Modelos Lineales y No Lineales para la Predicción de la Esperanza de Vida: Implementación Manual y Comparativa con Random Forest

De Mónica Monserrat Martínez Vásquez | A01710965

Inteligencia artificial avanzada para la ciencia de datos I (Gpo 101)

Fecha: 15/09/2025

1. INTRODUCCIÓN

El presente proyecto tiene como propósito desarrollar una técnica de machine learning desde cero, sin recurrir a frameworks especializados (como scikit-learn o TensorFlow), con el fin de reforzar los conocimientos prácticos adquiridos en clase. Para ello, se implementó un modelo de regresión lineal múltiple [1] optimizado mediante gradiente descendiente, orientado a predecir la esperanza de vida a partir de un conjunto de variables relacionadas con factores de salud, educación, inmunización y economía. El dataset utilizado fue el dataset de Life Expectancy [2], 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 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 desarrollo del proyecto incluyó varias etapas: la limpieza y normalización de los datos, la implementación manual del modelo con gradiente descendente, y la evaluación de su desempeño con métricas como MSE, MAE, R^2, sesgo (bias) y la varianza de los errores [3]. También se hizo un análisis de complejidad para ver cómo cambia el desempeño al usar más o menos variables, y se identificó que con aproximadamente 20 variables se obtiene el mejor resultado sin hacer el modelo innecesariamente complejo. Por último, se aplicaron técnicas de mejora como regularización L2 [4] y reducción de dimensionalidad por PCA [5], y se compararon los resultados de este modelo mejorado con los de un modelo basado en framework para evaluar sus diferencias y ventajas.

2. SELECCIÓN DEL DATASET

Uno de los primeros pasos de este proyecto consistió en elegir un conjunto de datos adecuado para implementar un modelo de IA sin el uso de frameworks. Como el modelo elegido fue una regresión lineal múltiple, el criterio principal para la selección fue que la variable objetivo tuviera una relación lineal significativa con otras variables numéricas del conjunto de datos.

En primera instancia, seleccione dos datasets que se alineaban con lo que estaba buscando y que me llamaron la atención:

- Parkinson's telemonitoring dataset [6].
- Life expectancy data (WHO + UN).

2.1 DATASET DE PARKINSON

En la figura Figura 1 y 2, indicaba que aunque había muchas variables, pocas tenían una fuerte relación lineal con la variable objetivo motor_UPDRS [7]. Por tanto este dataset, no era un conjunto de datos tan apropiado para una regresión lineal, ya que el modelo podría tener dificultades para generalizar o explicar la variabilidad de la variable objetivo.

```
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
```

Figura 1. Resultados de correlación del dataset de “Parkinson’s telemonitoring dataset”

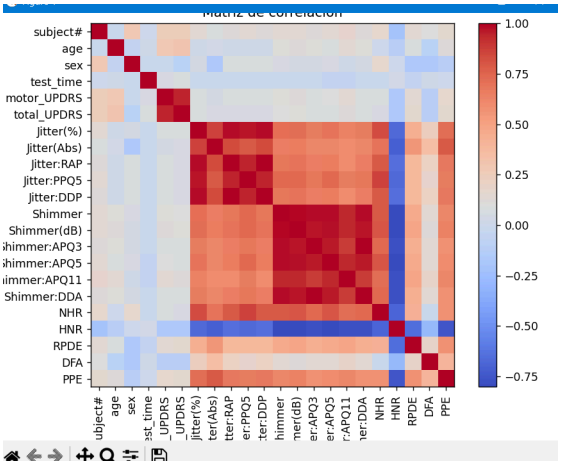

Figura 2. Mapa de correlación del dataset de “Parkinson’s telemonitoring dataset”

2.2 DATASET DE LIFE EXPECTANCY

Por otro lado, al analizar el dataset de Life Expectancy, vemos en la Figura 3 y 4, cómo es que se encontró una cantidad considerable de variables que sí tenían una correlación significativa con la esperanza de vida. Como se muestra en la matriz de correlación (véase Figura 1), variables como la escolaridad, la cobertura de vacunas o la mortalidad adulta mostraban relaciones lineales claras, lo que indicaba que este conjunto de datos era mucho más apropiado para el enfoque planteado.

```
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
```

Figura 3. Resultados de correlación del dataset de “Life expectancy data (WHO + UN)”

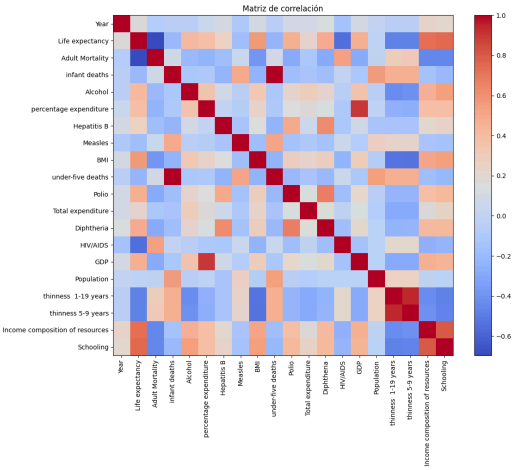

Figura 4. Mapa de correlación del dataset de “Life expectancy data (WHO + UN)”

En función de este análisis, decidí trabajar con el dataset de Life Expectancy, ya que ofrecía una mejor base para construir un modelo con buen ajuste y con potencial de mejora mediante regularización.

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 5 [8].

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

Figura 5. Fórmula de normalización (o estandarización)

Este paso fue importante para asegurarse de que todas las variables tuvieran la misma escala, evitando que aquellas con valores grandes dominaran el cálculo del gradiente. Gracias a esto, el entrenamiento se vuelve más estable y el algoritmo de optimización converge más fácilmente.

Además, el dataset se dividió en un 60% para entrenamiento, 20% para prueba y otros 20% para validación. Para esta etapa, se eligió una tasa de aprendizaje de 0.01, por ser un valor comúnmente utilizado y efectivo. Además, el entrenamiento se programó para realizar hasta 1000 épocas, pero con una condición de parada temprana si el error descendía por debajo de 1e-2, lo cual permite ahorrar tiempo de cómputo cuando el modelo ya ha aprendido lo suficiente.

4. IMPLEMENTACIÓN DEL MODELO, EJECUCIÓN Y ANÁLISIS DEL ENTRENAMIENTO (MAIN.PY)

El modelo desarrollado fue una regresión lineal múltiple implementada desde cero, optimizada mediante gradiente descendiente por lotes (batch gradient descent) [9]. La hipótesis lineal utilizada fue:

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

Figura 6. 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).

Una vez entrenado el modelo, se midió su desempeño utilizando varias 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, entre otras gráficas.

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

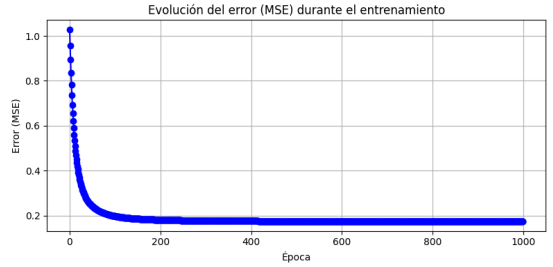

Figura 7. Gráfica del Error (MSE) por época durante el entrenamiento con todas las features

La figura 7 mostró una disminución constante durante las primeras aproximadamente 200 épocas, después de las cuales el error se estabilizó, indicando convergencia exitosa del modelo.

La versión final del entrenamiento usó las 20 variables numéricas más relevantes del dataset, seleccionadas por orden de correlación con la esperanza de vida. Esta configuración permitió alcanzar un equilibrio entre precisión y complejidad, evitando sobreajuste innecesario.

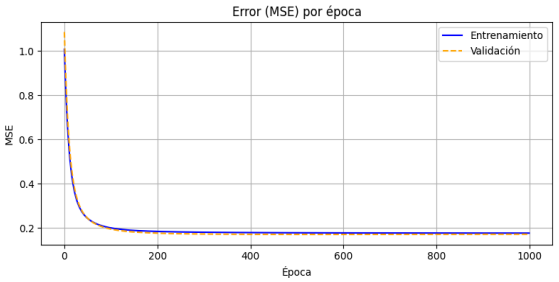

Figura 8. Gráfica del Error (MSE) por época durante el entrenamiento con 20 features

Como muestra la Figura 8, durante las primeras 200 épocas el error de entrenamiento (línea azul) y validación (línea naranja) disminuyó drásticamente, reflejando un aprendizaje eficiente. Posteriormente, ambos errores se estabilizaron sin señales de divergencia, lo que indica un ajuste adecuado (fit) y ausencia de sobreentrenamiento (overfitting).

Además, la figura 9 muestra la relación entre los valores reales de esperanza de vida y los valores predichos por el modelo. La mayoría de los puntos se agrupan cerca de la línea de identidad (roja punteada), lo cual indica un buen desempeño general.

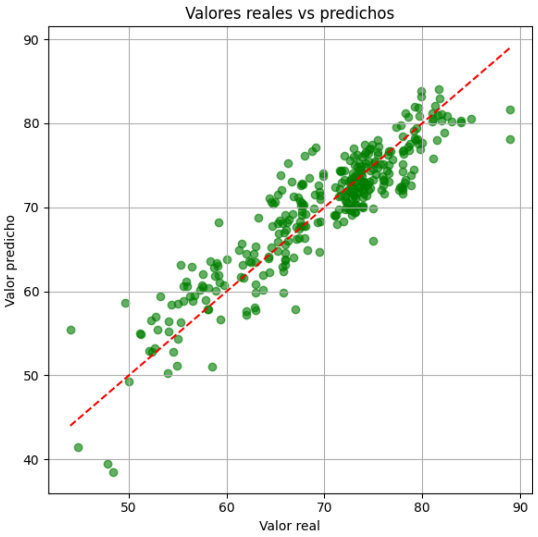

Figura 9. Gráfica valores reales vs. predichos del modelo de regresión lineal múltiple

Sin embargo, también se aprecian algunos errores mayores en los extremos del rango de esperanza de vida, lo que podría deberse a la presencia de valores atípicos o a relaciones no lineales que el modelo no puede capturar.

```
Theta final: [-0.06719344783546659, -0.22483380847912182, 0.0158
17734597123684, -0.06623170747972314, 0.0613669866921637, -0.013
51022641089524, 0.010501566274464155, 0.08826238531858113, -0.05
522250757731511, 0.02698253583282713, -0.0017636627310506944, 0.
05627709863237785, -0.32598226895476745, 0.03941637102115888, 0.
014084217702636981, -0.011268678982745686, -0.010181485236898538
, 0.20903466954833338, 0.3061901180207729]
Bias final: 0.009937461012539861
MAE (real): 2.6744
R² (real): 0.8307
Bias: -0.1382
Varianza: 11.6048
Real = 56.00 | Predicho = 60.58
Real = 59.20 | Predicho = 61.96
Real = 69.60 | Predicho = 71.01
Real = 76.00 | Predicho = 72.57
Real = 71.80 | Predicho = 74.43
```

Figura 10. Resultados del entrenamiento del modelo de regresión lineal múltiple

En cuanto a las métricas obtenidas (Figura 10), el modelo alcanzó un error absoluto medio (MAE) de 2.6744, 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.

Considerando que la esperanza de vida en el dataset oscila entre 36 y 90 años (Figura 11), este MAE representa un 3–5% del rango total, lo cual es razonable para un modelo lineal.

	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 11. Tabla de estadísticas descriptivas del dataset “Life expectancy data (WHO + UN)”

También, nos dio un coeficiente de determinación (R^2) de 0.8307 (ver Figura 10), significando que el modelo logra explicar aproximadamente el 83.07% 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.

En cuanto al bias final obtenido fue de −0.1382 (ver Figura 10), lo cual indica una ligera subestimación sistemática del modelo. Aunque el valor es pequeño, su signo negativo nos dice que las predicciones tienden a estar un poco por debajo de los valores reales. Este bias bajo, aunque negativo, es suficientemente cercano a cero como para considerar que el modelo está razonablemente bien calibrado.

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 11.6048, 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 sugiere una varianza moderada, lo que significa que el modelo mantiene una consistencia aceptable en sus predicciones sin generar errores extremos.

El modelo aprendió 19 coeficientes, correspondientes a las 20 variables predictoras (sin contar el bias). Estos coeficientes representan el peso de cada variable en la predicción, (ojo que esto es considerando que las variables fueron normalizadas previamente). Por ejemplo, podemos ver variables con

coeficiente positivo como escolaridad, ingresos y vacunaciones (polio, difteria, hepatitis B), lo cual nos dice que al aumentar estas variables incrementa la esperanza de vida. Y tambien podemos ver variables con coeficiente negativo como VIH/SIDA, mortalidad adulta y desnutrición infantil, las cuales sí incrementaran disminuiría la esperanza de vida.

Esto tiene sentido con el conocimiento general de salud pública y da evidencia de que el modelo está capturando relaciones reales en los datos, incluso sin técnicas avanzadas como árboles de decisión o redes neuronales.

5. EJECUCIÓN Y EVALUACIÓN CON DIFERENTES CANTIDADES DE FEATURES (MAIN_MR.PY)

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. El objetivo de este análisis fue identificar cómo la complejidad del modelo influye en su desempeño, y determinar si existe un punto óptimo donde se maximice el ajuste sin tener underfitting (modelo demasiado simple) o overfitting (modelo demasiado complejo).

Cada modelo fue evaluado tanto en el conjunto de prueba como en un conjunto de validación cruzada, registrando las métricas clave: MAE, R², Bias y Varianza. Estos resultados se registran en un archivo CSV para visualizarlos posteriormente.

```
N_Features,MAE_Test,R2_Test,MAE_Val,R2_Val,Bias,Varianza
2,4.459227926084685,0.47762616378665046,4.603707830188851,0.5169953964751642,0.05425560243751038,35.85596849739023
4,3.748524359699227,0.6088467754545899,4.026660622137051,0.5826443492416185,0.04849390500156828 4,26.848780903244787
8,3.493045730393171 2,0.6680039153216543,3.6344799535481513,0.6566077881905923,-0.08746483438472874,22.782577159969147
12,3.5063464452198843,0.6747815981358696,3.6349347176960 45,0.6615555514487099,-0.02299690712052705,22.324437094482963
16,3.2973946354780517,0.7435565045690573,3.4267858909666 58,0.7586191329101242,-0.04803190765374604,17.601531683434136
20,2.6743953726773 55,0.8306696571031257,2.7964406983958 08,0.8419200227482 27,-0.13822643649363 19,11.60475642612682
22,2.6743953726773 55,0.8306696571031257,2.7964406983958 08,0.8419200227482 27,-0.13822643649363 19,11.60475642612682
```

Figura 12. Métricas clave finales de múltiples modelos de regresión lineal múltiple de distintos números de variables (features)

Como podemos observar en la figura 12, 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).

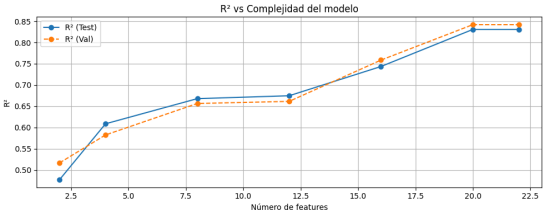

Figura 13. Desempeño del modelo con distinto número de variables (R2)

La Figura 13 muestra cómo el coeficiente de determinación R^2 mejora conforme se agregan más variables. Sin embargo, el crecimiento es rápido al inicio, pero luego se estabiliza: al pasar de 8 a 20 features, el incremento es moderado; y de 20 a 22, no hay mejora. Esto sugiere que las variables adicionales no aportan valor predictivo extra. Además, este patrón es característico de un modelo que va dejando atrás el underfit (modelo demasiado simple), se acerca al ajuste óptimo (fit) y evita el overfit al no incorporar variables irrelevantes.

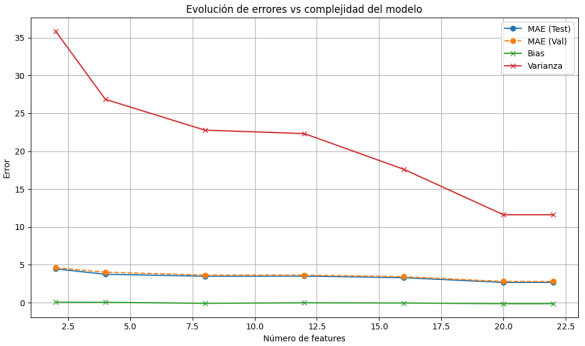

Figura 14. Comparación de métricas: MAE, Bias y Varianza vs # features

La Figura 14 complementa este análisis mostrando cómo disminuyen el MAE y la varianza al usar más features. En los modelos con 20 y 22 variables se observa un bias negativo, lo cual implica una subestimación sistemática de la variable objetivo. Pero a que me refiero cuando digo “Subestimación sistemática” bueno esto quiere decir que, en promedio, las predicciones tienden a estar por debajo del valor real de la esperanza de vida. Esta tendencia puede surgir cuando el modelo, aun teniendo buena capacidad de generalización, no logra capturar del todo ciertos patrones residuales del comportamiento de los datos. Sin embargo, dado que el margen de subestimación es pequeño y el resto de métricas como R^2 y MAE se mantienen en niveles óptimos, este sesgo leve es

considerado aceptable. Aunque este sesgo es pequeño, su presencia constante sugiere que algunos patrones residuales no están siendo completamente capturados por el modelo.

Este comportamiento es preferible frente a una alta varianza, ya que un bias leve y estable puede corregirse posteriormente mediante calibración o ajuste fino. En cambio, un modelo con alta varianza sería más inestable y difícil de corregir. Así, el modelo con 20 variables puede considerarse como el punto óptimo, logrando un buen ajuste (fit), bajo error, varianza aceptable y sin caer en complejidad innecesaria.

6. CONCLUSIÓN ANÁLISIS DE ALGORITMO DE MACHINE LEARNING SIN USO DE FRAMEWORK

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, incluir más variables (hasta 22) no mejora el desempeño, lo que sugiere que el modelo está en su punto ideal de ajuste sin incurrir en overfitting.. 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. Este análisis es fundamental, ya que nos proporciona una línea base sólida para aplicar posteriormente técnicas de regularización y comparar su efecto sobre el ajuste, sesgo y varianza del modelo.

7. IMPLEMENTACIÓN, EJECUCIÓN Y ANÁLISIS DEL MODELO RANDOM FOREST (IMPLEMENTACIÓN FRAMEWORK)

Después de evaluar el modelo lineal implementado desde cero, se entrenó un modelo con una técnica más avanzada: Random Forest Regressor, utilizando el módulo scikit-learn. Esta técnica de ensamblado se basa en la combinación de múltiples árboles de decisión entrenados sobre subconjuntos distintos de los datos para reducir la varianza del modelo y mejorar la precisión de las predicciones. A diferencia de una regresión lineal tradicional, Random Forest no asume relaciones lineales ni requiere normalización de datos, y tiene la capacidad de modelar relaciones complejas y no lineales entre las variables predictoras y la variable objetivo.

La evaluación se realizó sobre los mismos conjuntos de entrenamiento, validación y prueba que se usaron para el modelo manual, y con las mismas 20 variables seleccionadas por orden de correlación con la esperanza de vida.

```
--- Resultados Random Forest ---
MAE (test): 1.1746
R²  (test): 0.9555
MAE (val):  1.1848
R²  (val):  0.9603
Bias:       -0.1034
Varianza:   3.0425
```

Figura 15. Resultados del entrenamiento del modelo usando Random Forest

Como podemos observar en la figura 15, el Error absoluto medio (MAE), nos da el valor de 1.1746. Esto nos dice que, en promedio, las predicciones del modelo difieren del valor real en tan solo 1.17 años. Comparado con el rango del dataset (36 a 90 años de esperanza de vida), representa un error del aprox. 2%, que es bastante bajo. Este resultado demuestra una alta precisión del modelo.

El coeficiente de determinación (R^2) nos dio de 0.9555, lo que indica que el modelo es capaz de explicar el 95.55% de la variabilidad total de la esperanza de vida en los datos de prueba. Y en validación un 0.9603, lo cual confirma que el modelo tiene un ajuste muy alto (fit), sin señales de sobreajuste (overfitting).

En cuanto al sesgo (Bias) tenemos que es un sesgo negativo de -0.1034. El hecho de que el bias sea negativo nos dice que el modelo tiende a subestimar un poco los valores de esperanza de vida (por aproximadamente 0.1 años). Pero, al ser tan pequeño, no representa un problema de calibración significativo.

También, nos dió una varianza de los errores de 3.0425, lo cual indica que la dispersión de los errores alrededor de su media es baja. Como podemos recordar en el modelo lineal manual la varianza fue de 11.60, lo que expresa que Random Forest comete errores más consistentes y menos extremos.

La Figura 16 muestra la relación entre los valores reales y los predichos por el modelo. A diferencia del modelo manual, los puntos se alinean de forma muy cercana a la línea de identidad (roja punteada), lo que indica una alta precisión incluso en los casos extremos. No se observan errores sistemáticos ni agrupamientos dispersos, lo cual refuerza el diagnóstico de bajo bias y baja varianza.

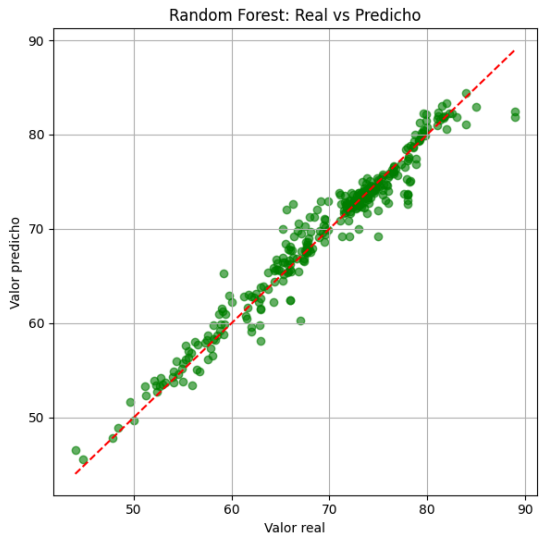

Figura 16. Gráfica valores reales vs. predichos del modelo usando Random Forest

En cuanto a la interpretación de las variables, la Figura 17 muestra la importancia relativa de cada una según el modelo de Random Forest. A diferencia de la regresión lineal, donde los coeficientes tienen un significado directo, en este caso la importancia se calcula a partir de la contribución de cada variable a la reducción del error.

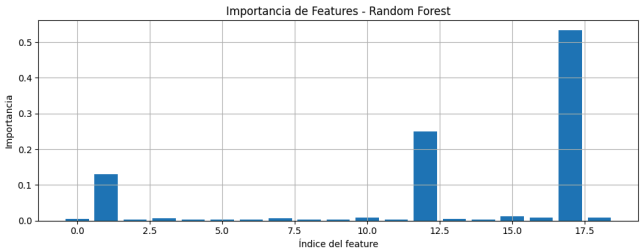

Figura 17. Importancia de las variables en Random Forest

El modelo asignó mucha importancia a solo tres variables:

- Schooling (Escolaridad): índice 17
- Income composition of resources (Composición del ingreso): índice 12
- BMI (Índice de masa corporal): índice 1

Las variables más importantes coinciden con las que habíamos visto anteriormente cuando checamos las correlaciones (analizar_correlaciones.py). Específicamente con las que tenían alta correlación positiva con la esperanza de vida.

El resto de las variables, aunque fueron utilizadas, aportan menos valor al modelo. Esto refuerza la idea de que una selección de variables más reducida, enfocada en las más predictivas, podría mantener buenos resultados con menor complejidad computacional.

8. CONCLUSIÓN ANÁLISIS DE ALGORITMO DE MACHINE LEARNING CON USO DE FRAMEWORK

El modelo Random Forest mostró un desempeño superior al modelo manual en todos los aspectos evaluados: menor MAE, mayor R^2, menor varianza de errores y un sesgo prácticamente nulo. Además, mantuvo consistencia entre los conjuntos de validación y prueba, lo que evidencia su capacidad de generalización. Si bien su principal desventaja es la menor interpretabilidad respecto a un modelo lineal, su habilidad para capturar relaciones no lineales y manejar automáticamente interacciones complejas lo convierte en una herramienta sumamente poderosa para este tipo de problemas.

9. COMPARATIVA ENTRE MODELO MANUAL Y MODELO CON FRAMEWORK

Después de analizar en profundidad ambos enfoques, se realizó una comparación directa entre el modelo manual (regresión lineal múltiple) y el modelo basado en framework (Random Forest), considerando métricas clave de desempeño, comportamiento visual de las predicciones y capacidad de generalización.

Métrica	Modelo manual (Regresión lineal)	Modelo Framework (Random Forest)
MAE (Test)	2.6744	1.1746
R^2(Test)	0.8307	0.9555
Bias (Test)	-0.1382	-0.1034
Varianza (Test)	11.6048	3.0425
MAE (Validación)	2.7964	1.1848
R^2(Validación)	0.8419	0.9603

Figura 18. Tabla comparativa de métricas modelo manual vs. modelo con framework

La Figura 14 muestra una tabla comparativa con los resultados obtenidos. En términos de precisión, el modelo con framework muestra una mejora considerable: el MAE en prueba se redujo de 2.67 a 1.17 años, lo que implica una disminución superior al 50% en el error promedio. Este resultado indica que el modelo Random Forest es más eficaz al predecir la esperanza de vida, cometiendo errores significativamente menores.

Además, mientras que el modelo manual explicó aprox. 83% de la variabilidad en los datos, el modelo con framework logró 95.5%, lo cual significa que detectó patrones complejos y no lineales que el modelo manual no pudo. Esto es bastante importante para este caso ya que el dataset que utilice puede tener relaciones no lineales o interacciones entre sí que pueden ser importantes.

Ahora, hablando de la varianza de los errores del modelo Random Forest fue de solo 3.04, en comparación con 11.60 del modelo manual. Esto significa que el modelo con framework comete errores más "homogéneos" o uniformes por así decirlo y menos extremos. En otras palabras, el modelo con framework es más estable.

Ambos modelos presentan un sesgo ligeramente negativo, lo que implica una tendencia leve a subestimar la esperanza de vida. Sin embargo, el sesgo es pequeño en ambos casos (menor a 0.14), lo que demuestra una buena calibración en general. En este aspecto, los dos están relativamente igual.

Otra cosa que podemos destacar es que comprando las gráficas de predicción real contra la predicha, en la regresión lineal manual, los puntos se alinean bien con la línea de identidad, pero hay más dispersión, especialmente en los extremos del rango. Y en el caso del modelo Random Forest, los puntos se alinean casi perfectamente con la línea de identidad, con mínima dispersión y errores reducidos, incluso en los casos extremos.

En conclusión, aunque el modelo manual fue útil para entender el problema, obtener un baseline interpretable y desarrollar desde cero los conceptos fundamentales, el modelo del framework presentó un desempeño superior en todos los aspectos evaluados: menor error, mayor ajuste (fit), bajo bias, baja varianza, y mejor generalización. Por tanto, para una aplicación práctica de predicción de esperanza de vida, el modelo basado en Random Forest sería el más adecuado.

10. MEJORA DEL MODELO: APLICACIÓN DE RIDGE (L2) Y EVALUACIÓN DE RESULTADOS

Para explorar oportunidades de mejora en el modelo manual de regresión lineal múltiple, se implementó la técnica de regularización Ridge (también conocida como L2). Esta técnica modifica la función de costo al agregar un término de penalización proporcional al cuadrado de los coeficientes, con el objetivo de reducir la varianza de los errores y mejorar la capacidad de generalización del modelo. La regularización L2 es especialmente útil cuando hay riesgo de sobreajuste (overfitting), ya que controla la magnitud de los coeficientes, evitando que el modelo dependa excesivamente de ciertas variables.

Aunque los resultados iniciales del modelo manual no mostraban señales claras de sobreajuste ni subajuste (underfitting), se decidió aplicar esta técnica como prueba diagnóstica, para confirmar el buen ajuste del modelo y evaluar si existía alguna mejora marginal posible.

Lo que se hizo fue una versión modificada de la función de costo (MSE_ridge) y del algoritmo de entrenamiento (update_ridge) que incorpora un término de penalización proporcional a la magnitud de los coeficientes. Se evaluó el desempeño del modelo entrenado con valores de λ (lambda) entre 0.01 y 10.0, y se compararon los resultados con el modelo original (sin regularización, λ=0.0).

A continuación se muestran los resultados obtenidos con λ = 10.0:

Métrica	**Sin regularización**	**Con L2 (λ = 10.0)**
MAE	2.6744	2.6744
R^2	0.8307	0.8308
Bias	−0.1382	−0.1373
Varianza	11.6048	11.5938

Figura 19. Tabla comparativa de métricas modelo manual sin vs con regularización L2

Los resultados muestran que las métricas se mantuvieron prácticamente constantes: el MAE permaneció en 2.6744, el R^2 apenas varió (de 0.8307 a 0.8308), el bias cambió ligeramente de −0.1382 a −0.1373, y la varianza de errores se redujo solo de 11.6048 a 11.5938. Estos cambios tan pequeños indican que el modelo original ya estaba adecuadamente ajustado y que no presentaba overfitting que pudiera corregirse mediante L2.

En conclusión, aunque la regularización L2 no mejoró significativamente las métricas del modelo, su aplicación permitió confirmar el buen ajuste alcanzado en la versión original, y validó que el comportamiento del modelo era estable y bien generalizado.

11. MEJORA DEL MODELO: APLICACIÓN DE HUBER + L2 + PCA Y EVALUACIÓN DE RESULTADOS

Como parte del proceso de mejora del modelo manual, se diseñó una versión más robusta que combinó tres técnicas: función de pérdida Huber, regularización

L2 y reducción de dimensionalidad mediante PCA. El objetivo de esta versión fue reducir errores extremos, mejorar la generalización del modelo y simplificar su complejidad sin perder capacidad predictiva.

En primer lugar, se reemplazó la función de costo MSE por la pérdida Huber, que combina lo mejor del MAE (robustez ante outliers) y del MSE (sensibilidad en errores pequeños), ayudando al modelo a no verse tan afectado por valores atípicos durante el entrenamiento. Esto es útil especialmente cuando se trabaja con fenómenos reales como la esperanza de vida, donde ciertos países pueden representar outliers importantes.

También se mantuvo la regularización L2 (Ridge) para penalizar coeficientes grandes, favoreciendo soluciones más simples y generalizables. Y finalmente, se aplicó una transformación PCA para reducir las 20 variables originales a 15 componentes principales, conservando el 98.36% de la varianza total. Esto permitió eliminar redundancias, acelerar el entrenamiento y evitar problemas de multicolinealidad.

```
--- Resultados ---

Entrenamiento (Train):
MAE: 2.8431 | R²: 0.8196 | Bias: -0.0211 | Varianza: 14.0605
Validación (Validation):
MAE: 2.7175 | R²: 0.8461 | Bias: -0.3281 | Varianza: 12.8051
Prueba (Test):
MAE: 2.6308 | R²: 0.8322 | Bias: -0.1886 | Varianza: 11.4854
Varianza explicada por los primeros 15 componentes: 98.36%
Varianza acumulada por cada componente:
PC1: 29.18%
PC2: 43.83%
PC3: 53.54%
PC4: 61.41%
PC5: 68.11%
PC6: 73.42%
PC7: 77.81%
PC8: 81.98%
PC9: 85.42%
PC10: 88.14%
PC11: 90.79%
PC12: 92.97%
PC13: 94.95%
PC14: 96.76%
PC15: 98.36%
Real = 56.00 | Predicho = 60.56
Real = 59.20 | Predicho = 61.31
Real = 69.60 | Predicho = 71.27
Real = 76.00 | Predicho = 72.39
Real = 71.80 | Predicho = 73.78
```

Figura 20. Resultados del modelo mejorado Huber + L2 + PCA

Los resultados finales pueden observarse en la Figura 20, donde se reportan las métricas obtenidas. El error absoluto medio (MAE) en el conjunto de prueba fue de 2.63 años, una mejora leve respecto al modelo manual original (que era de 2.67). El R^2 se mantuvo alto (alrededor de 0.83 en todos los conjuntos), lo que indica que el modelo sigue explicando una gran parte de la variabilidad de la esperanza de vida. Además, el bias fue negativo, aunque pequeño, lo que sugiere una ligera tendencia a subestimar las predicciones, algo que ya habíamos visto antes. En cuanto a la varianza de errores, se redujo de 11.60 a 11.48, reflejando una mejora en la estabilidad del modelo.

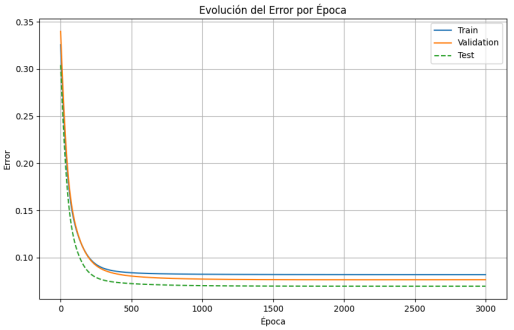

Figura 21. Evolución del error (MSE) por época

En la figura 21, la curva muestra una convergencia estable y progresiva tanto en entrenamiento como en validación, sin señales de overfitting. De hecho, el error de validación se mantiene apenas por debajo del de entrenamiento, lo cual es una buena señal de que el modelo está generalizando correctamente.

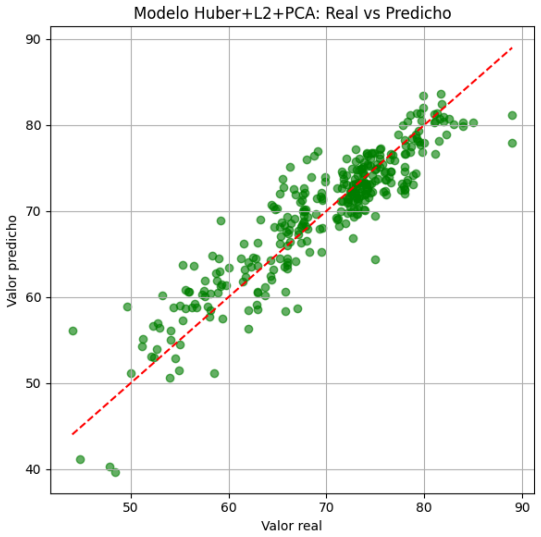

Figura 22. Gráfica valores reales vs. predichos del modelo mejorado Huber + L2 + PCA

Por último, la Figura 22 presenta la gráfica de valores reales vs. predichos. En ella se observa que la mayoría de las predicciones siguen de cerca la línea de referencia (y = x), con una alineación bastante buena en todo el rango de valores. Aunque persiste cierta subestimación en algunos casos, la tendencia general se mantiene bien capturada, lo que demuestra que el modelo logra una representación fiel del fenómeno, incluso después de aplicar técnicas de regularización y reducción de dimensionalidad.

12. COMPARATIVA Y ANÁLISIS FINAL

Métrica	Modelo Manual (Lineal)	Modelo Manual Mejorado (L2)	Modelo Manual Mejorado (Huber + L2 + PCA)	Modelo Framework (Random Forest)
MAE (Test)	2.6744	2.6744	2.6308	1.1746
R² (Test)	0.8307	0.8308	0.8322	0.9555
Bias (Test)	-0.1382	-0.1373	-0.1886	-0.1034
Varianza (Test)	11.6048	11.5938	11.4854	3.0425
MAE (Val)	2.7964	2.79 (similar al base)	2.7175	1.1848
R² (Val)	0.8419	0.842 (similar al base)	0.8461	0.9603

Figura 23. Tabla comparativa de métricas modelo manual vs modelo manual con L2 vs modelo manual con Huber + L2 + PCA vs Modelo Framework

Para cerrar el proyecto, se realizó una comparativa directa entre todas las versiones desarrolladas: el modelo lineal manual original, sus dos variantes mejoradas (una con regularización L2 y otra con Huber + L2 + PCA), y el modelo más avanzado basado en Random Forest usando un framework. En la Figura 23 se muestra una tabla resumen con las métricas clave de cada uno.

El modelo lineal manual sin modificaciones logró un MAE de 2.67 años y un R^2 de 0.8307 en el conjunto de prueba, lo cual es un resultado bastante bueno considerando su simplicidad. También reflejó un bias ligeramente negativo (−0.1382) y una varianza de errores relativamente alta (11.60), lo que sugiere que si bien el modelo generalizaba razonablemente bien, también presentaba una cierta dispersión en sus errores. En validación, sus métricas se mantuvieron estables, con un MAE de 2.79 y un R^2 de 0.8419.

Posteriormente, se aplicó la técnica de regularización L2 (Ridge), esperando reducir posibles signos de overfitting. Sin embargo, los resultados fueron prácticamente idénticos a los del modelo base: no hubo mejoras significativas ni en MAE, ni en R^2, ni en bias o varianza. Esto confirmó que el modelo original ya estaba bien ajustado y que no existía un problema de sobreajuste que pudiera ser corregido con regularización. En otras palabras, esta técnica no fue perjudicial, pero tampoco necesaria en ese momento.

La tercera versión exploró una estrategia más sofisticada, combinando la función de pérdida Huber, la regularización L2 y una reducción de variables mediante PCA (15 componentes que conservan el 98.36% de la varianza). Esta versión logró una mejora modesta pero clara: el MAE bajó a 2.63 años, el R^2 subió a 0.8322, y la varianza se redujo a 11.48, reflejando una mayor estabilidad en los errores. Aunque el bias se volvió un poco más negativo (−0.1886), se mantuvo dentro de un rango aceptable. Además, este modelo fue el que mejor generalizó entre los modelos manuales, como se observa también en sus métricas de validación (MAE de 2.71 y R^2 de 0.8461). El uso de PCA también ayudó a reducir redundancias y acelerar el entrenamiento, sin perder capacidad predictiva.

Por último, el modelo desarrollado con Random Forest dentro del framework scikit-learn fue el que obtuvo el mejor desempeño general. En el conjunto de prueba, logró un MAE de tan solo 1.17 años y un R^2 de 0.9555, lo cual implica que fue capaz de explicar casi el 96% de la variabilidad de la esperanza de vida. También redujo la varianza de los errores a 3.04, mostrando predicciones más consistentes y sin extremos. Su bias fue el más bajo (−0.1034), indicando una buena calibración. En validación, incluso mejoró ligeramente, con R^2 de 0.9603. Esto demuestra que el modelo no solo fue preciso, sino que generalizó muy bien a datos no vistos, lo cual es clave en problemas reales.

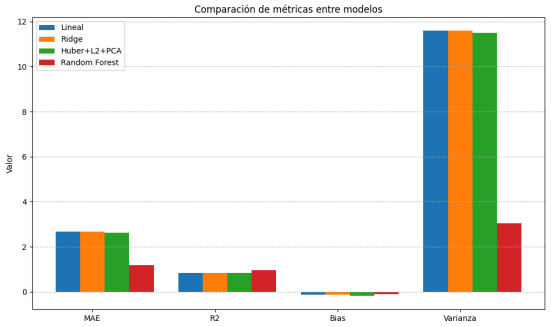

Figura 24. Gráfica comparativa de métricas modelo manual vs modelo manual con L2 vs modelo manual con Huber + L2 + PCA vs Modelo Framework

En conclusión (Figura 24), el análisis comparativo muestra que el modelo lineal manual es útil como punto de partida y para comprender los determinantes de la esperanza de vida. Sin embargo, las mejoras implementadas con Huber, L2 y PCA lograron una versión más robusta y ligeramente más precisa dentro de la misma familia de algoritmos. Aun así, el Random Forest se posiciona como la técnica con mejor desempeño en este caso, reduciendo el error más de un 50% y elevando la capacidad explicativa a niveles cercanos al 96%.

Referencias

[1] M. Fernández-Delgado, M. S. Sirsat, E. Cernadas, S. Alawadi, S. Barro, y M. Febrero-Bande, «An extensive experimental survey of regression methods», *Neural Netw.*, vol. 111, pp. 11-34, mar. 2019, doi: 10.1016/j.neunet.2018.12.010.

[2] S. Dattani, L. Rodés-Guirao, H. Ritchie, E. Ortiz-Ospina, y M. Roser, «Life Expectancy». Kaggle, 2023. [En línea]. Disponible en: https://www.kaggle.com/datasets/maryalebron/life-expectancy-data

[3] N. V. Thieu, «PerMetrics: A Framework of Performance Metrics forMachine Learning Models», *J. Open Source Softw.*, vol. 9, n.º 95, p. 6143, mar. 2024, doi: 10.21105/joss.06143.

[4] Y. Li, M. Li, y L. Zhang, «Evolutionary polynomial regression improved by regularization methods», *PLOS ONE*, vol. 18, n.º 2, p. e0282029, feb. 2023, doi: 10.1371/journal.pone.0282029.

[5] I. T. Jolliffe y J. Cadima, «Principal component analysis: a review and recent developments», *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.*, vol. 374, n.º 2065, p. 20150202, abr. 2016, doi: 10.1098/rsta.2015.0202.

[6] M. L. Athanasios Tsanas, «Parkinsons Telemonitoring». UCI Machine Learning Repository, 2009. doi: 10.24432/C5ZS3N.

[7] S. Fahn y R. L. Elton, «Unified Parkinson's disease rating scale». [En línea]. Disponible en: https://getm.sen.es/profesionales/escalas-de-valoracion/26-getm/escalas-de-valoracion/88-unified-parkinson-s-disease-rating-scale-updrs

[8] Google, «Numerical data: Normalization». [En línea]. Disponible en: https://developers.google.com/machine-learning/crash-course/numerical-data/normalization

[9] O. F. Razzouki, A. Charroud, Z. E. Allali, A. Chetouani, y N. Aslimani, «A Survey of Advanced Gradient Methods in Machine Learning», en *2024 7th International Conference on Advanced Communication Technologies and Networking (CommNet)*, Rabat, Morocco: IEEE, dic. 2024, pp. 1-7. doi: 10.1109/CommNet63022.2024.10793249.