MODELOS DE REGRESIÓN

¿Qué hace un algoritmo de aprendizaje para resolver una tarea de regresión?

El algoritmo a través de los datos construye una función que al aplicarla a un vector de entrada X, producirá como salida un estimado Y. Esta función es el modelo de regresión.

Objetivo: Que el estimado Y se aproxime al valor real tanto como sea posible.

¿Cuáles son los tipos de modelos de regresión que se pueden construir?

Modelos de Regresión

* Simple: Relación entre una variable independiente y la variable objetivo.
* Modelos Lineales
* Modelos No Lineales
* Múltiple: Relación entre mas de una variable independiente y la variable objetivo.
* Modelos Lineales
* Modelos No Lineales

Algoritmos para construir modelos de Regresión:

* Regresión Lineal
* Regresión de Lasso y Ridge
* Regresión Polinomial
* Arboles de decisión para regresión
* Maquinas de vectores de soporte para regresión
* Redes Neuronales para regresión
* Random Forest para regresión
* K-vecinos mas cercanos para regresión

Aplicaciones de los modelos de regresión

* Previsión de demandas

En retail, es importante determinar la demanda por parte de los usuarios para poder realizar una gestión eficiente del stock de productos. En este contexto, la regresión apoya en la construcción de modelos que permiten estimar la demanda futura (variable objetivo) con base en el comportamiento pasado. Este conocimiento es útil para optimizar la distribución de los productos, el control de inventarios y el diseño de ofertas tomando en cuenta los patrones de consumo de los usuarios.

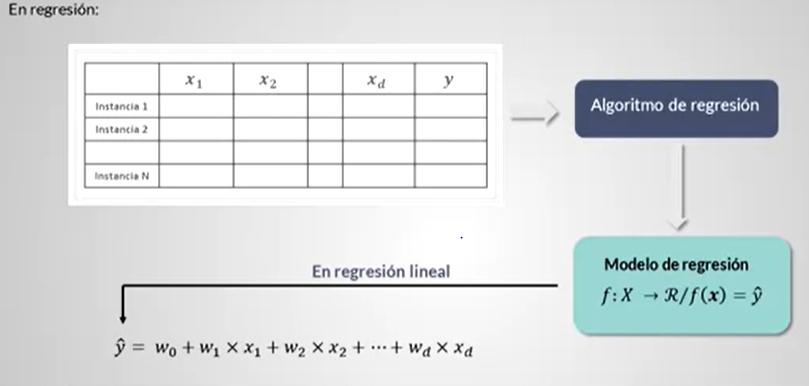
* Optimización de la producción

En agricultura, se utiliza la regresión para construir modelos que permiten estimar la producción (variable objetivo) dependiendo del área cultivada y el tipo de cultivo. En el conjunto de datos se incluyen también variables relacionadas con la ubicación y la estación del año, entre otras que, a juicio de los expertos, influyen en la producción. Este conocimiento apoya a los agricultores en la determinación del área a cultivar para obtener el rendimiento esperado, les permite monitorear los cultivos y tomar acciones oportunas para mitigar problemas con el fin de alcanzar los objetivos de producción, entre otras ventajas.

* Predicción del consumo energético

la predicción del consumo energético (variable objetivo) a partir de datos de sensores, con el fin de determinar hábitos de consumo por zonas, por tipos de usuario y de edificación. Así, se podría conocer cuánta energía consumimos y entender cómo lo hacemos, con el propósito de mejorar nuestros hábitos y lograr un consumo responsable y sostenible de la energía. A los responsables de la administración energética, este conocimiento les permitiría adecuar o mejorar los planes de gestión de las fuentes renovables y no renovables.

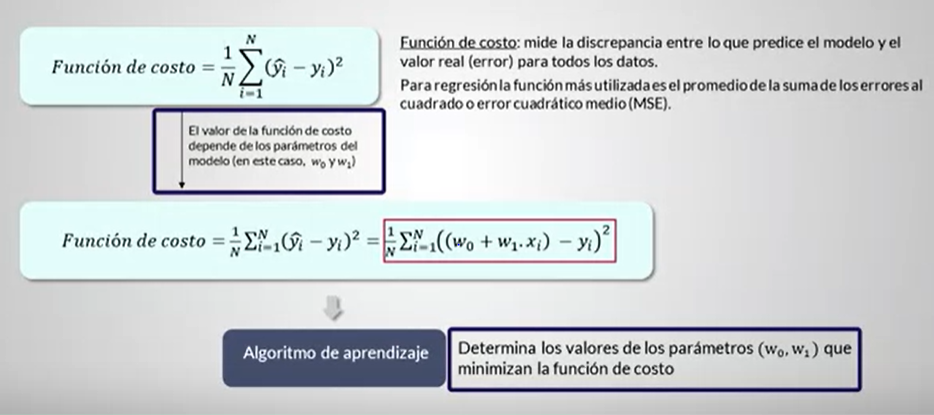
MODELOS DE REGRESIÓN LINEAL SIMPLE



El modelo de regresión se expresa como una combinación lineal de las variables de entrada (explicativas). En esta combinación lineal los coeficientes de regresión son los que determinan como cada variable afecta la predicción.

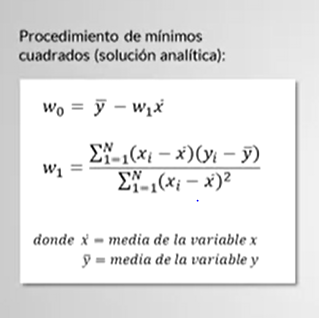
El algoritmo de este modelo nos determina la mejor recta lineal que se ajusta a los valores reales de Y, ósea, encuentra los coeficientes de regresión tal que la recta sea la que mejor representa al conjunto de datos.

El algoritmo de aprendizaje selecciona la solución que mejor se ajusta a los datos a través de una función de costo (minimiza la función de costo)



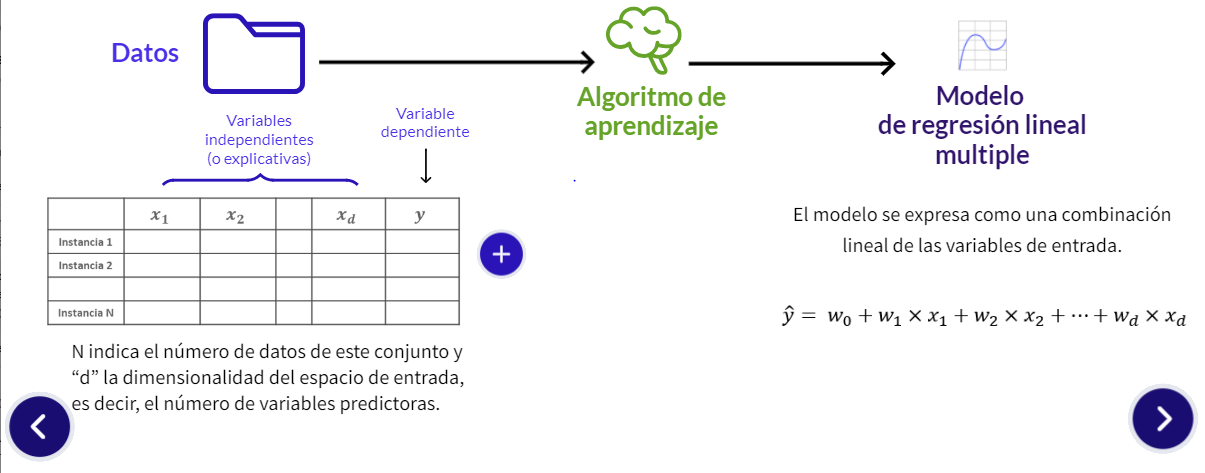
Este proceso de minimizar la función de costo se puede resolver de dos maneras, a través de un Enfoque analítico o un Enfoque basado en optimización numérica.

Para el Enfoque analítico se utiliza el procedimiento de mínimos cuadrados

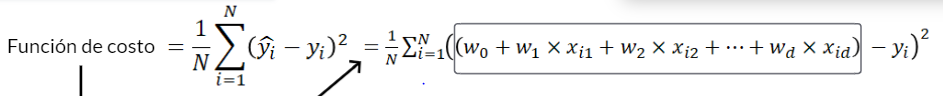


* El algoritmo de aprendizaje utiliza una función de costo que mide la discrepancia entre el estimado y el valor real.
* Las variables de esta función de costo son los parámetros del modelo (coeficientes de regresión).
* El algoritmo mediante un proceso analítico o de optimización, determina los valores de los parámetros del modelo que minimizan la función de costo.
* La función de costo para la regresión lineal es convexa y tiene un único mínimo.

MODELOS DE REGRESIÓN LINEAL MULTIPLE



Lo que hace el algoritmo que mejor selecciona la solución de ajuste a los datos, (los valores wo,w1,w2, que dan lugar al mejor hiperplano de mejor ajuste), es que determina los valores de dichos parámetros minimizando la función de costo, es decir, minimiza el error.



Este problema de minimizar la función de costo se puede resolver de dos maneras, a través de un enfoque analítico o un enfoque basado en optimización numérica (descenso por el gradiente).

A través del enfoque analítico:

Evaluar un modelo de regresión

Las Métricas más utilizadas para evaluar un modelo de regresión son:

* Coeficiente de Determinación – Rˆ2

Si es cercana a 1, el modelo se ajusta a los datos y lo contrario si se acerca a 0.

Proporciona una medida de la proporción de la variación en la variable objetivo que es explicada por el modelo. Si, por ejemplo, R2=86.3, podríamos afirmar: “86.3% de la variabilidad en la variable respuesta es tomada en cuenta por la combinación lineal entre esta y las variables predictoras”. Dependiendo del contexto, podría no obtenerse valores altos de R2.

* Error Cuadrático Medio

Entre mas cercana a cero, significa que el error del modelo es poco.

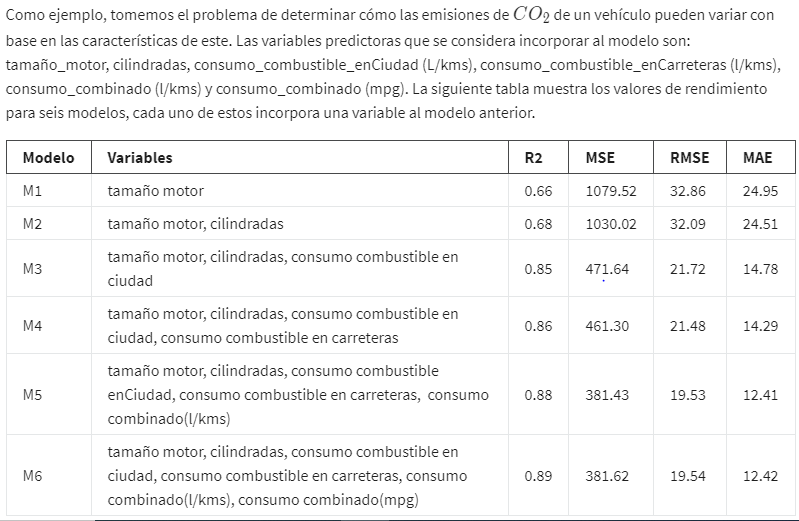
Para la interpretación, se suele utilizar la raíz cuadrada de este valor (RMSE) conocida como error típico o erros estándar. Una ventaja es que el valor se expresa en las unidades de medida de la variable objetivo. Al elevar al cuadrado el error de predicción se produce una sobreestimación cuando la diferencia entre el valor estimado y el valor real es grande. También es utilizada como función de costo por el algoritmo de aprendizaje ya que es derivable.

* Error Absoluto Medio

Entre mas cercana a cero, significa que el error es poco

Al igual que MSE, mide el error de predicción. Pero resulta más adecuada cuando en el conjunto de datos hay valores extremos.

Veamos un ejemplo del comportamiento de estas métricas en diferentes modelos cuando el numero de variables independientes aumenta, en teoría si la nueva variable es útil podría aumentar significativamente, para MSE y MAE debería entonces ser menor el error, aunque si la nueva variable no tiene poder de predicción estos valores podrían aumentar.

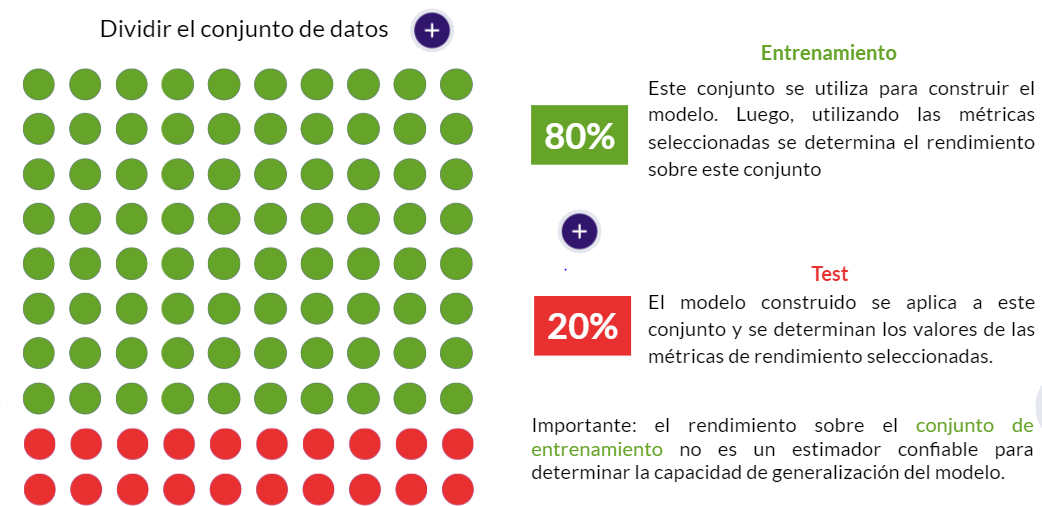


Para todos los modelos, los atributos resultaron estadísticamente significativos. Sin embargo, se observa que la incorporación de una tercera variable en el modelo M3, en este caso “consumo\_combustible\_enCiudad”, produce una mejora de rendimiento muy significativa, la cual se evidencia en todas las métricas; se puede afirmar entonces que esta variable es importante para la predicción. A partir de este modelo, la incorporación de más variables produce mejoras, pero no tan marcadas. Hasta llegar al último modelo con todos los atributos, en el cual la incorporación de la variable “consumo\_combinado(mpg)” produce un ligero empeoramiento en las métricas MSE, RMSE y MAE, más no así en R2.

CALIDAD DE LAS PREDICCIONES SOBRE DATOS NUEVOS

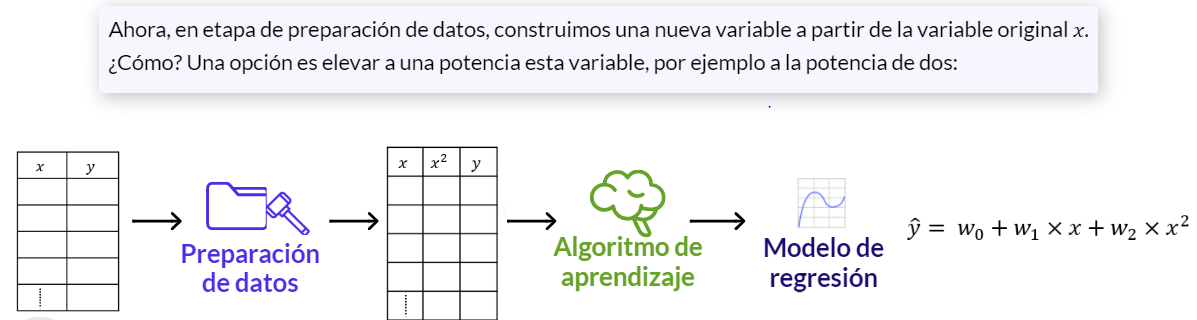


El objetivo es derivar un modelo que generalice bien y que pueda ser utilizado para realizar predicciones válidas para nuevos datos. Para determinar el rendimiento de generalización se utilizan métricas de rendimiento sobre datos no vistos durante el aprendizaje.

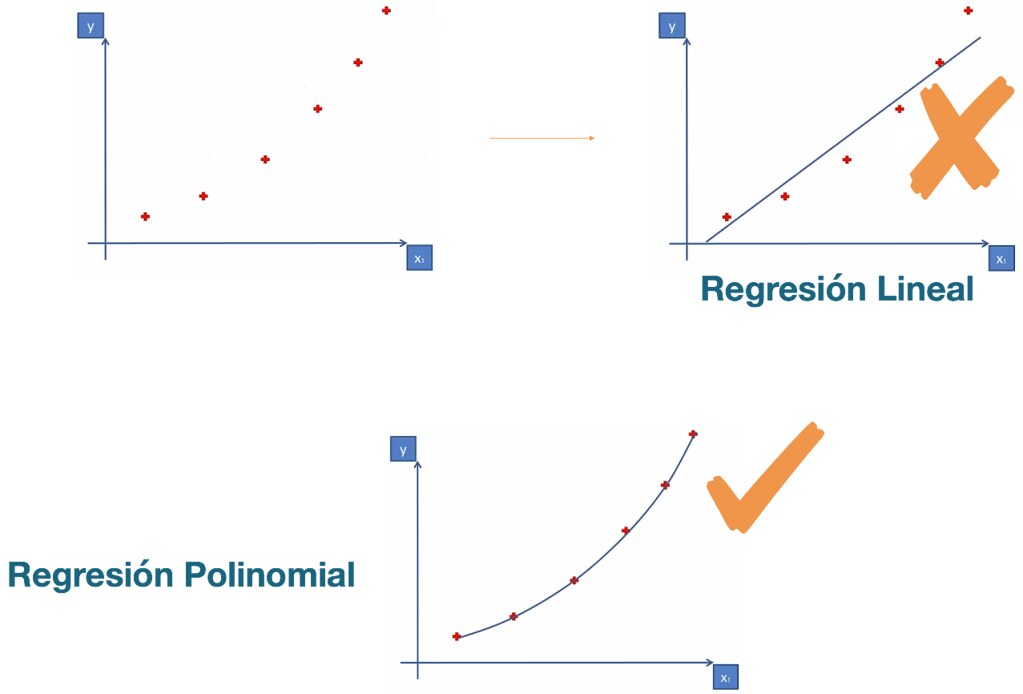


MODELOS DE REGRESION POLINOMICA

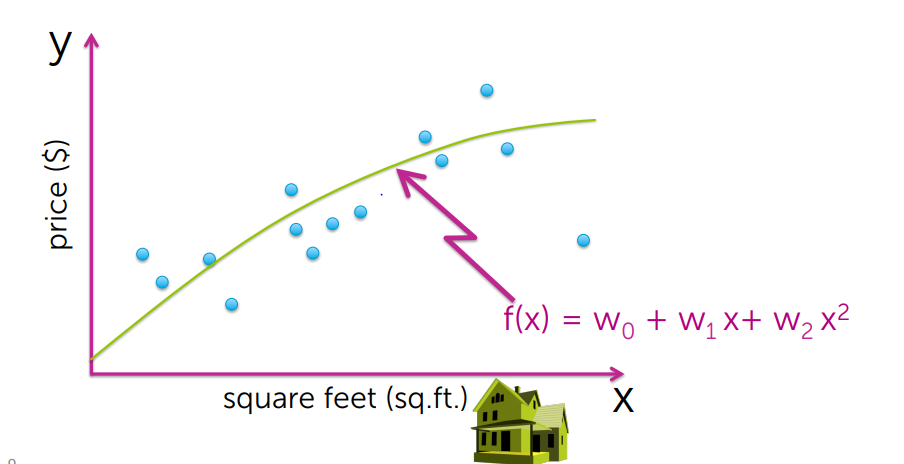
Se usan para construir modelos No Lineales: Relación No Lineal entre variables explicativas y la variable objetivo.

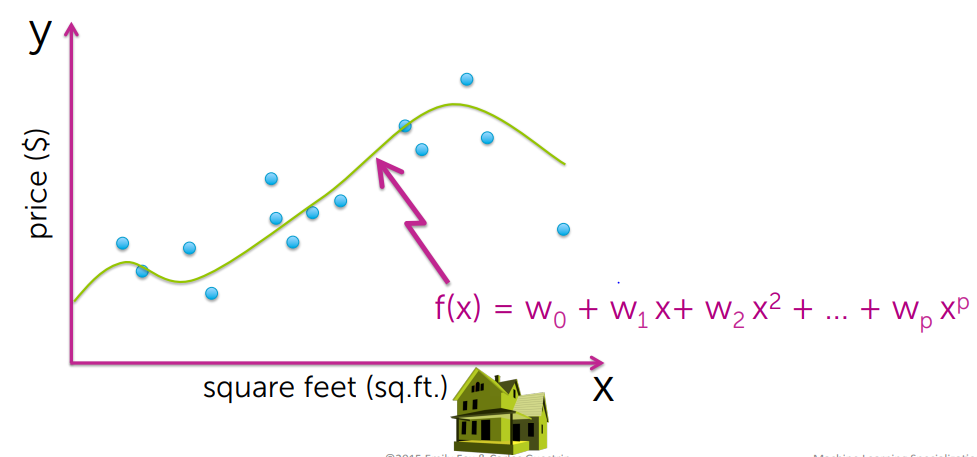


en la regresión polinomial la relación entre las variables predictoras y la variable objetivo se modela por un polinomio de grado p. Al aplicar la transformación polinomial a las variables de entrada se generan nuevas variables, cuyo número dependerá del grado del polinomio. A partir de estas, se define la combinación lineal que representa al modelo a través de los coeficientes de regresión. El algoritmo que nos permitirá encontrar los valores de estos coeficientes será el mismo de la regresión lineal. ¿Por qué? La razón de ello es la siguiente: aunque con la regresión polinomial se pueden modelar relaciones no lineales, el problema asociado con la estimación de parámetros sigue siendo lineal.



Se pueden considerar polinomios de orden aún más alto:





En nuestro modelo de regresión polinómica, tenemos todos esos “poderes” de x que aparecen, los podemos tratar como características del modelo, w0,w1,w2,w3 son los parámetros del modelo, asociadas a cada una de las características.

