# Tecnológico de Monterrey

**INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE MONTERREY, CAMPUS ESTADO DE MÉXICO.**

**Escuela de Ingenierías**

**Tarea 3 módulo 2**

***Inteligencia artificial avanzada para la ciencia de datos I (Gpo 101)***

Andrea Vianey Díaz Álvarez A01750147

**Fecha de entrega:** 15 de septiembre del 2022

# Análisis y Reporte sobre el desempeño del modelo

## Modelo de Regresión Lineal Múltiple.

El modelo de Regresión Lineal Múltiple es una técnica estadística que se usa para analizar por qué pasan las cosas o cuáles son las principales explicaciones de que algo ocurra, trata de ajustar modelos lineales entre una variable dependiente y una o más variables independientes. Suponemos que más de una variable está correlacionada con el valor de la variable dependiente y se busca predecir los valores de la variable dependiente.

$$y_j = b_0 + b_1 X_{1j} + b_2 X_{1j} + \ldots + b_k X_{kj} + \varepsilon_j$$

Donde y es la variable dependiente y X las variables independientes, Ɛ el error y b los coeficientes de regresión.

## Descripción del Dataset y su separación en conjunto de entrenamiento y prueba

En esta Actividad se estarán utilizando variables continuas. Antes de obtener los resultados se prepararon los datos:

1. Revisar que no haya datos incompletos.
2. Estandarización de datos. $z = \frac{x - \mu}{\sigma}$

En este caso no encontramos datos incompletos, pero si se realizó la estandarización de los datos para facilitar el aprendizaje.

El dataset que escogí para esta actividad se trata de la tasa de desempleo, donde por cada país se hizo una división del número de desempleados y el número total de personas que tienen la capacidad de trabajar.

Dataset: https://www.kaggle.com/datasets/pantanjali/unemployment-dataset

Donde como X se tomaron los años: 1991, 1992 y 1993 y como Y se tomó el año1994, para poder predecir el número de personas desempleadas por países en el año 1994.

Para poder probar el modelo de hizo la división de datos:

1. 80% datos de entrenamiento.
2. 20% datos de prueba.

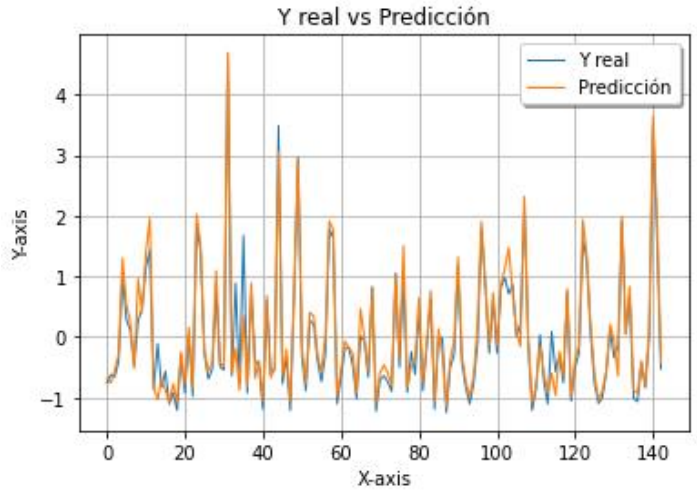
Estos porcentajes se hicieron para los dos entrenamientos y pruebas que se hicieron, pero con distintos datos.

## Diagnóstico y explicación el grado de bias o sesgo: bajo – medio - alto

Comprender como se genera el error dependiendo de diferentes fuentes ayuda a poder mejorar el proceso de ajuste de datos para que el modelo sea más preciso.

Error de bias: El error de bias es la diferencia entre la predicción esperada y los valores reales.

En la primera prueba se obtuvo un bias promedio de 0.13383211317685487 y en la segunda prueba se obtuvo 0.06617652826905399. Por lo que se podría concluir que el bias es bajo en este modelo.




## Diagnóstico y explicación el grado de varianza: bajo medio alto

Varianza: Se utiliza para determinar si las diferencias que existen entre las medias de muestreo exponen las diferencias que hay entre los valores medio.

En la primera prueba se obtuvo una varianza de 0.8253151396948403 y en la segunda se obtuvo 1.085266776126077. Por lo que se podría concluir que el modelo en general da una varianza baja que indica que los valores están próximos a la media.

## Diagnóstico y explicación el nivel de ajuste del modelo: underfit fit overfit

Después de analizar los resultados del bias y la varianza se puede concluir que el modelo no tiene ningún problema de underfit u overfit, ya que se hicieron los ajustes necesarios en las iteraciones y learning rate para que esto no ocurriera.

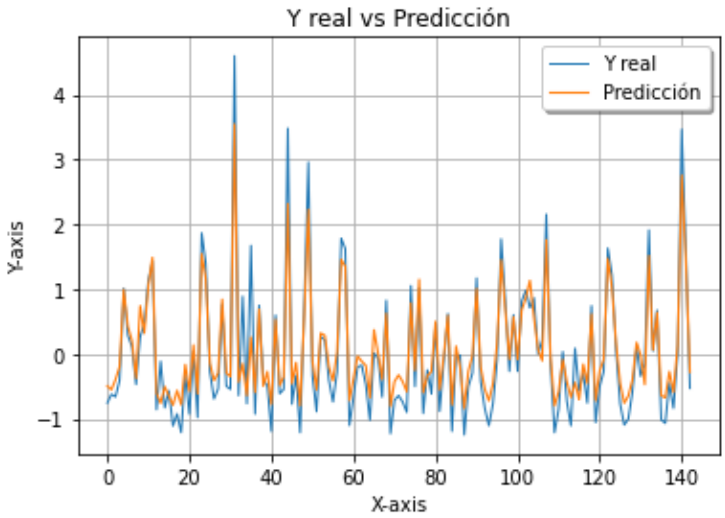
## Utilización de técnicas de regularización o ajuste de parámetros para mejorar el desempeño del modelo

Para la optimización del modelo se hicieron varias pruebas ajustando los valores del Alpha(learning rate) y las iteraciones.

En la gráfica de abajo se muestra el resultado obtenido después de hacer el cálculo del modelo con un Alpha = 1, y 100 iteraciones.

Se obtuvo como resultado:

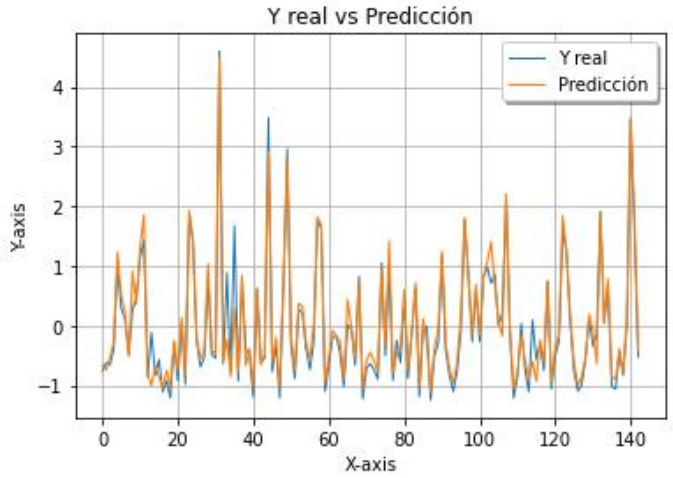
- R2: 0.890601901204356
- Bias: 0.07928864681311804
- Varianza: 0.604074589587426




En la gráfica de abajo se muestra el resultado obtenido después de hacer el cálculo del modelo con un Alpha = 0.001, y 5 iteraciones.

Se obtuvo como resultado:

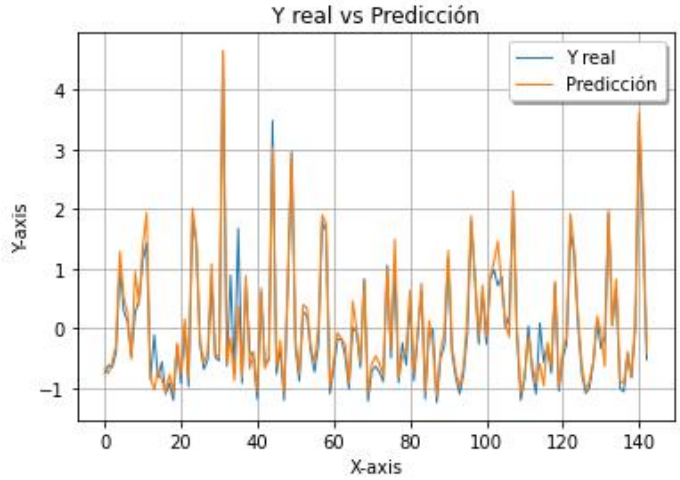
- $R^2$ = 0.9393987022062087
- Bias: 0.05408984113331557
- Varianza: 0.9945961415938158




En la gráfica de abajo se muestra el resultado obtenido después de hacer el cálculo del modelo con un Alpha = 0.01, y 10000 iteraciones.

Se obtuvo como resultado:

- R2: 0.9400528283355423
- Bias: 0.06257860324928416
- Varianza: 1.076048059524201




Después de hacer varias pruebas me di cuenta de que es necesario tener un valor bajo de Alpha para que fuera más preciso y un número “alto” de iteraciones para que pudiera terminar de aprender el modelo.