

Análisis de expectativa de vida en base a datos de salud, educación e ingreso PBI

CODER HOUSE

Contexto:

Aunque en el pasado se han realizado muchos estudios sobre los factores que afectan la esperanza de vida considerando variables demográficas, composición de ingresos y tasas de mortalidad. Se descubrió que en el pasado no se tenía en cuenta el efecto de la inmunización y el índice de desarrollo humano. Además, algunas de las investigaciones anteriores se realizaron considerando una regresión lineal múltiple basada en un conjunto de datos de un año para todos los países. Por lo tanto, esto motiva a resolver ambos factores establecidos anteriormente mediante la formulación de un modelo de regresión basado en un modelo de efectos mixtos y una regresión lineal múltiple considerando datos de un período de 2000 a 2015 para todos los países. También se considerarán vacunas importantes como la hepatitis B, la polio y la difteria. En pocas palabras, este estudio se centrará en los factores de inmunización, factores de mortalidad, factores económicos, factores sociales y también otros factores relacionados con la salud. Dado que las observaciones de este conjunto de datos se basan en diferentes países, será más fácil para un país determinar el factor de predicción que contribuye al menor valor de la esperanza de vida. Esto ayudará a sugerir a un país a qué área se le debe dar importancia para mejorar eficientemente la esperanza de vida de su población.

En base a esto. Podemos presentar las siguientes preguntas de interés:

- ¿Existe una correlación entre la tasa de inmunización (Hepatitis B, Polio, Difteria) y la esperanza de vida en diferentes países durante el período de 2000 a 2015?
- ¿Cómo se relaciona el índice de desarrollo humano (IDH) con la esperanza de vida y la mortalidad adulta a lo largo de los años estudiados?
- ¿Hay una conexión entre el gasto en salud (representado por el porcentaje del PIB dedicado a la salud y el gasto total) y la incidencia de enfermedades mortales como el VIH/SIDA en países de diferentes niveles económicos?
- ¿Cuál es la influencia de la situación económica (PIB per cápita) en la desnutrición (thinness) de diferentes grupos de edad (1-19 años y 5-9 años) en distintos países?
- ¿Cómo se correlacionan los niveles de educación (escolaridad) y la composición de ingresos con la esperanza de vida, considerando la distribución por países y el transcurso de los años?

Para esto, Utilizaremos el siguiente dataset, que contiene las siguientes columnas:

Variable	Descripción
Country	País
Year	Año
Status	Estado Desarrollado o en Desarrollo
Life expectancy	Esperanza de vida en años
Adult Mortality	Tasas de mortalidad adulta de ambos sexos (probabilidad de morir entre los 15 y 60 años por cada 1000 habitantes)
Infant deaths	Número de muertes infantiles por cada 1000 habitantes
Alcohol	Consumo de alcohol per cápita registrado (15+) (en litros de alcohol puro)
Percentage expenditure	Gasto en salud como porcentaje del Producto Interno Bruto per cápita (%)
Hepatitis B	Cobertura de vacunación contra la Hepatitis B (HepB) entre niños de 1 año (%)
Measles	Sarampión: número de casos reportados por cada 1000 habitantes
BMI	Índice de Masa Corporal promedio de toda la población
Under-five deaths	Número de muertes de menores de cinco años por cada 1000 habitantes
Polio	Cobertura de vacunación contra la Polio (Pol3) entre niños de 1 año (%)
Total expenditure	Gasto gubernamental general en salud como porcentaje del gasto gubernamental total (%)
Diphtheria	Cobertura de vacunación contra la Difteria, tétanos y tos ferina (DTP3) entre niños de 1 año (%)
HIV/AIDS	Muertes por VIH/SIDA por cada 1000 nacidos vivos (0-4 años)
GDP	Producto Interno Bruto per cápita (en USD)
Population	Población del país
Thinness 1-19 years	Prevalencia de delgadez entre niños y adolescentes de 10 a 19 años (%)
Thinness 5-9 years	Prevalencia de delgadez entre niños de 5 a 9 años (%)
Income composition of resources	Índice de Desarrollo Humano en términos de composición del ingreso de recursos (índice que va de 0 a 1)
Schooling	Número de años de escolarización (años)

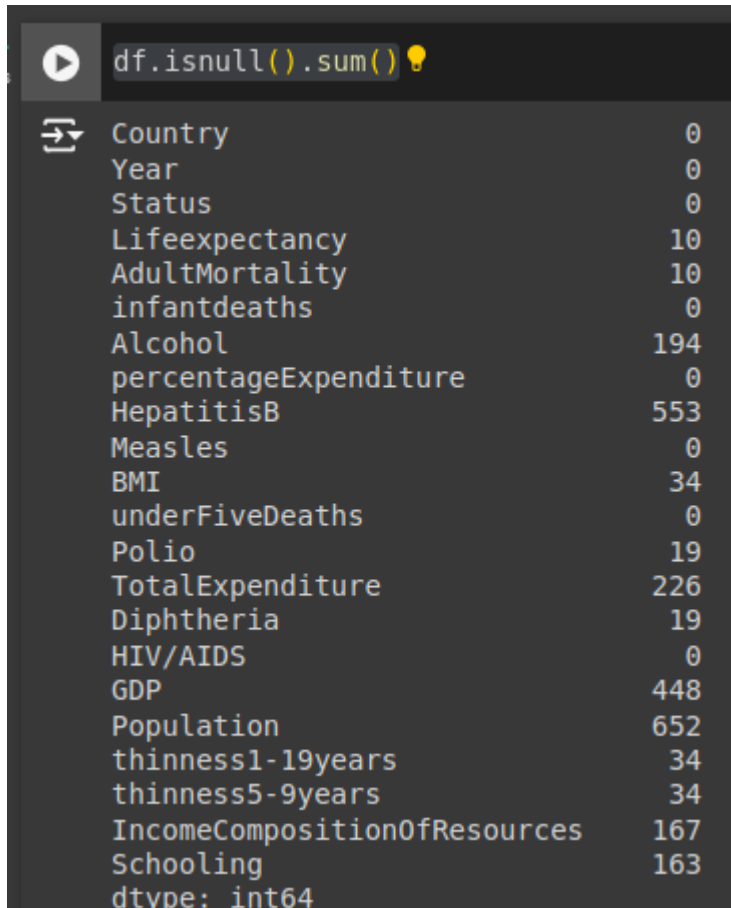
```
# Column Non-Null Count Dtype
---
0 Country 2938 non-null object
1 Year 2938 non-null int64
2 Status 2938 non-null object
3 Lifeexpectancy 2928 non-null float64
4 AdultMortality 2928 non-null float64
5 infantdeaths 2938 non-null int64
6 Alcohol 2744 non-null float64
7 percentageExpenditure 2938 non-null float64
8 HepatitisB 2385 non-null float64
9 Measles 2938 non-null int64
10 BMI 2904 non-null float64
11 underFiveDeaths 2938 non-null int64
12 Polio 2919 non-null float64
13 TotalExpenditure 2712 non-null float64
14 Diphtheria 2919 non-null float64
15 HIV/AIDS 2938 non-null float64
16 GDP 2490 non-null float64
17 Population 2286 non-null float64
18 thinness1-19years 2904 non-null float64
19 thinness5-9years 2904 non-null float64
```

20 IncomeCompositionOfResources 2771 non-null float64

21 Schooling 2775 non-null float64

Preparación y limpieza de datos:

1. El primer paso fue determinar cuales tuplas contengan valores nulos:



Country	0
Year	0
Status	0
Lifeexpectancy	10
AdultMortality	10
infantdeaths	0
Alcohol	194
percentageExpenditure	0
HepatitisB	553
Measles	0
BMI	34
underFiveDeaths	0
Polio	19
TotalExpenditure	226
Diphtheria	19
HIV/AIDS	0
GDP	448
Population	652
thinness1-19years	34
thinness5-9years	34
IncomeCompositionOfResources	167
Schooling	163
dtype: int64	

2. En este caso, al representar un poco porcentaje. Se opta por cambiar los valores nulos, al valor medio de cada campo. lo que deja un dataset para poder empezar a tener algunas visualizaciones de datos.

Objetivos:

El objetivo del proyecto es evaluar los factores que influyen en la expectativa de vida. Para ello, se analizarán cómo diferentes variables, tales como la cantidad de vacunas aplicadas, el salario y los años de estudio, afectan dicha expectativa.

En detalle, los objetivos específicos de este proyecto de data science son:

Identificar Factores Clave: Determinar cuáles son los factores más influyentes en la expectativa de vida a partir de un conjunto de variables socioeconómicas y de salud, como la mortalidad adulta, las muertes infantiles, el consumo de alcohol, el gasto en salud, la vacunación contra diversas enfermedades, el índice de masa corporal (BMI), el PIB per cápita, la población, la prevalencia de delgadez en diferentes rangos de edad, y el nivel educativo promedio.

Analizar Tendencias Temporales y Geográficas: Evaluar cómo estos factores varían a lo largo del tiempo y entre diferentes países, identificando patrones regionales y temporales que puedan explicar diferencias en la expectativa de vida.

Evaluar el Impacto de las Vacunas: Investigar específicamente el impacto de la vacunación (Hepatitis B, Polio, Difteria) en la expectativa de vida, considerando tanto la cobertura de vacunación como la incidencia de enfermedades prevenibles por vacunas.

Relacionar Indicadores Socioeconómicos y Salud: Examinar la relación entre indicadores socioeconómicos como el ingreso per cápita (GDP), la composición de recursos e ingresos, y el nivel educativo (Schooling) con la expectativa de vida, para entender cómo la riqueza y la educación contribuyen a una vida más larga y saludable.

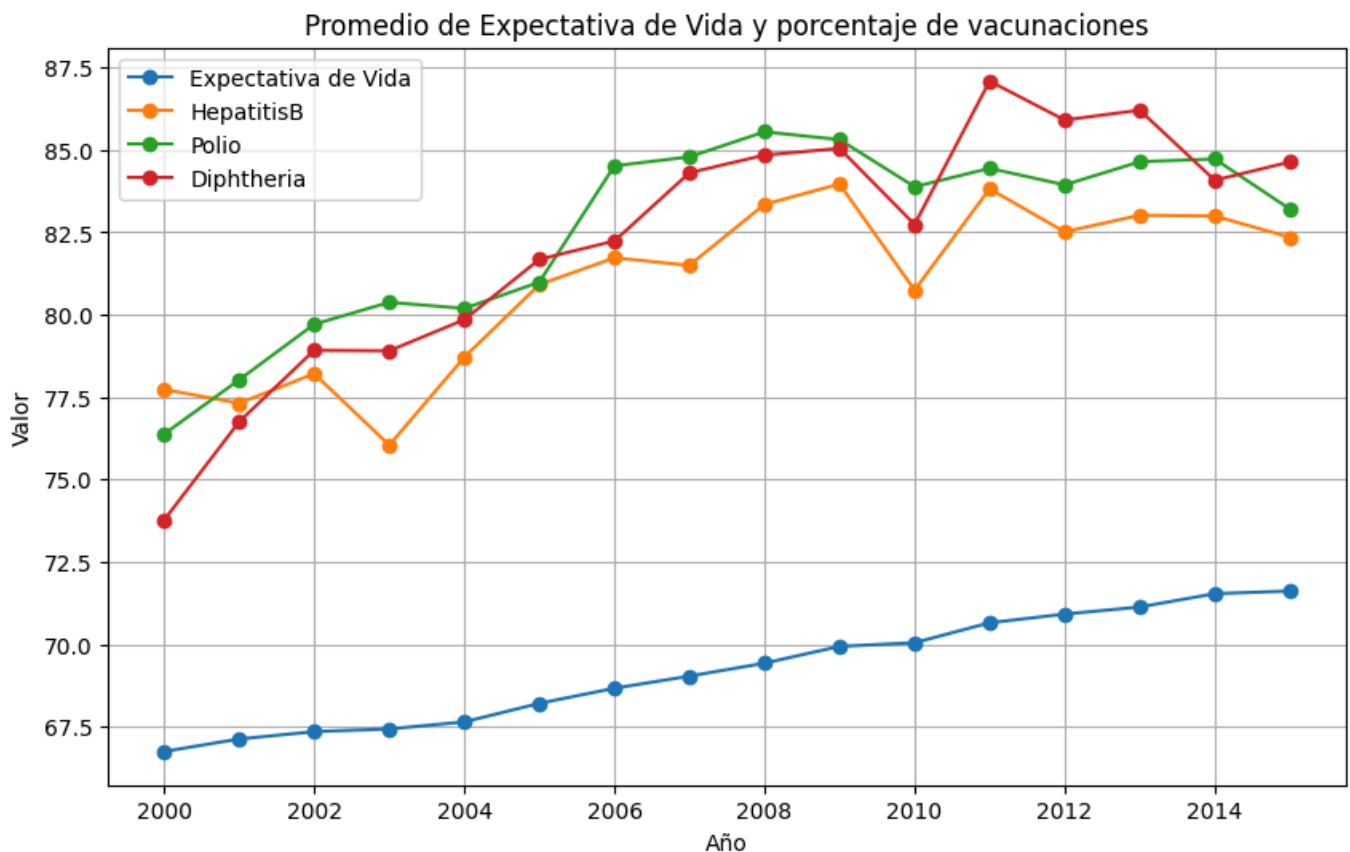
Impacto de las Enfermedades y Estilos de Vida: Evaluar cómo las enfermedades (HIV/AIDS) y los estilos de vida (consumo de alcohol, BMI) afectan la longevidad, proporcionando una visión integral de los factores que pueden ser modificables mediante políticas públicas y programas de salud.

Proporcionar Recomendaciones de Política: Basado en el análisis, ofrecer recomendaciones de políticas públicas y estrategias de intervención que puedan ayudar a mejorar la expectativa de vida, especialmente en países con indicadores de salud deficientes o en vías de desarrollo.

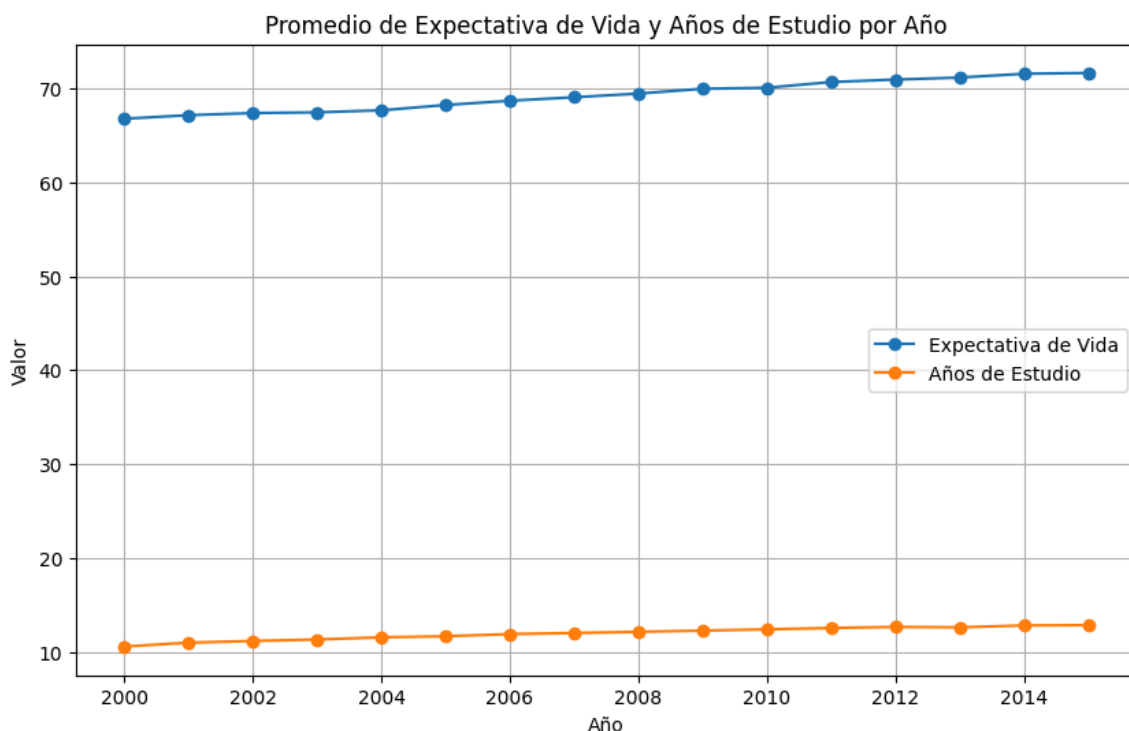
Desarrollar Modelos Predictivos: Construir modelos predictivos que puedan estimar la expectativa de vida basándose en las variables disponibles, proporcionando herramientas útiles para la planificación y evaluación de políticas de salud pública.

En resumen, este proyecto de data science tiene como objetivo no solo entender los factores que determinan la expectativa de vida, sino también ofrecer soluciones prácticas y basadas en datos para mejorarla a nivel global.

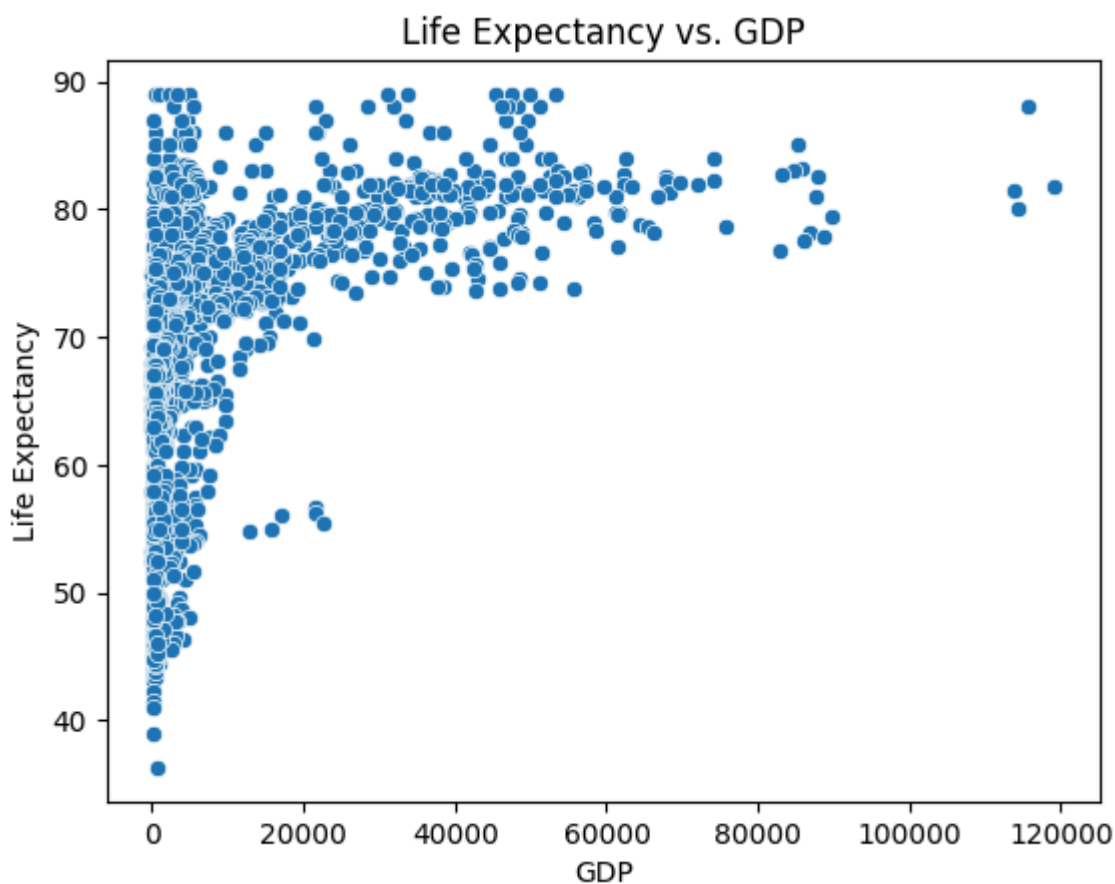
Insight y recomendaciones:



Un primer indicador que podemos encontrar. Es que a mayor es la tasa de vacunación de la población. Esta tiende a ser más duradera



De igual manera. Se puede relacionar los años de estudio con la expectativa de vida, El anterior gráfico corresponde a un promedio mundial. En donde se puede ver esta correlación, y que también año a año se aumentan estas variables



Este último gráfico. Tiene estrecha relación con el anterior gráfico. En donde se puede ver, Que a mayor ingreso per cápita. Mayor es la expectativa de vida de su población

Aplicando un modelo predictivo

Primer Modelo:

En este primer caso. Se mostrará un pequeño ejemplo para ver cuáles variables son relevantes

```
model2 = 'Lifeexpectancy~AdultMortality+infantdeaths+Alcohol+HepatitisB+Measles+BMI+underFiveDeaths'
lm2 = sm.ols(formula=model2, data=df).fit()
print(lm2.summary())
```

OLS Regression Results

	coef	std err	t	P> t	[0.025	0.975]
Dep. Variable:	Lifeexpectancy					
Model:	OLS					
Method:	Least Squares					
Date:	Sun, 26 May 2024					
Time:	19:00:56					
No. Observations:	2938					
Df Residuals:	2925					
Df Model:	12					
Covariance Type:	nonrobust					
R-squared:	0.770					
Adj. R-squared:	0.770					
F-statistic:	818.1					
Prob (F-statistic):	0.00					
Log-Likelihood:	-8623.2					
AIC:	1.727e+04					
BIC:	1.735e+04					
Intercept	54.7002	0.565	96.893	0.000	53.593	55.807
AdultMortality	-0.0302	0.001	-38.160	0.000	-0.032	-0.029
infantdeaths	0.1098	0.009	11.810	0.000	0.092	0.128
Alcohol	0.1230	0.025	4.826	0.000	0.073	0.173
HepatitisB	-0.0123	0.004	-2.799	0.005	-0.021	-0.004
Measles	-2.369e-05	8.59e-06	-2.757	0.006	-4.05e-05	-6.84e-06
BMI	0.0586	0.005	11.360	0.000	0.048	0.069
underFiveDeaths	-0.0823	0.007	-11.955	0.000	-0.096	-0.069
Polio	0.0282	0.005	5.606	0.000	0.018	0.038
Diphtheria	0.0406	0.005	7.677	0.000	0.030	0.051
GDP	5.396e-05	7.33e-06	7.365	0.000	3.96e-05	6.83e-05
IncomeCompositionOfResources	6.5176	0.710	9.179	0.000	5.125	7.910
Schooling	0.6431	0.047	13.735	0.000	0.551	0.735
Omnibus:	458.923		Durbin-Watson:	0.820		
Prob(Omnibus):	0.000		Jarque-Bera (JB):	1684.838		
Skew:	-0.747		Prob(JB):	0.00		
Kurtosis:	6.396		Cond. No.	1.29e+05		

En este caso se puede observar que las variables Population, TotalExpenditure, percentageExpenditure y Year, al Beta poder ser cero, No son de importancia para utilizar en el entrenamiento de un modelo, por lo cual se crea el siguiente modelo

Segundo Modelo:

OLS Regression Results

Dep. Variable:

Lifeexpectancy

R-squared:

0.770

Model:

OLS

Adj. R-squared:

0.770

Method:

Least Squares

F-statistic:

818.1

Date:

Sun, 26 May 2024

Prob (F-statistic):

0.00

Time:

19:00:56

Log-Likelihood:

-8623.2

No. Observations:

2938

AIC:

1.727e+04

Df Residuals:

2925

BIC:

1.735e+04

Df Model:

12

Covariance Type:

nonrobust

<

En este caso, al sacar las variables que no pasaban el test de beta, podemos observar que el R^2 disminuye muy poco(0.770), por lo cual se confirma que se pueden discriminar las variables eliminadas, al tener un porcentaje del 77% de predicción, se creará un tercer modelo para intentar reducir variables para disminuir el costo de computo del modelo.

Tercer Modelo:

OLS Regression Results

=====

Dep. Variable:

Lifeexpectancy

R-squared:

0.627

Model:

OLS

Adj. R-squared:

0.626

Method:

Least Squares

F-statistic:

547.8

Date:

Sun, 26 May 2024

Prob (F-statistic):

0.00

Time:

19:00:56

Log-Likelihood:

-9334.8

No. Observations:

2938

AIC:

1.869e+04

Df Residuals:

2928

BIC:

1.875e+04

Df Model:

9

Covariance Type:

nonrobust

=====

coef

std err

t

P>|t|

[0.025

0.975]

Intercept

42.7706

0.594

72.062

0.000

41.607

43.934

infantdeaths

0.1354

0.012

11.631

0.000

0.113

0.158

HepatitisB

-0.0099

0.006

-1.771

0.077

-0.021

0.001

Measles

-1.23e-05

1.09e-05

-1.129

0.259

-3.37e-05

9.07e-06

underFiveDeaths

-0.1022

0.009

-11.849

0.000

-0.119

-0.085

Polio

0.0415

0.006

6.499

0.000

0.029

0.054

Diphtheria

0.0501

0.007

7.449

0.000

0.037

0.063

GDP

8.716e-05

9.22e-06

9.452

0.000

6.91e-05

0.000

IncomeCompositionOfResources

11.3481

0.891

12.736

0.000

9.601

13.095

Schooling

1.0131

0.056

18.022

0.000

0.903

1.123

=====

Omnibus:

253.865

Durbin-Watson:

0.377

Prob(Omnibus):

0.000

Jarque-Bera (JB):

690.420

Skew:

-0.477

Prob(JB):

1.19e-150

Kurtosis:

5.175

Cond. No.

1.26e+05

=====

Al eliminar Population y TotalExpenditure, se llega al porcentaje de R^2 de 0.627, lo que representa una pérdida de la capacidad de predicción del 15%, por lo cual no se estima utilizarlo para entrenar el modelo

Generar nuevo dataset basado en variables de interés.

Para este caso. Y viendo los resultados de las pruebas anteriores. Se optó por no usar las siguientes Columnas:

- Country
- Status
- Population
- TotalExpenditure
- percentageExpenditure
- Year

Quedando formado el dataset, de la siguiente manera.

Data columns (total 16 columns):

#	Column	Non-Null Count	Dtype
0	Lifeexpectancy	2938 non-null	float64
1	AdultMortality	2938 non-null	float64
2	infantdeaths	2938 non-null	int64
3	Alcohol	2938 non-null	float64
4	HepatitisB	2938 non-null	float64
5	Measles	2938 non-null	int64
6	BMI	2938 non-null	float64
7	underFiveDeaths	2938 non-null	int64
8	Polio	2938 non-null	float64
9	Diphtheria	2938 non-null	float64
10	HIV/AIDS	2938 non-null	float64
11	GDP	2938 non-null	float64
12	thinness1-19years	2938 non-null	float64
13	thinness5-9years	2938 non-null	float64
14	IncomeCompositionOfResources	2938 non-null	float64
15	Schooling	2938 non-null	float64

dtypes: float64(13), int64(2)

Primer Modelo: SKLearn (Regresión Lineal)

Para este caso, se utilizó una regresión lineal, dando como resultado un error cuadrático medio (MSE) de alrededor de 17 para las predicciones de un modelo de regresión lineal sobre la esperanza de vida puede interpretarse como el promedio de los cuadrados de las diferencias entre las predicciones y los valores reales de la esperanza de vida en el conjunto de prueba.

Un MSE de 16.9 indica que, en promedio, los valores predichos por el modelo están a una distancia de aproximadamente 4.1 años (ya que la raíz cuadrada del MSE es la desviación estándar) de los valores reales de la esperanza de vida en tu conjunto de prueba.

Es importante tener en cuenta que la interpretación del MSE puede variar dependiendo del contexto del problema y el rango de valores de la variable objetivo. En general, cuanto menor sea el MSE, mejor será la capacidad predictiva del modelo. Es útil comparar este valor con otros modelos o realizar técnicas adicionales de validación para evaluar más a fondo el rendimiento del modelo de regresión.


```

X = new_df.drop('Lifeexpectancy', axis=1)
y = new_df['Lifeexpectancy']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
model = LinearRegression()
model.fit(X_train, y_train)
predictions = model.predict(X_test)

# Calcular el error cuadrático medio (MSE)
mse = mean_squared_error(y_test, predictions)
print("MSE:", mse)
rmse = np.sqrt(mse)
rmse_rounded = round(rmse, 2)
print("Error promedio de prediccion: ", rmse_rounded, "Años")

```

 MSE: 17.005774029925313
 Error promedio de prediccion: 4.12 Años

Segundo Modelo: SVM (Kernel Lineal)

En este segundo modelo, La variación con la prueba anterior de SKLearn, no fue muy diferente, Dando como resultado, Un error medio de 4.17 Años. Lo que al momento no es un buen resultado.

```

# Paso 1: Dividir el conjunto de datos en características (X) y la variable objetivo (y)
X = new_df.drop('Lifeexpectancy', axis=1)
y = new_df['Lifeexpectancy']

# Paso 2: Dividir los datos en conjuntos de entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)


# Paso 3: Escalar las características
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Paso 4: Entrenar el modelo SVM
svm_model = SVR(kernel='linear')
svm_model.fit(X_train_scaled, y_train)

# Paso 5: Hacer predicciones en el conjunto de prueba
y_pred = svm_model.predict(X_test_scaled)

# Paso 6: Calcular el Error Cuadrático Medio (MSE)
mse = mean_squared_error(y_test, y_pred)
print("MSE:", mse)
print("Años correspondientes al MSE:", round(mse**0.5, 2))

```

 MSE: 17.36561310130898
 Años correspondientes al MSE: 4.17

Tercer Modelo: SVM (Kernel RBF)

Este tercer modelo. Dio un valor mucho más aceptable, Como resultado, el mismo quedo dentro del rango de error cercano a un año, Lo que demuestra mucha mejor calidad de modelo.


```

[22] # Paso 4: Entrenar el modelo SVM con kernel radial (RBF)
svm_model_rbf = SVR(kernel='rbf')
svm_model_rbf.fit(X_train_scaled, y_train)

# Paso 5: Hacer predicciones en el conjunto de prueba
y_pred_rbf = svm_model_rbf.predict(X_test_scaled)

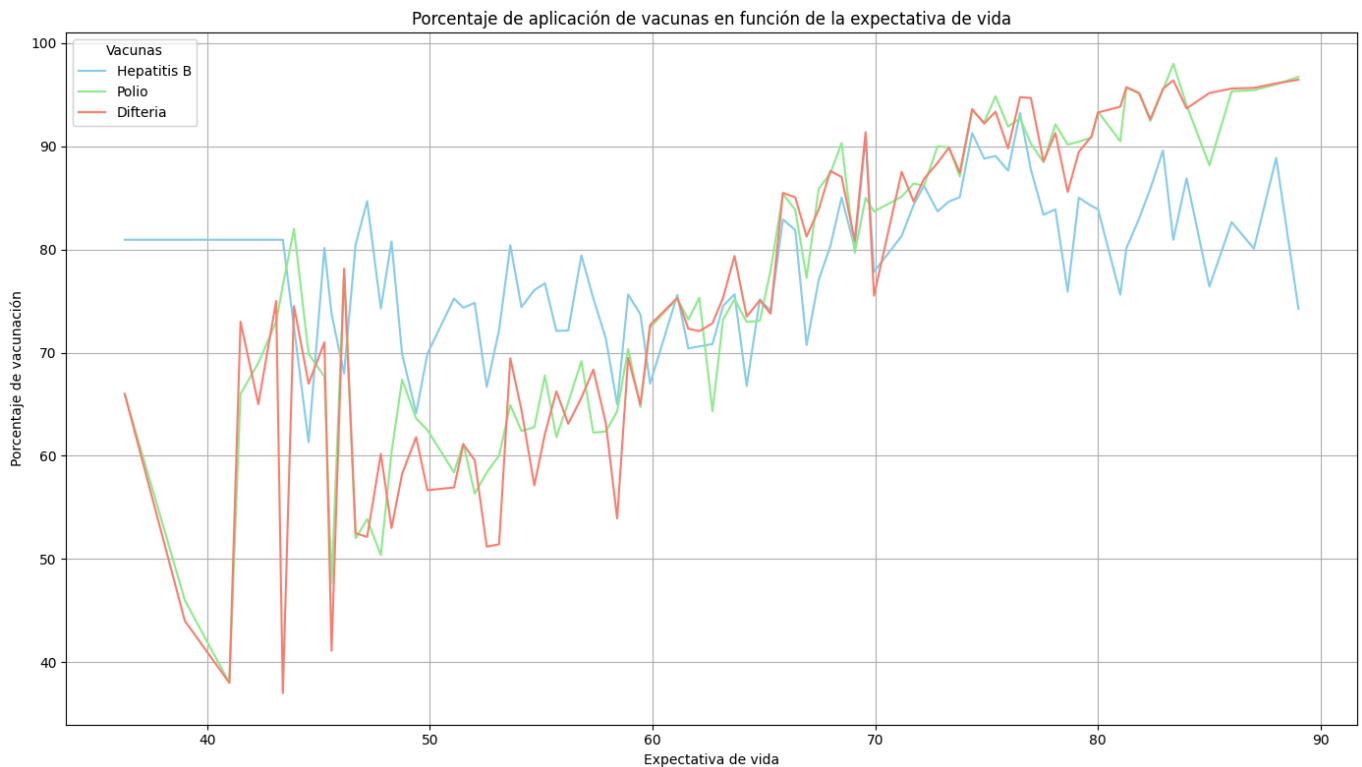
# Paso 6: Calcular el Error Cuadrático Medio (MSE) para el modelo con kernel radial
mse_rbf = mean_squared_error(y_test, y_pred_rbf)
print("MSE (Kernel Radial):", mse_rbf)
print("Años correspondientes al MSE (Kernel Radial):", round(mse_rbf**0.05, 3))

```

 MSE (Kernel Radial): 10.40872055253193

Conclusiones:

1. ¿Existe una correlación entre la tasa de inmunización (Hepatitis B, Polio, Difteria) y la esperanza de vida en diferentes países durante el período de 2000 a 2015?



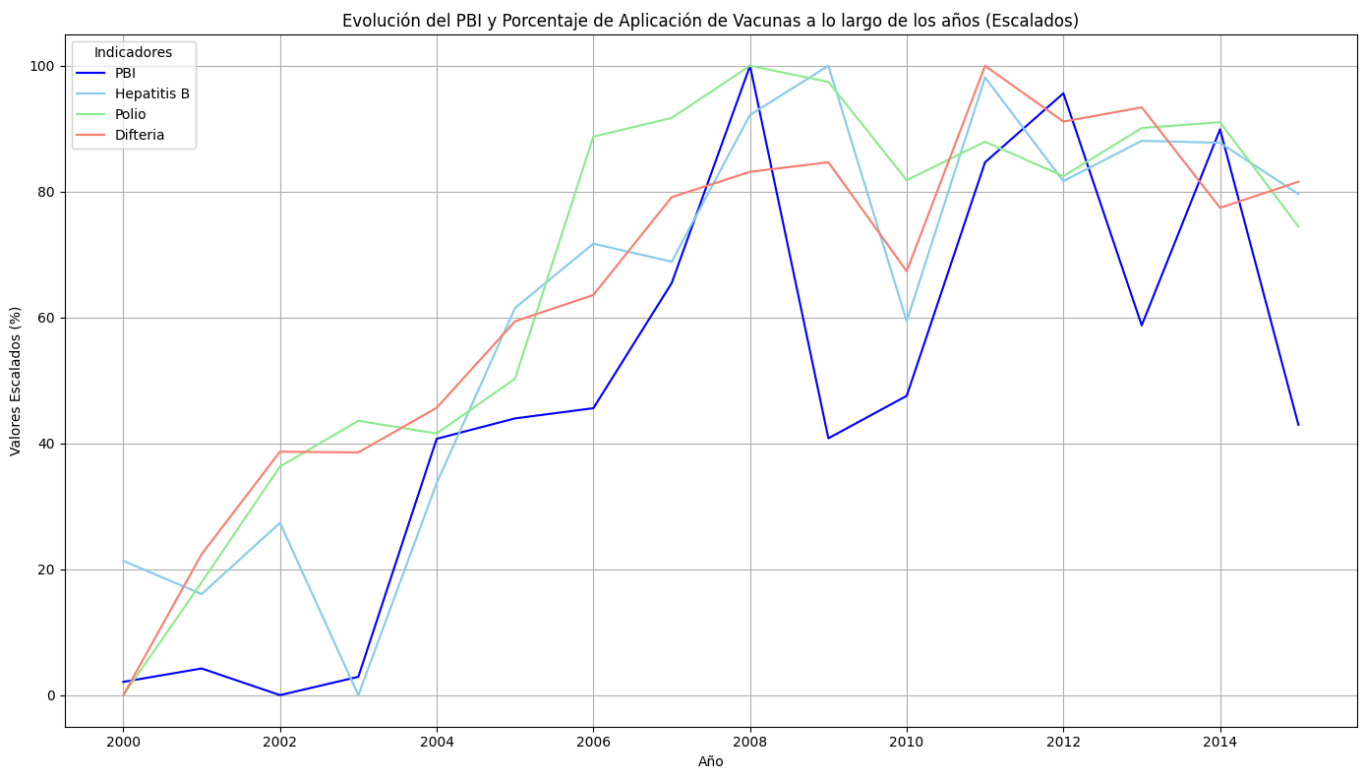
Si. Países con mayor tasa de aplicación de estas vacunas. Tienen una mayor expectativa de vida que países que las aplican en menor medida.

2. ¿Cómo se relaciona el índice de desarrollo humano (IDH) con la esperanza de vida y la mortalidad adulta a lo largo de los años estudiados?



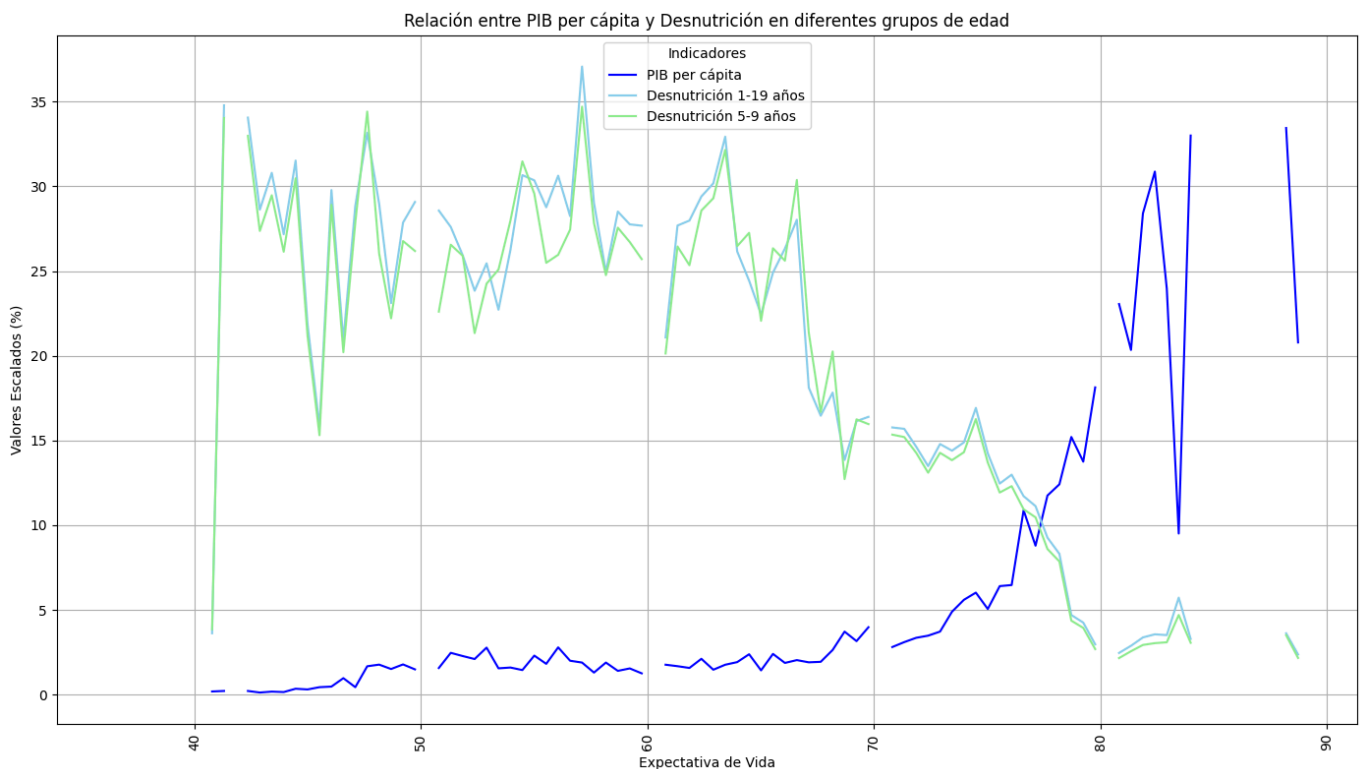
Esto se relaciona, de manera directa, con que los países con más desarrollo humano, tienden a tener mayor estudios, Mayor tasa de vacunación de enfermedades peligrosas, y también mayores ingresos medios.

3. ¿Hay una conexión entre el gasto en salud (representado por el porcentaje del PIB dedicado a la salud y el gasto total) y la incidencia de enfermedades mortales como el VIH/SIDA en países de diferentes niveles económicos?



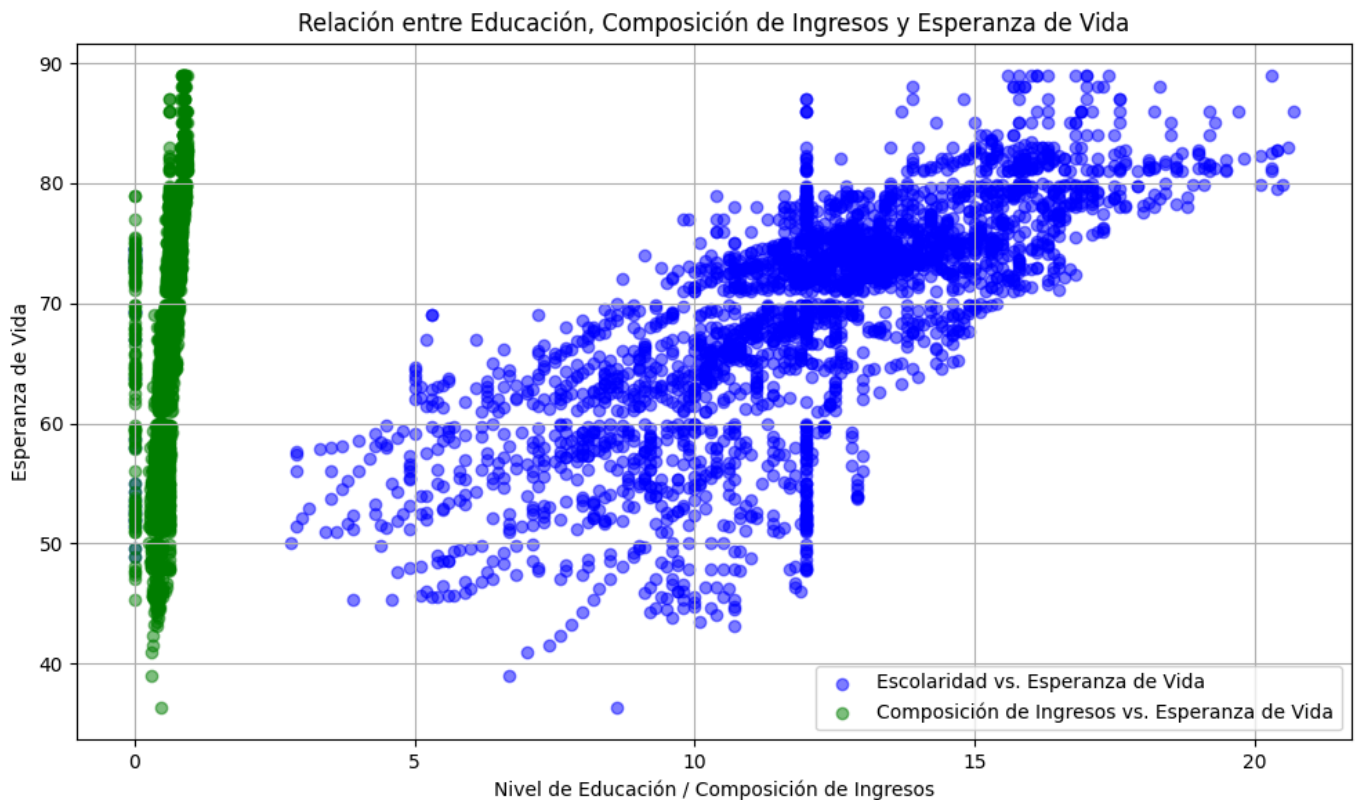
Si. Países con mayor gasto en salud, tienen mayor tasa de aplicación de vacunas de índole indispensable para tratar enfermedades peligrosas. De igual manera, tienen menor tasa de mortalidad adulta que la media mundial.

4. ¿Cuál es la influencia de la situación económica (PIB per cápita) en la desnutrición (thinness) de diferentes grupos de edad (1-19 años y 5-9 años) en distintos países?



La influencia de estas variables es directa. Se puede observar de manera muy clara, que países con un PBI muy bajo son los que mayor cantidad de desnutrición tienen. Dejando también expuesto que son los que tienen en base a los datos, el menor gasto en salud y educación. Lo que desencadena en una educación carente y una tasa de vacunación muy precaria.

5. ¿Cómo se correlacionan los niveles de educación (escolaridad) y la composición de ingresos con la esperanza de vida, considerando la distribución por países y el transcurso de los años?



En base a esta pregunta, también se puede ver que los niveles de escolaridad, Están atados directamente a la de los niveles de ingreso. Mientras mayores sean los niveles de escolaridad. Mayores son los ingresos de su población.