Universidad de La Habana

Predicción del coste de la prima "out of pocket" de seguro médico en MEPS 2022

Diego J. Puentes Fernandez
Ciencia de Datos

Contents

1 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.

2 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 de enfermedades crónicas.

3 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:

Métrica	**Valor**
Total de personas	22431
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
Condiciones médicas distintas	206
Media de condiciones por persona	4.80
Top 5 condiciones más comunes	CIR007: 5391, END010: 4268, MUS010: 3061, END002: 2

4 Descripción de la variable objetivo

La variable objetivo es `prima out of pocket editada`, que es la prima mensual que pagan las personas por mantener su seguro médico, arreglada por el MEPS después de corregir errores (por eso el "editada"). Sus detalles son:

- Entradas válidas: 10189
- Mínimo: 1.00
- Máximo: 4583.33
- Media: 358.37
- Mediana: 270.83
- Q1: 136.00
- Q3: 478.83
- Varianza: 113310.34
- Desviación estándar: 336.62
- Dispersión (máx - mín): 4582.33
- Cantidad de outliers (—z—¿3): 186

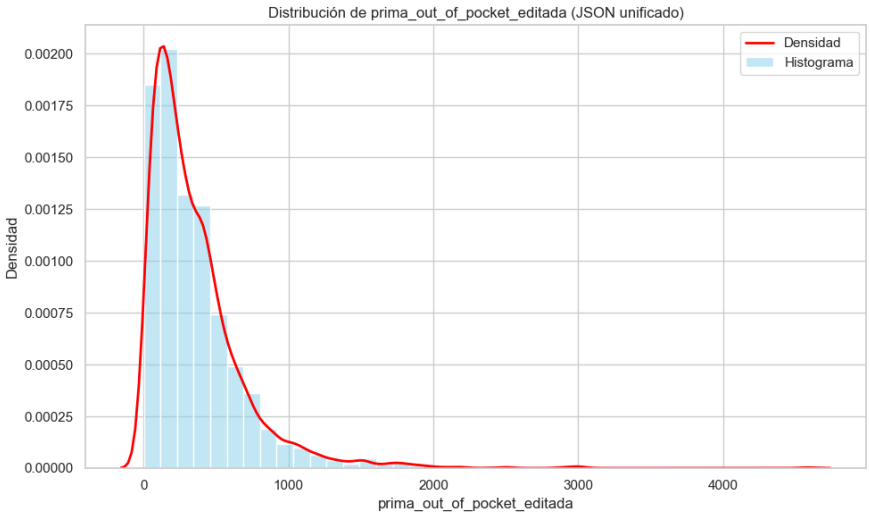

Figure 1: Distribución de la variable objetivo

5 Caracterización de outliers de la variable objetivo

A continuación se presenta un panorama general de las personas que son outliers en la variable objetivo, comparando sus características respecto al grupo general. Se observa que no presentan diferencias extremas ni especiales, por lo que se decidió prescindir de ellos para el entrenamiento de los modelos.

Métrica	**Outliers**	**Resto**
Cantidad de outliers (—z—¿3)	186	10003
Porcentaje de outliers	1.83%	98.17%
Media de edad	36.76	42.13
Media de ccsr_num_total	2.15	2.96
Media de ccsr_otra_condicion	1.78	2.26
Proporción sexo_Male	0.46	0.48
region_Midwest	0.27	0.23
region_Northeast	0.18	0.17
region_South	0.19	0.35
region_West	0.35	0.26
estado_civil_Married	0.55	0.48
estado_civil_Never married	0.20	0.23
estado_civil_Separated	0.00	0.01
estado_civil_Under 16 - not applicable	0.23	0.15
estado_civil_Widowed	0.01	0.04
raza_etnicidad_Non-Hispanic Asian only	0.06	0.07
raza_etnicidad_Non-Hispanic Black only	0.04	0.11
raza_etnicidad_Non-Hispanic Other race or multi-race	0.02	0.04
raza_etnicidad_Non-Hispanic White only	0.74	0.64
ccsr_Essential hypertension	0.13	0.21
ccsr_Disorders of lipid metabolism	0.11	0.16

ccsr_Diabetes mellitus without complication	0.02	0.08
ccsr_Bacterial infections	0.01	0.02
ccsr_Osteoarthritis	0.02	0.06
ccsr_Cataract and other lens disorders	0.01	0.02
ccsr_Esophageal disorders	0.02	0.06
ccsr_Retinal and vitreous conditions	0.01	0.02
ccsr_Other general signs and symptoms	0.02	0.03
ccsr_Abnormal findings without diagnosis	0.01	0.02

6 Primera Propuesta de Modelo

A partir de los datos obtenidos y procesados, se construyó un dataset final con las siguientes columnas principales. En la siguiente tabla se indica si la columna se mantuvo igual, se le aplicó codificación (encoding) y el tipo de transformación realizada, de acuerdo al script de procesamiento:

Columna	**Tipo de transformación**
edad	-
estado_salud_percibido	Label Encoding (ordinal)
ccsr_num_total	-
ccsr_otra_condicion	-
categoria_pobreza	Label Encoding (ordinal)
tiene_historial_empleo	Binaria (1 si tiene historial, 0 si no)
horas_por_semana	Media de horas, imputada si falta
sexo_Male	One-hot encoding (binaria)
raza_etnicidad_Non-Hispanic Asian only	One-hot encoding (binaria)
raza_etnicidad_Non-Hispanic Black only	One-hot encoding (binaria)
raza_etnicidad_Non-Hispanic Other race or multi-race	One-hot encoding (binaria)
raza_etnicidad_Non-Hispanic White only	One-hot encoding (binaria)
estado_civil_Married	One-hot encoding (binaria)
estado_civil_Never married	One-hot encoding (binaria)
estado_civil_Separated	One-hot encoding (binaria)
estado_civil_Under 16 - not applicable	One-hot encoding (binaria)
estado_civil_Widowed	One-hot encoding (binaria)
region_Midwest	One-hot encoding (binaria)
region_Northeast	One-hot encoding (binaria)
region_South	One-hot encoding (binaria)
region_West	One-hot encoding (binaria)
ccsr_Essential hypertension	One-hot encoding (binaria, top 10 CCSR)
ccsr_Disorders of lipid metabolism	One-hot encoding (binaria, top 10 CCSR)
ccsr_Diabetes mellitus without complication	One-hot encoding (binaria, top 10 CCSR)
ccsr_Bacterial infections	One-hot encoding (binaria, top 10 CCSR)
ccsr_Osteoarthritis	One-hot encoding (binaria, top 10 CCSR)
ccsr_Cataract and other lens disorders	One-hot encoding (binaria, top 10 CCSR)

ccsr_Esophageal disorders	One-hot encoding (binaria, top 10 CCSR)
ccsr_Retinal and vitreous conditions	One-hot encoding (binaria, top 10 CCSR)
ccsr_Other general signs and symptoms	One-hot encoding (binaria, top 10 CCSR)
ccsr_Abnormal findings without diagnosis	One-hot encoding (binaria, top 10 CCSR)
seguro_Public only	One-hot encoding (binaria)
seguro_Uninsured	One-hot encoding (binaria)
prima_out_of_pocket_editada	Variable objetivo (sin transformación)

El dataset resultante permite representar de forma numérica y categórica las principales características demográficas, socioeconómicas y de salud de cada individuo, facilitando su uso en modelos de aprendizaje automático supervisado.

La distribución de las principales variables del dataset final se muestra a continuación. Estas visualizaciones permiten apreciar la diversidad y balance de los datos empleados para el modelado:

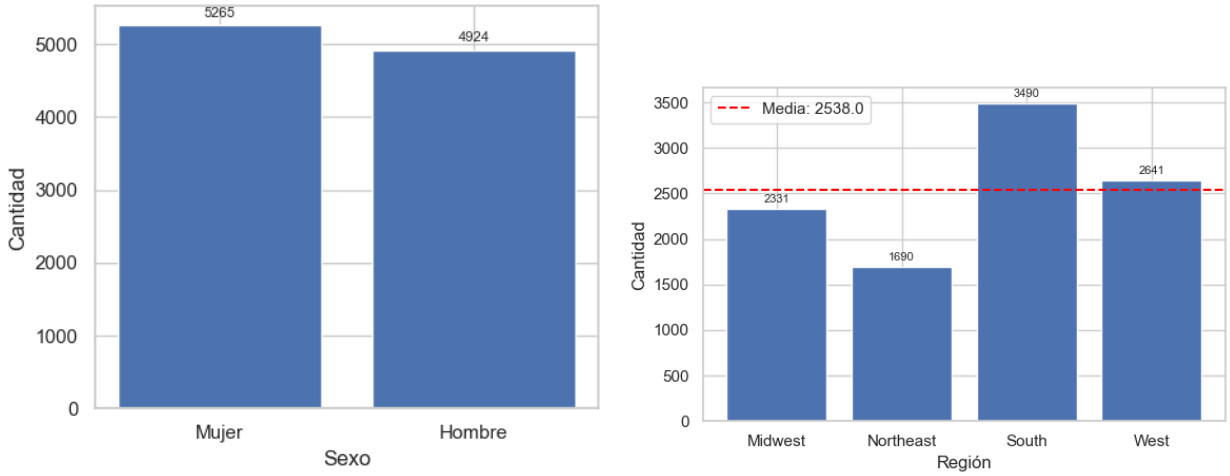

Figure 2: Distribución de sexo y región

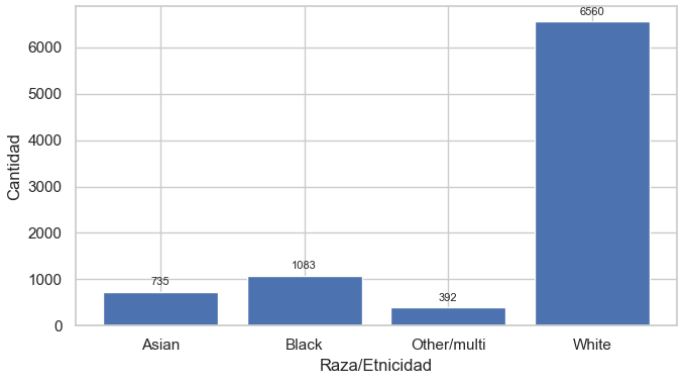

Figure 3: Distribución por etnia

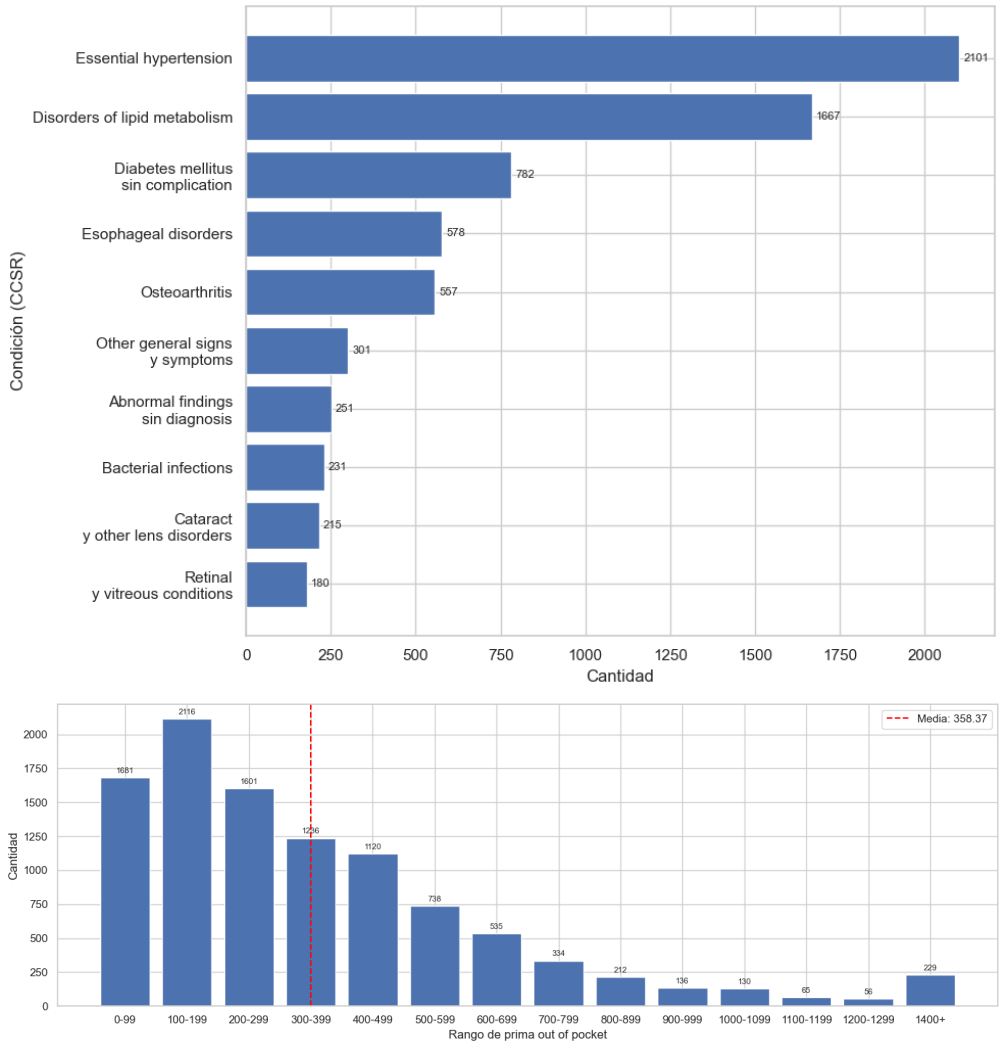

Figure 4: Distribución por enfermedades y variable objetivo

Justificación de la selección de modelo

La selección del modelo para la predicción de la prima "out of pocket" se fundamentó en la exploración de la relación entre las variables predictoras y la variable objetivo. El análisis exploratorio mostró que la relación entre las variables independientes y la prima no es lineal, lo que sugiere que modelos lineales simples pueden no capturar adecuadamente la complejidad de los datos.

No obstante, como punto de partida y para establecer una línea base de comparación, se implementó primero una regresión lineal. Esto permitió obtener una referencia inicial del desempeño y entender las limitaciones de los modelos lineales en este contexto.

Esta estrategia permitió comparar de manera objetiva los resultados y justificar la necesidad de emplear modelos no lineales para mejorar la capacidad predictiva sobre la variable objetivo.

Resultados del ajuste de modelos y discusión

Tras el ajuste de hiperparámetros y la comparación de varios modelos (regresión lineal, Random Forest, Gradient Boosting y XGBoost), se observó que ninguno de los modelos logró un desempeño satisfactorio respecto al dataset y la tarea planteada. A continuación se muestra la tabla de métricas obtenidas para cada modelo:

Modelo	MAE	RMSE	R^2	Bias	Varianza
LinearRegression	191.60	252.71	0.06	-0.15	4786.28
RandomForest	187.35	248.62	0.09	2.86	9098.07
GradientBoosting	186.89	247.47	0.09	0.84	7156.23
XGBoost	186.80	247.24	0.10	0.70	7099.15

Table 4: Comparación de modelos tras ajuste de hiperparámetros.

Como se aprecia en la tabla y en las gráficas, los valores de R^2 son muy bajos (cercanos a cero), y los errores absolutos y cuadráticos medios (MAE y RMSE) son elevados en todos los casos. Incluso tras el ajuste de hiperparámetros, los modelos no logran capturar la complejidad de la variable objetivo ni mejorar significativamente respecto a una predicción trivial. Esto indica que la relación entre las variables predictoras y la prima "out of pocket" es débil o está fuertemente condicionada por factores no presentes en el dataset.

Esta dificultad se explica, en parte, por la naturaleza de la variable objetivo: su distribución es altamente dispersa y presenta una gran cantidad de valores atípicos, como se mostró en el análisis exploratorio. Además, la variable objetivo depende de factores externos (como políticas de aseguradoras, subsidios, condiciones particulares de los individuos) que no están reflejados en las variables disponibles. La combinación de alta dispersión, presencia de outliers y falta de variables explicativas clave limita la capacidad predictiva de los modelos, incluso los más complejos y ajustados.

En conclusión, los resultados obtenidos muestran que, dadas las características del dataset y la variable objetivo, la predicción precisa de la prima "out of pocket" es una tarea desafiante y limitada con la información disponible.

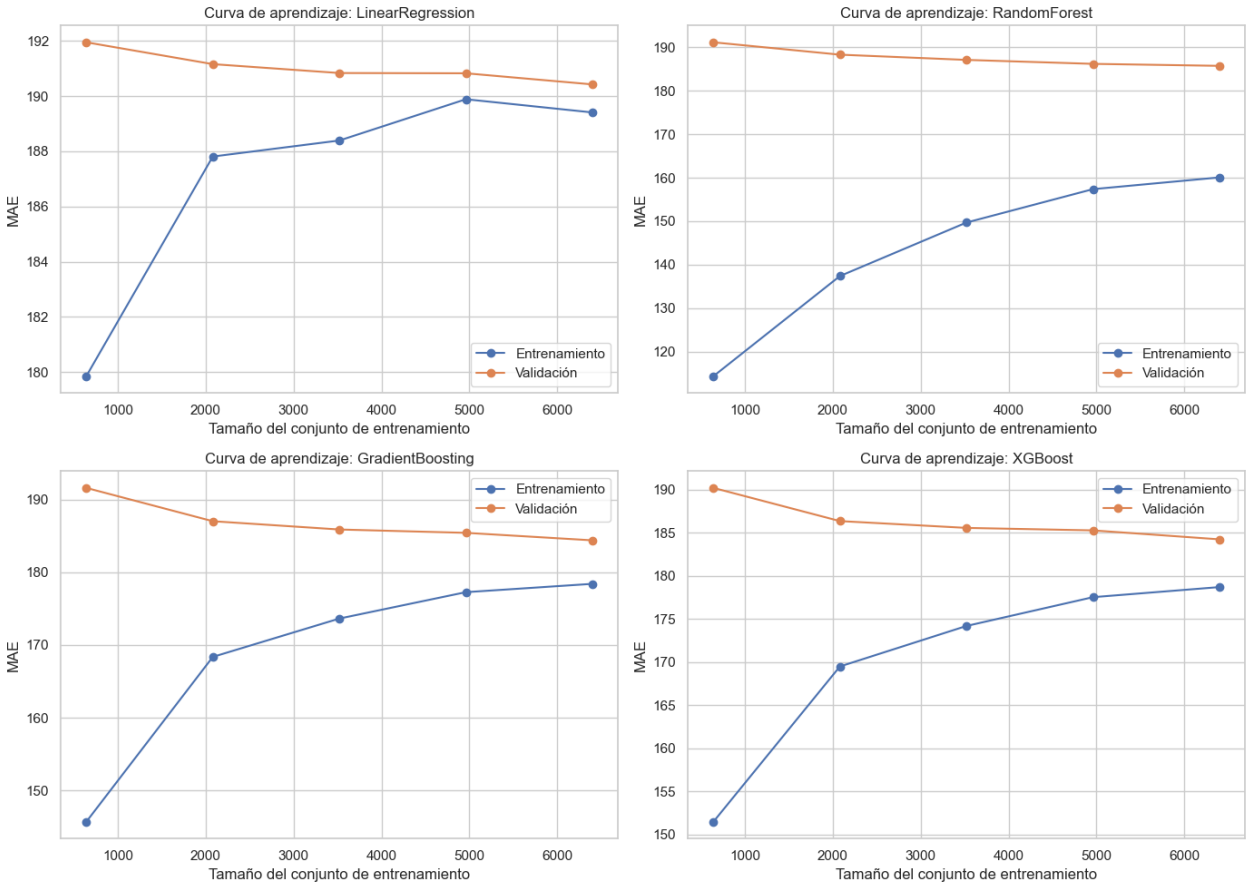

Figure 5: Curvas de aprendizaje y ajuste de los modelos principales.

7 Segunda Propuesta de Modelado: Cambiando la Pregunta

Ante los resultados insatisfactorios obtenidos al intentar predecir el valor exacto de la prima "out of pocket" para cada persona, se exploró un nuevo enfoque de modelado. En lugar de predecir un único valor, se propuso estimar, para cada individuo, un conjunto de tres valores que definan los límites superiores de rangos de conveniencia para la prima esperada, basados en personas similares.

Definición de los rangos de conveniencia

La idea central es que, para cada persona, el modelo devuelva una lista de tres valores:

- El primer valor corresponde al límite superior de la categoría **Excelente**.
- El segundo valor es el límite superior de la categoría **Bueno**.
- El tercer valor es el límite superior de la categoría **Regular**.

De este modo, se obtiene un rango personalizado de referencia para la prima, en vez de una predicción puntual, lo que resulta más robusto ante la alta dispersión y variabilidad de la variable objetivo.

Lógica y construcción de los rangos

Para construir estos límites, se siguió la siguiente lógica:

1. Para cada persona del conjunto de test, se identifican los individuos más similares en el conjunto de entrenamiento, utilizando las variables predictoras (excluyendo la variable objetivo para evitar fuga de información).
2. Sobre el conjunto de personas similares, se calcula la distribución de la prima "out of pocket" y se determinan los percentiles que definen los límites de las categorías: por ejemplo, el percentil 25 para el límite de **Excelente**, el percentil 50 para **Bueno** y el percentil 75 para **Regular**. Estos valores pueden ajustarse según la conveniencia o el criterio del análisis.
3. Así, para cada persona, el modelo devuelve una terna de valores que acotan los rangos de prima esperada según su perfil, en vez de un único valor.

Prevención de fuga de información

Un aspecto fundamental de este enfoque es evitar la fuga de información (*data leakage*). Para ello, la búsqueda de personas similares y el cálculo de los rangos se realiza siempre utilizando únicamente el conjunto de entrenamiento, sin acceder a los valores reales de la variable objetivo en el conjunto de test. De este modo, se garantiza que la estimación de los rangos es válida y no está contaminada por información futura o no disponible en un escenario real.

8 Evaluación de Modelos del Tercer Modelado: Predicción de Límites Personalizados

En este tercer enfoque, el objetivo fue predecir no un valor único de prima, sino tres límites personalizados (**Excelente**, **Bueno**, **Regular**) para cada persona, utilizando un modelo multisalida (*MultiOutputRegressor*) basado en Random Forest y otros algoritmos. El dataset se dividió en un 70% para entrenamiento y un 30% para prueba, asegurando que la construcción de los límites personalizados para cada persona del test se realizó **sin fuga de información**, es decir, usando sólo los datos de entrenamiento para definir los vecinos y los percentiles.

El ajuste de hiperparámetros se realizó mediante **GridSearchCV**, probando combinaciones de parámetros para cada modelo y seleccionando la mejor según validación cruzada. Además, se aplicó **cross-validation** (validación cruzada 5-fold) para estimar el desempeño real de los modelos y evitar sobreajuste.

Comportamiento general del modelo

El modelo multisalida mostró un desempeño muy superior al de la predicción directa de la prima. Los errores (MAE y RMSE) fueron considerablemente menores y los valores de R^2 mucho más altos, indicando que los modelos lograron capturar la estructura de los datos y predecir rangos personalizados de prima con alta precisión. Esto se debe a que el enfoque de rangos es más robusto ante la alta dispersión y variabilidad de la variable objetivo, y aprovecha la información de personas similares en el dataset.

Resultados por límite

Límite Excelente

Modelo	MAE	RMSE	R^2	Bias	Varianza
LinearRegression	34.89	44.95	0.23	-0.03	858.45
GradientBoosting	15.44	21.12	0.83	0.04	2144.26
RandomForest	8.74	14.28	0.92	-0.04	2455.73
XGBoost	15.12	20.43	0.84	0.09	2113.49

Table 5: Métricas para el límite Excelente

Límite Bueno

Modelo	MAE	RMSE	R^2	Bias	Varianza
LinearRegression	60.27	74.94	0.23	-0.01	1771.51
GradientBoosting	20.77	28.53	0.89	0.00	5222.17
RandomForest	18.10	26.27	0.91	0.11	5476.06
XGBoost	20.59	28.37	0.89	0.10	5173.39

Table 6: Métricas para el límite Bueno

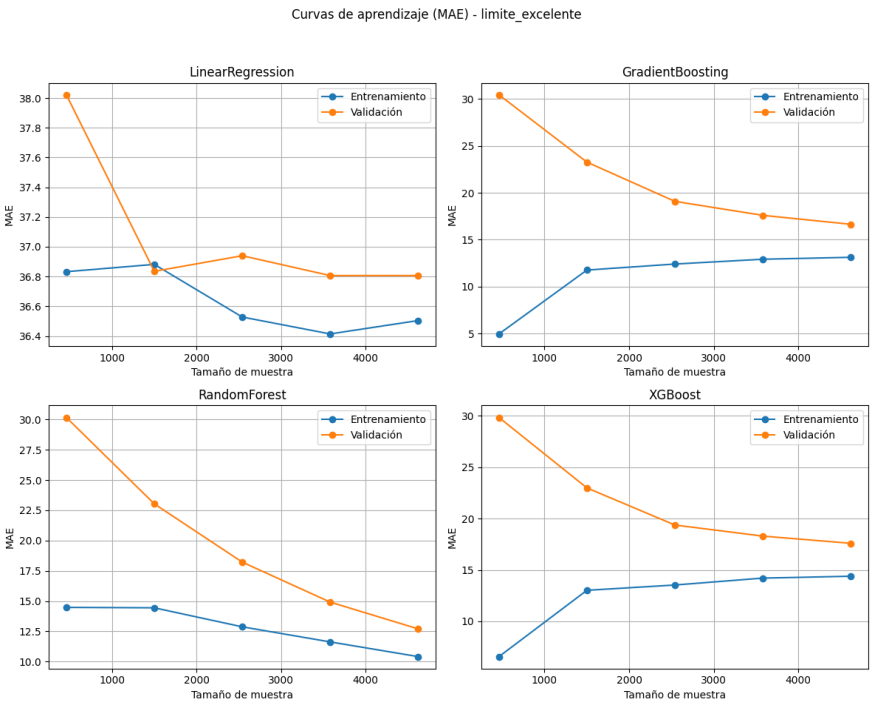

Figure 6: Curva de aprendizaje para el límite Excelente

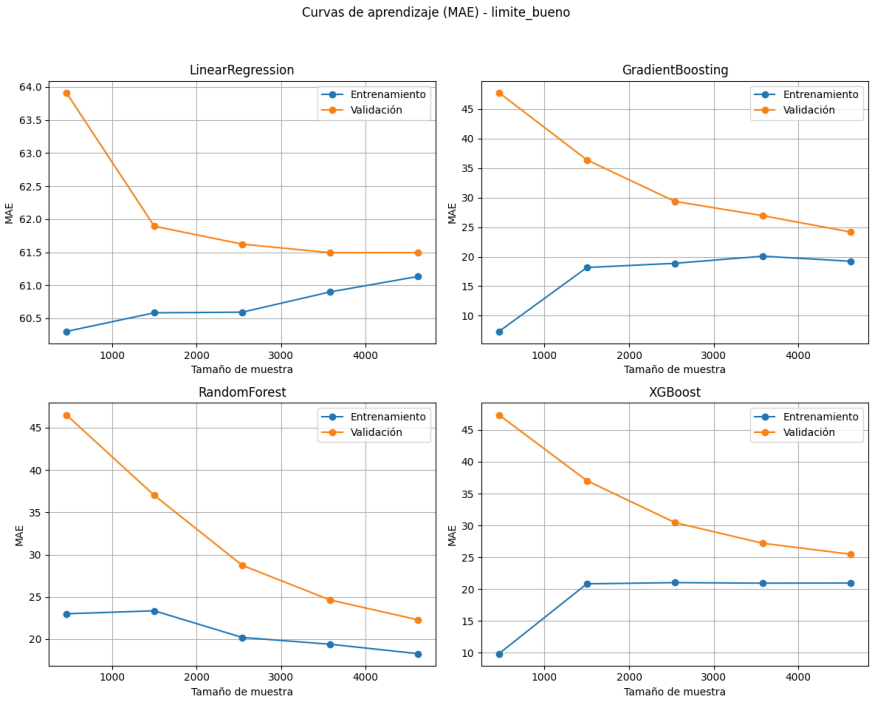

Figure 7: Curva de aprendizaje para el límite Bueno

Límite Regular

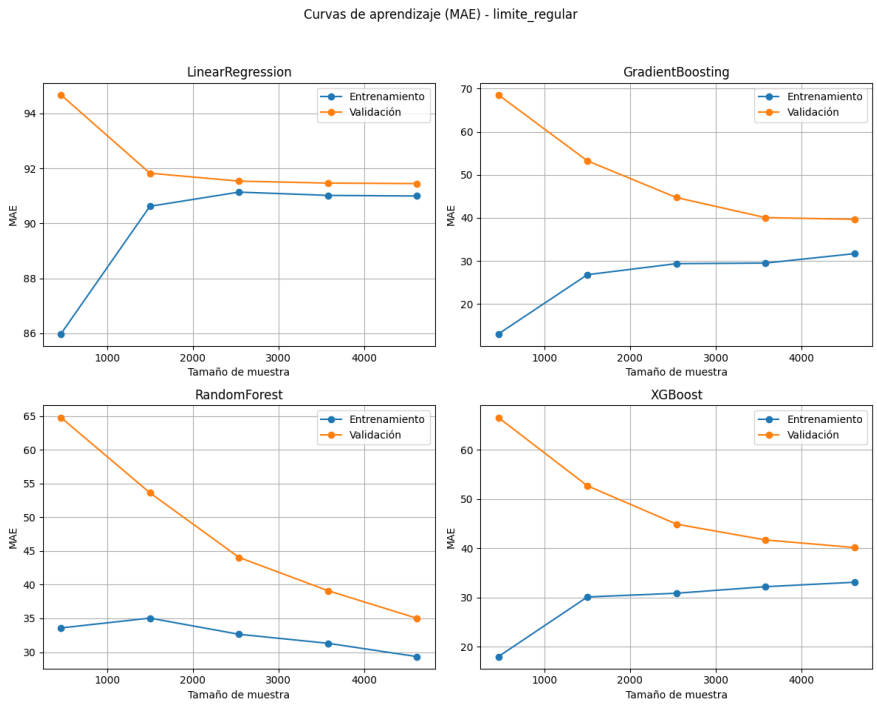

Figure 8: Curva de aprendizaje para el límite Regular

Modelo	MAE	RMSE	R^2	Bias	Varianza
LinearRegression	90.75	115.58	0.14	0.02	2354.56
GradientBoosting	34.72	49.76	0.84	0.16	10893.46
RandomForest	28.69	45.16	0.87	0.11	11379.11
XGBoost	34.47	49.43	0.84	0.09	10713.24

Table 7: Métricas para el límite Regular

Validación cruzada y ajuste de hiperparámetros

Para cada modelo y cada límite, se realizó validación cruzada 5-fold y ajuste de hiperparámetros. Las siguientes tablas muestran los resultados promedio de validación y entrenamiento para cada modelo y cada límite:

Límite Excelente

Modelo	MAE (val)	RMSE (val)	R^2 (val)	MAE (train)	RMSE (train)	R^2 (train)
LinearRegression	36.74	48.13	0.26	36.53	47.90	0.27
GradientBoosting	26.37	35.66	0.60	25.82	34.82	0.62
RandomForest	4.71	9.22	0.97	1.81	3.57	1.00
XGBoost	10.83	15.36	0.92	6.58	9.63	0.97

Table 8: Validación cruzada para el límite Excelente

Límite Bueno

Modelo	MAE (val)	RMSE (val)	R^2 (val)	MAE (train)	RMSE (train)	R^2 (train)
LinearRegression	61.51	75.97	0.23	61.17	75.54	0.24
GradientBoosting	42.26	54.59	0.60	41.28	53.29	0.62
RandomForest	7.97	15.08	0.97	3.08	6.00	1.00
XGBoost	14.59	21.02	0.94	8.94	13.05	0.98

Table 9: Validación cruzada para el límite Bueno

Límite Regular

Modelo	MAE (val)	RMSE (val)	R^2 (val)	MAE (train)	RMSE (train)	R^2 (train)
LinearRegression	91.54	115.86	0.14	91.01	115.26	0.15
GradientBoosting	59.37	79.21	0.60	58.05	77.47	0.62
RandomForest	15.31	27.96	0.95	5.79	10.69	0.99
XGBoost	25.43	37.33	0.91	15.57	23.25	0.97

Table 10: Validación cruzada para el límite Regular

Discusión de resultados y comparación de modelos

Los resultados muestran que **Random Forest** y **XGBoost** son los modelos con mejor desempeño en la predicción de los límites personalizados, alcanzando valores de R^2 superiores a 0.9 y errores absolutos (MAE) muy bajos en comparación con los otros modelos. El modelo Random Forest, en particular, logra el menor MAE y el mayor R^2 en los tres límites, lo que lo posiciona como la mejor opción para este problema.

Este enfoque de predicción de rangos funciona mejor que la predicción directa del valor único de la prima porque aprovecha la información de personas similares y reduce el impacto de la alta dispersión y los outliers en la variable objetivo. El dataset favorece este enfoque porque, aunque la relación entre las variables predictoras y la prima es débil para una predicción puntual, sí permite agrupar a las personas en perfiles similares y estimar rangos de referencia mucho más estables y útiles para la toma de decisiones.

Implementación técnica y despliegue

Para implementar este modelo, se utilizó **MultiOutputRegressor** con Random Forest, permitiendo predecir los tres límites de manera simultánea y eficiente. El modelo entrenado se exportó y se sirvió a través de una página web interactiva construida con **Streamlit**, donde los usuarios pueden ingresar sus características y obtener su rango personalizado de prima esperada de forma sencilla y visual.

9 Conclusiones

En este trabajo se exploraron diferentes enfoques para la predicción de la prima "out of pocket" de seguro médico usando datos de MEPS 2022. Se demostró que la predicción directa del valor único de la prima es una tarea muy difícil debido a la alta dispersión, presencia de outliers y falta de variables explicativas clave en el dataset. Sin embargo, el enfoque de predicción de límites personalizados basado en personas similares y modelos multisalida permitió obtener resultados mucho más robustos y útiles, con errores bajos y alta capacidad explicativa. La solución final, implementada y desplegada en una interfaz web, permite a los usuarios obtener rangos personalizados de referencia para su prima, facilitando la toma de decisiones informadas y realistas en un contexto de alta incertidumbre y variabilidad.