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



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:



# 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

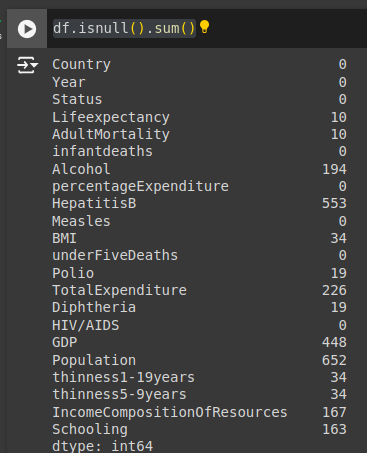
19 thinness5-9years 2904 non-null float64

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:



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.

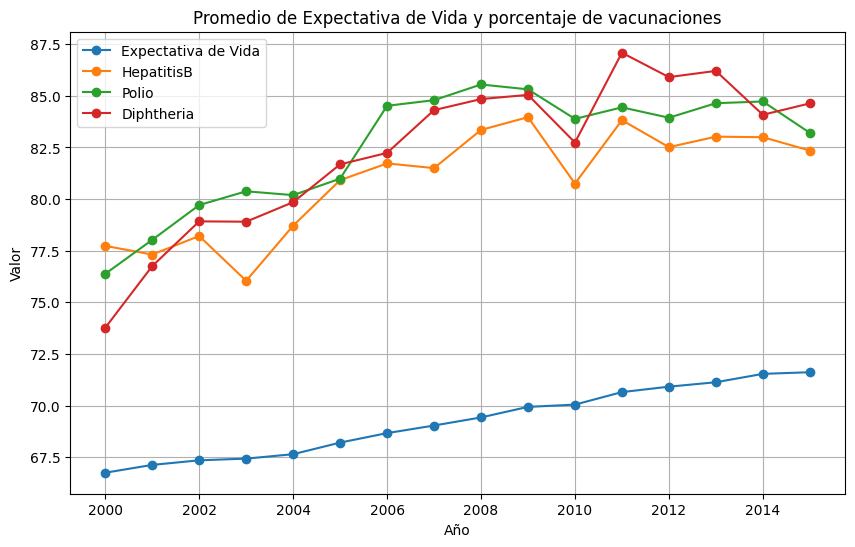
Impacto de las Enfermedades y Estilos de Vida: Evaluar cómo las enfermedades (HIV/AIDS) y los estilos de vida (consumo de alcohol, BMI) afectan la longevidad, proporcionando una visión integral de los factores que pueden ser modificables mediante políticas públicas y programas de salud.

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.

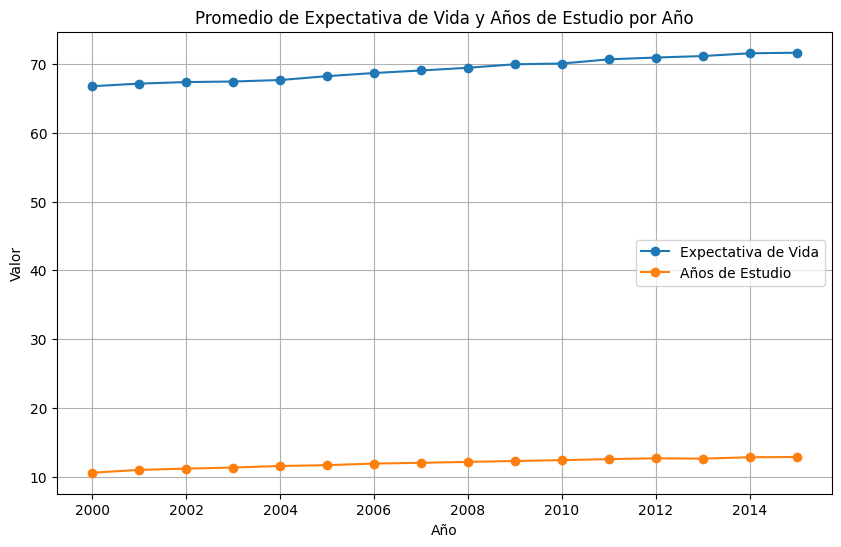
Desarrollar Modelos Predictivos: Construir modelos predictivos que puedan estimar la expectativa de vida basándose en las variables disponibles, proporcionando herramientas útiles para la planificación y evaluación de políticas de salud pública.

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.

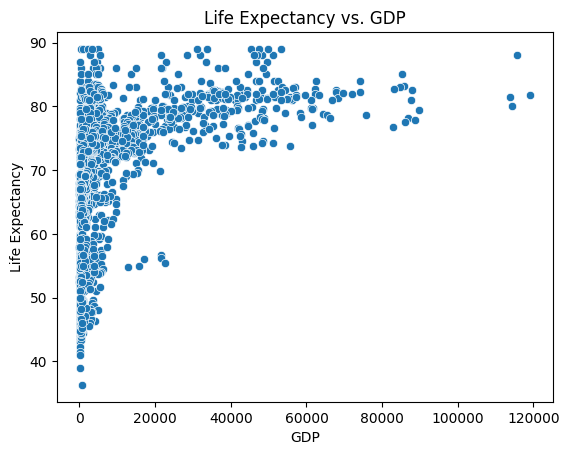
Insight y recomendaciones:



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

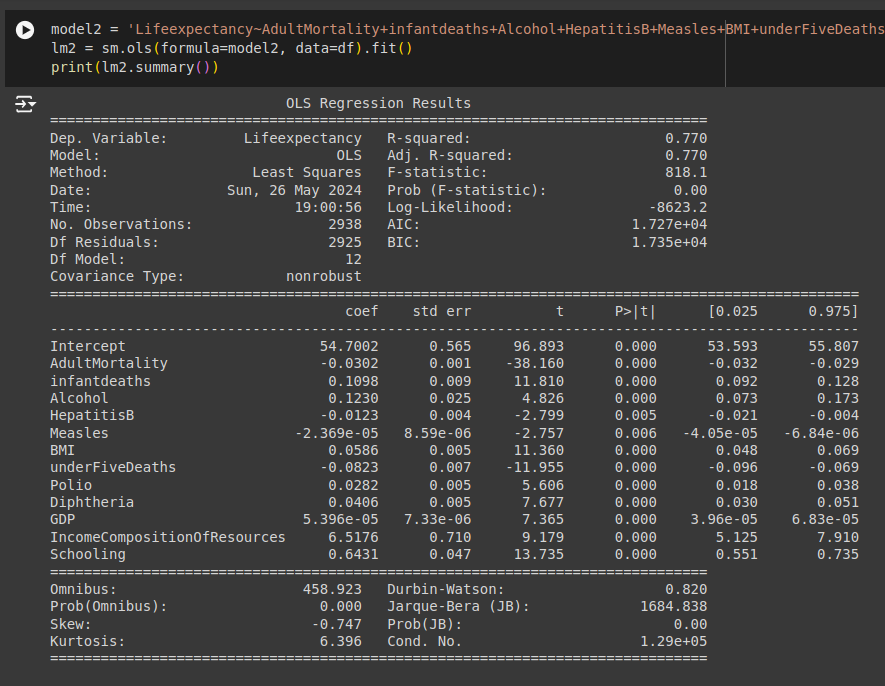


Este último gráfico. Tiene estrecha relación con el anterior gráfico. En donde se puede ver, Que a mayor ingreso per cápita. Mayor es la expectativa de vida de su población

Aplicando un modelo predictivo

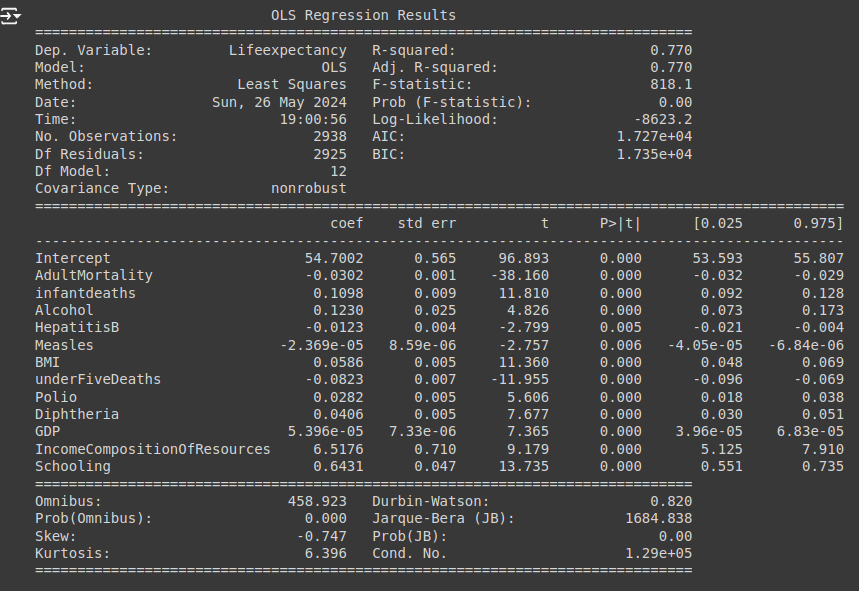
Primer Modelo:

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



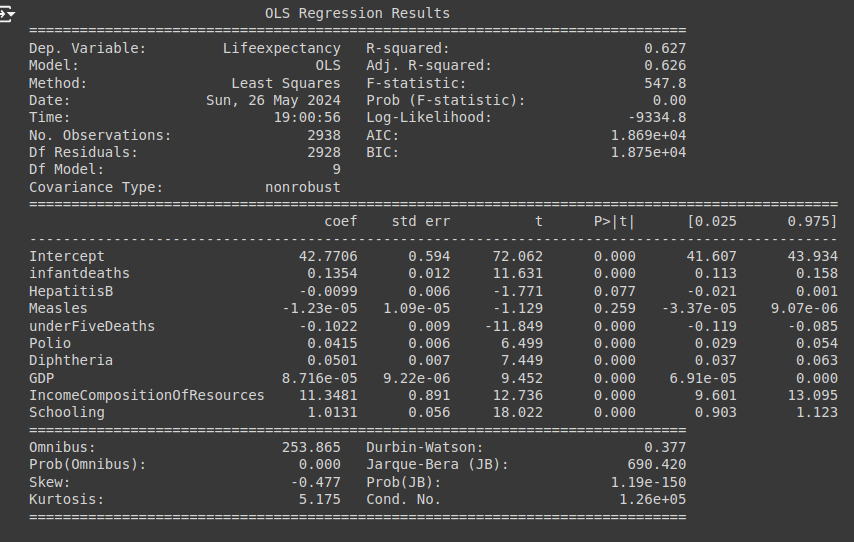
En este caso se puede observar que las variables Population, TotalExpenditure, percentageExpenditure y Year, al Beta poder ser cero, No son de importancia para utilizar en el entrenamiento de un modelo, por lo cual se crea el siguiente modelo

Segundo Modelo:



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:



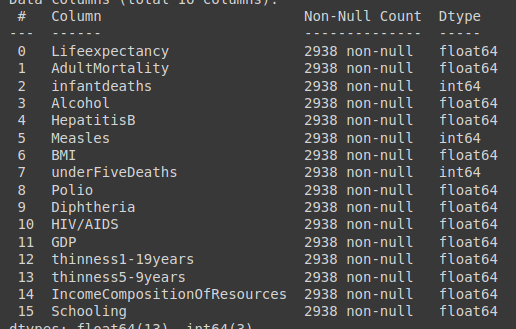
Al eliminar Population y TotalExpenditure, se llega al porcentaje de R^2 de 0.627, lo que representa una pérdida de la capacidad de predicción del 15%, por lo cual no se estima utilizarlo para entrenar el modelo

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.

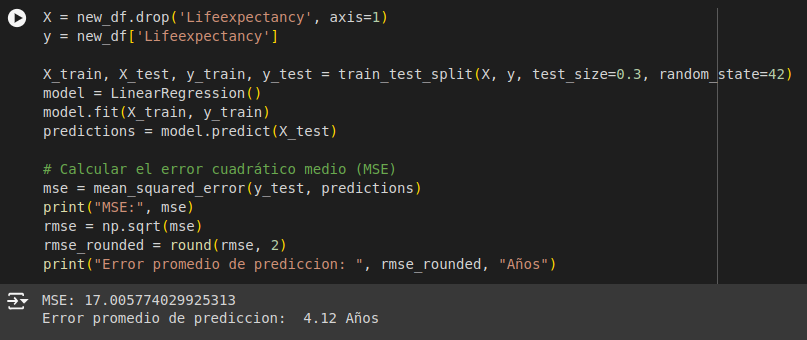


Primer Modelo: SKLearn (Regresión Lineal)

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.

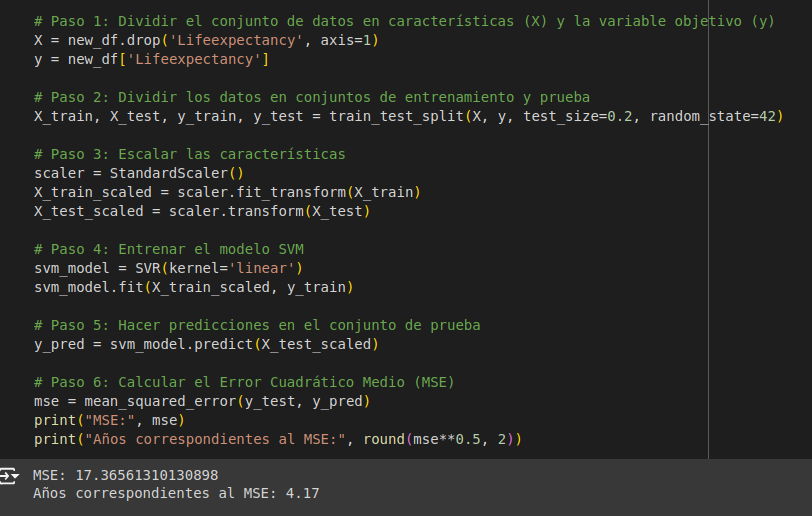
Un MSE de 16.9 indica que, en promedio, los valores predichos por el modelo están a una distancia de aproximadamente 4.1 años (ya que la raíz cuadrada del MSE es la desviación estándar) de los valores reales de la esperanza de vida en tu conjunto de prueba.

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.

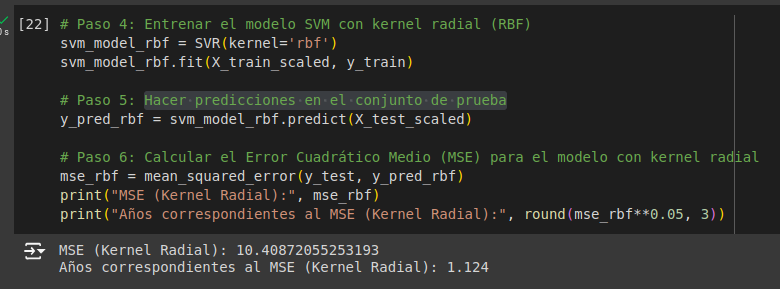


Segundo Modelo: SVM (Kernel Lineal)

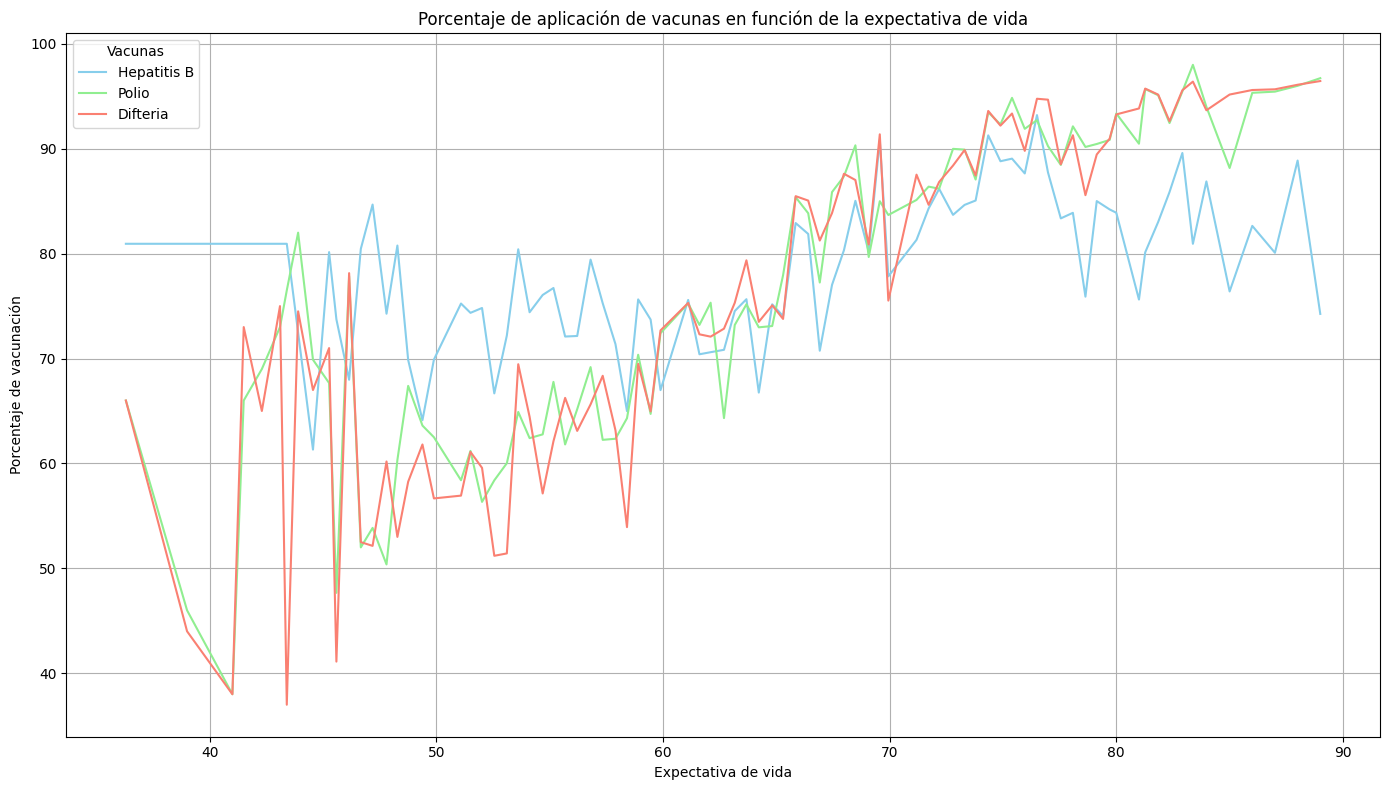
En este segundo modelo, La variación con la prueba anterior de SKLearn, no fue muy diferente, Dando como resultado, Un error medio de 4.17 Años. Lo que al momento no es un buen resultado.



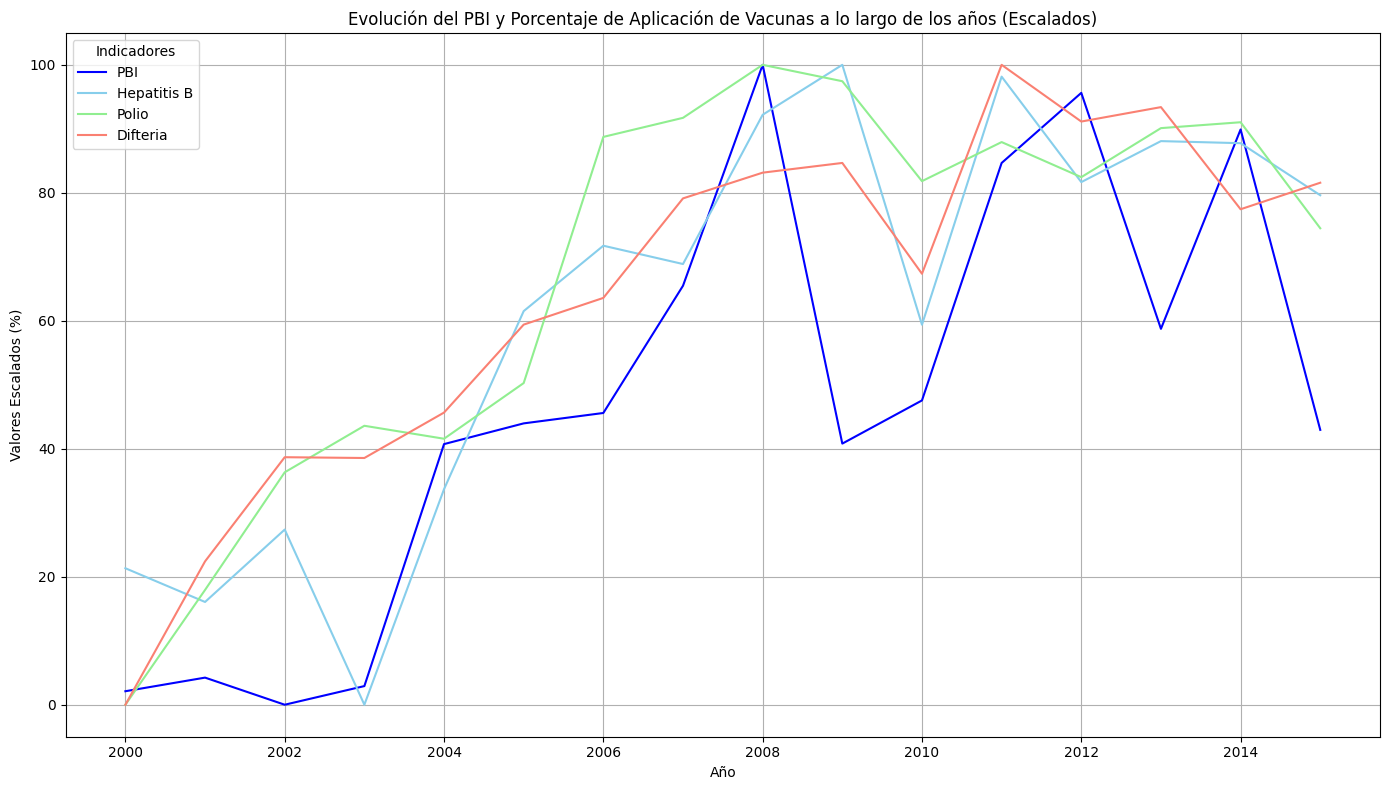
Tercer Modelo: SVM (Kernel RBF)

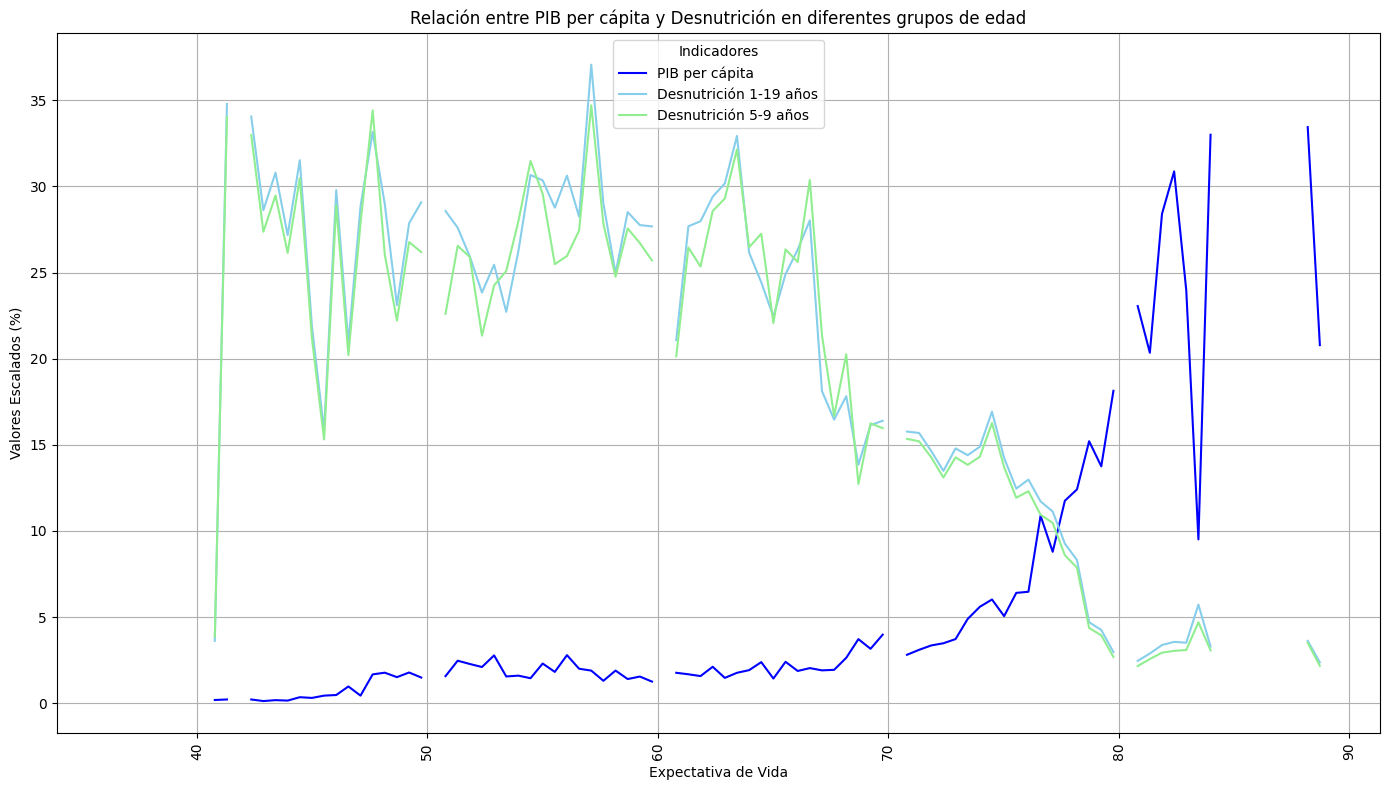
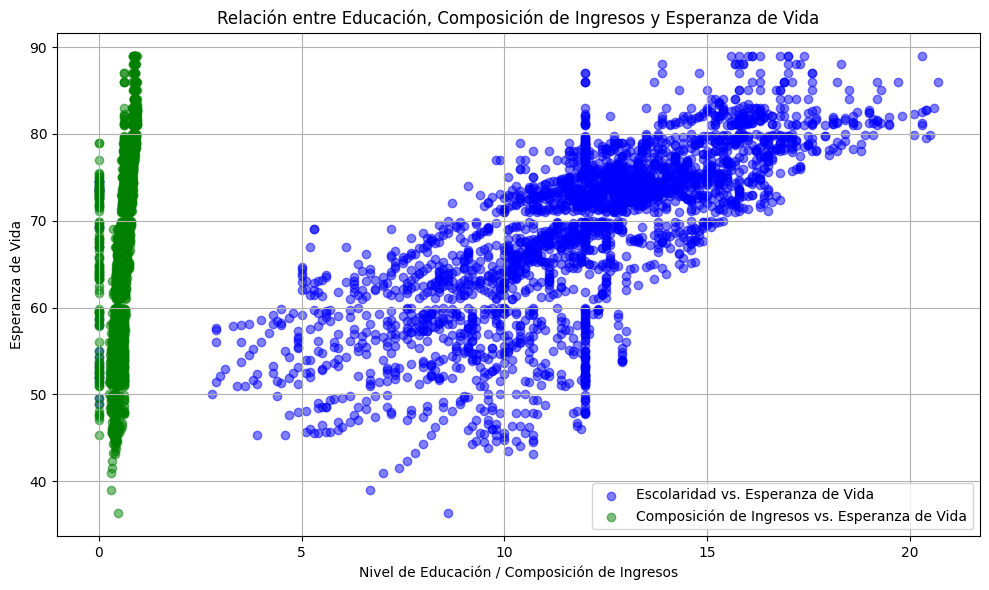
Este tercer modelo. Dio un valor mucho más aceptable, Como resultado, el mismo quedo dentro del rango de error cercano a un año, Lo que demuestra mucha mejor calidad de modelo.

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.

1. ¿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.

1. ¿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.
2. ¿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.