Tecnológico de Monterrey

**Campus Querétaro**

**Asignatura**

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

**Tema**

**Módulo 2 Análisis y Reporte sobre el desempeño del modelo**

**Alumno**

Carolina Arratia Camacho
A01367552

**Fecha**

10 de septiembre del 2023

Link al código:
https://colab.research.google.com/drive/1IkP_8bNV2hQ438d7I3qqGi8rrOXg_MuJ?usp=sharing

El modelo que se utilizó para el análisis que se presenta en este apartado es el modelo realizado sin el uso de un framework.

Este busca que, a través de la implementación de un modelo de regresión lineal, y con ayuda de la función de gradiente descendiente, se pueda predecir el índice de felicidad de un país dado, tomando como indicadores las variables de Economy (GDP per Capita), Family, Health (Life Expectancy), Freedom, Trust (Government Corruption), Generosity.

El dataset con el que se está trabajando cuenta con los datos del índice de felicidad de 158 países, es decir, cuenta con 158 registros. De estos el 90% están siendo usados para entrenar el modelo, y el otro 10% se usa para prueba.

Posteriormente se realizó cross-validation dividiendo el train en 4.

Dado que se trata de un modelo de regresión lineal, las métricas que se tomaron en cuenta para evaluarlo fueron la $R^2$ y el MSE.

La $R^2$ es útil ya que permite evaluar el rendimiento de mi modelo. Me dice qué tan bien mi modelo se ajusta a los datos, y la calidad de este para replicar los resultados. Po lo que, el obtener un valor de 1 sería lo deseado.

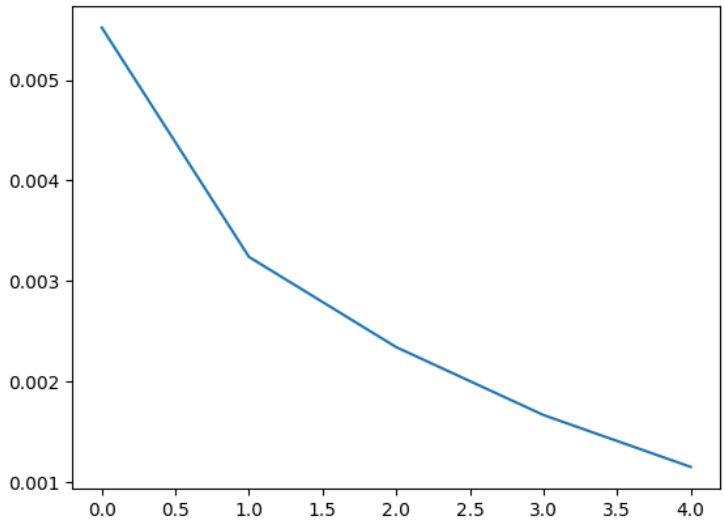
El Error Cuadrático Medio (MSE) también ayuda a medir la calidad del modelo ya que mide el promedio de los errores al cuadrado, es decir, la diferencia entre mi hipótesis de <y>, y su verdadero valor.

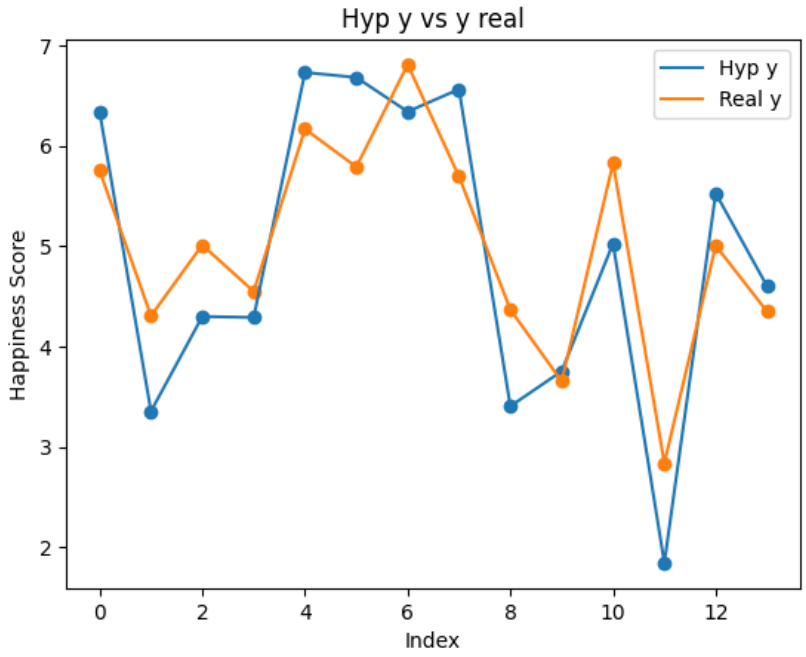
**Grado de bias o sesgo**

Medio

Al correr el modelo de regresión lineal se pueden observar las siguientes gráficas.

Gráfica de error






Donde se puede ver que la mayoría de los datos predichos, a pesar de que se acercan a los reales, y se cuenta con un error bajo, es posible observar que estos no son iguales.

Al calcular el error cuadrático medio de los datos de prueba se obtienen un 0.486, este es el promedio, al cuadrado de cuánto varían los datos reales de los hipotéticos.

**Grado de varianza**

Medio

Para evaluar qué tan bien funciona mi modelo, y de cierta manera determinar su accuracy, se hace uso de la $R^2$.

Para determinar la varianza, lo que se realizó fue comparar qué tanto cambia el valor de $R^2$ con respecto al train, el test y el cross-validation.

El valor de $R^2$ que se obtuvo para cada uno de estos fue:

Train: 0.4424

Test: 0.5458

Cross validation: 0.55

**Nivel de ajuste del modelo**

Underfitting

Considerando el accuracy ($R^2$) planteado anteriormente, tanto de train como de test, se puede determinar que estas son muy bajas. Además, que al ver si MSE se observan valores altos, especialmente el que describe al train ya que este es de 0.73, mientras que el del test es de 0.49.

Además de que, al presentarse mayor error en el training se llegó a esta conclusión.

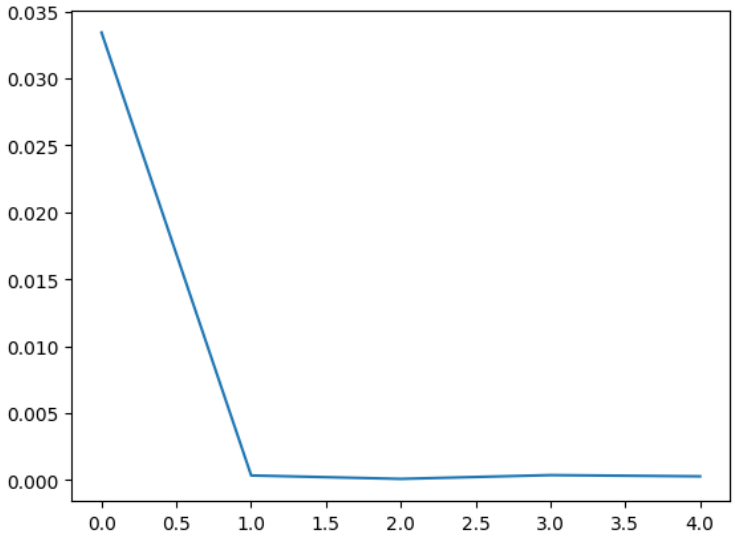
**Mejoramiento del modelo**

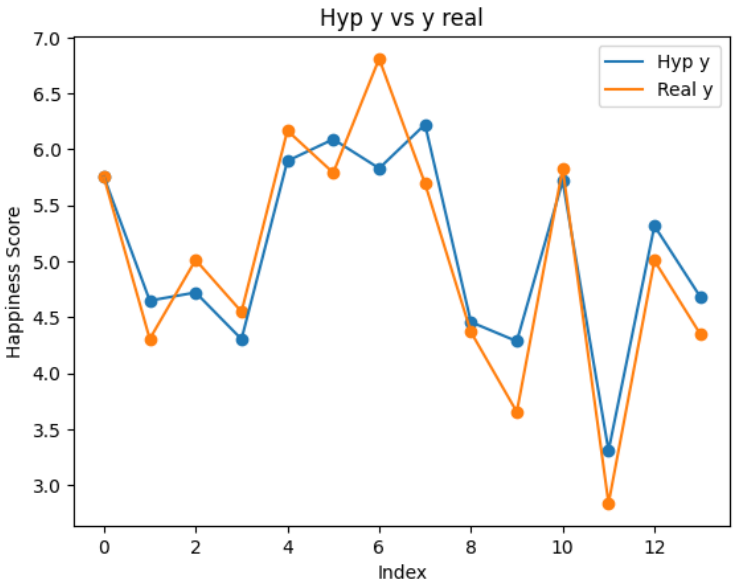
Con las pruebas anteriores se determinó que el modelo no era capaz de representar la relación entre las variables independientes y la dependente, y por ende no es muy preciso. Teniendo un rendimiento poco eficiente tanto en los datos de training como en los de test.

Con el propósito de mejorar el modelo se decidió hacer uso del Principal Component Analysis. Dado que es muy complicado el encontrar la relación entre las 6 variables independientes y la variable dependiente, al hacer uso del PCA se puede resolver este problema. El PCA permite encontrar las variables más significativas y así trabajar con menos variables.

Las gráficas obtenidas con esta mejora son las siguientes

Gráfica de error



Al analizar la $R^2$ de este nuevo modelo se obtiene que el train sube a un 0.76, mientras el test se desempeña en un 0.83, por lo que su accuracy mejoró considerablemente.