**Documentación del modelo de Machine Learning**

**Descripción del modelo:**

* Nombre del modelo: Regresión Lineal Múltiple para la Esperanza de Vida
* Objetivo del modelo: El objetivo de este modelo es predecir la esperanza de vida basándose en diferentes indicadores socioeconómicos.
* Breve resumen: Este modelo utiliza un enfoque de regresión lineal múltiple para predecir la esperanza de vida utilizando varios indicadores socioeconómicos, como el homicidio, el gasto en educación, la tasa de nacimientos, la tasa de mortalidad infantil y el PIB. Se recopilan datos de diferentes fuentes, como el Banco Mundial, y se realiza un preprocesamiento para calcular los promedios anuales de los indicadores. Luego, se entrena el modelo de regresión lineal múltiple y se evalúa su rendimiento utilizando métricas como el coeficiente de determinación (R²), el error cuadrático medio (MSE) y el error absoluto medio (MAE).

**El modelo de regresión lineal múltiple se selecciona en este caso debido a las siguientes razones:**

* Naturaleza del problema: Los datos son continuos y se presentan en función de los años. El objetivo es predecir la variable dependiente (esperanza general) utilizando múltiples variables independientes relacionadas con indicadores socioeconómicos y de salud. La regresión lineal múltiple es una opción adecuada para este tipo de problemas, ya que busca establecer una relación lineal entre las variables independientes y la variable dependiente.
* Alta correlación entre los datos: Existe una alta correlación entre los datos. El modelo de regresión lineal múltiple puede aprovechar esta correlación para modelar y predecir la variable dependiente de manera más precisa. Al incluir múltiples variables independientes correlacionadas, se pueden capturar las relaciones complejas y los efectos conjuntos de estas variables en la variable dependiente.
* Interpretabilidad: La regresión lineal múltiple es un modelo altamente interpretable. Los coeficientes estimados para cada variable independiente proporcionan información sobre la dirección y la magnitud de su influencia en la variable dependiente. Esto permite comprender y explicar cómo cada variable contribuye al resultado final.
* Eficiencia computacional: El modelo de regresión lineal múltiple es relativamente rápido de entrenar y calcular en comparación con otros modelos más complejos, como las redes neuronales. Esto puede ser beneficioso cuando se tienen grandes cantidades de datos o se requiere realizar cálculos rápidos en tiempo real.
* Suposiciones simplificadas: La regresión lineal múltiple asume una relación lineal entre las variables independientes y la variable dependiente. Si los datos cumplen con esta suposición y no hay violaciones graves de otras suposiciones, como la homocedasticidad o la ausencia de multicolinealidad, la regresión lineal múltiple puede proporcionar resultados sólidos y confiable

**Datos utilizados:**

* Fuente de datos: Los datos utilizados en este modelo fueron extraídos del Banco Mundial.
* Descripción de los datos: Se utilizaron varios conjuntos de datos relacionados con la esperanza de vida y diferentes indicadores socioeconómicos, como el gasto en educación la tasa de mortalidad infantil y el PIB. Cada conjunto de datos se almacena en un archivo CSV separado.
* Preprocesamiento de datos: Se realizaron los siguientes pasos de preprocesamiento de datos:
  + Cálculo de los promedios anuales de los indicadores utilizando la función **groupby** en pandas.
  + Transformación de la columna "x2" dividiendo los valores por 10 mil millones.
  + Transformación de la columna "x3" dividiendo los valores por 100 mil millones.

**Modelado:**

* Modelo utilizado: Regresión Lineal Múltiple
* Variables independientes (X): Los indicadores socioeconómicos promediados por año, incluyendo el gasto en educación (x1), la tasa de mortalidad infantil (x2) y el PIB (x3).
* Variable dependiente (y): Esperanza de vida

**Proceso de entrenamiento:**

* División de los datos: Los datos se dividieron en un conjunto de entrenamiento y un conjunto de prueba, utilizando un tamaño de prueba del 20% y una semilla aleatoria de 50 para mantener la reproducibilidad.
* Ajuste del modelo: Se utilizó el modelo de regresión lineal múltiple de scikit-learn (**LinearRegression()**) para ajustar el modelo a los datos de entrenamiento.
* Coeficientes e intercepto: Se obtuvieron los coeficientes de regresión y el intercepto del modelo ajustado.
* Evaluación del modelo: Se evaluó el rendimiento del modelo utilizando el coeficiente de determinación (R²) en el conjunto de prueba.

**Resultados y métricas de evaluación:**

* Coeficientes: Los coeficientes obtenidos para cada variable independiente fueron:
  + x1 (Gasto en Educación): 1.11838573
  + x2 (Tasa de Mortalidad Infantil): -0.30723327
  + x3 (PIB): -0.47070806
* Intercepto: El valor del intercepto fue 78.1986576140676.
* Score del modelo: El coeficiente de determinación (R²) del modelo en el conjunto de prueba fue de 0.9796946147223929, lo que indica un buen ajuste del modelo a los datos.
* Error cuadrático medio (MSE): El MSE entre las predicciones del modelo y los valores reales en el conjunto de prueba fue de 0.03262159381954741.
* Error absoluto medio (MAE): El MAE entre las predicciones del modelo y los valores reales en el conjunto de prueba fue de 0.1561371952185103.