# Predicción personalizada de primas de seguro médico usando Machine Learning

Diego J. Puentes Fernández

Ciencia de Datos

# Índice

# Predicción personalizada de primas de seguro médico usando Machine Learning

Diego Puentes, Universidad de La Habana

Julio 2025

## Descripción del problema

El objetivo de este trabajo es predecir, a partir de características personales y de salud de los individuos, el coste de la prima "out of pocket" de su seguro médico utilizando técnicas de aprendizaje automático.

## Origen y descripción de los datos

Los datos utilizados provienen de tres fuentes principales: el Medical Expenditure Panel Survey (MEPS), el Clinical Classifications Software Refined (CCSR) y el Crosswalk for Clinical Information Reporting (CCIR).

- **MEPS**: Encuesta nacional de gastos médicos en Estados Unidos, que recopila información detallada sobre el uso de servicios de salud, gastos y seguros médicos de la población.
- **CCSR**: Herramienta de clasificación que agrupa diagnósticos clínicos de acuerdo a códigos ICD-10, facilitando el análisis de condiciones de salud.
- **CCIR**: Tabla de correspondencia que permite mapear códigos y categorías clínicas entre diferentes sistemas de clasificación.

## Primer procesamiento y mapeo de los datos

Se implementó una función de mapeo para convertir las columnas de los archivos CSV originales en nombres más entendibles y descriptivos, utilizando los archivos de usuario SAS provistos por MEPS (archivos .txt). Además, se realizó un análisis exploratorio de los datos, cuyos resultados principales se resumen en la siguiente tabla:

| Indicador | Valor |
|---|---|
| **Demografía (fyc)** | |
| Columnas disponibles | dwelling_unit_id, person_id, person_unique_id, panel<br>marital_status_2022, region_2022, total_healthcare_ex<br>poverty_category_2022, insurance_coverage_2022, per |
| Edad (mín, máx, Q1, Q3, media, mediana) | 0.0, 85.0, 23.0, 64.0, 43.56, 45.0 |
| Cantidad de personas por raza | |
| Non-Hispanic White only | 12211 |
| Hispanic | 4883 |
| Non-Hispanic Black only | 3244 |
| Non-Hispanic Asian only | 1220 |
| Non-Hispanic Other/multi-race | 873 |
| Cantidad de personas por estado civil | |
| Married | 8602 |
| Never married | 5495 |
| Under 16 - not applicable | 3765 |
| Divorced | 2546 |
| Widowed | 1619 |
| Separated | 397 |
| -7 | 6 |
| -8 | 1 |
| Cantidad de personas por región | |
| South | 8602 |
| West | 5693 |
| Midwest | 4498 |
| Northeast | 3443 |
| Inapplicable | 195 |
| Cantidad de personas por categoría de pobreza | |
| High income | 8282 |
| Middle income | 6269 |
| Poor/negative | 3725 |
| Low income | 3105 |
| Near poor | 1050 |
| Cantidad de personas individuales | 22431 |
| **Condiciones médicas (cond)** | |
| Columnas disponibles | person_unique_id, condition_id, panel_number, condi |
| Condiciones médicas distintas | 206 |
| Media de condiciones por persona | 4.80 |
| Top 5 condiciones más comunes | |
| CIR007 | 5391 |
| END010 | 4268 |
| MUS010 | 3061 |
| END002 | 2334 |
| MBD005 | 2158 |
| **Primas y pagos (prpl)** | |
| Estadísticas out_of_pocket_premium_edited | mín: 0.0, máx: 4583.33, media: 306.68, mediana: 212 |
| Cantidad de valores válidos | 13075 |
| **Empleo (jobs)** | |
| Estadísticas hours_per_week | mín: 1.0, máx: 168.0, media: 35.83, mediana: 40.0, Q |

**Columnas disponibles en los datasets:**

- **Condiciones médicas (cond):** person_unique_id, condition_id, panel_number, condition_round, age_at_diagnosis, injury_flag, icd10_code, ccsr_category_1
- **Características personales y demográficas (fyc):** dwelling_unit_id, person_id, person_unique_id, panel_number, age_last_birthday, sex, race_ethnicity, marital_status_2022, region_2022, total_healthcare_exp_2022, total_out_of_pocket_exp_2022, poverty_category_2022, insurance_coverage_2022, perceived_health_status, person_weight_2022
- **Empleo (jobs):** person_unique_id, job_id, panel_number, round_number, insurance_offered, temporary_job, salaried_employee, hourly_wage, hours_per_week
- **Historial de seguros (prpl):** person_unique_id, panel_number, round_number, insurance_coverage, out_of_pocket_premium, out_of_pocket_premium_edited

No todas estas columnas ofrecían información relevante para el objetivo del trabajo. Para facilitar el análisis y la integración de la información, se decidió crear un archivo JSON unificado por persona, con la siguiente estructura anidada de campos principales:

```
{
  "edad": ,
  "sexo": ,
  "raza_etnicidad": ,
  "estado_civil": ,
  "region": ,
  "categoria_pobreza": ,
  "cobertura_seguro": ,
  "estado_salud_percibido": ,
  "condiciones_medicas_actuales": [
    {
      "descripcion_ccsr": ,
      "edad_diagnostico":
    },
    ...
  ],
  "condiciones_medicas_pasadas": [],
  "historial_empleo": [
    {
      "seguro_ofrecido": ,
      "trabajo_temporal": ,
      "empleado_asalariado": ,
      "salario_por_hora": ,
      "horas_por_semana":
    },
    ...
  ],
  "historial_seguros": [
    {
      "cobertura_seguro": ,
```

```
      "prima_out_of_pocket_editada":
    },
    ...
  ]
}
```

Para la generación del archivo JSON final, se aplicaron los siguientes filtros y criterios de limpieza:

- Se filtró la información de cada persona para conservar únicamente los registros correspondientes al máximo número de round reflejado en su historial de seguros. Esto garantiza que no exista desfase temporal entre la información demográfica, clínica y de seguros considerada para el análisis.
- Se seleccionaron únicamente aquellas personas que tuvieran al menos una entrada válida de la variable objetivo, es decir, un valor válido de `prima_out_of_pocket_editada` en su historial de seguros.
- Para cada persona, se eliminaron del historial de seguros los registros que no tuvieran un valor válido de `prima_out_of_pocket_editada`, de modo que sólo se conservaron las entradas relevantes para el modelado.

## Construcción del dataset para modelado

A partir del archivo JSON unificado, se construyó un dataset tabular donde cada fila representa un embedding asociado a una persona individual. El proceso de construcción y transformación de variables fue el siguiente:

- Se mantuvieron los valores originales de las variables numéricas.
- Se aplicó codificación one-hot (one hot encoding) a las variables categóricas principales (sexo, raza/etnicidad, estado civil, región).
- Se identificaron las 20 enfermedades (CCSR) con mayor correlación con la variable objetivo y se agregaron como features binarios (presencia/ausencia) mediante label encoding.
- Se agregó el número total de enfermedades por persona (`ccsr_num_total`) y el número de enfermedades no incluidas entre las 20 principales (`ccsr_otra_condicion`).
- Para la variable objetivo (`prima_out_of_pocket_editada`), si una persona tenía más de una entrada válida, se tomó la media de sus valores.

En resumen, las columnas seleccionadas y la transformación aplicada a cada una fueron:

| **Columna** | **Transformación aplicada** |
|---|---|
| edad | Se mantuvo igual (numérica) |
| estado_salud_percibido | Label encoding (1-4) |
| ccsr_num_total | Se mantuvo igual (numérica) |
| ccsr_otra_condicion | Se mantuvo igual (numérica) |
| sexo_Male | One hot encoding |
| raza_etnicidad_Non-Hispanic Asian only | One hot encoding |
| raza_etnicidad_Non-Hispanic Black only | One hot encoding |
| raza_etnicidad_Non-Hispanic Other race or multi-race | One hot encoding |
| raza_etnicidad_Non-Hispanic White only | One hot encoding |
| estado_civil_Married | One hot encoding |
| estado_civil_Never married | One hot encoding |
| estado_civil_Separated | One hot encoding |
| estado_civil_Under 16 - not applicable | One hot encoding |
| estado_civil_Widowed | One hot encoding |
| region_Midwest | One hot encoding |
| region_Northeast | One hot encoding |
| region_South | One hot encoding |
| region_West | One hot encoding |
| ccsr_Essential hypertension | One hot encoding (presencia/ause |
| ccsr_Disorders of lipid metabolism | One hot encoding (presencia/ause |
| ccsr_Diabetes mellitus without complication | One hot encoding (presencia/ause |
| ccsr_Bacterial infections | One hot encoding (presencia/ause |
| ccsr_Osteoarthritis | One hot encoding (presencia/ause |
| ccsr_Cataract and other lens disorders | One hot encoding (presencia/ause |
| ccsr_Esophageal disorders | One hot encoding (presencia/ause |
| ccsr_Retinal and vitreous conditions | One hot encoding (presencia/ause |
| ccsr_Other general signs and symptoms | One hot encoding (presencia/ause |
| ccsr_Abnormal findings without diagnosis | One hot encoding (presencia/ause |
| ccsr_Other specified bone disease and musculoskeletal deformities | One hot encoding (presencia/ause |
| ccsr_Otitis media | One hot encoding (presencia/ause |
| ccsr_Osteoporosis | One hot encoding (presencia/ause |
| ccsr_Thyroid disorders | One hot encoding (presencia/ause |
| ccsr_Neurodevelopmental disorders | One hot encoding (presencia/ause |
| ccsr_Other and ill-defined heart disease | One hot encoding (presencia/ause |
| ccsr_Other specified upper respiratory infections | One hot encoding (presencia/ause |
| ccsr_Nutritional deficiencies | One hot encoding (presencia/ause |
| ccsr_Other specified inflammatory condition of skin | One hot encoding (presencia/ause |
| ccsr_General sensation/perception signs and symptoms | One hot encoding (presencia/ause |
| prima_out_of_pocket_editada | Se mantuvo igual (media por pers |

Cuadro 2: Resumen de las columnas y transformaciones aplicadas en el dataset final para modelado.

# Visualización de variables principales

En esta sección se presentan las principales visualizaciones generadas durante el análisis exploratorio y la construcción del dataset. Todas las imágenes fueron exportadas desde el notebook y se encuentran en la raíz del proyecto.

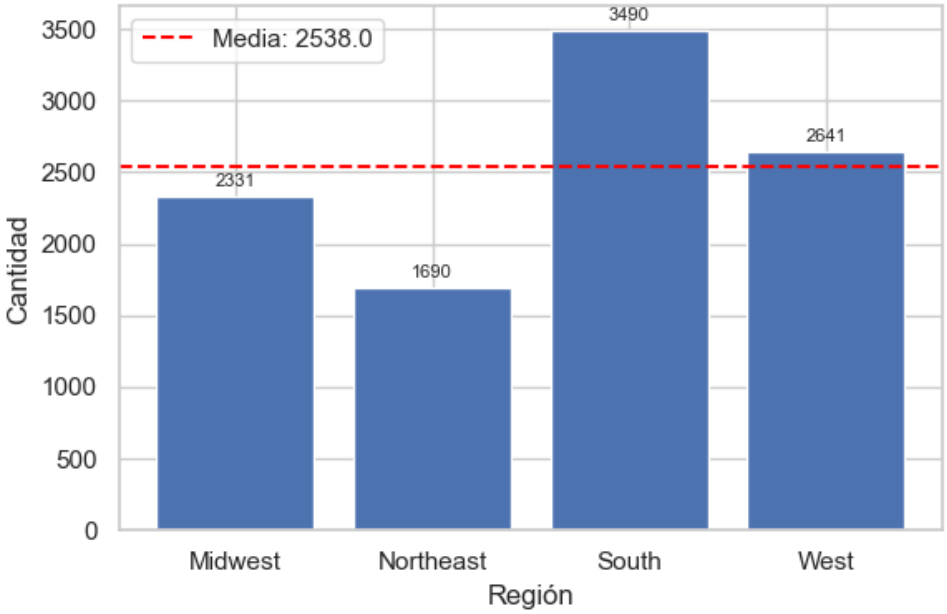


Figura 1: Cantidad de personas por región.

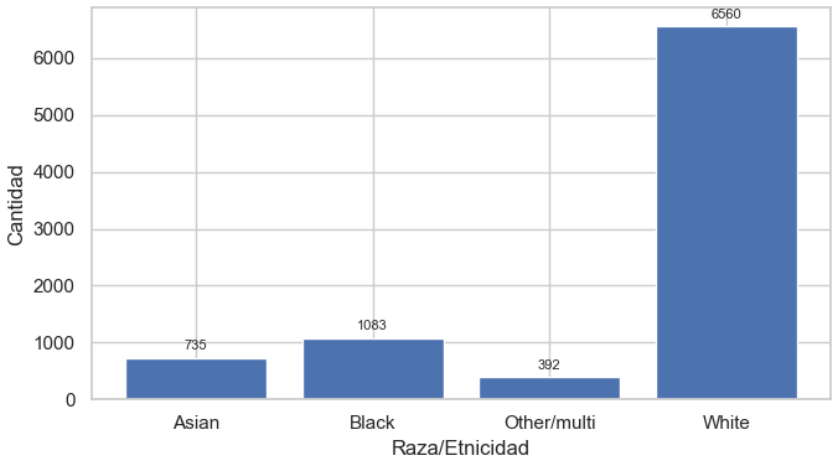


Figura 2: Cantidad de personas por raza/etnicidad.

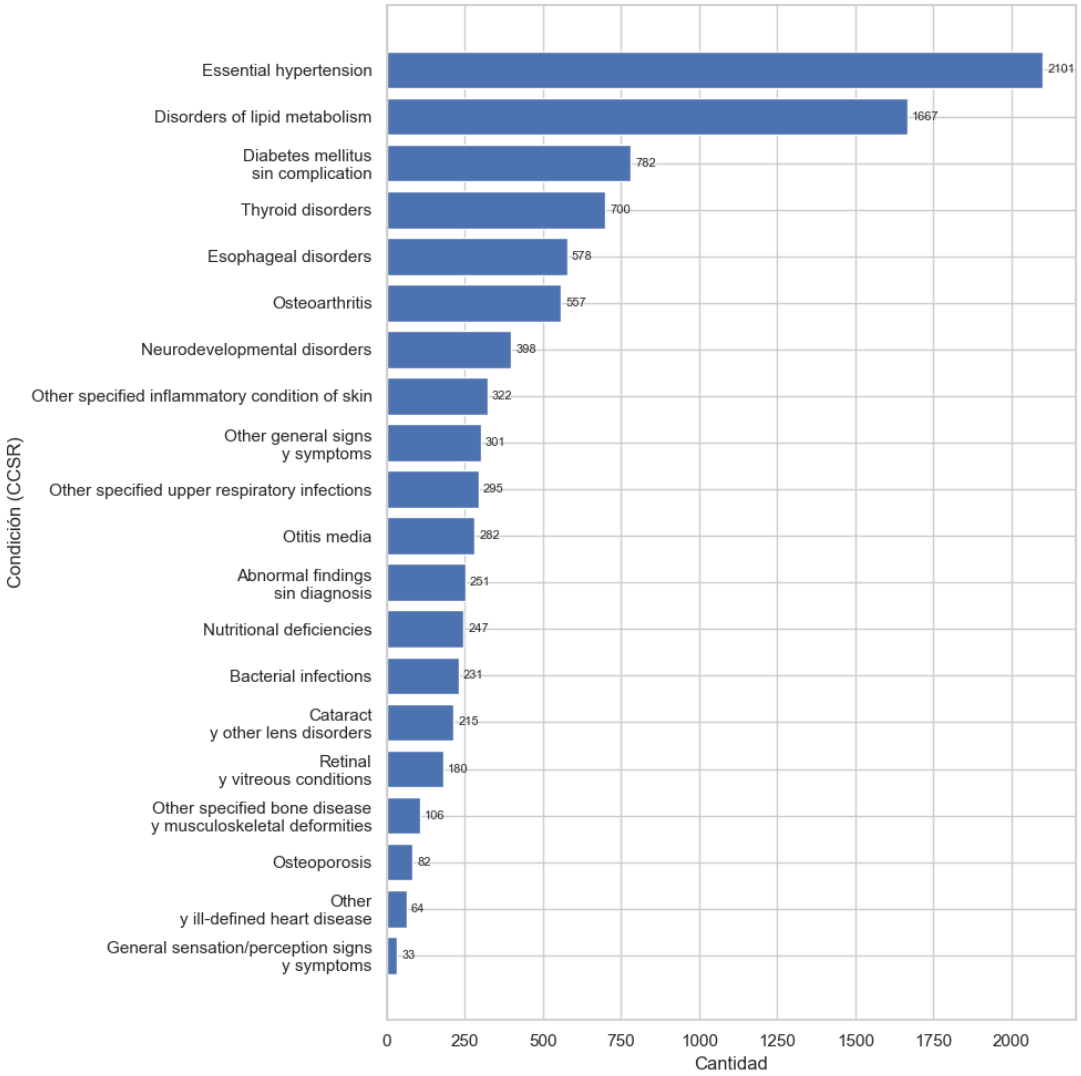


Figura 3: Cantidad de personas con cada una de las 20 enfermedades principales (CCSR).

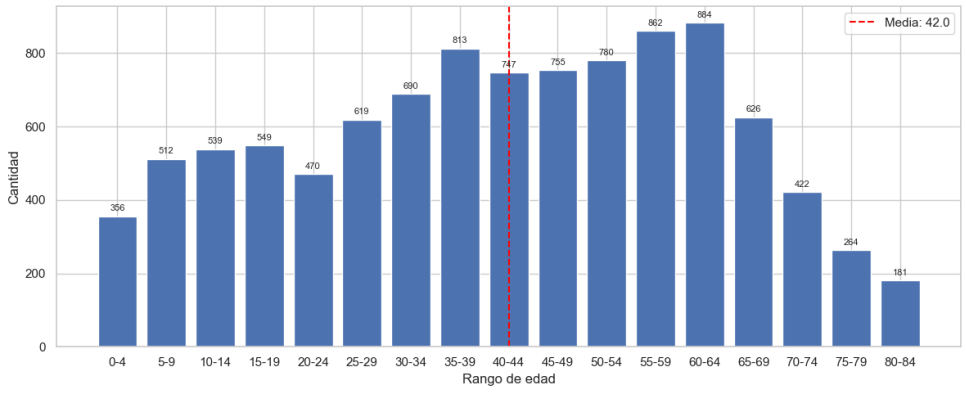


Figura 4: Cantidad de personas por rango de edad (5 años).

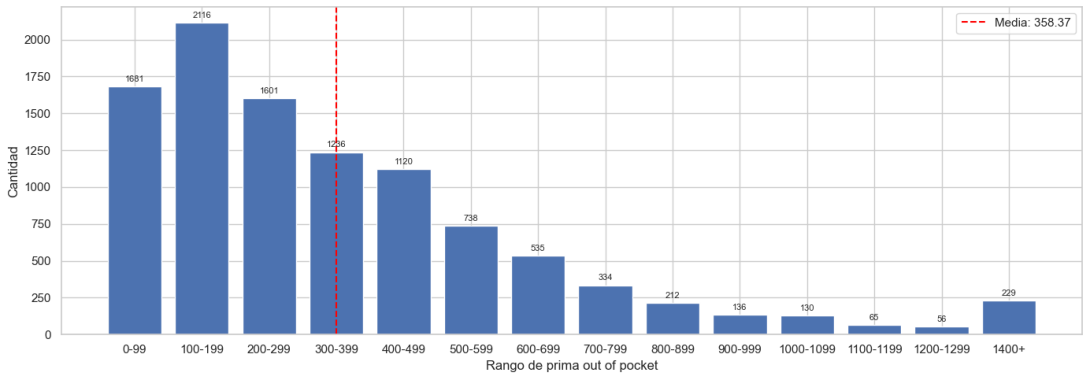


Figura 5: Cantidad de personas por rango de pagos de prima (de 100 en 100).

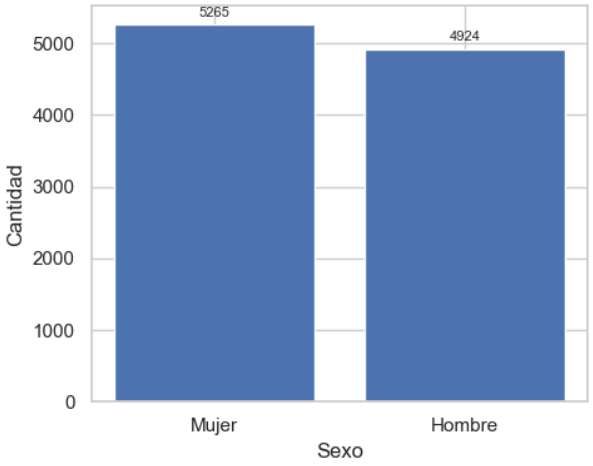


Figura 6: Cantidad de personas por sexo.

## Caracterización de Outliers en la variable objetivo

Se identificaron los outliers de la variable objetivo (`prima_out_of_pocket_editada`) como aquellos con $|z| > 3$ respecto a la media. A continuación se resumen sus características comparadas con el resto del dataset:

| **Característica** | **Outliers** | **Resto** |
|---|---|---|
| Cantidad de outliers | 186 | 10000+ |
| Porcentaje sobre el total | 1.83 % | 98.17 % |
| Edad (media) | 36.76 | 42.13 |
| Nº total de enfermedades (media) | 2.15 | 2.96 |
| Nº de otras condiciones (media) | 1.53 | 2.00 |
| Sexo masculino (proporción) | 0.46 | 0.48 |
| **Región (proporción)** | | |
| Midwest | 0.27 | 0.23 |
| Northeast | 0.18 | 0.17 |
| South | 0.19 | 0.35 |
| West | 0.35 | 0.26 |
| **Estado civil (proporción)** | | |
| Married | 0.55 | 0.48 |
| Never married | 0.20 | 0.23 |
| Separated | 0.00 | 0.01 |
| Under 16 - not applicable | 0.23 | 0.15 |
| Widowed | 0.01 | 0.04 |
| **Raza/etnicidad (proporción)** | | |
| Non-Hispanic Asian only | 0.06 | 0.07 |
| Non-Hispanic Black only | 0.04 | 0.11 |
| Non-Hispanic Other/multi-race | 0.02 | 0.04 |
| Non-Hispanic White only | 0.74 | 0.64 |
| **Enfermedades principales (proporción)** | | |
| Essential hypertension | 0.13 | 0.21 |
| Disorders of lipid metabolism | 0.11 | 0.16 |
| Diabetes mellitus without complication | 0.02 | 0.08 |
| Bacterial infections | 0.01 | 0.02 |
| Osteoarthritis | 0.02 | 0.06 |
| Cataract and other lens disorders | 0.01 | 0.02 |
| Esophageal disorders | 0.02 | 0.06 |
| Retinal and vitreous conditions | 0.01 | 0.02 |
| Other general signs and symptoms | 0.02 | 0.03 |
| Abnormal findings without diagnosis | 0.01 | 0.02 |
| Other specified bone disease and musculoskeletal deformities | 0.00 | 0.01 |
| Otitis media | 0.02 | 0.03 |
| Osteoporosis | 0.01 | 0.01 |
| Thyroid disorders | 0.05 | 0.07 |
| Neurodevelopmental disorders | 0.08 | 0.04 |
| Other and ill-defined heart disease | 0.00 | 0.01 |
| Other specified upper respiratory infections | 0.03 | 0.03 |
| Nutritional deficiencies | 0.02 | 0.02 |
| Other specified inflammatory condition of skin | 0.04 | 0.03 |
| General sensation/perception signs and symptoms | 0.00 | 0.00 |

Cuadro 3: Comparación de características entre outliers y el resto del dataset para la variable objetivo.

Los outliers representan el 1.8 % de la muestra y tienden a ser ligeramente más jóvenes, con menos enfermedades y menor proporción de condiciones crónicas que el resto. La mayoría son de raza blanca no hispana y residen en la región Oeste o Midwest, con menor presencia en el Sur. No se observa una diferencia marcada en sexo ni en estado civil, aunque hay una ligera mayor proporción de casados y menores de 16 años. En cuanto a enfermedades, los outliers presentan menor prevalencia de hipertensión, diabetes y dislipidemias.

En conclusión, los outliers no presentan características demográficas o clínicas radicalmente distintas al resto del dataset, lo que sugiere que los valores extremos de la variable objetivo pueden deberse a factores no capturados en las variables disponibles o a variabilidad natural en los datos.

# Primera Propuesta de Modelado

La primera propuesta de modelado consistió en construir un modelo capaz de predecir con exactitud el valor de la prima "out of pocket" que una persona debería pagar por su seguro médico, a partir de las variables disponibles en el dataset procesado.

Sin embargo, la predicción exacta de la prima "out of pocket" resulta un reto considerable con la información disponible, y requiere considerar enfoques alternativos de modelado, técnicas de manejo de outliers y posiblemente la incorporación de variables adicionales, como se muestra a continuación.

## Métricas de evaluación utilizadas

Para evaluar el desempeño de los modelos de regresión en la predicción de la prima "out of pocket", se emplearon las siguientes métricas:

- **MAE (Mean Absolute Error / Error Absoluto Medio):** Indica el promedio de las diferencias absolutas entre los valores predichos y los reales. Es una métrica intuitiva y robusta frente a outliers, ya que mide el error promedio en las mismas unidades que la variable objetivo.
- **RMSE (Root Mean Squared Error / Raíz del Error Cuadrático Medio):** Penaliza más fuertemente los errores grandes que el MAE, ya que eleva al cuadrado las diferencias antes de promediar. Es útil para identificar si el modelo comete errores grandes con frecuencia.
- **$R^2$ (Coeficiente de determinación):** Mide la proporción de la varianza de la variable objetivo explicada por el modelo. Un valor cercano a 1 indica buen ajuste, mientras que valores cercanos a 0 o negativos indican bajo poder predictivo.

Estas métricas son relevantes porque permiten comparar modelos de manera objetiva, considerando tanto la magnitud promedio del error (MAE), la sensibilidad a errores grandes (RMSE) y la capacidad explicativa global del modelo ($R^2$).

# Modelos para el primer problema

Para abordar la predicción de la prima "out of pocket" se seleccionaron cuatro modelos de regresión representativos:

- **Linear Regression**: Modelo lineal base, útil como referencia y para identificar relaciones lineales simples.
- **Random Forest**: Ensamble de árboles de decisión, robusto ante relaciones no lineales y capaz de manejar variables categóricas codificadas.
- **Gradient Boosting**: Ensamble secuencial que optimiza el error, adecuado para capturar patrones complejos y relaciones no lineales.
- **XGBoost**: Variante avanzada de boosting, eficiente y con regularización, ampliamente utilizada en competencias de ciencia de datos.

Estos modelos fueron elegidos por su complementariedad: permiten comparar desde un enfoque lineal simple hasta técnicas de ensamble y boosting que pueden capturar relaciones más complejas y no lineales presentes en los datos.

## Resultados sin ajuste de hiperparámetros

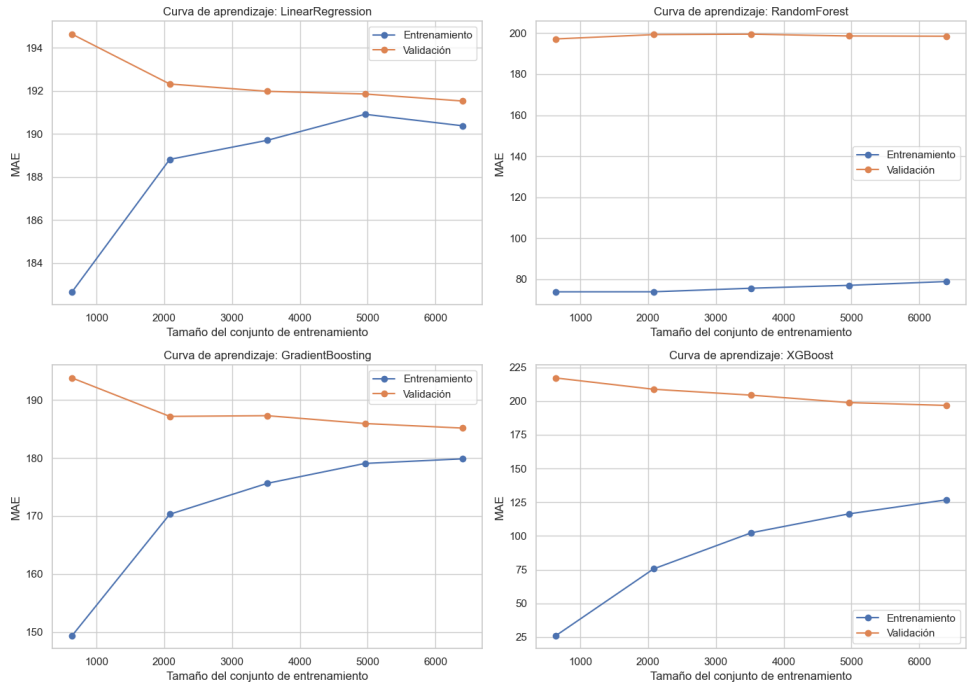


Figura 7: Curvas de aprendizaje de los 4 modelos sin ajuste de hiperparámetros.

| Modelo | MAE | RMSE | $R^2$ |
|---|---|---|---|
| LinearRegression | 191.64 | 252.57 | 0.06 |
| RandomForest | 203.96 | 269.36 | -0.07 |
| GradientBoosting | 187.45 | 248.91 | 0.08 |
| XGBoost | 198.06 | 260.85 | -0.01 |

Cuadro 4: Comparación de métricas de los modelos sin ajuste de hiperparámetros.

## Resultados con ajuste de hiperparámetros

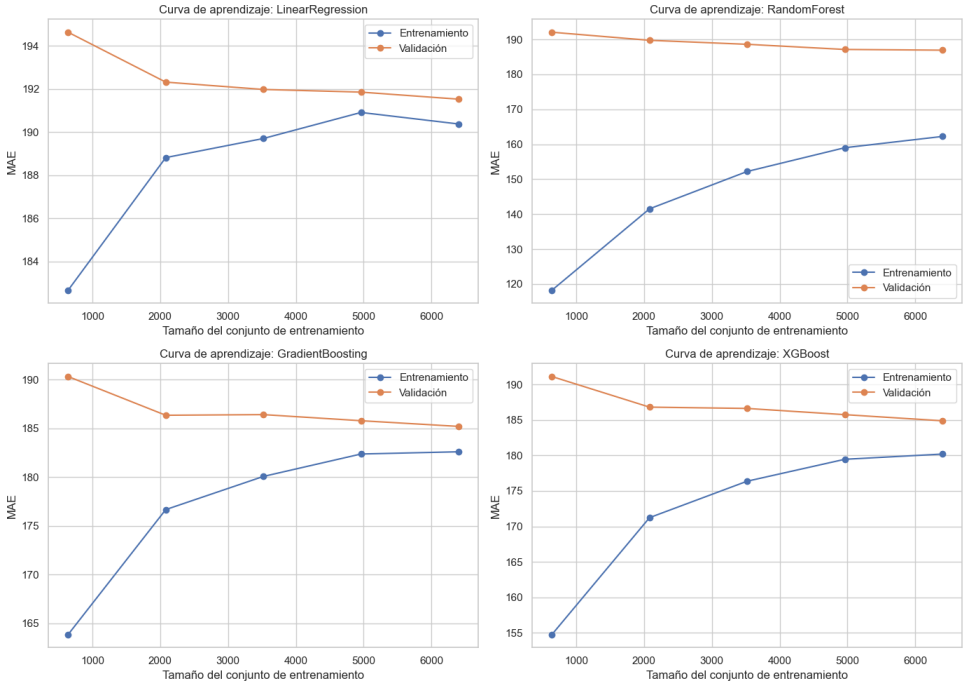


Figura 8: Curvas de aprendizaje de los 4 modelos tras ajuste de hiperparámetros.

| Modelo | MAE | RMSE | $R^2$ |
|---|---|---|---|
| LinearRegression | 191.64 | 252.57 | 0.06 |
| RandomForest | 188.90 | 250.27 | 0.07 |
| GradientBoosting | 187.30 | 248.44 | 0.09 |
| XGBoost | 187.36 | 248.76 | 0.08 |

Cuadro 5: Comparación de métricas de los modelos tras ajuste de hiperparámetros.

## Conclusiones sobre el desempeño de los modelos

En general, todos los modelos presentan un desempeño modesto, con valores de MAE y RMSE relativamente altos en comparación con la media de la variable objetivo, y valores de $R^2$ bajos (cercanos a 0). Esto indica que la capacidad explicativa de los modelos sobre la variabilidad de la prima es limitada.

El ajuste de hiperparámetros mejora ligeramente el desempeño de los modelos de ensamble (Random Forest, Gradient Boosting y XGBoost), pero la mejora es marginal. El modelo lineal sirve como referencia y muestra que la relación entre las variables disponibles y la prima no es predominantemente lineal.

Las razones de este desempeño limitado están asociadas a la estructura del dataset:

- La variable objetivo presenta una alta dispersión y asimetría, con una cola de valores extremos (outliers), lo que dificulta que los modelos ajusten correctamente tanto los valores típicos como los atípicos.
- Las variables explicativas disponibles muestran correlaciones muy bajas con la variable objetivo, tanto lineales como no lineales, y en conjunto los modelos lineales apenas explican un 5 % de la variabilidad de la prima ($R^2$ multivariable lineal $\approx 0.05$).
- Ninguna variable individualmente, ni siquiera con ajustes polinómicos, logra explicar una fracción significativa de la variabilidad de la prima (todas las correlaciones y $R^2$ individuales son cercanas a cero).
- Existen relaciones complejas y factores no observados que los modelos no pueden capturar completamente, lo que justifica la necesidad de modelos no lineales, a pesar del poco escazo éxito en esta tarea.

En conclusión, aunque los modelos avanzados de boosting y ensamble logran un desempeño ligeramente superior, la predicción exacta de la prima "out of pocket" sigue siendo un reto considerable con la información disponible.

# Segunda propuesta de Modelado: Replanteando el Problema

Ante la dificultad de predecir con precisión un único valor de prima "out of pocket" para cada individuo, se propone un replanteamiento del problema: en lugar de estimar un valor puntual, predecir **intervalos de primas** que sean relevantes y útiles para la toma de decisiones.

En este enfoque, para cada individuo se busca determinar una lista de tres valores que representen límites superiores de intervalos de prima:

- **Excelente:** Límite superior de lo que se consideraría una prima muy baja o especialmente favorable para el perfil del individuo.
- **Buena:** Límite superior de una prima razonable o aceptable para el perfil del individuo.
- **Regular:** Límite superior de una prima que, aunque no es óptima, sigue siendo aceptable dentro del contexto del mercado y las características del individuo.

De este modo, el modelo no intenta predecir un único número exacto, sino proporcionar una referencia personalizada de rangos de prima, facilitando la comparación y la toma de decisiones informadas.

**Ventajas de este enfoque:**

- La alta dispersión y asimetría de la variable objetivo, así como la baja capacidad explicativa de los modelos lineales, sugieren que predecir intervalos puede ser más robusto y útil que intentar ajustar un valor puntual.
- El enfoque por intervalos permite acomodar la variabilidad natural y los factores no observados presentes en el dataset, ofreciendo una predicción más realista y accionable.
- Los intervalos pueden adaptarse a la distribución real de primas observada en el dataset, reflejando mejor la incertidumbre y la heterogeneidad de los casos individuales.

En la siguiente sección se mostrarán los resultados obtenidos aplicando este segundo enfoque de modelado basado en intervalos personalizados de prima.

# Resultados del modelado por intervalos personalizados

Para cada uno de los tres límites de prima (excelente, buena y regular), se entrenaron los cuatro modelos principales tanto sin ajuste de hiperparámetros como con ajuste. A continuación se presentan primero los resultados sin ajuste y luego los resultados con ajuste de hiperparámetros.

## Límite Excelente

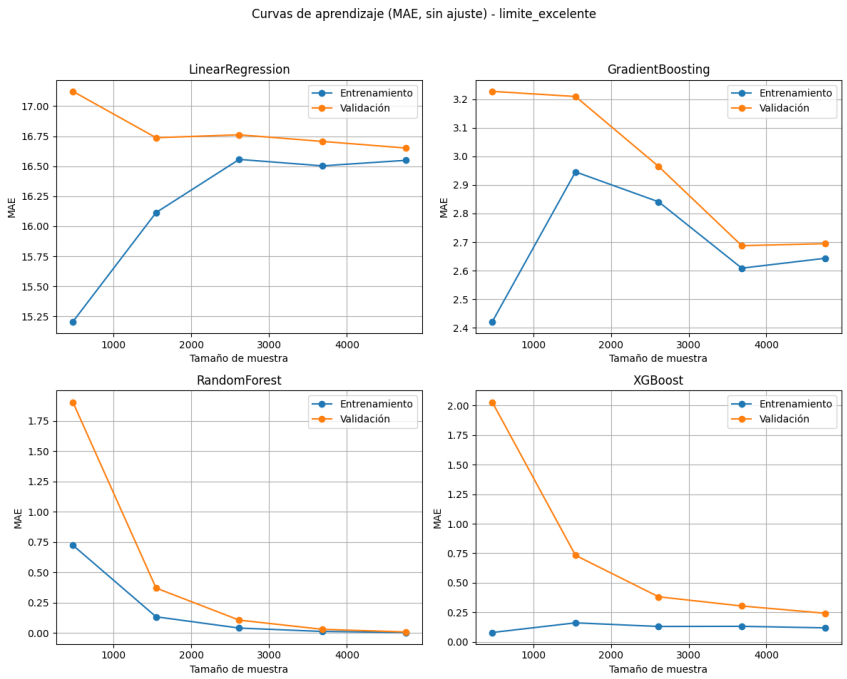


Figura 9: Curvas de aprendizaje para el límite excelente (sin ajuste de hiperparámetros).

| Modelo | MAE | RMSE | $R^2$ |
|---|---|---|---|
| LinearRegression | 16.68 | 21.89 | 0.72 |
| GradientBoosting | 2.65 | 3.36 | 0.9935 |
| RandomForest | 0.00 | 0.00 | 1.00 |
| XGBoost | 0.17 | 0.30 | 0.9999 |

Cuadro 6: Métricas para el límite excelente (sin ajuste de hiperparámetros).

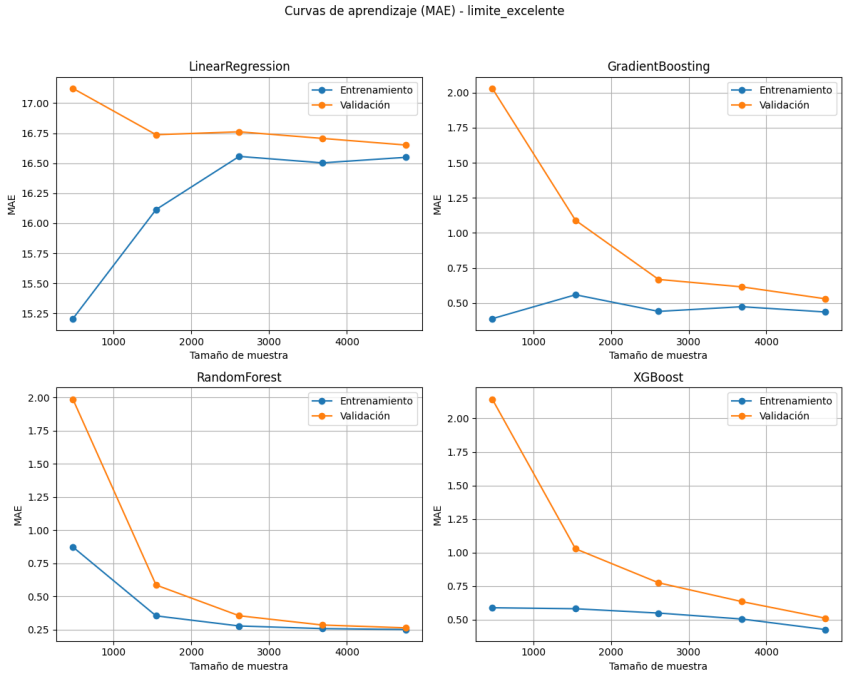


Figura 10: Curvas de aprendizaje para el límite excelente (con ajuste de hiperparámetros).

| Modelo | MAE | RMSE | $R^2$ |
|---|---|---|---|
| LinearRegression | 16.68 | 21.89 | 0.72 |
| GradientBoosting | 0.48 | 0.67 | 0.9997 |
| RandomForest | 0.26 | 0.52 | 0.9998 |
| XGBoost | 0.47 | 0.65 | 0.9998 |

Cuadro 7: Métricas para el límite excelente (con ajuste de hiperparámetros).

## Límite Bueno

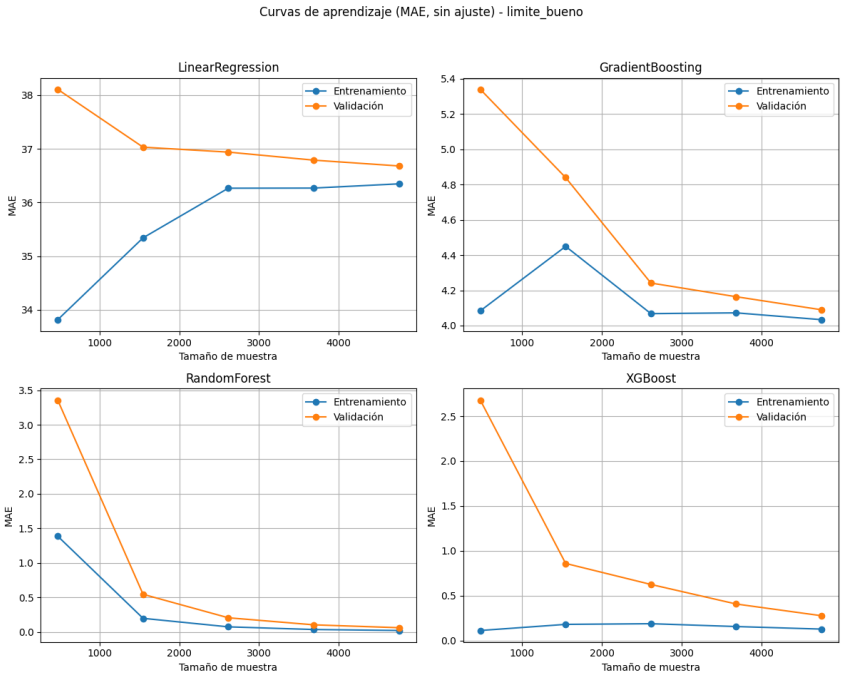


Figura 11: Curvas de aprendizaje para el límite bueno (sin ajuste de hiperparámetros).

| **Modelo** | **MAE** | **RMSE** | **$R^2$** |
|---|---|---|---|
| LinearRegression | 36.15 | 46.48 | 0.56 |
| GradientBoosting | 4.15 | 5.70 | 0.9933 |
| RandomForest | 0.01 | 0.13 | 1.00 |
| XGBoost | 0.25 | 0.53 | 0.9999 |

Cuadro 8: Métricas para el límite bueno (sin ajuste de hiperparámetros).

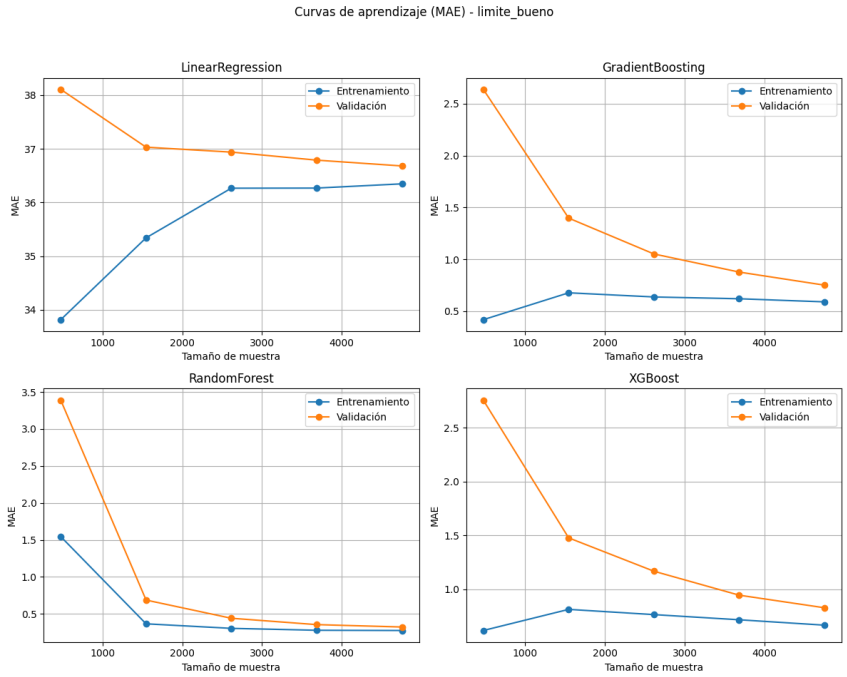


Figura 12: Curvas de aprendizaje para el límite bueno (con ajuste de hiperparámetros).

| **Modelo** | **MAE** | **RMSE** | $\mathbf{R^2}$ |
|---|---|---|---|
| LinearRegression | 36.15 | 46.48 | 0.56 |
| GradientBoosting | 0.73 | 1.07 | 0.9998 |
| RandomForest | 0.24 | 0.73 | 0.9999 |
| XGBoost | 0.71 | 1.03 | 0.9998 |

Cuadro 9: Métricas para el límite bueno (con ajuste de hiperparámetros).

## Límite Regular

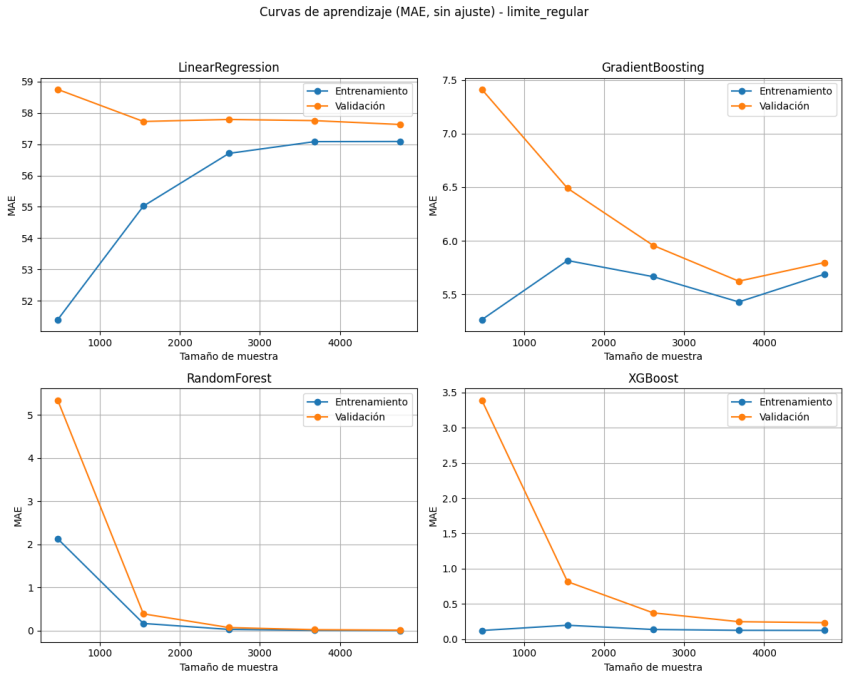


Figura 13: Curvas de aprendizaje para el límite regular (sin ajuste de hiperparámetros).

| **Modelo** | **MAE** | **RMSE** | **$R^2$** |
|---|---|---|---|
| LinearRegression | 56.67 | 69.25 | 0.35 |
| GradientBoosting | 5.86 | 7.79 | 0.9918 |
| RandomForest | 0.00 | 0.09 | 1.00 |
| XGBoost | 0.08 | 0.14 | 1.00 |

Cuadro 10: Métricas para el límite regular (sin ajuste de hiperparámetros).

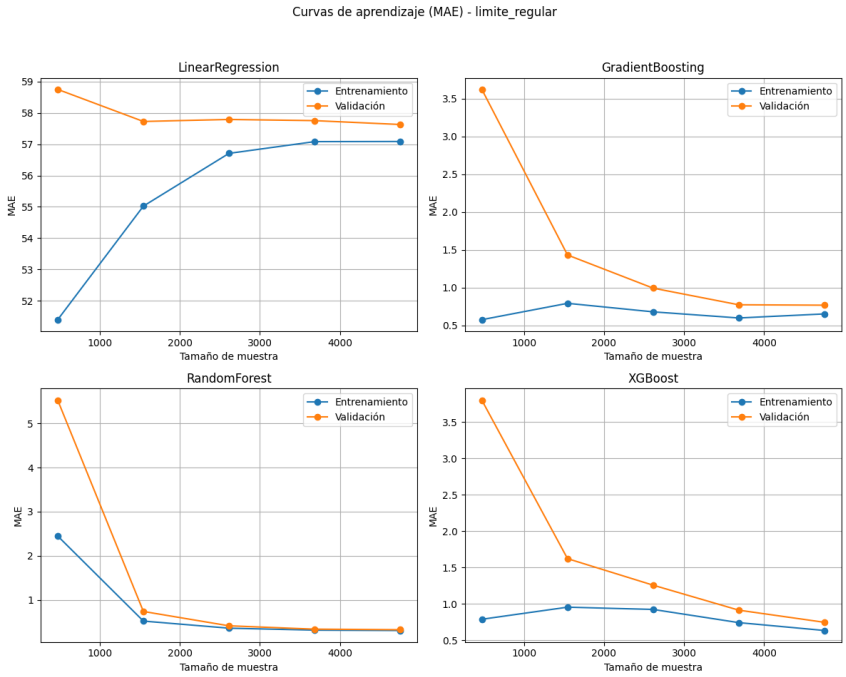


Figura 14: Curvas de aprendizaje para el límite regular (con ajuste de hiperparámetros).

| **Modelo** | **MAE** | **RMSE** | **$R^2$** |
|---|---|---|---|
| LinearRegression | 56.67 | 69.25 | 0.35 |
| GradientBoosting | 0.49 | 0.72 | 0.9999 |
| RandomForest | 0.29 | 1.08 | 0.9998 |
| XGBoost | 0.64 | 0.92 | 0.9999 |

Cuadro 11: Métricas para el límite regular (con ajuste de hiperparámetros).

## Conclusiones del modelado por intervalos

Los resultados muestran que el enfoque de predicción de intervalos personalizados permite obtener un ajuste mucho más preciso que la predicción de un valor único de prima. En todos los límites y para todos los modelos de ensamble (Random Forest, Gradient Boosting y XGBoost), los valores de MAE y RMSE son extremadamente bajos y los coeficientes de determinación $R^2$ se acercan a 1, indicando un ajuste casi perfecto.

Incluso el modelo lineal mejora notablemente su desempeño respecto al modelado de la prima puntual, aunque sigue siendo superado por los modelos de ensamble y boosting.

En conclusión, el modelado por intervalos personalizados es una alternativa mucho más robusta y útil para este tipo de datos, ya que permite acomodar la alta variabilidad y dispersión de la variable objetivo, proporcionando referencias realistas y personalizadas para la toma de decisiones en seguros médicos.