**Apuntes Primera regresión Lineal Platzy**

**Estandarizacion de datos**

<https://www.feedingthemachine.ai/estandarizar-informacion-machine-learning/>

ejemplo:

<https://github.com/feedingthemachine/exblog/blob/master/Standarizar/Estandarizar.ipynb>

Videos explicativos de escalado, normalizacion, estandarizacion

<https://www.youtube.com/watch?v=-VuR14Qyl7E> (canal youtube codigo maquina)

Escalado y Estandarización de Variables en python (Rocio Chavez)

<https://www.youtube.com/watch?v=YOyiAzJv8Qg>

Estandarizacion de variables-ejemplo numerico (Rocio Chavez)

[www.youtube.com/watch?v=Zujm13fhyI0](http://www.youtube.com/watch?v=Zujm13fhyI0)

L**ink a imagen De datos de RM Y MEDV ESTANDARIZADOS**

<https://ibb.co/17j0qmF>

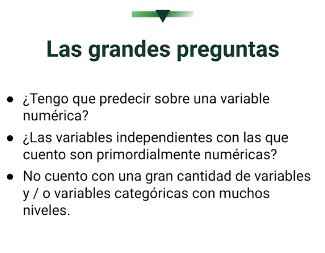
**Link a imagen De datos de RM Y MEDV NO ESTANDARIZ**ADOS

<https://ibb.co/56nStMH>

**Video de DOT Csv explicativo de Regresion Lineal con Boston Data Housing**

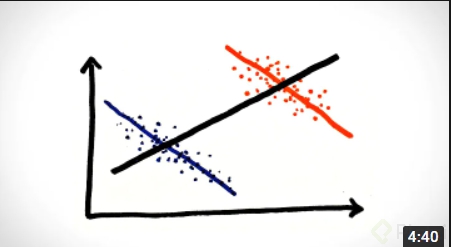
<https://www.youtube.com/watch?v=k964_uNn3l0&list=RDCMUCy5znSnfMsDwaLlROnZ7Qbg&start_radio=1&rv=k964_uNn3l0&t=132>

**Cuando Utilizar el Modelo de regresion Lineal**

****

* Las variables que puedo usar en este modelo deben ser Numericas
* Cuando tengo con gran cantidad de categorias ( por ejemplo perro, Gato, Pato…..) y a su vez estas categorias tienen subcategorias No es Recomendable usar este modelo

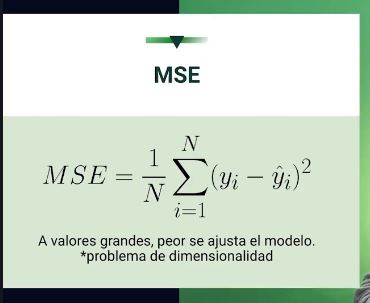
La razón por la que no es recomendable que usemos la regression linear cuando hay muchas categorías es la paradoja de simsom. les recomiendo este [video](https://platzi.com/new-home/clases/2739-regresion-lineal/46180-cuando-utilizar-un-modelo-de-regresion-lineal/url) pahttps://www.youtube.com/watch?v=ebEkn-BiW5k&tra que lo vean



**Metricas**

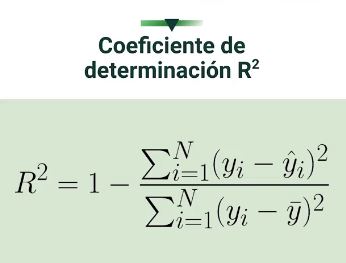
Las metricas las utilizamos para saber si un modelo de regresion se ajusta bien o mal a los datos reales

**Funcion de perdida Modelo de Regresión Lineal - Error Cuadrático medio MSE(Mean Square Error)**



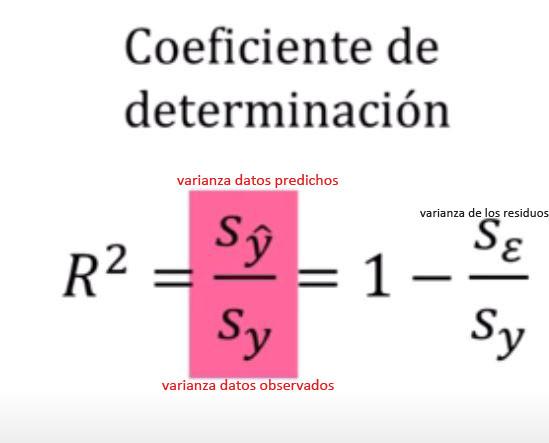
Comparamos los valores predichos en el modelo con los valores reales

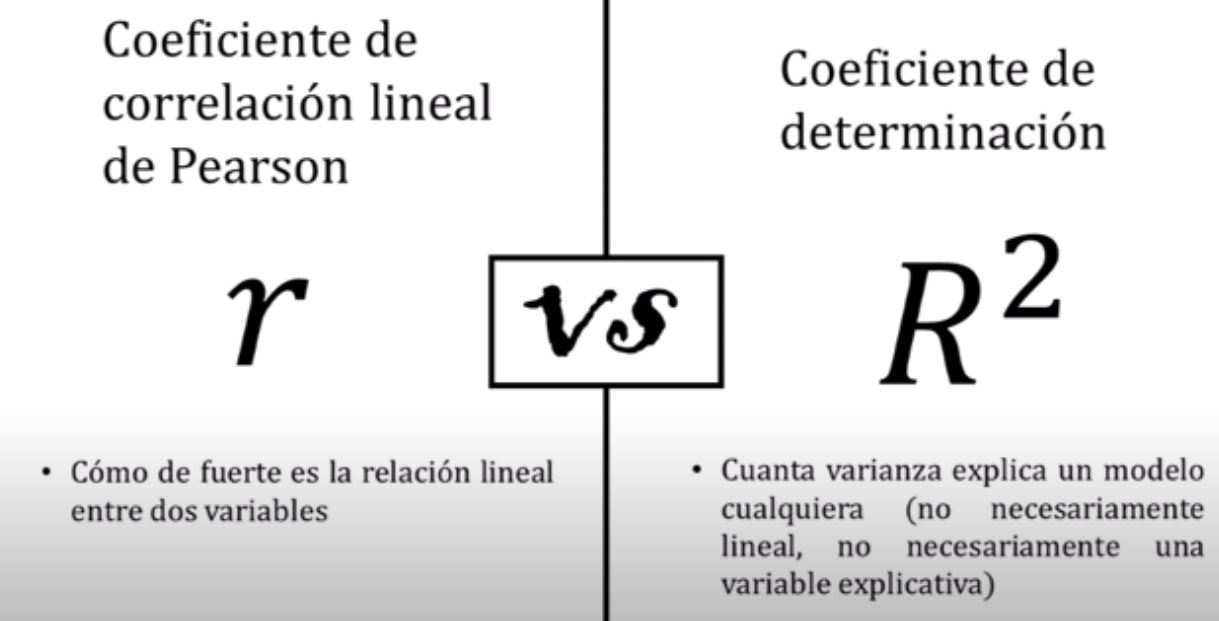
**Coeficiente de Determinacion R^2**

****

*El valor de* ***R2 varía entre 0 y 1.*** *Un valor de 0 indica que el modelo no puede explicar la variabilidad (varianza) de la variable dependiente, mientras que un valor de 1 indica que el modelo explica toda la variabilidad de la variable dependiente.*

Lo podemos tambien entender como relacion de varianzas

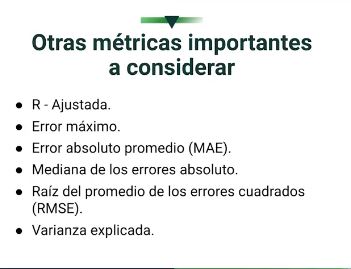




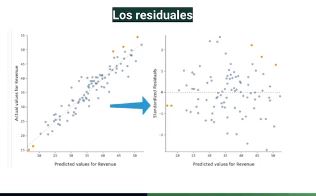
Aca un video que explicativo

<https://www.youtube.com/watch?v=MtDXpWC_RLc>

**Tambien existen otras métricas**

****

**Grafica de residuales - Analisis**

****

Los residuales son las diferencias entre los valores observados y los valores predichos por el modelo de regresión. La gráfica de residuales muestra estos residuales en el eje vertical (y) y **los valores predichos en el eje horizontal (x)**. Cada punto en la gráfica representa una observación del conjunto de datos.

El objetivo principal de la gráfica de residuales es analizar los patrones y estructuras que puedan indicar problemas en el modelo de regresión. Algunos patrones comunes que se pueden identificar en la gráfica de residuales son:

1. Patrón aleatorio: Los residuales se distribuyen aleatoriamente alrededor de cero, lo que indica que el modelo es apropiado y se ajusta bien a los datos.
2. Patrón en forma de U: Los residuales tienden a ser positivos para valores bajos y negativos para valores altos de los valores predichos. Esto puede indicar una no linealidad en la relación entre las variables.
3. Patrón en forma de embudo: La dispersión de los residuales aumenta o disminuye a medida que aumentan los valores predichos. Esto puede indicar una heterocedasticidad, es decir, una variabilidad no constante en los errores.
4. Patrón en forma de J: Los residuales tienen una forma de J o S, lo que puede indicar una violación de la suposición de linealidad en el modelo.

La gráfica de residuales también puede ayudar a identificar valores atípicos o influentes en el conjunto de datos. Los valores atípicos se representarán como puntos que se alejan significativamente de la línea central de residuales, mientras que los valores influentes pueden tener un impacto desproporcionado en los resultados del modelo.

Aca Video explicativo

<https://www.youtube.com/watch?v=rN9VPYjTIUU>

**¿Qué hace el one-hot encoding?**

El One-Hot Encoding es una técnica utilizada en el procesamiento de datos para convertir variables categóricas en una representación numérica adecuada para su uso en algoritmos de aprendizaje automático.

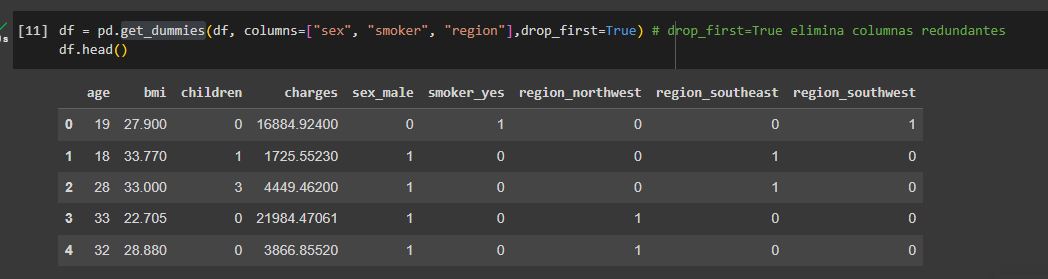
Cuando trabajamos con datos categóricos, como colores, categorías de productos o etiquetas, los algoritmos de aprendizaje automático generalmente requieren que los datos estén en formato numérico. Sin embargo, simplemente asignar números enteros a las categorías puede generar un orden o una relación implícita que no existe en realidad.

El One-Hot Encoding resuelve este problema al crear nuevas columnas (variables binarias) para cada categoría distinta en la variable original. Cada columna representará una categoría y contendrá un valor binario (1 o 0) que indica si la observación pertenece a esa categoría o no.

Por ejemplo, consideremos la variable categórica "Color" con tres posibles categorías: "Rojo", "Verde" y "Azul". Después de aplicar el One-Hot Encoding, se crearán tres nuevas columnas: "Color\_Rojo", "Color\_Verde" y "Color\_Azul". Si una observación tiene el valor "Rojo" en la variable original, la columna "Color\_Rojo" tendrá el valor 1 y las otras columnas tendrán el valor 0.

Esta representación permite que los algoritmos de aprendizaje automático utilicen la información categórica de manera adecuada, sin imponer ninguna relación numérica errónea. Además, el One-Hot Encoding evita la asignación de valores numéricos arbitrarios, lo que podría generar un sesgo en el modelo.

En nuestro Tp utilizamos la funcion de pandas **get\_dummies**

****