

Diplomado en Analítica y Ciencia de Datos

Caso Práctico Módulo 7

Python

Tasa de Mortalidad de Adultos 2019-2021

Luis Alberto Pineda García

1.- Descripción del problema

- **Justificación**

Revisando opciones entre distintos sitios de internet y diversos temas, se optó por elegir un dataset que contiene información sobre la tasa de mortalidad de adultos en un periodo de 2019 a 2021, de diferentes países con un total de 156 registros o países y 10 columnas o factores que de alguna manera impactan el resultado de la tasa.

Fuente de Datos:

<https://www.kaggle.com/datasets/mikhail1681/adult-mortality-rate-2019-2021>

Archivo CSV: _Adult mortality rate (2019-2021).csv

- **Objetivo**

Analizar la tasa de mortalidad de adultos y la relación con distintos factores como el PIB promedio del país y la Inversión en Salud. Realizar regresiones con un modelo XG-Boost para entrenar los datos con un 70% y predecir la Tasa Global de Mortalidad de Adultos.

Herramientas: Google Colab (Python) Librerías: seaborn, pandas, matplotlib, numpy, Scikit-learn

- **Hipótesis**

Se cree que debe haber una relación muy marcada entre la tasa de mortalidad *promedio*(*Average_CDR*) y el *PIB_promedio* del país en Millones de Dolares(*Average_GDP(M\$)*) ?

También que debe haber una relación muy evidente entre la inversión en salud con respecto de la *Tasa de Mortalidad* (*Average_CDR*) vs (*Average_HEXP(\$)*)

Variable Objetivo: *Average_CDR*

2.- Análisis Exploratorio de Datos (EDA):

Contexto del Conjunto de Datos:

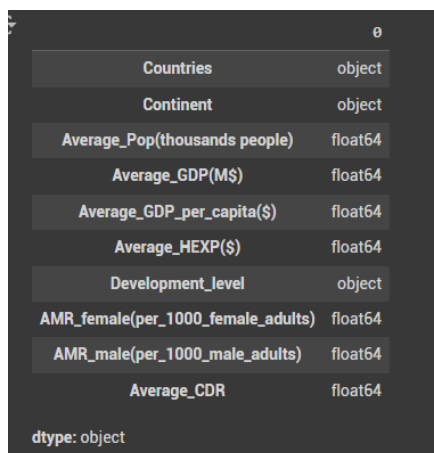
El Dataset elegido tiene información sobre la tasa de mortalidad de Adultos en distintos Países y algunas columnas o factores que afectan a ésta misma.

Las columnas del archivo analizado contiene son las siguientes:

- Países: País de estudio.
- Continente: Continente del Pais
- Average_Pop:: Población promedio en miles.
- PIB_promedio: PIB promedio del país en millones de dólares.
- PIB_per_cápita_promedio: PIB per cápita promedio en dólares.
- Average_HEXP: Gasto en Salud Per Cápita en dólares.
- Nivel de desarrollo: Nivel de desarrollo del estado en estudio
- AMR_female: Mortalidad promedio de mujeres adultas por cada 1000 por año
- AMR_male: Mortalidad promedio de hombres adultos por cada 1000 por año
- Average_CDR: Tasa bruta de mortalidad promedio

• Descripción de la Base de Datos

La Base de datos elegida (archivo CSV) cuenta con 156 registros y 10 columnas descriptivas de los equipos.

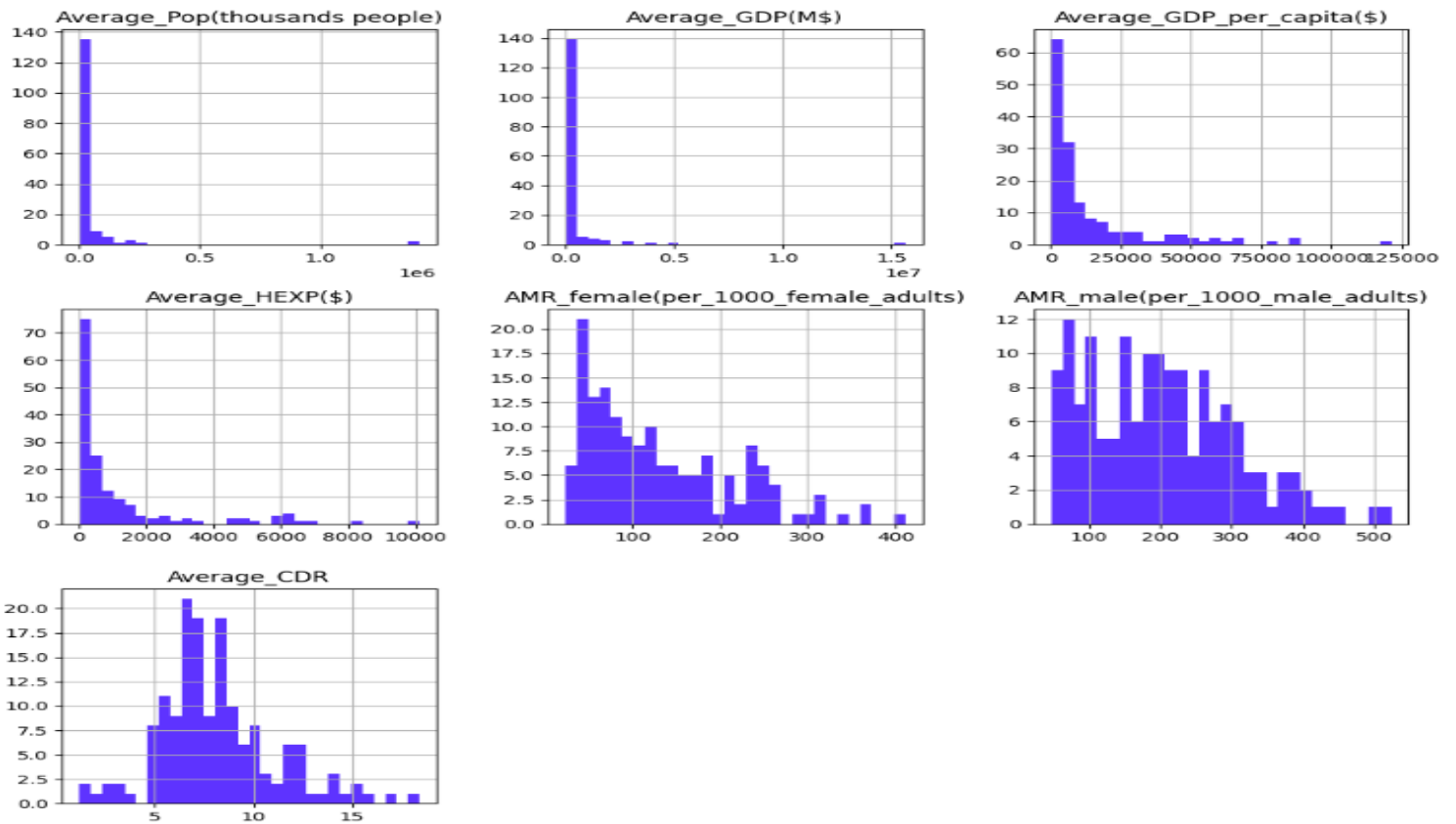


```
0
Countries      object
Continent      object
Average_Pop(thousands people)  float64
Average_GDP(M$)      float64
Average_GDP_per_capita($)      float64
Average_HEXP($)      float64
Development_level      object
AMR_female(per_1000_female_adults)  float64
AMR_male(per_1000_male_adults)      float64
Average_CDR      float64
dtype: object
```

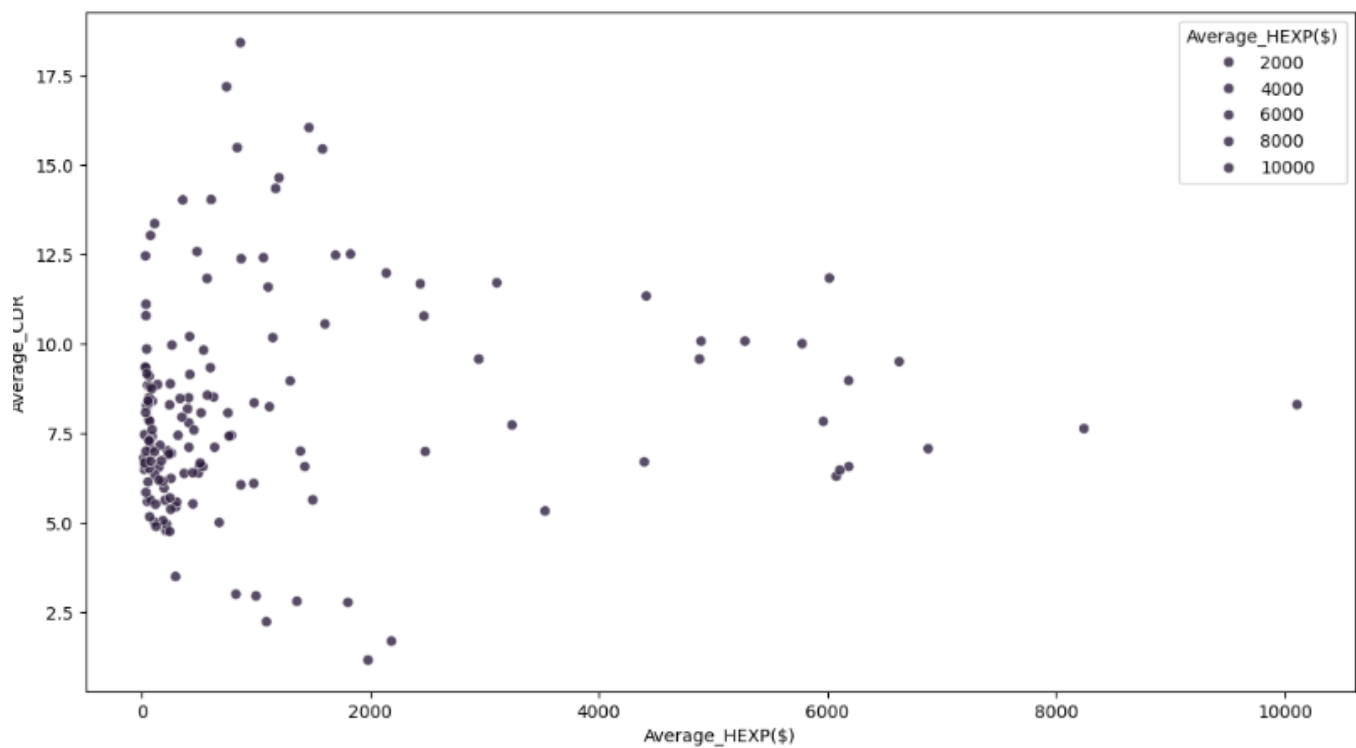
• Limpieza de los datos

No fue necesaria, el dataset esta totalmente limpio.

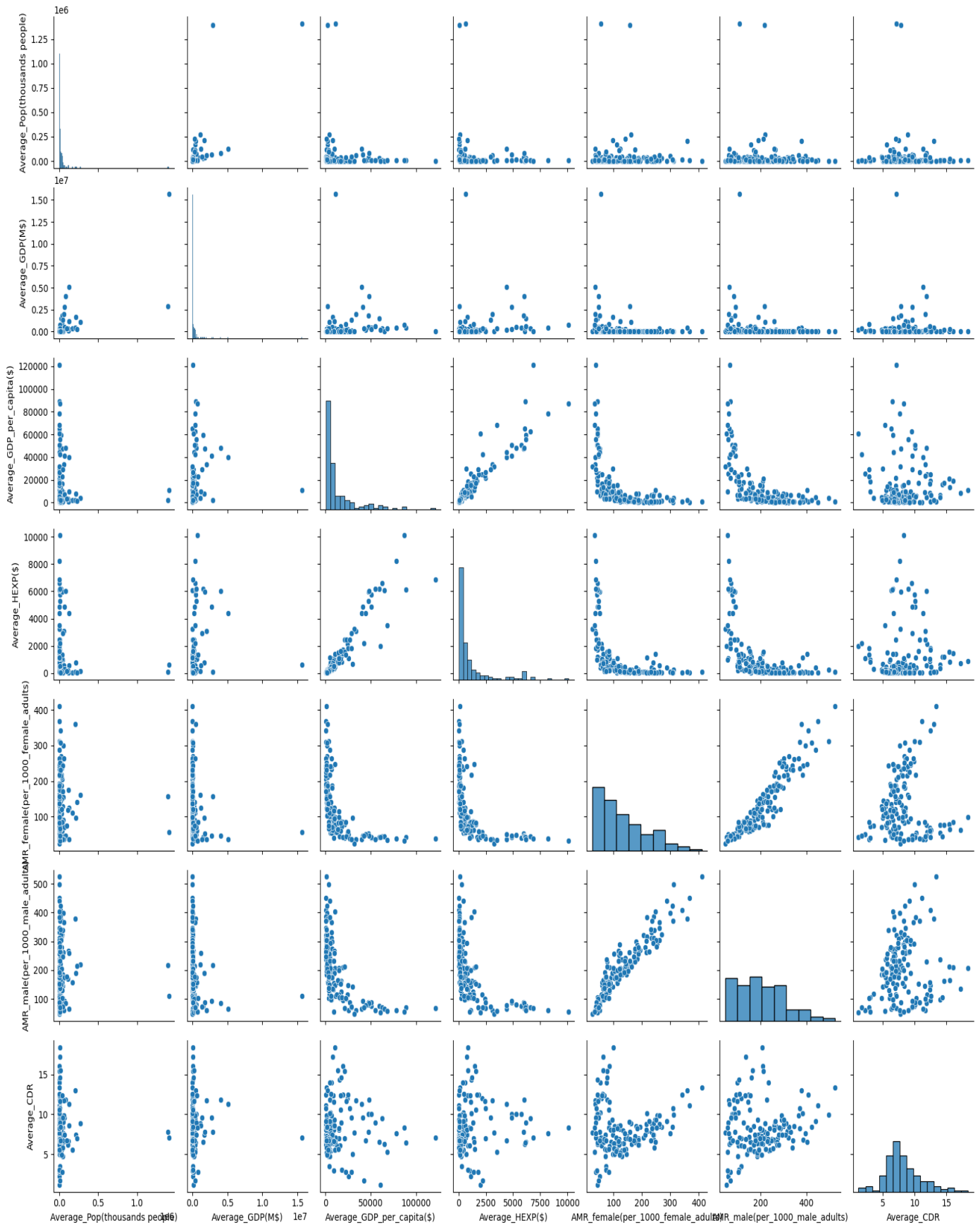
Histogr ma de todas las columnas num ricas



Gr fico scatterplot entre la variabe Objetivo: Tasa de Mortalidad y la Variable factoror Inversi n en Salud



Gráficos de dispersión entre todas las columnas y distribuciones



Estadística de los datos

```
#Consultando Estadística de los Datos
df.describe()
```

	Average_Pop(thousands people)	Average_GDP(¥\$)	Average_GDP_per_capita(\$)	Average_HEXP(\$)	AMR_female(per_1000_female_adults)	AMR_male(per_1000_male_adults)	Average_CDR
count	1.560000e+02	1.560000e+02	156.000000	156.000000	156.000000	156.000000	156.000000
mean	4.145031e+04	3.636484e+05	14009.898974	1165.549551	130.893590	202.383013	8.143654
std	1.616104e+05	1.402295e+06	20315.146615	1875.714534	84.503337	103.804217	2.953754
min	1.108000e+01	5.531000e+01	257.740000	18.920000	24.060000	46.790000	1.170000
25%	2.092582e+03	1.122715e+04	1987.862500	93.247500	62.440000	113.475000	6.397500
50%	8.778430e+03	3.613559e+04	5819.610000	408.170000	109.050000	195.000000	7.615000
75%	2.692418e+04	2.223658e+05	16348.190000	1154.520000	184.642500	269.897500	9.387500
max	1.410402e+06	1.565411e+07	121304.680000	10107.990000	411.090000	524.480000	18.400000

Hallazgos del Análisis exploratorio

Se observa visiblemente que las tasas registradas de la mayoría de los países, se concentran a la izquierda, es decir: La Mayoría de los países cuentan con un PIB promedio bajo y Gastos en Salud bajos con respecto de otros. Pero eso no quiere decir que la Tasa de Mortalidad disminuya en función de esa variable. Ya que se observa claramente que a pesar de ser un presupuesto bajo.

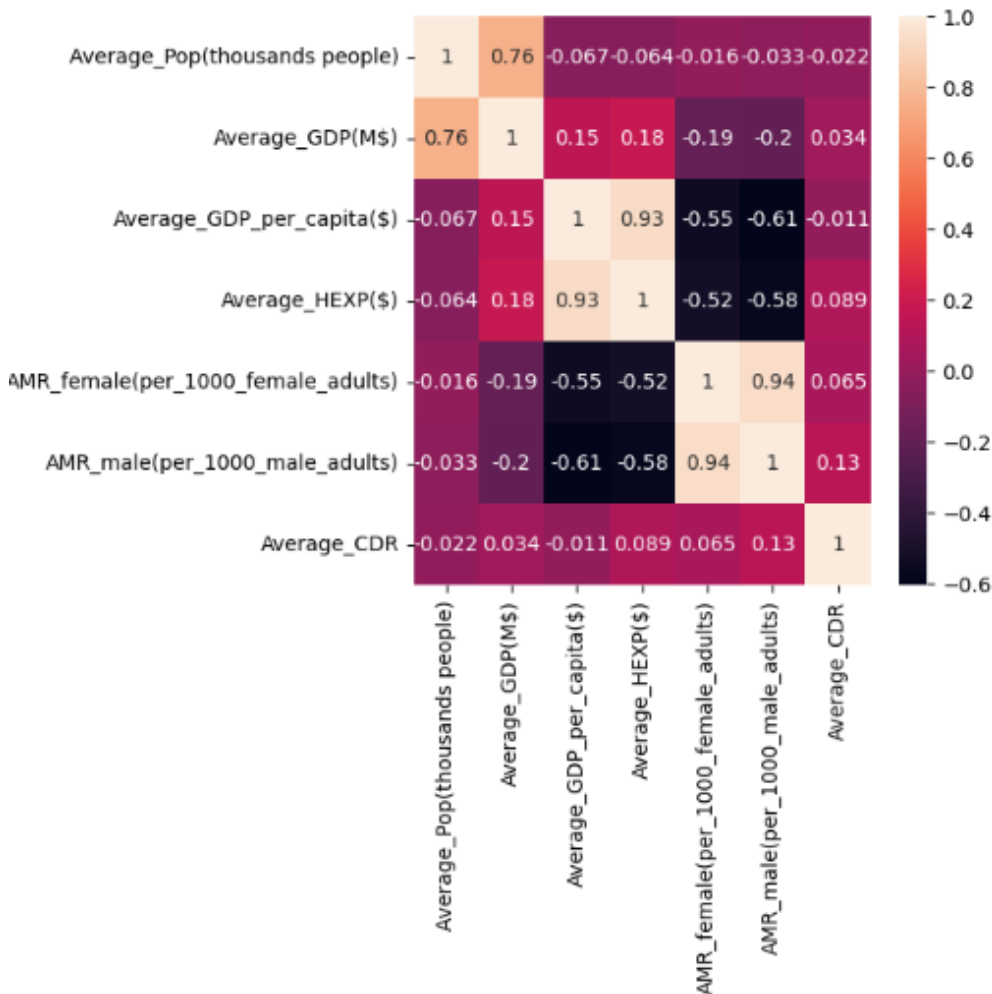
Hay Países con Tasas Altas, Medias y Bajas distribuidos uniformemente.

Aun no es posible confirmar la hipótesis, por lo tanto es necesario hacer un análisis de correlación para determinar el impacto en la variable objetivo.

3.- Preparación de los Datos:

- Limpiar los datos: No hubo necesidad de limpiar datos, el dataset no tiene nulos.
- Normalizar o Escalar los datos: No hubo necesidad de Escalar los Datos, todas las variables numéricas se pueden correlacionarse
- Convertir variables categóricas en variables numéricas
 - Se eliminaron variables de Tipo Object

Una vez eliminadas las variables object se procede a realizar la matriz de correlación



Hallazgos encontrados en matriz de correlación

Variable Objetivo: Average_CDR

Inversión en Salud Average_HEXP(\$) no tiene un impacto significativo en la variable objetivo con una medición de 0.089

*PIB Promedio Average_GDP(M\$) tiene una relación débil en la variable objetivo con una medición de 0.034

Por lo tanto: No se confirma la Hipótesis! Porque realmente el PIB de cada país y su inversión en Salud no tiene impacto considerable sobre la variable objetivo: Tasa bruta de mortalidad promedio Average_CDR

4.- Modelado Predictivo:

Entrenando modelo con librería **Scikit-learn**

Dividiendo los datos en un 30% de datos de prueba y un 70% de entrenamiento.

- Algoritmo utilizado: XG-Boost Regresión Lineal
- Métrica: Error al cuadrático (reg:squarederror)
- Tasa de aprendizaje: 0.1
- Estimadores: 400 árboles de decisión o estimadores

```
[30] from sklearn.model_selection import train_test_split
      X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.30)
```

```
import xgboost as xgb

#Aplicando Regresión utilizando métrica: Error al cuadrático
#con una tasa de aprendizaje de 0.1 y 400 árboles de decisión o estimadores
modelo = xgb.XGBRegressor(objective='reg:squarederror', learning_rate = 0.1, max_depth = 30, n_estimators = 400)
modelo.fit(X_train, y_train) #ajustando datos
```

Evaluando puntuación o certeza del modelo de entrenamiento

```
# Predecir puntuación o Tasa de Mortalidad del modelo entrenado utilizando el conjunto de datos de prueba
result = modelo.score(X_test, y_test) #le damos una puntuación según el modelo de entrenamiento
print("Certeza : {}".format(result))

Certeza : -0.21666472402827597
```

La certeza fue negativa la cual indica que según los datos entrenados la Tasa de Mortalidad es muy aleatoria y es muy poco impactada por las otras variables o factores elegidos

5.- Evaluación del Modelo

```
from sklearn.metrics import r2_score, mean_squared_error, mean_absolute_error
from math import sqrt
n = len(X_test)
k = X_test.shape[1]
print(k)
print("La cantidad de datos de prueba es: ", n)
print("La cantidad de datos de entrenamiento es: ", len(X_train))
RMSE = float(format(np.sqrt(mean_squared_error(y_test, y_predict)), '.3f')) #3 decimales
MSE = mean_squared_error(y_test, y_predict)
MAE = mean_absolute_error(y_test, y_predict)
r2 = r2_score(y_test, y_predict)
ajuste_r2 = 1 - (1-r2) * (n-1) / (n-k-1)

print('RMSE =', RMSE, '\nMSE =', MSE, '\nMAE =', MAE, '\nR2 =', r2, 'El valor ajustado de r2 = ', ajuste_r2)

##RMSE fue alto lo cual indica que el modelo tiene una buena predicción
#R2 0.16 indica un 16% en la variación de los datos

#Se concluye que el modelo de predicción es efectivo según la muestra entrenada
#Pero la Hipótesis planteada fue respondida desde la correlación.
```

6

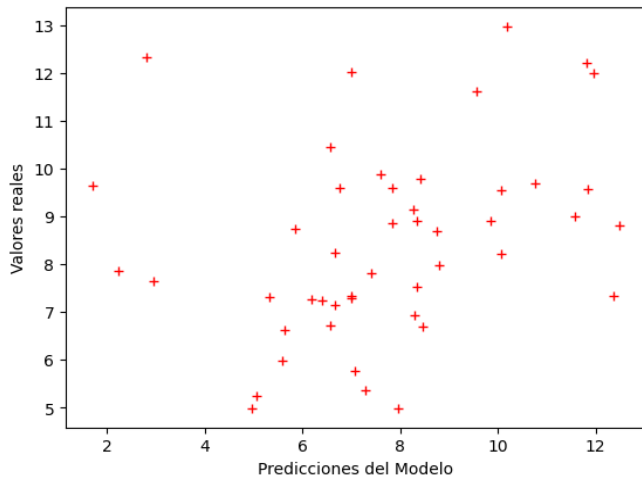
La cantidad de datos de prueba es: 47
La cantidad de datos de entrenamiento es: 109
RMSE = 2.825
MSE = 7.979743166461337
MAE = 1.9815516873623464
R2 = -0.21666472402827597 El valor ajustado de r2 = -0.39916443263251744

- RMSE fue alto lo cual indica que el modelo tiene una buena predicción acorde a los datos de prueba
- R2 0.16 indica un 16% en la variación de los datos
- Se concluye que el modelo de predicción es efectivo según la muestra entrenada, pero la Hipótesis planteada fue respondida desde la correlación.

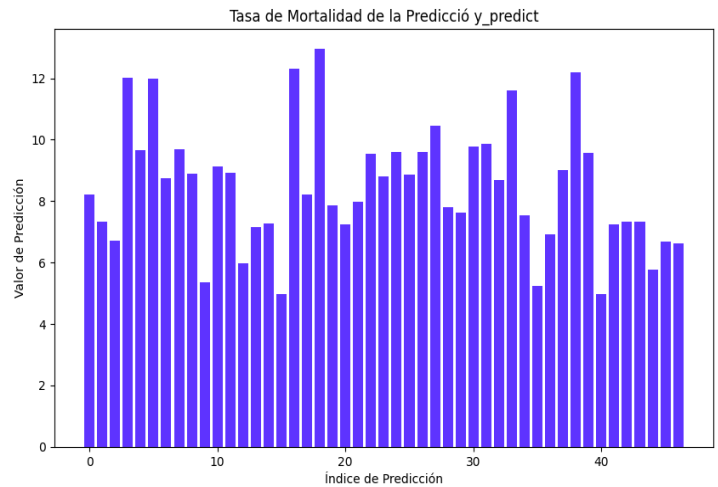
6.- Visualización de Resultados

Graficando datos de la predicción $y_{predict}$

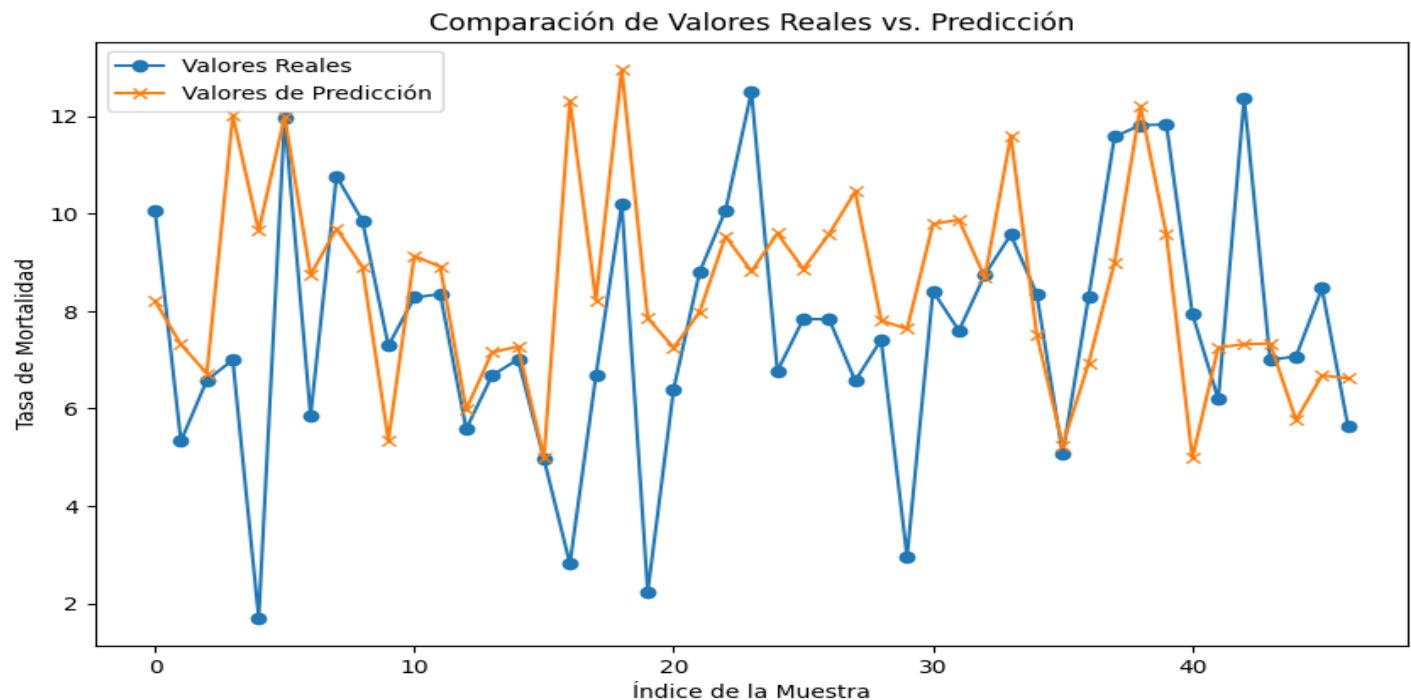
Graficando Dispersión de la Predicción



Graficando en Barras la Predicción



Comparativo de los datos de prueba con los datos de la predicción



Analizando el gráfico de la predicción vs los valores reales de muestra gráficamente se observa que la precisión de la predicción es aproximada, pero con ciertos márgenes de error, pero a final de cuentas el algoritmo cumplió con su objetivo, Se observa también que mantienen la misma aleatoriedad de las tasas entre los países independientemente de las variables factores planteadas en la hipótesis. Ya que anteriormente en la correlación se determinó que tienen un bajo impacto.

7.- Conclusiones y Recomendaciones

Respondiendo a la hipótesis planteada se logró determinar que la variable objetivo "Tasa de Mortalidad", es poco impactada por los factores elegidos en la hipótesis como: PIB del País e Inversión en Salud ya que los datos registrados en el dataset realmente están muy dispersos o aleatorios.

La correlación con las otras variables es muy baja, entendiendo que cada país por cuestiones de cultura, nivel de desarrollo económico, idiosincrasia, desarrollo social es diferente en cada uno de éstos y evidentemente causaran una aleatoriedad en la Tasa de Mortalidad

Quizá se lograría un resultado más certero haciendo predicciones de la tasa de mortalidad, con registros del mismo país durante varios años para poder tener una correlación con más impacto entre las otras variables factores.

8.- Despliegue del Modelo

El Modelo si puede ser implementado en un entorno productivo en una API o WEB SERVICE con los siguientes métodos básicos.

API /Web Service

1. Con un Método para la Alta o Captura de Registros de la Tasa de Mortalidad y sus factores de cada país, y por cada Año transcurrido
- 2.- Un Método de consulta para generar predicciones de Tasa de Mortalidad con parámetros para una consulta más dinámica como:
 - Continente (opcional)
 - País (opcional)
 - Métrica
 - Tasa Aprendizaje
 - No. Estimadores

Una vez construida la API o Webservice pueden integrarse en Páginas Web, Aplicaciones Móviles o Aplicaciones de escritorio para fines que al interesado convengan.