Life Expectany

Datos socioeconómicos y salud de diferentes naciones del mundo.



**Francisco Borgno**

CoderHouse - Data Science

# INTRODUCCIÓN

El repositorio de datos del Observatorio Mundial de la Salud (GHO) de la Organización Mundial de la Salud (OMS) realiza un seguimiento del estado de salud, así como de muchos otros factores relacionados, para todos los países. El conjunto de datos relacionados con la esperanza de vida y los factores de salud de 193 países se ha recopilado del mismo sitio web del repositorio de datos de la OMS y los datos económicos correspondientes se recopilaron del sitio web de las Naciones Unidas. Entre todas las categorías de factores relacionados con la salud, solo se eligieron los factores críticos que son más representativos.

# HIPÓTESIS

Los factores como el gasto administrativo en salud, el desarrollo de campañas de vacunación y prevención de enfermedades graves, así como también el gasto individual de los individuos, sus costumbres, y el contexto económico en el que se ven envueltos, son fundamentales a la hora de pensar cual sería la expectativa de vida de cada ser. Observaremos cómo los factores socioeconómicos de cada nación estructuran la esperanza de vida de los individuos.

# PROCEDIMIENTO

Nos estructuramos una serie de consultas

1. ¿Cómo influyen los factores socioeconómicos en la esperanza de vida de las diferentes naciones? ¿Cómo influyen en los países Latinoamericanos?
2. ¿Cómo afectan los niveles de vacunación infantil? Hepatitis B, Polio, Difteria ¿Como se observa en Latinoamérica?
3. ¿Cómo influye el consumo de alcohol y el IMC (índice de masa corporal) en la esperanza de vida? ¿Y su escolarización? ¿Cómo se observan en Latinoamérica?

**Por ultimo.  
Generaremos un modelo predictivo, que seleccionara las variables claves que afecten negativa y positivamente los niveles de Esperanza de vida con la idea de contrastar con el EDA realizado.**

# 

# RESULTADOS

1. Como factores socioeconómicos entendimos el “PIB per Cápita” (PIBPC) el “Gasto porcentual” (% de gasto en materia sanitaria que desprende del PIBPC) y “Gasto total” (% de gasto de las Administraciones Sanitarias de cada nación).  
   La relación en el caso del PIBPC, y el Gasto porcentual, nos permite observar que las naciones con mejores índices (o niveles más altos) tienen niveles de esperanza de vida superiores a la media, así como también, las naciones que hacen mayores gastos en materia de salud, poseen mejores estadísticas en relación a la Esperanza de vida.
2. Decidimos categorizar los niveles de inmunización en estas tres variantes (Hepatitis B, Polio y Difteria) para entender mejor su relación con la esperanza de vida, también, lo observamos con los niveles de mortalidad infantil.  
   Una vacunación temprana, evidencia una disminución de los niveles en las muertes infantiles, como también, una buena relación con la esperanza de vida.
3. En los usos sociales el alcohol, así como también, los niveles educativos de los individuos pueden determinar la esperanza de vida que estos presentan, analizamos Alcohol, IMC (índice de masa corporal) y escolaridad.   
   Si bien, el alcohol tiene una relación negativa con la Esperanza de vida, no representa un factor fundamental en la misma (no al menos en todos los casos), el IMC sin embargo, si, podemos entenderlo como una relación directa y un factor decisivo, un nivel muy inferior de IMC, puede observarse en una esperanza de vida baja.   
   La escolaridad por su parte (los años de estudios) tiene una relación positiva, entendiéndose como un factor para una buena longevidad (quizás, por las consecuencias propias).

# Feature Selection, Modelado y Resultados

Para identificar las características más relevantes que afectan la esperanza de vida, se utilizó el método de Lasso (Least Absolute Shrinkage and Selection Operator). Aquellas características cuyo coeficiente es reducido a cero son eliminadas, mientras que las que mantienen un coeficiente distinto de cero se consideran relevantes para el modelo

En el análisis, se encontraron varias variables clave que influyen significativamente en la esperanza de vida, entre las que destacan:

* **Escolaridad y Índice de Desarrollo Humano (IDH**) como las variables con el mayor impacto positivo, sugiriendo que un mayor acceso a la educación y un desarrollo integral mejoran notablemente la calidad de vida.
* **Mortalidad Adulta y VIH/SIDA**, que resultaron ser variables con un impacto negativo significativo, lo que indica que una alta mortalidad por enfermedades crónicas e infecciosas reduce considerablemente la esperanza de vida.

2. **Modelado Predictivo: Random Forest**

El modelo de Random Forest fue seleccionado para predecir la esperanza de vida en función de las características identificadas. Random Forest es un algoritmo de ensamble que crea múltiples árboles de decisión y promedia sus resultados para hacer predicciones robustas y precisas. Este método tiene la ventaja de manejar relaciones no lineales y proporcionar medidas de la importancia de las características, lo que permite una interpretación más profunda del modelo.

Para optimizar el rendimiento del modelo, se realizó un ajuste de hiperparámetros utilizando **GridSearchCV**. Los parámetros óptimos obtenidos fueron:

* **Número de árboles (n\_estimators): 200**
* **Profundidad máxima de los árboles (max\_depth): 30**
* **Número máximo de características a considerar en cada división (max\_features): sqrt**

3. **Resultados y Rendimiento del Modelo**

El modelo optimizado de **Random Forest mostró un R² de 0.95**0 durante la validación cruzada, lo que significa que el 95% de la variabilidad en la esperanza de vida se puede explicar a partir de las características seleccionadas. Este alto coeficiente de determinación indica que el modelo es bastante preciso y robusto.

En la evaluación final, se destacaron las siguientes conclusiones:

* Las variables con mayor influencia positiva en la esperanza de vida fueron **Escolaridad e Índice de Desarrollo Humano (IDH)**. Esto refuerza la importancia de la educación y el desarrollo social para mejorar la calidad de vida.
* Las variables con mayor influencia negativa fueron **Mortalidad Adulta y VIH/SIDA**, destacando la necesidad de políticas de salud pública efectivas para reducir la mortalidad asociada con enfermedades prevenibles y mejorar la expectativa de vida en las poblaciones afectadas.

En resumen, el modelo de Random Forest ha proporcionado una comprensión sólida de los factores que influyen en la esperanza de vida, permitiendo destacar áreas clave para intervenciones políticas y sociales.