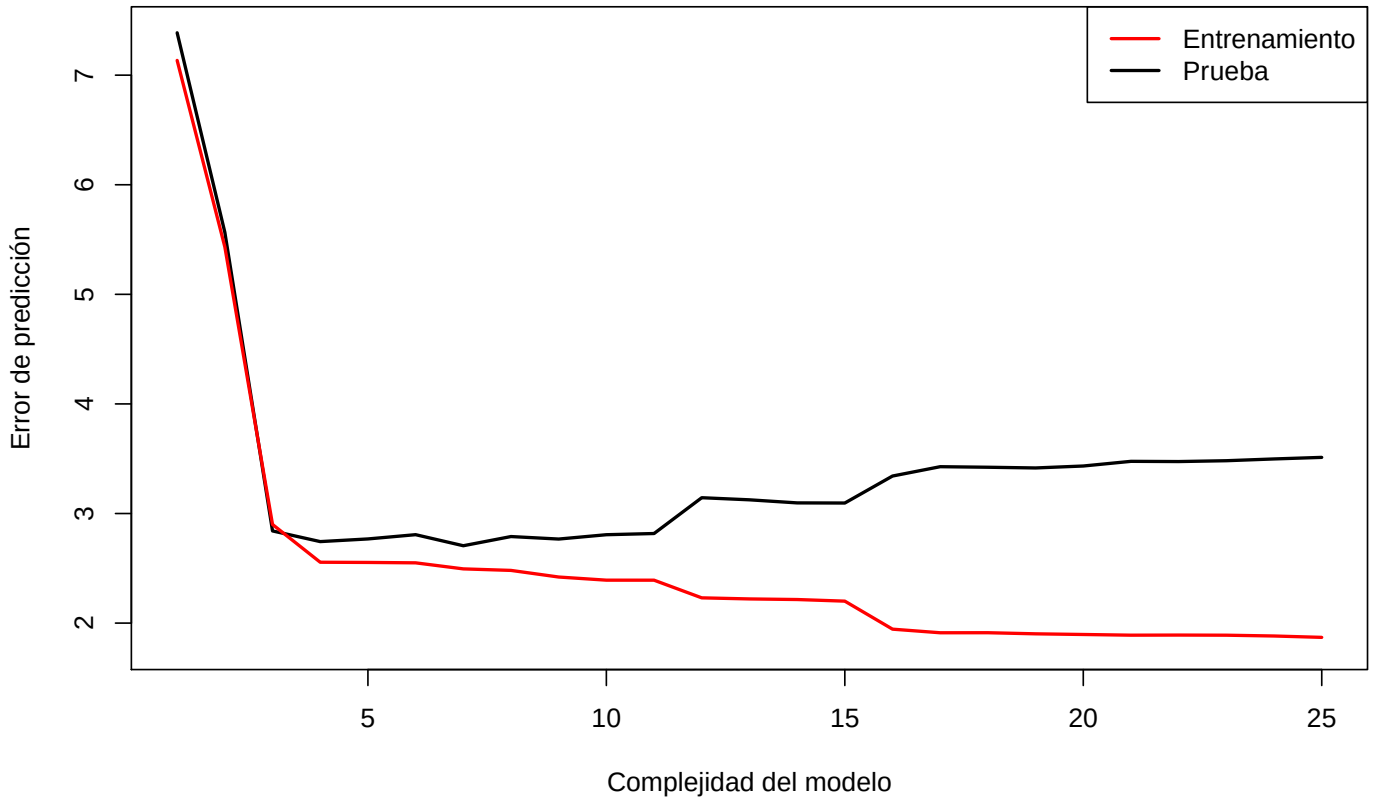
SESGO Y VARIANZA

- Lo ideal sería, por tanto, poder contar con modelos que tuvieran **poco sesgo y poca varianza** pero, generalmente, cuando disminuye uno, aumenta el otro, y viceversa.
- Los modelos más **complejos** suelen dar lugar a mejores resultados para el conjunto de entrenamiento pero, a cambio, suelen tener gran variabilidad, pues capturan las **especificidades** de dicho conjunto.

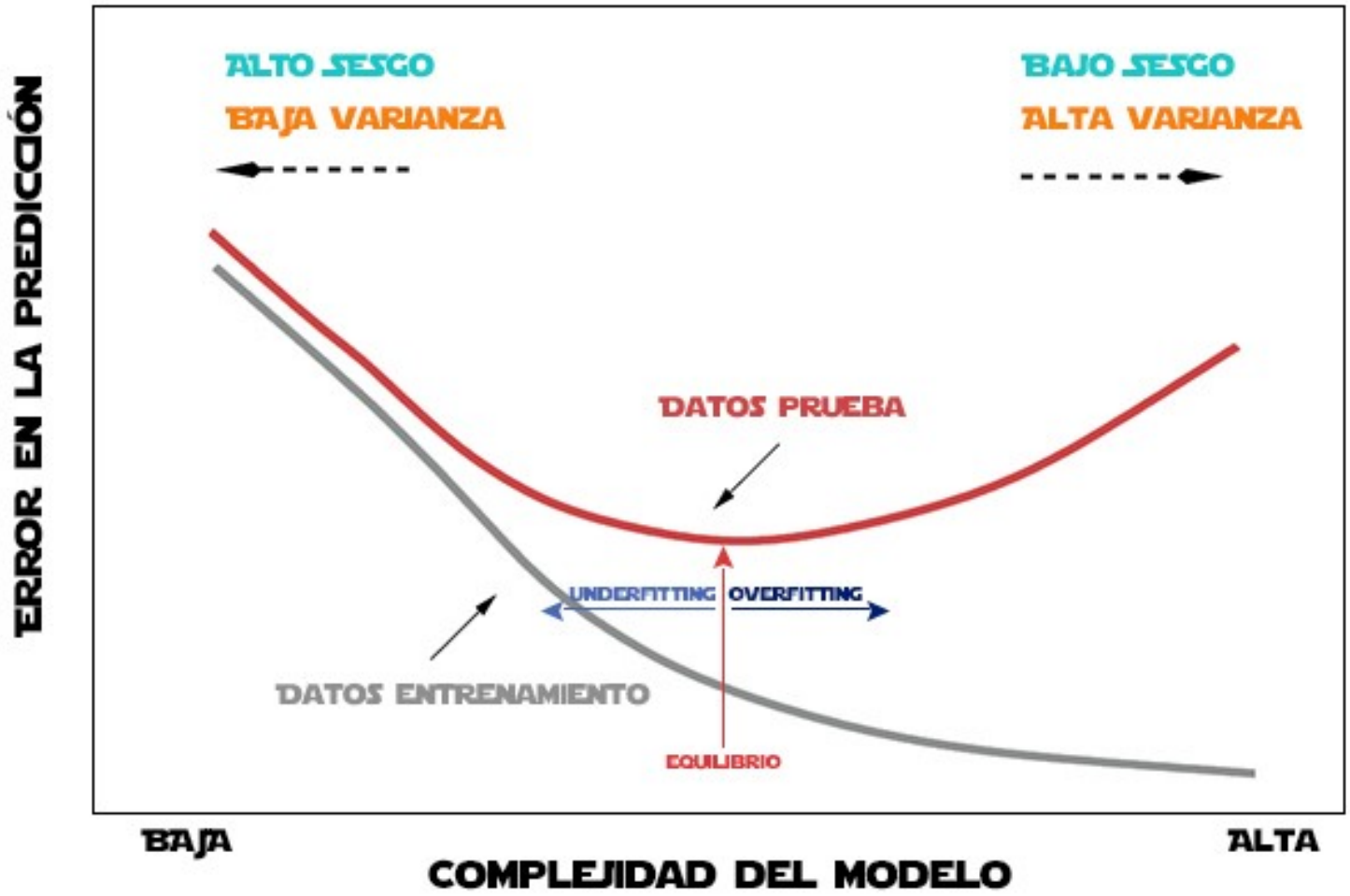


SESGO Y VARIANZA

- Lo ideal sería, por tanto, poder contar con modelos que tuvieran **poco sesgo y poca varianza** pero, generalmente, cuando disminuye uno, aumenta el otro, y viceversa.
- Los modelos más **complejos** suelen dar lugar a mejores resultados para el conjunto de entrenamiento pero, a cambio, suelen tener gran variabilidad, pues capturan las **especificidades** de dicho conjunto.



Modelos sencillos vs. modelos complejos



https://koldopina.com/equilibrio-varianza-sesgo/