Modelo para predicción de esperanza de vida

De Mónica Monserrat Martínez Vásquez | A01710965

Inteligencia artificial avanzada para la ciencia de datos I (Gpo 101)
Momento de Retroalimentación: Módulo 2 Implementación de una técnica de aprendizaje máquina sin el uso de un framework. (Portafolio Implementación)

Fecha: 31/08/2025

1. Introducción

El presente proyecto tiene el objetivo principal de desarrollar desde cero una técnica de aprendizaje máquina sin el uso de frameworks especializados.

Para ello, se implementó un modelo de regresión lineal múltiple optimizado mediante gradiente descendiente, con el propósito de predecir la esperanza de vida (Life expectancy) a partir de un conjunto de variables relacionadas con factores de salud, educación, inmunización y economía.

El dataset utilizado fue el "Life Expectancy Data", recopilado por la Organización Mundial de la Salud (WHO) y las Naciones Unidas (UN). Este conjunto de datos integra observaciones de 193 países entre los años 2000 y 2015, y está compuesto por 22 columnas y 2938 registros. La variable objetivo es la esperanza de vida al nacer (Life expectancy), mientras que las variables predictoras se dividen en cuatro grandes categorías:

- Factores de inmunización: cobertura de vacunas como Hepatitis B, Polio y Difteria.
- Factores de mortalidad: mortalidad adulta, mortalidad infantil y mortalidad en menores de cinco años.
- Factores económicos: Producto Interno Bruto (GDP), porcentaje de gasto en salud, composición del ingreso.
- Factores sociales: escolaridad promedio (Schooling), consumo de alcohol, índice de masa corporal (BMI), entre otros.

Este dataset es especialmente relevante ya que permite analizar cómo distintos determinantes de salud pública, condiciones socioeconómicas y políticas de vacunación influyen en la esperanza de vida de las poblaciones. Además, cuenta con suficientes variables fuertemente correlacionadas tanto positiva como negativamente con la variable objetivo, lo que lo convierte en un caso de estudio adecuado para el uso de regresión lineal.

El proyecto se desarrolló en fases que incluyeron la selección y análisis del dataset, la preparación de los datos mediante limpieza y normalización, la implementación del modelo desde cero con gradiente descendente, y la evaluación de métricas como MAE, R^2, bias y varianza de errores. Posteriormente, se analizó la complejidad del modelo comparando distintos números de variables para identificar el punto óptimo de predicción. Los resultados muestran que el modelo explica un 83% de la variabilidad de la esperanza de vida con un error promedio de 2.6 años, confirmando su efectividad, y que alrededor de 20 variables son suficientes para alcanzar el mejor desempeño sin incrementar innecesariamente la complejidad.

2. Selección del dataset

Para este proyecto, el objetivo fue implementar un modelo de aprendizaje máquina sin el uso de frameworks, yo escogí utilizar un modelo de regresión lineal múltiple para predecir una variable continua. Justamente por esta razón es que uno de los criterios más importantes para la selección del dataset fue que su variable objetivo tuviera una relación lineal significativa con varias de sus otras variables numéricas.

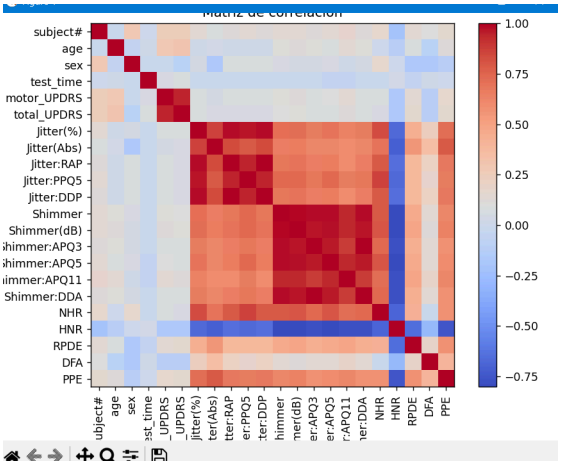
De mi primera seleccion de datasets, seleccione dos datasets que se alineaban con lo que estaba buscando y que sobretodo me llamaron la atención:

- Parkinson's telemonitoring dataset
- Life expectancy data (WHO + UN)

A través de un análisis de correlación de Pearson, observé lo siguiente:

- Dataset de Parkinson:

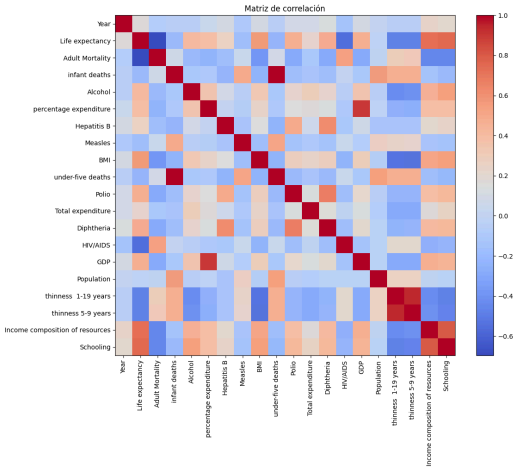
```
Correlación con la variable objetivo (motor_UPDRS):

motor_UPDRS        1.000000
total_UPDRS        0.947231
age                0.273665
subject#           0.252919
PPE                0.162433
Shimmer:APQ11      0.136560
RPDE               0.128607
Shimmer(dB)        0.110076
Shimmer            0.102349
Shimmer:APQ5       0.092105
Jitter(%)          0.084816
Shimmer:APQ3       0.084261
Shimmer:DDA        0.084260
Jitter:PPQ5        0.076291
NHR                0.074967
Jitter:DDP         0.072698
Jitter:RAP         0.072684
test_time          0.067918
Jitter(Abs)        0.050903
sex               -0.031205
DFA               -0.116242
HNR               -0.157029
```



Esto indicaba que aunque había muchas variables, pocas tenían una fuerte relación lineal con la variable objetivo (motor_UPDRS). Por tanto, era menos adecuado para un modelo de regresión lineal.

- Dataset de Life Expectancy:

```
Correlación con la variable objetivo:
Life expectancy                  1.000000
Schooling                        0.751975
Income composition of resources  0.724776
BMI                              0.567694
Diphtheria                       0.479495
Polio                            0.465556
GDP                              0.461455
Alcohol                          0.404877
percentage expenditure           0.381864
Hepatitis B                      0.256762
Total expenditure                0.218086
Year                             0.170033
Population                      -0.021538
Measles                         -0.157586
infant deaths                   -0.196557
under-five deaths               -0.222529
thinness 5-9 years              -0.471584
thinness  1-19 years            -0.477183
HIV/AIDS                        -0.556556
Adult Mortality                 -0.696359
```



Esto demostró que el conjunto de datos incluía múltiples variables fuertemente correlacionadas positiva o negativamente con la esperanza de vida, lo que lo hacía ideal para un modelo de regresión lineal. Por esta razón elegí este dataset.

3. Preparación de datos y normalización

Una de las primeras cosas que realice para preprocesar mi dataset fue eliminar las columnas no numéricas (Country y Status). Además, se eliminaron los registros con valores faltantes y se aplicó normalización estándar (z-score) a todas las variables predictoras y a la variable objetivo antes del entrenamiento. Se usó la fórmula que podemos observar en la Figura 1.

$$x_{norm} = \frac{x-\mu}{\sigma}$$

Figura 1. Fórmula de normalización (o estandarización)

Esto se hizo para que todas las variables contribuyeran al entrenamiento de manera igualitaria al calcular el gradiente. Además, ayuda a que las variables con mayor escala no afecten al entrenamiento facilitando así la convergencia del algoritmo de optimización.

Además, el dataset se dividió en un 80% para entrenamiento y 20% para prueba con unos hiperparámetros de:

- Tasa de aprendizaje (α) de 0.01, elegida por ser la más común.

- 1000 épocas, pero el entrenamiento para sí el error cae por debajo de $1e^{-2}$.

4. Implementación del modelo

El modelo desarrollado fue una regresión lineal múltiple implementada desde cero, optimizada mediante gradiente descendiente por lotes (batch gradient descent). La hipótesis lineal utilizada fue:

$$\hat{y} = \theta_0 x_0 + \theta_1 x_1 + \ldots + \theta_n x_n + b$$

Figura 2. Fórmula de hipótesis lineal múltiple

Donde $\hat{y}$ representa la predicción de la esperanza de vida a partir de las variables predictoras. Se definió como función de costo el error cuadrático medio (MSE), y el modelo se entrenó iterativamente ajustando los parámetros θ y el bias (b).

Evaluación y análisis

El desempeño se midió con métricas estándar:

- MAE (error absoluto medio), para cuantificar la magnitud promedio del error en años.
- R^2 (coeficiente de determinación), para medir la proporción de varianza explicada por el modelo.
- Bias y varianza, para diagnosticar la estabilidad y tipo de error del modelo.

Además, se implementaron herramientas gráficas de evolución del error (MSE) durante el entrenamiento y comparación entre valores reales y predichos. Como extras también se realizaron análisis de correlaciones iniciales para validar las relaciones de las variables del dataset.

5. Ejecución y análisis del entrenamiento con todas las variables (main.py)

La primera versión del entrenamiento que hice utilizó todas las variables numéricas del dataset para tener una línea base de desempeño.

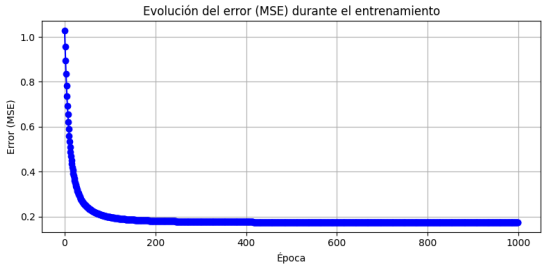


Figura 3. Gráfica del Error (MSE) por época durante el entrenamiento

La figura 3 mostró una disminución constante durante las primeras aproximadamente 200 épocas, después de las cuales el error se estabilizó, indicando convergencia exitosa del modelo.

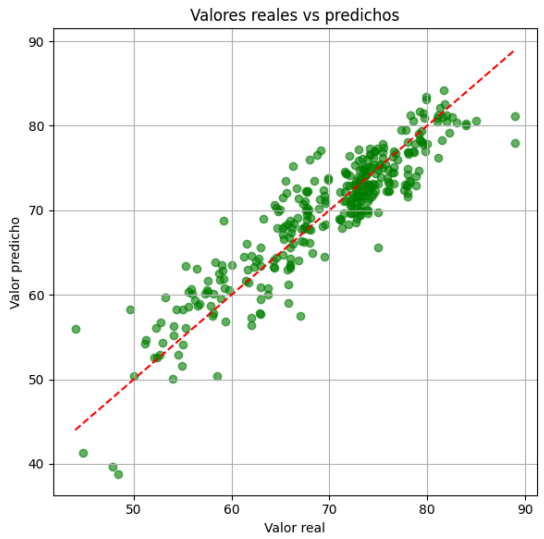


Figura 4. Gráfica valores reales vs. predichos

Además, en la figura 4 podemos observar que los puntos se alinean cercanamente a la línea de regresión, indicando un buen ajuste. Pero, se observan algunos errores que están rezagados en los valores bajos y altos de esperanza de vida, que podrían explicarse por outliers o factores no lineales.

```
Theta final: [-0.0619597010160012 3, -0.24921196023051498, 0.0
13834450449737013, -0.0667470319891927, 0.050419525982045466,
 -0.01558860412029421, 0.005281517578984783, 0.08197025662274
063, -0.05592567734447287, 0.01828981415393549, 0.01286247946
409671, 0.053113021274684855, -0.309139166969402, 0.044485504
09235253, 0.023549168640160538, -0.011138153721972079, -0.008
333329386376502, 0.23140840290317657, 0.2937983413498624]
Bias final: 0.0015898801083965488
MAE (real): 2.6390
R² (real): 0.8329
Real = 56.00 | Predicho = 60.11
Real = 59.20 | Predicho = 61.91
Real = 69.60 | Predicho = 70.87
Real = 76.00 | Predicho = 72.25
Real = 71.80 | Predicho = 74.38
```

Figura 5. Resultados del entrenamiento del modelo de regresión lineal múltiple

En cuanto al error absoluto medio (MAE) nos dio un valor de 2.6390 (ver Figura 5), indicándonos que en términos de esperanza de vida (años), en promedio, la diferencia entre los valores reales y las predicciones del modelo es de aproximadamente 2.6 años.

	Year	Life expectancy
count	2938.000000	2928.000000
mean	2007.518720	69.224932
std	4.613841	9.523867
min	2000.000000	36.300000
25%	2004.000000	63.100000
50%	2008.000000	72.100000
75%	2012.000000	75.700000
max	2015.000000	89.000000

Figura 6. Tabla de estadísticas descriptivas del dataset "Life expectancy data (WHO + UN)"

Como podemos ver en la Figura 6 la esperanza de vida oscila entre 36 y 90 años, por lo que este MAE representa solo un 3–5% del rango de los valores reales, lo que puede considerarse muy aceptable para un modelo lineal.

También, nos dio un coeficiente de determinación (R^2) de 0.8329 (ver Figura 5), significando que el modelo logra explicar aproximadamente el 83.29% de la variabilidad total de la esperanza de vida a partir de las variables predictoras. Siendo un nivel bastante alto de ajuste considerando que la relación de la esperanza de vida con los demás factores no eran perfectamente lineales. Que el valor sea alto significa que el modelo ha aprendido los patrones importantes y generaliza bien. Aun así, nos queda

aproximadamente el 17% de la variabilidad sin explicar, que puede ser por factores de dataset o interacciones no lineales entre las variables que no captura el modelo.

El modelo arrojó 19 coeficientes, es decir, los "Theta final" (ver Figura 5), donde cada uno refleja el impacto de una variable sobre la esperanza de vida al variar en una unidad (cabe destacar que es en escala normalizada) manteniendo las demás constantes. Algunos coeficientes fueron positivos, como escolaridad, ingreso y vacunaciones, lo que indica que su incremento se asocia con una mayor esperanza de vida; mientras que otros resultaron negativos, como mortalidad adulta, VIH/SIDA y desnutrición infantil, lo que nos dice que su aumento reduce la expectativa de vida.

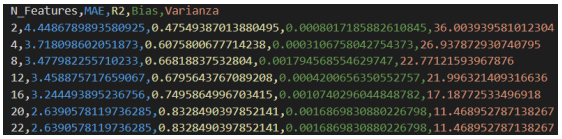
En cuanto al bias final obtenido fue de 0.0016 (ver Figura 5), como es un valor que se acerca al 0 nos respalda que el modelo está bien calibrado tras la normalización. Esto demuestra que el modelo no está desplazando las predicciones para compensar errores, sino que el aprendizaje se concentra en los coeficientes de las variables.

La varianza de los errores refleja qué tanto se dispersan las diferencias entre valores reales y predichos alrededor de su media. En este modelo, la varianza fue 11.47, lo que equivale a una desviación más o menos de 3.4 años: aunque el MAE es de 2.6, algunos errores alcanzan 3 a 4 años. Esto es un buen resultado porque indica consistencia en las predicciones, evitando errores muy grandes en unos casos y muy pequeños en otros.

6. Ejecución y evaluación con diferentes cantidades de features

Para entender mejor la relación entre la complejidad del modelo y su capacidad de predicción, implementé un segundo archivo llamado "main_mr.py". Este archivo entrena múltiples modelos de regresión lineal múltiple aumentando gradualmente el número de variables (features) utilizadas en el entrenamiento. El objetivo de la prueba fue analizar cómo se comporta el modelo al incorporar más características, buscando un equilibrio entre simplicidad y desempeño (con menores errores y mayor R^2) , además de evaluar métricas clave como MAE, R^2, bias y varianza de los errores. Otra cosa muy importante del porqué se realizó esta prueba fue para determinar si había un punto óptimo de complejidad en el modelo (por ejemplo para evitar overfitting o underfitting).

Estos resultados se registran en un archivo CSV para visualizarlos posteriormente.



N_Features	MAE	R2	Bias	Varianza
2	4.4486789893580925	0.4754938701388049 5	0.0008017185882610845	36.003939581012304
4	3.718098602051873	0.6075800677714238	0.0003106758042754373	26.937872930740795
8	3.477982255710233	0.6681883753280 4	0.0017945685546297 47	22.771215939678 76
12	3.458875717659067	0.6795643767089208	0.0004200656350552757	21.996321409316636
16	3.244493895236756	0.7495864996703415	0.0010740296044848782	17.187725334969 18
20	2.6390578119736285	0.8328490397852141	0.0016869830880226798	11.468952787138267
22	2.6390578119736285	0.8328490397852141	0.0016869830880226798	11.468952787138267

Figura 7. Métricas clave finales de múltiples modelos de regresión lineal múltiple de distintos números de variables (features)

Como podemos observar en la figura 7, al aumentar el número de variables el MAE disminuye y el R^2 aumenta, reflejando una mejora en el ajuste. Sin embargo, la ganancia en R^2 se vuelve marginal después de las 20 features, y el modelo con 22 ofrece el mismo desempeño que con 20. Esto indica que las últimas variables no aportan valor y que un modelo con 20 características logra igual poder predictivo con menor complejidad. Usando solo de 2 y 4 features contábamos con un nivel de ajuste bajo (underfit).

Conclusión

En este proyecto comprobé que la regresión lineal múltiple funciona bien para explicar la esperanza de vida a partir de distintas variables. El modelo que implementé alcanzó un R^2 de 0.83 y un MAE de 2.6 años, lo que significa que logra predecir con precisión el fenómeno. Al probar con diferentes cantidades de variables noté que con unas 20 se obtiene el mejor equilibrio entre precisión y simplicidad, sin necesidad de aumentar más la complejidad. Además, los coeficientes ajustados tuvieron sentido: escolaridad, ingreso y vacunación se relacionan con más años de vida, mientras que mortalidad adulta, VIH/SIDA y desnutrición reducen la esperanza de vida. En resumen, el modelo no solo predice bien, sino que también ayuda a entender qué factores influyen más, aunque reconozco que no alcanza a capturar relaciones no lineales o cosas más complejas.

Referencias

Life expectancy (WHO). (2018, 10 febrero). Kaggle. https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who/data

UCI Machine Learning Repository. (s. f.). https://archive.ics.uci.edu/datasets?Task=Regression&skip=30&take=10&sort=desc&orderBy=NumHits&search=&Area=Biology&Area=Climate+and+Environment&Area=Health+and+Medicine&Area=Physics+and+Chemistry

UCI Machine Learning Repository. (s. f.-b). https://archive.ics.uci.edu/dataset/189/parkinsons+telemonitoring

GeeksforGeeks. (2025, 4 agosto). *AUC ROC Curve in Machine Learning*. GeeksforGeeks. https://www.geeksforgeeks.org/machine-learning/auc-roc-curve/

1Aria - Parkinson. Escala de Evaluación Unificada (UPDRS). (s. f.). https://1aria.com/entrada/parkinson-escala-de-evaluacion-unificada-updrs

Khan Academy. (s. f.). https://www.khanacademy.org/math/multivariable-calculus/applications-of-multivariable-derivatives/optimizing-multivariable-functions/a/what-is-gradient-descent#:~:text=Gradient%20descent%20is%20an%20algorithm,like%20we've%20seen%20before.

GeeksforGeeks. (2025a, julio 23). *What is Gradient descent?* GeeksforGeeks. https://www.geeksforgeeks.org/data-science/what-is-gradient-descent/

GeeksforGeeks. (2025a, julio 23). *Model Complexity & Overfitting in Machine Learning*. GeeksforGeeks. https://www.geeksforgeeks.org/machine-learning/model-complexity-overfitting-in-machine-learning/

Model complexity influence. (s. f.). Scikit-learn. https://scikit-learn.org/stable/auto_examples/applications/plot_model_complexity_influence.html