Análisis de la Esperanza de Vida y Variables Asociadas

Santiago Flynn
Julio 2023

Contexto Comercial

En el contexto comercial actual, el análisis de datos y la aplicación de técnicas de aprendizaje automático son fundamentales para comprender mejor a los clientes, optimizar las operaciones y tomar decisiones estratégicas. En nuestro proyecto, nos enfocamos en el análisis de variables relacionadas con la esperanza de vida en diferentes países. Este análisis tiene una gran relevancia, ya que los gobiernos y las organizaciones buscan constantemente mejorar la calidad de vida de las personas y aumentar la esperanza de vida. Al utilizar el poder de los datos y los modelos de regresión, podemos identificar los factores que influyen en la esperanza de vida y tomar decisiones informadas para impulsar mejoras significativas en diferentes aspectos de la sociedad.

Motivación y Audiencia

La motivación detrás de este proyecto radica en la necesidad de comprender las variables que influyen en la esperanza de vida de las personas y determinar qué países ofrecen mejores condiciones para un estilo de vida saludable y prolongado. En este sentido, la audiencia objetivo de este proyecto abarca tanto a individuos que están tomando decisiones sobre dónde vivir como a gobiernos y entidades que buscan implementar políticas y programas para mejorar la esperanza de vida de su población. Al proporcionar información y conocimientos basados en datos, este proyecto tiene como objetivo brindar una guía valiosa para la toma de decisiones informadas tanto a nivel individual como a nivel gubernamental.

Problema Comercial

El problema comercial que abordamos en este proyecto se centra en la necesidad de contar con un modelo predictivo que pueda orientar la toma de decisiones adecuadas en la implementación de políticas sociales de desarrollo. Con la gran cantidad de datos disponibles sobre variables relacionadas con la esperanza de vida, buscamos identificar las variables más relevantes y comprender cómo influyen en la calidad de vida de las personas. Al desarrollar un modelo predictivo preciso, podemos proporcionar a los responsables de la toma de decisiones una herramienta valiosa que les permita evaluar el impacto potencial de diferentes políticas y medidas en la esperanza de vida de la población. Esto les permite tomar decisiones más informadas y diseñar estrategias más efectivas para mejorar la calidad de vida de las personas y promover un desarrollo social sostenible.

Contexto Analítico

Nos basamos en el conjunto de datos proporcionado por el Global Health Observatory (GHO) de la Organización Mundial de la Salud (OMS). Este conjunto de datos abarca una amplia gama de variables relacionadas con la salud y el desarrollo socioeconómico de los países.

Este conjunto de datos proporciona una amplia gama de variables que nos permiten analizar y comprender mejor los factores que influyen en la esperanza de vida de las personas en diferentes países. Utilizando técnicas de análisis de datos y modelos de regresión, podemos explorar las relaciones entre estas variables y desarrollar un entendimiento más profundo de los determinantes de la esperanza de vida. Esto nos permite tomar decisiones informadas y desarrollar estrategias efectivas para mejorar la calidad de vida de las personas y promover el desarrollo socioeconómico sostenible.

Data Understanding - Campos

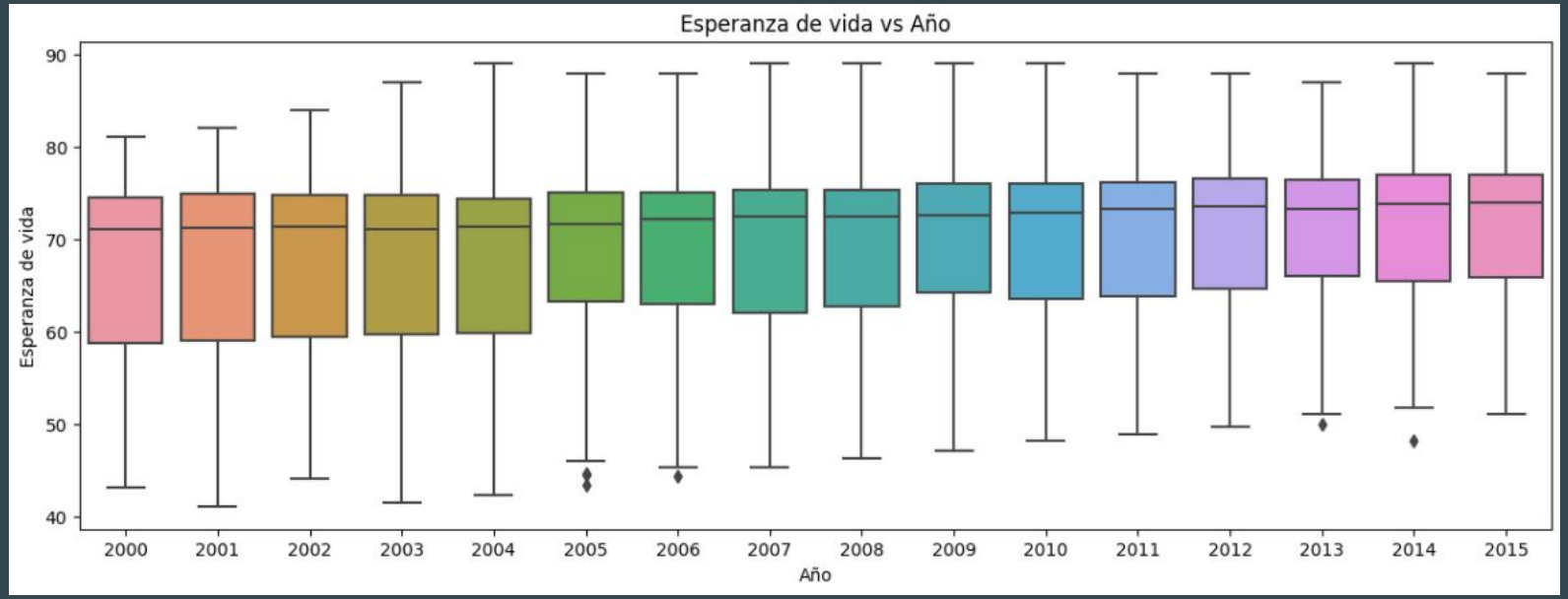
- **Country** (País)
- **Year** (Año)
- **Life expectancy** (Esperanza de vida)
- **Adult Mortality** (probabilidad de morir entre los 15 y 60 años por cada 1000 habitantes)
- **Infant deaths** (Número de muertes de infantes por cada 1000 habitantes)
- **Alcohol** (Consumo de alcohol per cápita registrado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

Data Understanding - Preguntas claves

- ¿Afectan realmente a la esperanza de vida varios factores de predicción que se han elegido inicialmente?
- ¿Cuáles son las variables de predicción que realmente afectan la esperanza de vida?
- ¿Debería un país con un valor de esperanza de vida más bajo (<65) aumentar su gasto sanitario para mejorar su esperanza de vida media?
- ¿Cómo afectan las tasas de mortalidad infantil y adulta a la esperanza de vida?
- ¿La esperanza de vida tiene una correlación positiva o negativa con los hábitos alimenticios, el estilo de vida, el ejercicio, el tabaquismo, el consumo de alcohol, etc.?
- ¿Cuál es el impacto de la escolarización en la esperanza de vida de los seres humanos?
- ¿La esperanza de vida tiene una relación positiva o negativa con el consumo de alcohol?
- ¿Los países densamente poblados tienden a tener una esperanza de vida más baja?
- ¿Cuál es el impacto de la cobertura de vacunación en la esperanza de vida?

EDA - Exploratory Data Analysis

Para el análisis de datos usamos las librerías Matplotlib y Seaborn.

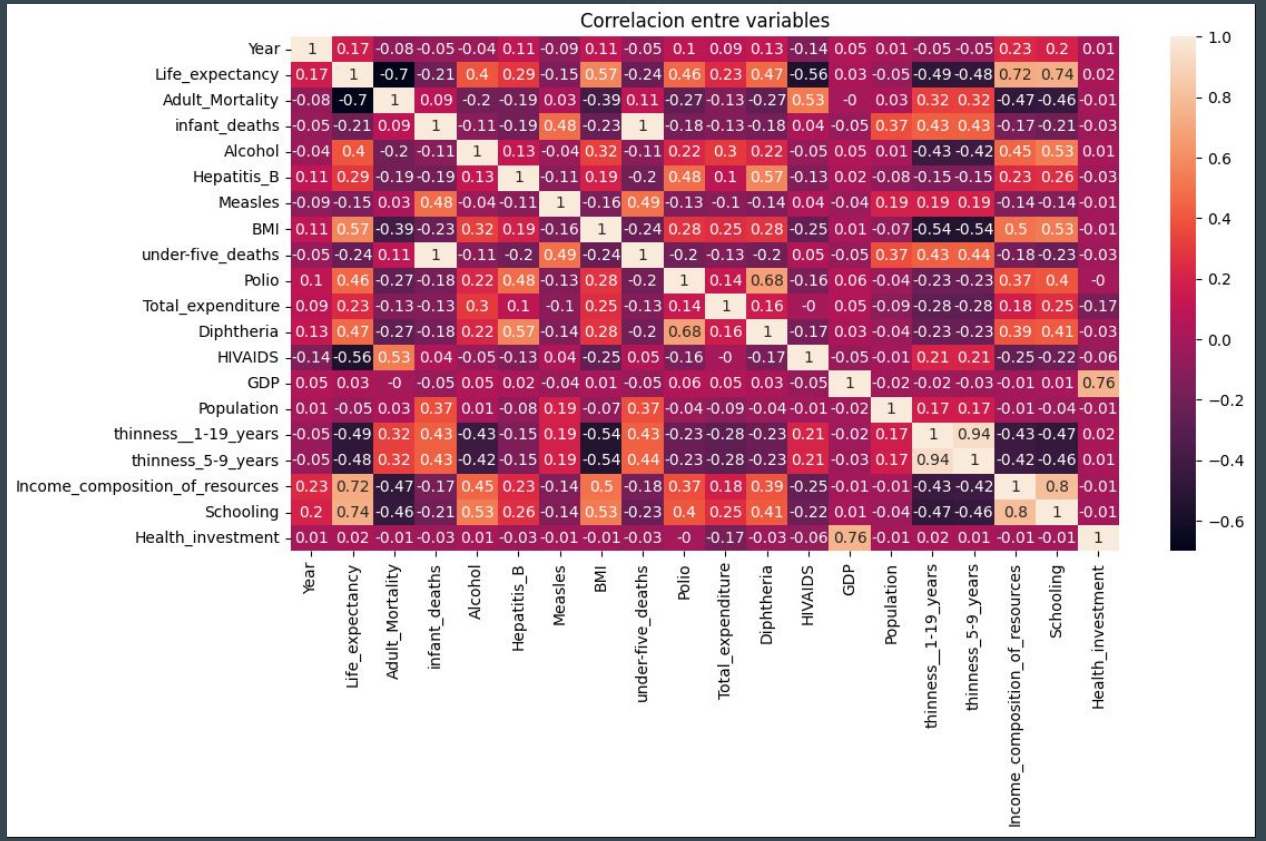


En este gráfico se puede observar la distribución de la distintas expectativas de vida para cada año.

EDA - Exploratory Data Analysis

En el siguiente análisis podemos observar la correlación entre variables.

Es importante analizar si las variables tienen mucha relación entre sí para no repetir información ni “hablar mucho de lo mismo”



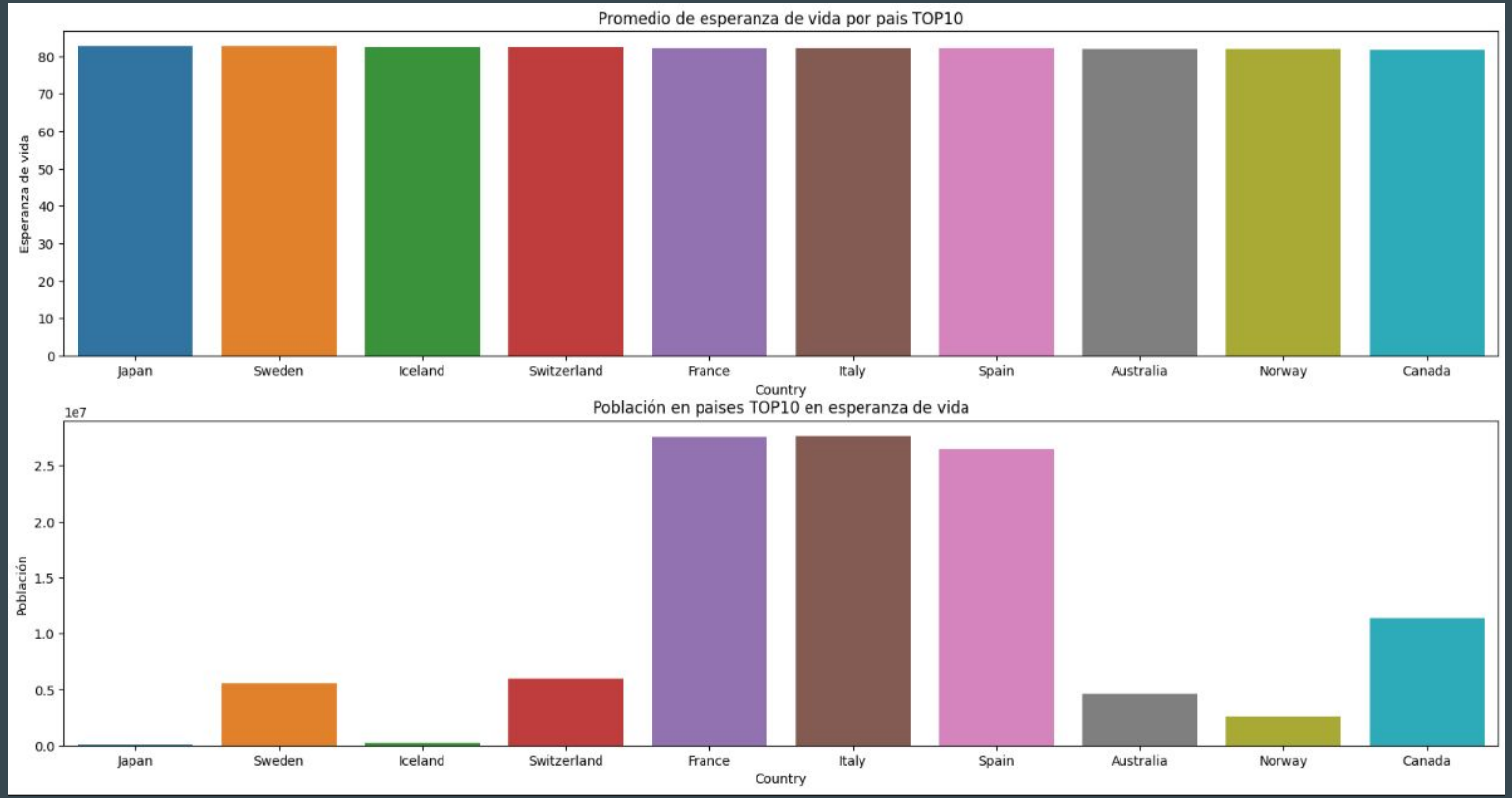
EDA - Exploratory Data Analysis

Para continuar con el análisis, se presentaron dos hipótesis:

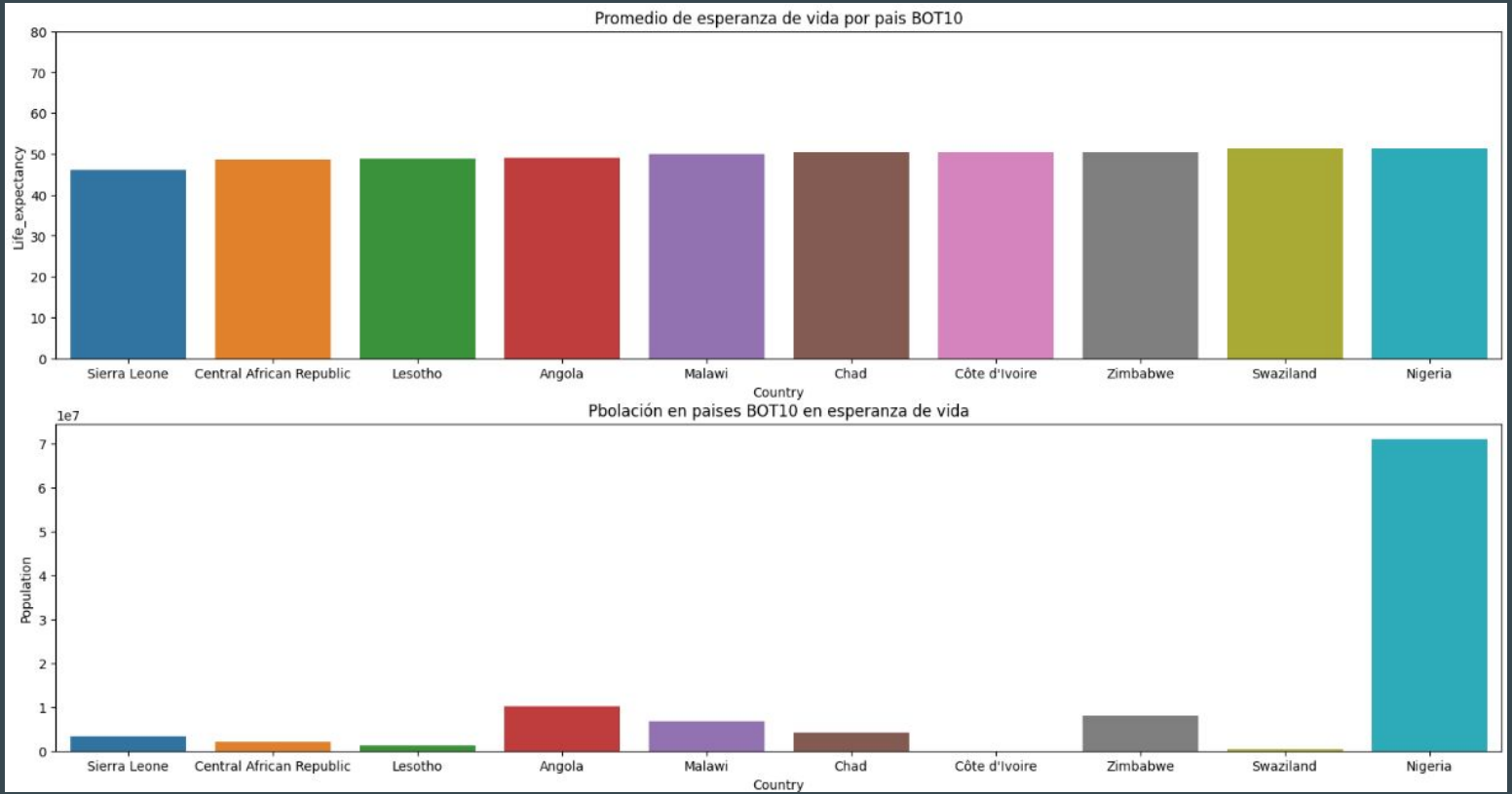
- Los países desarrollados son los que más esperanza de vida tienen, y los subdesarrollados son los que menos tienen
- En los países con mayor mortalidad son los en que hay más casos de HIV/Hepatitis B

Para la primera, agrupamos el dataframe por país y analizamos qué pasa en los 10 países con mayor expectativa de vida y en los 10 con peor. Y para la segunda, calculamos la inversión en salud dividiendo el GDP por el Total_expenditure

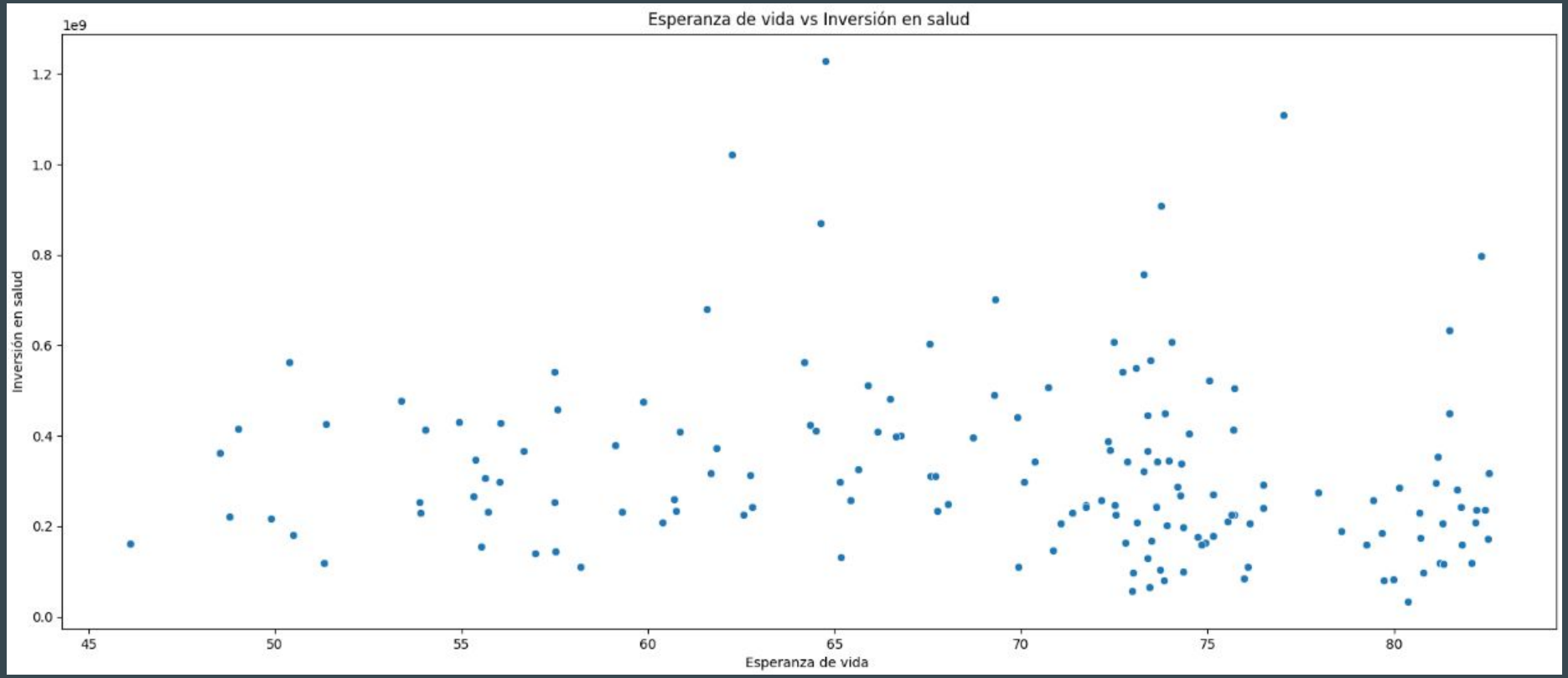
EDA - Exploratory Data Analysis



EDA - Exploratory Data Analysis



EDA - Exploratory Data Analysis



Data Wrangling

Posterior al análisis de la base de datos continuamos con la preparación del dataset observando la distribución de datos así como la existencia de campos vacíos (nan) o de datos duplicados que debamos tratar para no tener un impacto negativo en nuestro análisis para posteriormente volver a analizar cuanto se modificó su comportamiento.

La variable calculada que usamos para la Inversión en salud (GDP/Total_expenditure) la llamamos “Health_investment”

Data Wrangling

Nan y datos duplicados

Dado que tenemos una gran cantidad de valores nulos en algunos campos. Decidimos utilizar el método KNNimputer(). Con este método lo que obtenemos es un valor para cada dato faltante tomando en cuenta la media de los valores de los vecinos.

Distribución e identificación de Outliers

La correcta identificación y manejo de los outliers es esencial para garantizar la calidad y la confiabilidad de los resultados analíticos. Eliminar o tratar adecuadamente los outliers puede ayudar a mejorar la precisión de los modelos de aprendizaje automático, la estimación de parámetros y la interpretación de los resultados obtenidos.

Evaluando Modelos de Machine Learning

En el campo del Data Science, la selección y entrenamiento de modelos de aprendizaje automático efectivos es fundamental para lograr predicciones precisas y generalizadas. Para evaluar y comparar el rendimiento de diferentes modelos, es esencial utilizar métricas adecuadas que reflejen la calidad de las predicciones realizadas.

En este proyecto, nos enfocaremos en la comparación de cuatro modelos de regresión ampliamente utilizados: RandomForestRegressor, XGBRegressor, Support Vector Machine (SVM) y KNeighborsRegressor. Nuestro objetivo es determinar cuál de estos modelos ofrece el mejor rendimiento en términos de capacidad predictiva y generalización para predecir la longevidad de las personas utilizando diversas variables relacionadas.

Evaluando Modelos de Machine Learning

Al considerar las métricas para evaluar el rendimiento de los modelos, hemos optado por utilizar el error absoluto medio (MAE). Esta métrica mide la magnitud promedio de los errores en las predicciones, proporcionando una visión clara de la diferencia entre los valores reales y las predicciones del modelo. Elegimos el MAE en lugar del error cuadrático medio (MSE) porque el MAE es más intuitivo y menos sensible a valores atípicos en comparación con el MSE.

En este proyecto, hemos aplicado el PCA para justificar la selección de las columnas utilizadas en el modelo. Al realizar el análisis exploratorio de datos, hemos identificado variables que presentaban una alta correlación y, por lo tanto, podrían introducir información redundante en el modelo. Al aplicar el PCA, hemos podido seleccionar las componentes principales más relevantes que explican la mayor parte de la variabilidad de los datos y, por lo tanto, hemos optado por utilizar esas componentes en lugar de las variables originales.

Evaluando Modelos de Machine Learning - Validación simple

- Random Forest Regression Performance en el test set: MAE = 2.8594
- Support Vector Machine Regression Performance en el test set: MAE = 3.0208
- K-Nearest Neighbors Regression Performance en el test set: MAE = 3.1344
- XGBoost Reggressor Performance en el test set: MAE = 2.8903

Se observa que los modelos Random Forest Regression y XGBoost Regressor muestran un rendimiento superior en comparación con los modelos Support Vector Machine Regression y K-Nearest Neighbors Regression, en términos de la métrica de evaluación utilizada, el error absoluto medio (MAE).

Evaluando Modelos de Machine Learning - Validación cruzada

- RandomForestRegressor 0.62 de r2 promedio con una desviación estándar de 0.04
- XGBoost Reggressor 0.69 de r2 promedio con una desviación estándar de 0.06
- Support Vector Machine Regression 0.62 de r2 promedio con una desviación estandar de 0.11
- K-Nearest Neighbors Regression 0.68 de r2 promedio con una desviación estándar de 0.06

En base a estos resultados, se reafirma la elección del modelo XGBoost Regressor como el más adecuado para la predicción de la expectativa de vida en este proyecto. Su mayor rendimiento en términos de R2 promedio y desviación estándar demuestra su capacidad para capturar patrones y relaciones más precisas en los datos, lo que se traduce en mejores predicciones y una mayor capacidad de generalización.

Hiperparámetros

La optimización de hiperparámetros es un paso clave para afinar aún más el rendimiento del XGBoost Regressor. Además, la consideración de nuevas variables relevantes o la exploración de técnicas avanzadas de preprocesamiento de datos pueden contribuir a mejorar aún más la precisión de las predicciones.

Utilizamos dos métodos para obtener los hiperparámetros:

- Manual.
- RandomizedSearchCV

Utilizando estos hiperparámetros óptimos obtenidos, se entrenó un nuevo modelo XGBRegressor. Al evaluar este modelo en el conjunto de prueba, se obtuvo un MAE de 1.726. Esto indica que, en promedio, las predicciones del modelo difieren en aproximadamente 1.73 unidades de la variable objetivo real.

Conclusión

Este proyecto se ha centrado en el objetivo de predecir la expectativa de vida de un país utilizando variables relacionadas con la salud y la economía. A través del análisis y la aplicación de diversos modelos de regresión, así como técnicas de validación simple y validación cruzada, hemos obtenido resultados significativos y concluyentes.

Estos resultados tienen implicaciones significativas en la toma de decisiones y el diseño de políticas relacionadas con la calidad de vida y la salud pública. La capacidad de predecir la expectativa de vida de un país utilizando variables de salud y economía puede brindar información valiosa para la planificación de recursos, la identificación de áreas prioritarias de intervención y el monitoreo de los indicadores de bienestar.