Inteligencia artificial avanzada para la ciencia de datos I

Momento de Retroalimentación: Módulo 2 Análisis y Reporte sobre el desempeño del modelo. (Portafolio Análisis)

Erick Hernández Silva A01750170

Sobre el código

Modelo realizado

Para esta entrega se realizó un modelo de regresión lineal múltiple que puede recibir una cantidad n de características de entrada y una de salida. Se utiliza un formato csv para cargar los datos. El usuario puede elegir cuál es su característica de salida, y seleccionar cuáles características quiere usar como entrada (o usar todas las restantes). También se le permite seleccionar el porcentaje de datos que quiere usar para pruebas. El resto de datos se utiliza para entrenamiento.

Hiperparámetros ajustables

El único hiperparámetro ajustable es la cantidad de datos que quieres usar para pruebas y entrenamiento.

Pruebas

Datasets de pruebas

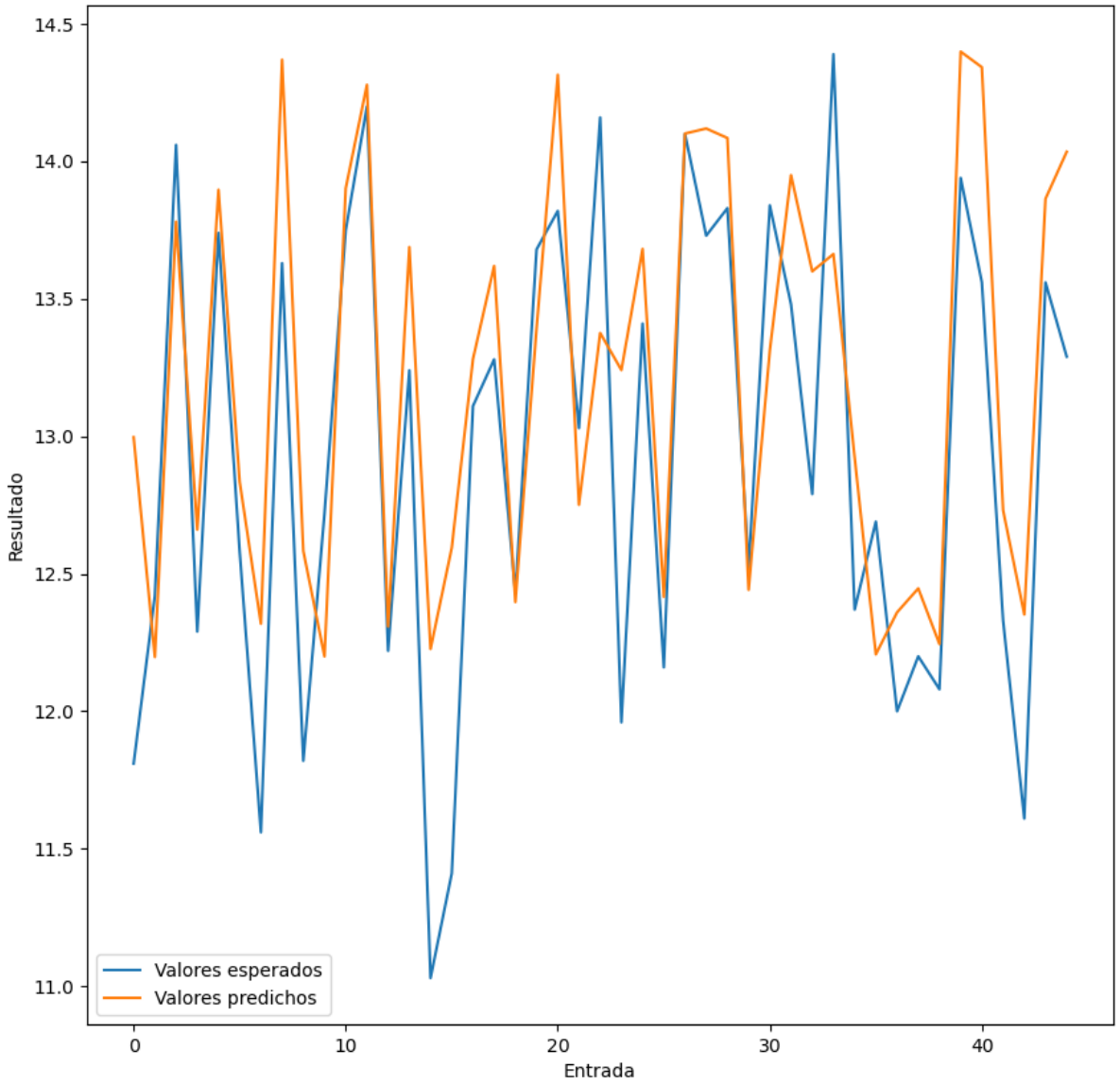
Para las pruebas se incluyó un dataset de prueba llamado wine.csv que son datos limpios.

Prueba con distintas columnas de entrada

El hecho de que nosotros podamos elegir las características de entrada nos permite obtener diferentes modelos con distinto nivel de accuracy o menor mse. Para estas pruebas vamos a tratar de predecir la columna *Alcohol*.

Prueba con todas las entradas

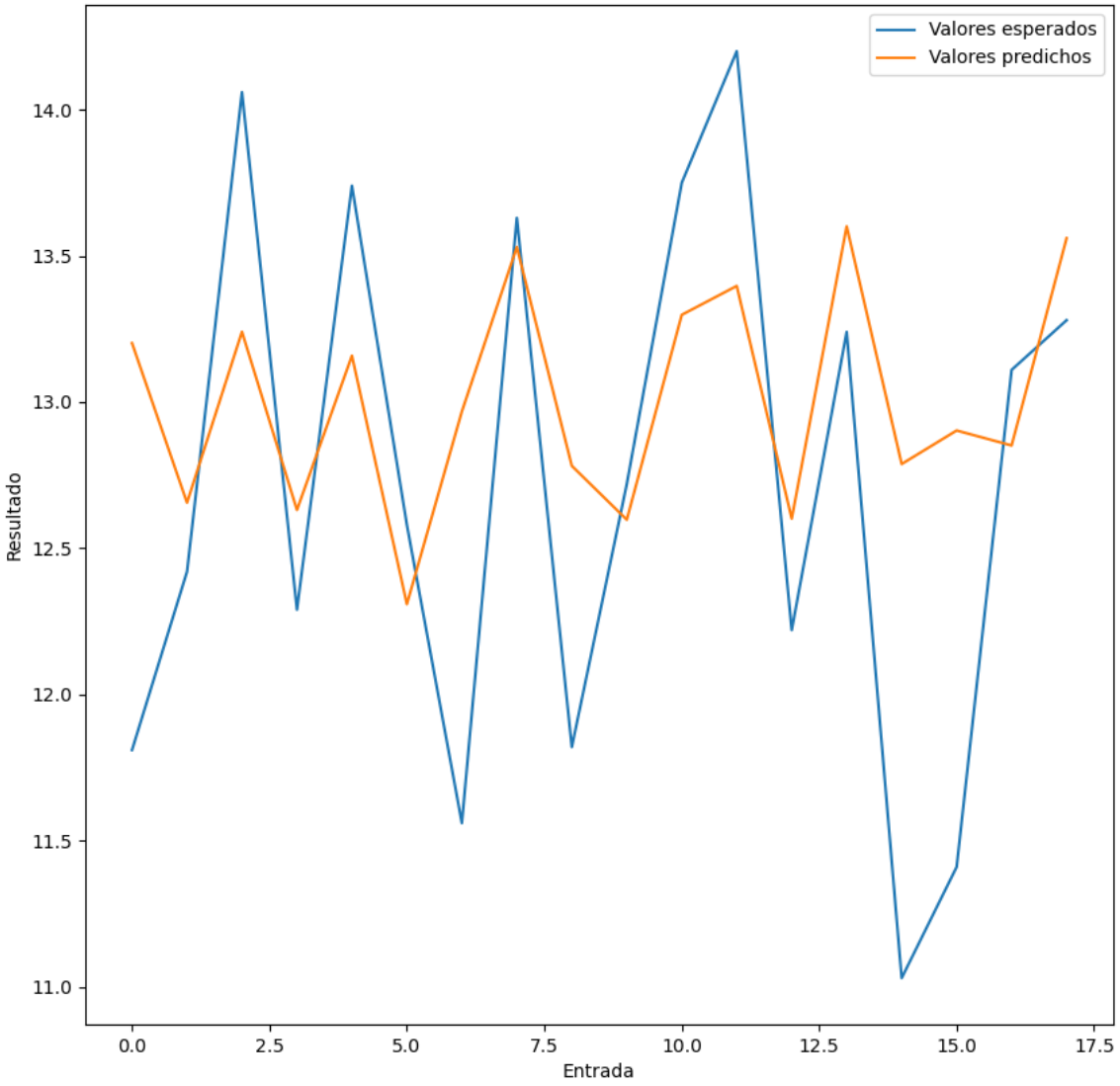
Se utilizó como característica de salida la columna *Alcohol* y se utilizaron las demás características de entrada.



Lo cual nos dió un modelo con una R cuadrada ajustada de 0.5739570080342566 y un MSE de 0.302, un Bias de 0.261 y una varianza de 0.042.
Como podemos ver en la gráfica (y en los datos de bias y varianza), tenemos un bias un poco elevado y una varianza muy baja. La varianza baja nos indica que estamos lejos de hacer overfitting ya que no estamos tratando de tocar todos los puntos. Sin embargo, nuestro bias es un poco elevado, lo cual nos podría indicar que nuestro modelo está tratando de generalizar demasiado, con lo que podríamos caer en un underfitting.

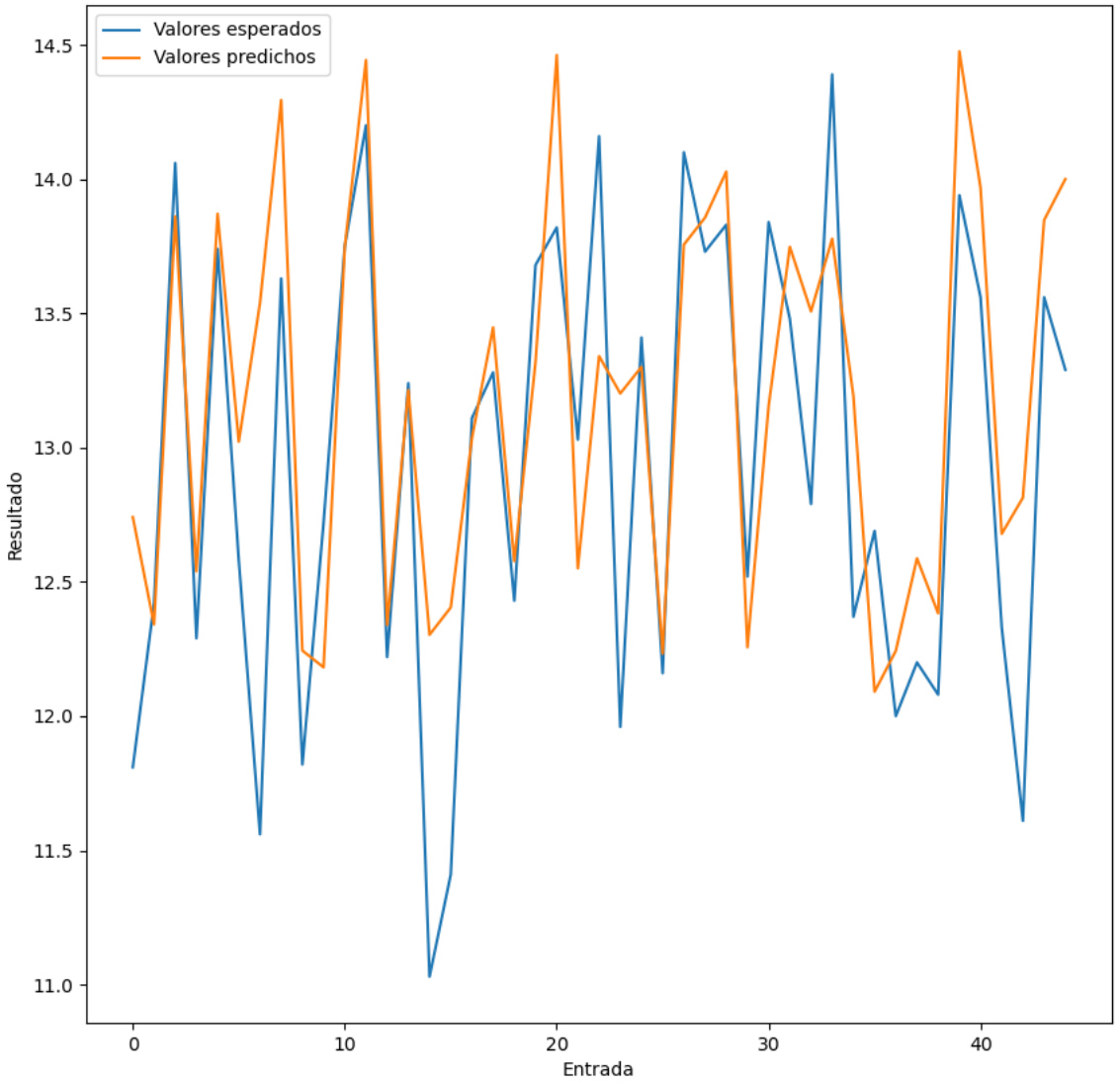
Mejora del modelo

Prueba con un dataset de pruebas del 10% y 4 entradas



Lo cual nos dió un modelo con una R cuadrada ajustada de 0.19288992756107504 y un MSE de 0.711, un Bias de 0.683 y una varianza de 0.029.
Como podemos ver en la gráfica (y en los datos de bias y varianza), tenemos un bias muy elevado y una varianza muy baja. La varianza baja nos indica que estamos lejos de hacer overfitting ya que no estamos tratando de tocar todos los puntos. Sin embargo, nuestro bias es muy elevado, y este modelo está tratando de generalizar demasiado.

Pruebas con un dataset de pruebas del 25 y 10 columnas



Lo cual nos dió un modelo con una R cuadrada ajustada de 0.49305253398088134 y un MSE de 0.306, un Bias de 0.278 y una varianza de 0.028.
Como podemos ver en la gráfica (y en los datos de bias y varianza), tenemos un bias mucho más bajo y una varianza muy baja. La varianza baja nos indica que estamos lejos de hacer overfitting ya que no estamos tratando de tocar todos los puntos. Y nuestro bias más bajo nos indica que estamos generalizando muchísimo menos. Con esto logramos tener un balance entre over y under fitting, dándonos un posible buen modelo para predecir los grados de *Alcohol*.

Análisis de accuracy y error

Información otorgada por el programa al finalizar

Al finalizar una corrida, siempre se nos otorgan datos como la R cuadrada, el MSE, una estimación del Bias y la varianza general para poder realizar un análisis sobre qué tan bueno es el modelo y poder hacer ajustes como lo hicimos anteriormente.

Conclusiones importantes sobre el modelo

Al modelo hay que introducir datos que ya hayan sido previamente tratados, de lo contrario podría comenzar a realizar predicciones inexactas o simplemente no funcionar.

Además, nos indica datos importantes como el bias y la varianza para asegurarnos que no caigamos en underfitting o en overfitting.