Parte I:
Modelos paramétricos predictivos

Aprendizaje Estadístico

Junio 2025

Aida Calviño

Índice

El concepto de Aprendizaje Estadístico se refiere al **conjunto de técnicas y modelos derivados de la estadística y la informática**, utilizados para comprender relaciones entre variables y/o predecir resultados a partir de datos.

Estas técnicas son pilares clave en áreas como la inteligencia artificial y la ciencia de datos.

Algunos conceptos claves en este área son:

- Modelos supervisados vs. no supervisados (aquí nos centramos en los primeros).
- Modelos paramétricos vs. no paramétricos.
- Métodos de remuestreo (entrenamiento/prueba y validación cruzada).
- Selección de variables.
- Regularización.
- Interpretabilidad (equilibrar precisión predictiva y complejidad del modelo).

CLASIFICACIÓN DE LAS TÉCNICAS

Las técnicas de aprendizaje estadístico pueden ser **clasificadas** como supervisadas o no supervisadas.

En el caso de las **técnicas supervisadas**, ha de existir una variable, que recibe el nombre de variable objetivo o respuesta, que se desee **predecir de manera precisa** a partir de la información proporcionada por el resto de variables. En ocasiones, este tipo de técnicas recibe el nombre de técnicas predictivas.

Por el contrario, en las **técnicas no supervisadas**, no existe ninguna variable respuesta, por lo que el objetivo consiste en estudiar las **relaciones** existentes entre las variables y las observaciones del conjunto de datos. Reciben este nombre pues no existe ninguna variable que permita "supervisar" los resultados.

EJEMPLOS DE TÉCNICAS:

- Supervisadas: regresión lineal y logística, árboles de decisión, *KNN*, *random forest*, *gradient boosting*, redes neuronales, etc.
- No supervisadas: análisis de componentes principales, análisis cluster, escalamiento multidimensional, etc.

CLASIFICACIÓN DE VARIABLES SEGÚN SU FUNCIÓN

- Identificativas: Sirven para **identificar observaciones** y su utilidad es limitada.
- Objetivo o dependiente (solo en técnicas supervisadas): Es la variable que se pretende **predecir/modelizar**.
- Input, de entrada o independientes: Se trata del resto de variables a analizar (en el contexto supervisado, también se llaman predictoras).
- rechazadas: Son variables que no se van a poder utilizar para el objetivo de estudio y, por ende, son **eliminadas**.

CLASIFICACIÓN DE VARIABLES SEGÚN SU TIPOLOGÍA

- Cuantitativas: Se trata de variables asociadas a **cantidades numéricas**. En ocasiones, se hace la distinción entre continuas y discretas.
- Cualitativas, nominales o categóricas: Toman un número **finito** de valores.
- Dicotómicas o binarias: Variables nominales que toman sólo **dos** valores.
- Ordinales: Variables nominales cuyas categorias pueden **ordenarse** (aunque la distancia entre dos valores no ha de ser necesariamente la misma).

Índice

La gran diferencia entre el Aprendizaje Estadístico y la Estadística "clásica" es que el tamaño de muestra suele ser muy grande, lo que hace imposible la aplicación de inferencia en la predicción, debido a que la gran **potencia del test** hace que toda hipótesis sea rechazada.

Alternativamente, es habitual realizar una **partición** de los datos en dos submuestras: entrenamiento y prueba.

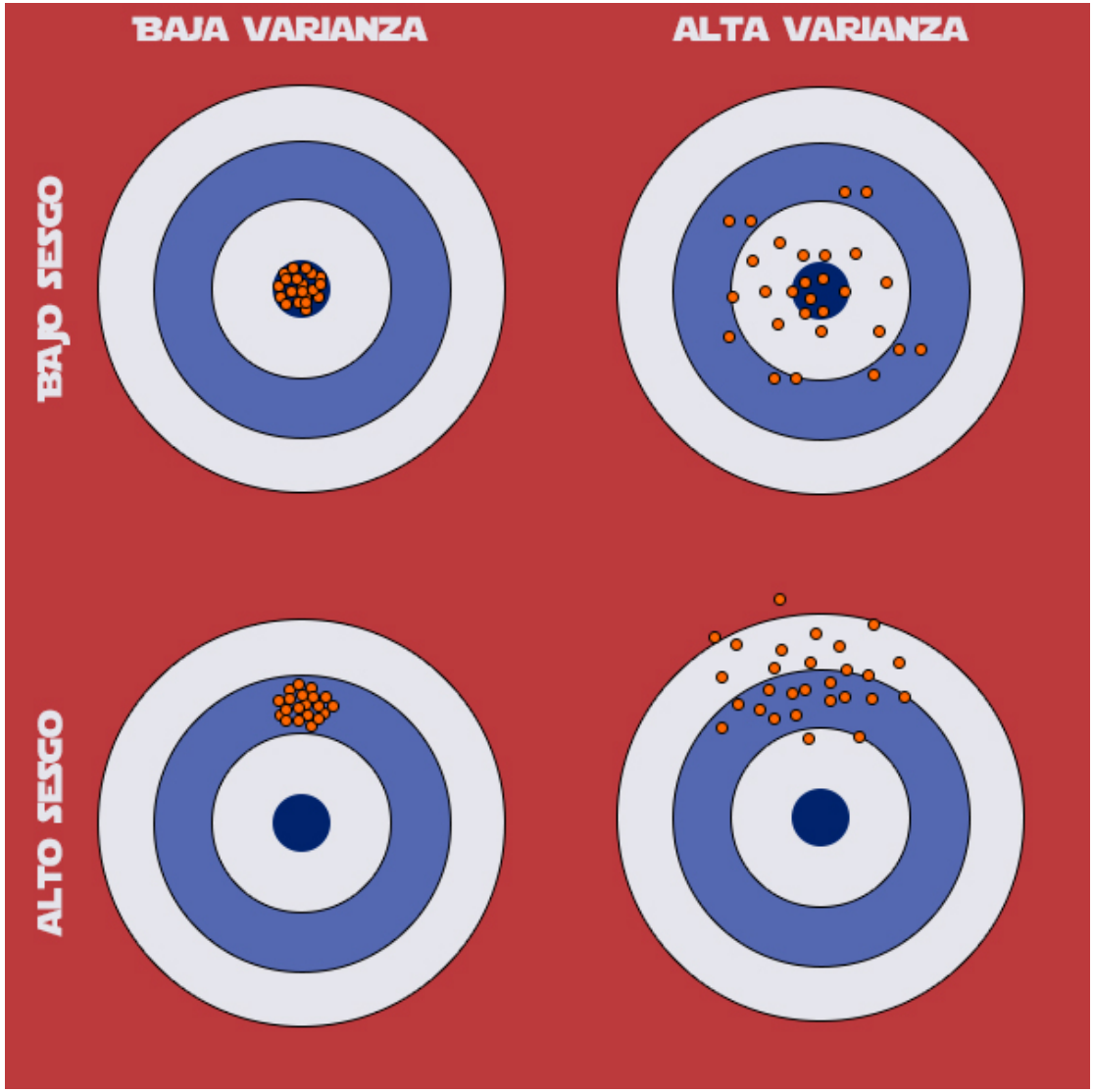
1. **Entrenamiento** (*training*): Sirve para **generar** y definir el modelo.
2. **Prueba** (*test*): Permite **evaluar** la bondad del modelo de manera independiente.

Cabe recordar que los modelos se construyen para ser aplicados en datos "desconocidos", por lo que es importante tener una **idea realista de su calidad** antes de aplicarlos.

Si este análisis de calidad se realiza sobre el mismo conjunto de datos que se ha utilizado para construir el modelo, puede dar lugar a métricas **artificialmente optimistas**, por lo que es preferible "probar" los modelos sobre un conjunto de datos distinto.

Así mismo, resulta de utilidad saber **si los modelos son estables**, en el sentido de que son igual de "buenos" cuando se aplican a distintos datos o si, por el contrario, a veces son "buenos" y otras, "malos".

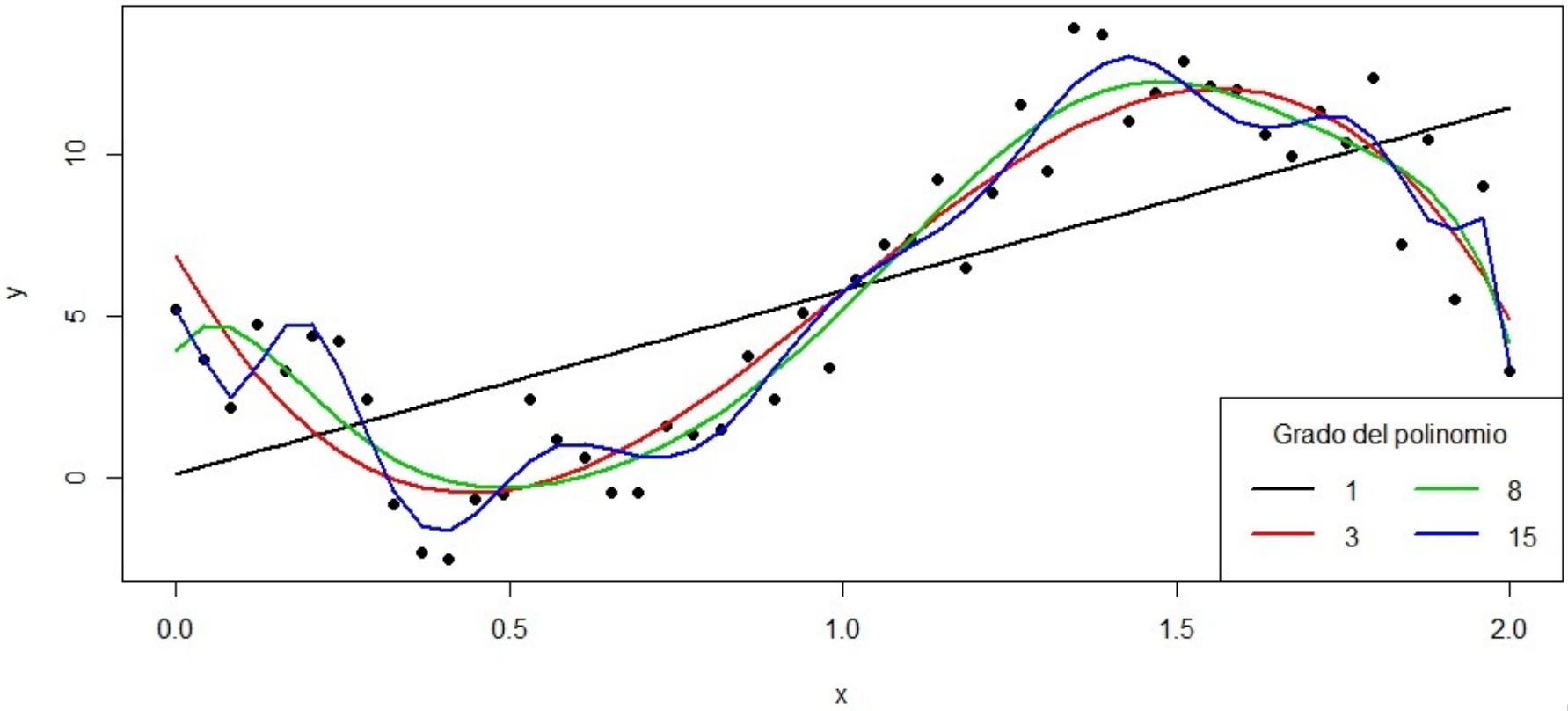
Esto nos lleva a la definición de dos conceptos clave: **el sesgo** de un modelo de predicción, que mide el error de predicción medio, y **la varianza**, que es la cantidad en la que cambia dicho error si aplicamos el modelo sobre conjuntos de datos diferentes.



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

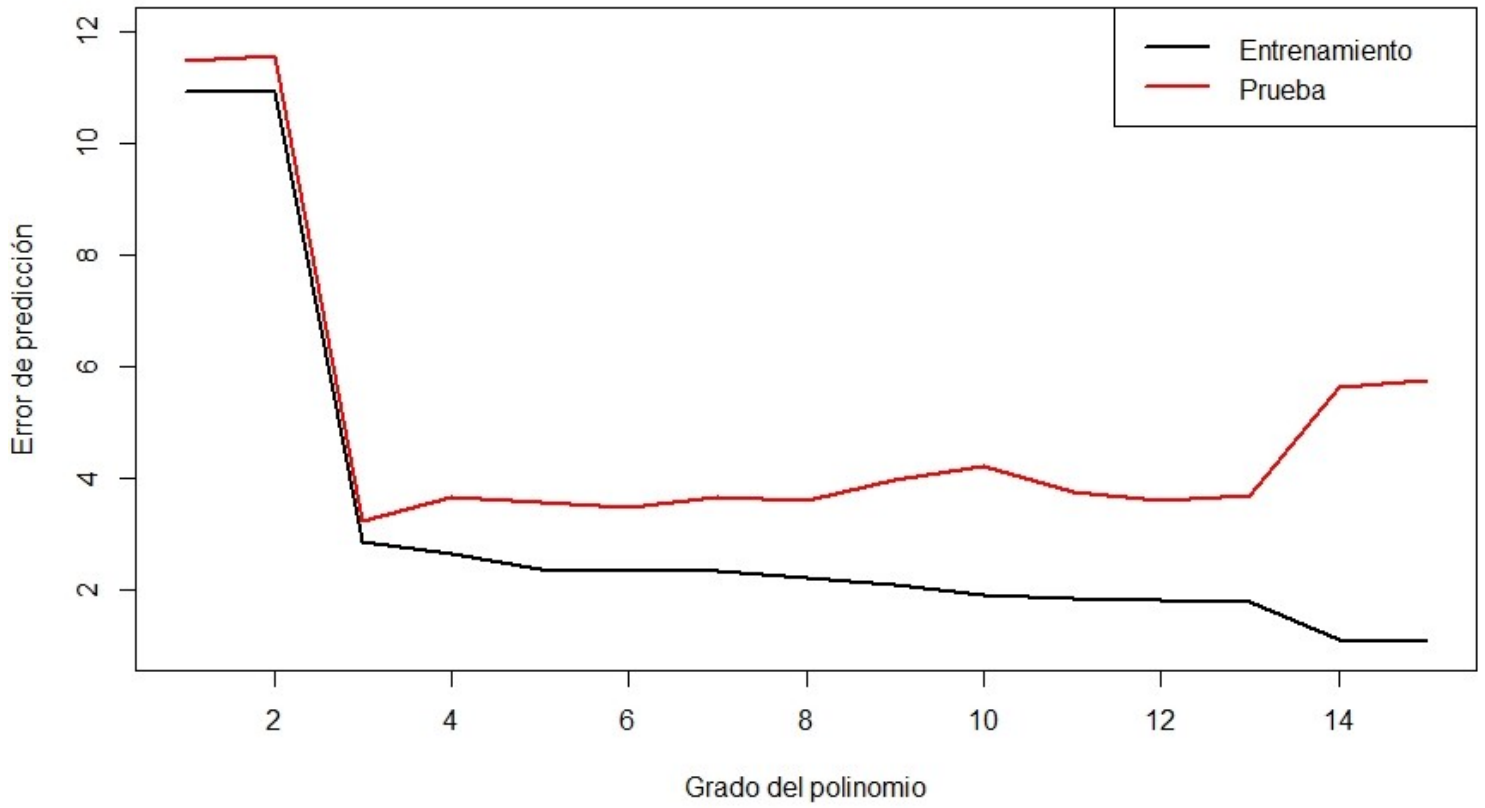
SESGO Y VARIANZA

- 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.

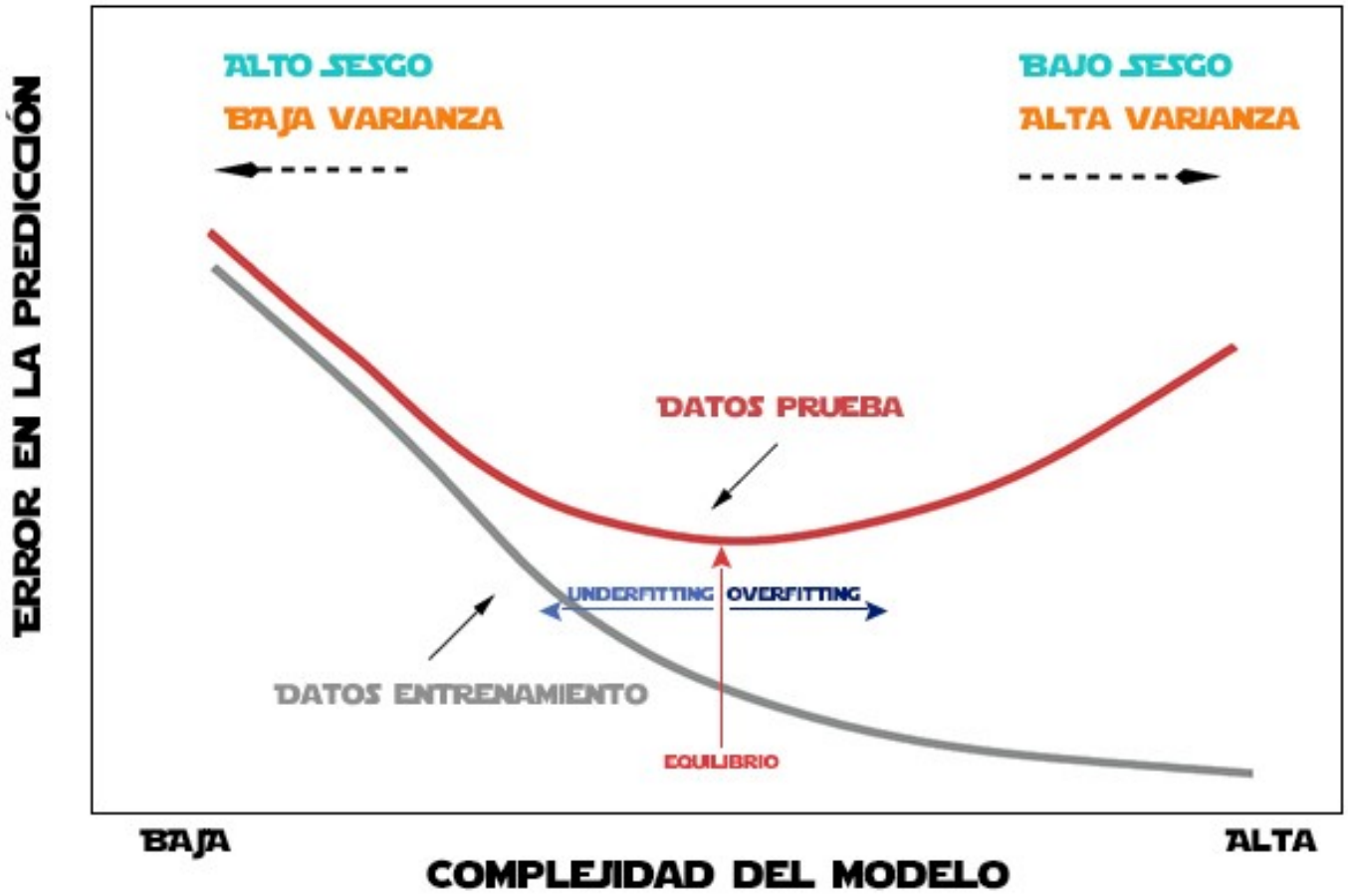


SESGO Y VARIANZA

- 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** (más variables y/o parámetros) 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 (sobreajuste).



Modelos sencillos vs. modelos complejos



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

Los modelos supervisados tienen por objetivo **predecir** una variable y (que recibe el nombre de dependiente u objetivo) a partir de un conjunto de m variables x_i (llamadas independientes o *input*) a través de una **ecuación**:

$$y = f(x_1, x_2, \ldots, x_m) + \epsilon,$$

donde ϵ representa el **error cometido** o, equivalentemente, la parte de la variable dependiente **no explicada** a partir de las variables independientes.

En el caso particular de la **regresión lineal**, la ecuación anterior se reduce a, pues se asume que la relación entre las variables input y la objetivo es de tipo lineal:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m + \epsilon,$$

donde y es una variable aleatoria **cuantitativa**, β_0 representa el valor que toma la variable objetivo cuando todas las variables input toman el valor 0 y los parámetros restantes representan **cuánto aumenta o disminuye** la variable objetivo por cada **incremento unitario** de las variables input.

A partir del modelo, es posible predecir el valor de y para un determinado individuo, **conocidos** los valores que toman las variables **input**:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$$

ESTIMACIÓN DE LOS PARÁMETROS

Para poder llevar a cabo la predicción, es necesario conocer el valor de los **parámetros**. Como no es así, debemos **estimarlos** a partir de los datos disponibles.

Para ello, buscaremos para qué valor de los parámetros se **minimiza el error** cometido por el modelo, que viene dado por: $y - \hat{y}$.

El **estimador de mínimos cuadrados**, aquel que minimiza la suma de cuadrados de los errores, viene dado por:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y},$$

donde

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_m \end{pmatrix} \quad \boldsymbol{Y} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{21} & \dots & x_{m1} \\ 1 & x_{12} & x_{22} & \dots & x_{m2} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & x_{1n} & x_{2n} & \dots & x_{mn} \end{pmatrix}$$

El modelo de regresión lineal clásico se sustenta en **las hipótesis** de normalidad de los residuos y homocedasticidad. No obstante, nótese que hasta el momento no se ha utilizado **ninguna**, por lo que los estimadores minimo cuadráticos son **siempre válidos**.

Con lo que sí hay que tener cuidado es con el uso de inferencia en este contexto, pues la validez de los p-valores sí depende del cumplimiento de dichas hipótesis.

Sin embargo, aunque no se imponga ninguna hipótesis, es importante realizar algunos comentarios sobre la **estimación de los parámetros**:

- Debido a la forma en la que se obtienen los estimadores de los parámetros, ni la variable objetivo ni las input pueden contener **datos ausentes**.
- Así mismo, es importante que no existan **datos atípicos** en las variables input, pues pueden desvirtuar los resultados.
- No se pueden incluir en el modelo dos variables independientes que estén **muy correlacionadas** pues, en ese caso, no es posible invertir la matriz $\boldsymbol{X'X}$.
- El número de parámetros incluidos en el modelo ha de ser **muy inferior** al de observaciones, para evitar problemas en la estimación y de sobreajuste.
- Para poder incluir variables independientes **categóricas** en el modelo deben transformarse previamente en variables *dummy*, fijando uno de sus niveles como referencia. En esos casos, la interpretación de los parámetros asociados a dicha variable consiste en el **aumento/disminución** de la variable objetivo entre los individuos de la categoría en cuestión y la **de referencia**.

Una forma de **evaluar el modelo** es a partir de la suma de cuadrados de los errores ***SSE*** (que mide el error cometido al usar el modelo), de la suma de cuadrados del modelo ***SSM*** (que mide la información contenida en el modelo) y la suma de cuadrados total ***SST*** (que mide el error en el que se incurre si no hay modelo).

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{SST} = \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{SSE} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{SSM}$$

Los valores de la *SSE* **no están limitados** y, por ende, resulta complicado saber si el modelo está aportando predicciones de calidad o no. Por ello, se suele recurrir al R^2, que mide la proporción de información explicada por el modelo:

$$R^2 = \frac{SSM}{SST} = 1 - \frac{SSE}{SST}$$

Este indicador **está acotado** y toma valores entre 0 (la calidad del modelo coincide con la del "no modelo") y 1 (el modelo es perfecto y no comete errores).

- Es importante destacar que, desde el punto de vista de la partición de **entrenamiento**, incluir variables en el modelo siempre supone una **mejora en la SSE**. No obstante, esta mejora puede no reflejarse en la partición de test/prueba, indicando **sobreajuste** (el modelo ha recogido "demasiada" información de los datos disponibles).
- Por ello, es recomendable evaluar los modelos siempre en las **dos particiones** pues, si existen grandes diferencias entre ellos, el modelo puede estar **sobreajustado**.
- Por otro lado, a la hora de determinar la **utilidad de las variables input**, puede resultar de utilidad obtener lo que se conoce como **suma de cuadrados de tipo III**.
- Esta suma de cuadrados permite cuantificar **cuánto aumenta la suma de cuadrados de los errores** (en el conjunto de datos de entrenamiento) debido a la eliminación de cada variable input. La suma de cuadrados de tipo III se obtiene para cada variable input y, por tanto, nos permite hacernos una idea sobre la utilidad de las mismas.
- En particular, cuando esta cantidad es **pequeña**, la variable correspondiente **carece de poder predictivo** pues el modelo no empeora significativamente al eliminarla del mismo.

Uno de los pasos más importantes en la fase de modelización es la **selección de las variables** que van a formar parte del modelo, pues esto será lo que determine su calidad.

Dado que generalmente **resulta inviable probar todos los modelos** que se pueden construir con las combinaciones de todas las variables independientes (tomándolas de una en una o varias al mismo tiempo), se hace necesario disponer de una metodología que nos ayude a la hora de tomar la decisión sobre **cuántas y qué variables explicativas utilizar**.

Aunque no son la única opción, nos centramos en los $\mathbf{3}$ **métodos de selección de variables** más utilizados:

- **Backward o hacia atrás**: Este método consiste en, partiendo del modelo que contiene todos los posibles efectos (variables y/o interacciones), ir eliminando **una a una** las variables que menos influyan en el modelo hasta que el modelo empeore con la eliminación de cualquiera de las variables restantes.

  El principal **inconveniente** que tiene este método es que, si el modelo "con todo" es muy grande, puede tardar mucho en completarse o, incluso, ser infactible (si $n < p$ o hay variables linealmente dependientes).

  Una vez que un efecto se elimina, **no puede volver a entrar en el modelo**.

- **Forward o hacia delante**: Este método consiste en, partiendo desde cero, ir introduciendo **uno a uno** los efectos que mayor mejora produzcan en el modelo hasta que no haya ningún efecto más fuera del modelo que **aporte información**.

  Una vez que un efecto entra en el modelo, **no puede salir**.

- **Stepwise o paso a paso**: Este método es una **mezcla** de los anteriores. El método es similar al forward, salvo por que **se pueden eliminar** los efectos que han entrado en el modelo (ya que la entrada de alguno puede hacer no significativo el aporte de otro). La eliminación de los efectos se hace de acuerdo al método backward.

  Por lo tanto, en cada paso se evaluan todos los posibles efectos a eliminar y a introducir y se selecciona aquella acción que mayor mejora produzca en el modelo.

  Es recomendable incluir un número **máximo de iteraciones** para paliar la posible aparición de bucles de entrada-salida de un mismo efecto.

La mejora/empeoramiento en los modelos se puede medir de distintas formas.

CÁLCULO DE LA MEJORA EN LOS MODELOS DEL PROCESO ITERATIVO

Como ya se ha comentado previamente, desde el punto de vista del conjunto de datos de **entrenamiento**, cuantas **más variables** tiene un modelo, menor error comete. Por lo tanto, si evaluamos los modelos en dicho conjunto de datos en cada paso del proceso iterativo, la conclusión siempre será que el **mejor modelo** es el que más variables contiene.

No obstante, como ya hemos visto, que una variable disminuya el *SSE* del conjunto de datos de entrenamiento, **no siempre implica una mejora** en la capacidad predictiva del modelo. Así pues, lo que se suele hacer en esos casos es recurrir a **medidas de evaluación alternativas** que, por un lado, evalúen la mejora producida en la *SSE* y, por otro, tengan en cuenta la mayor **complejidad del modelo** (solo se consideran útiles aquellas variables que mejoren "mucho" la calidad del modelo). Las más habituales son:

- **AIC** (Akaike information criterion): $n \ln(\frac{SSE}{n}) + 2\tau$
- **BIC ó SBC** (Bayesian o Schwarz information criterion): $n \ln(\frac{SSE}{n}) + \tau \ln(n)$

Nótese que el **primer sumando coincide** para los dos criterios, por lo que la diferencia entre ellos se reduce a la **penalización del número de parámetros** (τ), siendo el BIC/SBC el que más penaliza, dando lugar, por tanto, a modelos con menos parámetros.

Existen argumentos a favor y en contra de unos y otros criterios en la literatura, por lo que lo recomendable es **utilizarlos todos** para generar modelos potencialmente buenos y después compararlos con otra estrategia.

La división *train/test* puede resultar insuficiente si, además de construir y evaluar modelos, se desean **comparar varios entre sí** (tal y como acabamos de comentar) pues, de nuevo, se pueden obtener medidas de evaluación artificialmente optimistas.

Para solventar este problema se podría recurrir a una **tercera partición de los datos**, lo que implicaría reducir la cantidad de observaciones disponibles para el entrenamiento.

Por ello, la técnica habitual para comparar modelos es la **validación cruzada repetida**.

VALIDACIÓN CRUZADA REPETIDA EN K-PARTES

Este método consiste en dividir el conjunto de datos de entrenamiento en **k submuestras** e iterativamente construir el modelo con todas las observaciones menos las de una submuestra y evaluarlo a continuación con las observaciones **excluidas**.

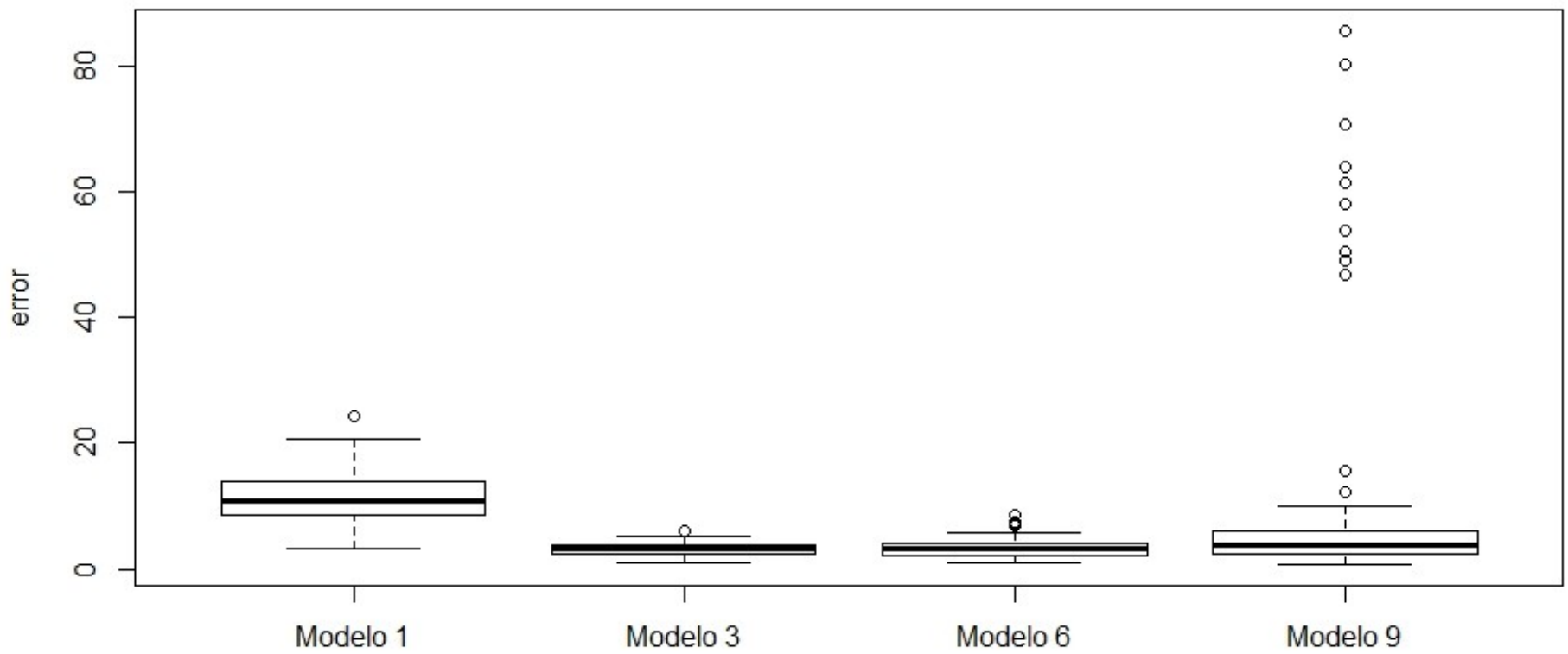
Tiene la ventaja de que **todas las observaciones son predichas una vez** sin formar parte de la construcción del modelo pero **también contribuyen a la construcción** del modelo en el resto de iteraciones.

Este proceso se repite con **distintas semillas** para evitar el posible efecto de la aleatoriedad.

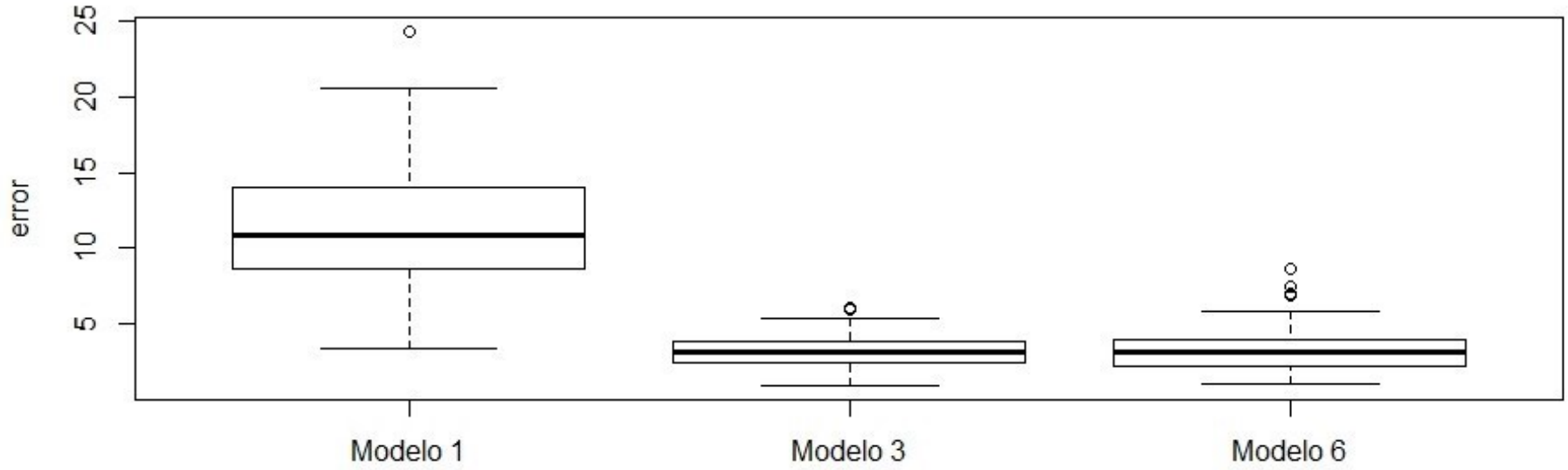
Es el método de remuestreo **más fiable y utilizado**, aunque requiere **bastante esfuerzo computacional**.

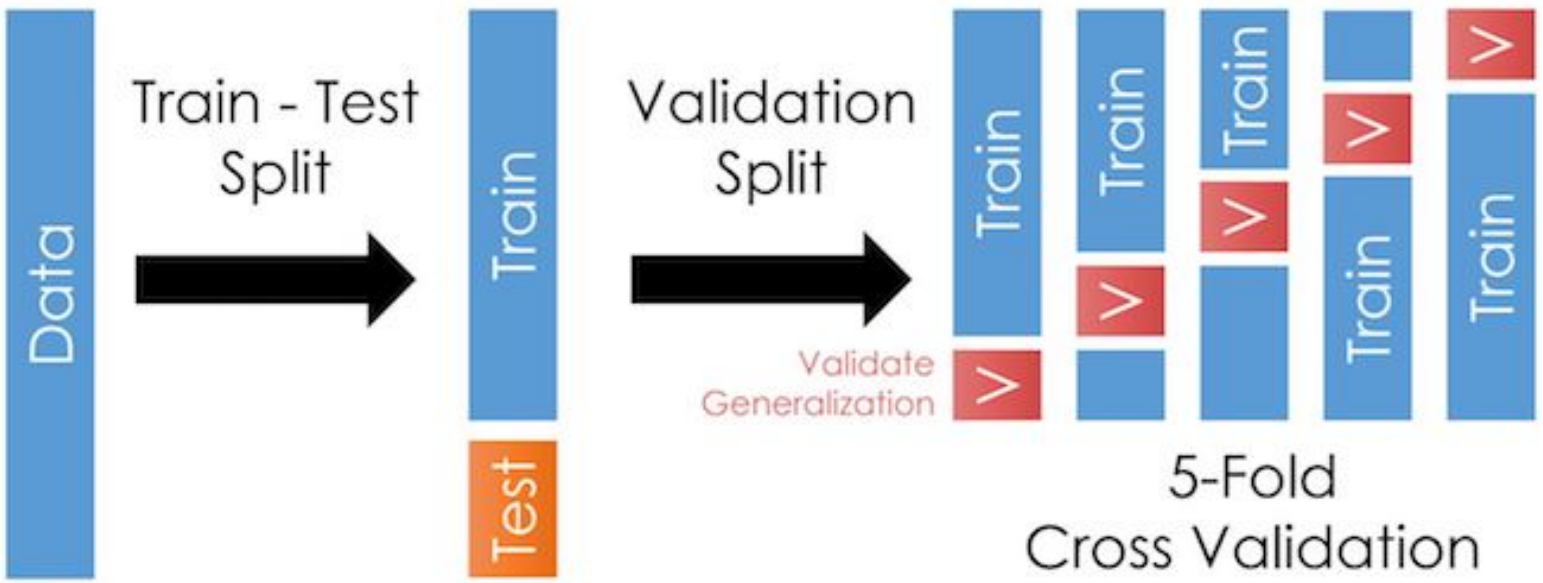
A la hora de **comparar modelos**, es habitual representar en un *boxplot* los resultados para evaluar el **comportamiento medio** (sesgo) de cada modelo, junto con su **variabilidad**.

Comparación de modelos con validación cruzada



Comparación de modelos con validación cruzada





Fuente: https://www.samlau.me/test-textbook/ch/15/bias_cv.html

En ocasiones, principalmente en presencia de multicolinealidad y cuando el número de observaciones no es suficientemente grande, el enfoque anterior puede no ser válido a la hora de seleccionar variables. Veamos una alternativa a dichos algoritmos:

REGRESIÓN PENALIZADA

Los modelos de **regresión penalizada** se basan en la misma formulación del modelo de regresión lineal, salvo que los parámetros se obtienen como:

$$\boldsymbol{\beta} = \arg\min \left\{ \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \|\boldsymbol{\beta}\| \right\},$$

donde $\hat{y}_i$ es la predicción de la observación i con el modelo de regresión lineal, $\lambda \geq 0$ es el parámetro de penalización y $\|\boldsymbol{\beta}\|$ es una norma aplicada al vector de parámetros del modelo.

El modelo penalizado más utilizado es el llamado ***Elastic Net*** (ENet), que viene dado por:

$$\boldsymbol{\beta} = \arg\min \left\{ \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \left(\alpha \sum_{i=1}^{t} |\beta_i| + (1-\alpha) \sum_{i=1}^{t} \beta_i^2 \right) \right\},$$

con t el número de parámetros del modelo saturado, el que cuenta con todos los posibles efectos pero sin contar con el término independiente, y $\alpha \in [0, 1]$.

Cuando $\alpha = 0$, el modelo recibe el nombre de ***Ridge*** (penalización cuadrática), mientras que si $\alpha = 1$, nos encontramos con el modelo ***LASSO*** (penalización absoluta). Valores intermedios proporcionan modelos con penalización mixta.

REGRESIÓN PENALIZADA

- Los modelos penalizados persiguen **reducir la magnitud de los parámetros** estimados y, por ende, el posible sobreajuste.
- La magnitud de la **penalización** viene dada por el valor del parámetro λ, dando lugar al modelo clásico si $\lambda = 0$ y al *no modelo* si $\lambda = \infty$.
- La **penalización absoluta** propicia una selección de variables "real", pues permite obtener coeficientes nulos (si la penalización es suficientemente grande).
- Las variables *input* han de **estandarizarse** para evitar que haya variables "dominantes" debido a la diferencia de escala. Esto hace que se pierda cierta interpretabilidad.
- Dependiendo de las características concretas de los datos, los **valores óptimos de λ y α** varían, por lo que se debe probar con varios valores y seleccionar el que mejores resultados ofrezca en la validación cruzada.
  - Este proceso recibe el nombre de afinación (aunque se suele recurrir al anglicismo *tuneo*).
  - Dado que la validación cruzada es un proceso aleatorio, para reducir aún más las posibilidades de sobreajuste, se puede recurrir a la **regla 1se** (*one standard error rule*), que consiste en seleccionar el modelo más sencillo que proporcione una métrica de calidad que se encuentre a menos de una desviación típica del óptimo.