



¿Qué podemos hacer para aumentar la esperanza de vida de las personas?

Influencia de los indicadores socioeconómicos y demográficos globales en la esperanza de vida al nacer.

Autor: Diego Leiva

Agenda

- 1 Encuadre
- 2 Preguntas de Interés
- 3 Metadata
- 4 EDA - Exploratory Data Analysis
- 5 Data Wrangling
- 6 Aplicación de Algoritmos de ML
- 7 Reducción de Dimensionalidad (PCA)
- 8 Resultados y Conclusiones
- 9 Limitaciones



1

Encuadre

La **esperanza de vida al nacer**, según la Organización Mundial de la Salud (OMS), es el número **promedio de años** que un recién nacido puede esperar **vivir**, si las tasas de mortalidad por edad actuales se mantienen constantes a lo largo de su vida.

La esperanza de vida varía significativamente entre países y regiones, influenciada por diversos factores económicos, sociales y demográficos. Este proyecto busca analizar estas variables para proporcionar información valiosa que permita:

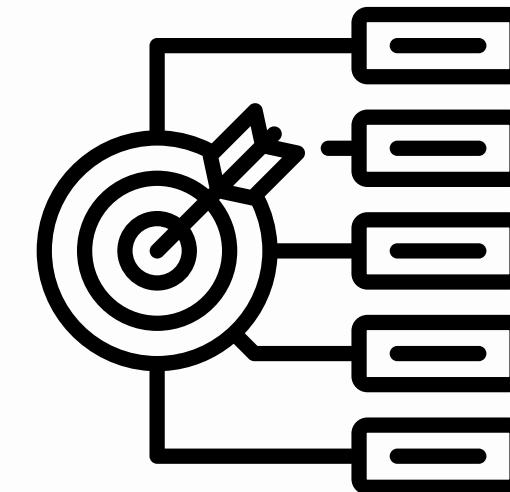
- Implementar políticas públicas efectivas para mejorar la salud y el bienestar.
- Desarrollar estrategias de prevención para enfermedades.
- Ayudar a las personas y a tomar decisiones informadas sobre su salud y estilo de vida.



Objetivos

Objetivo General

Explorar y comprender las relaciones entre variables económicas, sociales y demográficas para predecir la esperanza de vida al nacer utilizando técnicas de aprendizaje automático.



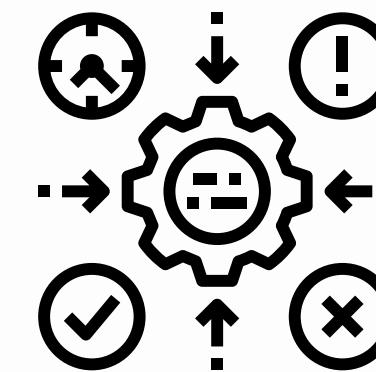
Objetivos específicos

Analizar tendencias y evolución de variables a lo largo del tiempo.
Identificar correlaciones entre esperanza de vida y otras variables.
Evaluuar el impacto económico en factores sociales y demográficos.
Comprender patrones geoespaciales en indicadores demográficos.

Contexto y Problema Comercial

Contexto Comercial

La esperanza de vida al nacer es un indicador clave del bienestar y desarrollo de las naciones, utilizado por la ONU y la OMS para formular políticas de salud y desarrollo sostenible.



Problema Comercial:

Predecir la esperanza de vida permite a gobiernos y empresas tomar decisiones estratégicas para mejorar condiciones de vida y planificar el desarrollo sostenible. Por ejemplo, empresas del sector salud pueden identificar mercados emergentes y gobiernos pueden diseñar políticas públicas eficaces.

2

Preguntas de Interés

Preguntas que se abordarán:

- ¿Qué factores socioeconómicos y demográficos influyen en la esperanza de vida?
- ¿Cómo han cambiado estos factores a lo largo del tiempo?
- ¿Qué estrategias se pueden implementar para aumentar la esperanza de vida a nivel global considerando sus indicadores más influyentes

Aplicaciones:

- Toma de decisiones estratégicas en empresas e instituciones públicas.
- Planificación de inversiones y expansión de operaciones.
- Diseño de políticas públicas para el desarrollo social y económico.
- Investigación en áreas como salud pública, economía y ciencias sociales.



3

Metadata

Metadata

201

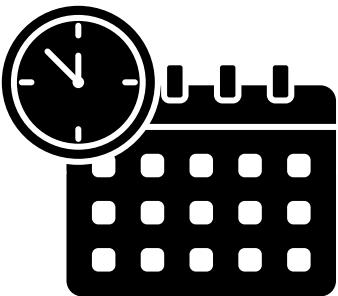
Países



Naciones de
todo el
planeta, de
todos los
continentes.

48

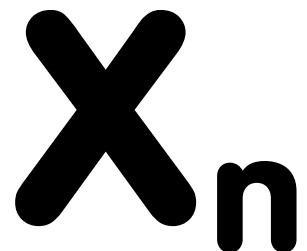
Años de estudio



Datos entregados de
forma anual, desde el
año 1973 hasta el
año 2021.

32

Indicadores



Donde encontramos
variables económicas
y demográficas.

~9700

Registros



Más de 9 mil
registros, de todos
los países en los
años de estudio.

Dataset: Global Socio-Economic & Demographic Insights

Fuente: Banco Mundial. URL: https://www.kaggle.com/datasets/samybaladram/databank-world-development-indicators?select=world_development_data_interpolated.csv

4

EDA - Exploratory Data Analysis

Columna	Non-Null Count	Dtype
Year	9947	int64
Country	9947	object
Region	9947	object
SubRegion	9947	object
IntermRegion	4214	object
SurfAreaSqKm	9900	float64
PopDens	9297	float64
PopGrowth%	9929	float64
GDP	8668	float64
GDPGrowth%	8424	float64
AdolFertRate	9947	float64
AgriValAdd%GDP	7496	float64
DomCredit%GDP	1185	float64
Exports%GDP	7434	float64
FertRate	9746	float64
FDINetBoP	8557	float64
GNI/CapAtlas	7867	float64
GNIAtlas	7868	float64
GrossCapForm%GDP	7179	float64
Imports%GDP	7434	float64
IndValAdd%GDP	7445	float64
InflConsPric%	8397	float64
LifeExpBirth	9740	float64
MerchTrade%GDP	8333	float64
MilExp%GDP	6433	float64
MobileSubs/100	9196	float64
MortRateUS	9133	float64
NetMigr	9947	float64
PopTotal	9930	float64
RevenueExGrants%GDP	4174	float64
SchEnrollPrim%	7930	float64
TaxRevenue%GDP	4228	float64
UrbanPopGrowth%	9929	float64

> Dataset:

Global Socio-Economic & Demographic Insights 

> Variable de Interés:

'LifeExpBirth'

> Registros:

9947

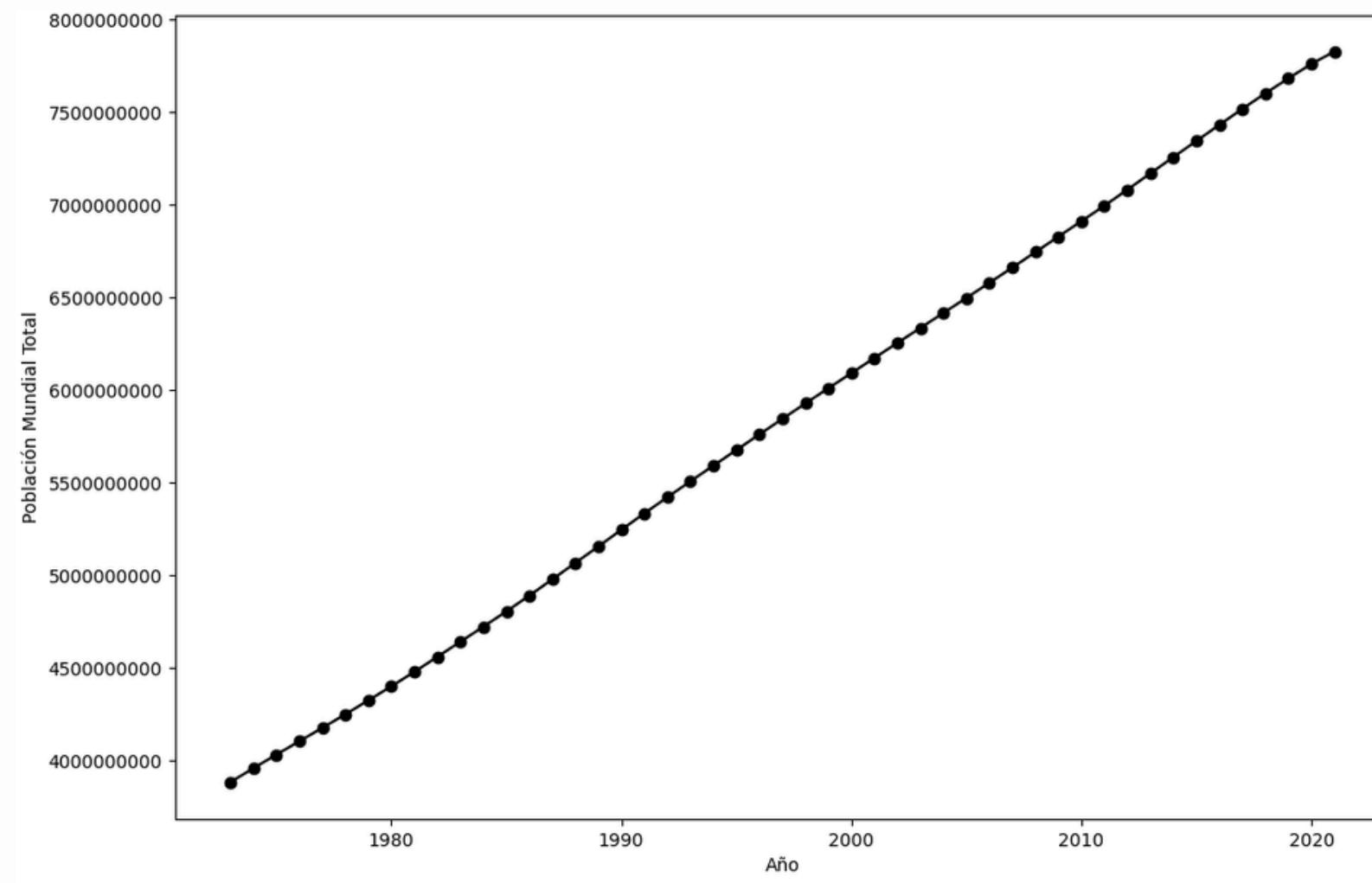
> Columnas:

33

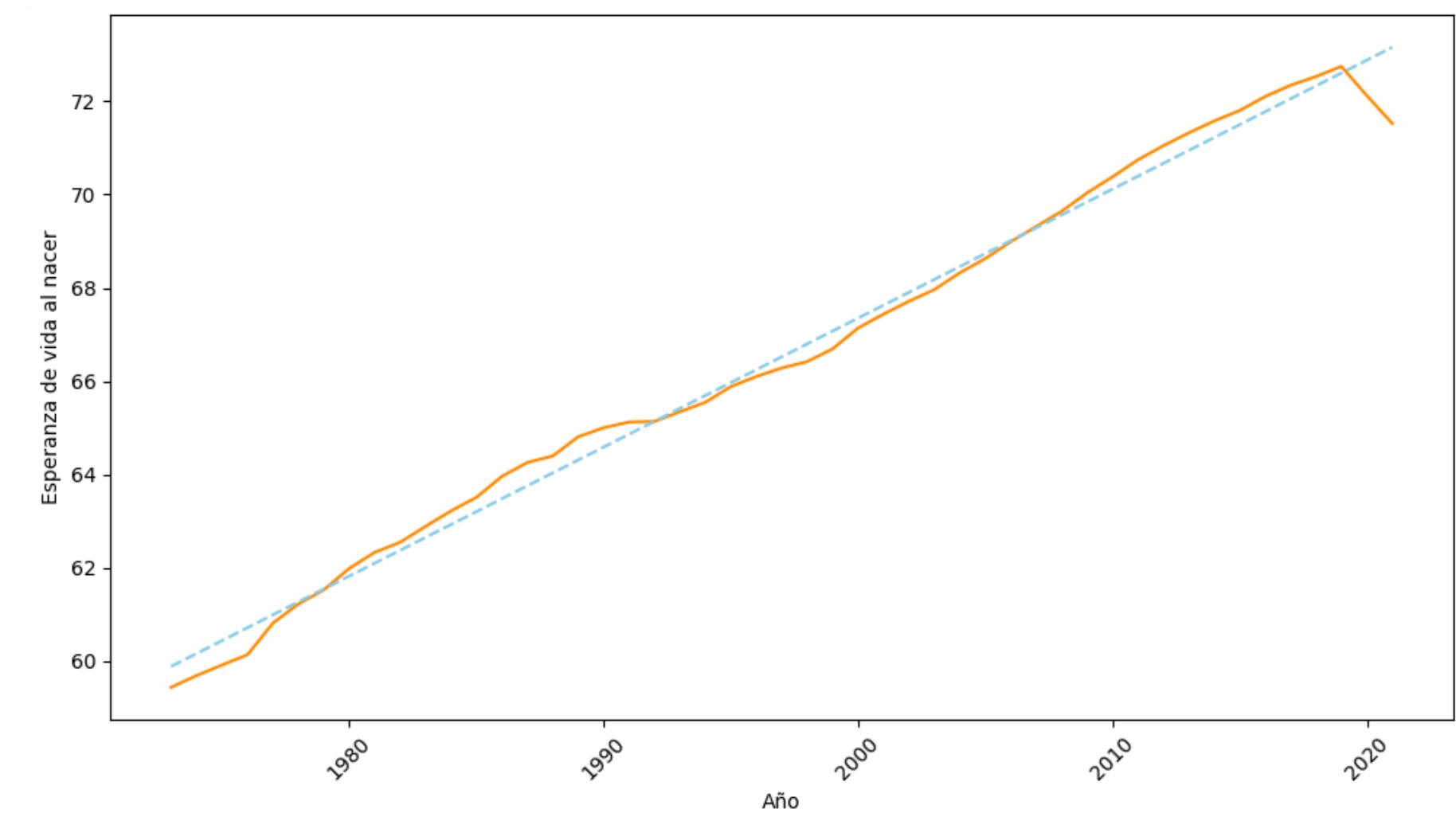
> Cantidad de variables por tipo:

- int64: 1
- object: 4
- float64: 28

Población Mundial Total por Año



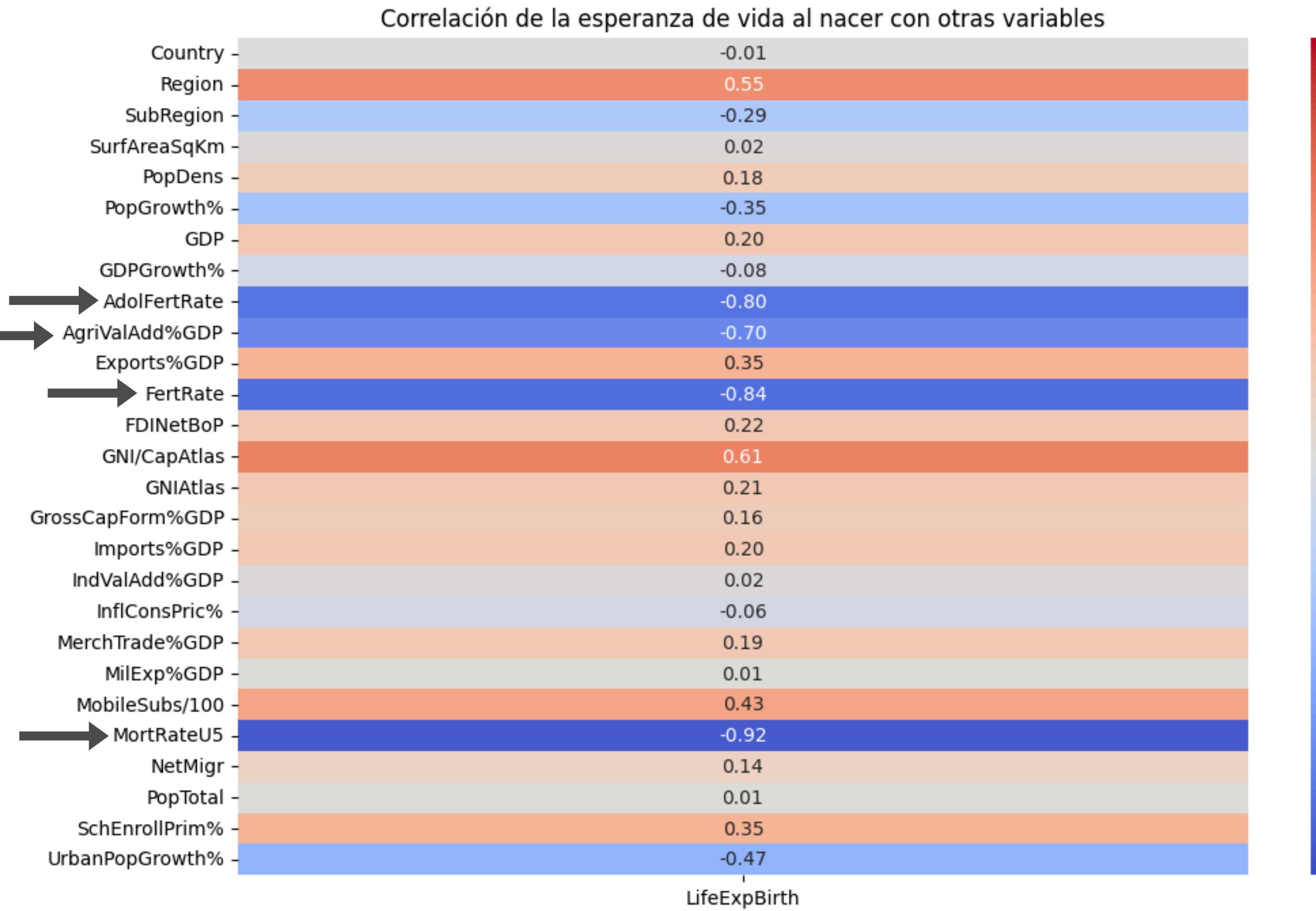
Promedio de Esperanza de Vida al Nacer por Año



La población mundial ha experimentado un notable crecimiento desde 1973 hasta 2021, pasando de 3.8 a 7.8 billones de habitantes en ese período, casi se ha duplicado en menos de cuatro décadas.

El gráfico muestra que la esperanza de vida promedio ha aumentado a través del tiempo. Esto coincide con el crecimiento de la población mundial.

¿Qué variables se relacionan mejor con la esperanza de vida al nacer?

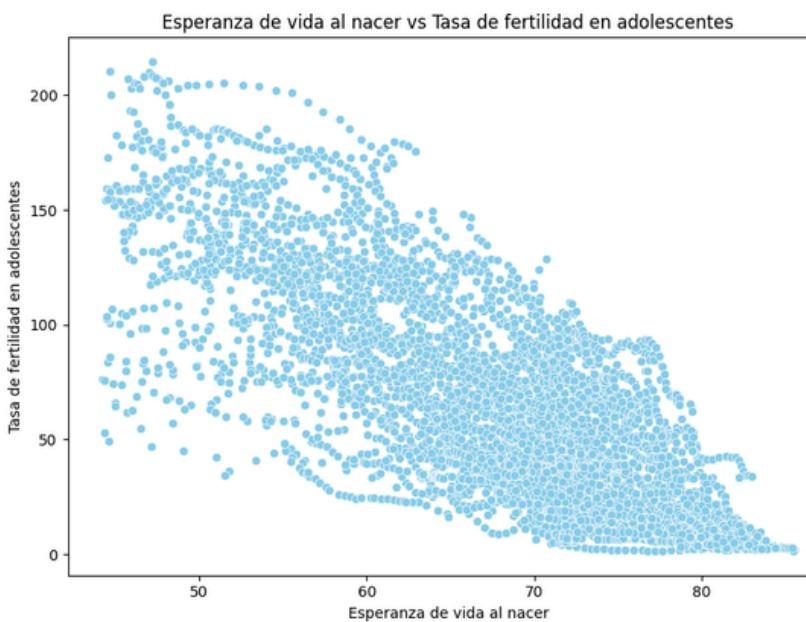


Se observan **4 indicadores** que se relacionan fuertemente con la esperanza de vida al nacer:

- **AdolFertRate:** Tasa de fertilidad en adolescentes.
- **AgriValAdd%GDP:** Valor agregado de la agricultura (% del PIB).
- **FertRate:** Tasa de fertilidad total.
- **MortRateU5:** Tasa de mortalidad de menores de cinco años.

Insights sobre indicadores que mayor influencia tienen sobre la esperanza de vida al nacer

1. Tasa de fertilidad en adolescentes



Correlación: Negativa y fuerte (índice de correlación: **-0.8**)

Insights:

- Las madres adolescentes enfrentan mayores riesgos durante el embarazo y el parto, lo que puede aumentar la mortalidad tanto para la madre como para el bebé.
- Los hijos de madres adolescentes tienen un mayor riesgo de nacer con bajo peso o con problemas de salud, lo que puede afectar su desarrollo y supervivencia.
- La falta de recursos económicos y educativos entre las madres adolescentes puede dificultarles brindar la atención médica y la nutrición adecuadas a sus hijos.

2. Valor agregado de la agricultura (% del PIB)



Correlación: Negativa y fuerte (índice de correlación: **-0.7**)

Insights:

- Países con un sector agrícola dominante tienden a tener un menor desarrollo económico y menos inversión en salud, educación e infraestructura.
- La falta de diversificación económica puede aumentar la pobreza y la desigualdad, limitando el acceso a atención médica y alimentación de calidad.
- Un sector agrícola fuerte puede impactar negativamente el medio ambiente, causando deforestación, contaminación del agua y uso excesivo de pesticidas.

3. Tasa de fertilidad total

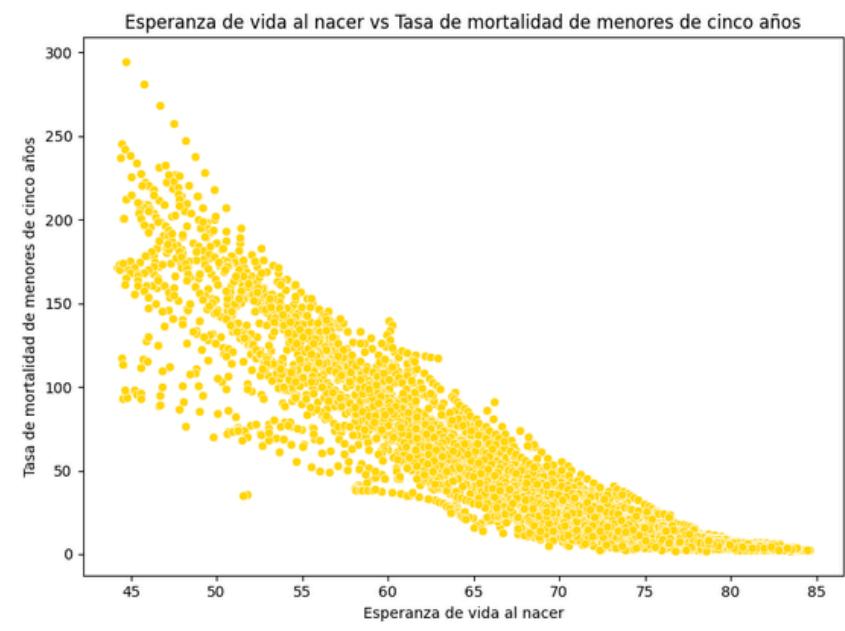


Correlación: Negativa y muy fuerte (índice de correlación: **-0.84**)

Insights:

- Una **alta tasa de fertilidad** puede poner una **presión sobre los recursos del país**, como la atención médica, la educación y la infraestructura.
- Cuando hay **muchos niños**, puede ser **más difícil** para los padres brindarles a todos la **atención y el cuidado** que necesitan.
- Las **familias numerosas** pueden tener **menos recursos económicos**, lo que puede afectar la salud y el bienestar de sus miembros.

4. Tasa de mortalidad de menores de cinco años



Correlación: Negativa y muy fuerte (índice de correlación: **-0.92**)

Insights:

- Una alta tasa de mortalidad infantil indica problemas en el sistema de salud, como acceso limitado a atención médica de calidad y agua potable.
- La mortalidad infantil puede impactar significativamente la esperanza de vida general de la población.

5

Data Wrangling

Valores nulos por década

Estadísticas de valores nulos y no nulos por década:				
Decada	Total Nulos	Total No Nulos	Total	Porcentaje Nulos
1970	14295.0	34019.0	48314.0	29.587697
1980	16651.0	52369.0	69020.0	24.124891
1990	11415.0	57605.0	69020.0	16.538684
2000	7979.0	61041.0	69020.0	11.560417
2020	1566.0	12238.0	13804.0	11.344538
2010	6597.0	62423.0	69020.0	9.558099

- Las décadas de los 70s y 80s concentraban casi el 50% de los valores perdidos. Se decidió eliminarlas del DataFrame.

Valores nulos por columna:

	Nulos	No Nulos	Total	Porcentaje Nulos
DomCredit%GDP	8762	1185	9947	88.086860
RevenueExGrants%GDP	5773	4174	9947	58.037599
IntermRegion	5733	4214	9947	57.635468
TaxRevenue%GDP	5719	4228	9947	57.494722
MilExp%GDP	3514	6433	9947	35.327234
GrossCapForm%GDP	2768	7179	9947	27.827486
Imports%GDP	2513	7434	9947	25.263899

- Para el caso de los valores perdidos por columna, se decidió eliminar aquellas con un numero mayor a 50% de valores perdidos.

El DataFrame quedaría de la siguiente forma:

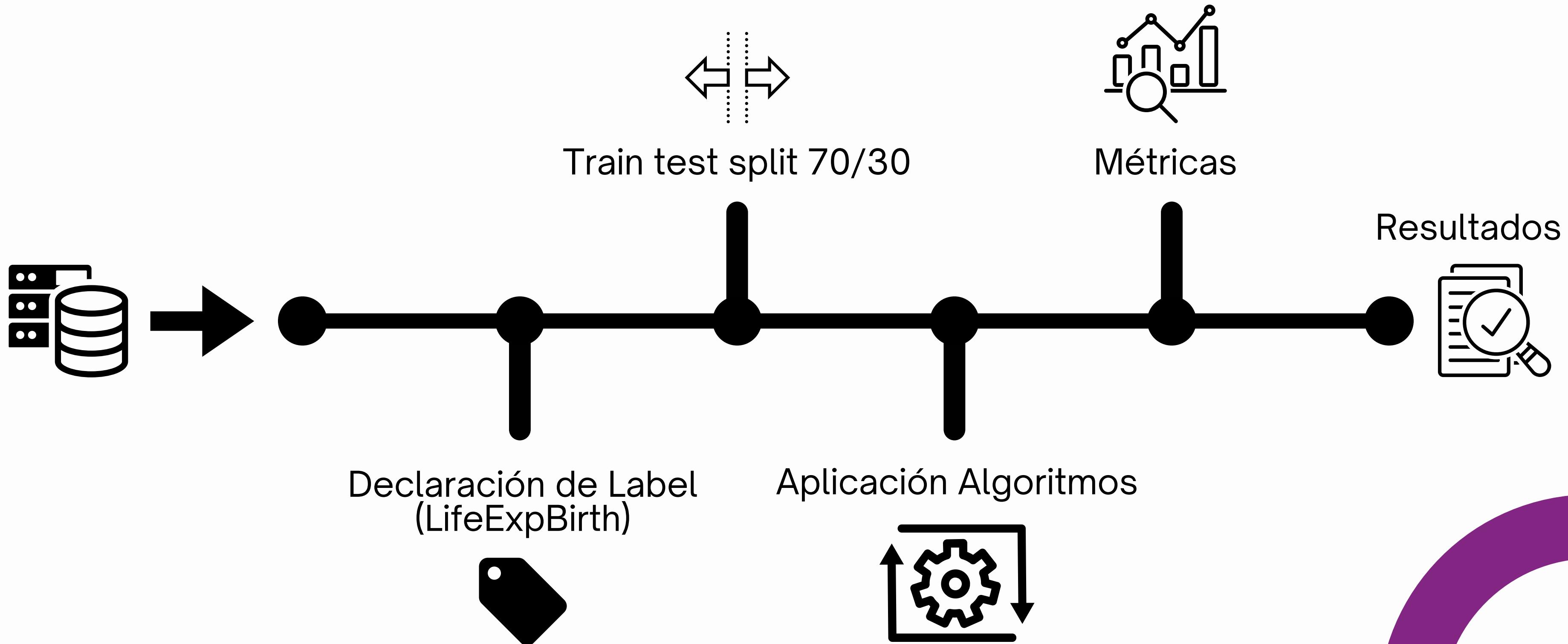
Registros: 6318

Columnas: 29

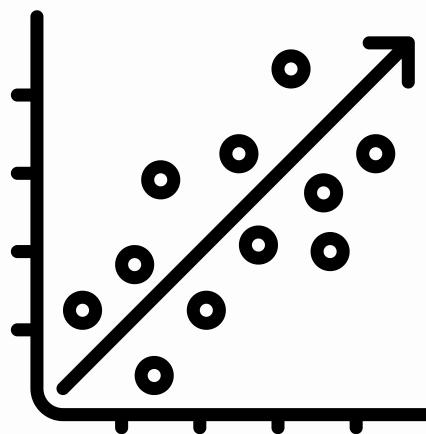
6

Aplicación de Algoritmos de ML

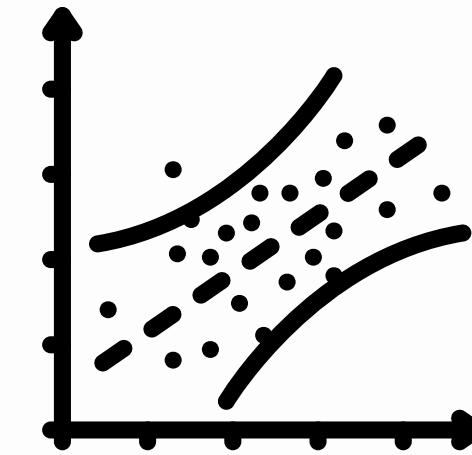
Etapas



Algoritmos Empleados



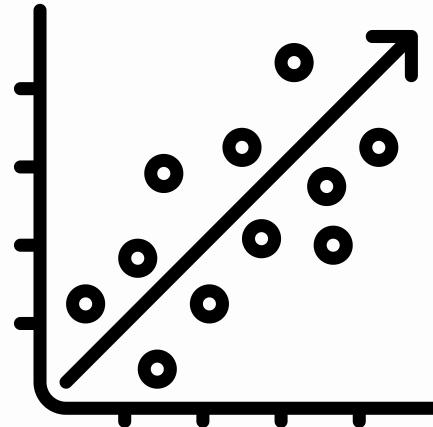
Regresión
Lineal



Support Vector
Regression

Regresión Lineal: Resumen Métricas

A continuación se presenta un resumen con los principales resultados obtenidos para cada modelo aplicado al dataset:



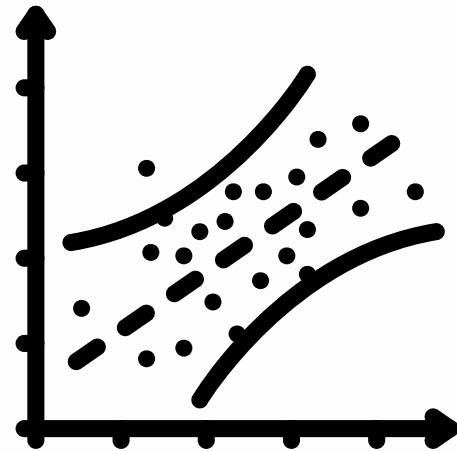
Métrica	Regresión Lineal
<hr/>	
MAE en conjunto de prueba	2.13
MAE promedio usando 10-fold cross-validation	2.24
RMSE en conjunto de prueba	2.96
RMSE promedio usando 10-fold cross-validation	3.04

Observación

El modelo de regresión lineal ha mostrado un rendimiento sólido con diferencias mínimas entre los MAE y RMSE obtenidos mediante validación cruzada y en el conjunto de prueba. Esto indica que el modelo está bien ajustado y es capaz de realizar predicciones precisas sobre la esperanza de vida al nacer en datos no vistos. (Posteriormente se aplicó Optimización de parámetros, sin embargo, arrojó que los mejores eran los entregados por defecto)

SVR: Resumen Métricas

A continuación se presenta un resumen con los principales resultados obtenidos para SVR aplicado al dataset:



Métrica	SVR
MAE en conjunto de prueba	6.27
MAE promedio usando 10-fold cross-validation	6.42
RMSE en conjunto de prueba	8.28
RMSE promedio usando 10-fold cross-validation	8.37

Observación

El modelo SVR tuvo un rendimiento aceptable, pero con MAE y RMSE más altos respecto a la regresión lineal, por lo que no fue tan efectivo para predecir la esperanza de vida.

7

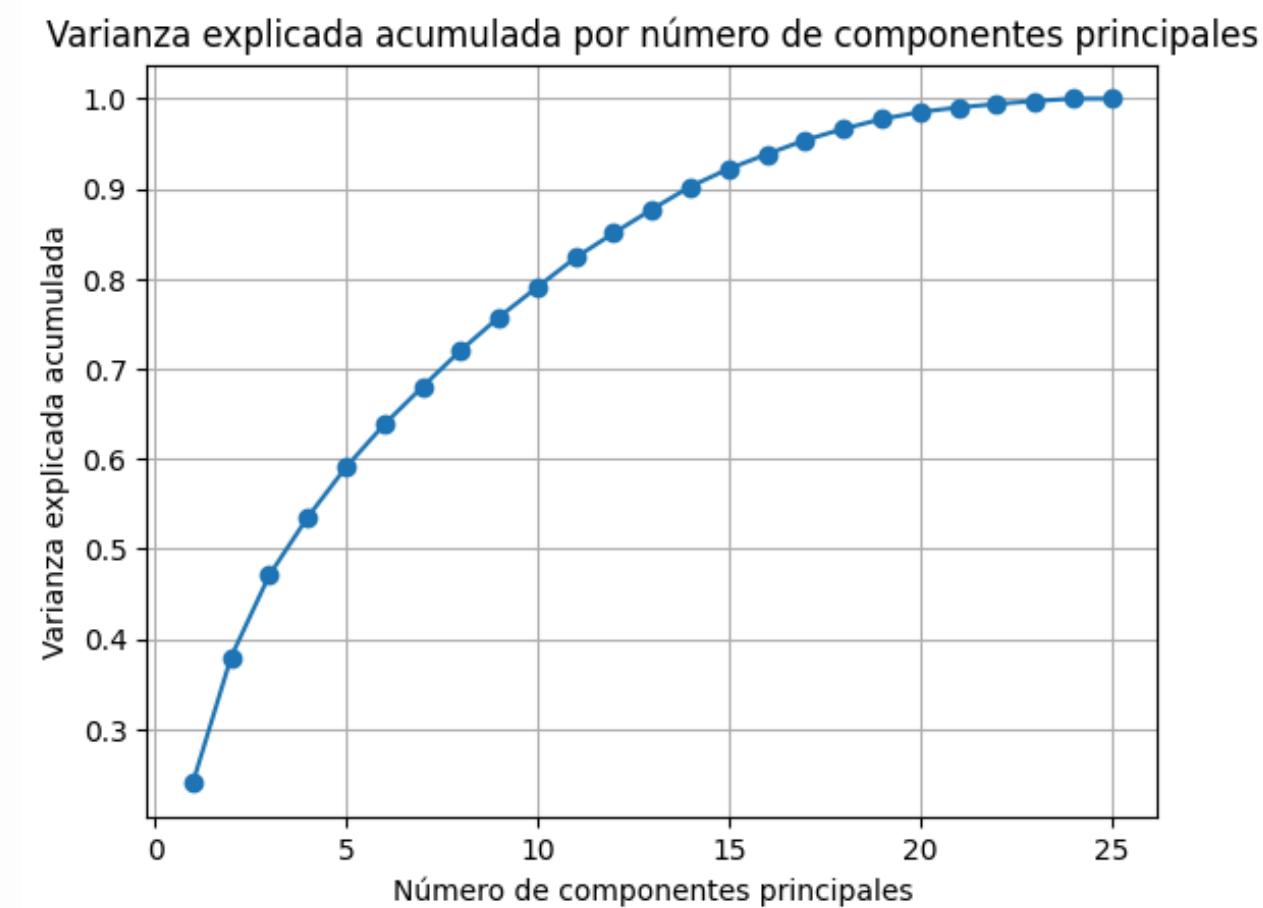
Reducción de Dimensionalidad (PCA)

Aplicación de PCA

1. Estandarización de variables

```
# Inicializar el objeto StandardScaler  
sc = StandardScaler()  
  
# Estandarizar las variables en df3_pca  
df3_pca_sc = sc.fit_transform(df3_pca)  
  
# Crear un nuevo DataFrame con las variables estandarizadas  
df3_pca_sc = pd.DataFrame(df3_pca_sc, columns=df3_pca.columns)
```

2. Selección de Componentes Principales: 14



Varianza explicada: 0.9

8

Resultados y Conclusiones

Resultados



Métrica	Regresión Lineal sin PCA	Regresión Lineal con PCA	SVR sin PCA	SVR con PCA
MAE en conjunto de prueba	2.13	0.21	6.27	0.17
MAE promedio usando 10-fold cross-validation	2.24	0.22	6.42	0.17
RMSE en conjunto de prueba	2.96	0.29	8.28	0.25
RMSE promedio usando 10-fold cross-validation	3.04	0.30	8.37	0.24

Conclusión Final

La aplicación de PCA ha mejorado significativamente el rendimiento de los modelos de regresión lineal y SVR.

- SVR con PCA mostró una notable reducción en errores absolutos y cuadráticos medios, siendo especialmente efectivo con la reducción de dimensionalidad.
- Recomendación: Utilizar SVR con PCA para predecir la esperanza de vida al nacer debido a su mayor precisión y consistencia.

Este modelo proporciona predicciones más precisas y eficientes, permitiendo a las organizaciones tomar decisiones mejor informadas y desarrollar estrategias para mejorar la calidad de vida y el bienestar de la población.

9

Limitaciones

Limitaciones



- **Datos Incompletos:** Algunos países y períodos tienen datos faltantes, lo que puede afectar la precisión de las predicciones y análisis.
- **Simplificación del Modelo:** Aunque PCA ayuda a reducir la dimensionalidad, puede perderse información valiosa durante el proceso de simplificación.
- **Variabilidad de los Datos:** Los datos socioeconómicos y demográficos pueden tener alta variabilidad y estar sujetos a cambios abruptos que el modelo podría no capturar adecuadamente.
- **Suposiciones del Modelo:** Los modelos de regresión lineal y SVR tienen suposiciones inherentes que podrían no reflejar completamente la realidad compleja y multifacética de las variables analizadas.
- **Limitaciones Temporales:** El análisis está limitado por el rango temporal de los datos (1973-2021), lo que puede no capturar las tendencias futuras o los cambios recientes.
- **Falta de Variables Contextuales:** Factores contextuales importantes, como políticas específicas de salud, conflictos o desastres naturales, no están incluidos en los datos, lo que podría influir en la esperanza de vida.



Gracias