Modelo para predicción de esperanza de vida

De Mónica Monserrat Martínez Vásquez | A01710965

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

Fecha: 11/09/2025

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

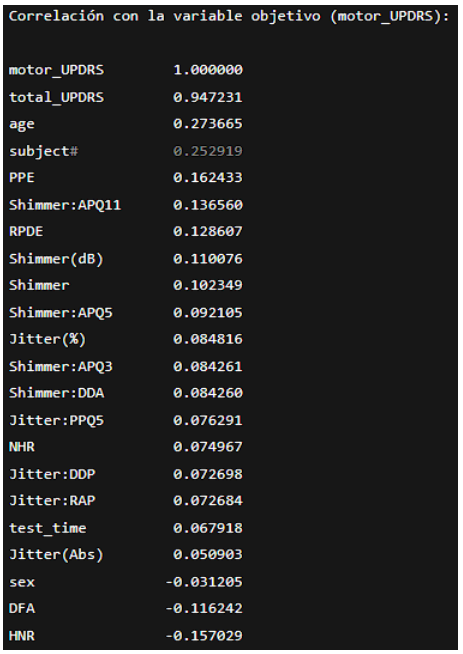
Para este proyecto, el objetivo fue implementar un modelo de aprendizaje máquina sin el uso de frameworks, yo escogí utilizar un modelo de regresión lineal múltiple para predecir una variable continua. Justamente por esta razón es que uno de los criterios más importantes para la selección del dataset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

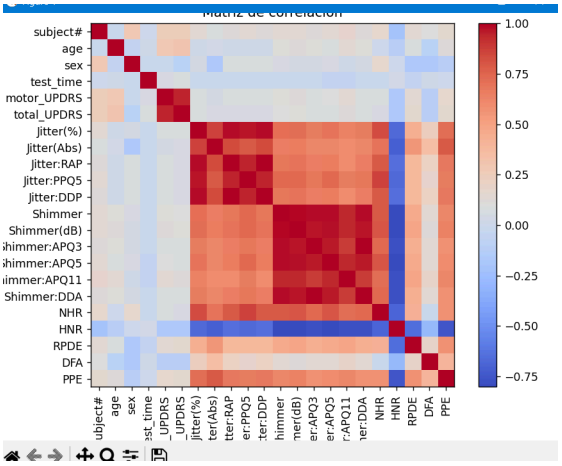
De mi primera seleccion de datasets, seleccione dos datasets que se alineaban con lo que estaba buscando y que sobretodo me llamaron la atención:

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

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

- Dataset de Parkinson:

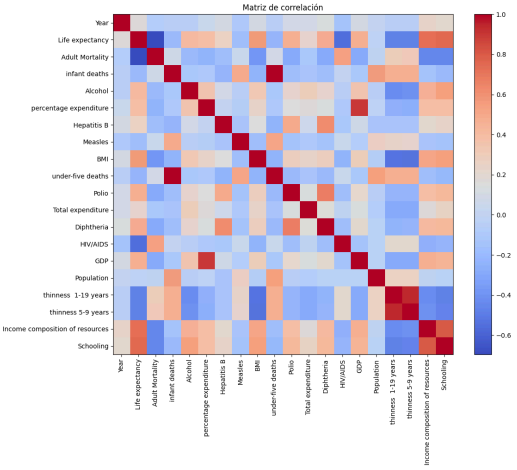




Esto indicaba que aunque había muchas variables, pocas tenían una fuerte relación lineal con la variable objetivo (motor_UPDRS). Por tanto, era menos adecuado para un modelo de regresión lineal.

- 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:

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

Figura 2. Fórmula de hipótesis lineal múltiple

Donde $\widehat{y}$ 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 (main.py)

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

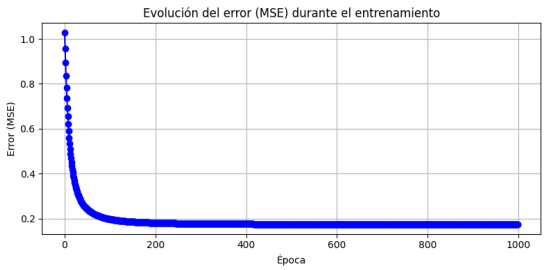


Figura 3. Gráfica del Error (MSE) por época durante el entrenamiento con todas las features

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.

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.

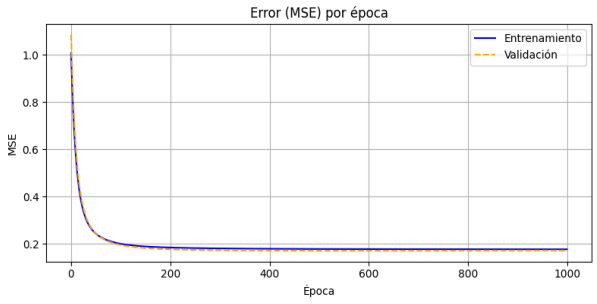


Figura 4. Gráfica del Error (MSE) por época durante el entrenamiento con 20 features

Como muestra la Figura 4, 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 una convergencia adecuada del modelo con la tasa de aprendizaje utilizada y sin evidencia de sobreajuste. Esto sugiere que el modelo no memorizó los datos, sino que generalizó correctamente los patrones.

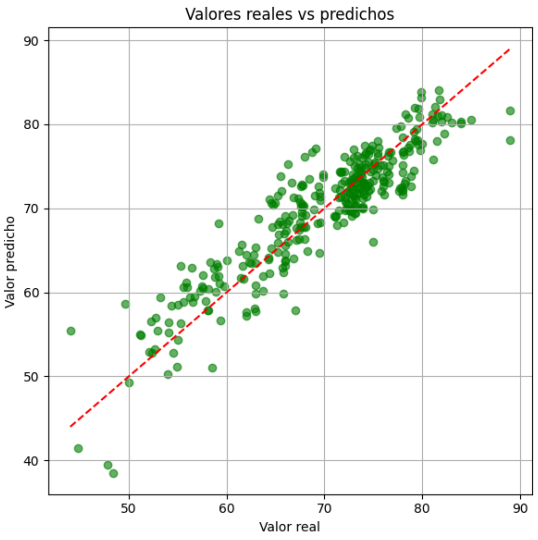


Figura 5. Gráfica valores reales vs. predichos

Además, la figura 5 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 en torno a la línea de identidad (roja punteada), lo cual evidencia un buen ajuste global. 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.

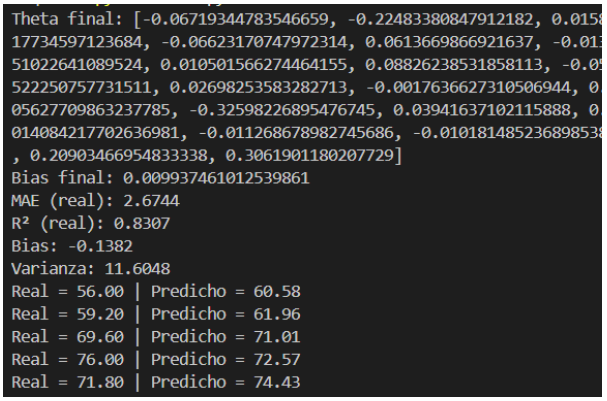


Figura 6. Resultados del entrenamiento del modelo de regresión lineal múltiple

En cuanto al error absoluto medio (MAE) nos dio un valor de 2.6744 (ver Figura 6), 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 7. Tabla de estadísticas descriptivas del dataset "Life expectancy data (WHO + UN)"

Como podemos ver en la Figura 7 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 aceptable para un modelo lineal.

También, nos dio un coeficiente de determinación (R^2) de 0.8307 (ver Figura 6), 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 6), 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. Esto es coherente con lo observado en la Figura 5.

El hecho de que el bias esté cerca de cero, no obstante, muestra 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 es un buen resultado porque indica consistencia en las predicciones, evitando errores muy grandes en unos casos y muy pequeños en otros.

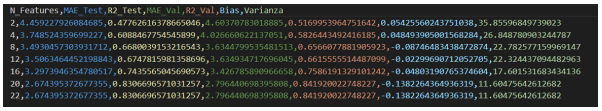
El modelo aprendió 19 coeficientes, correspondientes a las 20 variables predictoras (sin contar el bias). Estos coeficientes representan el peso de cada variable en la predicción, (ojo que esto es considerando que las variables fueron normalizadas previamente). Por ejemplo, podemos ver variables con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

6. 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
2,4.459227926084685,0.4776261637866304,4.607078301885,0.5169953964751642,0.05425560243751038,35.85596849739823
4,3.7485243596699227,0.6088467754545899,4.026660622137051,0.5826443492416185,0.04849390500156828 4,26.848780903244787
8,3.4930457303931712,0.6680039153216543,3.6344799535481513,0.6566077881905923,-0.08746483438472874,22.782577159969147
12,3.5063464452198843,0.6747815981358696,3.6349347176960045,0.6615555514487099,-0.02299690712052705,22.324437094482963
16,3.2973046354780517,0.7435565045690573,3.4267858090966658,0.7586191329101242,-0.04803190765374604,17.601531683434136
20,2.6743953726773555,0.8306696571031257,2.7964406983995808,0.841920022748227,-0.13822643649363 19,11.60475642612682
22,2.6743953726773555,0.8306696571031257,2.7964406983995808,0.841920022748227,-0.13822643649363 19,11.60475642612682
```

Figura 8. Métricas clave finales de múltiples modelos de regresión lineal múltiple de distintos números de variables (features)

Como podemos observar en la figura 8, 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).

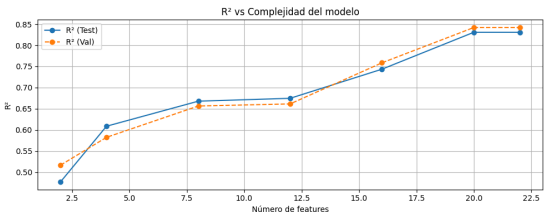


Figura 9. Desempeño del modelo con distinto número de variables (R^2)

La Figura 9 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.

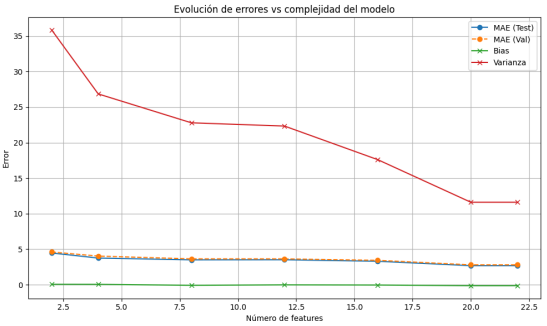


Figura 10. Comparación de métricas: MAE, Bias y Varianza vs # features

La Figura 10 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. De hecho, es preferible a un modelo con alta varianza, ya que una pequeña subestimación constante podemos corregirla después con técnicas de calibración o ajustes de post procesamiento.

7. Conclusión análisis de algoritmo de aprendizaje máquina 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.

8. 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 11. Resultados del entrenamiento del modelo usando Random Forest

Como podemos observar (Figura 11), 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. Es un resultado muy bueno, que demuestra que el modelo ha logrado capturar casi todos los patrones relevantes en los datos sin sobreajustar (lo cual se verifica al observar que el R^2 en validación es un poco superior, con 0.9603).

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.

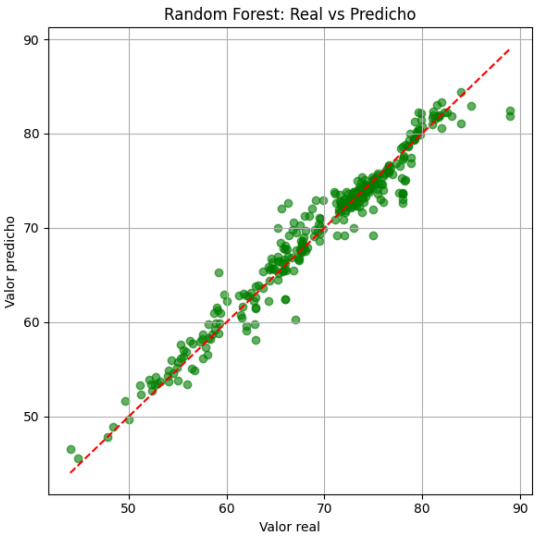


Figura 12. Resultados del entrenamiento del modelo usando Random Forest

En la figura 12, se puede observar cómo las predicciones del modelo se alinean casi perfectamente sobre la línea de identidad (línea roja discontinua). Esta alineación tan precisa indica que el modelo predice correctamente tanto para casos de esperanza de vida baja como alta, algo que

el modelo lineal no lograba tan bien, especialmente en los outliers del rango. A diferencia del modelo manual, no se observan rezagos sistemáticos ni agrupamientos con alta dispersión, lo que nos dice que este modelo tiene una excelente capacidad.

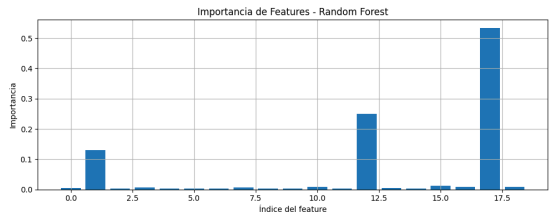


Figura 13. Importancia de las variables en Random Forest

Esta figura (Figura 13) muestra la importancia que el modelo asignó a cada una de las 20 variables utilizadas durante el entrenamiento. A diferencia de los coeficientes de una regresión lineal, que tienen una interpretación más directa (pendiente), la importancia en Random Forest indica cuánto contribuye cada variable a la reducción del error en los árboles del bosque.

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, aportaron 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.

9. Conclusión análisis de algoritmo de aprendizaje máquina con uso de framework

El modelo Random Forest no solo alcanzó mejores métricas que el modelo lineal, sino que también logró una generalización más precisa, como lo evidencian las métricas casi idénticas en validación y prueba.

Si bien su desventaja es la pérdida de interpretabilidad directa (no tenemos coeficientes como en la regresión lineal), la capacidad del modelo para capturar relaciones no lineales y manejar automáticamente interacciones complejas

lo convierten en una opción sumamente efectiva para este tipo de problema.

10. Comparativa entre modelo manual y modelo con framework

Tras analizar a profundidad ambos modelos, se realizó una comparación directa con base en las métricas clave de desempeño, el comportamiento visual de sus predicciones y su 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 14. Tabla comparativa de métricas modelo manual vs. modelo con framework

La figura 14, es una tabla comparativa de las métricas de cada modelo que hicimos, en este caso el modelo manual con regresión lineal y el modelo con framework.

Hablando primero de la precisión, podemos decir que el modelo Random Forest supera ampliamente al modelo lineal manual. El MAE se reduce más de un 50%, pasando de 2.67 a 1.17 años, lo que indica que el modelo del framework comete errores menores al predecir la esperanza de vida.

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” 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 vs 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, el modelo manual es valioso para entender el problema y tener un baseline interpretable, pero el modelo Random Forest es claramente en desempeño, precisión, robustez y generalización. Por lo tanto, para una aplicación real en predicción de esperanza de vida, el modelo del framework sería el más recomendable.

11. Ajuste de parámetros y regularización del modelo

Una vez comparados ambos enfoques se identificaron oportunidades para mejorar aún más el rendimiento del modelo en términos de precisión, robustez y generalización.

Para el modelo con framework (Random Forest):

- Tuning de hiperparámetros utilizando GridSearchCV o RandomizedSearchCV:
    - n_estimators: número de árboles
    - max_depth: profundidad máxima de cada árbol
    - min_samples_split: mínimo de muestras para dividir un nodo
    - max_features: número de features evaluados por split
    - bootstrap: si usar o no muestreo con reemplazo
- Validación cruzada (k-fold cross-validation) para evaluar la estabilidad de los resultados en múltiples subconjuntos de los datos.

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