Reporte de desempeño modelo de regresión lineal con dataset Student

Autor: Fermín Méndez García A01703366
Repositorio:
https://github.com/FerminMendez/ModuleAI/tree/main/M2_ML_model_using_frameworks

En este caso implementamos dos versiones de una regresión lineal para predecir las calificaciones de los estudiantes a partir de algunas de las 33 variables reportadas en el dataset con 649 instancias.

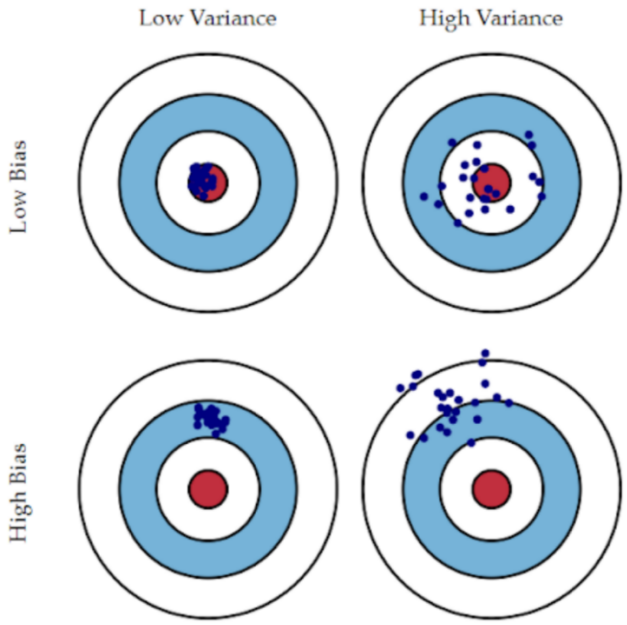
En este caso vamos a hablar de el desempeño de estas dos versiones de regresión lineal que implementamos.

1. Separación de datos en entrenamiento y test.

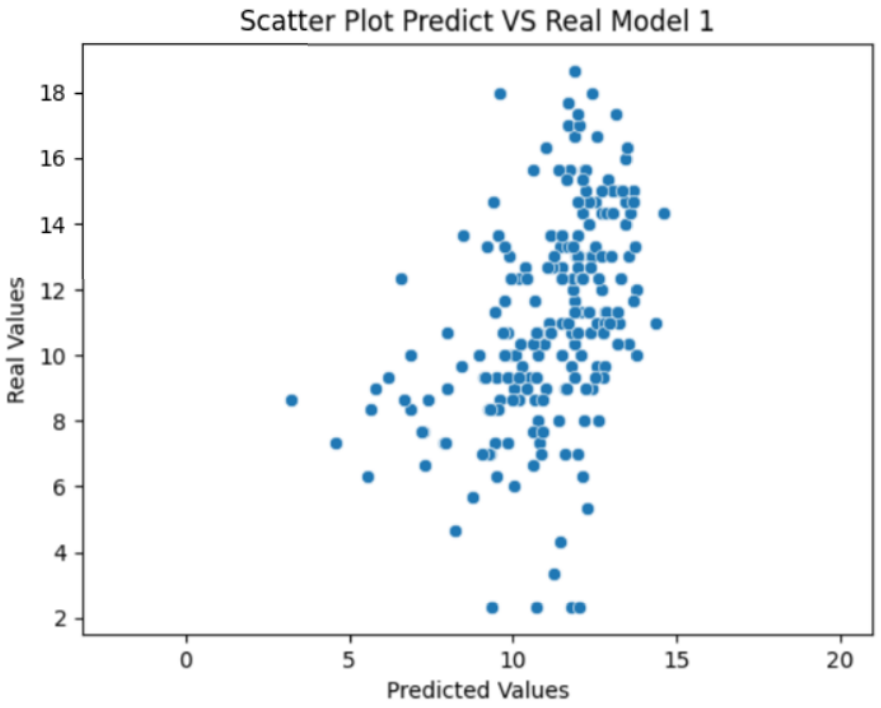
Con ayuda de la librería sklearn separamos el dataset en variables dependientes e independientes donde tomamos el 20% de los datos y los separamos para entrenamiento ya que tenemos menos de 1000 instancias esta suele ser un proporción adecuada,

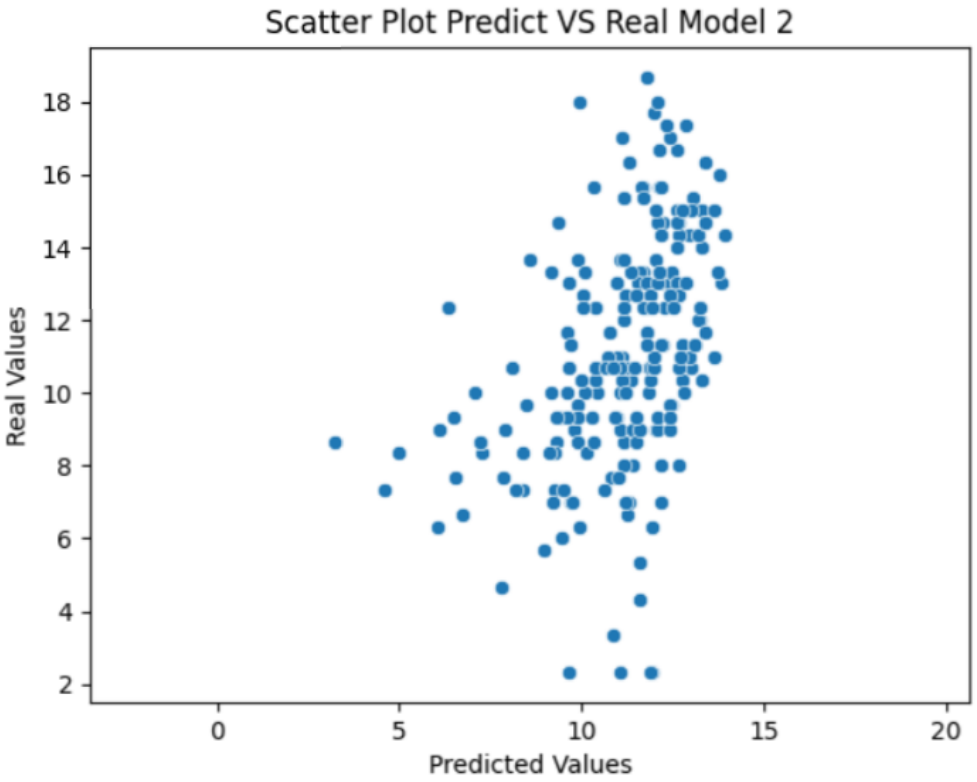
2. Grado de sesgo y varianza:

A nice metaphor of Bias and Variance from



Para explicar el sesgo y la varianza tomaremos la metáfora anterior y vamos a traer algunas gráficas de los resultados de nuestro modelo.





Nuestros modelos deberían aspirar a una línea recta de 45° ya que estamos comparando con los resultados esperados.
Un modelo de predicción con mucho sesgo mostraría una línea con mucha densidad
Un modelo de predicción con mucha varianza mostrará la línea con mucha dispersión.
En este caso y tomando en cuenta los resultados de nuestro modelo podemos concluir que este modelo tiene alto sesgo y alta varianza.

1. Regresión lineal 1

* TRAIN RESULTS:
MSE: 7.1524760537283365
R^2: 0.30791877023035896

* TEST RESULTS:
MSE: 8.663829527330574
R^2: 0.1637783640586179

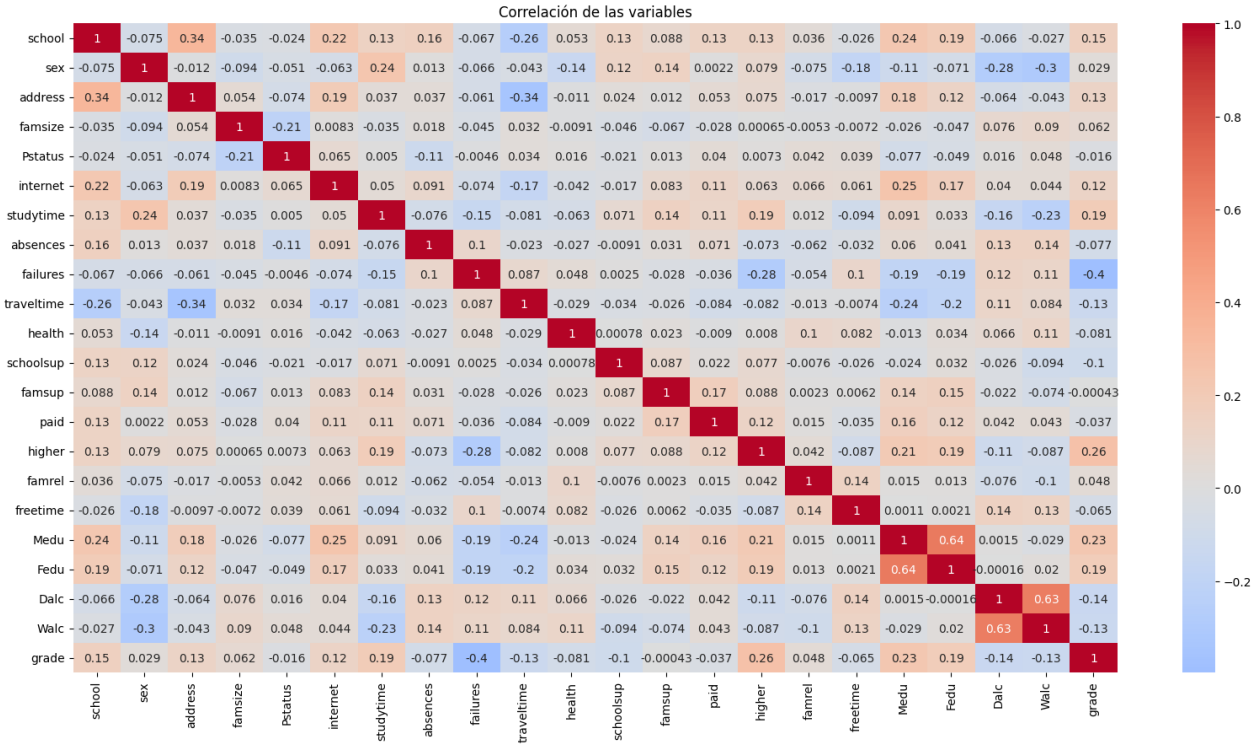
2. Modelo de regresión lineal 2

* TRAIN RESULTS:
MSE: 7.309731260988808
R^2: 0.2927025882521228

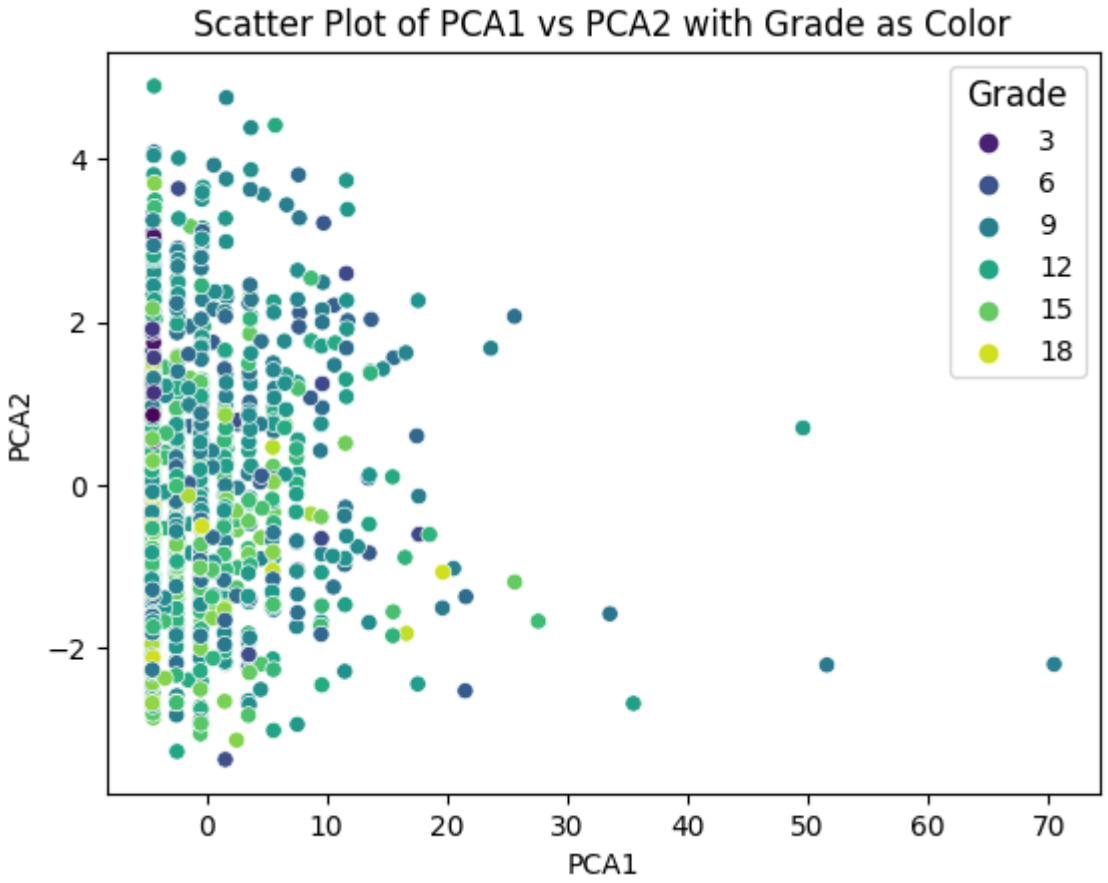
* TEST RESULTS:
MSE: 8.630070998710131
R^2: 0.1670366936391927

Por otro lado podemos ver que ni siquiera comparado con los datos de prueba el coeficiente R^2 se acerca al 1, no pasa de 0.3. Es un claro indicador que está underfit. Nuestro modelo es demasiado simple para describir el comportamiento que queremos predecir.

Algunos ajustes que podemos hacer es mejorar la calidad de los datos que le damos de acuerdo al índice de correlación o añadir PCA para ver si podemos encontrar patrones significativos. Eso fue lo que se intentó con el segundo modelo y mejoró de 0.163 a 0.167 casi nada. Veamos las gráficas para entender mejor este resultado.



En este caso no hay fuertes correlaciones con la variable independiente. Los datos no son muy buenos para predecir la calificación de un estudiante.



Por otra parte, los componentes principales explican apenas el 38% y 2% de la varianza. Por eso no podemos encontrar ningún patrón con la reducción de dimensionalidad.

En el repositorio podemos encontrar algunos otros intentos con redes neuronales y árboles de decisión y parece indicar que el problema principal es que los datos no son suficientes en cantidad, ni en significancia para hacer un mejor modelo.

Buscando otros análisis de este mismo dataset no encontré ninguna solución. Es un dataset pequeño que o no ha sido explorado en estos 9 años que se publicó o no hay algún modelo que pueda predecir calificaciones con alta precisión en este conjunto de datos.