Memorias del Proyecto Final de Nuclio

Título del Proyecto: Evaluación de cómo factores socioeconómicos y de salud pública impactan la esperanza de vida en distintos países**,** utilizando análisis de datos y modelado predictivo para identificar estrategias de salud pública.

Autor(es): Albert Llobet y Arnau Martínez.
Institución: Nuclio Digital School.
Fecha: Setiembre 2024.
Curso: Master en Data Analytics.
Tutor o director del proyecto: Pedro y Tamara.

Abstract

En un contexto global de creciente interés por la salud pública y el bienestar socioeconómico, esta investigación se centra en la evaluación de cómo factores socioeconómicos y de salud pública impactan la esperanza de vida en distintos países. El estudio se basa en análisis de datos y modelado predictivo para identificar estrategias efectivas de salud pública.

El problema central abordado es entender cómo variables como el ingreso per cápita, el acceso a servicios de salud, la educación y otros determinantes sociales afectan la esperanza de vida en diferentes contextos nacionales. Investigaciones previas han demostrado que estos factores tienen un impacto significativo en la salud y longevidad de las poblaciones, pero aún queda por determinar cómo estas relaciones varían según el país y su desarrollo económico.

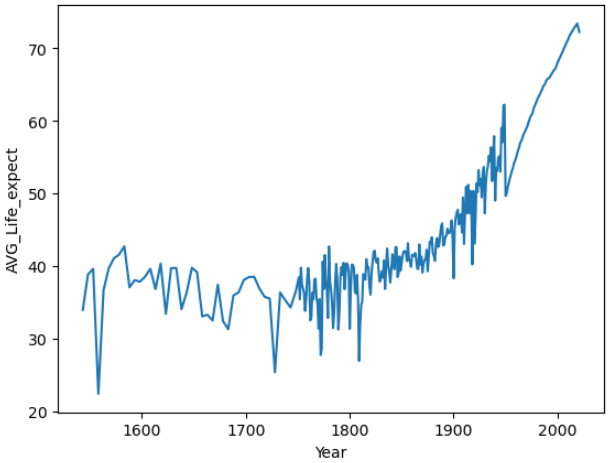
El objetivo principal de esta investigación es explorar y modelar estas relaciones de manera más precisa y detallada mediante el empleo de métodos analíticos avanzados. Los hallazgos de este estudio proporcionarán insights importantes para diseñar políticas públicas efectivas orientadas a mejorar la esperanza de vida y reducir las disparidades de salud entre diferentes países.

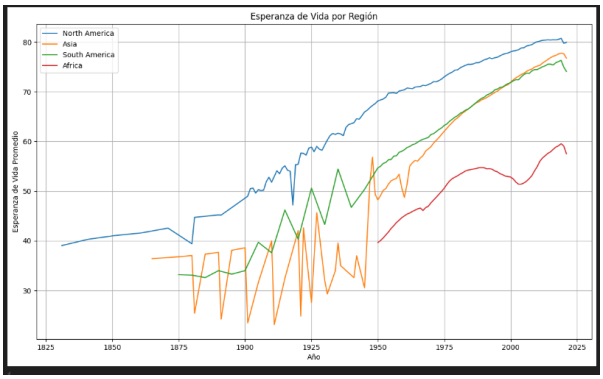
Los resultados preliminares sugieren que ciertos factores socioeconómicos y de salud pública tienen efectos variables en la esperanza de vida según el contexto nacional, destacando la necesidad de estrategias adaptadas y específicas para cada país. Estas conclusiones tienen implicaciones significativas para la formulación de políticas globales y locales que buscan mejorar la salud pública y el bienestar a nivel internacional.

Índice

Metodología

El primer pasó ha sido una exploración inicial de los datos por separado. Se han hecho distintos gráficos para tener un primer conocimiento de las características del conjunto de datos disponibles.





Todas las tablas comparten los siguientes atributos: “Entity” (nombre del país), “Code” (identificador del país), y “Year” (año del registro del dato).
Tras analizar los datos, vemos que cada registro pertenece a una “Entity” y “Year” distinto, por lo que construiremos las “primary key” alrededor de estos dos atributos. El atributo “Code” queda descartado, ya que múltiples “Entity” comparten un mismo “Code”.

Los distintos conjuntos de datos contienen registros empezando en distintos años, y todos finalizan en 2021. El posterior tratamiento de los datos acotará el alcance del análisis a un período definido del cual se dispongan datos suficientes para construir un análisis concluyente.

Con una visión más amplia del alcance de los datos y las primeras hipótesis empezando a formularse, necesitábamos filtrar la enorme cantidad de datos que teníamos para obtener los datos definitivos con los que íbamos a trabajar. Esto implicaba una nueva serie de pasos: analizar qué variables podían ser útiles, identificar y tratar los valores nulos de las tablas e identificar los outliers de cada tabla y pensar qué hacer con ellos.
Por otra parte, empezamos a documentar todo el proceso y ser conscientes de la necesidad de dejar por escrito todo lo que hacíamos para poder ir haciendo un registro del proceso.

Tratamiento de los datos

Combinar datos

Los datos de cada tabla muestran los valores de distintas variables socio-económicas por distintos países (“Entidad”) y años (“Año”).
Por lo tanto, podemos combinar todas las tablas mediante una columna primaria (“Primary Key”) que sea la combinación de estas dos columnas.

```
df_poverty["ID"] = df_poverty["Country"] + "_" + df_poverty["Year"].astype(str)

df_poverty.head(2)
```

Primary Key del dataset “Poverty”

De esta manera, obtenemos una tabla que contiene todos los datos posibles por cada país y año.

Agrupar países por Región Geográfica

Definiendo en una tabla separada a que continente pertenece cada país, podemos agrupar las variables de nuestra tabla principal por continentes, y así facilitar los posteriores análisis y conclusiones ya que pasamos de 261 entidades a 7 continentes.

- Asia: 52 entities
- Africa: 59 entities
- Europe: 54 entities
- Oceania: 24 entities
- South America: 16 entities
- North America: 42 entities
- Other: 14 entities

Dentro del continente “Other” quedan agrupados distintas entidades no pertenecientes a una región geográfica. Algunos ejemplos son:

‘High-income countries’, ‘Land-locked Developing Countries (LLDC)’, 'Least developed countries', 'Less developed regions', 'Less developed regions, excluding China'.

Agrupar variables por temática

Para el posterior análisis de correlaciones, agrupamos los distintos atributos por categoría, pudiendo así distinguir las siguientes. Esta agrupación ha sido definida estudiando los conceptos de los distintos atributos disponibles.

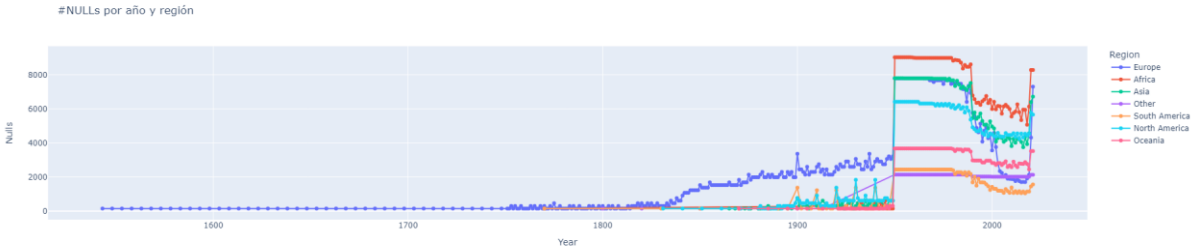
- Categorical
- Deaths
- Economic
- Consumption

- Income
- Vaccination

Valores nulos

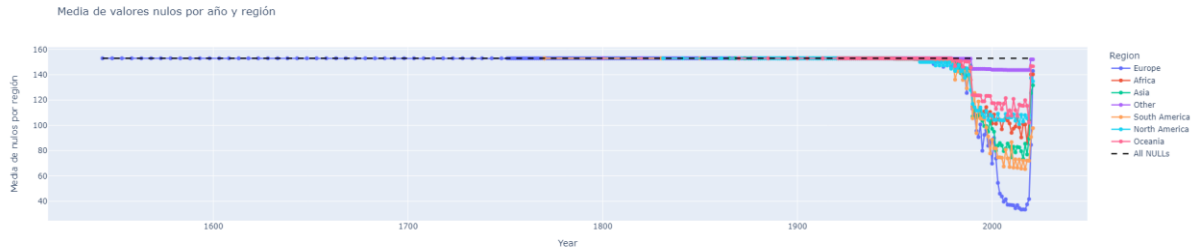
Analizando la cantidad de nulos en nuestro dataset final, vemos que la variable numérica con menor cantidad de nulos posee un 51% de nulos, lo que requiere un análisis y tratamiento de estos. Los porcentajes de nulos se pueden encontrar en la tabla '*Cantidad de nulos previos y después del tratamiento (%)*' de los anexos.

Si graficamos directamente los nulos por año y región, observamos mayores cantidades en aquellos años en que tenemos datos de "life expectancy", por lo estaríamos sesgando por aquellos años en que no disponemos la información de esta.



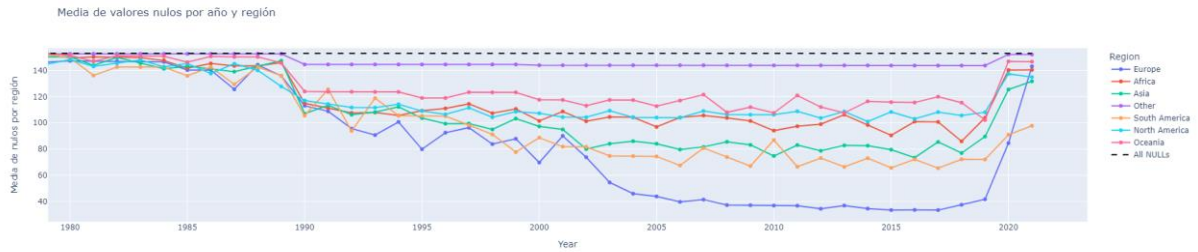
Cantidad de valores nulos por año y región (conjunto de datos sin modificar).

Para evitar este resultado, contamos no la cantidad de nulos sino la proporción de nulos media de cada región.
Al tener 153 variables numéricas, cuanto más cercano sea el valor de media de nulos a 153, mayor proporción de nulos tendremos para aquel año y región.



Media valores nulos por año y región (conjunto de datos sin modificar).

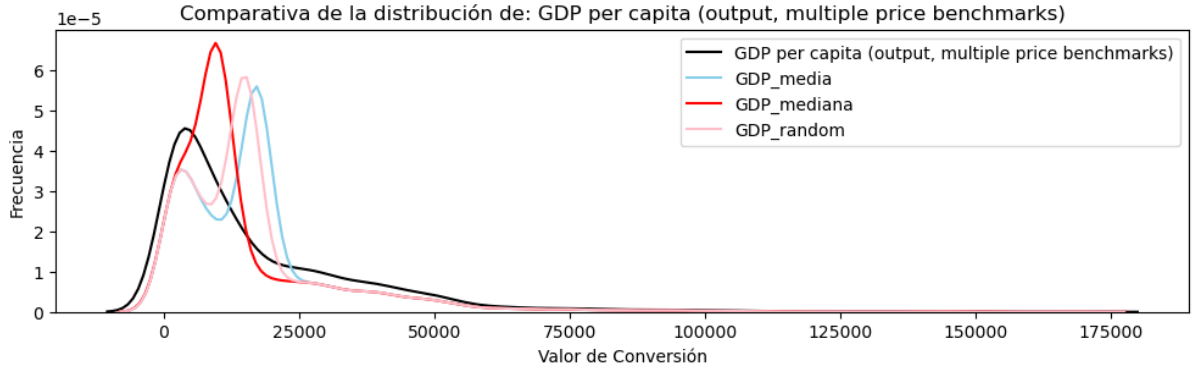
Como se puede ver en la gráfica anterior, no es hasta 1990 que la proporción de nulos se reduce.



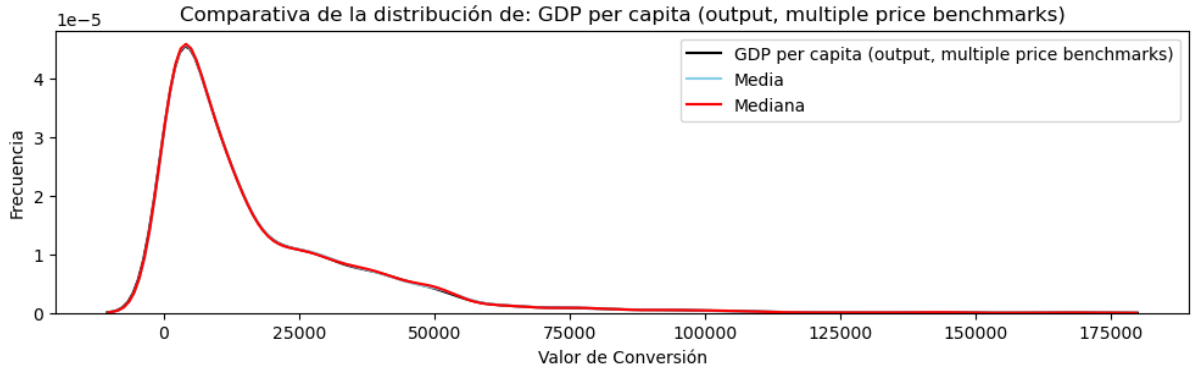
Media valores nulos por año y región a partir de 1990.

De la misma manera, la proporción de nulos para la región “Other” se mantiene prácticamente constante por encima de 140, lo que muestra la baja cantidad de información disponible para las entidades de esta región.
Es por estos dos motivos, que hemos decidido acotar el análisis de los datos al período 1990 - 2020 y excluir aquellas entidades presentes dentro de la región “Other”. De esta manera, reducimos la media de nulos del 82% al 60%.

Para tratar el resto de nulos, que representan un 60% de los datos disponibles, comparamos la distribución de estos imputando los valores nulos por la media, la mediana, y valores aleatorios.
Ejemplo con la variable “GDP per cápita”.



Ninguna de las tres imputaciones se parece a la distribución original, por lo que se estudia imputar los valores utilizando las medias, medianas, y valores aleatorios del país del que se tiene el nulo.

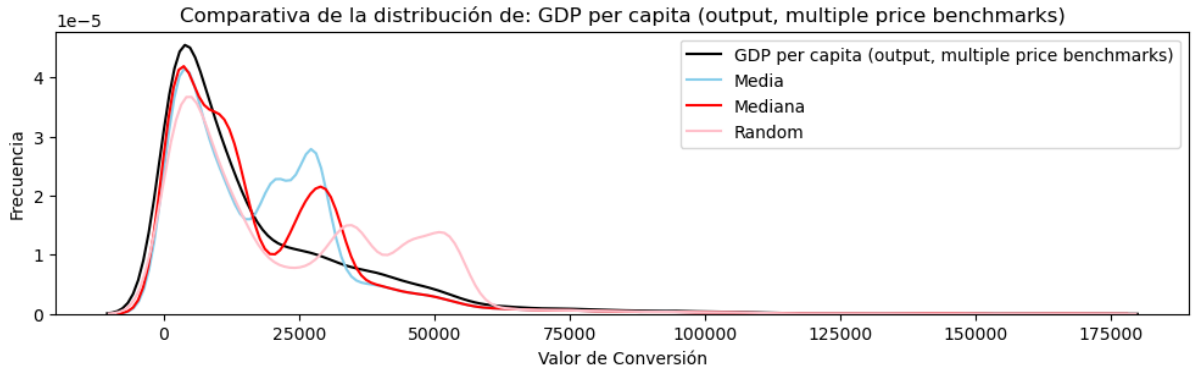


Vemos que cualquier imputación se parece a la distribución de los datos. Si nos fijamos en la cantidad de nulos por país para la GDP per cápita, observamos que en 61 países no tenemos ningún valor de GDP, por lo que no se está imputando los nulos.

Cantidad de nulos por país (%)	# Paises
6%	162
9%	3
13%	3
16%	3

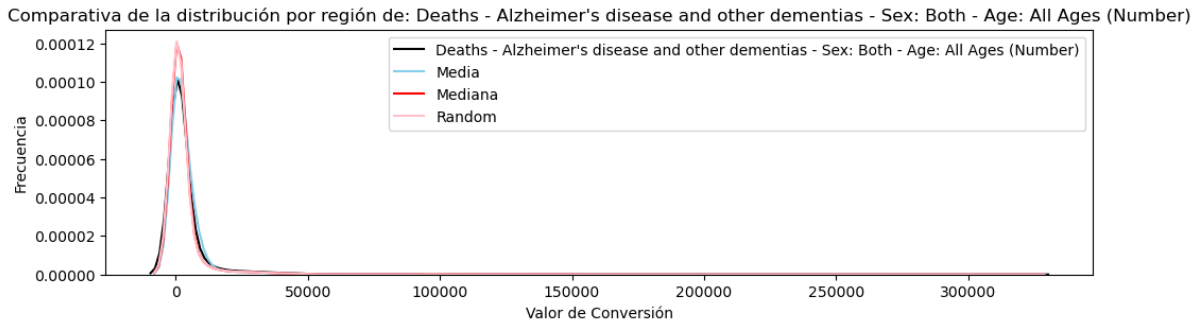
22%	2
25%	1
31%	2
34%	3
53%	2
59%	1
100%	61

Para evitar tener nulos no imputados, hacemos el análisis imputando los nulos por los valores medios, medianos, y aleatorios por región.



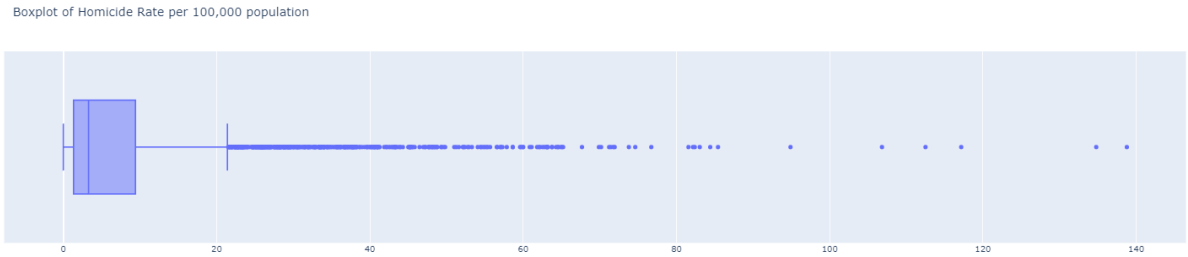
Vemos que la mediana es la mejor forma de imputar los nulos para la GDP.

Si hacemos el análisis para otra variable, por ejemplo, las muertes por Alzheimer, encontramos que la media es la mejor forma de imputar los nulos.



Outliers

Para tratar con los outliers, primero había que identificar qué valores de cada país y variable se correspondían con outliers. Para ello se hizo un primer intento para graficar la cantidad de outliers de cada tabla y luego sacar un df con esos valores:



Boxplot con los outliers de la tabla Homicide Rate.

Outliers del documento de Homicide Rate (Top 10)

```python
# Renombrar la columna para facilitar el acceso
homicide_rate.rename(columns={'Homicide rate per 100,000 population - Both sexes - All ages': 'Homicide_rate'}, inplace=True)

# Calcular el IQR (rango intercuartílico) para la tasa de homicidios
Q1 = homicide_rate['Homicide_rate'].quantile(0.25)
Q3 = homicide_rate['Homicide_rate'].quantile(0.75)
IQR = Q3 - Q1

# Definir los límites para los outliers
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

# Filtrar los outliers
outliers = homicide_rate[(homicide_rate['Homicide_rate'] < lower_bound) |
                         (homicide_rate['Homicide_rate'] > upper_bound)]

# Ordenar los outliers y seleccionar los diez últimos
last_ten_outliers = outliers.sort_values(by='Homicide_rate', ascending=False).tail(10)

# Mostrar los diez últimos outliers
last_ten_outliers
```

	Entity	Code	Year	Homicide_rate
4108	Venezuela	VEN	1996	21.978450
180	Antigua and Barbuda	ATG	2017	21.949320
4106	Venezuela	VEN	1994	21.870693
1093	Dominica	DMA	2010	21.816595
382	Bahamas	BHS	2007	21.808025
2757	Nigeria	NGA	2019	21.740790
3093	Puerto Rico	PRI	2005	21.678812
3105	Puerto Rico	PRI	2017	21.625423
1501	Greenland	GRL	1993	21.540892
3243	Saint Kitts and Nevis	KNA	2003	21.537567

Código para sacar en formato tabla los outliers del boxplot anterior.

Sin embargo, era necesario ser aún más directos con lo que se llegó a una solución: una función que funcionara con todas las tablas. De este modo, llamando la función con una tabla se podían ver, ordenados por país y año, todos los outliers de esa tabla:

```python
def obtener_outliers(df):
    # Obtener la lista de todos los países en el conjunto de datos
    paises = df['Entity'].unique()

    # Obtener las columnas que contienen datos de muertes
    columns_to_analyze = df.columns.difference(['Entity', 'Code', 'Year'])

    # Crear una lista para almacenar los outliers de cada país por cada causa
    outliers_list = []

    # Iterar sobre cada país
    for pais in paises:
        # Filtrar los datos para el país especificado y solo a partir del año 1990
        pais_data = df[(df['Entity'] == pais) & (df['Year'] >= 1990)]

        # Verificar si hay datos para el país especificado
        if pais_data.empty:
            print(f"No data found for {pais} from 1990 onwards.")
            continue

        # Iterar sobre cada columna (causa)
        for column in columns_to_analyze:
            # Calcular el IQR (rango intercuartílico) para la causa específica
            Q1 = pais_data[column].quantile(0.25)
            Q3 = pais_data[column].quantile(0.75)
            IQR = Q3 - Q1

            # Definir los límites para los outliers
            lower_bound = Q1 - 1.5 * IQR
            upper_bound = Q3 + 1.5 * IQR

            # Filtrar los outliers
            outliers = pais_data[(pais_data[column] < lower_bound) |
                                 (pais_data[column] > upper_bound)]

            if not outliers.empty:
                # Ordenar los outliers y seleccionar los diez últimos
                last_ten_outliers = outliers.sort_values(by=column, ascending=False).tail(10)
                for _, outlier in last_ten_outliers.iterrows():
                    outliers_list.append({
                        'pais': pais,
                        'variable': column,
                        'year': outlier['Year'],
                        'outlier_value': outlier[column]
                    })

    df_outliers = pd.DataFrame(outliers_list)
    df_outliers = df_outliers.sort_values(by=['pais', 'year'])

    return df_outliers
```

```
df_outliers_gvc = obtener_outliers(global_vaccionation_coverage)
df_outliers_gvc
```

	pais	variable	year	outlier_value
0	Afghanistan	PCV3 (% of one-year-olds immunized)	2014	49.0
1	Afghanistan	RotaC (% of one-year-olds immunized)	2018	45.0
11	Albania	Pol3 (% of one-year-olds immunized)	1990	89.0
4	Albania	BCG (% of one-year-olds immunized)	1991	80.0
5	Albania	DTP3 (% of one-year-olds immunized)	1991	78.0
...	...	...	...	...
2205	Zimbabwe	HepB3 (% of one-year-olds immunized)	1999	31.0
2210	Zimbabwe	Hib3 (% of one-year-olds immunized)	2008	75.0
2211	Zimbabwe	Hib3 (% of one-year-olds immunized)	2009	73.0
2212	Zimbabwe	PCV3 (% of one-year-olds immunized)	2012	21.0
2213	Zimbabwe	RotaC (% of one-year-olds immunized)	2014	48.0

2214 rows × 4 columns

Ejemplo de la función con la tabla de "global vaccination coverage"

Al guardar el resultado de la función en una variable, luego si se quería ordenar por cantidad de outliers en orden ascendente, era más fácil trabajar a partir de un df con todos los outliers ya dentro:

```
df_outliers_gvc.groupby(["pais", "variable"])["year"].count().reset_index().sort_values(by = "year")
```

	pais	variable	year
0	Afghanistan	PCV3 (% of one-year-olds immunized)	1
541	Marshall Islands	Pol3 (% of one-year-olds immunized)	1
542	Mauritania	HepB3 (% of one-year-olds immunized)	1
543	Mauritania	IPV1 (% of one-year-olds immunized)	1
546	Mauritania	Pol3 (% of one-year-olds immunized)	1
...	...	...	...
165	Central African Republic	Pol3 (% of one-year-olds immunized)	9
214	Cyprus	MCV1 (% of one-year-olds immunized)	10
813	Singapore	RCV1 (% of one-year-olds immunized)	10
811	Singapore	MCV1 (% of one-year-olds immunized)	10
215	Cyprus	RCV1 (% of one-year-olds immunized)	10

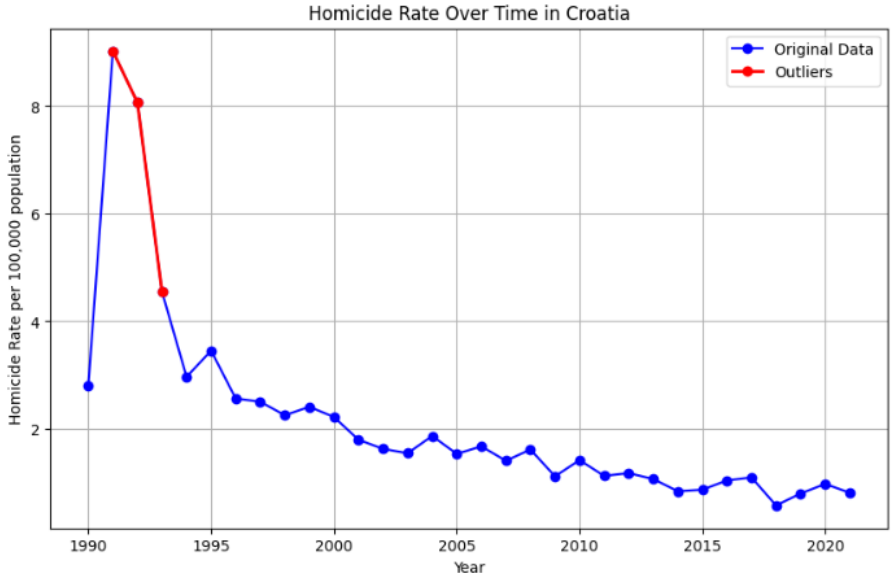
1003 rows × 3 columns

Se pueden encontrar más opciones de graficar los outliers en la sección de "Anexos". Estos intentos nos han servido para llegar a la función previamente mostrada.

Imputación de outliers

Lo normal en un proyecto como este sería imputar los outliers siguiendo la mediana de los valores ya existentes. Sin embargo, ir país por país por cada variable de cada tabla, aunque se hubiera hecho de una sola región, nos hubiera tomado un tiempo demasiado valioso. Por ello, debido a la falta de tiempo y de recursos humanos, nos hemos visto obligados a no imputar los outliers.

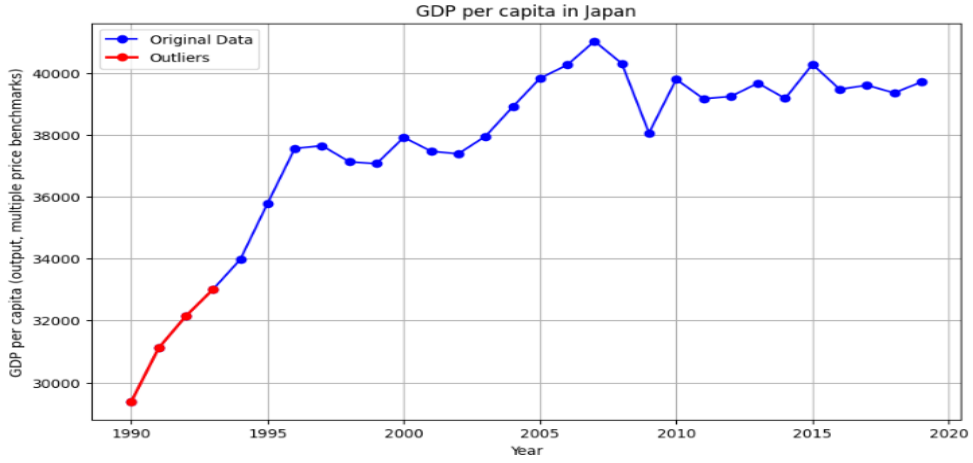
Asimismo, en caso de proceder con la imputación habría que tener en cuenta aquellos valores que según los gráficos aparecen como outliers pero que realmente no lo son. Entonces habría que ir valor por valor y ver cada caso específicamente.



Evolución de la ratio de homicidios en Croacia destacando los registros considerados como outliers

En este caso vemos que el segundo valor (1991) está muy alejado de lo que le correspondría, seguido del tercero (1992) y el cuarto valor (1993). Como se ha dicho antes, un análisis más exhaustivo podría identificar si hay una causa concreta de este pico o es un error de los datos.

Por contra, hay el caso opuesto de outliers que, según el gráfico, aparecen como tal y no parece que realmente lo sean:

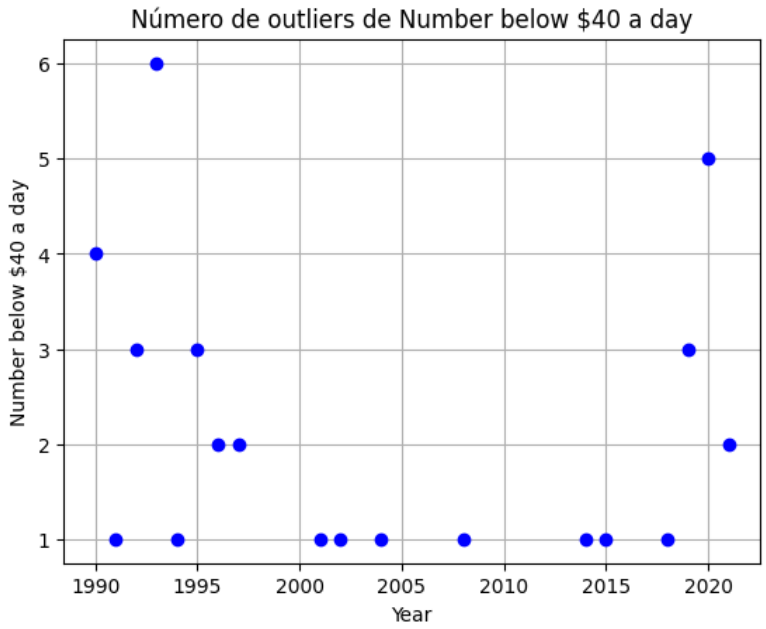


Evolución de la GDP per cápita en Japón destacando los registros considerados como outliers

En este gráfico vemos que el código marca como outliers los cuatro primeros valores. Pero viendo la progresión de los valores vemos que hay una progresión ascendente que termina en el valor correspondiente a 1996. Entonces, ¿son realmente outliers? Para responder se requeriría un análisis más exhaustivo que no ha entrado en el alcance del proyecto actual.

Con tal de dar los primeros pasos para este análisis se ha planteado la siguiente hipótesis. Si los outliers están cerca de los límites de "Year" [1990-2020], puede significar que el modelo está interpretando como outliers las tendencias crecientes o decrecientes de las variables socioeconómicas, como sucede en la gráfica anterior para la evolución de la GDP en Japón.

Nos disponemos a graficar en qué años encontramos los valores estadísticamente considerados como outliers.



Plot de los outliers de "Number below 40$ a day"

Para la variable "Number below $40 a day" se puede ver cómo efectivamente la mayoría de outliers están en estos extremos, por lo que haría falta mirar si estos valores son outliers reales.

Análisis de los datos

Todos los atributos en el conjunto de datos son numéricos, por lo que el análisis de estos se centrará en la correlación de atributos.

Análisis de correlación

Para el análisis de correlación, se ha escogido la correlación de Spearman.
La correlación de Spearman consiste en medir la correlación entre dos variables, pero, en lugar de comparar los valores reales de las variables, se compara el orden de sus valores. Esta medida está entre -1 y 1. Cuanto más cerca esté de estos extremos, más relación se detecta en los datos y cuanto más próxima a 0 menos evidencias hay.

Formalmente, la fórmula para la correlación de Spearman es la siguiente:

$$\rho(X,Y) = 1 - 6\Sigma {d_i}^2 / n(n^2 - 1)$$

donde:

- ρ es la correlación de Spearman
- d_i es la diferencia entre los rangos de los pares de datos, es decir $x\ index_i - y\ index_i$
- n es el número total de observaciones

La correlación de Spearman es robusta ante los outliers, ya que toma los índices en lugar de los valores reales de la variable. Por los motivos anteriormente mencionados, en el análisis actual no se han imputado los valores estadísticamente definidos como outliers, por lo que utilizar esta correlación protege el siguiente análisis de la influencia de estos.

El análisis está compuesto por los siguientes 5 apartados.

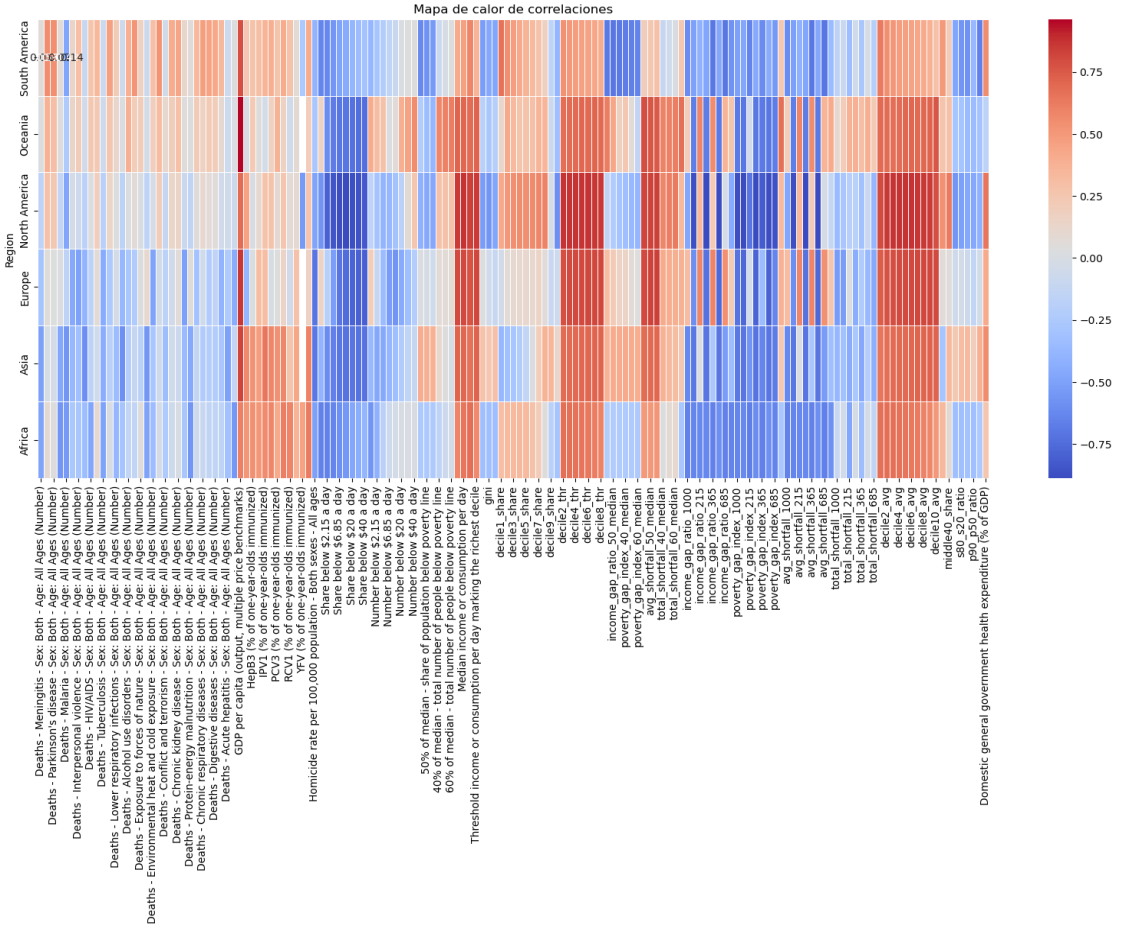
1. Análisis de correlación respecto a la “Life Expectancy” a nivel global.
2. Análisis de correlación respecto a las variables con mayor correlación a nivel global.
3. Análisis de correlación respecto a la “Life Expectancy” por región.
4. Análisis de correlación respecto a las variables con mayor correlación por región.
5. Hipótesis Healthcare Expenditure per cápita.

En los apartados 1 y 3 analizamos los atributos con mayor correlación con la esperanza de vida a nivel global y por región, respectivamente. A estas variables se les referirá como “atributos principales”.
En los apartados 2 y 4 analizamos los atributos con mayor correlación con respecto a las variables encontradas en los apartados anteriores. De esta manera, podemos entender el porqué de la influencia de los atributos principales en la esperanza de vida. A estas variables se les referirá como “atributos secundarios”.

Para evaluar cuáles son aquellos atributos principales y secundarios, se evaluará en cuántos países el valor de la correlación de cada atributo es superior a 0.95. De esta manera, centramos el análisis en aquellas variables que tienen una alta correlación en una mayor cantidad de países.

1. Análisis de correlación respecto la “Life Expectancy” a nivel global

Se toma una primera imagen de la correlación con la esperanza de vida a nivel global con el siguiente heatmap, en que se agrupan los valores por región.

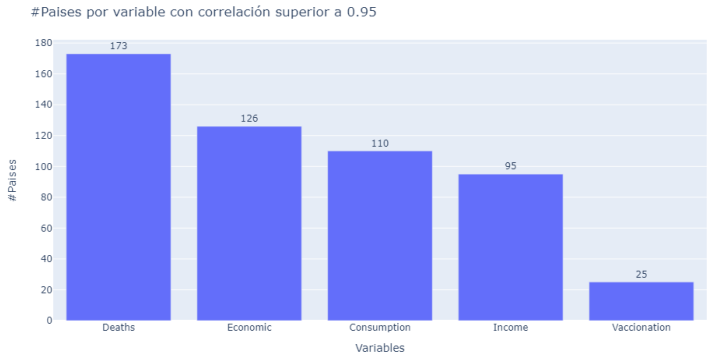


Agrupación por país de aquellas variables con correlación superior a ths=0.95

Existen un total de 6577 combinaciones de variable-país con correlación superior a 0.95 con respecto a la esperanza de vida. Graficamos en cuántos países cada variable tiene una correlación superior al ths para encontrar los atributos principales.

Análisis de los atributos con mayor correlación con la esperanza de vida (ths = 0.95)

Los atributos con mayor correlación son distintos tipos de muertes. Para confirmar que efectivamente es el grupo de atributos con mayor influencia, graficamos Agrupamos los atributos por grupo.



Análisis de los grupos de atributos con mayor correlación con la esperanza de vida (ths = 0.95)

Aquel grupo de variables con mayor correlación con la life expectancy es el de muertes.

Nos queda entender qué explica la influencia de las muertes, ya que esto nos puede llevar a otras variables que de forma indirecta estén influenciando a la esperanza de vida.
Nos disponemos a analizar las correlaciones de los atributos principales.

2. Análisis de atributos secundarios correlación respecto las variables con mayor correlación a nivel global

Las variables con correlación superior a 0.95 en mayor cantidad de países son:

1. “Deaths - Alzheimer's disease and other dementias - Sex: Both - Age: All Ages (Number)”
2. “Deaths - Parkinson's disease - Sex: Both - Age: All Ages (Number)”
3. “Deaths - Neoplasms - Sex: Both - Age: All Ages (Number)”
4. “Deaths - Chronic kidney disease - Sex: Both - Age: All Ages (Number)”
5. “Number below $40 a day”

También corresponden con las variables con un valor de correlación absoluto medio mayor, como se puede ver en los Apéndices (“Valores medios de la correlación con la “life expectancy” global”).

Para analizar los atributos secundarios, seguiremos los siguientes pasos:

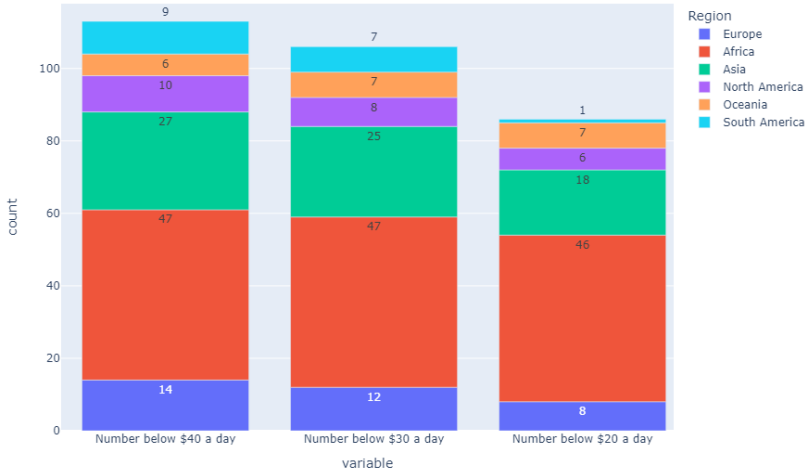
1. Definir variable a correlacionar: tomando como referencia las 5 variables con mayor correlación.
2. Definir variables con las que analizar la correlación:
    a. Eliminar variables categóricas: Entidad, Año, ID
    b. Eliminar variable “life expectancy”, al partir de la base de que tiene alta correlación con las variables anteriores.
    c. Hipótesis: analizar si queremos eliminar también del análisis aquellas variables del mismo grupo (i.e. si estamos buscando qué variables tienen mayor correlación con las muertes por Alzheimer, eliminar del análisis cualquier otra variable del grupo de muertes).
3. Calcular correlación: método de Spearman para eliminar la influencia de outliers en esta. Definido un threshold, filtrar aquellas variables con mayor correlación a este.
4. Graficar correlación por variable y región.

Para los pasos 2, 3, y 4 del proceso descrito, se han creado funciones que, dados los valores de entrada necesarios, permitan la elaboración de las gráficas correspondientes.

```
# 1. Definir variable a correlacionar
var_correlacion = variables_mayor_correlacion[0]

# 2. Definir amb quines variables correlacionar-la
variables = variables_a_correlazionar(df,var_correlacion, "no incluir")

# 3. Fer la correlacio i filtrar aquelles >ths
df_subcorr_interest = subcorrelaciones_analisis(df, variables, var_correlacion, 0.75)

# 4. Plot
subcorrelaciones_analisis_plot(df_subcorr_interest, "variable", "Region")
```

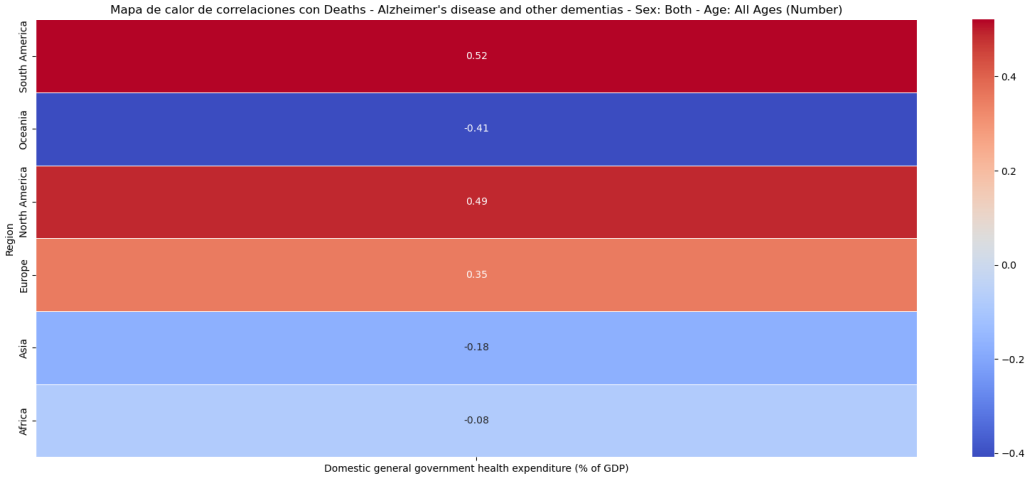
“Deaths - Alzheimer's disease and other dementias - Sex: Both - Age: All Ages (Number)”

Asumiendo un **ths = 0.95**, las tres variables con mayor influencia son las siguientes.



Análisis de correlación de las muertes por Alzheimer (ths = 0.95)

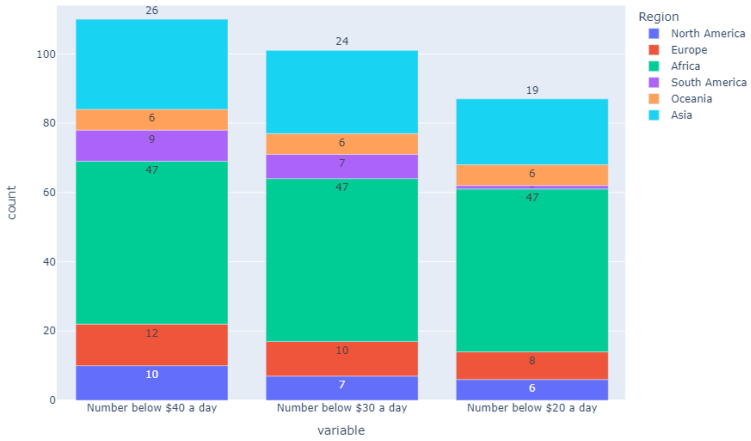
Aparece una nueva variable, la GDP per cápita. Al ser el Alzheimer una enfermedad que es tratada y acompañada en los hospitales, planteamos la hipótesis de si la HC Expenditure (attributo: “Domestic general government health expenditure (% of GDP)”) está detrás de esta influencia en las muertes por Alzheimer.
Fijándonos en exclusiva en esta variable, vemos que la influencia no es significativa.



Heatmap de la correlación de muertes por Alzheimer con la HC Expenditure por región

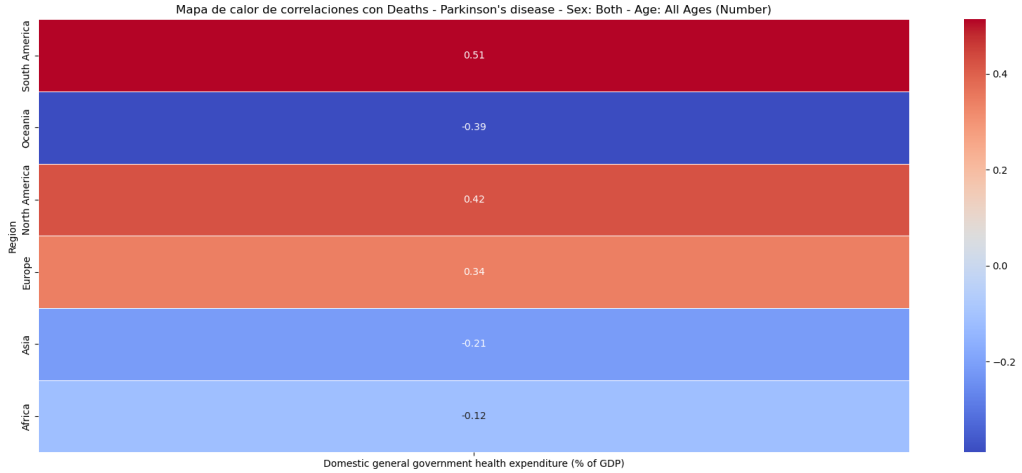
“Deaths - Parkinson's disease - Sex: Both - Age: All Ages (Number)”

Asumiendo un **ths = 0.95**, las tres variables con mayor influencia son las siguientes.



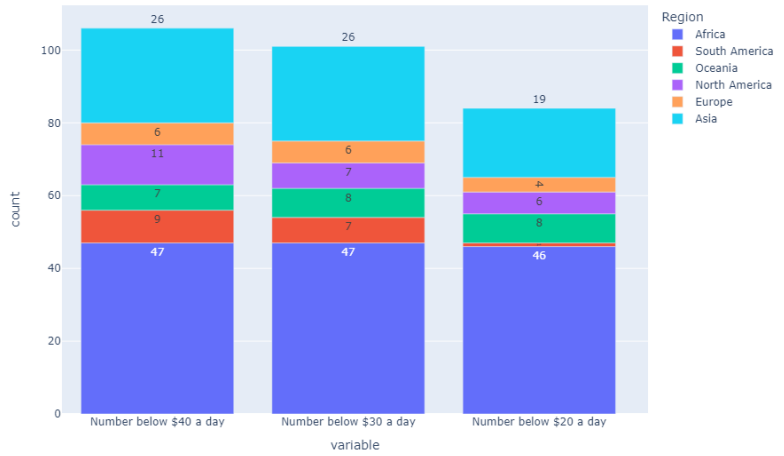
Análisis de correlación de las muertes por Parkinson (ths = 0.95)

Obtenemos los mismos resultados que para el Alzheimer. Al ser el Parkinson una enfermedad que es tratada y acompañada en los hospitales, planteamos de nuevo la hipótesis de si la “Domestic general government health expenditure (% of GDP)” está detrás de esta influencia en las muertes por Parkinson.
Fijándonos en exclusiva en esta variable, vemos que la influencia vuelve a no ser significativa.

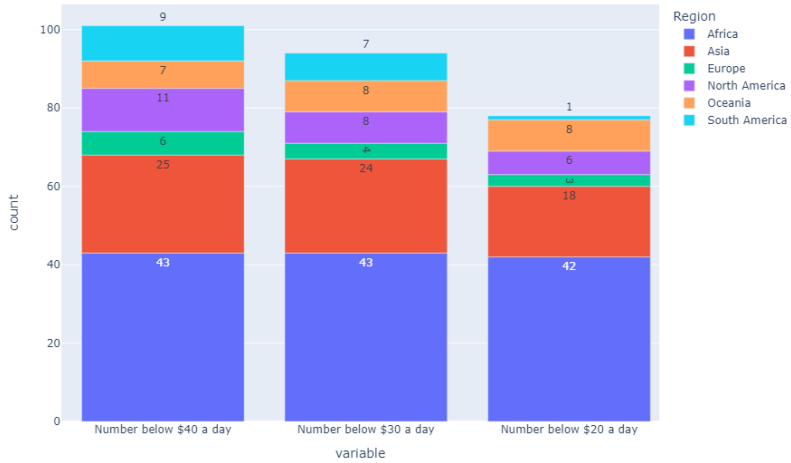


Heatmap de la correlación de muertes por Parkinson con la HC Expenditure por región

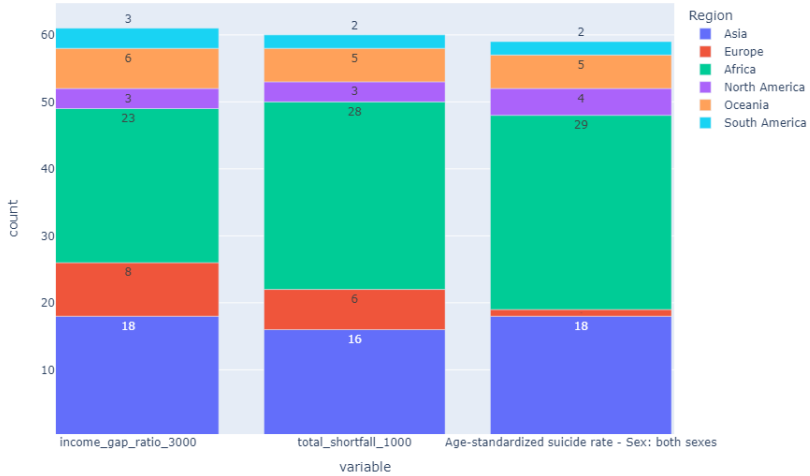
Analizando los siguientes atributos principales del TOP5, obtenemos los mismos atributos secundarios, lo que nos plantea la siguiente hipótesis: ¿obtendríamos distintos atributos principales y secundarios si estudiamos las correlaciones individualmente por región?
Los apartados 3 y 4 buscan dar respuesta a esta hipótesis.



Análisis de correlación de las muertes por "Neoplasms" (ths = 0.95)



Análisis de correlación de "Number below $40 a day" (ths = 0.95)



Análisis de correlación de las muertes por "Kidney disease" (ths = 0.95)

3. Análisis de correlación respecto la “Life Expectancy” por región

Resumen de los atributos principales por región:

	Top1	**Top2**	**Top3**	**Top4**	**Top5**
ALL regions	**Deaths Alzheimer**	**Deaths Parkinson**	**Deaths Neoplasms**	**Deaths Kidney disease**	**Number below $40 a day**
Europe (54)	Deaths Alzheimer (33)	Deaths Parkinson (30)	Deaths road injuries (28)	Deaths Kidney disease (25)	
North America (42)	Deaths Kidney disease (15)	Deaths Neoplasms (14)	Deaths Alzheimer (14)	Deaths Parkinson (12)	
South America (16)	Deaths Alzheimer (11)	Deaths Parkinson (10)	Deaths Neoplasms (10)	Deaths Kidney disease (10)	Deaths meningitis (10)
Asia (52)	Deaths Alzheimer (33)	Deaths Parkinson (33)	Deaths Neoplasms (33)	Deaths Kidney disease (29)	
Africa (59)	Number below $30 a day (39)	Number below $20 a day (38)	Number below $40 a day (38)	Deaths Alzheimer (37)	Total Shortfall (37)
Oceania (24)	Number below $30, $20 a day (8)	Total shortfall 2000, 3000, 4000, 60 median (8)	Decile avg, 2, 5, 6 thr (8)	P90 P10 ratio (8)	Deaths Neoplasms (8)

Tabla de atributos principales por región y cantidad de países en que la correlación es >0.95”

La tabla anterior permite encontrar particularidades dentro de las distintas regiones. Las principales conclusiones a destacar por región son:

- Europe, North America, South America, Asia: los atributos principales están alineados con los que encontramos a nivel global, relacionados con el grupo de muertes.

- Africa, Oceania: mayor influencia de atributos de consumo e ingresos por país que de mortalidad.

Ninguna de los atributos anteriores está presente en más del 70% de los países de la región con un valor correlación >0.95. Para el resto del análisis se ha mantenido este threshold para no añadir un sesgo a los datos. Una propuesta de mejora sería valorar distintos valores de ths para el análisis de atributos principales y secundarios.

4. Análisis de correlación respecto las variables con mayor correlación por región

Con tal de estudiar los atributos secundarios de las variables mencionadas en la anterior tabla, se han tomado tres atributos como referencia del análisis: mortalidad por Alzheimer, Parkinson, y “Kidney disease”, ya que son atributos principales para 4 de las 6 regiones.

En la siguiente tabla, se muestran, para las distintas regiones, los tres atributos con correlación superior al 0.95 presentes en la mayoría de países de esta, así como el porcentaje de países para cada atributo.

Atributos secundarios por región	Porcentaje de países en la región en los que existe una altra correlación con las variables		
	Alzheimer	Kidney	Parkinson
Africa			
Number below $20 a day	78%	71%	80%
Number below $30 a day	80%	73%	80%
Number below $40 a day	80%	73%	80%
Asia			
Number below $30 a day	48%	46%	46%
Number below $40 a day	52%	48%	50%
Europe			
GDP per cápita	35%	24%	39%
HC Expenditure (€/person)	35%	22%	30%
North America			
GDP per cápita	24%	21%	
Income & poverty gap			14%
Number below $10 a day	24%		
Number below $40 a day		26%	
Total shortfall 60 median			17%
South America			
Income & poverty gap	44%	44%	
Number below $30 a day			44%
Number below $40 a day	56%	56%	56%

La primera conclusión que extraemos es que para los tres atributos escogidos: muertes por alzheimer, parkinson, y kidney disease; obtenemos los mismos atributos con mayor correlación. De nuevo, esto puede indicar que el threshold escogido está limitando los resultados del análisis y camuflando otras variables.

Si analizamos las diferencias entre regiones, observamos la alta influencia de la riqueza del país y el gasto sanitario per cápita en Europa y Norte América.
Para regiones con mayores diferencias económicas y altas diferencias de riqueza, como África, South América, o Asia, vemos que la economía individual es el principal factor que

explica las mortalidades anteriores, representando los distintos “Number below $ a day” la cantidad de dinero diario con el que vive la población. Cabe destacar que esta variable también se hace presente en Norte América, donde la sanidad es principalmente privada y no es accesible a toda la población.

5. Hipótesis Healthcare Expenditure per cápita

Las variables con mayor influencia en la esperanza de vida están relacionadas con distintos tipos de mortalidad prematura. En el TOP5 global encontramos la mortalidad por Alzheimer, Parkinson, neoplasmas, y enfermedades renales. Todas ellas requieren de tratamiento y la esperanza y calidad de vida del paciente puede mejorar si se tienen los recursos. Por este motivo planteamos la hipótesis de que la "Healthcare Expenditure" está influyendo principalmente estas mortalidades y consecuentemente en marcando la esperanza de vida del país.

En el apartado anterior se muestra que la "HC Expenditure (%)" no aparece como variable con alta correlación para la mortalidad por Alzheimer. Lo mismo sucede con el resto de variables en el TOP5. Sí aparece la "GDP per cápita". Utilizando estas dos variables, podemos calcular el la "HC Expenditure per cápita", lo que nos dará un valor absoluto de la cantidad de dinero que cada país destina al cuidado de la salud por ciudadano.

Cabe mencionar que se tienen datos de "Healthcare Expenditure" disponibles solo a partir del año 2000, por lo que faltan valores para los diez primeros años del análisis. Por los motivos anteriormente mencionados, se ha pospuesto la imputación de estos nulos a futuro. El cálculo de esta nueva variable es el siguiente:

$$HC\ Expenditure_{per\ capita} = GDP_{per\ capita} \cdot HC\ Expenditure_{\%GDP}$$

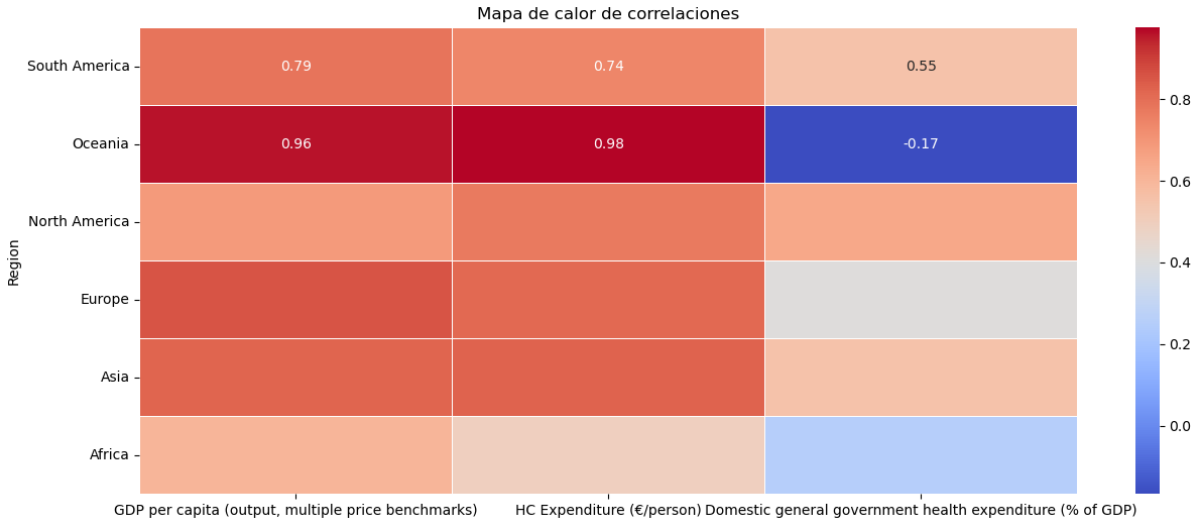
Analizamos la correlación de las tres variables siguientes con respecto a la "Life Expectancy". En la tabla siguiente, se puede ver las características de la correlación tratando el valor absoluto de los datos, y así convertir el intervalo [-1,1] a [0,1].

Attribute	mean	std	Q0	Q25	Q50	Q75	Q100
HC Expenditure (%GDP)	0.590	0.267	0.023	0.361	0.650	0.808	0.998
GDP per cápita	0.787	0.25	0.026	0.759	0.875	0.959	1
HC Expenditure per cápita	0.769	0.246	0	0.652	0.871	0.958	1

Correlación de "HC Expenditure" con "Life Expectancy".

Podemos ver como la HC Expenditure per cápita y la GDP tienen una mayor correlación con esta que la HC Expenditure (%).

La "HC Expenditure per cápita" pasa a ser la variable con mayor influencia en mayor cantidad de países #60, teniendo una correlación >0.95 en 48 países.
La "GDP per cápita" ocupa la posición #50 teniendo una correlación >0.95 en 53 países.



Para comprobar si la HC Expenditure per cápita también aparece como atributo secundario, tomamos como referencia las muertes por Alzheimer. La tabla siguiente muestra los atributos secundarios con respecto al Alzheimer por región.

Europe (54)	GDP: 19 países **HC Expenditure: 19 países** Decile: 17 países
NA (42)	Number below $10 a day: 10 países GDP: 10 países **HC Expenditure: 8 países**
SA (16)	Number below $40 a day: 9 países Income & poverty gap: 7 países **HC Expenditure: 7 países**
Asia (52)	Number below $40 a day: 27 países Number below $30 a day: 25 países
Africa (59)	Number below $30 a day: 47 países Number below $40 a day: 47 países Number below $20 a day: 46 países

Destacamos tres comportamientos distintos:

- Europe & North America: tanto la GDP como la HC Expenditure per cápita influyen en la mortalidad por Alzheimer.
- South America: la HC Expenditure per cápita está en el TOP3 de atributos secundarios, pero ni la GDP per cápita ni la HC Expenditure (%) lo están.
- Asia & Africa: ninguna de las tres variables aparece en el TOP3 de atributos secundarios.

La siguiente tabla compara por región en cuántos países las distintas variables son atributos secundarios por lo que respecta la mortalidad por Alzheimer.

Attribute	Europe (54)	North A. (42)	South A. (16)	Asia (52)	Africa (59)
HC Expenditure (%GDP)	2	3	1	4	4
GDP per cápita	19	10	4	16	10
HC Expenditure per cápita	19	8	7	15	6

Modelo predictivo

Random Forest

El objetivo es identificar qué características se asocian con una esperanza de vida superior al percentil 90 (Life Expectancy > q90). Esta es una tarea de clasificación binaria en la que estamos interesados en etiquetar las variables según si la esperanza de vida de un país está por encima de este umbral (q90) o no.

Aunque la esperanza de vida es una variable continua (lo que lo convierte en un problema de regresión en su forma más básica), hemos transformado el problema en una tarea de clasificación. Un modelo de clasificación, como Random Forest, nos permite enfocarnos específicamente en la identificación de países con esperanzas de vida muy altas (por encima del percentil 90), lo cual es útil para entender qué factores contribuyen a que un país pertenezca a este grupo. En lugar de predecir un valor exacto de esperanza de vida, estamos más interesados en comprender los patrones que distinguen a este grupo de países.

Ventajas del Random Forest en Clasificación

-Robustez a Outliers: El modelo Random Forest es robusto frente a outliers, lo cual es beneficioso en problemas donde algunos países pueden tener características extremas que podrían sesgar los resultados en un modelo de regresión.

-Importancia de Características: Random Forest proporciona una medición de la importancia de las características, lo que nos ayuda a identificar cuáles son las variables más influyentes en determinar si un país tiene una esperanza de vida superior al percentil 90.

```python
# Unificar las tablas en una sola utilizando "Entity", "Year" como claves
merged_df = dataframes['life_expectancy']

# Uniendo las demás tablas
for name, df in dataframes.items():
    if name != 'life_expectancy':
        # Eliminar o renombrar columnas duplicadas antes de fusionar
        columns_to_remove = [col for col in df.columns if col in merged_df.columns and col not in ["Entity", "Year"]]
        df = df.drop(columns=columns_to_remove)

        # Realizar la fusión
        merged_df = pd.merge(merged_df, df, on=["Entity", "Year"], how="left")

# Renombrar la columna de esperanza de vida para mayor claridad
merged_df = merged_df.rename(columns={"Period life expectancy at birth - Sex: all - Age: 0": "Life Expectancy"})

# Calcular el percentil 90 de la esperanza de vida
q90 = merged_df["Life Expectancy"].quantile(0.90)

# Crear la variable objetivo
merged_df['Life Expectancy > q90'] = (merged_df["Life Expectancy"] > q90).astype(int)

# Eliminar columnas no deseadas
columns_to_drop = ["Year", "Entity", "Code", "Life Expectancy"]
X = merged_df.drop(columns=columns_to_drop + ['Life Expectancy > q90'], errors='ignore')
y = merged_df['Life Expectancy > q90']

# Verificamos las primeras filas del dataframe preparado
X.head(), y.head(), q90
```

Código para preparar el Random Forest

Tabla de la matriz de confusión de todos los modelos

Modelo	Accuracy	Precision_0	Recall_0	F1_0	Precision_1	Recall_1	F1_1
RF-q90-median	0,92	0,92	0,99	0,95	0,80	0,33	0,46
RF-q90-mean	0,92	0,92	0,99	0,95	0,81	0,32	0,46
RF-q90-noImputacion	0,92	0,92	0,99	0,96	0,86	0,33	0,48
RF-q90-median-MaxFeat75	0,92	0,92	0,99	0,96	0,89	0,33	0,48
RF-q90-median-MaxFeat2	0,91	0,92	0,99	0,95	0,79	0,30	0,44
RF-q10-median	0,94	0,95	0,99	0,97	0,83	0,39	0,53
RF-q10-mean	0,94	0,94	0,99	0,97	0,80	0,42	0,55
RF-q10-noImputacion	0,95	0,95	0,99	0,97	0,88	0,50	0,64

A la hora de hacer los modelos nos planteamos dos hipótesis:

- ¿Imputar o no los nulos afecta los resultados?
- ¿El parámetro max_features afecta los resultados?

Para comprobar la primera hipótesis hicimos tres modelos tanto del percentil Q90 como del percentil Q10 imputando por la mediana, la media y sin imputarlos y dos modelos del Q90 con usando dos valores extremos del parámetro max_features, 2 y 75.

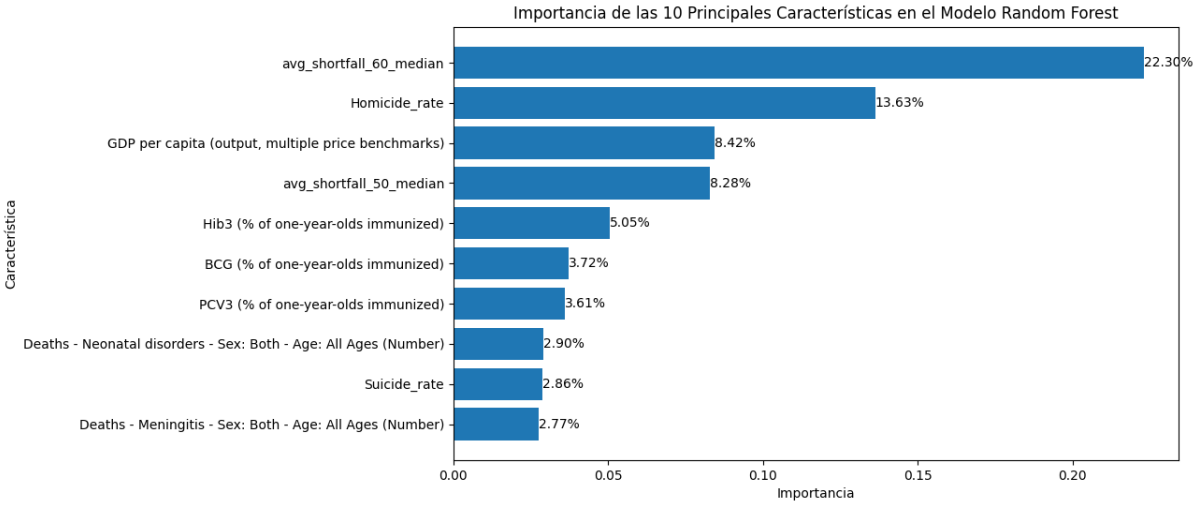
Los resultados obtenidos se pueden leer detalladamente en el apartado de '*Modelo Predictivo'*.

Tabla con los valores con más peso de los modelos

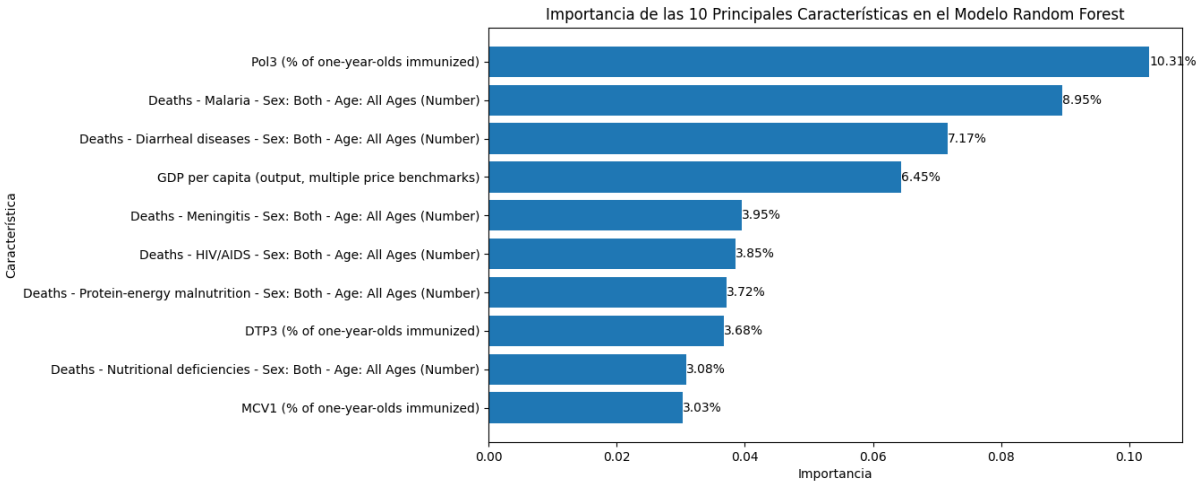
Top3 Features	rf_q90_median	rf_q90_mean	rf_q90_NoImputacion	rf_q90_median_MaxFeat75	rf_q90_median_MaxFeat2	rf_q10_median	rf_q10_mean	rf_q10_NoImputacion
Homicide Rate	6.62%	6.52%	6.23%	13.63%	"	"	"	"
avg_shortfall_60_median	6.37%	6.13%	6.03%	22.30%	"	"	"	"
decile8_thr	5.85%	5.82%	5.32%	"	4.19%	"	"	"
GDP per cápita (output, multiple price benchmarks)	"	"	"	8.42%	"	"	5.97%	"
avg_shortfall_40_median	"	"	"	"	2.77%	"	"	"
avg_shortfall_4000	"	"	"	"	2.65%	"	"	"
Pol3 (% of one-year-olds immunized)	"	"	"	"	"	9.98%	10.72%	10.31%
Deaths - Malaria - Sex: Both - Age: All Ages (Number)	"	"	"	"	"	9.44%	10.37%	8.95%
Deaths - Diarrheal diseases - Sex: Both - Age: All Ages (Number)	"	"	"	"	"	6.82%	"	7.17%

Una vez obtenidas las principales características, quisimos ver qué peso tenía cada una de ellas respecto del total. Por ello sacamos estos gráficos de barras para tener una visión más amplia.

Muestra de la importancia de todas las variables para el modelo Q90 con max_features = 75:



Muestra de la importancia de todas las variables para el modelo Q10 sin imputar:



Conclusiones

Correlación

Alta cantidad de valores nulos. Mejorar recolección de datos en las variables con mayor correlación.
Hace falta un estudio detallado de cada país y variable para entender si los outliers estadísticos son efectivamente outliers.
Es necesario separar el estudio de la esperanza de vida por regiones para obtener conclusiones específicas de cada región. Aquellas variables con mayor correlación a nivel global se mantienen constantes para el resto de regiones, pero para cada región aparecen focos distintos que explican la influencia de estas en la esperanza de vida. El resumen de sugerencias por región se puede encontrar en *'Análisis de atributos secundarios correlación respecto las variables con mayor correlación a nivel global'*.

Se ha tomado como ejemplo la influencia de las muertes por Alzheimer en la esperanza de vida, ya que es una variable con alta correlación para la mayoría de países y regiones. Aquellas variables con mayor correlación con la mortalidad por Alzheimer varían según la región que se estudie, como queda resumido en el apartado '*Análisis de correlación respecto las variables con mayor correlación* '.

Vemos la gran influencia de la capacidad de generar riqueza de los países en regiones como Europa o NorteAmérica, mientras que en el resto de regiones es la pobreza o diferencia de riqueza de la población lo que influye más en la esperanza de vida.

La hipótesis de que la influencia de la variable GDP es en realidad causada por la HC Expenditure ha quedado negada para las regiones de Asia y África, mientras que para el resto de regiones tiene una influencia directa cuando se mide el valor por cápita en variables como la mortalidad por Alzheimer, las cuales influyen directamente en la esperanza de vida.

Modelo Predictivo

Las matrices de confusión de los 8 modelos dan resultados similares para todos ellos, por lo que podemos entender que los distintos modelos tienen capacidades similares de predecir si los datos están en el P90 o P10.

En los tres modelos de Q90 y Q10 en que variamos la imputación de los datos de entreno y validación aparecen las mismas variables con pesos similares dentro de los modelos. Concluímos entonces que la imputación o no de los datos no es un factor determinante en las características del modelo entrenado.

Sin embargo, en los modelos con los parámetros de "max_features" sí que se aprecian valores destacables. Poniendo un valor de "max_features=75" se obtienen unas características con porcentajes relativamente altos comparándolos con los otros modelos. A su vez, ese modelo tiene los valores más altos de precisión, con lo que indicaría que, al permitir que los árboles tomen más características, también sube el riesgo de sobreajuste.

En cambio, el modelo con el parámetro de “max_features=2” no solo tiene unos porcentajes más bajos que el resto de modelos, pero unos valores de precisión más bajos que los demás, si no que aparecen variables nuevas no presentes en los otros modelos. Concluimos entonces que valores extremos de max features sí que afectan a las características del modelo, aunque no obtengamos valores distintos para la matriz de confusión.

Un motivo que puede explicar por qué no observamos diferencias en las distintas matrices de confusión, aun teniendo modelos como maxfeatures75 o maxfeatures2 que dan pesos y variables distintos, puede ser el tamaño del conjunto de datos utilizado para entrenar el modelo.
Hay que tener en cuenta que, al decidir utilizar un modelo de clasificación basado en el P10 y P90 de los datos, estamos limitando el número de positivos al 10% del conjunto de datos, que representan 823 filas, y se están clasificando como negativos 7401. Basado en los resultados de los modelos, dejamos abierta la cuestión de si faltan datos para entrenar los modelos propuestos.

Finalmente, se observan dos tendencias entre todos los modelos. Todos los modelos de Q90 tienen, en su mayoría, factores económicos y la tasa de homicidios en el top 10. Por tanto, se asocia que el nivel económico de un país está asociado a una esperanza de vida superior y que una tasa de homicidios baja aporta más seguridad para la población.
En cambio, los modelos de Q10 tienen en su mayoría factores de salud. Ello significa que aquellos países con un nivel de sanidad y vacunación deficitarios tienen una esperanza de vida más baja.

Propuestas de mejora

1. Recolección de datos: reducir la cantidad de nulos en los atributos principales y secundarios.

Pese al filtrado de datos por año y región geográfica, aún tenemos porcentajes grandes de nulos en los atributos con mayores correlaciones. Una primera propuesta para los países es la de mejorar la recolección de datos en las variables que más influencian la esperanza de vida en el país o región. El dashboard es una herramienta que permite identificar estas variables por país.

Attributes	After treatment (%)	Before treatment (%)
Deaths - Alzheimer's disease and other dementias - Sex: Both - Age: All Ages (Number)	21,3	70,37
Deaths - Chronic kidney disease - Sex: Both - Age: All Ages (Number)	21,3	70,37
Deaths - Neoplasms - Sex: Both - Age: All Ages (Number)	21,3	70,37
Deaths - Parkinson's disease - Sex: Both - Age: All Ages (Number)	21,3	70,37
GDP per cápita (output, multiple price benchmarks)	31,38	51,3
Number below $30 a day	75,08	89,68
Number below $40 a day	75,08	89,68

2. Análisis de distintos ths. Tomar como ths el valor medio de la correlación de los 10 atributos principales (ths = 0.85).

Para el análisis de correlación se ha tomado como ths que la correlación entre los atributos y el atributo a correlacionar sea >0.95. La media de correlación para los 20 atributos principales es de 0.85, por lo que estudiar qué variables aparecen en mayor cantidad de países a este ths podría dar resultados distintos para el análisis.

3. Imputación de valores nulos.

Evaluar para cada variable por qué valor es mejor imputar los nulos de cada país: media país, mediana país, valor aleatorio país, media región, mediana región, valor aleatorio región, media global, mediana global, valor aleatorio global. Crear función que evalúa las distintas distribuciones y tomé aquella que más se parezca al conjunto de datos original.

4. Imputación de valores outliers.

Evaluar para cada conjunto de valores atributo-país, si los datos considerados estadísticamente outliers son en realidad outliers o máximos/mínimos relativos al período del que se toman los datos que forman parte de la progresión de estos. Propuesta de función que detecte aquellos valores outliers en aquellos años cercanos a los límites del análisis: 1999 & 2020. Estos valores son potencialmente los inicios y finales de tendencias crecientes o decrecientes en los distintos países.
Otra posibilidad es la de adaptar el IQR para que el cálculo de outliers muestre valores más precisos.

5. Creación de un modelo de regresión complementario al análisis de correlación.

El modelo creado transforma un problema de regresión a uno de clasificación. Para encontrar el efecto conjunto de las distintas variables en la esperanza de vida, un modelo de regresión hubiera sido más adecuado.

Annexos

Cantidad de nulos previos y después del tratamiento (%)

Attribute	%NULLs after	%NULLs before
Entity	0	0
Year	0	0
Region	0	0
LifeExpectancy	0	0
ID	0	0
GDP per cápita (output, multiple price benchmarks)	31,38	51,3
DTP3 (% of one-year-olds immunized)	21,42	63,37
Pol3 (% of one-year-olds immunized)	21,48	63,37
MCV1 (% of one-year-olds immunized)	21,53	63,98
BCG (% of one-year-olds immunized)	35,53	69,91
Deaths - Malaria - Sex: Both - Age: All Ages (Number)	21,3	70,37
Deaths - Parkinson's disease - Sex: Both - Age: All Ages (Number)	21,3	70,37
Deaths - Diarrheal diseases - Sex: Both - Age: All Ages (Number)	21,3	70,37

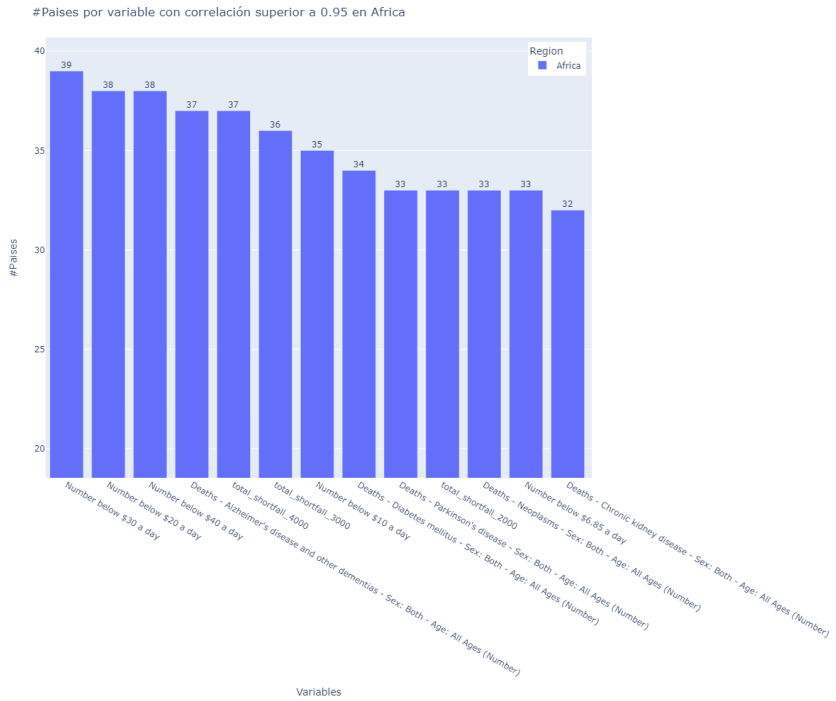
Valores medios de la correlación con la “life expectancy” global

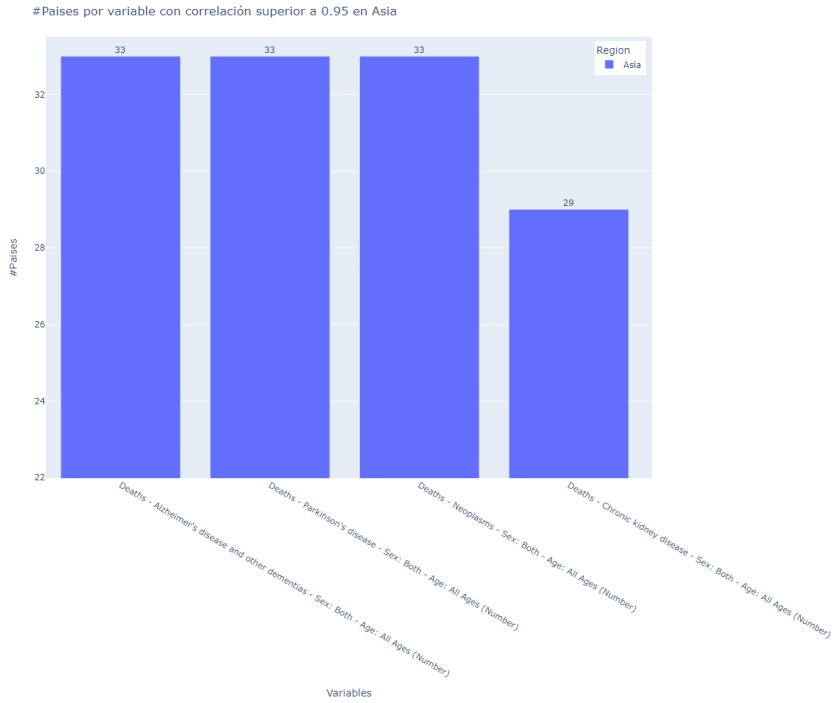
Attribute	Atributos_grupo	value_abs
Deaths - Alzheimer's disease and other dementias - Sex: Both - Age: All Ages (Number)	Deaths	0,940
Deaths - Parkinson's disease - Sex: Both - Age: All Ages (Number)	Deaths	0,929
Deaths - Neoplasms - Sex: Both - Age: All Ages (Number)	Deaths	0,916
Deaths - Chronic kidney disease - Sex: Both - Age: All Ages (Number)	Deaths	0,916
Number below $30 a day	Consumption	0,886
Number below $40 a day	Consumption	0,880
GDP per cápita (output, multiple price benchmarks)	Economic	0,878
HC Expenditure (€/person)	Economic	0,875
decile6_thr	Income	0,854
decile7_avg	Income	0,854
decile4_avg	Income	0,854
Deaths - Diarrheal diseases - Sex: Both - Age: All Ages (Number)	Deaths	0,851
decile6_avg	Income	0,847
Deaths - Meningitis - Sex: Both - Age: All Ages (Number)	Deaths	0,845
Deaths - Digestive diseases - Sex: Both - Age: All Ages (Number)	Deaths	0,841
decile3_avg	Income	0,841
decile5_avg	Income	0,841
Deaths - Diabetes mellitus - Sex: Both - Age: All Ages	Deaths	0,837

(Number)		
decile4_thr	Income	0,835
decile3_thr	Income	0,835
decile5_thr	Income	0,835
poverty_gap_index_3000	Income	0,835
decile8_thr	Income	0,829
Median income or consumption per day	Consumption	0,829
decile8_avg	Income	0,828
avg_shortfall_60_median	Economic	0,828
decile9_avg	Income	0,827
Deaths - Road injuries - Sex: Both - Age: All Ages (Number)	Deaths	0,825
decile7_thr	Income	0,823
Deaths - Protein-energy malnutrition - Sex: Both - Age: All Ages (Number)	Deaths	0,822
Mean income or consumption per day	Consumption	0,822
decile2_thr	Income	0,822
Deaths - Tuberculosis - Sex: Both - Age: All Ages (Number)	Deaths	0,820
Deaths - Cardiovascular diseases - Sex: Both - Age: All Ages (Number)	Deaths	0,818
Deaths - Neonatal disorders - Sex: Both - Age: All Ages (Number)	Deaths	0,817
Deaths - Alcohol use disorders - Sex: Both - Age: All Ages (Number)	Deaths	0,817
poverty_gap_index_2000	Income	0,816

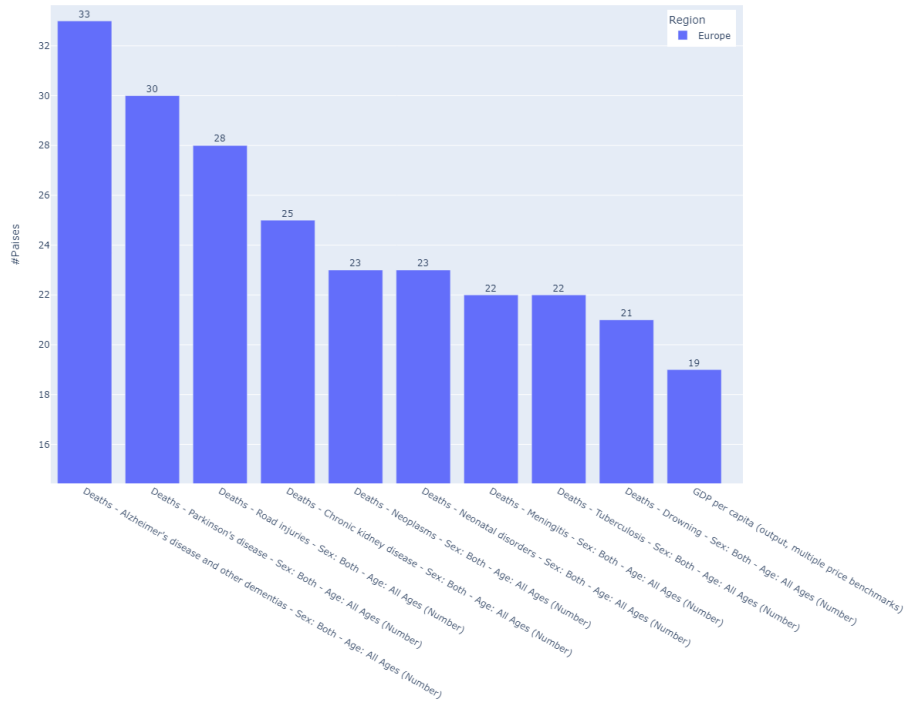
poverty_gap_index_4000	Income	0,816
Deaths - Cirrhosis and other chronic liver diseases - Sex: Both - Age: All Ages (Number)	Deaths	0,814
Number below $10 a day	Consumption	0,810
income_gap_ratio_3000	Income	0,810
avg_shortfall_3000	Economic	0,810
decile2_avg	Income	0,809
Threshold income or consumption per day marking the richest decile	Consumption	0,809
Threshold income or consumption per day marking the poorest decile	Consumption	0,809
total_shortfall_3000	Economic	0,804
Number below $20 a day	Consumption	0,804
Deaths - Chronic respiratory diseases - Sex: Both - Age: All Ages (Number)	Deaths	0,803
total_shortfall_60_median	Economic	0,791
total_shortfall_4000	Economic	0,785
Domestic general government health expenditure (% of GDP)	Economic	0,656

Análisis de correlación con “life expectancy” por región

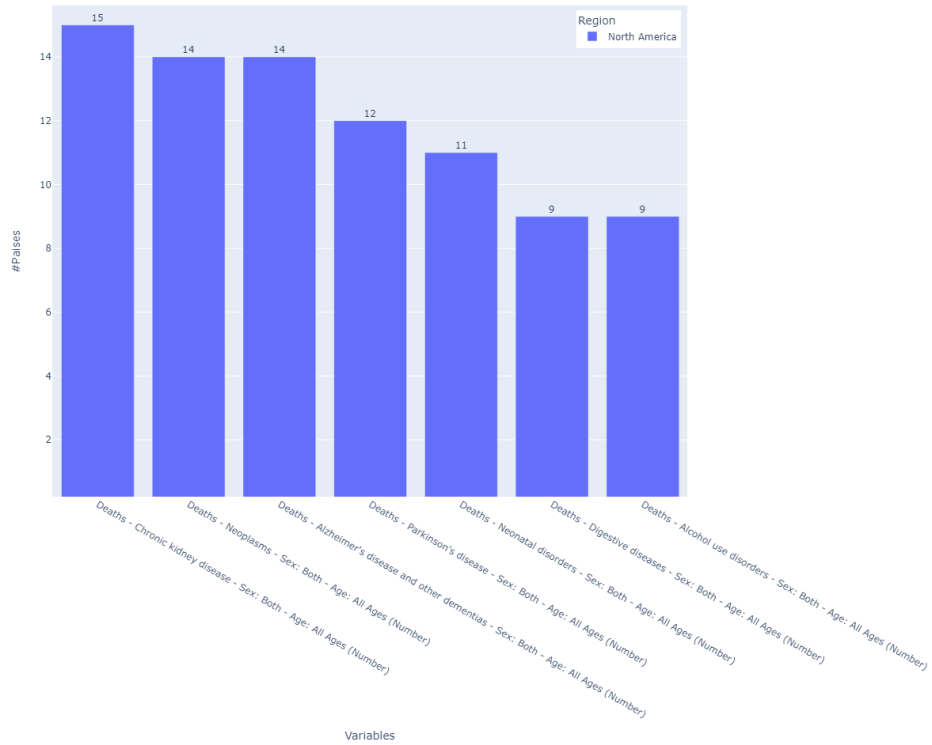


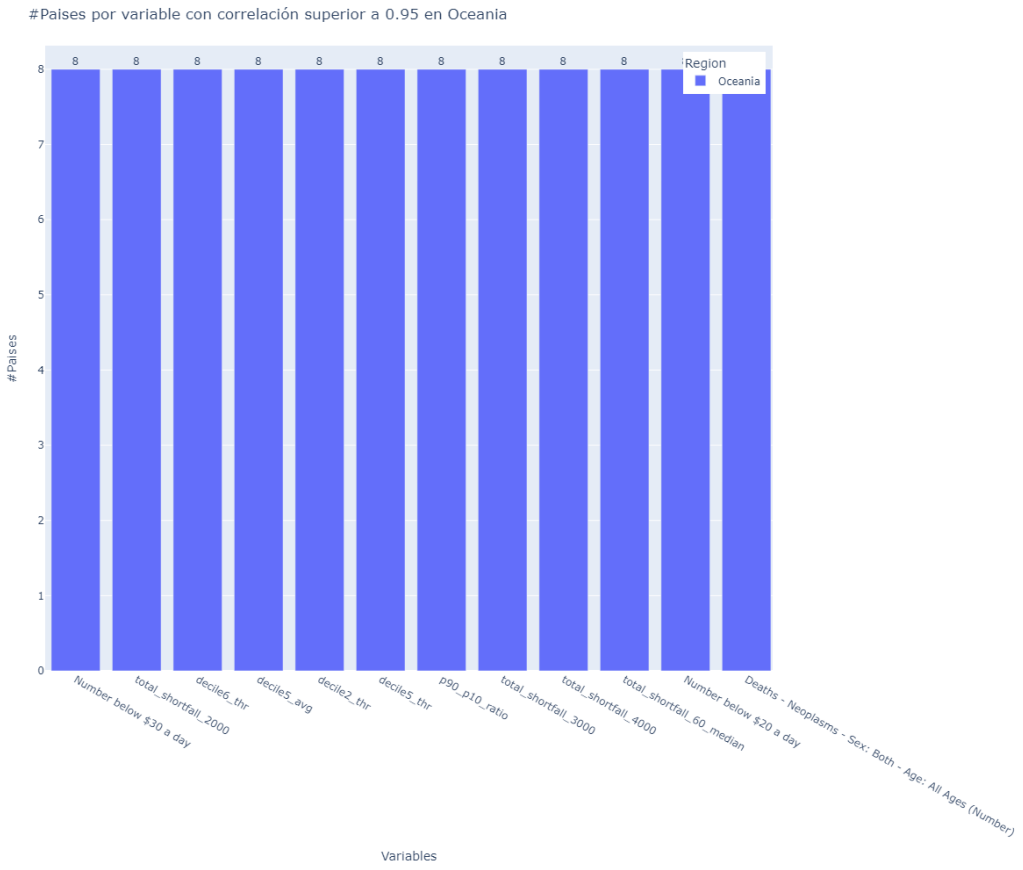


#Paises por variable con correlación superior a 0.95 en Europe



#Paises por variable con correlación superior a 0.95 en North America

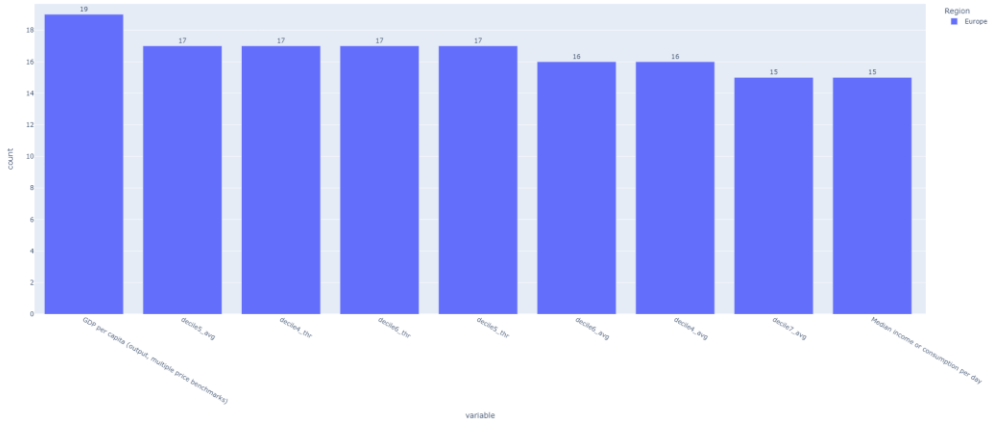



#Paises por variable con correlación superior a 0.95 en Oceania
Region
Oceania
#Paises
Variables
Number below $30 a day
total_shortfall_2000
decile6_thr
decile5_avg
decile2_thr
decile5_thr
p90_p10_ratio
total_shortfall_3000
total_shortfall_4000
total_shortfall_60_median
Number below $20 a day
Deaths - Neoplasms - Sex: Both - Age: All Ages (Number)

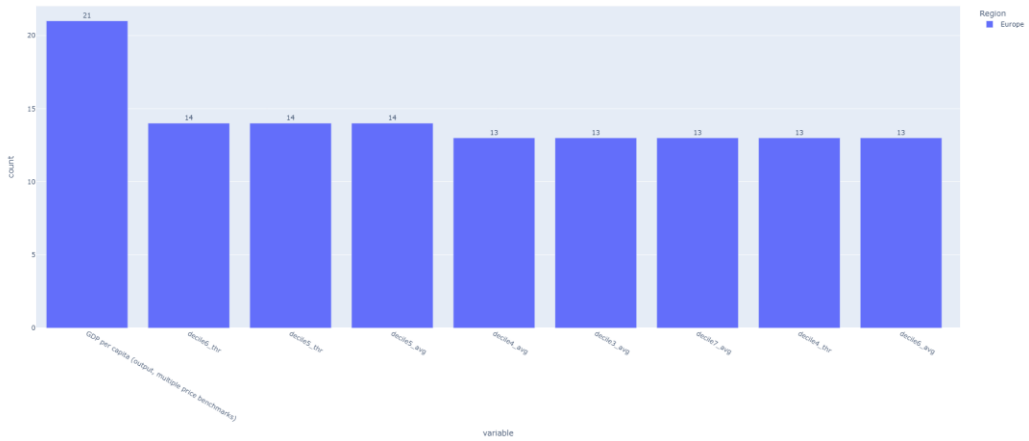
Análisis de correlación con atributos secundarios por región

Correlación de les variables de mayor correlación con la Life Expectancy en Europe

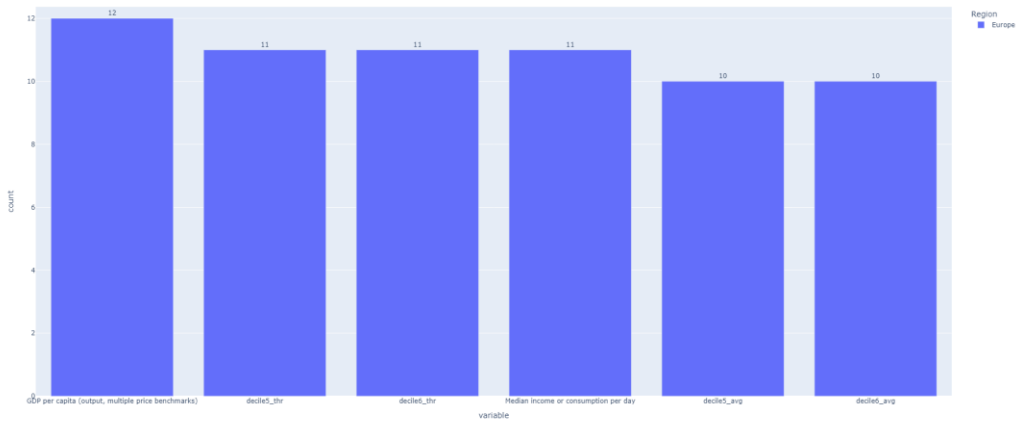
- Alzheimer



- Parkinson

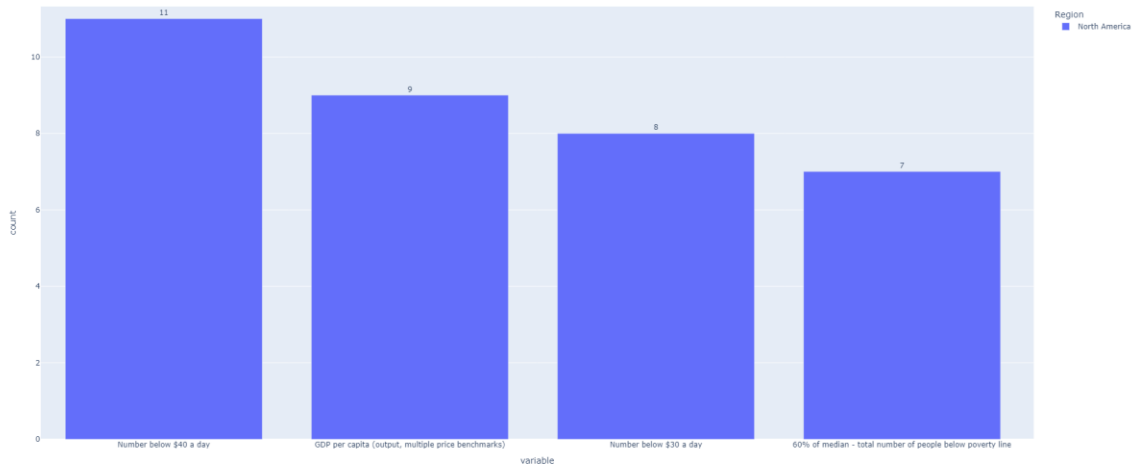


- Road injuries

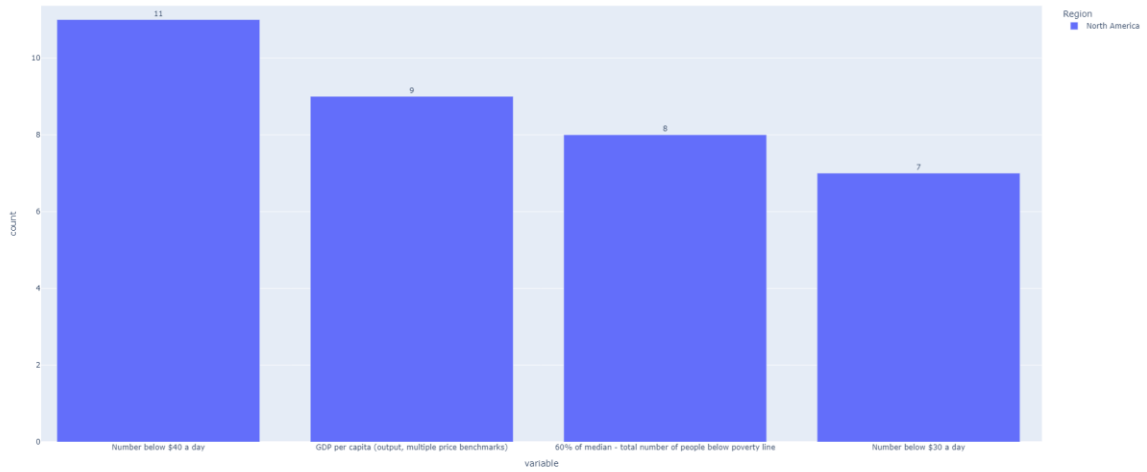


Correlación de les variables de mayor correlación con la Life Expectancy en North America

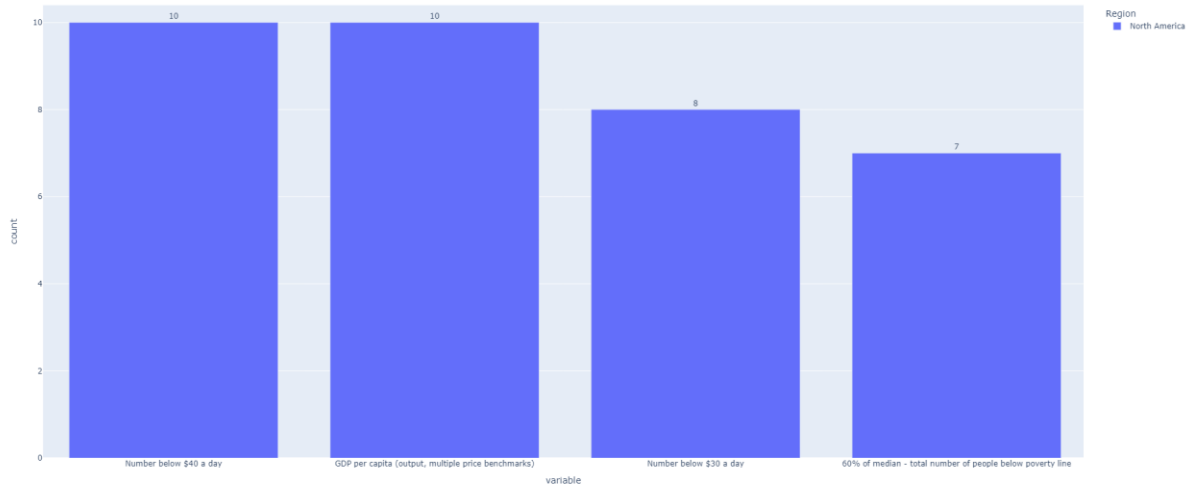
- Kidney disease



- Neoplasms

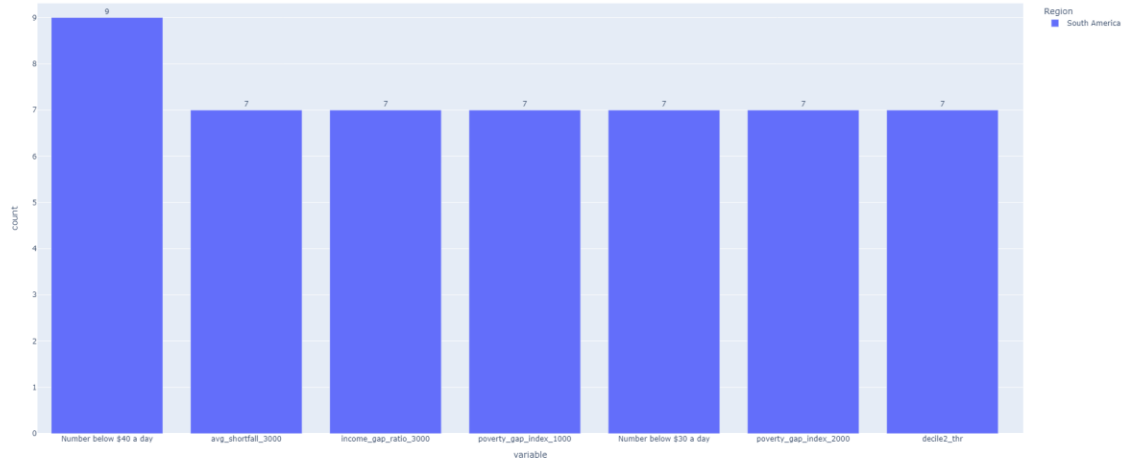


- Alzheimer



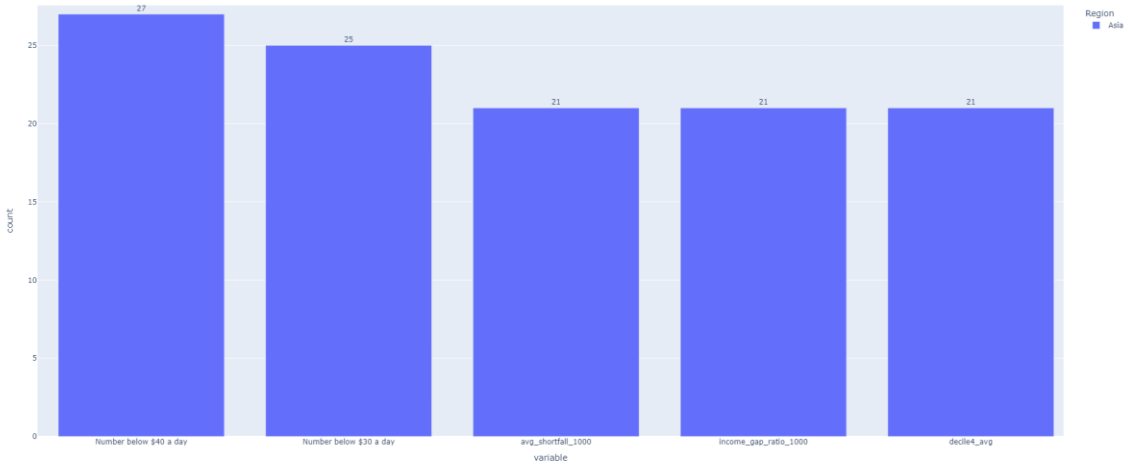
Correlación de les variables de mayor correlación con la Life Expectancy en South America

- Alzheimer



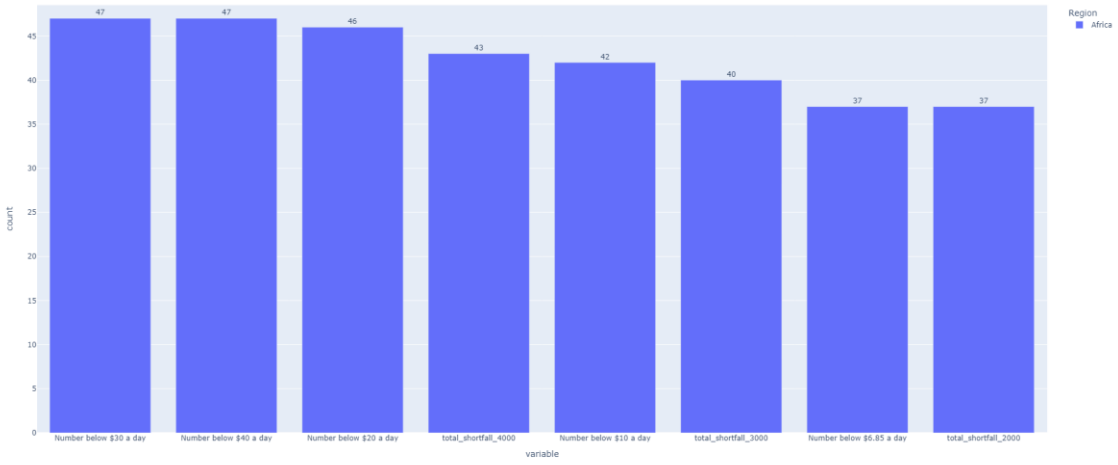
Correlación de les variables de mayor correlación con la Life Expectancy en Asia

- Alzheimer



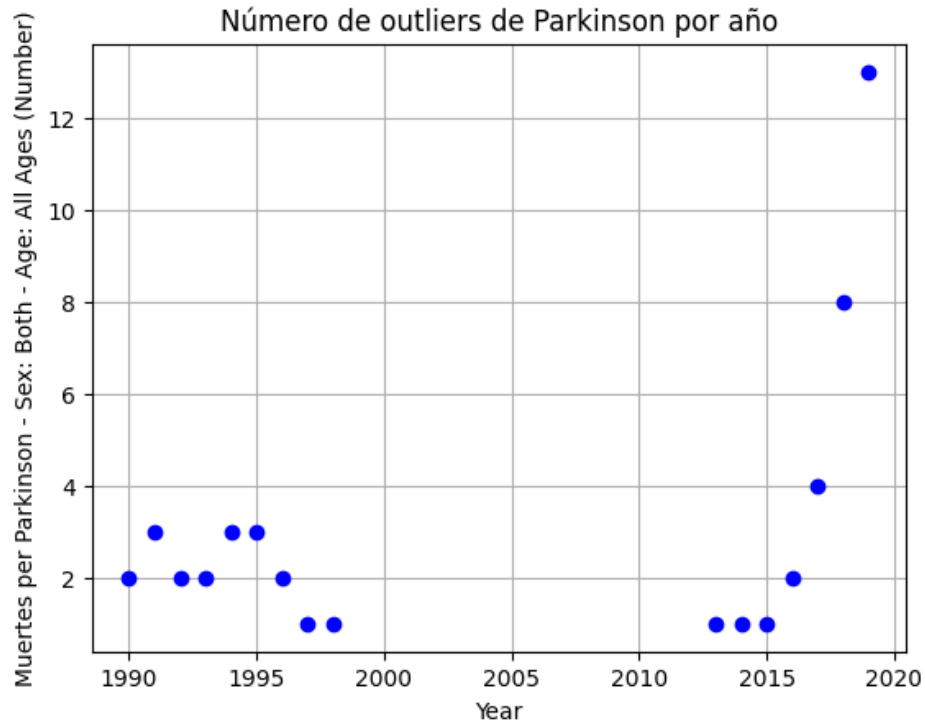
Correlación de les variables de mayor correlación con la Life Expectancy en África

- Alzheimer

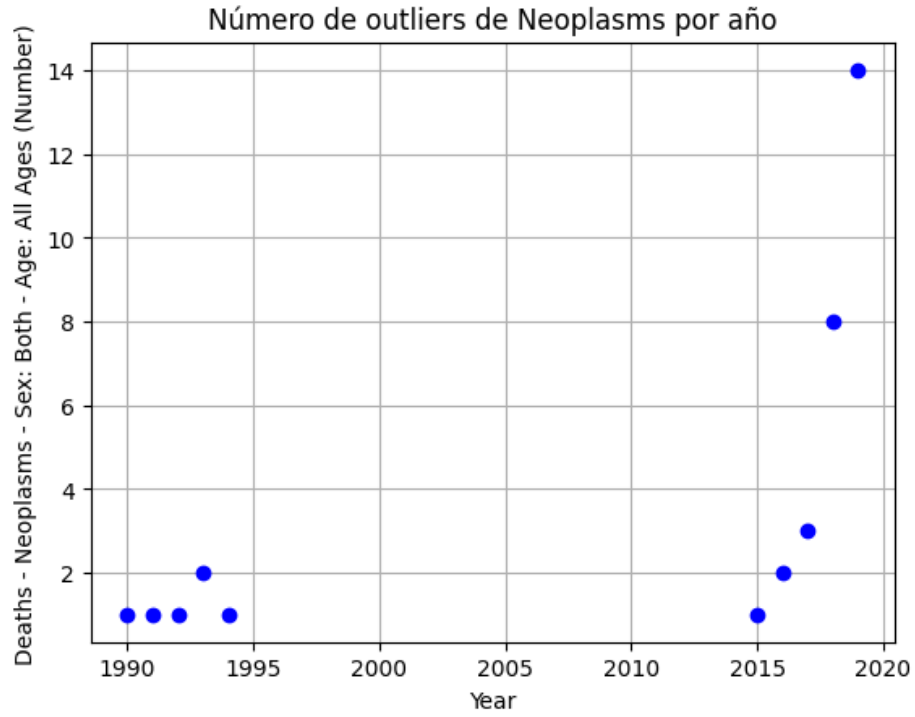


Outliers con mayor correlación

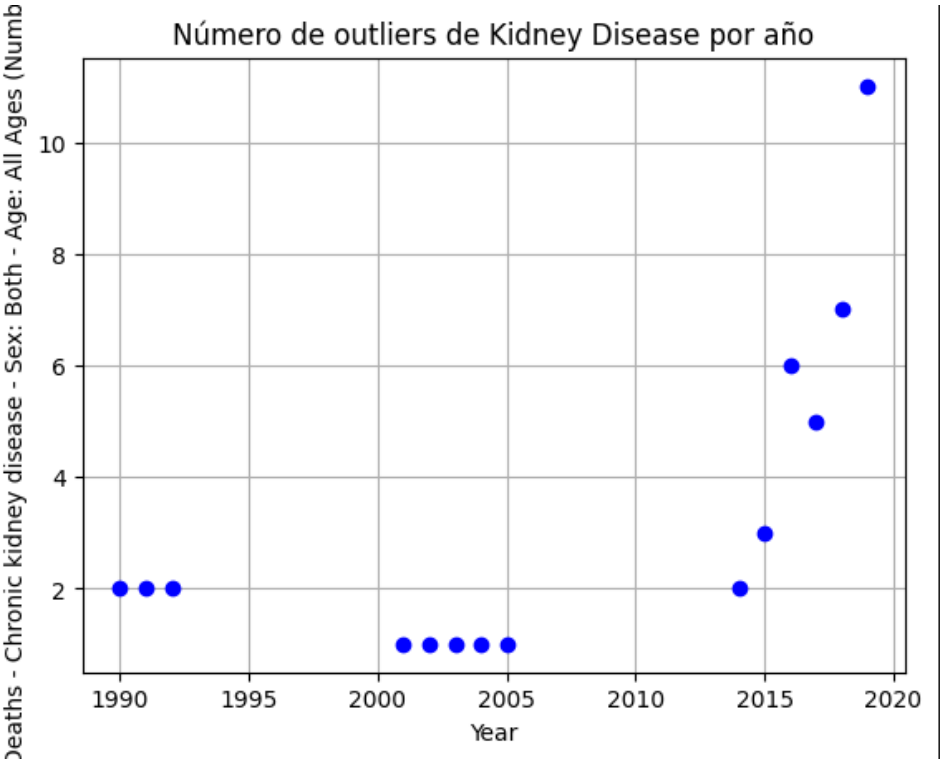
a) Parkinson



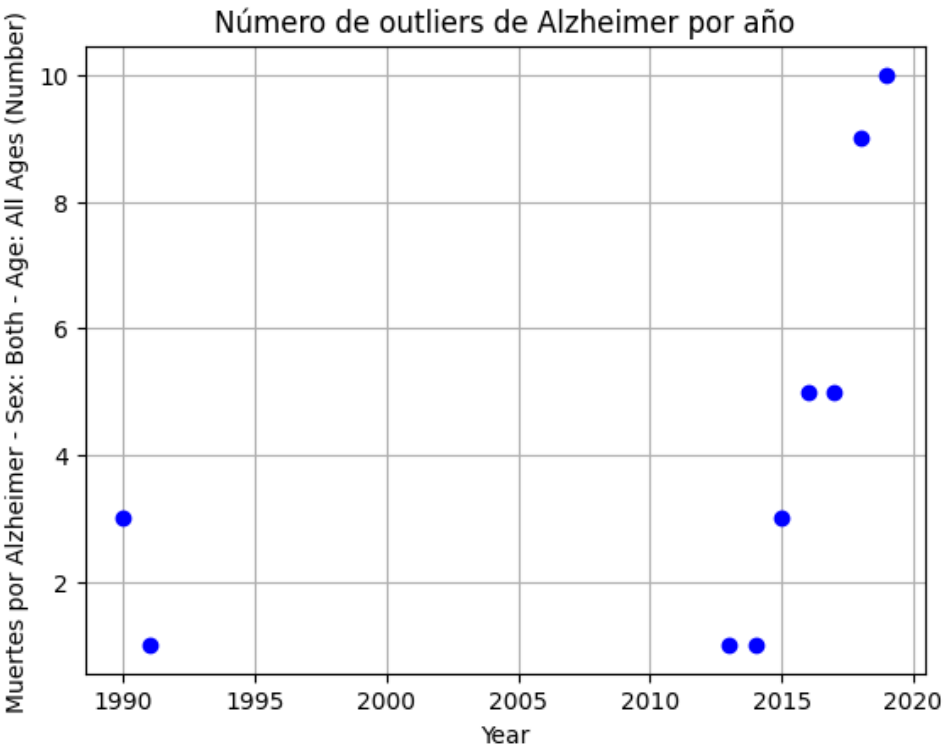
b) Neoplasms



c) Kidney disease



d) Alzheimer



Graficación de los outliers de muertes por Alzheimer de un país

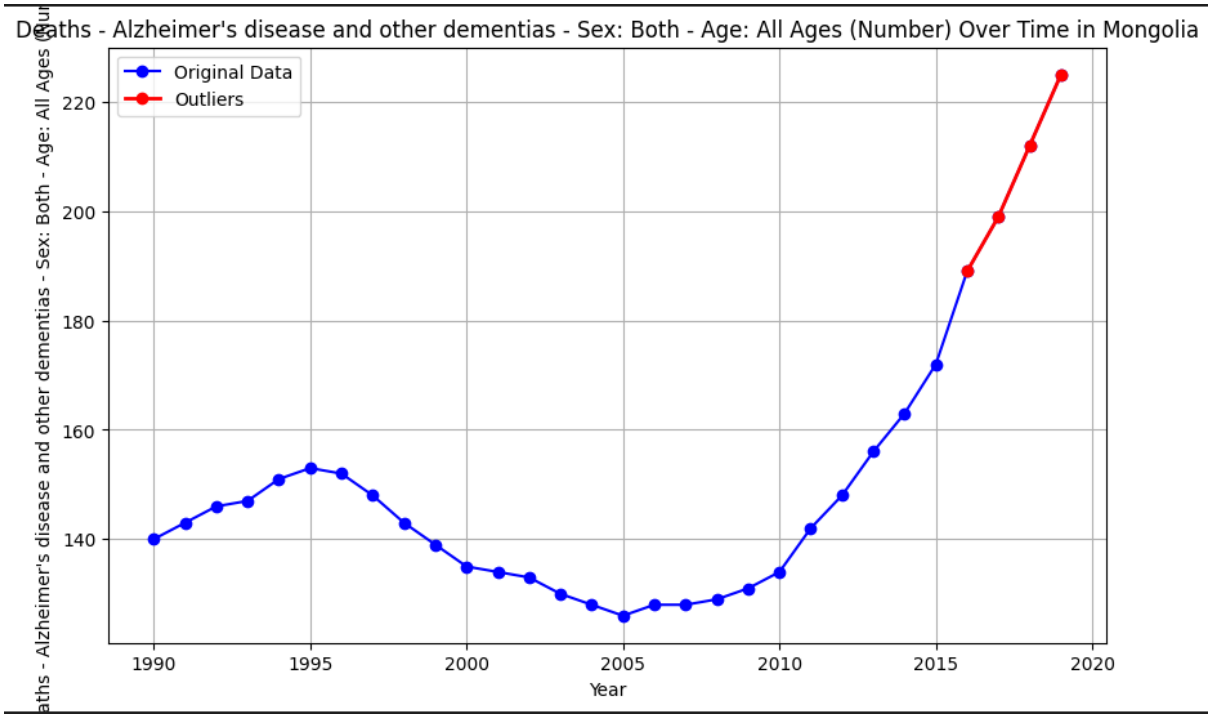
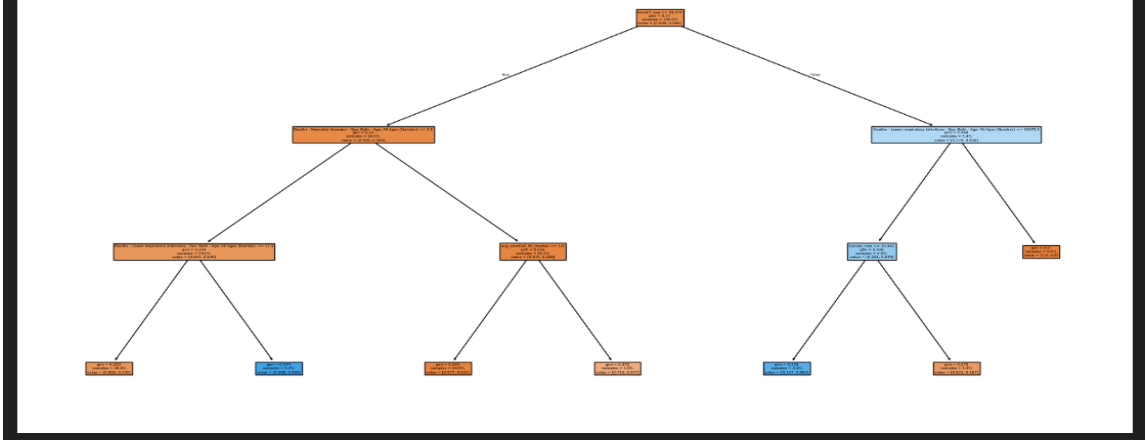


Tabla completa de los modelos de Random Forest

Modelo	rf_q90_median	rf_q90_mean	rf_q90_NoImputacion	rf_q90_median_MaxFeat75	rf_q90_median_MaxFeat2	rf_q10_median	rf_q10_mean	rf_q10_NoImputacion
Top_Feature_1	Homicide rate per 100	Homicide rate per 100	Homicide rate per 100	avg_shortfall_60_median (22.30%)	decile8_thr (4.19%)	Pol3 (% of one-year-olds immunized) (9.98%)	Pol3 (% of one-year-olds immunized) (10.72%)	Pol3 (% of one-year-olds immunized) (10.31%)
Top_Feature_2	000 population - Both sexes - All ages (6.62%)	000 population - Both sexes - All ages (6.52%)	000 population - Both sexes - All ages (6.23%)	Homicide rate per 100	avg_shortfall_40_median (2.77%)	Deaths - Malaria - Sex: Both - Age: All Ages (Number) (9.44%)	Deaths - Malaria - Sex: Both - Age: All Ages (Number) (10.37%)	Deaths - Malaria - Sex: Both - Age: All Ages (Number) (8.95%)
Top_Feature_3	avg_shortfall_60_median (6.37%)	avg_shortfall_60_median (6.13%)	avg_shortfall_60_median (6.03%)	000 population - Both sexes - All ages (13.63%)	avg_shortfall_4000 (2.65%)	Deaths - Diarrheal diseases - Sex: Both - Age: All Ages (Number) (6.82%)	GDP per capita (output	Deaths - Diarrheal diseases - Sex: Both - Age: All Ages (Number) (7.17%)
Top_Feature_4	decile8_thr (5.85%)	decile8_thr (5.82%)	decile8_thr (5.32%)	GDP per capita (output	income_gap_ratio_2000	GDP per capita (output	multiple price benchmarks) (5.97%)	GDP per capita (output
Top_Feature_5	Median income or consumption per day	Median income or consumption per day	Median income or consumption per day	multiple price benchmarks) (8.42%)	decile10_avg	multiple price benchmarks)	Deaths - Diarrheal diseases - Sex: Both - Age: All Ages (Number)	multiple price benchmarks)
Top_Feature_6	avg_shortfall_50_median	avg_shortfall_50_median	decile6_avg	avg_shortfall_50_median	Share below $30 a day	Deaths - HIV/AIDS - Sex: Both - Age: All Ages (Number)	Deaths - Meningitis - Sex: Both - Age: All Ages (Number)	Deaths - Meningitis - Sex: Both - Age: All Ages (Number)
Top_Feature_7	decile4_avg	decile4_avg	avg_shortfall_50_median	Hib3 (% of one-year-olds immunized)	avg_shortfall_60_median	Deaths - Meningitis - Sex: Both - Age: All Ages (Number)	Deaths - Nutritional deficiencies - Sex: Both - Age: All Ages (Number)	Deaths - HIV/AIDS - Sex: Both - Age: All Ages (Number)
Top_Feature_8	decile6_avg	decile6_avg	decile4_avg	BCG (% of one-year-olds immunized)	poverty_gap_index_1000	DTP3 (% of one-year-olds immunized)	Deaths - HIV/AIDS - Sex: Both - Age: All Ages (Number)	Deaths - Protein-energy malnutrition - Sex: Both - Age: All Ages (Number)
Top_Feature_9	decile9_avg	decile9_avg	decile9_avg	PCV3 (% of one-year-olds immunized)	poverty_gap_index_3000	Deaths - Protein-energy malnutrition - Sex: Both - Age: All Ages (Number)	MCV1 (% of one-year-olds immunized)	DTP3 (% of one-year-olds immunized)
Top_Feature_10	decile8_avg	decile5_avg	Share below $40 a day	Deaths - Neonatal disorders - Sex: Both - Age: All Ages (Number)	Median income or consumption per day	Deaths - Nutritional deficiencies - Sex: Both - Age: All Ages (Number)	DTP3 (% of one-year-olds immunized)	Deaths - Nutritional deficiencies - Sex: Both - Age: All Ages (Number)

Plot del primer modelo:

```
# Obtener el primer árbol del Random Forest
arbol_forest = model.estimators_[0]

# Hacer el plot del árbol de decisión
plt.figure(figsize=(20, 10))
plot_tree(arbol_forest, feature_names = X.columns, filled=True, proportion=True)
plt.show()
```



Dashboard

https://lookerstudio.google.com/u/0/reporting/6db41173-60ac-46a2-88e5-c82512f80aa1/page/JDpAE