



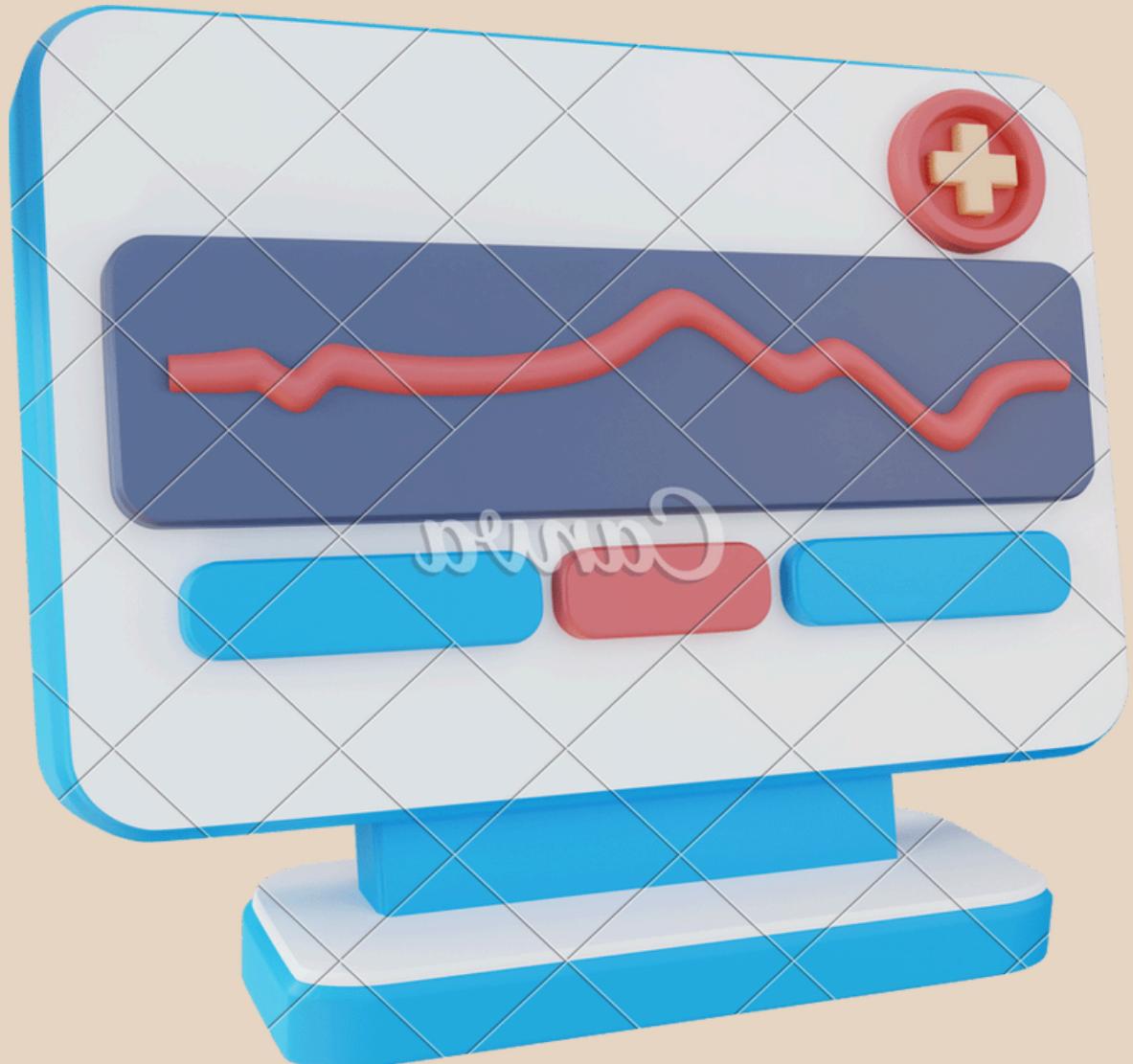
P03 - LIFE EXPECTANCY

Gonzalo Cano Padilla
André Esteban Vera
Nicolás Martínez Gutiérrez

OBJETIVO DEL PROYECTO

Predecir la esperanza de vida usando Random Forest y XGBoost, optimizando hiperparámetros con validación cruzada.

- Ambos modelos se basan en **árboles de regresión**, los cuales dividen los datos en regiones condicionales.
- En cada región, el modelo estima la respuesta esperada, mejorando la capacidad predictiva mediante el ensamble de múltiples árboles.

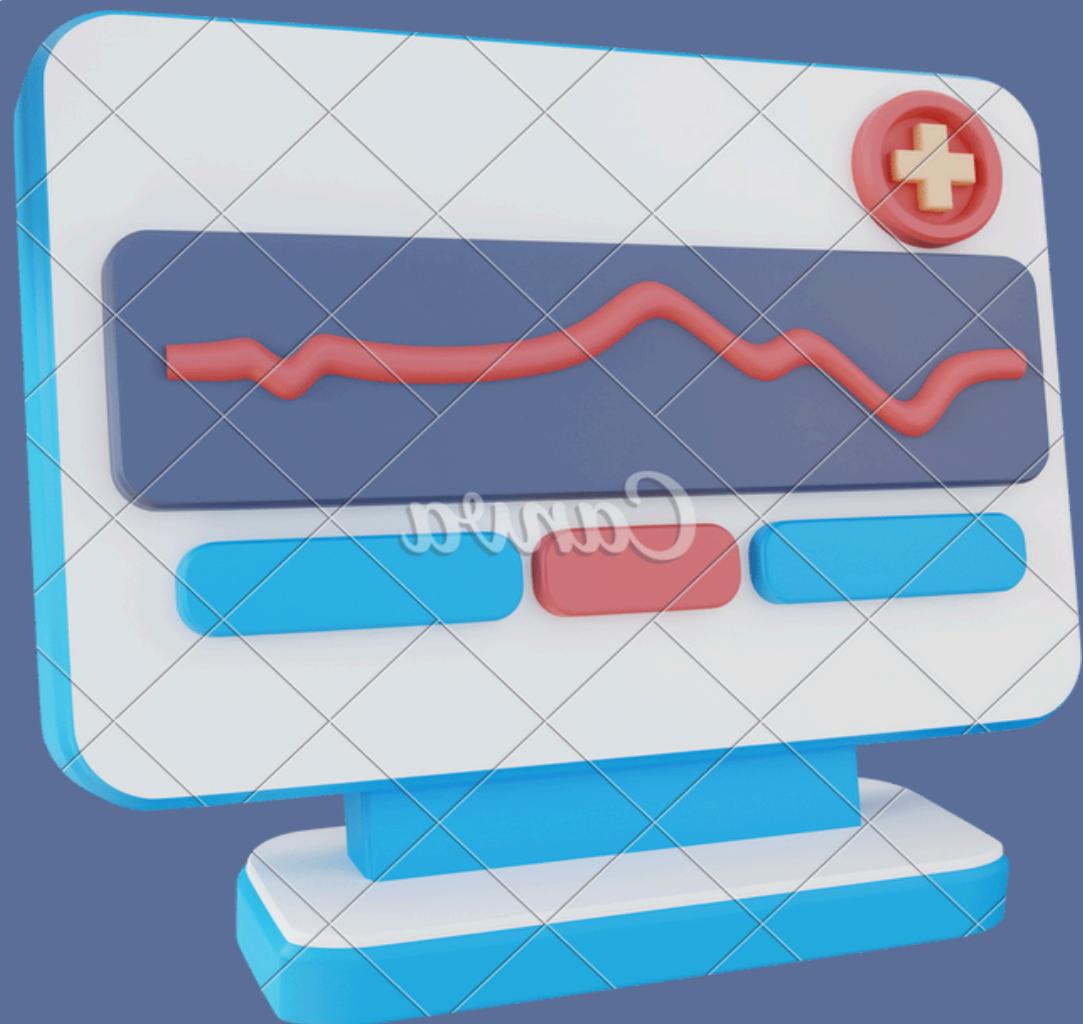


DATASET

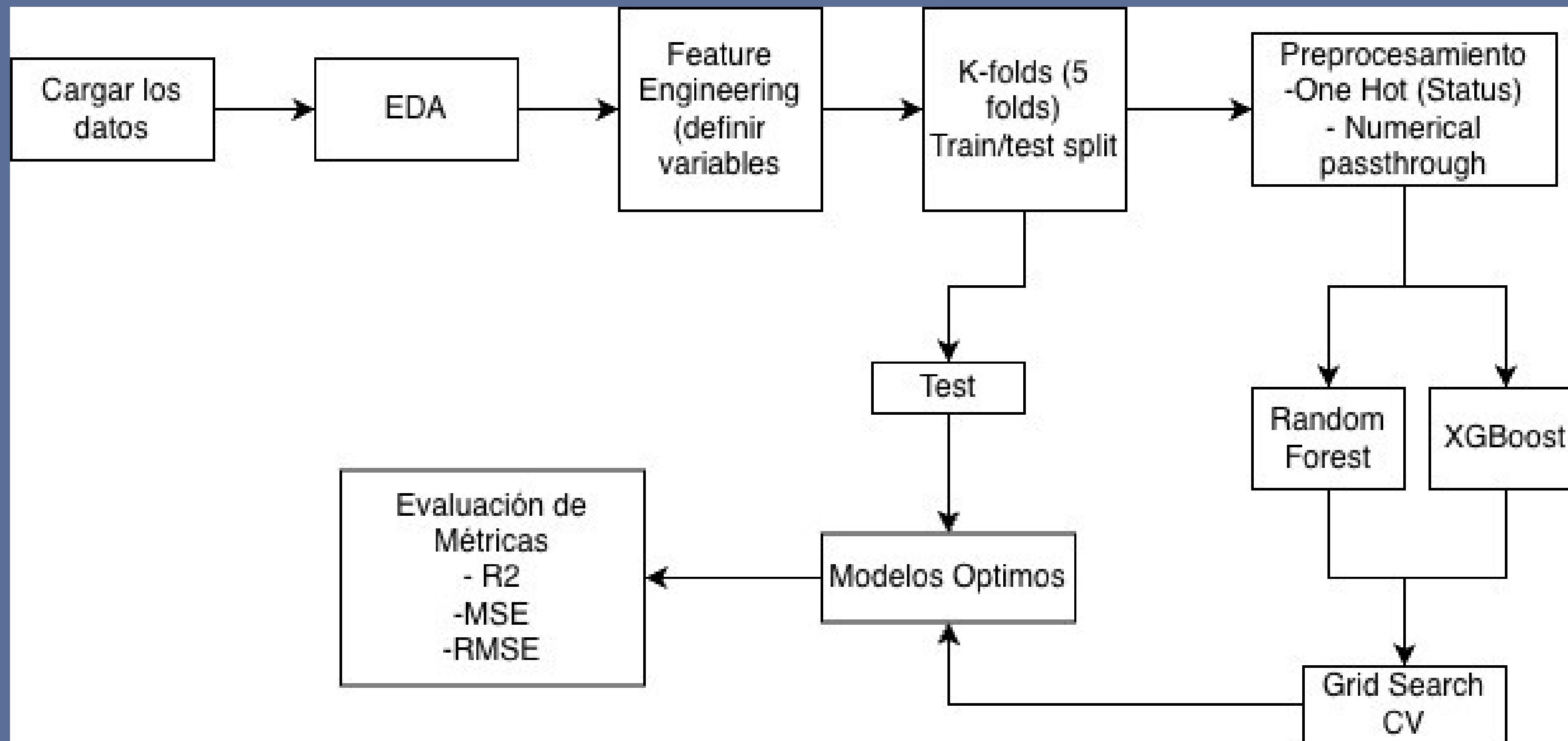
- **Salud:** mortalidad infantil, muertes menores de cinco años, vacunación, delgadez infantil, mortalidad adulta.
- **Economía:** PIB per cápita, porcentaje de gasto en salud, composición del ingreso.
- **Demografía:** población, año.
- **Factores sociales:** estatus del país (desarrollado o en desarrollo).
- **Variable objetivo:** *Life Expectancy*

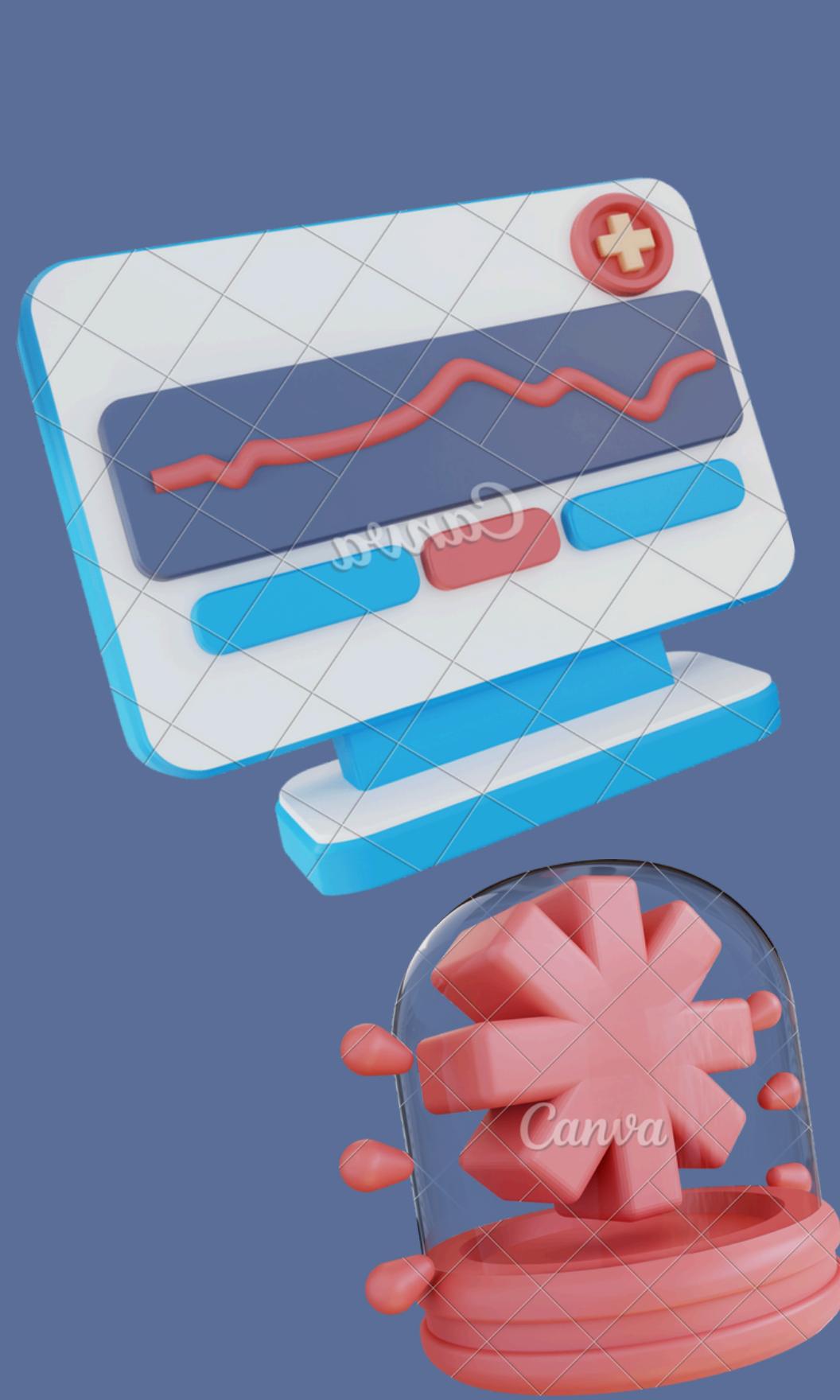
Preguntas clave

- ¿Qué factores influyen más en la esperanza de vida?
- ¿Qué modelo predice mejor?



Pipeline





MODELOS, MÉTRICAS Y COMPARACIÓN FINAL

Modelos evaluados:

- Random Forest Regressor
- XGBoost Regressor

Evaluación de modelos:

- Se aplicó k-fold cross-validation.
- Se calcularon MAE, RMSE y R^2 , reportando media y desviación estándar de cada métrica.
- El modelo final se seleccionó optimizando principalmente el MAE (menor error absoluto medio).

Meta final:

Elegir el modelo más preciso y estable para predecir la esperanza de vida.

RANDOM FOREST REGRESSOR

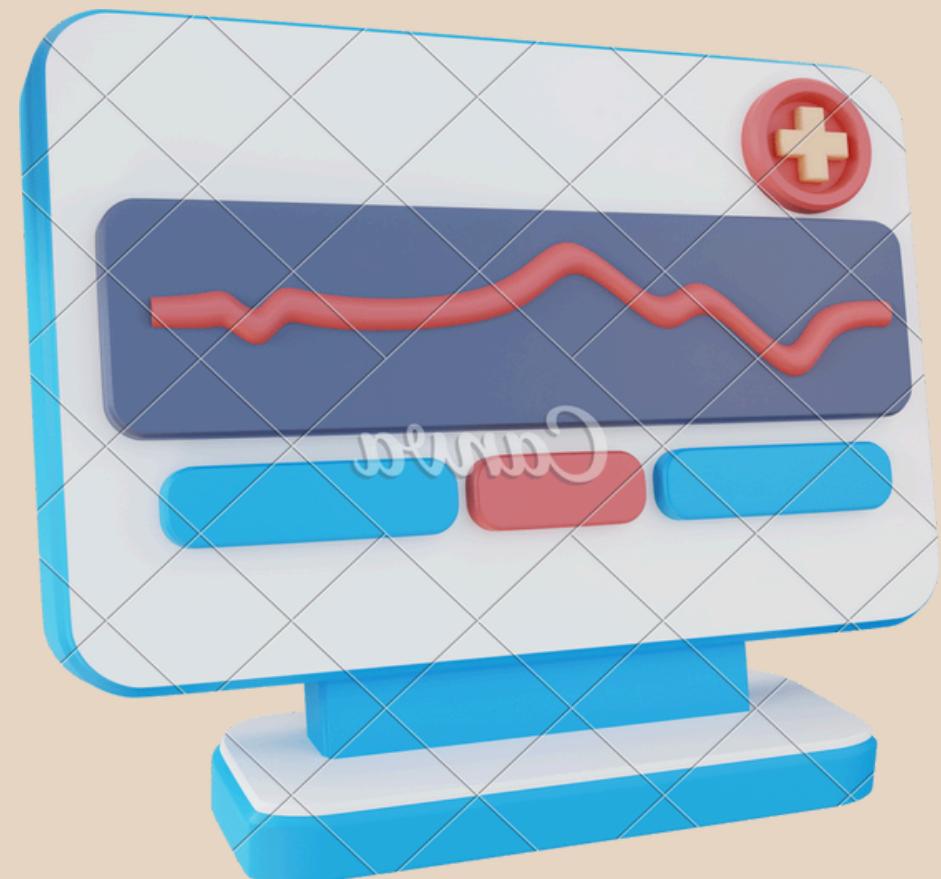
Un Random Forest para regresión combina muchos árboles entrenados con:

- **Bootstrap con reemplazo**
- **Selección aleatoria de variables en cada división**

Esto reduce la varianza y mejora la estabilidad del modelo frente a ruido.

Hiperparámetros clave a optimizar:

- n_estimators (número de árboles)
- max_depth
- min_samples_split
- min_samples_leaf
- max_features
- bootstrap (si usar o no bootstrap)



Se usará un random_state fijo para asegurar resultados consistentes.

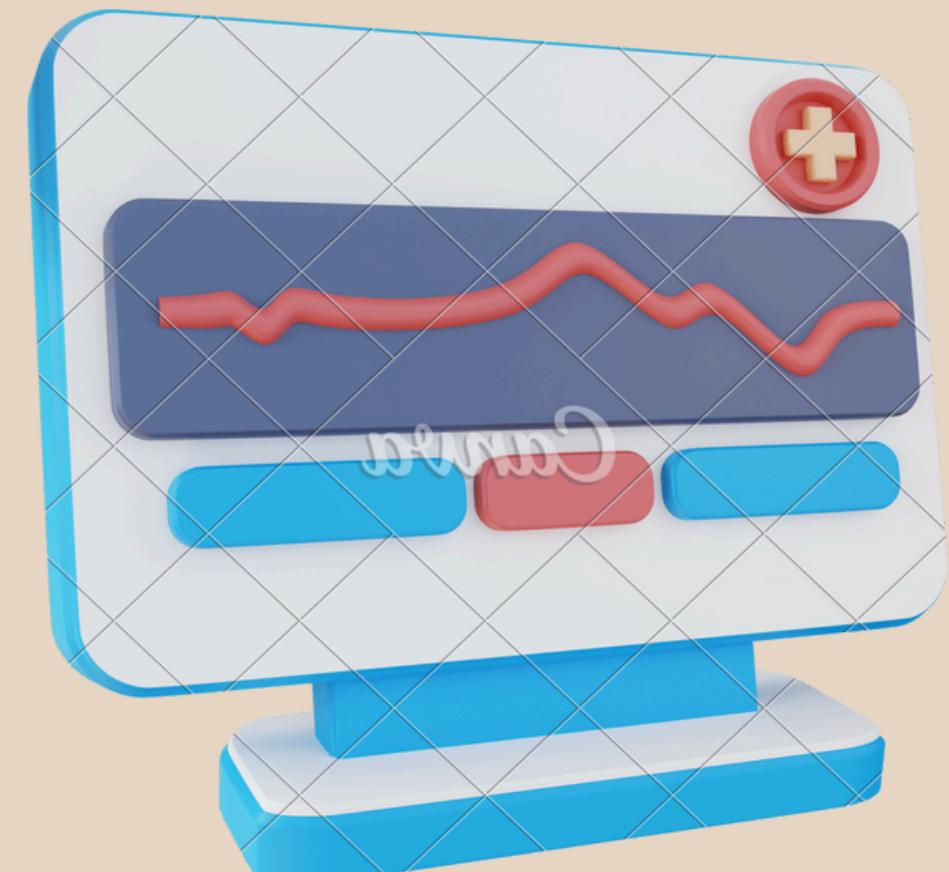
XGBOOST REGRESSOR

Modelo de boosting que construye árboles secuenciales en donde cada nuevo árbol corrige los errores del modelo anterior.

- Muy potente para relaciones no lineales
- Incluye regularización
- Control fino de complejidad del modelo

Hiperparámetros clave a optimizar:

- n_estimators
- learning_rate
- max_depth
- subsample
- colsample_bytree
- gamma
- reg_lambda, reg_alpha (regularización)



También se define un random_state fijo para reproducibilidad.

EDA (ANÁLISIS EXPLORATORIO DE DATOS)

Dataset original: 2,938 registros y 22 variables.

Se detectaron **valores faltantes** en varias columnas (Alcohol, Hepatitis B, Total expenditure, GDP, Population, Schooling, entre otras).

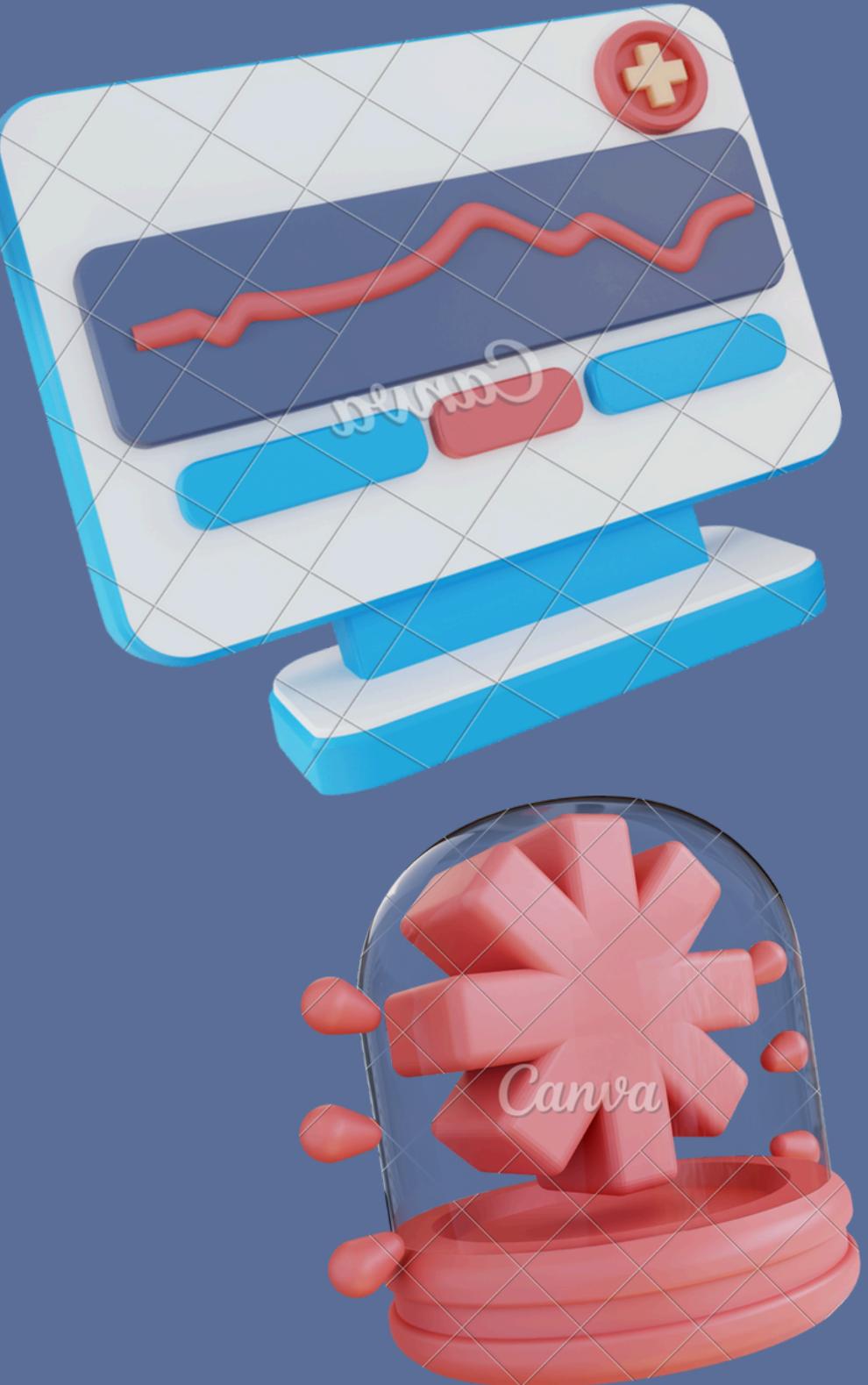
Después de **eliminar registros incompletos**, el dataset final quedó en 1,649 observaciones completas.

Se revisó la estructura del dataset y las distribuciones estadísticas generales



FEATURE ENGINEERING / PREPROCESAMIENTO

- Se eliminaron las columnas Country y Year por no aportar valor predictivo directo.
- Variable objetivo: Life expectancy.
- La variable categórica Status se transformó en variables dummies mediante OneHotEncoder.
- Todas las demás variables se mantuvieron como numéricas para el modelo.
- Se definió un ColumnTransformer para integrar el preprocessamiento (dummies + variables numéricas) dentro del pipeline de modelado.



RANDOM FOREST REGRESSOR

Mejores hiperparámetros (GridSearchCV)

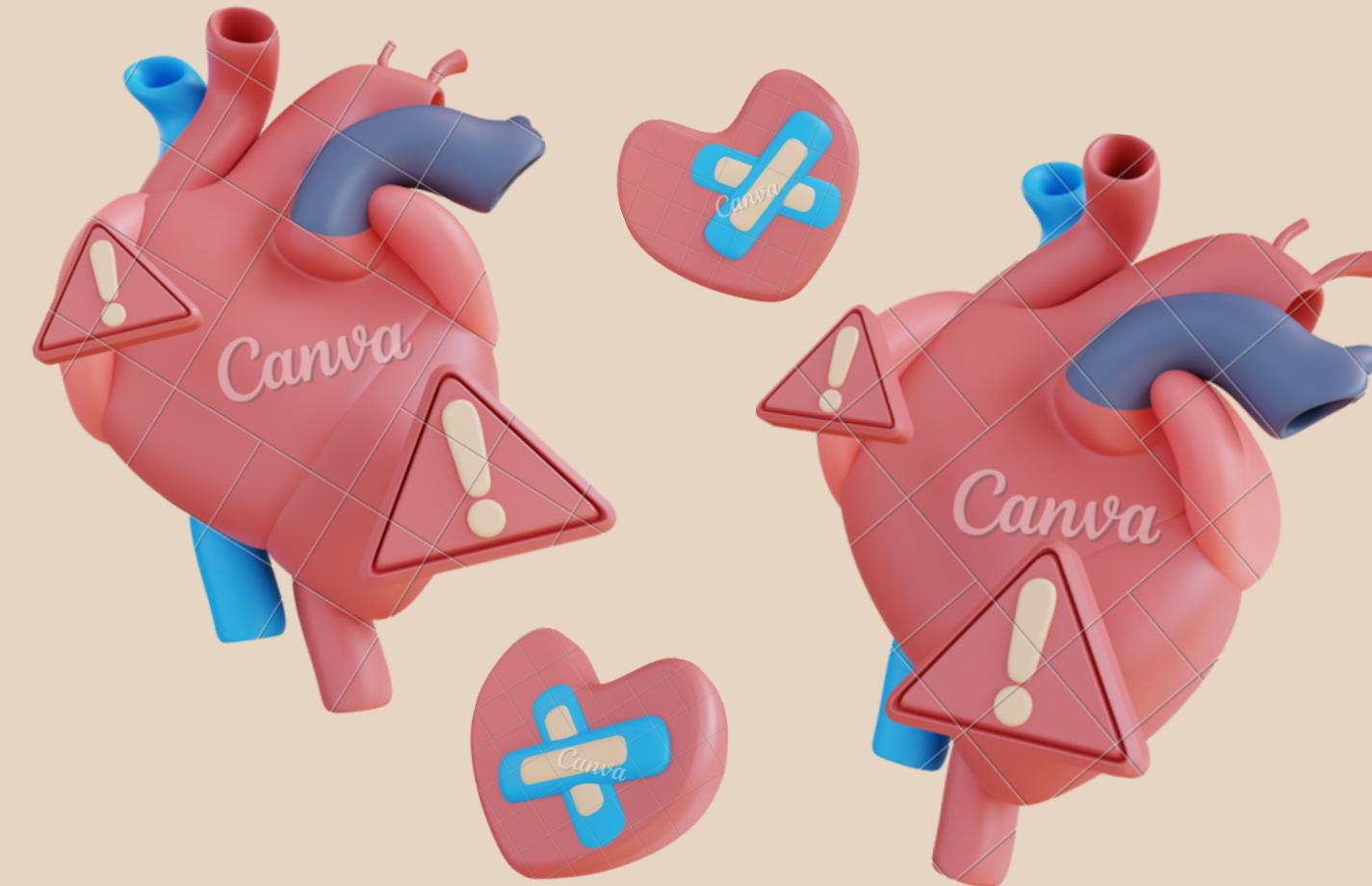
- n_estimators: 400
- max_depth: None
- min_samples_split: 2
- min_samples_leaf: 1

Desempeño (Cross-Validation, 5 folds)

- MAE: 1.27 ± 0.08
- RMSE: 1.93 ± 0.15
- R²: 0.950 ± 0.004

Desempeño en Test

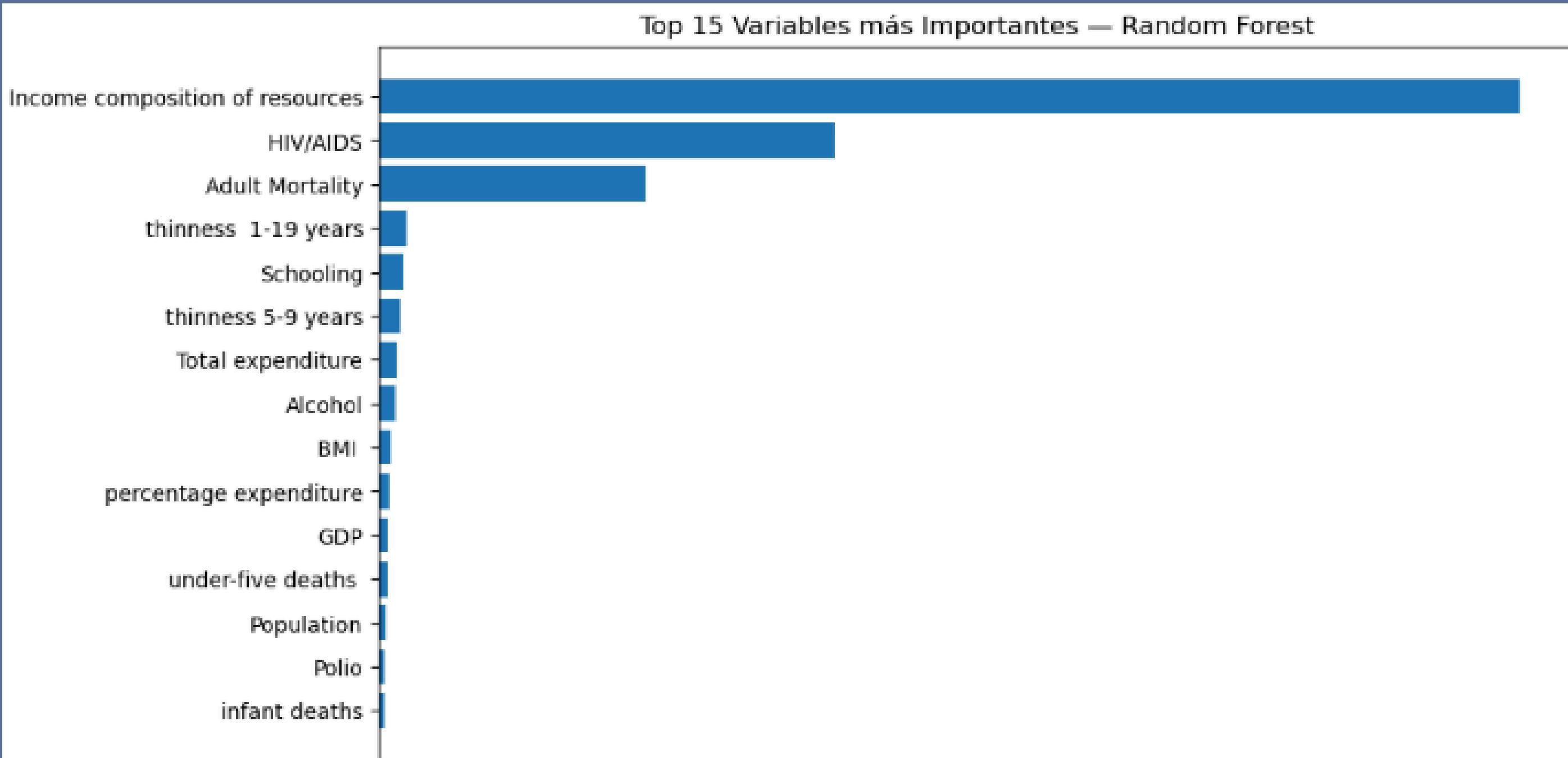
- MAE: 1.25
- RMSE: 2.00
- R²: 0.949



Conclusión clave

El Random Forest mostró alta precisión y estabilidad, explicando alrededor del 95% de la variabilidad y con errores apenas superiores a un año. Sus resultados confirman una excelente capacidad de generalización sin sobreajuste.

IMPORTANCIA DE LAS VARIABLES — RANDOM FOREST



- **Principales factores:** Income composition, HIV/AIDS y Adult Mortality.
- **Estas tres variables explican casi 90% de la importancia del modelo.**
- Nutrición, escolaridad y gasto en salud aportan menor influencia.

XGBOOST REGRESSOR

Mejores hiperparámetros (GridSearchCV)

- `n_estimators`: 500
- `max_depth`: 7
- `learning_rate`: 0.1
- `subsample`: 0.8
- `colsample_bytree`: 1.0

Desempeño (Cross-Validation, 5 folds)

- MAE: 1.19 ± 0.09
- RMSE: 1.85 ± 0.17
- R^2 : 0.955 ± 0.005

Desempeño en Test

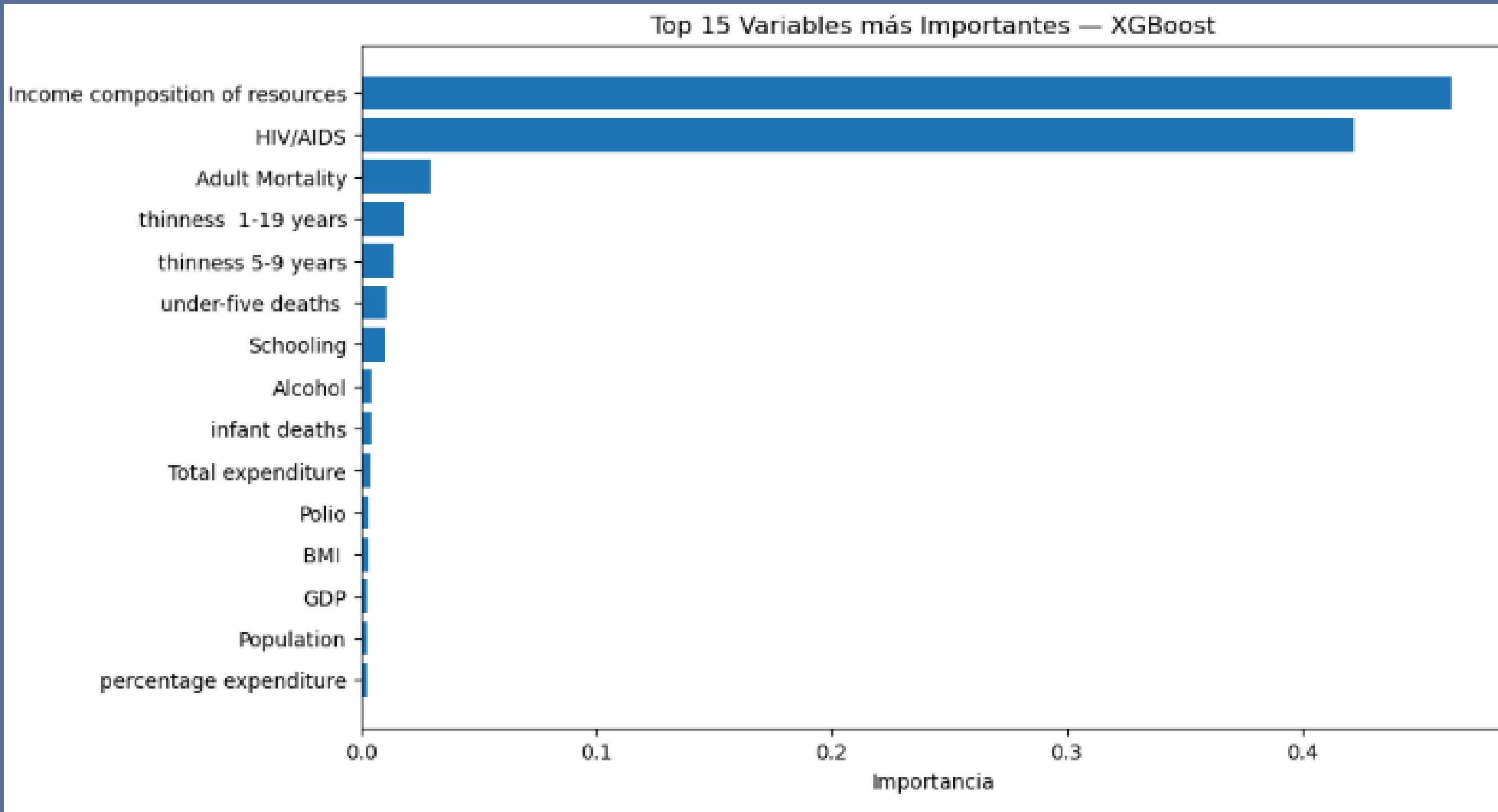
- MAE: 1.21
- RMSE: 2.05
- R^2 : 0.947



Conclusión clave

XGBoost logró el mejor desempeño promedio en CV, con la mayor capacidad explicativa (~95.5%). Es un modelo muy estable y ligeramente superior a Random Forest en validación.

PRINCIPALES DETERMINANTES DE LA ESPERANZA DE VIDA (XGBOOST)



- **Principales factores:** Income composition, HIV/AIDS y Adult Mortality.
- **Estos explican gran parte de la variación en la esperanza de vida.**
- A comparación del RF, distribuye un poco más el peso hacia factores de nutrición y mortalidad infantil.

CONCLUSIÓN

Tanto Random Forest como XGBoost predicen la esperanza de vida con gran precisión ($\approx 95\% R^2$). XGBoost tuvo un rendimiento ligeramente mejor en validación, aunque ambos modelos generalizan bien. Los factores más influyentes fueron la composición del ingreso y la prevalencia de VIH/SIDA, junto con la mortalidad adulta, nutrición y nivel educativo. En conjunto, los modelos confirman que la esperanza de vida depende principalmente de condiciones socioeconómicas y sanitarias.

