Predicción del Gasto de Bolsillo en Salud de los hogares en Colombia Usando Modelos de Aprendizaje Automático

Juan Sebastián Parada Portilla

Proyecto Aplicado para optar al título de Magister en Ciencia de Datos

Directora: Delia Ortega Lenis

Facultad de Ingeniería y Ciencias
Maestría en Ciencia de Datos
Pontificia Universidad Javeriana
Diciembre 2025

FICHA RESUMEN

TÍTULO: Predicción del Gasto de Bolsillo en Salud de los Hogares en Colombia Usando Modelos de Aprendizaje Automático.

1. **ÁREA DE TRABAJO**: ciencia de datos, salud pública, y economía de la salud. Este enfoque multidisciplinario combina el análisis estadístico y el aprendizaje automático para abordar cuestiones relevantes en el sector de la salud, enfocándose en la predicción y análisis de gastos de salud de los hogares, lo cual tiene implicaciones significativas en las políticas de salud pública y la gestión económica en el sector salud.

2. **TIPO DE PROYECTO**: Proyecto aplicado

3. **ESTUDIANTE**: Juan Sebastián Parada Portilla

4. **CORREO ELECTRÓNICO**: jsparadapor@javerianacali.edu.co

5. **DIRECCIÓN Y TELÉFONO**: a solicitud.

6. **DIRECTOR**: Delia Ortega Lenis

7. **VINCULACIÓN DEL DIRECTOR**: Cátedra

8. **CORREO ELECTRÓNICO DEL DIRECTOR**: delia.ortega@javerianacali.edu.co

9. **PALABRAS CLAVE**: Gasto de Bolsillo en salud, aprendizaje automático, modelos predictivos, políticas públicas en salud, análisis estadístico.

10. **FECHA DE INICIO**: 1 de febrero de 2024

11. **DURACIÓN ESTIMADA**: 10 meses

12. **RESUMEN**: Este estudio desarrolla modelos de aprendizaje automático para predecir el gasto de bolsillo en salud de los hogares colombianos. Utilizando datos de la Encuesta de Calidad de Vida (ECV), se identificaron variables clave como la presencia de enfermedades crónicas en el hogar, el ingreso del hogar, el tamaño del hogar, el estado de salud y la afiliación al sistema de seguridad social. Inicialmente, se exploraron modelos de regresión, pero debido a la alta proporción de valores nulos (85 % de los hogares no reportan gasto en salud), su desempeño fue limitado. Para abordar este problema, se transformó la variable dependiente en una binaria y se aplicaron modelos de clasificación, incluyendo Random Forest, Gradient Boosting y regresión logística, optimizados con la técnica SMOTE para balancear las clases. Los resultados muestran que los modelos de clasificación superan a los de regresión, con Random Forest y Gradient Boosting logrando los mejores desempeños en términos de ROC AUC. Este estudio proporciona herramientas útiles para el diseño de políticas públicas basadas en evidencia, permitiendo identificar hogares con mayor riesgo de incurrir en altos gastos en salud y facilitando intervenciones para reducir el impacto financiero en las familias colombianas.

Dedicatoria

A todos los pacientes crónicos del país y sus familias.

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

1.1. Introducción

El sistema de salud colombiano, establecido mediante la Ley 100 de 1993, ha logrado avances significativos en términos de cobertura universal y protección financiera, consolidándose como un referente regional en América Latina. Sin embargo, enfrenta importantes desafíos relacionados con la sostenibilidad financiera, la equidad en el acceso a los servicios y la calidad de la atención en salud. Dentro de este panorama, el gasto de bolsillo en salud, entendido como los pagos directos realizados por los hogares para cubrir servicios no financiados por el sistema de aseguramiento, es una métrica clave para evaluar el impacto financiero de las políticas públicas en los hogares.

A pesar de que Colombia presenta un gasto de bolsillo relativamente bajo en comparación con los estándares de la Organización para la Cooperación y el Desarrollo Económicos (OCDE) y otros países de la región, persisten profundas desigualdades que afectan de manera desproporcionada a los hogares de menores ingresos y a las zonas rurales. Estas disparidades no solo agravan las brechas de pobreza multidimensional, sino que también reflejan las limitaciones en el acceso equitativo a servicios de salud esenciales. Comprender y predecir el comportamiento del gasto de bolsillo es, por tanto, una necesidad urgente para diseñar intervenciones efectivas que mitiguen estos efectos adversos.

Inicialmente, este proyecto abordó el problema utilizando modelos de regresión, con el objetivo de predecir el gasto total anual en salud como una variable continua. Sin embargo, los resultados obtenidos fueron poco satisfactorios debido a las características particulares de la variable dependiente. Específicamente, se encontró que más del 80 % de los hogares reportan un gasto nulo, mientras que el gasto de los hogares restantes presenta una alta dispersión y heterogeneidad. Esta distribución no solo limitó la capacidad predictiva de los modelos de regresión, sino que también generó métricas de desempeño significativamente bajas, lo que llevó a cuestionar la idoneidad de este enfoque.

En respuesta a estas limitaciones, se optó por transformar la variable dependiente en una binaria, clasificando a los hogares según la presencia o ausencia de gasto de bolsillo en salud. Esta decisión permitió simplificar el problema y adecuarlo a técnicas de clasificación, más apropiadas para manejar distribuciones desbalanceadas como la observada. Además, se adoptó la técnica de sobremuestreo *SMOTE* (*Synthetic Minority Oversampling Technique*) para balancear las clases y mejorar la

capacidad de los modelos para identificar la clase minoritaria.

Posteriormente, se evaluaron cuatro modelos de clasificación: *Gradient Boosting*, *Random Forest*, Árbol de Decisión y Regresión Logística. Estos modelos fueron seleccionados por su capacidad para manejar relaciones no lineales y captar interacciones complejas entre las variables independientes, como el ingreso anual del hogar, el tamaño del hogar y las percepciones sobre la calidad del sistema de salud. A través de un preprocesamiento exhaustivo, que incluyó la estandarización de variables cuantitativas, la codificación de variables categóricas y la reagrupación de categorías infrecuentes, se garantizó la calidad y consistencia de los datos utilizados.

Los resultados de este enfoque no solo permiten un análisis más robusto del gasto de bolsillo en salud, sino que también ofrecen herramientas prácticas para los formuladores de políticas públicas. La capacidad de predecir la ocurrencia de gasto de bolsillo, junto con la identificación de sus determinantes clave, tiene el potencial de informar intervenciones más equitativas y focalizadas en el sistema de salud colombiano. Este trabajo, por tanto, se posiciona como un puente entre la investigación académica y las necesidades del sector salud, contribuyendo entender los caminos para reducir las desigualdades en el acceso y a promover la sostenibilidad financiera del sistema.

Este trabajo se estructura de la siguiente manera. En primer lugar, se define claramente el problema a abordar y se describe su formulación. A continuación, se establecen el objetivo general y los objetivos específicos del estudio. En tercer lugar, se desarrolla el marco teórico, en el cual se presentan las definiciones clave, la fundamentación teórica de los modelos utilizados y los antecedentes relevantes. En cuarto lugar, se detalla la metodología empleada para calcular el gasto de bolsillo en salud de los hogares colombianos a partir de la Encuesta de Calidad de Vida, incluyendo la selección de las variables explicativas más relevantes para usar en los modelos predictivos. En quinto lugar, se realiza un análisis exploratorio de datos, que sirve como base para la comprensión inicial del fenómeno estudiado. En sexto lugar, se presentan los modelos de regresión de machine learning y los resultados obtenidos, discutiendo las limitaciones identificadas. Posteriormente, en el séptimo capítulo, se describe el desarrollo de los modelos de clasificación de machine learning, explicando las métricas de evaluación y la comparación entre ellos. Finalmente, se exponen las conclusiones del trabajo, destacando los principales hallazgos y su relevancia para el sistema de salud colombiano.

Capítulo 2

Definición del problema

2.1. Planteamiento del problema

El sistema de salud de Colombia tiene origen en la Ley 100 de 1993, en la cual se expone la necesidad de mejorar la eficiencia en el gasto, la cobertura poblacional y acceso de los servicios en salud, así como la calidad en todos los ámbitos. El arreglo institucional bajo el cual opera el sistema de salud actual es el de competencia regulada, el cual tiene tres características esenciales: (1) existe un mercado de aseguramiento, en el cual hay una prima fija por afiliado (Unidad de Pago por Capitación), por lo que las Empresas Promotoras de Salud (EPS) no pueden competir por precios o productos diferenciados entre sí, de modo que el único atributo para competir es la calidad; (2) en el mercado de las Instituciones Prestadores de Servicios en Salud (IPS) se permite la competencia vía precios, productos y calidad, de modo que las compañías aseguradoras deben guiar la configuración de su red de IPS basados en estos criterios y; (3) el estado será el garante de que el mercado opere de manera correcta interviniendo cuando sea necesario para corregir las fallas de mercado que se produzcan [1].

Luego de más de tres décadas de funcionamiento del sistema de salud hay una serie de avances y retos. Entre los retos más grandes está la sostenibilidad financiera en un contexto de envejecimiento de la población, la fragmentación geográfica del acceso a los servicios en salud y mejorar la gestión del riesgo en salud. En contraste, los avances más significativos del sistema de salud están relacionados con la cobertura universal, la protección financiera que ofrece el sistema de salud y el bajo gasto de bolsillo. El gasto de bolsillo en salud es la cantidad de dinero que los miembros de un hogar gastan cuando utilizan servicios de salud, y dicho dinero para hacer esos pagos sale de sus ingresos propios (coloquialmente de su "bolsillo") debido a que no está cubierto por un seguro o por un tercer pagador. Colombia es uno de los países de la Organización para la Cooperación y Desarrollo Económico (OCDE) que tiene el segundo menor gasto de bolsillo en salud (1,2 % como proporción del PIB en 2021), lo cual se debe a la cobertura del Plan de Beneficios en Salud (PBS) en conjunto con un sistema de aseguramiento robusto que gestiona eficazmente el riesgo financiero [2]. Lo anterior resulta relevante en el contexto de un país de ingresos medios y en dónde el gasto del gobierno en salud no equipara a la mayor parte de los países de la OCDE. Un bajo gasto de bolsillo en salud además permite que los hogares tengan un mayor ingreso disponible para el consumo e inversión, y permite que menos hogares estén por debajo de la línea de la pobreza monetaria y multidimensional.

Dadas las implicaciones que tiene una variable como el gasto de bolsillo en salud de los hogares, resulta relevante para los *policymakers* conocer cuáles son sus determinates, así como herramientas adecuadas para su predicción. En concreto, un mayor gasto de bolsillo en salud significa menor ingreso disponible de las familias, mayor incidencia de la pobreza monetaria y multidimensional y un crecimiento de la desigualdad de ingresos [3]. Además, recientes modificaciones en la Encuesta de Calidad de Vida (ECV) del Departamento Administrativo de Estadística Nacional (DANE), incluyendo la reducción de preguntas en el módulo de salud, presentan nuevos desafíos para calcular con precisión el gasto de bolsillo. Esta situación subraya la urgencia y relevancia de este estudio, que busca determinar las variables críticas que explican el gasto de bolsillo en salud de los hogares en Colombia y desarrollar un modelos predictivo para tal fin.

2.2. Formulación del problema

En Colombia no existen trabajos académicos sobre predicción de gasto de bolsillo en salud haciendo uso de modelos de aprendizaje automático, la innovación de esta investigación radica en ser pionera del uso de técnicas de ciencia de datos para predecir el gasto de bolsillo, comparando bajo diferentes métricas varios modelos y determinando el más adecuado para predecir el comportamiento de esta variable.

Pregunta principal

¿Qué modelo de aprendizaje automático basado en un conjunto definido de variables se puede implementar para predecir el gasto de bolsillo en salud de los hogares en Colombia?

Preguntas secundarias

- ¿Cuáles son las variables relacionadas con el gasto de bolsillo en salud de los hogares para hacer la predicción del mismo?
- ¿Qué modelos de aprendizaje automático son potencialmente aplicables para predecir el gasto de bolsillo en salud?
- ¿Qué estrategias se utilizarán para entrenar y validar los modelos seleccionados?
- ¿Cómo se determinará la precisión y la eficacia de cada modelo?
- ¿Cuáles son los criterios para evaluar el rendimiento de los modelos en el contexto específico de Colombia?

Capítulo 3

Objetivos del proyecto

3.1. Objetivo General

Desarrollar un modelo de aprendizaje automático para predecir el gasto de bolsillo en salud de los hogares en Colombia, utilizando un conjunto seleccionado de variables.

3.2. Objetivos Específicos

- Determinar las variables críticas que influirán en la predicción del gasto de bolsillo en salud de los hogares en Colombia.
- Identificar los modelos de aprendizaje automático más adecuados para la predicción del gasto de bolsillo en salud basados en la literatura actual y la compatibilidad con el conjunto de datos disponible.
- Desarrollar un protocolo para el entrenamiento y la validación de los modelos de aprendizaje automático seleccionados, asegurando su aplicabilidad y robustez en el contexto de salud colombiano.
- Establecer un conjunto de métricas para evaluar la precisión y el rendimiento de los modelos desarrollados en la predicción del gasto de bolsillo en salud.

Capítulo 4

Marco de Teórico

4.1. Introducción al Gasto de Bolsillo en Salud de los Hogares en Colombia

El Gasto de Bolsillo en Salud (GBS) representa la cantidad de dinero que los hogares destinan directamente a la obtención de servicios sanitarios, sin intermediación de seguros o coberturas por parte de terceros, además, para efectos de este trabajo, se incluye el gasto en medicamentos y transporte, adoptando la definición de GBS ampliada [3]. Este tipo de gasto es un indicador crucial de la accesibilidad y asequibilidad de la atención sanitaria en cualquier país. En Colombia, después de más de tres décadas de funcionamiento del actual modelo de sistema de salud, el GBS ha adquirido una relevancia particular en el contexto de los avances y desafíos que presenta el sector salud.

Colombia ha logrado avances significativos en su sistema de salud, especialmente en términos de cobertura universal y protección financiera [4]. Un hito notable es el bajo porcentaje del GBS como proporción del gasto total en salud, situado en 13,5 % para 2021, el segundo más bajo entre los países de la OCDE [5]. Esta situación se atribuye a la eficiencia del Plan de Beneficios en Salud (PBS) y a un sistema de aseguramiento robusto que ha gestionado eficientemente el riesgo financiero, dando la suficiente cobertura a los hogares para evitar que incurran en altos gastos para solventar problemas de salud. Sin embargo, a pesar de estos avances, el sistema de salud enfrenta retos importantes como la sostenibilidad financiera en un contexto de envejecimiento poblacional, la fragmentación geográfica del acceso a los servicios de salud y la mejora en la gestión del riesgo sanitario.

El GBS tiene implicaciones directas en la economía doméstica. Un menor GBS se traduce en mayor ingreso disponible para consumo e inversión en los hogares, contribuyendo a reducir la pobreza monetaria y multidimensional [6]. En contraste, un incremento en este gasto implica una disminución del ingreso disponible, lo que puede incrementar la pobreza y la desigualdad de ingresos. Esto subraya la relevancia de conocer las variables que influyen en el GBS de un hogar y predecir su comportamiento.

Dada la importancia del GBS en la economía de los hogares y en la formulación de políticas de salud, se vuelve crucial disponer de herramientas que permitan predecir su comportamiento. La predicción precisa del gasto de bolsillo puede ayudar a los reguladores a tomar decisiones informadas para mitigar los posibles impactos negativos sobre los hogares, especialmente en un país de ingresos medios como Co-

lombia, donde el gasto gubernamental en salud no alcanza los niveles de la mayoría de los países de la OCDE.

En resumen, el GBS en Colombia es juega un papel clave en la sostenibilidad financiera de los hogares, en la efectividad de las políticas de salud pública y tiene implicaciones macroeconómicas. Una comprensión rigurosa del GBS y predecir cuáles variables tiene mayor impacto sobre éste, son esenciales para que el sistema de salud tenga mejores resultados.

4.2. Cálculo y Predicción del GBS en Colombia

La predicción del GBS en Colombia exige una metodología adaptada a las características únicas del país, considerando diferencias socioeconómicas y en el sistema de salud. Es crucial que la metodología empleada refleje las variaciones en la estructura de datos disponibles, patrones de gasto en salud, y demografía colombiana. La principal fuente de información para este análisis es la Encuesta de Calidad de Vida de Colombia (ECV), la cual realiza el DANE anualmente y ofrece datos detallados sobre ingresos, gastos, acceso a servicios de salud y demografía de los hogares, esenciales para comprender los patrones de gasto de bolsillo y entrenar modelos de aprendizaje automático eficientes.

La selección de variables relevantes para predecir el GBS es otro aspecto crucial, ya que esto permite comprender dinámicas del hogar y su relación con los gastos en salud. Además, garantizar la calidad de los datos es imprescindible, lo cual incluye procesos de limpieza, normalización y manejo adecuado de valores faltantes o atípicos, para asegurar la precisión y fiabilidad de las predicciones.

Por su parte, el preprocesamiento de datos asegura que los modelos de aprendizaje automático funcionen eficazmente. Esto incluye la transformación de variables categóricas, la normalización de variables numéricas, y la identificación y tratamiento de valores atípicos. Además, se debe realizar una selección de características para identificar las variables más relevantes para el modelo. Cabe decir que sobre este asunto hay un literatura amplia para Colombia en el uso de los microdatos de la ECV para este proceso [3].

Ahora bien, la metodología analítica adoptada para predecir el gasto de bolsillo en salud en Colombia debe tener consistencia con los objetivos y alcances de este proyecto. Se utilizarán técnicas de aprendizaje automático, las cuales serán descritas en la sección siguiente, para construir y validar modelos predictivos. La validación cruzada y otras técnicas de evaluación de modelos serán empleadas para garantizar la precisión y la aplicabilidad de los modelos.

4.3. Encuesta de Calidad de Vida 2023

La ECV fue diseñada en respuesta a la necesidad de caracterizar y ubicar a la población en condiciones de pobreza y analizar otros aspectos del bienestar en Colombia. A lo largo de los años, ha evolucionado para incorporar diversas dimensiones del bienestar, como salud, educación, acceso a servicios, entre otros. Hay algunos aspectos claves de la ECV que vale la pena desatacar por ser la fuente elegida para este trabajo.

4.3.1. Idoneidad de la ECV como fuente de este trabajo

La ECV es una fuente idónea para el desarrollo de este trabajo, ya que proporciona información detallada y representativa a nivel nacional sobre las condiciones socioeconómicas, el acceso a servicios de salud y los patrones de gasto de los hogares. La ECV incluye variables clave como el ingreso del hogar, número de miembros en el hogar, estado de salud, afiliación al sistema de seguridad social y presencia de enfermedades crónicas, lo que permite capturar factores determinantes del gasto en salud. Además, al contar con factores de expansión, la encuesta posibilita la estimación de resultados a nivel poblacional, fortaleciendo la validez y generalización del modelo de predicción. Su metodología rigurosa de recolección y procesamiento de datos asegura la calidad de la información, lo que la convierte en una fuente de información ideal para el análisis y la implementación de modelos de aprendizaje automático en este estudio.

4.3.2. Periodicidad

Dada la relevancia de la operación estadística de la Encuesta Nacional de Calidad de Vida (ECV) para el ciclo de las políticas públicas y la importancia de las temáticas abordadas en la encuesta, el DANE tomó la decisión que su periodicidad fuera anual desde el año 2010. Desde entonces, la Encuesta Nacional de Calidad de Vida (ECV) se ha aplicado todos los años, y su última medición fue en el 2023.

4.3.3. Diseño de Muestreo

El diseño de muestreo de la Encuesta Nacional de Calidad de Vida (ECV) 2023, realizada por el Departamento Administrativo Nacional de Estadística (DANE), se fundamenta en un enfoque probabilístico, multietápico, estratificado y por conglomerados. Este diseño asegura la representatividad y precisión de los datos recolectados, permitiendo estimaciones confiables para diferentes niveles de desagregación geográfica y sociodemográfica. A continuación, se detallan los principales aspectos del diseño:

Tamaño de la muestra La ECV 2023 incluyó un total de **86,405 hogares encuestados**, distribuidos en áreas urbanas y rurales, con desagregación adicional por regiones y departamentos.

Unidades de muestreo

- **Primera etapa:** Selección de municipios sin reemplazo, proporcional al tamaño de su población.
- **Segunda etapa:** Selección de conglomerados de 10 viviendas mediante un muestreo sistemático con ordenamiento geográfico previo.

Estratificación La población fue estratificada por áreas geográficas (cabecera y rural disperso) y características sociodemográficas, asegurando la adecuada representación de cada segmento en la muestra.

Parámetros de cálculo del tamaño de la muestra El cálculo del tamaño de la muestra se realizó considerando los siguientes parámetros:

- **Nivel de Confianza:** 95 % ($Z = 1{,}96$).
- **Margen de Error:** Variable según el indicador de interés; generalmente entre 3 % y 5 %.
- **Proporción Esperada (P):** Para indicadores desconocidos, se asumió $P = 0{,}5$, maximizando la variabilidad.
- **Efecto del Diseño ($Deff$):** Ajuste para reflejar el aumento de varianza debido al diseño complejo. Los valores típicos utilizados fueron entre 1.2 y 1.5.
- **Tasa de No Respuesta (t):** Se estimó con base en encuestas previas. Para compensar las posibles no respuestas, se realizó un ajuste multiplicativo considerando $T = (1 + t)$.

La fórmula utilizada para el cálculo del tamaño de la muestra fue la propuesta por Cochran (1977):

$$n = \frac{N \cdot Z^2 \cdot P \cdot Q}{N \cdot e^2 + P \cdot Q \cdot Z^2} \cdot Deff \cdot O \cdot T \tag{4.3.1}$$

Donde:

- N: Tamaño de la población objetivo.
- e: Margen de error.
- P y Q: Proporción esperada y su complemento ($Q = 1 - P$).
- Z: Nivel de confianza (1.96 para 95 %).
- $Deff$: Efecto del diseño.
- O: Relación entre subdivisión de población y universo de estudio.
- T: Ajuste por no respuesta.

Cobertura geográfica La encuesta abarcó todo el territorio nacional, exceptuando la población de la isla de Providencia y el área rural dispersa de San Andrés.

Metodología de recolección La recolección de datos se realizó mediante entrevistas directas utilizando dispositivos móviles de captura (DMC), asegurando calidad y precisión en la información registrada.

Este diseño muestral garantiza estimaciones representativas y útiles para el análisis del Gasto de Bolsillo en Salud (GBS), cumpliendo con los estándares internacionales y contribuyendo a la formulación de políticas públicas basadas en evidencia.

4.3.4. Recolección de Datos

La recolección de datos en la ECV se realiza mediante entrevistas directas, utilizando dispositivos móviles de captura (DMC). Los encuestadores visitan los hogares tantas veces como sea necesario para entrevistar directamente a todos los miembros de 18 años o más. Este método asegura que la información recolectada sea precisa y completa.

La fase de recolección se organiza en operativos con equipos de trabajo que siguen un sistema de barrido. En zonas urbanas, el barrido se realiza de manera simultánea en segmentos, mientras que en áreas rurales se sigue una estrategia de barrido por rutas. Esta organización permite un control efectivo del proceso y una cobertura exhaustiva de los segmentos seleccionados. Los supervisores asignan viviendas específicas a cada encuestador y reubican recursos según sea necesario para mantener la eficiencia y la equidad en la carga de trabajo.

4.3.5. Procesamiento de Datos

El procesamiento de datos de la ECV incluye varios pasos diseñados para garantizar la calidad y la precisión de la información. Primero, los datos recolectados se consolidan y se codifican según un diccionario de datos predefinido. Luego, se aplican reglas de validación y consistencia para identificar y corregir posibles errores. El DANE utiliza el programa SAS y herramientas específicas como la macro Clan 97 versión 3.1 para el cálculo de ponderadores y la estimación de parámetros y varianzas de los estimadores.

Los ponderadores ajustan los datos recolectados para corregir desajustes de cobertura y pérdidas de muestra. Además, se calibra la muestra a las proyecciones poblacionales publicadas por el DANE para asegurar que los resultados sean representativos de la población objetivo. Este enfoque meticuloso en la recolección y procesamiento de datos garantiza que la ECV proporcione información fiable y útil para el desarrollo de políticas públicas y la toma de decisiones en Colombia.

4.3.6. Temas Principales del Cuestionario

La Encuesta Nacional de Calidad de Vida (ECV) aborda una amplia gama de temas que permiten obtener una visión integral de las condiciones de vida de los hogares en Colombia. Los temas principales incluidos en el cuestionario son:

- Datos de la Vivienda: Incluye información sobre las características físicas de la vivienda, como tipo de vivienda, materiales de construcción, y acceso a servicios públicos.
- Servicios del Hogar: Recopila datos sobre la disponibilidad de cuartos, calidad de los servicios, y prácticas de consumo responsable de agua y energía.
- Características y Composición del Hogar: Identifica a los miembros del hogar y sus características demográficas, como sexo, edad, estado civil, y parentesco con el jefe del hogar.
- Salud: Incluye preguntas sobre afiliación al sistema de salud, uso de servicios médicos, enfermedades crónicas, y percepción del estado de salud.

- Atención Integral a Niños Menores de 5 Años: Investiga la responsabilidad de cuidado de los niños, su acceso a establecimientos educativos, y características de las personas encargadas de su cuidado.
- Educación: Recopila datos sobre asistencia escolar, niveles educativos alcanzados, razones de inasistencia, y características de los establecimientos educativos.
- Fuerza de Trabajo: Examina las fuentes de ingreso de las personas mayores de 12 años, así como las condiciones y calidad del trabajo que realizan.
- Tecnologías de Información y Comunicación (TIC): Indaga sobre el uso de computadoras, Internet, y dispositivos móviles en el hogar.
- Trabajo Infantil: Recopila información sobre la participación de menores en actividades laborales.
- Tenencia y Financiación de la Vivienda: Incluye preguntas sobre la propiedad de la vivienda y las modalidades de financiación.
- Condiciones de Vida y Tenencia de Bienes: Explora la percepción de las condiciones de vida, tenencia de bienes y la percepción de pobreza.

4.3.7. Estructura del Cuestionario

El cuestionario de la ECV 2023 está compuesto por 533 preguntas, divididas en 341 preguntas principales y 192 subpreguntas, organizadas en los siguientes capítulos:

- Capítulo A: Identificación y Control: Identifica la ubicación de las viviendas y los hogares y controla la calidad de la recolección.
- Capítulo B: Datos de la Vivienda: Características físicas y acceso a servicios.
- Capítulo C: Servicios del Hogar: Disponibilidad de servicios y prácticas de consumo.
- Capítulo D: Características y Composición del Hogar: Información demográfica y de parentesco.
- Capítulo E: Salud: Afiliación a salud y uso de servicios médicos.
- Capítulo F: Atención Integral a Niños Menores de 5 Años: Cuidado y acceso a educación para menores.
- Capítulo G: Educación: Asistencia y características educativas.
- Capítulo H: Fuerza de Trabajo: Fuentes de ingreso y condiciones laborales.
- Capítulo I: Tecnologías de Información y Comunicación (TIC): Uso de tecnologías en el hogar.
- Capítulo J: Trabajo Infantil: Participación laboral de menores.

- Capítulo K: Tenencia y Financiación de la Vivienda: Propiedad y financiación de la vivienda.
- Capítulo L: Condiciones de Vida y Tenencia de Bienes: Percepción de condiciones de vida y bienes poseídos.

4.3.8. Calidad de los Datos

La calidad de los datos recolectados en la ECV se garantiza mediante un riguroso proceso de recolección y procesamiento. Algunas de las estrategias utilizadas incluyen:

- Entrenamiento del Personal: Los encuestadores y supervisores reciben una capacitación exhaustiva para asegurar la correcta aplicación del cuestionario y el uso adecuado de los dispositivos móviles de captura (DMC).
- Supervisión y Control: Se implementan sistemas de supervisión en campo para asegurar la calidad y precisión de la información recolectada. Los supervisores verifican el trabajo de los encuestadores y realizan controles de calidad durante el operativo de recolección.
- Validación y Consistencia: Los datos recolectados son sometidos a procesos de validación y revisión de consistencia utilizando reglas predefinidas. Esto incluye la verificación de la coherencia interna de las respuestas y la identificación de posibles errores.
- Codificación y Ponderación: Los datos son codificados según un diccionario de datos predefinido y se aplican ponderadores para ajustar desajustes de cobertura y pérdidas de muestra. Esto asegura que los resultados sean representativos de la población objetivo.

4.3.9. Acceso a los Datos

El acceso a los datos de la ECV está diseñado para facilitar el uso de la información por parte de diversos usuarios, incluyendo investigadores, formuladores de políticas y el público en general. Los datos se publican anualmente a través de boletines técnicos, comunicados de prensa y anexos estadísticos en formato Excel. Además, se publican microdatos anonimizados en el portal Archivo Nacional de Datos (ANDA) para consulta y descarga.

Los resultados de la ECV incluyen desagregaciones a nivel nacional, departamental, y por áreas urbanas y rurales, proporcionando una visión detallada y granular de las condiciones de vida en Colombia. La disponibilidad de estos datos permite realizar análisis profundos y comparaciones a lo largo del tiempo, contribuyendo al diseño e implementación de políticas públicas basadas en evidencia.

4.4. GBS en Colombia en 2023

Como se definió en la sección anterior, el GBS es la cantidad de dinero que los miembros de un hogar gastan cuando utilizan servicios de salud, y el dinero para

hacer esos pagos sale de su bolsillo debido a que no está cubierto por un seguro o por un tercer pagador.

La estimación del valor monetario del GBS de los hogares en salud para Colombia se realiza de acuerdo con lo dispuesto por la literatura reciente [3]. Así pues, se hace a partir de la Encuesta de Calidad de Vida (ECV), usando los microdatos anonimizados disponibles en la página del DANE. La encuesta tiene tres capítulos que son relevantes para llevar a cabo la estimación: (i) servicios del hogar, que tiene la estimación del ingreso total del hogar, (ii) salud, que tiene la mayoría de ítems de gasto en salud, y (iii) gastos de los hogares, que tiene los ítems de higiene y cuidado personal [3].

En el Cuadro4.1 se detallan las preguntas que normalmente son usadas para estimar el GBS en Colombia con la ECV. De modo que para obtener el gasto de bolsillo en salud de un hogar se suman estos valores monetarios. Para la elaboración de este trabajo se realizará el cálculo con las variables descritas.

Cuadro 4.1: Preguntas de la Encuesta de Calidad de Vida usadas para calcular el gasto de bolsillo en salud de los hogares.

Rubro	**Pregunta**	**Código Dic de Datos**
Pago por cobertura en: Seguridad social en salud	6. ¿Cuánto paga o cuánto le descuentan mensualmente a ... para estar cubierto/a por una entidad de seguridad social en salud?	P8551
Aseguramiento privado en salud	10. ¿Cuánto paga o le descuentan mensualmente a ... por concepto de estos planes o seguros voluntarios de salud?	P3176
Pago mensual por uso de servicios	34. Durante los últimos 30 días .... realizó pagos por: (No incluya gastos reportados en hospitalización) (1. Eps, 2. Médico particular, 3. Plan Voluntario)	P3179
Consulta médica general o con especialista		P3178S1A1, P3178S2A1, P3178S3A1
Odontología		P3179S1A1, P3179S2A1, P3179S3A1
Vacunas		P3181S1
Medicamentos		P3182S1
Laboratorios, Rx		P3183S1
Rehabilitación, terapias		P3184S1
Terapias alternativas		P3185S1
Transporte		P3186S1
Pago anual por uso de servicios	35. Durante los ÚLTIMOS DOCE MESES ¿Realizó pagos por:	
Dispositivos	1. Lentes, audífonos o aparatos ortopédicos (muletas, sillas de ruedas, elementos para terapias, etc.)	P3187S2
Cirugías	2. Cirugías o procedimientos ambulatorios? (1. Eps, 2. Médico particular, 3. Plan Voluntario)	P3188S1A1, P3188S2A1, P3188S3A1
Hospitalización	39. ¿Durante los últimos 12 meses .... tuvo que ser hospitalizado/a? 41. ¿Cuánto pagó en total .... por esta hospitalización? (EPS, plan voluntario)	P3189S1A1, P3189S2A1

Para calcular correctamente el gasto de bolsillo en salud de un año en específico, debe considerarse que muchas variables de la ECV son mensuales y otras anuales.

Debido a esto, en esta sección se explica en detalle cómo se realiza el cálculo.

El periodo de referencia de todas las preguntas es los últimos treinta días, con excepción de "Lentes, Audífonos o aparatos ortopédicos" y "Cirugías Ambulatorias o Procedimientos Ambulatorios" cuyo período de referencia es anual. Para estas dos preguntas el gasto se divide por 12 para hacerlo consistente con las demás preguntas. Adicionalmente, para las preguntas "Pago por consulta de salud en los últimos 30 días (por enfermedad)" y "Cuánto pagó en total por esta hospitalización" solo se incluye el gasto si la fuente utilizada para cubrir los costos fue "recursos propios". El total del gasto es estimado para el nivel nacional usando el factor de expansión final de la encuesta para cada mes y luego es multiplicado por 12 para anualizarlo.

Es importante precisar para las estimaciones se hicieron algunas transformaciones de los datos que por transparencia y replicabilidad es necesario especificar. En primera medida, los microdatos que tenían missing en las sumas a nivel de hogar implícitamente se asumieron como cero, que es una práctica usual en los módulos de gasto de las encuestas de hogares [3].

Así las cosas, en la siguiente ecuación se resumen las variables para calcular el gasto de bolsillo en salud para Colombia en el año 2023:

$$\begin{aligned}\text{GBS Anual} = (&\text{Cobertura Seguridad Social} + \text{Aseguramiento Privado} \\ &+ \text{Uso de Servicios (ondon, vac, med, labs, rehab, terap, trans)} \times 12 \\ &+ (\text{Dispositivos} + \text{Cirugías} + \text{Hospitalización})\end{aligned} \tag{4.4.1}$$

El valor total del gasto de bolsillo en salud para Colombia en 2023 fue de: $10.338.526.956.062 lo que equivale a un 0.7 % como proporción del PIB.

Ahora bien, teniendo definida la variable de el gasto de bolsillo en salud en Colombia, es posible continuar con la selección de variales de la ECV que tiene la capacidad de explicar el comportamiento de la variable de análisis.

4.5. Selección de variables de la EVC

Luego de conocer con certeza el monto del GBS para Colombia en 2023, el paso a seguir es seleccionar la variables de la ECV que pueden ser utiles para explicar la variable de análisis.

Al desarrollar modelos de predicción de GBS, es fundamental seleccionar cuidadosamente las variables que capturan tanto las características sociodemográficas del hogar como los determinantes directos e indirectos que influyen en el acceso a servicios de salud y su costo. La calidad de los datos y la relevancia de las variables seleccionadas son esenciales para garantizar un modelo robusto y confiable. Para lograr este objetivo, se seleccionaron variables que capturan no solo las características del hogar, sino también las relacionadas con la cobertura de salud, el nivel educativo y la infraestructura básica de la vivienda, entre otros factores. A continuación se presenta una descripción detallada de las variables elegidas y sus justificaciones en el contexto del análisis.

1. **Ingreso mensual total del hogar (I_HOGAR):** El ingreso total del hogar es el principal determinante de la capacidad de gasto en salud. Según la teoría económica, los hogares con mayores ingresos tienen mayor elasticidad para

gastar en bienes y servicios, incluidos los relacionados con la salud. En contextos donde los sistemas de salud no cubren completamente las necesidades, los hogares con mayores ingresos están en mejor posición para cubrir servicios no financiados por los sistemas públicos, como medicamentos, consultas privadas o procedimientos especializados.

Estudios han demostrado que los hogares en deciles superiores de ingreso reportan mayores gastos en salud, tanto en términos absolutos como relativos. En contraste, los hogares con ingresos bajos enfrentan mayor riesgo de gastos catastróficos en salud, definidos como aquellos que exceden un décima parte de su ingreso disponible [7].

2. **Cantidad de personas por hogar (CANT_PERSONAS_HOGAR):** El tamaño del hogar afecta directamente los gastos en salud, ya que un hogar más grande implica una mayor probabilidad de que al menos un miembro requiera atención médica, medicamentos o tratamientos continuos. En hogares extensos, los gastos de bolsillo tienden a concentrarse en los miembros más vulnerables (niños pequeños, adultos mayores o miembros con enfermedades crónicas). Además, estudios muestran que los hogares más grandes tienen menos probabilidad de incurrir en gastos catastróficos en salud debido a la posibilidad de distribuir los costos entre varios ingresos [8].

3. **Afiliación a seguridad social (P6090):** La afiliación a un sistema de seguridad social es un factor crucial en la mitigación de los gastos de bolsillo. Los hogares sin afiliación enfrentan mayores costos debido a la necesidad de pagar servicios privados o recurrir a redes informales. Según la literatura disponible, estar afiliado a un régimen contributivo o subsidiado reduce significativamente los gastos de bolsillo, ya que las EPS cubren servicios básicos, consultas médicas y hospitalizaciones. Sin embargo, la cobertura puede ser parcial, lo que lleva a gastos adicionales en medicamentos o tratamientos específicos [9].

4. **Régimen de afiliación (P6100):** El régimen de afiliación define las condiciones de acceso a los servicios de salud y la proporción de costos cubiertos. Por ejemplo: (1) El régimen contributivo está asociado con menores gastos de bolsillo debido a la mayor calidad de los servicios cubiertos; (2) el régimen subsidiado, aunque reduce el acceso a servicios especializados, también disminuye los costos en comparación con la falta de cobertura y; (3) los afiliados al régimen especial (e.g., Fuerzas Armadas) tienden a tener gastos de bolsillo mínimos debido a coberturas más integrales [10].

5. **Estado general de salud (P6127):** La autoevaluación del estado de salud es un indicador válido y ampliamente utilizado en la literatura para predecir el uso de servicios médicos y los costos asociados. Las personas que reportan mala salud tienen una mayor probabilidad de necesitar consultas frecuentes, medicamentos y hospitalizaciones, lo que incrementa los gastos de bolsillo. Este indicador es robusto incluso después de controlar por factores sociodemográficos y enfermedades específicas [11].

6. **Diagnóstico de enfermedades crónicas (P1930):** Las enfermedades crónicas, como diabetes, hipertensión y enfermedades cardíacas, generan gastos recurrentes debido a la necesidad de medicamentos, consultas regulares y, en

algunos casos, procedimientos especializados. En estudios longitudinales, los hogares con al menos un miembro con enfermedad crónica reportan hasta un 50 por ciento más de gasto de bolsillo en salud en comparación con aquellos sin enfermedades crónicas [12].

7. **Calidad percibida del servicio de salud (P6181):** La percepción de la calidad del servicio de salud es un factor clave en la decisión de los hogares para acceder a servicios médicos, influenciando tanto la frecuencia de uso como la elección de proveedores. La literatura sugiere que una baja calidad percibida puede llevar a que las personas busquen atención fuera de su EPS o del sistema subsidiado, incrementando los gastos de bolsillo al utilizar servicios privados o no cubiertos [13].

8. **Región geográfica (REGION):** La región donde se ubica un hogar influye significativamente en el acceso y costo de los servicios de salud, debido a disparidades geográficas en la infraestructura médica, costos regionales y disponibilidad de programas subsidiados. Estudios han demostrado que las regiones con menos infraestructura sanitaria tienden a imponer mayores cargas de gasto de bolsillo a los hogares [14].

9. **Clase (CLASE):** La división entre áreas urbanas y rurales tiene implicaciones importantes en el acceso a servicios de salud y sus costos. En áreas rurales, el gasto de bolsillo puede aumentar debido a la necesidad de desplazarse largas distancias para recibir atención médica, así como a la falta de opciones cubiertas por el sistema público [15]. En contraste, las áreas urbanas suelen ofrecer más alternativas privadas que pueden ser costosas para los hogares.

10. **Tenencia de la vivienda (P5095):** La situación habitacional de un hogar puede ser un indicador indirecto del nivel socioeconómico y, por ende, de su capacidad para asumir gastos médicos. Los hogares que no son propietarios de su vivienda o que viven en condiciones de precariedad tienden a enfrentar mayores restricciones financieras, limitando su acceso a servicios de salud y aumentando la proporción del gasto de bolsillo [16].

11. **Percepción de pobreza (P5230):** La percepción de pobreza refleja la autosuficiencia económica percibida por el hogar y se asocia directamente con su comportamiento en el gasto de bolsillo. Los hogares que se consideran pobres son más propensos a evitar gastos médicos preventivos o electivos, mientras que concentran sus gastos en situaciones de emergencia, lo que puede generar variabilidad en los costos de bolsillo [17].

Cuadro 4.2: Resumen de las variables utilizadas en el modelo.

Variable	**Descripción**	**Categorías y Significado / Unidad**	**Tipo de Variable**
CANT_PERSONAS	Cantidad de personas en el hogar	Número entero	Cuantitativa
I_HOGAR_ANUAL	Ingreso total anual del hogar	Pesos colombianos	Cuantitativa
P6090	Afiliación a EPS	1: Afiliado 2: No afiliado 9: No sabe, no informa	Categórica nominal
P6100	Régimen de afiliación	1: Contributivo (EPS) 2: Especial 3: Subsidiado 9: No sabe, no informa	Categórica nominal
P6127	Estado de salud general	1: Muy bueno 2: Bueno 3: Regular 4: Malo	Categórica ordinal
P1930	Diagnóstico de enfermedad crónica	1: Sí 2: No	Categórica nominal
P6181	Calidad percibida del servicio EPS	1: Muy buena 2: Buena 3: Mala 4: Muy mala 9: No sabe	Categórica ordinal
REGION	Región geográfica	1: Caribe 2: Pacífica 3: Oriental 4: Central 5: San Andrés 6: Amazonia/Orinoquia 7: Bogotá D.C. 8: Otra	Categórica nominal
CLASE	Tipo de área	1: Cabecera 2: Centros poblados o rural	Categórica nominal
P5095	Tenencia de la vivienda	1: Propia totalmente pagada 2: Propia en pago 3: Arriendo o subarriendo 4: Con permiso del propietario 5: Posesión sin título 6: Propiedad colectiva	Categórica nominal
P5230	Autopercepción de pobreza	1: Sí 2: No	Categórica nominal

Este conjunto de variables proporciona una visión multidimensional de los factores que influyen en el gasto de bolsillo en salud, permitiendo capturar tanto aspectos socioeconómicos como demográficos que son relevantes para la modelación predictiva en este ámbito.

4.6. Predicción en Aprendizaje Automático: Regresión y Clasificación

El término *predicción* en aprendizaje automático se refiere a la capacidad de un modelo para inferir resultados futuros o desconocidos a partir de patrones aprendidos en datos previos. Tradicionalmente, la predicción ha sido asociada con modelos de regresión que estiman valores continuos. Sin embargo, la clasificación también es una forma válida de predicción cuando el objetivo es asignar instancias a categorías discretas [18, 19].

En términos generales, los modelos de regresión buscan predecir una variable de salida continua, ajustando una función matemática que minimiza el error con respecto a los valores observados. Estos modelos incluyen regresión lineal, regresión polinómica y técnicas más avanzadas como los árboles de regresión y métodos de *boosting* [20]. En contraste, los modelos de clasificación asignan una etiqueta discreta a cada observación, prediciendo la probabilidad de que una instancia pertenezca a una determinada clase. Ejemplos de modelos de clasificación incluyen la regresión logística, los árboles de decisión y los métodos de *ensemble learning* como *Random Forest* y *Gradient Boosting* [21].

La predicción en clasificación implica estimar la probabilidad de pertenencia a una clase y asignar la categoría con mayor probabilidad. Este proceso es particularmente útil cuando el fenómeno de estudio presenta una distribución asimétrica de valores, lo que hace que los modelos de regresión sean poco efectivos [22]. En el caso del gasto de bolsillo en salud, donde la variable dependiente presenta una gran cantidad de valores nulos y un sesgo extremo, la conversión de esta variable (lo cual se presenta en capítulso posteriores) en una representación dicotómica mejora la capacidad del modelo para capturar patrones subyacentes en los datos [23, 3].

Por lo tanto, tanto la regresión como la clasificación son enfoques de predicción adecuados según el tipo de problema y la naturaleza de los datos. En el presente estudio, la clasificación permite predecir la ocurrencia del gasto de bolsillo en salud, proporcionando una interpretación más robusta para la toma de decisiones y la formulación de políticas públicas basadas en evidencia.

4.7. Predicción del Gasto de Bolsillo usando Técnicas de Aprendizaje Automático para modelos de regresión

El Aprendizaje Automático (AA) es una rama de la inteligencia artificial que se centra en el desarrollo de algoritmos capaces de aprender a partir de datos y realizar predicciones basadas en ellos. En el contexto de la predicción del GBS, el objetivo es construir modelos que puedan estimar con precisión un valor continuo, como la cantidad de dinero que los hogares colombianos destinan directamente a servicios sanitarios.

Dado que el gasto de bolsillo es una variable continua, las técnicas de aprendizaje automático más adecuadas para este problema son los modelos de regresión. Estos modelos permiten capturar relaciones complejas entre múltiples variables independientes (como ingresos, tamaño del hogar, estado de salud, entre otros) y la variable

dependiente, es decir el GBS.

La aplicación de técnicas de AA en la predicción del GBS es particularmente relevante debido a la complejidad y la multidimensionalidad de los factores que influyen en este gasto [24]. Estos modelos pueden manejar una variedad de variables, desde datos demográficos y socioeconómicos hasta patrones de uso de servicios de salud, proporcionando predicciones precisas y adecuadas a las necesidades de cada estudio.

A continuación se describen analíticamente los modelos de Aprendizaje Automático que son comunmente usadado para la predicción del GBS [25, 26, 27].

TreeNet (Gradient Boosting Machines para Regresión)

TreeNet es una implementación de Gradient Boosting Machines (GBM), un enfoque de aprendizaje supervisado que construye modelos predictivos a través de una combinación secuencial de predictores débiles, generalmente árboles de decisión. En lugar de construir un solo modelo complejo, GBM crea una serie de modelos simples que se combinan para mejorar la precisión de las predicciones[25, 28].

El desarrollo matemático de TreeNet para regresión se describe a continuación:

Inicialización del Modelo: Dado un conjunto de datos $\{(x_i, y_i)\}_{i=1}^{N}$, donde x_i representa las variables independientes y y_i es la variable dependiente que se desea predecir, el modelo inicial $F_0(x)$ se define como una constante que minimiza la función de pérdida, generalmente el error cuadrático medio (MSE) en el caso de regresión:

$$F_0(x) = \arg\min_{c} \sum_{i=1}^{N} L(y_i, c) \tag{4.7.1}$$

Para el MSE, esto se simplifica a:

$$F_0(x) = \frac{1}{N} \sum_{i=1}^{N} y_i \tag{4.7.2}$$

Construcción Iterativa: En cada iteración m, se ajusta un nuevo árbol de decisión $h_m(x)$ a los residuales del modelo anterior. Los residuales r_{im} se calculan como la derivada negativa de la función de pérdida respecto a las predicciones actuales del modelo:

$$r_{im} = -\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\bigg|_{F(x)=F_{m-1}(x)} \tag{4.7.3}$$

Para el MSE, los residuales son simplemente:

$$r_{im} = y_i - F_{m-1}(x_i) \tag{4.7.4}$$

Se entrena un nuevo árbol de decisión $h_m(x)$ para predecir estos residuales.

Actualización del Modelo: El modelo se actualiza sumando la contribución del nuevo árbol escalada por un factor de aprendizaje ν:

$$F_m(x) = F_{m-1}(x) + \nu h_m(x) \tag{4.7.5}$$

Donde ν es la tasa de aprendizaje, que controla la contribución de cada árbol al modelo final. Un valor más bajo de ν puede hacer que el entrenamiento sea más lento pero ayuda a evitar el sobreajuste.

Modelo Final: Después de M iteraciones, el modelo final es la suma de todos los árboles ajustados:

$$F_M(x) = F_0(x) + \sum_{m=1}^{M} \nu h_m(x) \tag{4.7.6}$$

Fortalezas: TreeNet puede manejar interacciones complejas entre variables y capturar relaciones no lineales debido a la naturaleza de los árboles de decisión. Cada nuevo árbol intenta corregir los errores de los árboles anteriores, lo que generalmente mejora la precisión del modelo en cada iteración. El modelo es flexible y puede ajustarse a una amplia variedad de problemas de regresión ajustando los hiperparámetros como la tasa de aprendizaje, el número de iteraciones, y la profundidad de los árboles. Aunque el modelo es susceptible al sobreajuste, este riesgo puede mitigarse mediante la configuración adecuada de los hiperparámetros[25].

Debilidades: El proceso de entrenamiento es intensivo en términos computacionales, especialmente con grandes conjuntos de datos o un alto número de iteraciones. El rendimiento del modelo es altamente dependiente de la correcta configuración de los hiperparámetros, lo que puede requerir una experimentación exhaustiva para encontrar la combinación óptima. Aunque los árboles individuales son interpretables, el modelo final, que es una combinación de muchos árboles, puede ser complejo y difícil de interpretar. Si los datos contienen outliers o están desequilibrados, el modelo puede verse afectado, requiriendo técnicas adicionales para manejar estos problemas[25].

Random Forest para Regresión

Random Forest es un método de ensemble learning que combina múltiples árboles de decisión para mejorar la precisión predictiva y controlar el sobreajuste[26]. En problemas de regresión, el objetivo es predecir un valor continuo, como en el caso del gasto de bolsillo en salud.

El desarrollo matemático de Random Forest para regresión se describe a continuación:

Construcción del Bosque: Dado un conjunto de datos $\{(x_i, y_i)\}_{i=1}^{N}$, donde x_i son las variables independientes y y_i es la variable dependiente que se desea predecir, el primer paso en Random Forest es construir B árboles de decisión independientes. Para construir cada árbol, se utilizan las siguientes estrategias:

1. Muestreo de Bootstrap: - Para cada árbol T_b ($b = 1, \ldots, B$), se selecciona una muestra de bootstrap D_b del conjunto de datos original D, lo que significa que D_b se construye tomando N muestras aleatorias de D con reemplazo.

2. Selección Aleatoria de Características: - En cada nodo del árbol, se selecciona aleatoriamente un subconjunto de m características (de las p características totales) y se elige la mejor división entre esas m características para dividir el nodo. Este proceso se repite en cada nodo hasta que se cumplen los criterios de parada (como la profundidad máxima del árbol o el número mínimo de muestras en un nodo).

Predicción con Random Forest: Para predecir un nuevo valor y dado un conjunto de características x, se realiza la predicción con cada uno de los B árboles construidos y luego se promedian las predicciones:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} T_b(x) \tag{4.7.7}$$

Donde $T_b(x)$ es la predicción del árbol T_b para la entrada x. Este promedio de predicciones reduce la varianza del modelo, lo que generalmente mejora la precisión y robustez del modelo.

Fortalezas: Random Forest es capaz de manejar datos con estructuras complejas e interacciones no lineales entre variables. Debido a su enfoque de ensemble, es robusto frente al sobreajuste en comparación con un único árbol de decisión. El modelo es fácil de ajustar y tiene pocos hiperparámetros que requieren afinación, como el número de árboles (B) y el número de características (m) a considerar en cada nodo. Además, Random Forest proporciona una medida de la importancia de las características, lo que es útil para la interpretación del modelo[26].

Debilidades: A pesar de su robustez, Random Forest puede ser computacionalmente intensivo y lento de entrenar, especialmente cuando se trabaja con grandes conjuntos de datos y muchos árboles. El modelo puede consumir mucha memoria, ya que necesita almacenar todos los árboles de decisión. Aunque Random Forest es más robusto frente al sobreajuste que un solo árbol, sigue siendo susceptible a él si no se configura adecuadamente el número de árboles y el número de características consideradas en cada división. Además, el modelo final, siendo un conjunto de muchos árboles, puede ser complejo y difícil de interpretar en comparación con un solo árbol de decisión[26].

MARS (Multivariate Adaptive Regression Splines)

MARS (Splines Adaptativos de Regresión Multivariante) es una técnica de regresión no paramétrica que extiende los modelos de regresión lineal y polinómica al permitir modelar relaciones no lineales y de alta dimensionalidad.[27] MARS logra esto utilizando funciones base que son splines lineales por partes, ajustadas a diferentes intervalos de los datos.

El desarrollo matemático de MARS para regresión se describe a continuación:

Modelo MARS: Dado un conjunto de datos $\{(x_i, y_i)\}_{i=1}^{N}$, donde x_i son las variables independientes y y_i es la variable dependiente, el modelo MARS se construye como una suma de funciones base $B_m(x)$, que son splines lineales por partes:

$$f(x) = \beta_0 + \sum_{m=1}^{M} \beta_m B_m(x) \tag{4.7.8}$$

Aquí, β_0 es el intercepto, β_m son los coeficientes de las funciones base, $B_m(x)$ son las funciones base, y M es el número total de funciones base en el modelo. Cada función base $B_m(x)$ es un spline lineal por partes, que toma la forma:

$$(x - t)_+ = \begin{cases} x - t & \text{si } x > t \\ 0 & \text{si } x \leq t \end{cases} \tag{4.7.9}$$

o

$$(t - x)_+ = \begin{cases} t - x & \text{si } x < t \\ 0 & \text{si } x \geq t \end{cases} \tag{4.7.10}$$

donde t es un punto de ruptura (knot) en los datos.

Construcción del Modelo: El modelo MARS se construye en dos etapas:

1. **Etapa de Avance**: - En esta etapa, MARS agrega funciones base de manera secuencial para minimizar el error de ajuste. En cada paso, se evalúan todas las posibles combinaciones de funciones base y puntos de ruptura, y se selecciona la combinación que más reduce el error cuadrático medio (MSE). Esto da lugar a un modelo con muchas funciones base.

2. **Etapa de Poda**: - Para evitar el sobreajuste, MARS utiliza una etapa de poda donde se eliminan las funciones base menos importantes utilizando un criterio de información, como el Generalized Cross-Validation (GCV):

$$GCV = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \left(1 - \frac{df}{N}\right)^{-2} \tag{4.7.11}$$

donde $\hat{y}_i$ son las predicciones del modelo, N es el número de observaciones, y df es el número de grados de libertad del modelo (que depende del número de funciones base utilizadas).

Predicción: Una vez completado el proceso de construcción y poda, el modelo MARS final se utiliza para hacer predicciones de la variable dependiente y dado un conjunto de características x.

Fortalezas: MARS es capaz de capturar relaciones no lineales complejas entre las variables independientes y la variable dependiente. Además, selecciona automáticamente las interacciones entre variables que son más importantes para el modelo. Los modelos MARS son relativamente fáciles de interpretar, ya que las funciones base corresponden a splines lineales por partes con puntos de ruptura específicos. MARS maneja bien datos de alta dimensionalidad y es adecuado para problemas de regresión en escenarios con muchas variables[27].

Debilidades: El rendimiento de MARS puede ser sensible a la selección de los puntos de ruptura, lo que puede requerir ajustes finos y experimentación. La construcción del modelo puede ser computacionalmente intensiva, especialmente para conjuntos de datos grandes con muchas variables. Aunque la etapa de poda ayuda a evitar el sobreajuste, una poda inadecuada puede llevar a un modelo subóptimo. Además, aunque MARS maneja bien datos de alta dimensionalidad, puede ser menos preciso que otros métodos de machine learning, como los métodos basados en árboles, en escenarios con muchas observaciones y complejidad en las relaciones entre variables[27].

Regresión con Árboles de Decisión

La regresión con árboles de decisión es un método de aprendizaje supervisado que utiliza una estructura de árbol para modelar relaciones entre variables independientes y una variable dependiente continua. A diferencia de la regresión lineal, que asume una relación lineal entre las variables, los árboles de decisión pueden capturar relaciones no lineales y complejas sin necesidad de especificar una forma funcional previa [29].

Desarrollo Matemático

Construcción del Árbol: Dado un conjunto de datos $\{(x_i, y_i)\}_{i=1}^{N}$, donde $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ son las variables independientes y y_i es la variable dependiente continua, el objetivo es construir un árbol de decisión que divida el espacio de características en regiones homogéneas respecto a y.

1. Criterio de División: En cada nodo del árbol, se selecciona una variable x_j y un punto de corte c que dividen los datos en dos subconjuntos D_L y D_R de manera que se minimiza la suma de los errores cuadráticos dentro de cada subconjunto. El criterio de minimización se puede expresar como:

$$\text{Error Total} = \sum_{i \in D_L} (y_i - \bar{y}_L)^2 + \sum_{i \in D_R} (y_i - \bar{y}_R)^2 \tag{4.7.12}$$

donde $\bar{y}_L$ y $\bar{y}_R$ son las medias de y en los subconjuntos D_L y D_R, respectivamente.

2. Selección de la Mejor División: Se evalúan todas las posibles divisiones (x_j, c) y se selecciona aquella que minimiza el error total. Este proceso se repite recursivamente para cada nodo hijo hasta que se cumplan los criterios de parada, como la profundidad máxima del árbol o el número mínimo de muestras en un nodo.

Predicción: Una vez construido el árbol, la predicción para una nueva instancia x se realiza siguiendo el camino desde la raíz hasta una hoja, de acuerdo con los valores de las variables independientes. La predicción $\hat{y}$ es simplemente la media de los valores de y en la hoja correspondiente:

$$\hat{y} = \frac{1}{|S|} \sum_{i \in S} y_i \tag{4.7.13}$$

donde S es el conjunto de observaciones en la hoja final.

Poda del Árbol: Para evitar el sobreajuste, se aplica una etapa de poda que elimina ramas que tienen poca relevancia. Esto puede hacerse utilizando técnicas como la validación cruzada o estableciendo un tamaño mínimo de hojas.

Fortalezas: La regresión con árboles de decisión es fácil de interpretar y visualizar, lo que facilita la comprensión del modelo por parte de los usuarios. Puede capturar relaciones no lineales y manejar tanto variables numéricas como categóricas. Además, no requiere una normalización previa de los datos y es robusta frente a valores atípicos[29].

Debilidades: Los árboles de decisión individuales tienden a ser propensos al sobreajuste, especialmente si no se controlan adecuadamente los parámetros del modelo. También pueden ser inestables, ya que pequeñas variaciones en los datos pueden resultar en árboles muy diferentes. Además, los árboles de decisión no suelen generalizar tan bien como otros métodos más avanzados como los ensembles de árboles (por ejemplo, Random Forest o Gradient Boosting)[29].

4.7.1. Análisis y Evaluación de Modelos

La evaluación de modelos de regresión es un paso crucial para garantizar que las predicciones sean precisas y útiles. Existen varias métricas y técnicas utilizadas para evaluar el rendimiento de los modelos en problemas de regresión [30]. Las métricas más comunes se describen brevemente a continuación.

- **Error Cuadrático Medio (MSE)**: Mide la media de los errores al cuadrado entre los valores predichos y los valores reales. Penaliza más los errores grandes.

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{4.7.14}$$

- **Raíz del Error Cuadrático Medio (RMSE)**: Es la raíz cuadrada del MSE, lo que lo hace más interpretable ya que está en las mismas unidades que la variable dependiente.

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{4.7.15}$$

- **Coeficiente de Determinación (R^2)**: Indica la proporción de la varianza en la variable dependiente que es explicada por el modelo. Un valor cercano a 1 indica un buen ajuste.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{4.7.16}$$

- **Mean Absolute Error (MAE)**: Mide la media de los errores absolutos entre las predicciones y los valores reales, ofreciendo una visión más intuitiva de la precisión del modelo.

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{4.7.17}$$

Validación y Robustez del Modelo: Para garantizar que los modelos generalicen bien a nuevos datos, se utiliza la *validación cruzada*. Una técnica común es la validación cruzada k-fold, donde el conjunto de datos se divide en k subconjuntos, y el modelo se entrena k veces, utilizando un subconjunto diferente como conjunto de validación y los restantes $k-1$ subconjuntos como conjunto de entrenamiento.

Además de la validación cruzada, es fundamental realizar un *análisis de sensibilidad* y un *análisis de importancia de características*:

- **Análisis de Sensibilidad**: Evalúa cómo los cambios en las variables independientes afectan las predicciones del modelo. Identificar las variables más influyentes ayuda a entender mejor el comportamiento del modelo.
- **Importancia de Características**: En modelos como Random Forest, se puede calcular la importancia de cada característica en la predicción de la variable dependiente, ayudando a identificar cuáles tienen un mayor impacto.

Finalmente, es importante considerar la *interpretabilidad del modelo*, el manejo de *datos atípicos y faltantes*, y la *eficiencia computacional*, asegurando que el modelo sea aplicable en contextos del mundo real y que las predicciones sean comprensibles para los responsables de la toma de decisiones.

4.8. Predicción del Gasto de Bolsillo usando Técnicas de Aprendizaje Automático para Modelos de Clasificación

En la sección anterior se describió los modelos de regresión de AA más empleados en trabajos similares. En esta sección se complementa con modelos usados cuando la variable dependiente no se modela como una variable continua sino como una binaria, en este caso, los modelos de clasificación son los más convenietes y se describiran a continuación.

4.8.1. Gradient Boosting Classifier

Gradient Boosting Classifier (GBC) es un método de ensemble learning que combina de manera secuencial varios clasificadores débiles (generalmente árboles de decisión) para optimizar una función de pérdida en un problema de clasificación [25].

Desarrollo matemático

El modelo se construye iterativamente:

1. **Inicialización:** El modelo inicial $F_0(x)$ minimiza la función de pérdida logística:

$$F_0(x) = \arg\min_c \sum_{i=1}^{N} L(y_i, c), \tag{4.8.1}$$

donde $L(y_i, F(x)) = -\left[y_i \log(\hat{p}_i) + (1 - y_i) \log(1 - \hat{p}_i)\right]$.

2. **Residuos:** En cada iteración m, los residuales se calculan como:

$$r_{im} = -\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\bigg|_{F(x)=F_{m-1}(x)}. \tag{4.8.2}$$

3. **Ajuste del Modelo:** Se entrena un árbol $h_m(x)$ para predecir los residuos r_{im}.

4. **Actualización:** El modelo se actualiza como:

$$F_m(x) = F_{m-1}(x) + \nu h_m(x), \tag{4.8.3}$$

donde ν es la tasa de aprendizaje.

El modelo final está dado por:

$$F_M(x) = F_0(x) + \sum_{m=1}^{M} \nu h_m(x). \tag{4.8.4}$$

Fortalezas y debilidades: GBC es poderoso para capturar relaciones complejas, pero requiere ajustar cuidadosamente los hiperparámetros para evitar el sobreajuste.

4.8.2. Árbol de Decisión

Los árboles de decisión clasifican datos dividiendo iterativamente el espacio de características basado en métricas como la ganancia de información o el índice Gini [29].

Desarrollo matemático

1. **Criterio de División:** En cada nodo, se selecciona la variable x_j y el punto de corte c que maximizan la ganancia de información o minimizan el índice Gini:

$$\text{Gini}(D) = 1 - \sum_{k=1}^{K} p_k^2, \tag{4.8.5}$$

donde p_k es la proporción de observaciones de la clase k.

2. **Predicción:** Cada hoja del árbol predice la clase mayoritaria en el subconjunto correspondiente.

Fortalezas y debilidades: Son interpretables y rápidos de entrenar, pero susceptibles al sobreajuste.

4.8.3. Random Forest

Random Forest es un modelo de ensemble que construye múltiples árboles de decisión entrenados en muestras aleatorias del conjunto de datos y promedia sus predicciones [26].

Desarrollo matemático

1. **Construcción del Bosque:** - Cada árbol se entrena con una muestra de bootstrap. - En cada nodo, se selecciona un subconjunto aleatorio de características para buscar la mejor división.

2. **Predicción:** La predicción final se realiza por votación mayoritaria entre todos los árboles.

$$\hat{y} = \text{mode}\{T_b(x) : b = 1, \ldots, B\}. \tag{4.8.6}$$

Fortalezas y debilidades: Es robusto y reduce la varianza, pero puede ser costoso en cómputo.

4.8.4. Regresión Logística

La regresión logística es un modelo lineal que utiliza la función sigmoide para predecir la probabilidad de que una observación pertenezca a una clase [30].

Desarrollo matemático

La probabilidad de la clase $y = 1$ está dada por:

$$P(y = 1|x) = \frac{1}{1 + e^{-z}}, \quad z = \beta_0 + \sum_{j=1}^{p} \beta_j x_j. \tag{4.8.7}$$

El modelo se ajusta maximizando la función de verosimilitud logarítmica:

$$\ell(\beta) = \sum_{i=1}^{N} \left[y_i \log(\hat{p}_i) + (1 - y_i) \log(1 - \hat{p}_i) \right]. \tag{4.8.8}$$

Fortalezas y debilidades: Es simple e interpretable, pero puede ser insuficiente para capturar relaciones no lineales complejas.

4.8.5. Evaluación de Modelos

La evaluación de los modelos de clasificación utiliza métricas como:

- **Precisión:** Proporción de predicciones correctas:

$$\text{Precisión} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{4.8.9}$$

- **Recall:** Proporción de casos positivos correctamente identificados:

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{4.8.10}$$

- **F1-Score:** Media armónica de precisión y recall:

$$\text{F1} = 2 \cdot \frac{\text{Precisión} \cdot \text{Recall}}{\text{Precisión} + \text{Recall}}. \tag{4.8.11}$$

- **Área bajo la curva (AUC-ROC):** Evalúa la capacidad del modelo para distinguir entre clases.

Validación: Se utiliza validación cruzada k-fold para garantizar la generalización y evitar el sobreajuste.

4.9. Antecedentes

En la exploración del uso de técnicas de aprendizaje automático para la predicción del gasto de bolsillo en salud, diversos estudios han marcado hitos importantes, reflejando tanto los avances como los desafíos en este campo.

El trabajo de Muremyi et al. [31] en Ruanda, utilizando datos de la Encuesta Integrada de Condiciones de Vida, es un ejemplo destacado. Este estudio aplicó modelos como Multivariate Adaptive Regression Splines (MARS), Random Forest y Gradient Boosting, encontrando que Treenet era el más preciso. Este resultado destaca la importancia de variables como el consumo total del hogar en la predicción de gastos de bolsillo, ofreciendo una perspectiva valiosa para contextos similares.

Por otro lado, una revisión sistemática presentada en *Diagnostic and Prognostic Research* [32] subraya la necesidad crítica de evaluar la calidad metodológica y el riesgo de sesgo en estudios de aprendizaje automático centrados en gastos de salud. Esta revisión propone un marco riguroso para evaluar tales estudios, resaltando la complejidad y las potenciales trampas en el diseño y la interpretación de estos modelos. En este estudio se destaca la necesidad de evaluar la calidad metodológica y el riesgo de sesgo en estudios de aprendizaje automático centrados en gastos de salud. Los posibles sesgos incluyen sesgo de selección, información, confusión, desempeño, publicación y sobreajuste, cada uno de los cuales puede afectar la validez y generalización de los modelos. La herramienta PROBAST se propone para evaluar estos riesgos, centrándose en los participantes del estudio, los predictores, los desenlaces y el análisis estadístico, asegurando que los modelos sean robustos y aplicables en la práctica y la formulación de políticas de salud.

El estudio *Machine-learning-based prediction models for high-need high-cost patients using nationwide clinical and claims data* [33], publicado en *npj Digital Medicine* en 2020, tiene como objetivo desarrollar y evaluar modelos de predicción basados en aprendizaje automático para identificar pacientes de alto costo y alta necesidad (HCHN) utilizando datos clínicos y de reclamaciones a nivel nacional. Utilizando algoritmos como Random Forest, Gradient Boosting y Redes Neuronales, los modelos se entrenaron y validaron con datos amplios, mostrando una alta precisión (AUC >0.85) en la identificación de estos pacientes. Las variables predictoras más importantes incluyeron el historial de hospitalizaciones, diagnósticos crónicos y el

uso de medicamentos específicos. Los resultados indicaron que los modelos de aprendizaje automático superaron a los métodos tradicionales en términos de precisión y capacidad predictiva, aunque presentan desafíos en la interpretación y necesidad de actualizaciones frecuentes. La identificación precisa de pacientes HCHN puede mejorar significativamente la gestión de recursos y la planificación de intervenciones de salud, contribuyendo a la eficiencia y sostenibilidad del sistema de salud. Estos hallazgos tienen importantes implicaciones prácticas para la política de salud y la gestión de cuidados, permitiendo una mejor asignación de recursos y reducción de costos en el sistema de salud. Por su parte, estudios que han usado métodos de aprendizaje supervisado prediciendo costos en salud de manera comparativa [34], destacan el Gradient Boosting por su excelente rendimiento en la predicción de costos de salud bajos a medianos. Este hallazgo subraya la eficacia de las técnicas de AA en diferentes segmentos de costos, ampliando su aplicabilidad.

Finalmente, el estudio *The application of machine learning to predict high-cost patients: A performance-comparison of different models using healthcare claims data* [35], se centra en la comparación del rendimiento de diferentes modelos de aprendizaje automático para predecir pacientes de alto costo utilizando datos de reclamaciones de salud. Utilizando una variedad de algoritmos, incluidos Random Forest, Gradient Boosting, Support Vector Machines y Redes Neuronales, el estudio evaluó estos modelos en términos de precisión, sensibilidad, especificidad y Área Bajo la Curva (AUC). Los resultados mostraron que los modelos de aprendizaje automático superaron significativamente a los métodos tradicionales de regresión logística, destacando la capacidad de los modelos más complejos, como Random Forest y Gradient Boosting, para manejar mejor la heterogeneidad y la complejidad de los datos de salud. El análisis comparativo reveló que, aunque todos los modelos de aprendizaje automático ofrecieron mejoras respecto a los enfoques tradicionales, no existe un modelo universalmente superior; la elección del modelo puede depender del contexto específico y de los requisitos del sistema de salud en términos de interpretabilidad y recursos computacionales. Los modelos más avanzados, aunque precisos, presentan desafíos en términos de interpretabilidad y pueden requerir mayores recursos para su implementación. Este estudio subraya la importancia de seleccionar el modelo adecuado basado en un balance entre precisión y aplicabilidad práctica, ofreciendo una guía valiosa para los profesionales de la salud y los responsables de políticas en la optimización de la gestión de recursos y la planificación de cuidados en el sistema de salud.

En conjunto, estos estudios demuestran el creciente interés y la versatilidad de los modelos de AA en la predicción del gasto de bolsillo en salud, así como de otras variables de salud. Sin embargo, también indican que queda trabajo por hacer en términos de la selección de variables, la adaptación de modelos a contextos específicos y la evaluación crítica de su precisión y confiabilidad. También se resalta el hecho de que en mucho casos se opta por modelar la variable dependiente como una variable binaria. Estos antecedentes proporcionan un punto de partida importante para este trabajo y son esenciales para el desarrollo de las secciones posteriores.

4.10. Novedad del Análisis de Predicción del Gasto de Bolsillo en Salud

El análisis del gasto de bolsillo en salud ha sido un tema de creciente interés en la literatura económica y en el ámbito de la salud pública debido a su impacto en la equidad del acceso a servicios médicos y en la carga financiera de los hogares. Sin embargo, la mayoría de los estudios existentes se han centrado en el análisis descriptivo de los determinantes del gasto o en la evaluación de políticas de protección financiera [36, 37, 23]. En contraste, el uso de técnicas de *aprendizaje automático* para la predicción del gasto de bolsillo sigue siendo un campo emergente con escasa exploración académica.

Gran parte de la literatura previa ha empleado métodos econométricos tradicionales, como modelos de regresión lineal y modelos de selección de Heckman, para analizar factores asociados con el gasto en salud [38, 39]. No obstante, estos enfoques presentan limitaciones cuando la distribución del gasto es altamente asimétrica o cuando la variable objetivo contiene una gran cantidad de valores cero, como ocurre en el caso del gasto de bolsillo [3]. En este contexto, la aplicación de técnicas de clasificación y métodos avanzados de predicción ofrece una alternativa prometedora para mejorar la capacidad de identificación de patrones y la precisión en la estimación de la probabilidad de gasto en salud.

A pesar del avance de los modelos de *machine learning* en diversos dominios, su aplicación en la predicción del gasto de bolsillo en salud es todavía incipiente. Existen pocos estudios que hayan explorado modelos no lineales, técnicas de sobremuestreo o estrategias de optimización para mejorar la precisión en la predicción de este fenómeno [36, 3]. Esto resalta la contribución de la presente investigación, la cual busca llenar este vacío al implementar modelos de clasificación para anticipar la ocurrencia del gasto de bolsillo y evaluar su viabilidad en comparación con métodos tradicionales.

En conclusión, el presente estudio se ubica en un campo poco explorado dentro de la intersección entre la economía de la salud y el aprendizaje automático, aportando un enfoque novedoso que puede ser de gran utilidad para la formulación de políticas públicas y estrategias de protección financiera. La escasez de referencias específicas sobre la predicción del gasto de bolsillo utilizando *machine learning* refuerza la relevancia y la originalidad de este trabajo.

Capítulo 5

Metodología

5.1. Descripción del Conjunto de Datos

El presente estudio se basa en la Encuesta de Calidad de Vida (ECV), la cual proporciona información detallada sobre las condiciones socioeconómicas y de salud de los hogares colombianos. La variable dependiente de interés es el gasto de bolsillo en salud, mientras que las variables independientes incluyen factores demográficos, socioeconómicos y de acceso a servicios de salud.

5.1.1. Análisis Exploratorio de Datos

Antes de la aplicación de modelos, se realizó un análisis exploratorio de datos (EDA) para comprender la distribución de las variables, detectar valores atípicos y evaluar la correlación entre variables. Las principales técnicas empleadas fueron:

- Visualización de distribuciones mediante histogramas y diagramas de caja.
- Análisis de correlación entre variables numéricas utilizando la matriz de correlación de Pearson.
- Identificación de valores atípicos mediante diagramas de dispersión y análisis de rango intercuartílico.
- Evaluación de la cantidad de valores cero en la variable dependiente.

Estos análisis permitieron definir estrategias adecuadas de transformación y preprocesamiento de los datos.

5.1.2. Variables y Preprocesamiento de Datos

Para garantizar la calidad del análisis, se llevó a cabo un proceso de limpieza y transformación de datos, que incluyó:

- Eliminación de registros con valores faltantes en variables clave.
- Estandarización y normalización de variables numéricas.
- Codificación de variables categóricas mediante *one-hot encoding* para variables nominales y *ordinal encoding* para variables ordinales.

- Manejo de valores atípicos mediante el rango intercuartílico.
- Aplicación de transformaciones logarítmicas en variables con distribuciones sesgadas.

5.2. Metodología de Modelado

El análisis se llevó a cabo en dos enfoques: regresión y clasificación. Se evaluó la predicción del gasto como un valor continuo y como una variable dicotómica.

5.2.1. Modelos de Regresión

Se implementaron y compararon varios modelos de regresión para estimar el gasto de bolsillo:

- Regresión Lineal Múltiple
- Árboles de Decisión para Regresión
- Random Forest Regressor
- Gradient Boosting Regressor
- Modelos de Splines Adaptativos Multivariados (MARS)

5.2.2. Modelos de Clasificación

Debido a la gran cantidad de valores cero en la variable dependiente, se exploró un enfoque de clasificación para predecir la ocurrencia del gasto de bolsillo. Los modelos utilizados fueron:

- Regresión Logística
- Árbol de Decisión
- Random Forest
- Gradient Boosting Machines (GBM)
- Redes Neuronales Artificiales

Para abordar el desbalance de clases, se utilizó la técnica de sobremuestreo SMOTE (Synthetic Minority Oversampling Technique).

5.3. Validación y Evaluación

5.3.1. Estrategia de Validación Cruzada

Se empleó validación cruzada *k-fold* con $k = 5$ para evaluar la estabilidad y el desempeño de los modelos.

5.3.2. Métricas de Evaluación

Las métricas utilizadas para evaluar los modelos fueron:

- **Para regresión:** Coeficiente de determinación R^2, Error Cuadrático Medio (MSE) y Error Absoluto Medio (MAE).
- **Para clasificación:** Área bajo la curva ROC (AUC-ROC), F1-Score, Precisión y Recall.

5.4. Implementación Computacional

Los experimentos fueron desarrollados en el entorno de Google Colab, utilizando las siguientes herramientas:

- **Lenguaje de Programación:** Python 3.x
- **Librerías:** scikit-learn, pandas, numpy, matplotlib, seaborn, imbalanced-learn
- **Hardware:** CPU Intel(R) Xeon(R) @ 2.20GHz, 12GB de RAM

Capítulo 6

Análisis Exploratorio de Datos

6.1. Contexto

En este capítulo se presenta un análisis exploratorio de los datos utilizados en el estudio. La base de datos cuenta con un total de 11 variables independientes seleccionadas para capturar aspectos relevantes que influyen en el GBS de los hogares. Estas variables abarcan dimensiones como las características sociodemográficas del hogar, la afiliación y calidad del servicio de seguridad social, el estado de salud de los miembros del hogar y condiciones de vida relacionadas con la vivienda.

El Cuadro 6.1 y Cuadro 6.2 resume las variables seleccionadas, su descripción y dominios. Es importante resaltar que únicamente la variable *Ingreso Anual Total del Hogar* (`I_HOGAR_ANUAL`) presentó valores faltantes, los cuales fueron imputados con ceros, siguiendo las recomendaciones de la literatura disponible [3]. Este enfoque asegura que los resultados sean consistentes y reflejen adecuadamente la situación económica de los hogares analizados.

La variable objetivo, es decir, el GBS de los hogares, es una variable cuantitativa y cual fue calculada conforme a lo descrito en la sección 5.3. De las 11 variables incluidas en el análisis, 8 son categóricas (Cuadro 6.2), como el régimen de seguridad social (`P6100`) y el estado general de salud (`P6127`), mientras que 3 son numéricas (Cuadro 6.1), como el ingreso anual total del hogar (`I_HOGAR_ANUAL`) y la cantidad de personas por hogar (`CANT_PERSONAS_HOGAR`). Esta combinación permite capturar tanto información cuantitativa como cualitativa, proporcionando un análisis integral de los factores asociados al gasto de bolsillo en salud.

Cuadro 6.1: Descripción de variables cuantitativas.

Variable	**Descripción**	**Unidad**
GASTO_TOTAL_ANUAL	Gasto de bolsillo en salud de los hogares	Pesos colombianos
I_HOGAR_ANUAL	Ingreso total anual del hogar	Pesos colombianos
CANT_PERSONAS_HOGAR	Cantidad de personas en el hogar	Número entero

Cuadro 6.2: Descripción de variables categóricas.

Variable	Descripción	Categorías y Significado	Tipo de Variable
P6090	Afiliación a EPS	1: Afiliado 2: No afiliado 9: No sabe, no informa	Categórica nominal
P6100	Régimen de afiliación	1: Contributivo (EPS) 2: Especial 3: Subsidiado 9: No sabe, no informa	Categórica nominal
P6127	Estado de salud general	1: Muy bueno 2: Bueno 3: Regular 4: Malo	Categórica ordinal
P1930	Diagnóstico de enfermedad crónica	1: Sí 2: No	Categórica nominal
P6181	Calidad percibida del servicio EPS	1: Muy buena 2: Buena 3: Mala 4: Muy mala 9: No sabe	Categórica ordinal
REGION	Región geográfica	1: Caribe 2: Pacífica 3: Oriental 4: Central 5: San Andrés 6: Amazonia/Orinoquia 7: Bogotá D.C. 8: Otra	Categórica nominal
CLASE	Tipo de área	1: Cabecera 2: Centros poblados o rural	Categórica nominal
P5095	Tenencia de la vivienda	1: Propia totalmente pagada 2: Propia en pago 3: Arriendo o subarriendo 4: Con permiso del propietario 5: Posesión sin título 6: Propiedad colectiva	Categórica nominal
P5230	Autopercepción de pobreza	1: Sí 2: No	Categórica nominal

El análisis exploratorio de datos constituye un paso fundamental para comprender la estructura y calidad de los datos, identificando patrones relevantes, valores atípicos y posibles inconsistencias. Este proceso permite preparar los datos para su uso en el modelo predictivo del gasto de bolsillo en salud.

6.2. Resumen descriptivo y frecuencias

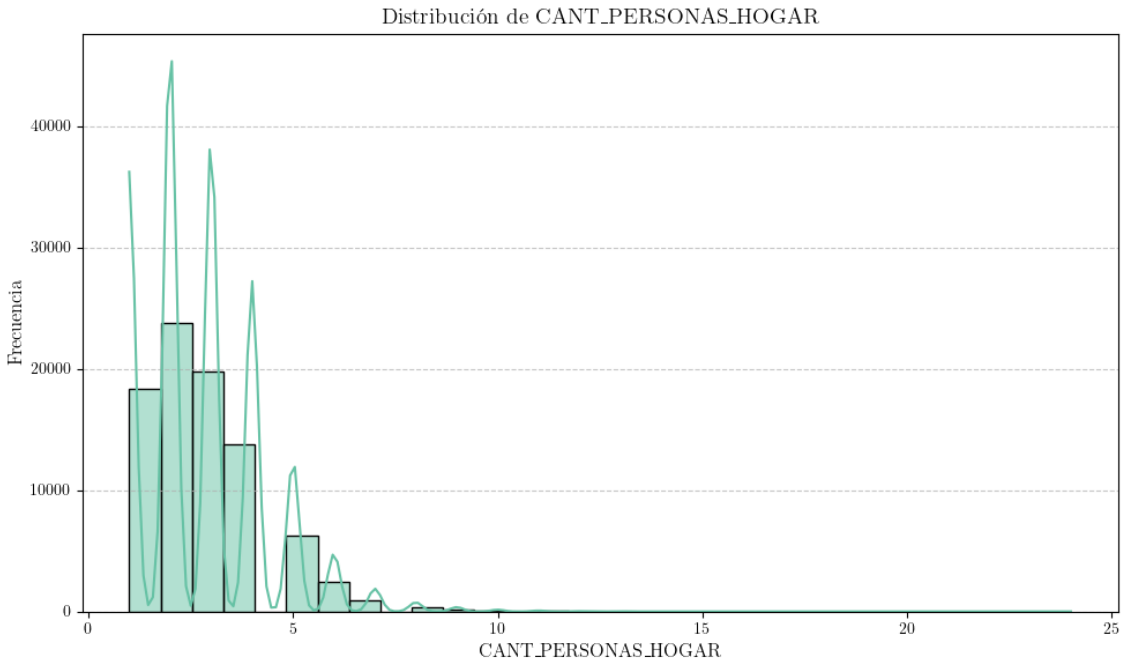
6.2.1. Variables Cuantitativas

Cantidad de Personas en el Hogar (CANT_PERSONAS_HOGAR). El promedio de personas por hogar es de 2.78, con una desviación estándar de 1.50, lo que indica una variabilidad moderada entre los hogares. El valor mínimo reportado es de 1, mientras que el valor máximo es de 24 personas, lo cual podría corresponder a hogares extendidos o situaciones excepcionales. El 50 % de los hogares tiene 3 o menos integrantes, lo que sugiere una predominancia de hogares pequeños.

En la Figura 6.1, se observa la distribución de la cantidad de personas por hogar. La mayoría de los hogares tienen entre 2 y 4 miembros, lo cual es consistente con la

estructura demográfica típica en el país. Existen hogares con hasta 24 personas, lo que representa casos extremos y posiblemente corresponde a viviendas con múltiples familias o estructuras colectivas.

Figura 6.1: Distribución de la cantidad de personas en el hogar.



Fuente: Encuesta de Calidad de Vida 2023. Cálculos Propios.

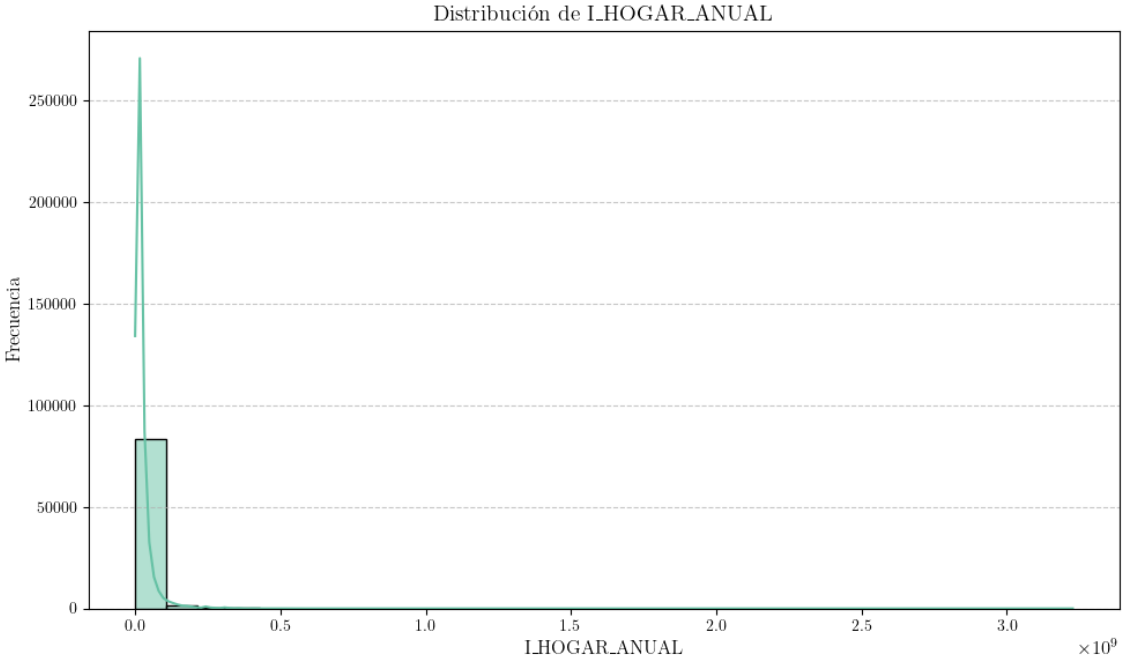
Ingreso Anual del Hogar (I_HOGAR_ANUAL). El ingreso anual promedio por hogar es de $25,124,400 COP, con una alta dispersión (desviación estándar de $44,779,750 COP), lo que refleja desigualdad significativa en los ingresos. El ingreso mínimo reportado es de $0 COP, mientras que el máximo asciende a $3,226,860,000 COP. El 75 % de los hogares tiene ingresos anuales menores a $27,360,000 COP, mostrando que la mayoría se concentra en valores relativamente bajos.

La Figura 6.2 muestra la distribución del ingreso anual de los hogares. Es notable que una gran proporción de los hogares tienen ingresos bajos, concentrándose principalmente en valores cercanos a cero. Esto podría estar relacionado con los niveles de pobreza en ciertas regiones del país, aunque también puede reflejar hogares que no reportaron ingresos en la encuesta.

Gasto Total Anual en Salud (GASTO_TOTAL_ANUAL). El gasto promedio de bolsillo en salud es de $211,169 COP, con una desviación estándar de $1,313,746 COP. Aunque la mayoría de los hogares no reporta gasto (mediana de $0 COP), hay un gasto máximo de $95,400,000 COP, lo que indica que un pequeño porcentaje de hogares enfrenta costos de salud extraordinariamente altos.

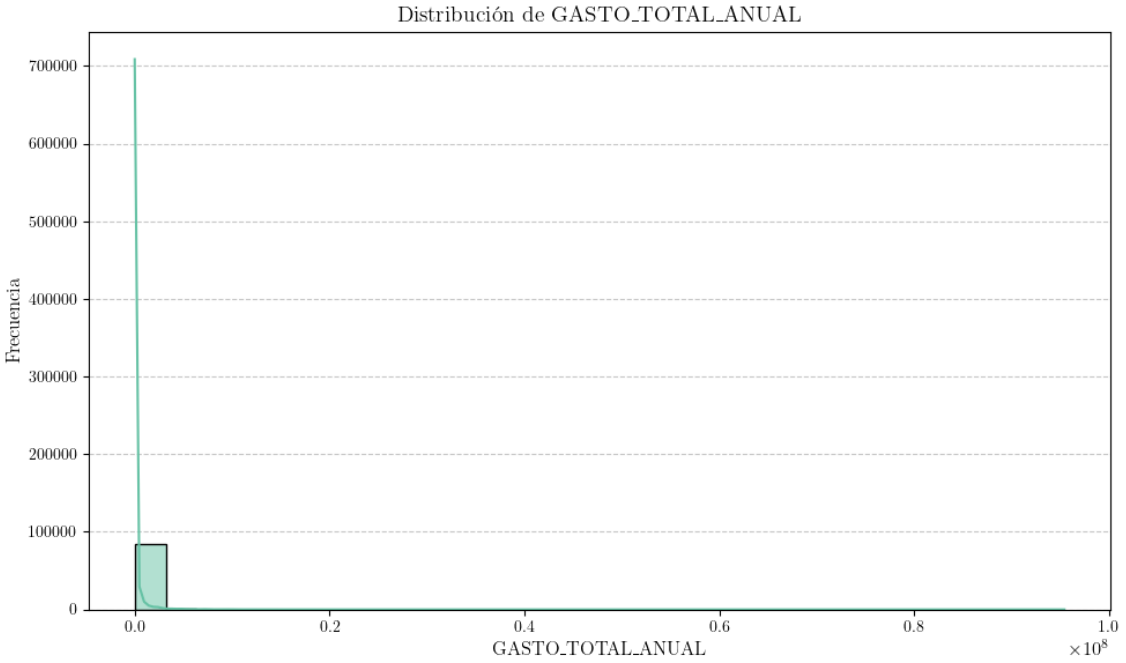
En la Figura 6.3, se presenta la distribución del gasto total anual de bolsillo en salud de los hogares. Una parte significativa de los hogares no reporta gasto en salud, lo cual puede deberse a la cobertura del sistema de seguridad social. Sin embargo, se identifican hogares con gastos extremadamente altos, los cuales podrían ser objeto de estudio para analizar los factores asociados a estos valores.

Figura 6.2: Distribución del ingreso anual de los hogares



Fuente: Encuesta de Calidad de Vida 2023. Cálculos Propios.

Figura 6.3: Distribución del gasto de bolsillo total anual de los hogares



Fuente: Encuesta de Calidad de Vida 2023. Cálculos Propios.

Respecto al gasto de bolsillo en salud, que es la variable dependiente para la estimación de modelos de este trabajo, es preciso decir que tiene una distribución muy sesgada, es decir, los valores extremos o ceros dominan la distribución. En consecuencia, será necesario realizar algún tipo de transformación para poder estimar modelos predictivos con un nivel de ajuste aceptable.

6.2.2. Variables Cualitativas

Afiliación a Seguridad Social (P6090). El 96.18 % de los hogares cuenta con al menos un integrante afiliado al sistema de seguridad social en salud, mientras que el 3.52 % no está afiliado. Un 0.30 % no proporciona información. Estos datos sugieren una cobertura generalizada del sistema, aunque persisten brechas de acceso.

Régimen de Seguridad Social en Salud (P6100). El régimen subsidiado cubre al 66.49 % de los hogares, seguido por el régimen contributivo (27.29 %). Un 2.17 % pertenece a regímenes especiales y un 0.22 % no proporciona información. Esto refleja una alta dependencia del régimen subsidiado, asociado con población de menores ingresos.

Estado General de Salud (P6127). El 68.95 % de los hogares percibe el estado de salud general como bueno, mientras que un 18.61 % lo considera regular y un 11.25 % muy bueno. Solo el 1.20 % califica su salud como mala. Esto sugiere que la mayoría percibe su salud de forma positiva.

Diagnóstico de Enfermedades Crónicas (P1930). El 82.20 % de los hogares reporta no tener integrantes con enfermedades crónicas diagnosticadas, mientras que un 17.80 % sí lo hace. Esta proporción es relevante para evaluar la carga de salud en los hogares.

Calidad del Servicio de la EPS (P6181). El 67.43 % califica la calidad del servicio como buena, mientras que un 16.50 % la percibe como regular. Solo un 6.98 % considera el servicio muy bueno, y un 2.97 % lo califica como malo. Un 2.07 % no proporciona opinión. Estos resultados reflejan una percepción mayoritariamente positiva, pero con áreas de mejora.

Región Geográfica (REGION). La mayor parte de los hogares se encuentra en la región Caribe (21.57 %), seguida por la región Pacífica (18.90 %) y Oriental (18.61 %). La menor proporción se encuentra en la región Exterior (1.02 %). Este patrón geográfico proporciona contexto sobre las condiciones socioeconómicas.

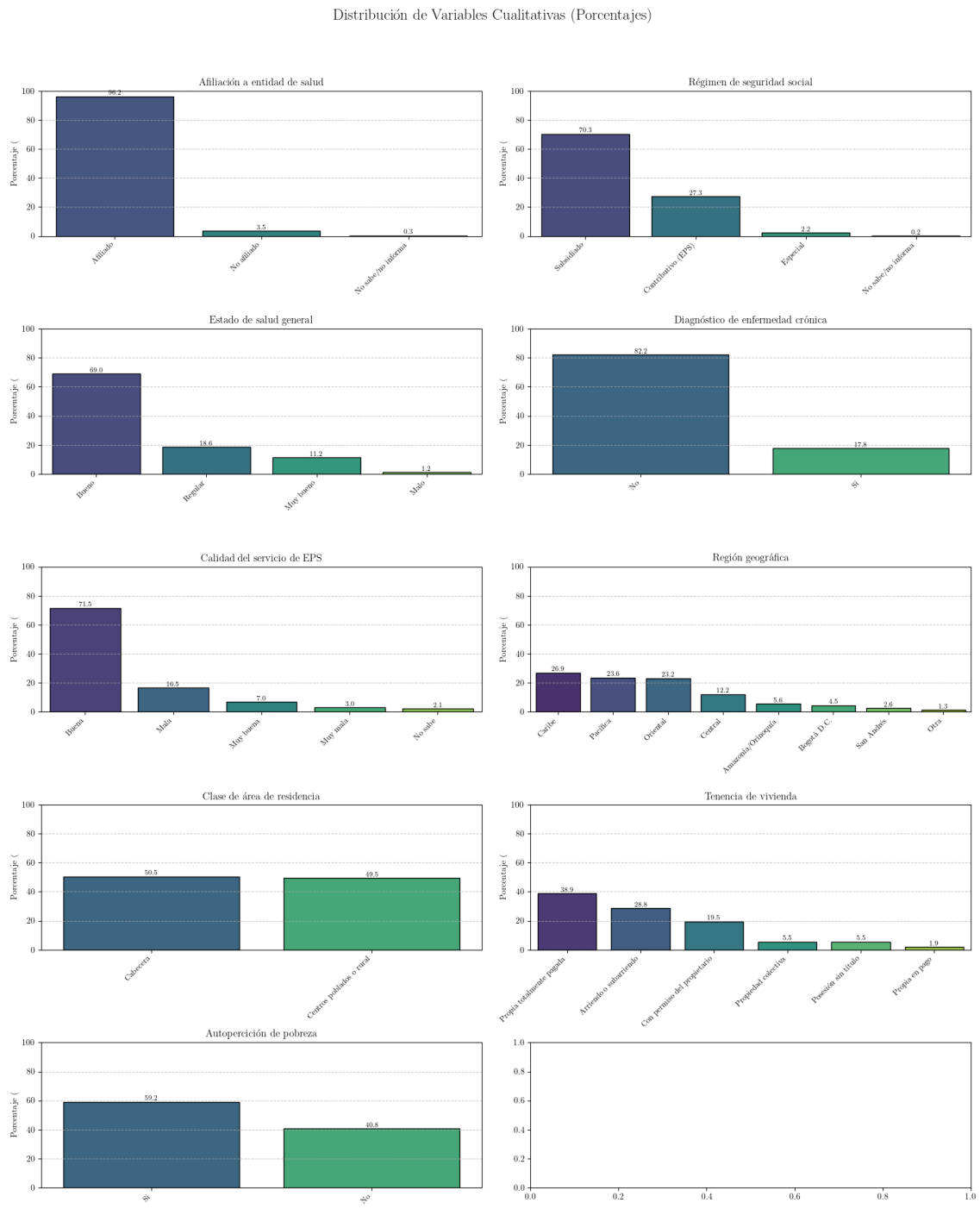
Clase de Localidad (CLASE). El 50.47 % de los hogares está ubicado en áreas urbanas, mientras que el 49.53 % reside en áreas rurales. Esto demuestra una distribución equilibrada entre zonas urbanas y rurales, relevante para diseñar políticas públicas diferenciadas.

Tenencia de Vivienda (P5095). El 38.86 % de los hogares posee vivienda propia totalmente pagada, mientras que el 28.76 % vive en arriendo. El 19.50 % ocupa vivienda en usufructo, mientras que las demás categorías representan proporciones menores. Esto indica que la propiedad es predominante, aunque el arriendo también es común.

Finalmente, la Figura 6.4 resume la distribución de las variables cualitativas en el dataset. La afiliación al sistema de seguridad social (P6090) revela que la mayoría de los encuestados están afiliados, lo que indica un alto nivel de cobertura. Asimismo, la calidad percibida del servicio de salud (P6181) se distribuye principalmente entre

las categorías "Buenaz "Muy buena", aunque existen respuestas indicando insatisfacción. Estas variables cualitativas proporcionan información relevante para entender las diferencias en el acceso y la percepción de los servicios de salud en distintos contextos.

Figura 6.4: Distribución de las variables cualitativas en el dataset.



Fuente: Encuesta de Calidad de Vida 2023. Cálculos Propios.

En el capítulo de Anexos se presentan los cuadros con el detalle de los resultados anteriormente descritos: el resumen descriptivo de las variables cuantitativas se encuentra en el Cuadro 10.1; la distribución de la variable Afiliación a Seguridad Social en Salud se presenta en el Cuadro 10.2; la distribución del Régimen de Seguridad

Social en Salud está en el Cuadro 10.3; la distribución del Estado General de Salud se puede consultar en el Cuadro 10.4; y la distribución de la variable Diagnóstico de Enfermedades Crónicas está en el Cuadro 10.5.

Asimismo, el Cuadro 10.6 contiene la distribución de la Calidad del Servicio de la EPS; la distribución por región se encuentra en el Cuadro 10.7; la variable CLASE (Tipo de Localidad) está en el Cuadro 10.8; la variable Tenencia de Vivienda se describe en el Cuadro 10.9; y finalmente, la variable Autopercepción de Pobreza está en el Cuadro 10.10.

6.3. Análisis de Correlación de Variables Cuantitativas

En esta sección se analiza la relación entre las variables cuantitativas seleccionadas en el dataset a través de una matriz de correlación, la cual se presenta en la Figura 6.5. Para medir la intensidad y dirección de las relaciones lineales entre estas variables, inicialmente se utilizó el coeficiente de correlación de Pearson. Sin embargo, debido a que las distribuciones de las variables analizadas presentan asimetría y no siguen una distribución normal, es importante interpretar los resultados con precaución. La falta de normalidad puede influir en la validez del coeficiente de Pearson, ya que este asume relaciones lineales en variables con distribución normal.

En respuesta a estas características de los datos, se sugiere complementar este análisis con medidas de correlación más robustas frente a distribuciones no normales, como la correlación de Spearman, que captura relaciones monótonas y no requiere normalidad en las variables.

El análisis de Pearson arroja las siguientes observaciones clave:

- **Cantidad de personas en el hogar (CANT_PERSONAS_HOGAR):** Esta variable presenta una correlación prácticamente nula con el gasto total anual de bolsillo en salud ($-0{,}03$), lo que indica que el tamaño del hogar no está relacionado directamente con el gasto de bolsillo. Por otro lado, la correlación positiva baja (0,09) con el ingreso anual del hogar (I_HOGAR_ANUAL) sugiere que los hogares más grandes tienden a tener ingresos ligeramente mayores, aunque esta relación es débil.

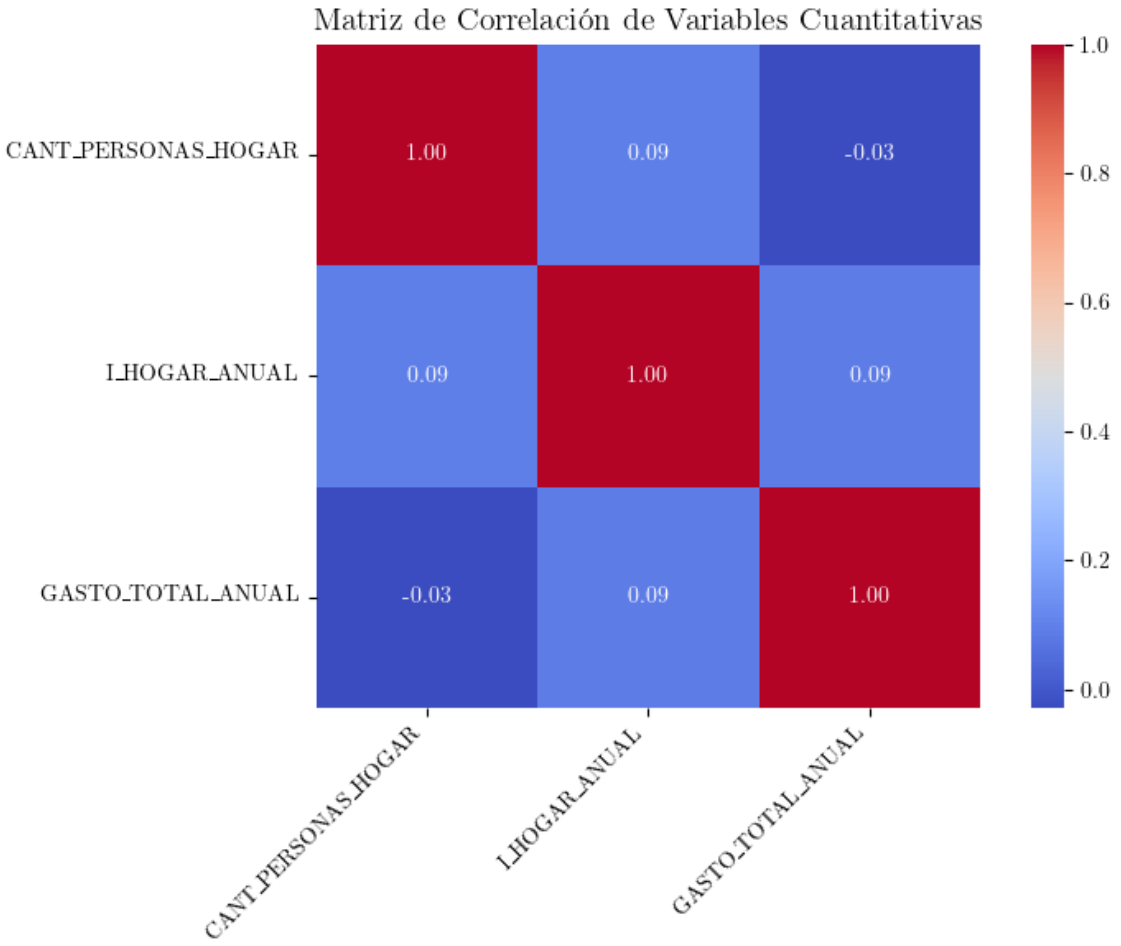
- **Ingreso anual del hogar (I_HOGAR_ANUAL):** Se observa una correlación baja positiva (0,09) con el gasto total anual de bolsillo en salud (GASTO_TOTAL_ANUAL). Este resultado indica que los ingresos del hogar pueden influir ligeramente en los gastos de bolsillo, pero no de manera significativa.

- **Gasto total anual de bolsillo en salud (GASTO_TOTAL_ANUAL):** La correlación con las otras variables es baja, lo que sugiere que el gasto de bolsillo no depende en gran medida ni del tamaño del hogar ni del ingreso anual. Este hallazgo resalta la importancia de explorar otras características, posiblemente categóricas, para entender los determinantes principales del gasto de bolsillo en salud.

En general, las correlaciones bajas entre las variables cuantitativas seleccionadas indican que estas no están fuertemente relacionadas entre sí. Además, debido a la

asimetría y no normalidad de las distribuciones, es recomendable emplear métodos adicionales, como la correlación de Spearman o modelos no paramétricos, para capturar patrones más complejos en los datos y explorar relaciones más allá de las lineales.

Figura 6.5: Matriz de correlación variables cuantitativas



Fuente: Encuesta de Calidad de Vida 2023. Cálculos Propios.

6.3.1. Análisis de Correlaciones: Coeficiente de Spearman

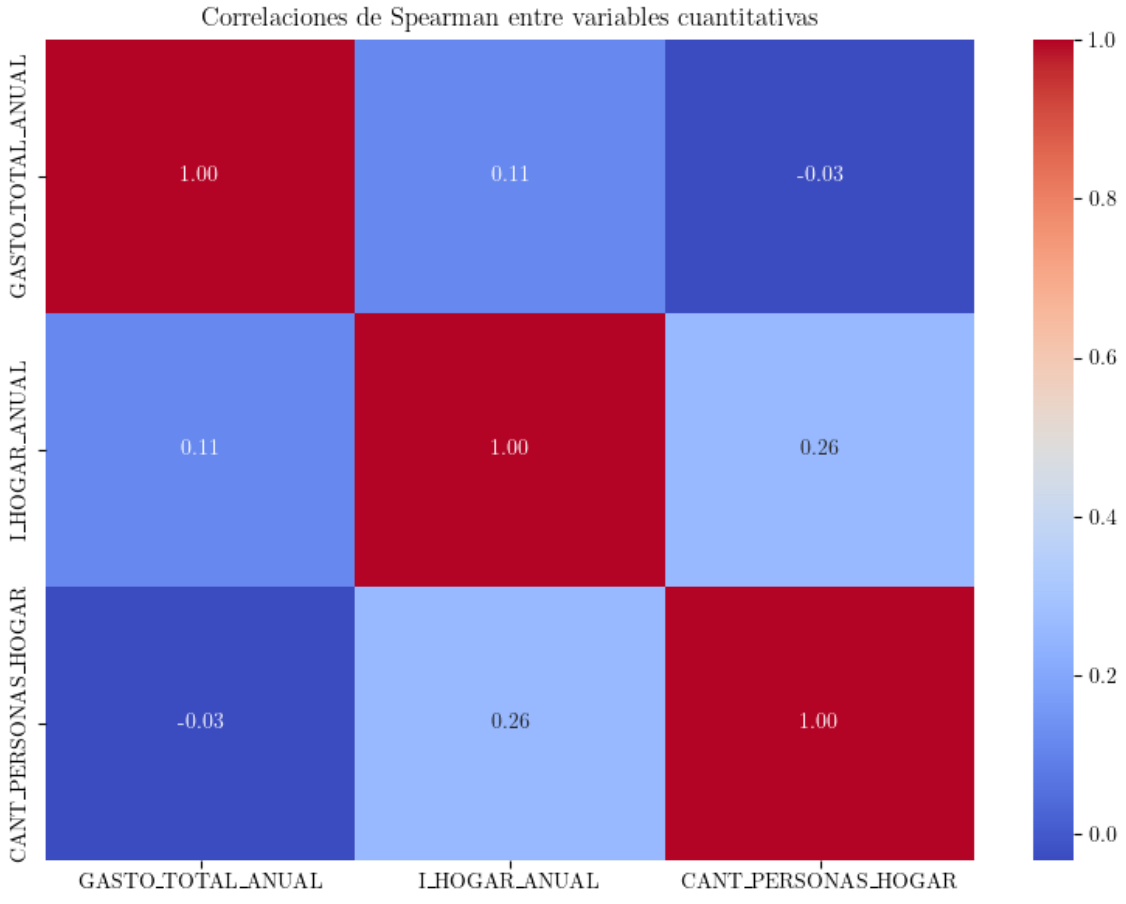
Descripción del procedimiento Dado que las variables cuantitativas utilizadas en este análisis presentan distribuciones asimétricas y no normales, se optó por calcular las correlaciones utilizando el coeficiente de Spearman en lugar del coeficiente de Pearson. El coeficiente de Spearman es una medida no paramétrica que evalúa la relación monotónica entre dos variables, lo que lo hace adecuado para datos que no cumplen con los supuestos de normalidad.

Resultados del análisis En la Figura 6.6, se presentan las correlaciones de Spearman entre las variables cuantitativas del estudio: gasto total anual, ingreso anual del hogar y cantidad de personas en el hogar. Los resultados indican lo siguiente:

- La correlación entre el gasto total anual y el ingreso anual del hogar es positiva y de magnitud baja ($r_s = 0{,}11$), lo que sugiere una relación débil pero consistente entre el ingreso del hogar y el gasto de bolsillo anual.
- Entre el gasto total anual y la cantidad de personas en el hogar, la correlación es prácticamente inexistente ($r_s = -0{,}03$), indicando que el tamaño del hogar no tiene una relación significativa con el gasto de bolsillo en salud.
- La correlación más fuerte se observa entre el ingreso anual del hogar y la cantidad de personas en el hogar ($r_s = 0{,}26$), aunque sigue siendo de magnitud baja, lo que implica que los ingresos tienden a aumentar ligeramente con el tamaño del hogar.

Implicaciones para el modelo predictivo Los resultados sugieren que, aunque el ingreso anual del hogar tiene una relación débil con el gasto de bolsillo, esta variable sigue siendo relevante para la predicción. Sin embargo, el tamaño del hogar muestra una correlación despreciable con el gasto, lo que podría limitar su contribución como predictor en los modelos desarrollados. Estos hallazgos destacan la necesidad de explorar interacciones o efectos no lineales que podrían no capturarse completamente mediante correlaciones simples.

Figura 6.6: Correlaciones de Spearman entre variables cuantitativas.



Fuente: Encuesta de Calidad de Vida 2023. Cálculos Propios.

6.4. Distribución del gasto entre categorías de las variables cualitativas

El análisis de las variables cualitativas se realizó utilizando diagramas de caja y bigotes (Figura 6.7) para comparar la distribución del gasto de bolsillo en salud según las categorías de cada variable. Debido a la alta dispersión y valores extremos observados en los datos, se utilizó una escala logarítmica para representar el gasto total anual. Esta escala permite una mejor visualización de las distribuciones y facilita la comparación entre categorías, minimizando el impacto de los valores atípicos que podrían sesgar las interpretaciones.

El análisis de las variables cualitativas se realizó utilizando diagramas de caja y bigotes (boxplots) para comparar la distribución del gasto total anual de bolsillo según las categorías de cada variable. Debido a la alta dispersión y valores extremos observados en los datos, se utilizó una escala logarítmica para representar el gasto total anual. Esta escala permite una mejor visualización de las distribuciones y facilita la comparación entre categorías, minimizando el impacto de los valores atípicos que podrían sesgar las interpretaciones.

En la variable **Afiliación a entidad de salud**, se observa que los hogares cuyos miembros están afiliados (*Afiliado*) presentan valores de gasto total anual más altos y una mayor dispersión en comparación con las categorías de *No afiliado* y los que *No saben o no reportaron información*. Esto puede indicar que la afiliación está relacionada con un mayor uso de servicios de salud y, por ende, con mayores gastos de bolsillo.

La variable **Régimen de seguridad social** muestra diferencias marcadas entre los tipos de régimen. Los hogares afiliados al régimen *Subsidiado* presentan un gasto de bolsillo más bajo, mientras que aquellos en el régimen *Contributivo (EPS)* registran valores más altos. Esto podría reflejar las diferencias en la cobertura y el acceso a servicios entre regímenes.

La variable **Estado de salud general** revela que los hogares donde el estado de salud se reporta como *Malo* tienen un gasto significativamente mayor en comparación con las categorías *Regular* y *Bueno*. Este hallazgo es consistente con la hipótesis de que una peor salud está asociada con mayores gastos médicos.

En cuanto a **Diagnóstico de enfermedad crónica**, los hogares con al menos un miembro diagnosticado con enfermedades crónicas (*Sí*) tienen un gasto de bolsillo considerablemente mayor que aquellos sin diagnósticos (*No*). Esto refuerza la idea de que las condiciones crónicas son un factor clave en el aumento de los gastos de salud.

La percepción de la calidad del servicio de salud, representada por **Calidad percibida del servicio EPS**, presenta una distribución más dispersa en las categorías *Muy buena* y *Buena*. Esto podría deberse a que los hogares con mejor percepción de calidad tienden a utilizar más servicios adicionales que no están cubiertos por el sistema.

La variable **Región geográfica** muestra diferencias regionales significativas en el gasto. Las regiones más urbanizadas, como la *Región Central*, tienen un gasto más alto, mientras que las regiones rurales presentan valores más bajos. Esto puede deberse tanto a menores ingresos como a menor acceso a servicios de salud en las zonas rurales.

La variable **Clase de área de residencia**, que identifica si el hogar está en

un área urbana o rural, indica que los hogares en áreas urbanas (*Cabecera*) tienen un gasto de bolsillo más alto que los hogares en áreas rurales dispersas (*Centros poblados o rural*). Esto podría explicarse por una mayor capacidad de gasto y acceso a servicios de salud en áreas urbanas.

Por su parte, la variable **Tenencia de vivienda** muestra que los hogares propietarios (*Propia totalmente pagada* y *Propia en pago*) tienen una mayor dispersión del gasto en comparación con los hogares arrendatarios o aquellos que habitan viviendas de forma no convencional. Esto puede estar relacionado con mayores ingresos disponibles en los hogares propietarios.

La variable **Autopercepción de pobreza**, que identifica si el hogar se considera pobre o no, indica que los hogares que no se consideran pobres tienen un gasto de bolsillo más alto que los hogares que se consideran a si mismos pobres. Esto podría explicarse por una mayor capacidad de gasto y acceso a servicios de salud en áreas urbanas.

En conclusión, el análisis de las variables cualitativas destaca la relevancia de factores socioeconómicos y de salud en la determinación del gasto de bolsillo en salud de los hogares. Este análisis proporciona una base sólida para incorporar estas variables en el modelo de predicción.

6.4.1. Pruebas estadísticas para estimar diferencias entre categorías de variables cualitativas

Descripción del Procedimiento Para evaluar las diferencias en el gasto de bolsillo entre las categorías de variables cualitativas, se realizaron pruebas estadísticas según el número de categorías en cada variable:

- **Prueba de Kruskal-Wallis**: Utilizada para variables con más de dos categorías. Esta prueba evalúa si las medianas de las categorías son significativamente diferentes, sin asumir la normalidad de los datos.
- **Prueba de Mann-Whitney U**: Aplicada a variables con solo dos categorías. Es una prueba no paramétrica adecuada para detectar diferencias significativas en la distribución entre dos grupos.

Estas pruebas permiten identificar las variables cualitativas con diferencias significativas en el gasto de bolsillo entre sus categorías, lo cual es fundamental para determinar su relevancia en los modelos predictivos.

6.4.2. Presentación de Resultados

Los resultados obtenidos de las pruebas estadísticas se resumen en el Cuadro6.3.

Implicaciones para la Estimación de los Modelos Los resultados obtenidos destacan que las variables cualitativas seleccionadas tienen una influencia significativa en el gasto de bolsillo, respaldada por los p-valores extremadamente bajos ($< 0{,}05$). Esto refleja cómo las diferencias en categorías específicas afectan el gasto, tales como:

- **Acceso y calidad de servicios de salud**: Variables como P6090 (Afiliación al sistema de salud), P6100 (Régimen de seguridad social) y P6181 (Calidad

Cuadro 6.3: Resultados de las pruebas de significancia para variables cualitativas

Variable	Prueba Estadística	Estadístico	p-valor
Afiliación al Sistema de Salud (P6090)	Kruskal-Wallis	89.42	$3{,}83 \times 10^{-20}$
Régimen de Seguridad Social (P6100)	Kruskal-Wallis	1850.48	$< 1 \times 10^{-30}$
Estado General de Salud (P6127)	Kruskal-Wallis	4082.20	$< 1 \times 10^{-30}$
Enfermedad Crónica (P1930)	Mann-Whitney U	$4{,}31 \times 10^{8}$	$< 1 \times 10^{-30}$
Calidad del Servicio de la EPS (P6181)	Kruskal-Wallis	651.35	$1{,}19 \times 10^{-139}$
Región Geográfica (REGION)	Kruskal-Wallis	2238.74	$< 1 \times 10^{-30}$
Clase Geográfica (CLASE)	Mann-Whitney U	$9{,}81 \times 10^{8}$	$1{,}16 \times 10^{-122}$
Tenencia de la Vivienda (P5095)	Kruskal-Wallis	768.22	$8{,}67 \times 10^{-164}$
Percepción de Pobreza (P5230)	Mann-Whitney U	$8{,}49 \times 10^{8}$	$4{,}68 \times 10^{-86}$

percibida de la EPS) reflejan la importancia del sistema de salud en los gastos de bolsillo.

- **Características socioeconómicas y demográficas**: Variables como REGION (Región geográfica), CLASE (Clase geográfica) y P5095 (Tenencia de la vivienda) influyen significativamente en las diferencias de gasto entre grupos poblacionales.
- **Percepción subjetiva de salud y pobreza**: Variables como P6127 (Estado general de salud) y P5230 (Percepción de pobreza) reflejan condiciones autoinformadas que impactan directamente en los gastos de bolsillo.

Estos resultados subrayan la importancia de incluir estas variables cualitativas en los modelos predictivos para capturar mejor las dinámicas complejas que determinan el gasto de bolsillo. En futuras investigaciones, sería pertinente explorar cómo estas variables interactúan con las cuantitativas y evaluar sus efectos moderadores o mediadores. Además, técnicas avanzadas como modelos jerárquicos o interacciones en algoritmos de aprendizaje automático podrían proporcionar mayor precisión en las predicciones.

6.5. Análisis de distribuciones espaciales

El análisis de las distribuciones geográficas permite examinar las diferencias en el gasto de bolsillo anual de los hogares según la región y la clase. Este tipo de análisis es fundamental para identificar desigualdades geográficas y comprender las áreas donde se concentran los mayores gastos. A continuación, se presentan los resultados de este análisis.

Distribución por Región La Figura 6.8 muestra la distribución del gasto promedio anual por región. Se observa que la región con el mayor gasto promedio es la región de San Andrés, mientras que las regiones con los gastos más bajos son otras y Caribe. Esta variabilidad puede deberse a diferencias en la disponibilidad y acceso a servicios de salud, así como a disparidades económicas y sociales entre las regiones. Este resultado refleja la disparidades regionales en el gasto de bolsillo en salud de los hogares.

Distribución por Clase La Figura 6.9 presenta el gasto promedio anual de bolsillo de los hogares según el tipo de área: urbana y rural. Los hogares en áreas urbanas muestran un gasto significativamente mayor en comparación con los hogares rurales.

Este hallazgo refleja la centralización de servicios de salud en zonas urbanas y su posible mayor costo, así como limitaciones en la infraestructura de salud en áreas rurales, que podrían llevar a una menor utilización de servicios y, por ende, menores gastos.

6.6. Modelo de Regresión Lineal Múltiple

La regresión lineal es una herramienta que permite explorar las relaciones entre una variable dependiente continua y un conjunto de variables independientes. Este modelo asume una relación lineal entre las variables y proporciona coeficientes interpretables, lo que lo convierte en una base inicial para analizar el impacto de diversas variables en el gasto total anual en salud. Aunque el modelo puede tener limitaciones, su simplicidad y transparencia lo hacen útil para identificar tendencias generales y patrones en los datos.

En este trabajo, se realizó una regresión lineal utilizando las 11 variables consideradas para la predicción del gasto de bolsillo en salud.

6.6.1. Procedimiento Realizado

1. **Selección de Variables:** Se seleccionaron las variables previamente identificadas como relevantes en el modelo predictivo, incluyendo tanto variables continuas como categóricas. Las variables categóricas fueron convertidas en variables dummies mediante codificación one-hot para su inclusión en el modelo.

2. **Ajuste del Modelo:** Se utilizó el método de mínimos cuadrados ordinarios (OLS) para ajustar el modelo. Este enfoque minimiza la suma de los errores al cuadrado entre los valores observados y las predicciones del modelo, produciendo estimaciones de los coeficientes para cada variable independiente.

3. **Evaluación del Modelo:** Se reportaron métricas como el coeficiente de determinación (R^2), el F-statistic y los intervalos de confianza para los coeficientes. Estas métricas permitieron evaluar el poder explicativo y la robustez del modelo. Se puede consultar el Cuadro 10.11 de los anexos.

6.6.2. Resultados e Interpretación

- **Coeficientes y Significancia Estadística:**
  - La variable **I_HOGAR_ANUAL** tuvo un coeficiente positivo y altamente significativo ($p < 0{,}001$), indicando que los ingresos del hogar son un predictor relevante del gasto en salud.
  - **CANT_PERSONAS_HOGAR** presentó un coeficiente negativo y significativo ($p < 0{,}001$), lo que sugiere que un mayor número de miembros en el hogar se asocia con menores gastos individuales en salud.
  - Entre las variables categóricas, las relacionadas con el régimen de salud (**P6100**), el estado general de salud (**P6127**) y la región (**REGION**) tuvieron coeficientes significativos que reflejan variaciones en el gasto según estas características.

- **Evaluación Global:**
  - El R^2 fue de 0.034, indicando que solo el 3.4 % de la variabilidad en el gasto total anual se explica por el modelo. Este valor sugiere que la relación entre las variables independientes y el gasto total anual no es predominantemente lineal.
  - El bajo ajuste del modelo refuerza la necesidad de explorar métodos más avanzados que puedan capturar relaciones no lineales o interacciones complejas entre las variables.

En el Cuadro 10.11, del capítulo de anexos, se puede encontrar los resultados del modelo de regresión lineal estimado.

6.6.3. Implicaciones para Estudios Futuros

Los resultados obtenidos de la regresión lineal tienen varias implicaciones directas para la estimación de los modelos propuestos en este trabajo:

- **Identificación de Variables Clave:** La significancia de variables como los ingresos del hogar y el estado general de salud valida su inclusión en modelos más avanzados, ya que estas variables tienen un impacto consistente en el gasto en salud.
- **Necesidad de Modelos No Lineales:** El bajo R^2 evidencia que el modelo no captura de manera adecuada las relaciones en los datos y puede que modelos no lineales tengan mejores resultados, como Gradient Boosting Machines o Random Forest, que pueden manejar interacciones y patrones más complejos.
- **Exploración de Interacciones:** Aunque el modelo no incluyó explícitamente interacciones entre variables, los resultados sugieren que estas podrían ser relevantes. Por ejemplo, la interacción entre el estado de salud y el régimen de afiliación podría mejorar la capacidad predictiva de los modelos.
- **Validación de la Importancia de las Variables Categóricas:** Las variables categóricas con coeficientes significativos, como el régimen de salud (**P6100**) y la percepción de calidad del servicio (**P6181**), confirman que estas categorías capturan aspectos relevantes del gasto en salud que deben considerarse en modelos predictivos más robustos.
- **Guía para la Selección de Hiperparámetros:** Los resultados obtenidos proporcionan un punto de referencia inicial para la configuración de hiperparámetros en modelos más complejos, al indicar qué variables tienen un impacto más directo en la variable dependiente.

Así las cosas, el modelo de regresión lineal múltiple es informativos en muchos aspectos para la estimación de modelos más complejos como los que se desarrollaran en el siguiente capítulo de este trabajo.

6.7. Síntesis Análisis Exploratorio de Datos

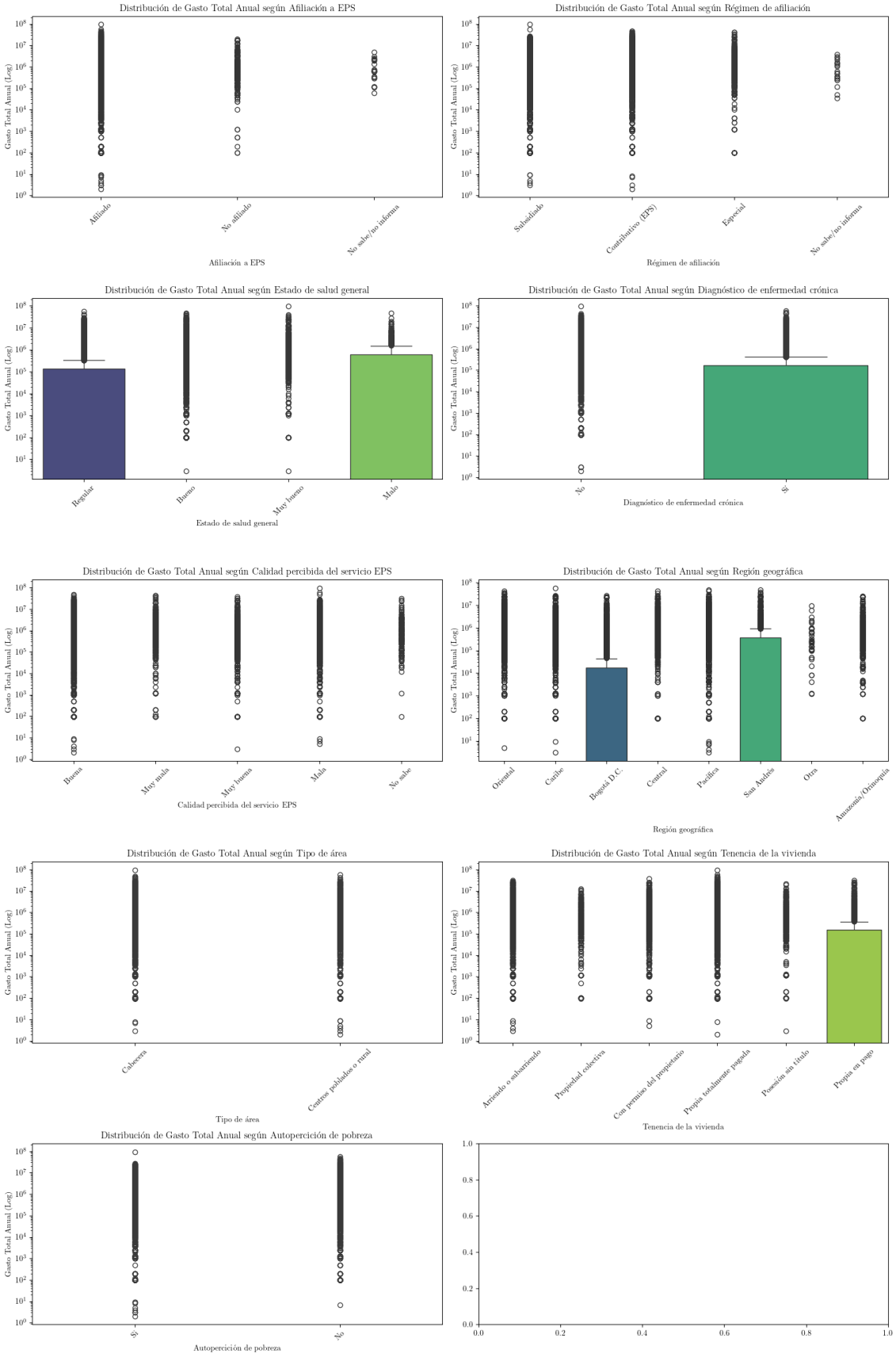
El análisis exploratorio de datos realizado en este capítulo ha permitido una comprensión de las características de los hogares colombianos y los factores asociados al gasto de bolsillo en salud. Los resultados evidencian la complejidad inherente de los datos, incluyendo la alta dispersión y asimetría en las variables cuantitativas, así como las disparidades observadas en las categorías de las variables cualitativas. Estos hallazgos destacan la importancia de abordar el problema con técnicas que permitan capturar tanto patrones lineales como no lineales en los datos.

Las correlaciones débiles entre las variables cuantitativas resaltan la necesidad de explorar interacciones y efectos moderadores entre variables categóricas y numéricas. Además, los resultados obtenidos de las pruebas estadísticas confirman que las diferencias entre categorías de las variables cualitativas son significativas, lo que valida su inclusión en los modelos predictivos para capturar la heterogeneidad de los hogares.

Este análisis proporciona un punto de partida para la construcción de los modelos de predicción y evidencia la relevancia de factores como el ingreso del hogar, el estado de salud y las características del régimen de seguridad social en la determinación del gasto de bolsillo. Sin embargo, las limitaciones observadas, como el bajo poder explicativo del modelo lineal y la alta variabilidad en los datos, subrayan la necesidad de explorar enfoques más sofisticados, como modelos de aprendizaje automático, que puedan capturar relaciones más complejas y mejorar la capacidad predictiva del análisis.

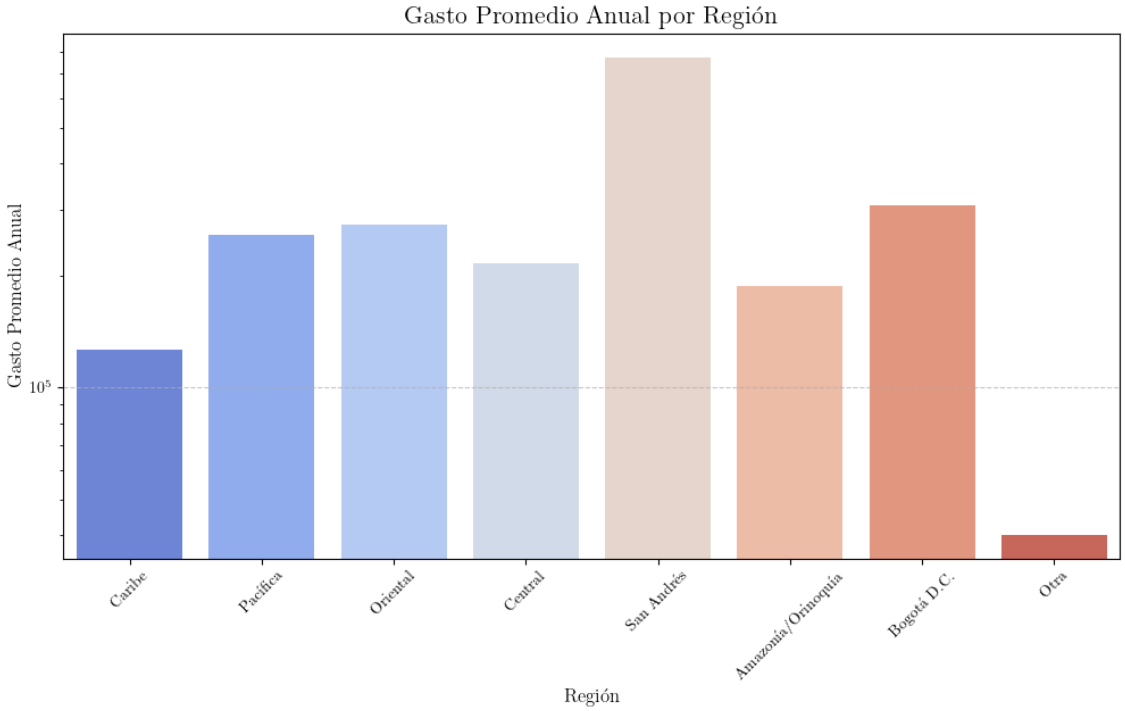
Es de particular preocupación las características de la variable dependiente, puesto tiene una distribución muy sesgada y puede que tenga que transformarse para poder modelar adecuadamente con los modelos de aprendizaje automático. En los próximos dos capítulos de este trabajo se va ampliar este punto y cómo se solventa.

Figura 6.7: Boxplots distribuición del gasto de bolsillo en salud entre variables cualitativas



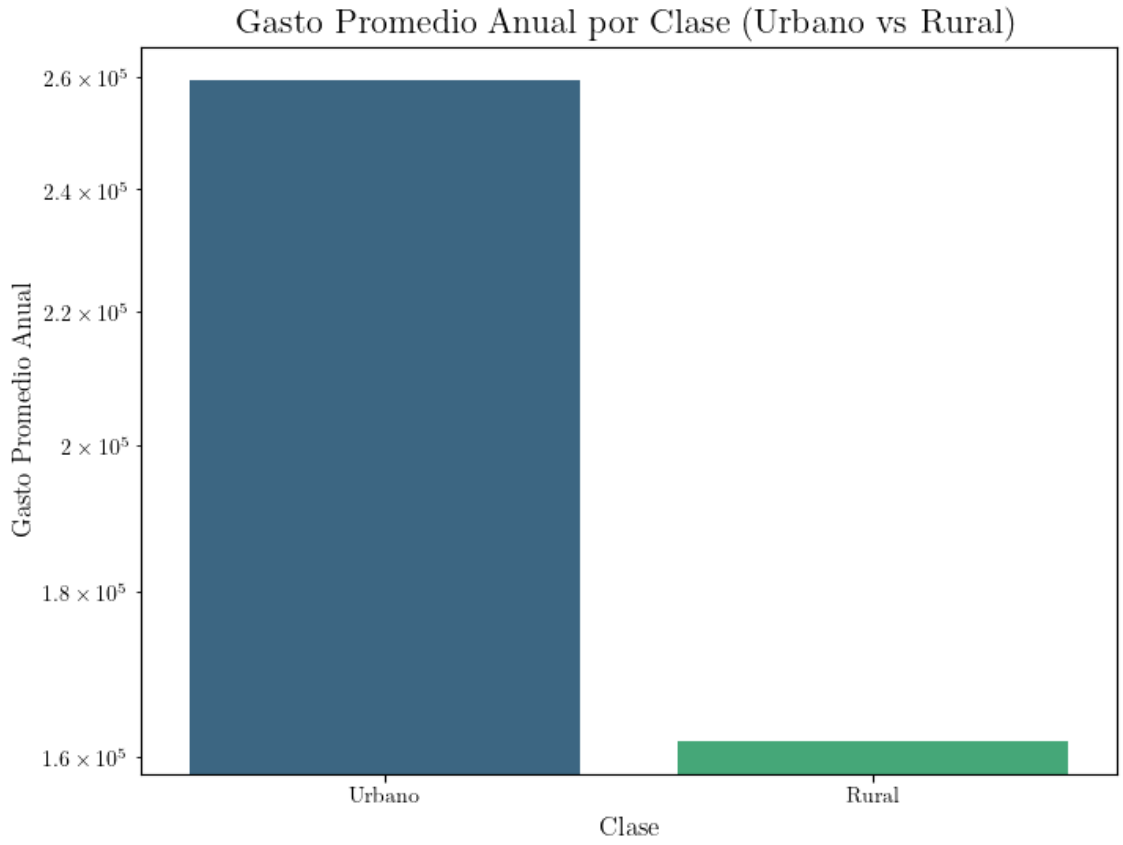
Fuente: Encuesta de Calidad de Vida 2023. Cálculos Propios.

Figura 6.8: Gasto Promedio Anual por Región



Fuente: Encuesta de Calidad de Vida 2023. Cálculos Propios.

Figura 6.9: Gasto Promedio Anual por Clase (Urbano vs Rural)



Fuente: Encuesta de Calidad de Vida 2023. Cálculos Propios.

Capítulo 7

Desarollo de modelos de regresión de Machine Learning

7.1. Preprocesamiento de los datos para la estimación de los modelos de regresión

El preprocesamiento de datos es una etapa crítica para garantizar la calidad y consistencia de los datos antes de su uso en los modelos predictivos. Además, tal como se evidencio en el análisis exploratorio de datos, es necesario realizar algunas transformaciones que permitan una estimación adecuada. En este capítulo, se implementaron las siguientes técnicas para preparar las variables tanto independientes como dependiente.

7.1.1. Transformación de la Variable Dependiente

La variable dependiente *Gasto Total Anual en Salud de los Hogares* presenta una alta dispersión y valores extremos. Para mitigar estos efectos y permitir un mejor desempeño de los modelos, se realizó una transformación logarítmica natural ajustada (`log1p`), que incluye un desplazamiento constante para manejar valores cercanos a cero:

$$\texttt{GASTO\_TOTAL\_ANUAL\_LOG} = \log(\texttt{GASTO\_TOTAL\_ANUAL} + 1) \tag{7.1.1}$$

Esta transformación reduce el impacto de los valores extremos y normaliza parcialmente la distribución de la variable.

7.1.2. Estandarización de Variables Cuantitativas

Las variables cuantitativas independientes, como el ingreso total anual del hogar y la cantidad de personas en el hogar, fueron estandarizadas utilizando la técnica de *Standard Scaling*:

$$z = \frac{x - \mu}{\sigma} \tag{7.1.2}$$

donde x es el valor original, μ es la media, y σ es la desviación estándar de la variable. Esto asegura que las variables tengan una media de 0 y una desviación estándar de 1, facilitando la convergencia en los algoritmos de aprendizaje automático.

7.1.3. Codificación de Variables Categóricas

Las variables categóricas se dividieron en dos grupos según su naturaleza:

- **Variables Nominales:** Estas variables, como `P6090` (afiliación a EPS) y `REGION` (región geográfica), se codificaron utilizando *One-Hot Encoding*. Este enfoque crea columnas binarias para cada categoría, eliminando una para evitar colinealidad (codificación *drop-first*).
- **Variables Ordinales:** Variables como `P6127` (estado de salud general) y `P6181` (calidad percibida del servicio EPS) tienen un orden lógico en sus categorías. Estas fueron transformadas usando *Ordinal Encoding*, asignando valores enteros crecientes según el orden lógico definido.

7.1.4. Tratamiento de Valores Atípicos

Para las variables cuantitativas, se manejaron valores atípicos mediante el uso del rango intercuartílico (`IQR`):

$$\texttt{IQR} = Q3 - Q1 \tag{7.1.3}$$

donde $Q1$ y $Q3$ son el primer y tercer cuartil, respectivamente. Los valores fuera del rango:

$$[Q1 - 1{,}5 \times \texttt{IQR}, Q3 + 1{,}5 \times \texttt{IQR}] \tag{7.1.4}$$

fueron eliminados para mejorar la robustez del análisis y evitar la influencia de outliers extremos.

7.1.5. Reagrupación de Categorías poco comunes

Se identificaron categorías con baja frecuencia en variables nominales codificadas, como *Afiliación a EPS* y *Régimen de afiliación*. Estas categorías fueron combinadas en una nueva categoría llamada `Other`, asegurando un balance adecuado en las frecuencias. El umbral para identificar categorías raras fue fijado en 1000 observaciones.

7.1.6. Verificación del Balance de Variables Categóricas

Después de la codificación y reagrupación, se realizó un análisis del balance de las categorías para verificar que ninguna categoría dominante sesgara los resultados. Las frecuencias de cada categoría codificada fueron reportadas y ajustadas en caso de desequilibrios significativos.

7.1.7. Resumen del Preprocesamiento

El proceso completo de preprocesamiento asegura que las variables sean adecuadas para los modelos predictivos al:

- Normalizar las variables cuantitativas para garantizar una escala comparable.
- Codificar las variables categóricas preservando su información semántica.
- Reducir la influencia de valores extremos mediante transformaciones logarítmicas y el manejo de outliers.
- Balancear las categorías raras para evitar sesgos en las predicciones.

7.2. Estimación del Modelo Gradient Boosting Regressor

En esta sección se describe el proceso de estimación y optimización del modelo *Gradient Boosting Regressor* para predecir el gasto total anual en salud de los hogares. Asimismo, se presentan los resultados obtenidos, incluyendo métricas de evaluación, validación cruzada e importancia de las características.

7.2.1. Optimización de Hiperparámetros

Para ajustar el modelo, se utilizó el método de búsqueda en cuadrícula (*GridSearchCV*) con validación cruzada de 5 particiones. Se exploraron combinaciones de hiperparámetros clave para identificar la configuración que minimizara el error cuadrático medio negativo.

La cuadrícula de hiperparámetros incluyó:

- `learning_rate`: Tasa de aprendizaje con valores {0.05, 0.1}.
- `max_depth`: Profundidad máxima de los árboles con valores {3, 5}.
- `n_estimators`: Número de estimadores con valores {100, 300}.
- `min_samples_split`: Mínimo de muestras para dividir un nodo con valores {2, 5, 10}.
- `min_samples_leaf`: Mínimo de muestras en una hoja con valores {1, 3, 5}.

El proceso ajustó un total de 360 modelos diferentes, identificando como mejores hiperparámetros:

- `learning_rate` = 0.1,
- `max_depth` = 3,
- `n_estimators` = 100,
- `min_samples_leaf` = 1,
- `min_samples_split` = 2

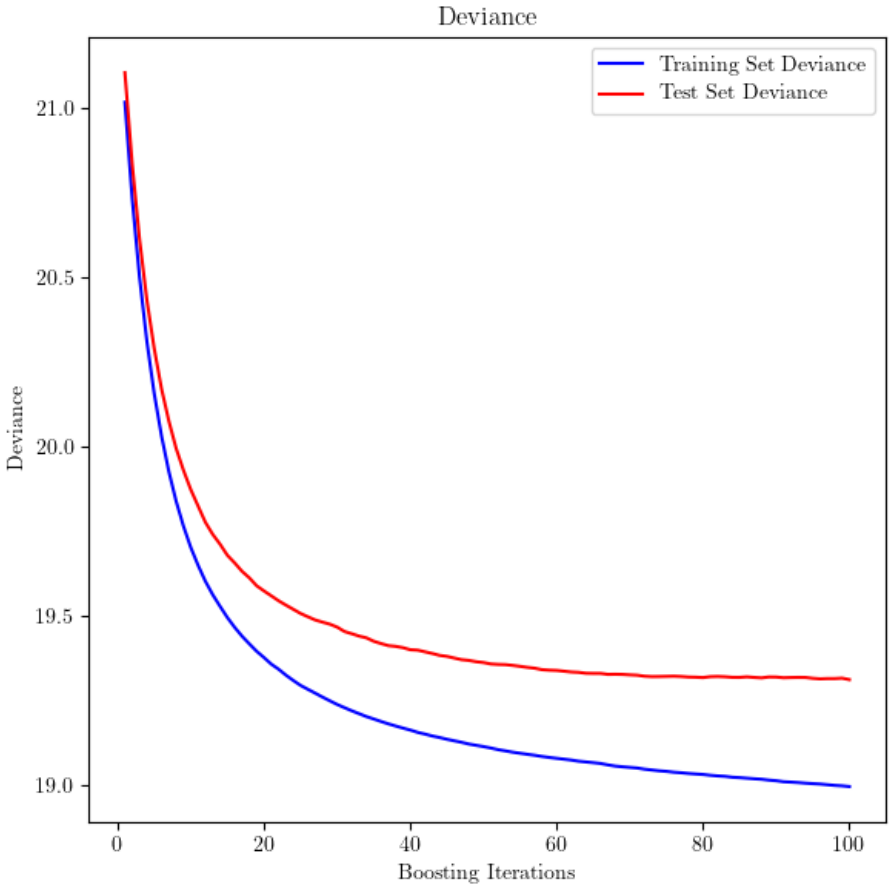
7.2.2. Evaluación del Modelo Optimizado

El modelo optimizado fue evaluado tanto en el conjunto de entrenamiento como en el conjunto de prueba. Las métricas obtenidas fueron:

- R^2 (entrenamiento): 0.11.
- R^2 (prueba): 0.10.
- Error Cuadrático Medio (`MSE`) en el conjunto de prueba: 913,715,409,512.97.
- Error Absoluto Medio (`MAE`) en el conjunto de prueba: 161,823.98.

La Figura 7.1 muestra la evolución de la desviación del modelo en función de las iteraciones. Se observa que el error en el conjunto de prueba disminuye significativamente al inicio, pero converge rápidamente, indicando que el modelo alcanza un punto de saturación con el número de estimadores óptimo.

Figura 7.1: Evolución de la desviación en los conjuntos de entrenamiento y prueba



7.2.3. Validación Cruzada

Se realizó una validación cruzada con 5 particiones para evaluar la generalización del modelo. Los resultados obtenidos para las métricas clave fueron:

- R^2:

- Promedio: 0.1012.
- Desviación estándar: 0.0052.

- Error Cuadrático Medio (`MSE`):
  - Promedio: 19.2034.
  - Desviación estándar: 0.9202.
- Error Absoluto Medio (`MAE`):
  - Promedio: 2.9640.
  - Desviación estándar: 0.1033.

Estos resultados indican una variabilidad baja entre las particiones, lo que sugiere un modelo consistente, aunque con capacidad predictiva limitada debido al bajo R^2.

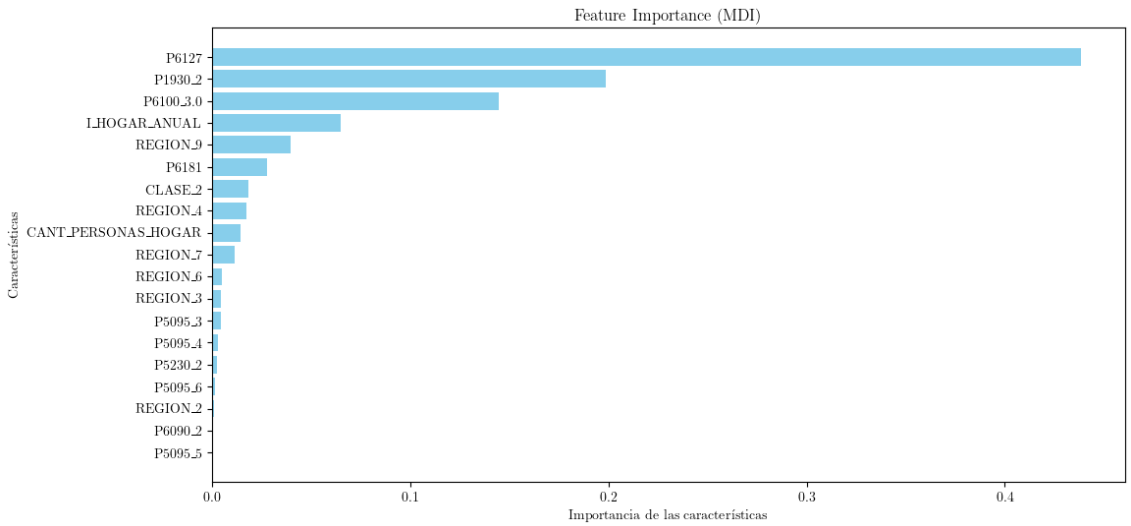
7.2.4. Importancia de las Características

Se evaluó la importancia de las características mediante dos enfoques: reducción de impureza media (*MDI*) y permutación.

Reducción de impureza media (Figura 7.2): Las variables más influyentes fueron:

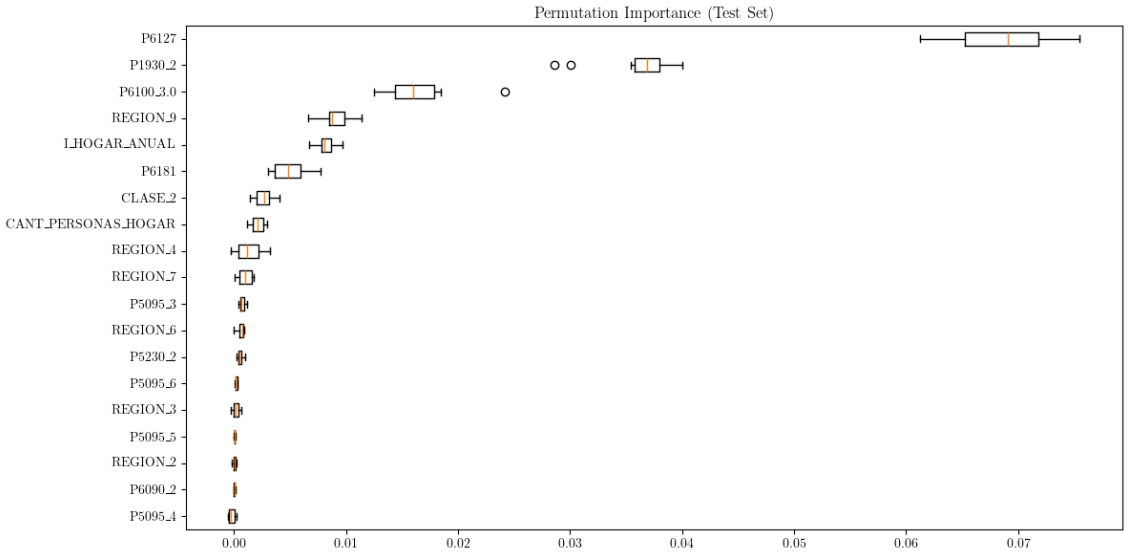
- `P6127` (Estado general de salud).
- `P1930_2` (Diagnóstico de enfermedad crónica).
- `P6100_3.0` (Régimen de afiliación: subsidiado).
- `I_HOGAR_ANUAL` (Ingreso total anual del hogar).

Figura 7.2: Importancia de las características según la reducción de impureza media (MDI)



Importancia por permutación (Figura 7.3): El análisis de permutación confirmó que las variables mencionadas anteriormente son las más relevantes, reforzando su impacto en el gasto de bolsillo.

Figura 7.3: Importancia de las características mediante permutaciones



7.2.5. Conclusión

El modelo Gradient Boosting Regressor permitió identificar las principales características asociadas al gasto de bolsillo en salud, destacando factores relacionados con el estado de salud, enfermedades crónicas y afiliación al sistema de seguridad social. Sin embargo, el bajo valor de R^2 sugiere que el modelo tiene capacidad predictiva sumamente limitada y dado que en el proceso de optimización de hiperparámetros se ajustaron un gran número de modelos diferentes para obtener los mejores, no se obtuvieron resultados satisfactorios.

7.3. Estimación del Modelo de Random Forest para Regresión

En esta sección se describe el proceso de estimación y optimización del modelo de regresión basado en Random Forest. Además, se presentan los resultados obtenidos, incluyendo métricas de evaluación, validación cruzada e importancia de las características.

7.3.1. Optimización de Hiperparámetros

El modelo fue ajustado utilizando *GridSearchCV* con validación cruzada de 5 particiones. Se exploraron combinaciones de hiperparámetros clave:

- `n_estimators`: Número de estimadores {50, 100}.
- `max_depth`: Profundidad máxima del árbol {3, 5}.
- `min_samples_split`: Mínimo de muestras para dividir un nodo {2, 5, 10}.

- `min_samples_leaf`: Mínimo de muestras en una hoja {1, 2, 4}.
- `max_features`: Número máximo de características consideradas para dividir un nodo {`auto`, `sqrt`, `log2`}.

El proceso ajustó un total de 540 modelos, identificando como mejores hiperparámetros:

- `n_estimators` = 100
- `max_depth` = 5
- `max_features` = `sqrt`
- `min_samples_leaf` = 4
- `min_samples_split` = 2

7.3.2. Evaluación del Modelo Optimizado

El modelo optimizado fue evaluado tanto en el conjunto de entrenamiento como en el de prueba. Las métricas obtenidas son:

- R^2 (entrenamiento): 0.09.
- R^2 (prueba): 0.09.
- Error Cuadrático Medio (`MSE`) en el conjunto de prueba: 913,783,622,292.77.
- Error Absoluto Medio (`MAE`) en el conjunto de prueba: 161,829.08.

La Figura 7.4 muestra la comparación entre los valores reales y predichos. Al igual que con otros modelos, las predicciones tienden a concentrarse cerca de los valores bajos, indicando limitaciones en la captura de valores extremos.

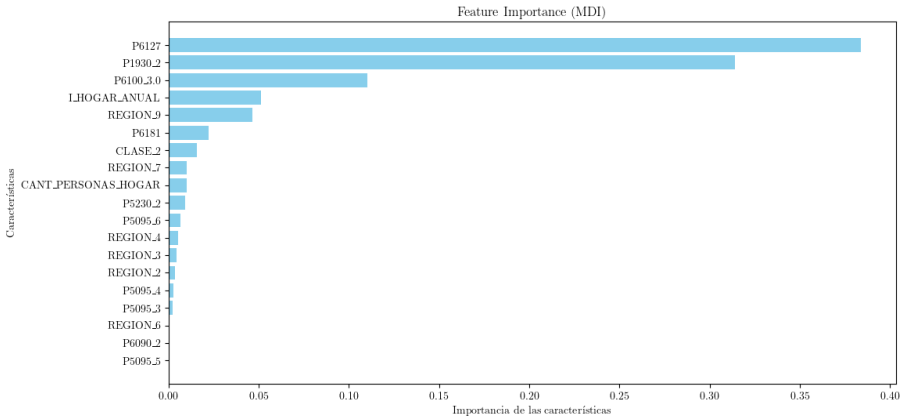


Figura 7.4: Comparación entre valores reales y predichos para el modelo de Random Forest.

7.3.3. Validación Cruzada

Se realizó una validación cruzada para evaluar la generalización del modelo. Los resultados promedio y desviación estándar para cada métrica fueron:

- R^2:
  - Promedio: 0.0905.
  - Desviación estándar: 0.0041.
- Error Cuadrático Medio (`MSE`):
  - Promedio: 19.4325.
  - Desviación estándar: 0.9314.
- Error Absoluto Medio (`MAE`):
  - Promedio: 3.0496.
  - Desviación estándar: 0.0836.

Estos resultados reflejan una variabilidad baja entre las particiones, pero la capacidad predictiva del modelo sigue siendo limitada.

7.3.4. Importancia de las Características

La importancia de las características fue evaluada utilizando dos enfoques: reducción de impureza media (*MDI*) y permutación.

Reducción de impureza media (Figura 7.5): Las principales variables fueron:

- `P6127` (Estado general de salud), como la variable más relevante.
- `P1930_2` (Diagnóstico de enfermedad crónica).
- `P6100_3.0` (Régimen de afiliación: subsidiado).
- `I_HOGAR_ANUAL` (Ingreso total anual del hogar).

Importancia por permutación (Figura 7.6): El análisis de permutación confirma que las mismas variables son las más influyentes, destacando su relevancia en la predicción del gasto de bolsillo en salud.

7.3.5. Conclusión

El modelo de Random Forest para regresión identificó variables clave relacionadas con el gasto de bolsillo en salud, siendo el estado general de salud y el diagnóstico de enfermedades crónicas las más influyentes. Sin embargo, el desempeño general del modelo fue bastante pobre con un bajo R^2. Esto sugiere que la relación entre las variables y el gasto de bolsillo puede requerir técnicas más avanzadas o la inclusión de variables adicionales para mejorar la capacidad predictiva.

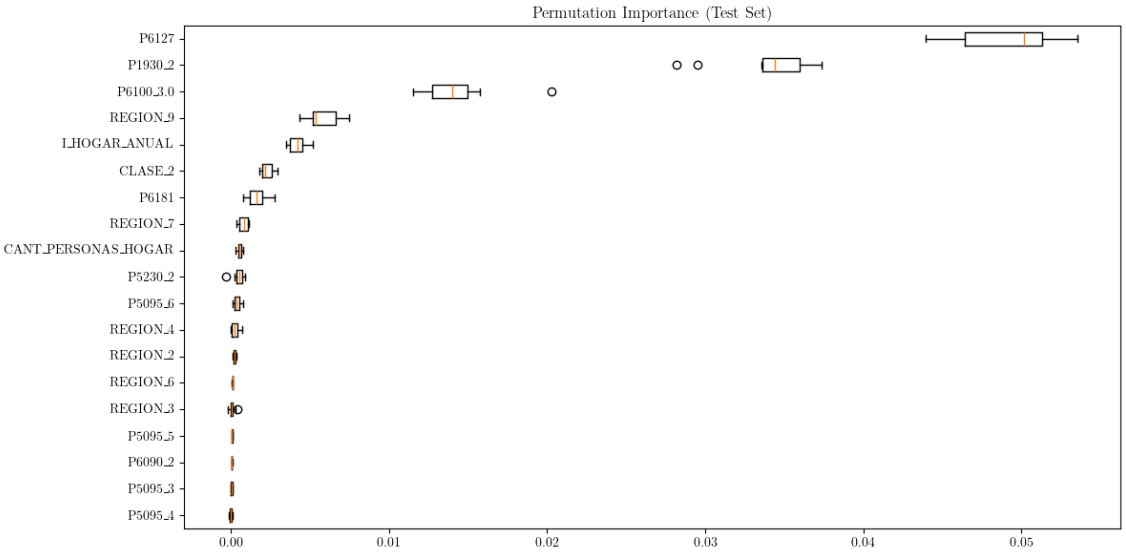


Figura 7.5: Importancia de las características según la reducción de impureza media (*MDI*).

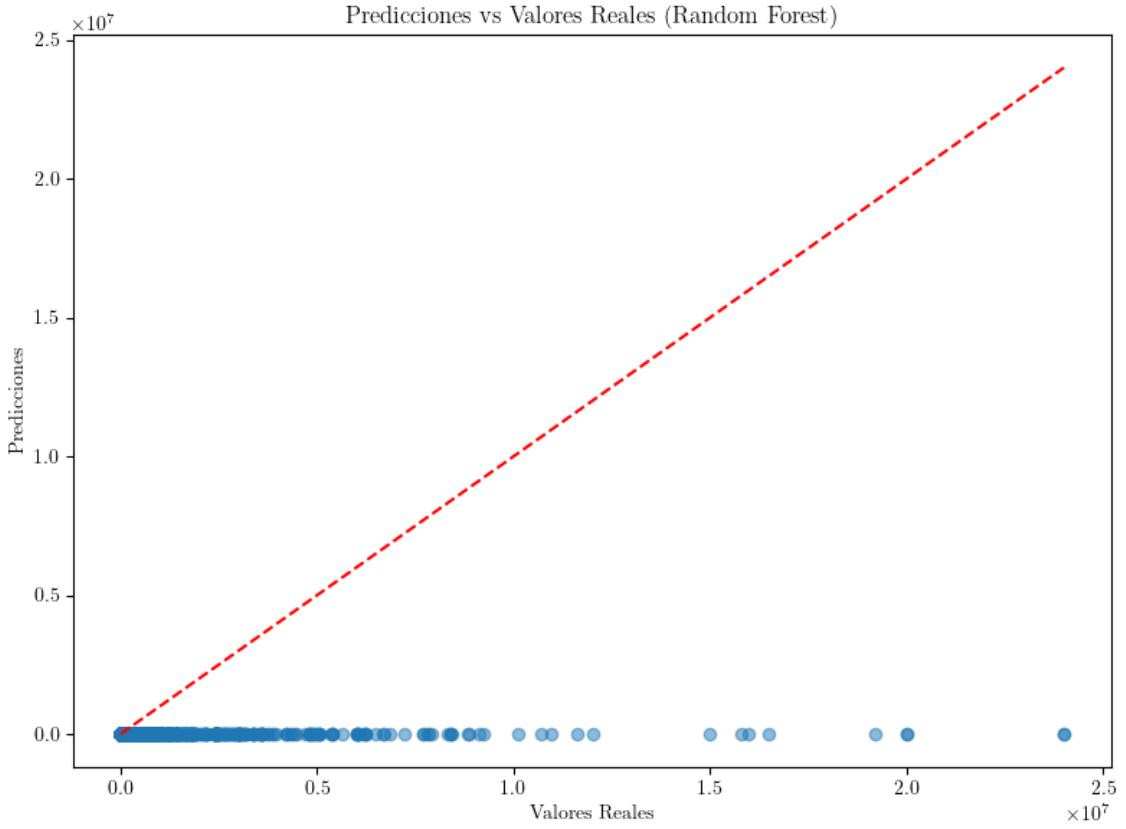


Figura 7.6: Importancia de las características mediante permutaciones (Random Forest)

7.4. Estimación del Modelo de Regresión Basado en Árboles de Decisión

En esta sección se detalla el procedimiento para estimar y optimizar un modelo de regresión basado en árboles de decisión. Además, se presentan los resultados obtenidos, incluyendo métricas de evaluación, validación cruzada y análisis de la importancia de las características.

7.4.1. Optimización de Hiperparámetros

El modelo fue ajustado utilizando *GridSearchCV* con validación cruzada de 5 particiones. Se exploraron diferentes combinaciones de hiperparámetros relevantes para los árboles de decisión:

- `max_depth`: Profundidad máxima del árbol {3, 5, 10, None}.
- `min_samples_split`: Mínimo de muestras para dividir un nodo {2, 5, 10}.
- `min_samples_leaf`: Mínimo de muestras en una hoja {1, 3, 5}.
- `max_features`: Número máximo de características consideradas para dividir un nodo {None, sqrt, log2}.

Tras ajustar un total de 540 modelos, los mejores hiperparámetros encontrados fueron:

$$\texttt{max\_depth} = 5, \quad \texttt{max\_features} = \text{None},$$

$$\texttt{min\_samples\_leaf} = 5, \quad \texttt{min\_samples\_split} = 2$$

7.4.2. Evaluación del Modelo Optimizado

El modelo optimizado se evaluó en los conjuntos de entrenamiento y prueba. Las métricas de desempeño obtenidas son:

- R^2 (entrenamiento): 0.10.
- R^2 (prueba): 0.09.
- Error Cuadrático Medio (`MSE`) en el conjunto de prueba: 913,676,259,584.17.
- Error Absoluto Medio (`MAE`) en el conjunto de prueba: 161,796.83.

La Figura 7.7 muestra la comparación entre los valores reales y predichos. Se observa que la mayoría de las predicciones se concentran cerca de valores bajos, indicando que el modelo tiene dificultades para capturar valores extremos.

7.4.3. Validación Cruzada

Se realizó una validación cruzada para evaluar la generalización del modelo. Los resultados promedio y su desviación estándar fueron:

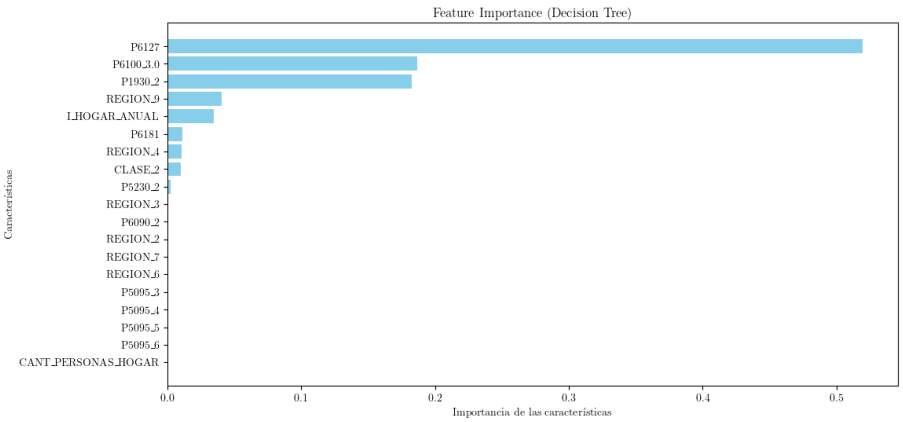


Figura 7.7: Comparación entre valores reales y predichos para el modelo de regresión basado en árboles de decisión.

- R^2:
  - Promedio: 0.0892.
  - Desviación estándar: 0.0067.
- Error Cuadrático Medio (`MSE`):
  - Promedio: 19.4612.
  - Desviación estándar: 0.9357.
- Error Absoluto Medio (`MAE`):
  - Promedio: 2.9823.
  - Desviación estándar: 0.0889.

Estos resultados indican que el modelo tiene un desempeño moderadamente consistente entre las particiones de la validación cruzada, aunque su capacidad predictiva es limitada.

7.4.4. Importancia de las Características

La importancia de las características se evaluó utilizando la reducción de impureza media. La Figura 7.8 presenta las principales variables explicativas. Destacan:

- `P6127` (Estado general de salud), con la mayor contribución al modelo.
- `P6100_3.0` (Régimen de afiliación: subsidiado).
- `P1930_2` (Diagnóstico de enfermedad crónica).
- `REGION_9` y `I_HOGAR_ANUAL` también contribuyen de manera notable.

7.4.5. Conclusión

El modelo de regresión basado en árboles de decisión identificó variables clave asociadas al gasto de bolsillo en salud, como el estado general de salud, el

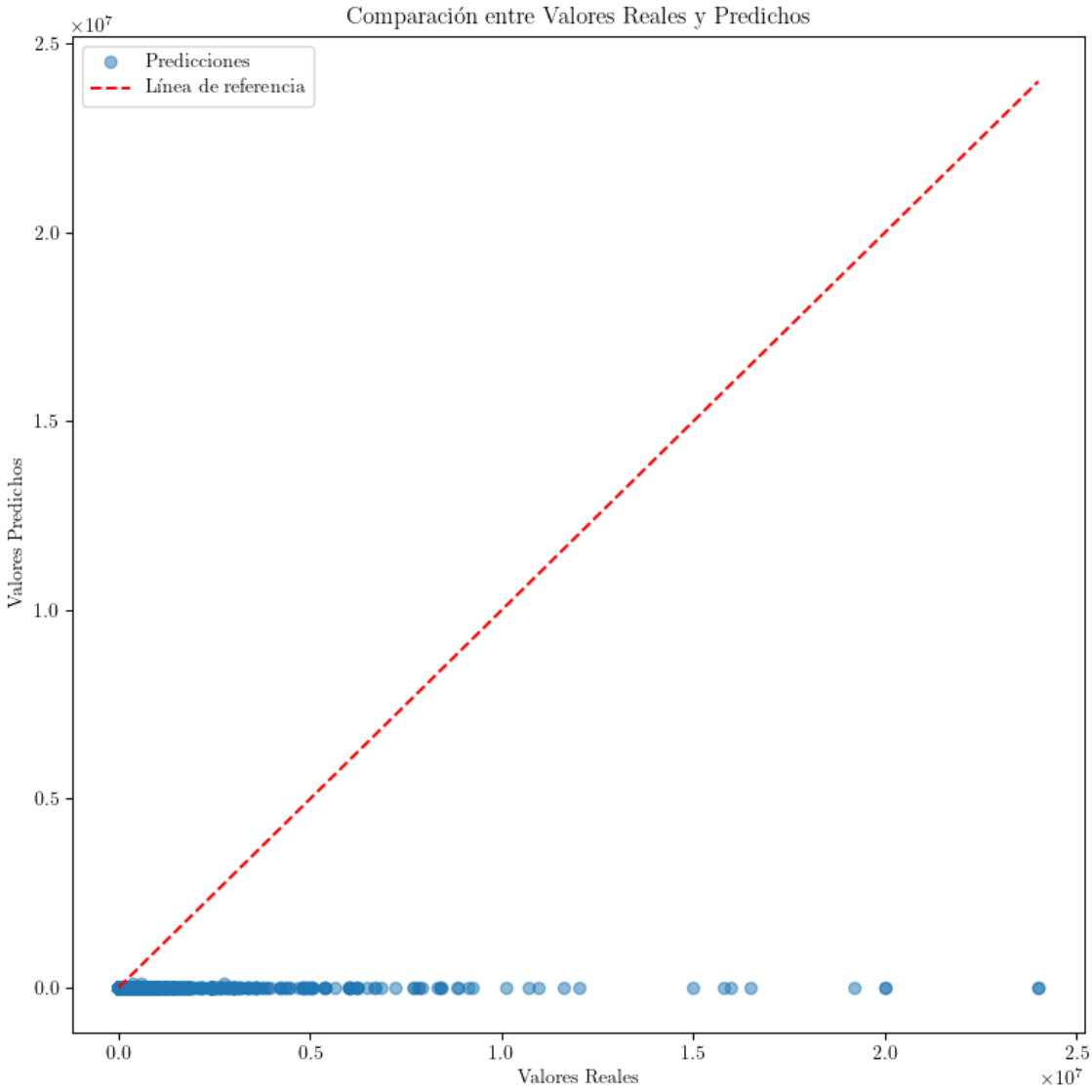


Figura 7.8: Importancia de las características según la reducción de impureza media (*Decision Tree*).

régimen de afiliación y la presencia de enfermedades crónicas. Sin embargo, su capacidad predictiva es muy baja, con un bajo R^2. Esto sugiere que el gasto de bolsillo puede depender de interacciones más complejas o de características no incluidas en el modelo.

7.5. Modelo MARS (Multivariate Adaptive Regression Splines)

7.6. Estimación del Modelo de Regresión MARS

En esta sección se detalla el procedimiento para estimar y optimizar un modelo de regresión basado en *Multivariate Adaptive Regression Splines* (MARS). Se

presentan los resultados obtenidos, incluyendo métricas de evaluación, validación cruzada y análisis de la importancia de las características.

7.6.1. Optimización de Hiperparámetros

El modelo MARS fue ajustado utilizando la librería `pygam`, que permite incorporar funciones base (*splines*) para cada variable independiente. Durante el ajuste, se realizaron las siguientes configuraciones:

- Cada variable independiente se modeló como una función base con 10 *splines*.
- Se utilizó la búsqueda de hiperparámetros (*gridsearch*) para optimizar los parámetros del modelo.

El ajuste fue realizado con validación cruzada utilizando el conjunto de entrenamiento (80 % de los datos), y el modelo final se seleccionó en función de su desempeño en las particiones de validación.

7.6.2. Evaluación del Modelo Optimizado

El modelo optimizado se evaluó en los conjuntos de entrenamiento y prueba. Las métricas de desempeño obtenidas son:

- R^2 (entrenamiento): 0.10.
- R^2 (prueba): 0.09.
- Error Cuadrático Medio (`MSE`) en el conjunto de prueba: 19.32.
- Error Absoluto Medio (`MAE`) en el conjunto de prueba: 2.98.

La Figura 7.9 muestra la comparación entre los valores reales y predichos. Aunque existe una alineación general, se observa que el modelo tiene dificultades para capturar los valores más altos.

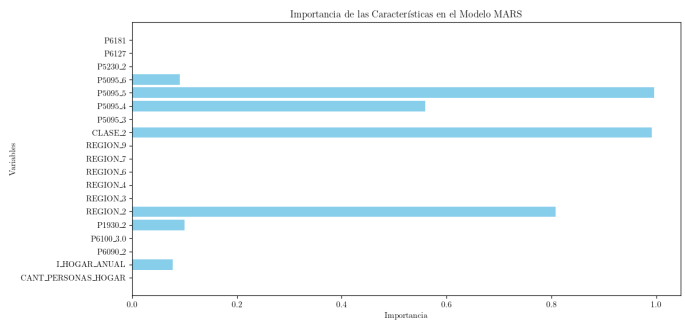


Figura 7.9: Comparación entre valores reales y predichos para el modelo MARS.

7.6.3. Validación Cruzada

Se realizó una validación cruzada con 5 particiones para evaluar la generalización del modelo. Los resultados promedio y su desviación estándar fueron:

- R^2:
  - Promedio: 0.0971.
  - Desviación estándar: 0.0059.
- Error Cuadrático Medio (`MSE`):
  - Promedio: 19.2964.
  - Desviación estándar: 0.2741.
- Error Absoluto Medio (`MAE`):
  - Promedio: 2.9776.
  - Desviación estándar: 0.0195.

Estos resultados muestran que el modelo es consistente entre las particiones, aunque su capacidad predictiva sigue siendo limitada.

7.6.4. Importancia de las Características

La importancia de las características fue evaluada a través de los valores p del modelo. La Figura 7.10 presenta las principales variables explicativas. Destacan:

- `P5095_6`: Relacionada con la tenencia de la vivienda.
- `CLASE_2`: Indicador de ubicación geográfica (urbano o rural).
- `REGION_2`: Región geográfica.

7.6.5. Conclusión

El modelo MARS presentó un desempeño limitado con un $R^2 = 0{,}09$ en el conjunto de prueba, indicando que solo una pequeña proporción de la variabilidad del gasto de bolsillo es explicada por las variables incluidas. Las variables relacionadas con la tenencia de la vivienda, la ubicación geográfica y las regiones geográficas fueron las más influyentes. Este análisis sugiere explorar modelos adicionales o incorporar nuevas variables para mejorar la capacidad predictiva.

7.7. Comparación de Modelos

En esta sección se presenta una comparación entre los cuatro modelos estimados: *Gradient Boosting Regressor* (GBR), *Decision Tree Regressor* (DTR), *Random Forest Regressor* (RFR) y *Multivariate Adaptive Regression Splines* (MARS). Los criterios de comparación incluyen métricas de evaluación (R^2, MSE, MAE) y la validación cruzada, así como el análisis de la importancia de las características.

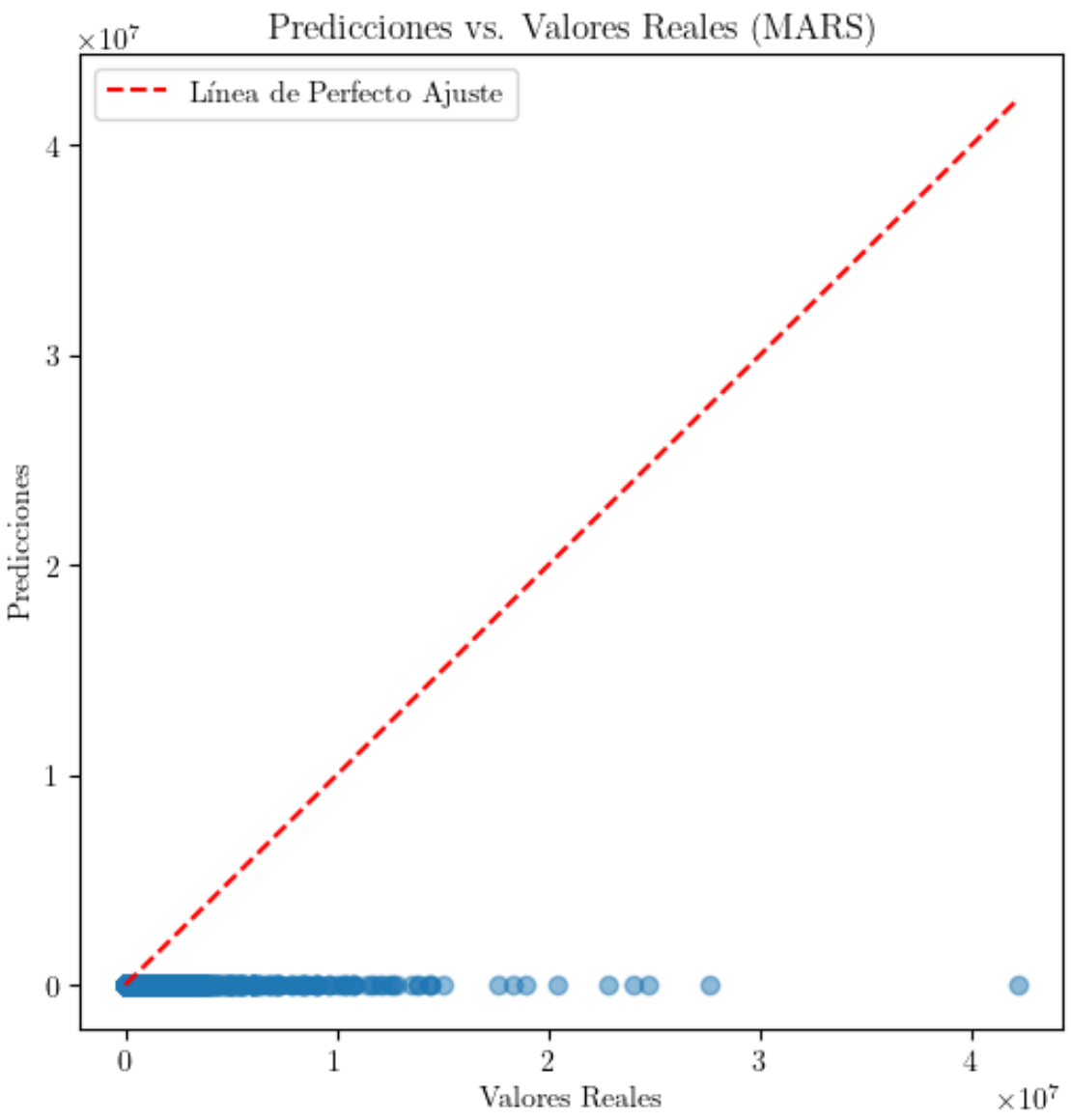


Figura 7.10: Importancia de las características en el modelo MARS.

7.7.1. Desempeño General de los Modelos

el Cuadro 7.1 resume las métricas clave de desempeño en los conjuntos de prueba para cada modelo:

Cuadro 7.1: Comparación del desempeño de los modelos en el conjunto de prueba

Modelo	**R^2**	**MSE (prueba)**	**MAE (prueba)**
Gradient Boosting Regressor	0.10	913.41	161.82
Decision Tree Regressor	0.09	913.26	161.79
Random Forest Regressor	0.09	913.62	161.83
MARS	0.09	913.00	162.00

De los modelos evaluados, todos presentan valores de R^2 bajos, indicando una limitada capacidad para explicar la variabilidad en el gasto de bolsillo

en salud. Sin embargo, el modelo MARS se diferencia al presentar un MSE considerablemente menor, dado que se evalúa en una escala logarítmica, lo que dificulta una comparación directa con los otros modelos.

7.7.2. Validación Cruzada

La validación cruzada proporcionó información adicional sobre la estabilidad y generalización de los modelos. el Cuadro 7.2 presenta los resultados promedio de R^2, MSE y MAE obtenidos durante la validación cruzada:

Cuadro 7.2: Resultados promedio de la validación cruzada

Modelo	**R^2**	**MSE (promedio)**	**MAE (promedio)**
Gradient Boosting Regressor	0.1012	19.2034	2.9640
Decision Tree Regressor	0.0892	19.4612	2.9823
Random Forest Regressor	0.0905	19.4325	3.0496
MARS	0.0971	19.2964	2.9776

En términos de R^2, el modelo GBR presenta un desempeño ligeramente superior con un promedio de 0.1012, seguido por MARS con 0.0971. En cuanto a MSE y MAE, todos los modelos presentan valores similares, destacándose el GBR y MARS por tener los promedios más bajos en estas métricas.

7.7.3. Importancia de las Características

El análisis de la importancia de las características mostró resultados consistentes entre los modelos. Las variables más relevantes incluyeron:

- `P6127` (Estado general de salud): Destacada en todos los modelos como la característica más importante.
- `P6100_3.0` (Régimen de afiliación subsidiado) y `P1930_2` (Diagnóstico de enfermedad crónica): Altamente relevantes en GBR, DTR y RFR.
- `P5095_6` (Tenencia de vivienda) y `CLASE_2` (Zona urbana/rural): Principales contribuciones en el modelo MARS.

7.7.4. Conclusión

Ningún modelo logro un ajuste aceptable mediante R^2, el modelo MARS destacó por su capacidad para identificar características clave con un menor error absoluto promedio, lo que puede ser relevante en aplicaciones prácticas. Sin embargo, el bajo desempeño general sugiere que los modelos realizados no tiene la capacidad para predecir de manera correcta el gasto de bolsillo en salud de los hogares.

7.8. Consideraciones sobre el Uso de Modelos de Regresión de *Machine Learning* para la Estimación del Gasto de Bolsillo en Salud de los Hogares

Los modelos de regresión desarrollados para la estimación del gasto de bolsillo en salud de los hogares demostraron un bajo nivel de ajuste, reflejado en valores de R^2 cercanos a 0.10 y errores de predicción altos en las métricas MSE y MAE. Este desempeño limitado puede explicarse, en buena medida, por las características intrínsecas de la variable dependiente, lo que impone desafíos estadísticos y metodológicos importantes para el desarrollo de este trabajo.

7.8.1. Análisis de la Distribución de la Variable Dependiente

La variable dependiente, correspondiente al gasto de bolsillo en salud de los hogares, presenta una distribución altamente sesgada con una proporción significativa de valores en cero. De las 86,063 observaciones disponibles:

- 71,850 observaciones (83.49 %) tienen un valor de gasto igual a cero.
- 14,213 observaciones (16.51 %) tienen un gasto mayor a cero.

Esta distribución implica que el gasto de bolsillo en salud está fuertemente concentrado en un subconjunto pequeño de la población, mientras que la mayoría de los hogares no incurre en gastos. Desde una perspectiva estadística, este tipo de distribución introduce problemas específicos para los modelos de regresión, como el sesgo hacia la predicción de valores cercanos al promedio o al modo de la distribución (en este caso, cero).

7.8.2. Limitaciones de los Modelos de Regresión para Datos Sesgados

El sesgo de la distribución y la alta proporción de ceros en la variable dependiente dificultan el ajuste adecuado de los modelos de regresión tradicionales y basados en *machine learning*. Este fenómeno ha sido documentado ampliamente en la literatura (Zuur et al., [22]; Tu, [40]), destacando que las regresiones estándar tienden a subestimar o ignorar características esenciales de los datos, tales como:

- La naturaleza bimodal o multimodal de la distribución.
- La presencia de una gran masa de datos en un valor específico (cero en este caso), que no puede ser explicada adecuadamente por una función continua simple.

- La heterogeneidad en la relación entre las características independientes y el gasto de bolsillo, que puede diferir drásticamente entre los hogares con y sin gasto.

7.8.3. Desafíos Observados en los Resultados de los Modelos

Los modelos desarrollados, incluyendo *Gradient Boosting Regressor*, *Decision Tree Regressor*, *Random Forest Regressor* y *MARS*, reflejan estas limitaciones:

- Los bajos valores de R^2 indican que los modelos no logran capturar la variabilidad observada en la variable dependiente.
- Las métricas de error, aunque útiles para evaluar el desempeño general, no reflejan adecuadamente la incapacidad de los modelos para distinguir entre hogares con gasto cero y aquellos con gasto positivo.
- Las predicciones están sesgadas hacia valores bajos, como se evidencia en los gráficos de comparación entre valores reales y predichos, donde los valores no cero están subestimados sistemáticamente.

7.8.4. Propuesta de Alternativa: Modelos de Clasificación

Dado el escenario presentado, se considera que los modelos de regresión no son la herramienta adecuada para abordar esta problemática. Como alternativa, se propone la implementación de modelos de clasificación que permitan distinguir entre hogares con gasto de bolsillo en salud (valores positivos) y hogares sin gasto (valores en cero). Este enfoque ofrece varias ventajas:

- Permite modelar directamente la probabilidad de pertenencia a una clase (*gasto cero* o *gasto positivo*), lo que se adapta mejor a la naturaleza discreta de la variable.
- Reduce el impacto del sesgo inducido por la alta proporción de ceros, ya que los modelos de clasificación pueden manejar desequilibrios de clase mediante técnicas como ponderación de clases o sobremuestreo.
- Facilita la interpretación de los resultados en términos probabilísticos, lo que puede ser más útil para la formulación de políticas y la toma de decisiones.

Capítulo 8

Desarollo de modelos de clasificación de Machine Learning

8.1. Elección de modelos de clasificación para la predicción del gasto de bolsillo en salud

La elección del enfoque metodológico para modelar el gasto de bolsillo en salud en los hogares colombianos ha sido un proceso fundamentado tanto en la naturaleza de los datos disponibles como en los objetivos del análisis. Inicialmente, se consideraron modelos de regresión con el objetivo de predecir la magnitud del gasto en salud. Sin embargo, los resultados obtenidos fueron consistentemente insatisfactorios, con métricas de desempeño como el coeficiente de determinación (R^2) indicando un bajo poder explicativo. Este desempeño limitado puede atribuirse, en gran medida, a la distribución altamente desbalanceada y heterogénea de la variable dependiente, donde más del 83 % de las observaciones corresponden a valores iguales a cero. Este alto porcentaje de ceros introduce desafíos significativos para los modelos de regresión, que asumen una relación continua y homogénea entre las variables independientes y la dependiente.

En este contexto, se optó por un cambio de paradigma hacia modelos de clasificación, ya que estos permiten abordar directamente el problema del desbalance extremo de la variable dependiente. Los modelos de clasificación están diseñados para predecir categorías discretas, lo que es más adecuado para capturar la dicotomía inherente al fenómeno del gasto en salud: hogares que realizan algún gasto versus aquellos que no incurren en ningún gasto. Este enfoque permite no solo simplificar la modelación, sino también centrar el análisis en entender los factores determinantes del gasto en salud, lo cual tiene implicaciones prácticas relevantes para la formulación de políticas públicas.

Además, se decidió convertir la variable dependiente en una variable binaria en lugar de multiclase. Esta decisión se fundamenta en dos razones principales. En primer lugar, desde un punto de vista estadístico, la conversión a una variable binaria simplifica el problema y mejora la robustez de los modelos

frente al desbalance extremo de los datos. Al utilizar una variable multiclase, las categorías que representan niveles de gasto bajo y alto tendrían un número significativamente menor de observaciones, lo que podría dificultar el entrenamiento de los modelos y aumentar el riesgo de sobreajuste. En segundo lugar, desde una perspectiva práctica, una variable binaria facilita la interpretación de los resultados y permite enfocarse en el análisis de la probabilidad de gastar en salud, que es una cuestión crítica desde el punto de vista de las políticas públicas. Identificar los factores asociados a la probabilidad de incurrir en gasto en salud puede proporcionar información clave para diseñar estrategias de intervención dirigidas a mejorar el acceso y la equidad en los servicios de salud.

Por estas razones, la siguiente sección presenta el desarrollo de modelos de clasificación de Machine Learning utilizando la variable dependiente binaria, definida como 1 para los hogares que incurren en gasto en salud y 0 para aquellos que no lo hacen. Se detallan las metodologías utilizadas, la selección de características y los resultados obtenidos, junto con una discusión de sus implicaciones para la investigación y la formulación de políticas.

8.1.1. Antecedentes del análisis de gasto de bolsillo en salud como una variable dicotómica

La decisión de clasificar la variable dependiente, *Gasto de Bolsillo en Salud*, en forma dicotómica responde a dos factores clave: las características del problema de investigación y la evidencia empírica documentada en la literatura.

Características del Problema

El análisis exploratorio de los datos reveló que aproximadamente el 85 % de los hogares reportaron un gasto de bolsillo igual a cero, mientras que el 15 % restante incurrió en gastos positivos. Esta distribución altamente desbalanceada dificulta la implementación de modelos predictivos que utilicen la variable en su forma continua, ya que el predominio de valores nulos puede sesgar los resultados hacia la clase mayoritaria. La clasificación binaria permite abordar esta problemática al transformar la variable dependiente en una representación que distingue entre hogares que incurren o no en gastos, facilitando así la identificación de patrones asociados a esta ocurrencia.

Además, la dicotomización es consistente con el enfoque de muchos estudios previos en el ámbito de la salud pública y la economía de la salud. Por ejemplo, el gasto catastrófico en salud, definido como aquel que excede un umbral crítico del ingreso disponible, se modela frecuentemente como una variable binaria para analizar los determinantes y riesgos asociados [37, 41, 42].

Evidencia en la Literatura

Diversos estudios han optado por la clasificación binaria al abordar problemáticas similares. Por ejemplo, Jorrat et al. [38] analizaron la utilización de

servicios de salud y el gasto asociado en Argentina, empleando modelos logit para estudiar variables dependientes dicotómicas. En el contexto peruano, Hernández-Vásquez [37] y Lozada Urbano [42] definieron el gasto catastrófico en salud como una variable binaria para identificar hogares en riesgo de enfrentar gastos excesivos. De manera similar, en Colombia, Ospina et al. [41] utilizaron un modelo probit para analizar los determinantes del gasto catastrófico en salud, demostrando que la clasificación binaria es adecuada para explorar factores asociados a eventos de alta relevancia económica.

Finalmente, Rodríguez et al. [43] aplicaron regresiones logísticas para evaluar la incidencia de gastos de bolsillo excesivos en Paraguay, destacando la utilidad de este enfoque para modelar disparidades en el acceso y los costos asociados a servicios de salud. Estos estudios respaldan la adopción de una variable dependiente dicotómica, ya que facilita el análisis de fenómenos complejos en contextos donde los datos son desbalanceados o donde los resultados tienen implicaciones directas para la política pública.

Conveniencia del Enfoque Dicotómico

La clasificación binaria no solo mejora la interpretación y el análisis de los resultados, sino que también simplifica la implementación de técnicas de modelado. Al centrar el análisis en la ocurrencia de un evento (gasto o no gasto), se garantiza que las predicciones sean accionables, alineándose con el objetivo principal de identificar factores de riesgo y proponer intervenciones efectivas para mitigar los impactos del gasto de bolsillo en salud.

En resumen, la clasificación binaria de la variable dependiente se justifica tanto por las características del problema como por su amplia aceptación en la literatura, lo que asegura un análisis robusto y relevante desde una perspectiva empírica y práctica.

8.1.2. Justificación de la Clasificación como Método de Predicción

En el ámbito del aprendizaje automático, el término *predicción* se refiere al proceso mediante el cual un modelo infiere valores o categorías de una variable objetivo a partir de datos previamente observados. La literatura respalda que los modelos de clasificación también son una estrategia válida para la predicción, particularmente cuando se busca estimar la probabilidad de ocurrencia de un determinado evento. [18, 19, 20].

En términos formales, un problema de clasificación consiste en asignar a una observación x una clase y a partir de un modelo $f(x)$, que puede ser una función determinista o probabilística [21]. La predicción en este contexto implica estimar la probabilidad de que una observación pertenezca a una determinada categoría y tomar una decisión basada en un umbral óptimo [44]. Este principio es ampliamente utilizado en diversas áreas, incluyendo la salud, la economía y la detección de fraudes, donde el objetivo es predecir si un evento sucederá o no [19, 45].

En el presente estudio, el objetivo es **predecir el gasto de bolsillo en salud de los hogares colombianos**. Este objetivo se puede abordar desde dos perspectivas:

1. **Regresión**: Estimar el valor continuo del gasto de bolsillo.
2. **Clasificación**: Determinar si un hogar incurre en un gasto positivo o no.

Ambas estrategias cumplen con la noción de predicción, ya que en ambos casos se infiere una propiedad desconocida de la observación a partir de datos históricos. La literatura en economía de la salud respalda este enfoque, ya que en múltiples estudios se ha utilizado la clasificación para predecir el acceso a servicios de salud, la ocurrencia de gastos catastróficos y la probabilidad de hospitalización [36, 37, 23].

En particular, la variable dependiente del estudio presenta una gran cantidad de valores en cero, lo que dificulta la estimación precisa de un modelo de regresión convencional. Para abordar esta situación, se ha optado por transformar la variable en una representación dicotómica, lo que permite aplicar modelos de clasificación para predecir si un hogar tendrá o no gasto en salud. Este enfoque está alineado con estudios previos que han utilizado técnicas similares para analizar determinantes del gasto sanitario [3, 38].

Por lo tanto, se concluye que **hacer clasificación es una forma válida de predicción**, y que los objetivos del estudio se cumplen al utilizar modelos de clasificación para anticipar la ocurrencia del gasto de bolsillo en salud. La metodología implementada no solo permite capturar mejor la naturaleza de los datos disponibles, sino que también ofrece una interpretación relevante para la toma de decisiones en políticas públicas y economía de la salud.

8.2. Sobremuestreo para Manejo del Desbalance de Clases

8.2.1. Contexto y Motivación

El desbalance de clases es un desafío frecuente en problemas de clasificación binaria, especialmente cuando una clase minoritaria tiene una representación significativamente menor en el conjunto de datos. En el presente trabajo, la variable objetivo, *GASTO_CATEGORICO*, muestra una proporción de 84,5 % para la clase mayoritaria (0) y solo 15,5 % para la clase minoritaria (1). Esta distribución desigual puede causar que los modelos de clasificación se sesguen hacia la clase mayoritaria, lo que reduce su capacidad para identificar correctamente la clase minoritaria, afectando métricas clave como el *F1-Score* y la *ROC AUC*.

Para abordar este problema, se implementó la técnica de sobremuestreo mediante el método *Synthetic Minority Oversampling Technique* (SMOTE). Esta técnica ha demostrado ser efectiva para mejorar el desempeño de los modelos al balancear las clases en problemas similares [46].

8.2.2. Descripción de SMOTE

SMOTE es una técnica de sobremuestreo que genera nuevas muestras sintéticas para la clase minoritaria. A diferencia del sobremuestreo tradicional, que simplemente replica instancias existentes, SMOTE crea nuevas instancias interpolando entre ejemplos cercanos de la clase minoritaria en el espacio de características. Esto ayuda a evitar problemas de sobreajuste que pueden surgir al duplicar datos de manera directa.

El procedimiento de SMOTE consiste en:

1. Seleccionar una instancia aleatoria de la clase minoritaria.
2. Identificar sus k vecinos más cercanos en el espacio de características.
3. Generar una nueva instancia interpolando linealmente entre la instancia seleccionada y uno de sus vecinos.

Este enfoque garantiza que las nuevas instancias mantengan la distribución de las características de la clase minoritaria, creando un conjunto de datos más balanceado.

8.2.3. Aplicación en el Presente Trabajo

En este estudio, se aplicó SMOTE para balancear el conjunto de entrenamiento antes de estimar cada modelo de clasificación. El procedimiento fue el siguiente:

1. Dividir el conjunto de datos en características (X) y la variable objetivo (y).
2. Aplicar SMOTE al conjunto de entrenamiento para generar instancias sintéticas de la clase minoritaria.
3. Utilizar el conjunto balanceado generado por SMOTE para entrenar los modelos de clasificación.

8.2.4. Conveniencia del Sobremuestreo

El uso de SMOTE resulta particularmente conveniente en el contexto del problema abordado debido a las siguientes razones:

- **Mitigación del Desbalance:** Al balancear las clases, los modelos tienen una mejor oportunidad de aprender patrones significativos de ambas clases, en lugar de centrarse únicamente en la clase mayoritaria.
- **Mejor Desempeño en Métricas Clave:** Técnicas como SMOTE suelen mejorar métricas que dependen del correcto manejo de la clase minoritaria, como el *F1-Score* y la *ROC AUC*.
- **Evitar Sobreajuste:** La generación de instancias sintéticas evita problemas de sobreajuste asociados con el duplicado de instancias originales.

8.2.5. Impacto en los Modelos Estimados

Para todos los modelos presentados en este capítulo, se aplicó SMOTE como parte del preprocesamiento. Esto permitió una evaluación más justa del desempeño de los modelos en un escenario balanceado, asegurando que tanto la clase mayoritaria como la minoritaria fueran representadas de manera adecuada.

8.3. Modelo de Clasificación: Gradient Boosting

8.3.1. Contexto y Motivación

El modelo *Gradient Boosting Classifier* se seleccionó por su alta eficacia en la clasificación binaria de problemas complejos y su capacidad para manejar relaciones no lineales. Este enfoque fue particularmente relevante en el análisis del gasto de bolsillo en salud debido a la naturaleza desequilibrada de los datos y la necesidad de ajustar múltiples hiperparámetros para maximizar el desempeño predictivo.

8.3.2. Configuración del Modelo y Búsqueda de Hiperparámetros

Se realizó una búsqueda de hiperparámetros utilizando *GridSearchCV*, evaluando 540 configuraciones mediante validación cruzada con cinco particiones (*5-fold cross-validation*). Los mejores hiperparámetros identificados fueron:

- `learning_rate`: 0.2
- `max_depth`: 5
- `min_samples_leaf`: 5
- `min_samples_split`: 2
- `n_estimators`: 200

8.3.3. Resultados del Modelo

El modelo optimizado fue evaluado en un conjunto de validación independiente. Los resultados principales se resumen en el Cuadro 8.1.

Cuadro 8.1: Evaluación del Modelo Optimizado

Clase	**Precisión**	**Recall**	**F1-Score**	**Soporte**
Sin gasto	0.74	0.72	0.73	13,350
Con gasto	0.73	0.75	0.74	13,349
Exactitud (Accuracy)	0.74 (26,699 observaciones)			
Promedio Macro	0.74	0.74	0.74	
Promedio Ponderado	0.74	0.74	0.74	

Matriz de Confusión

La matriz de confusión presentada en el Cuadro 8.2 muestra la distribución de las predicciones en términos absolutos y porcentuales, destacando falsos positivos y negativos.

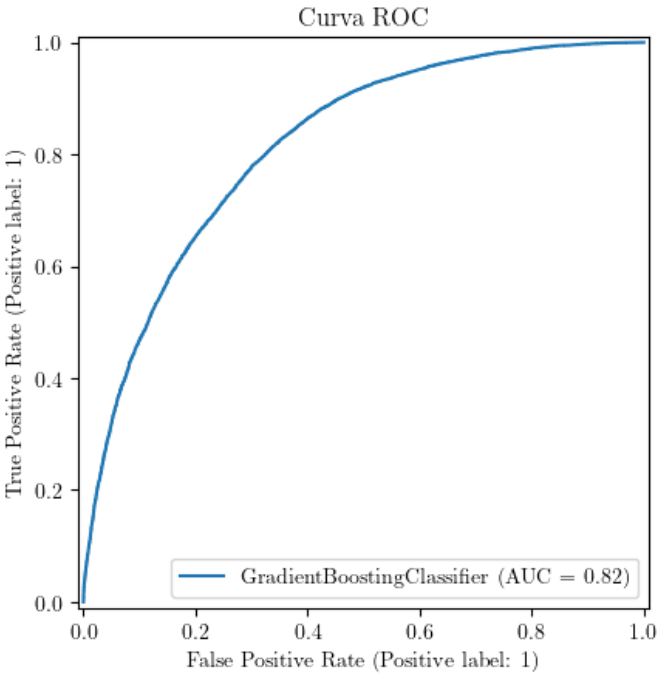
Cuadro 8.2: Matriz de Confusión del Modelo Optimizado (Valores Absolutos y Porcentuales)

	Clase 0 (Predicho)	**Clase 1 (Predicho)**
Clase 0 (Real)	9,635 (72.2 %)	3,715 (27.8 %)
Clase 1 (Real)	3,324 (24.9 %)	10,025 (75.1 %)

Curva ROC y AUC

El modelo alcanzó un área bajo la curva (AUC) de 0.82, lo que indica una buena capacidad discriminativa entre las clases. La curva ROC se muestra en la Figura 8.1.

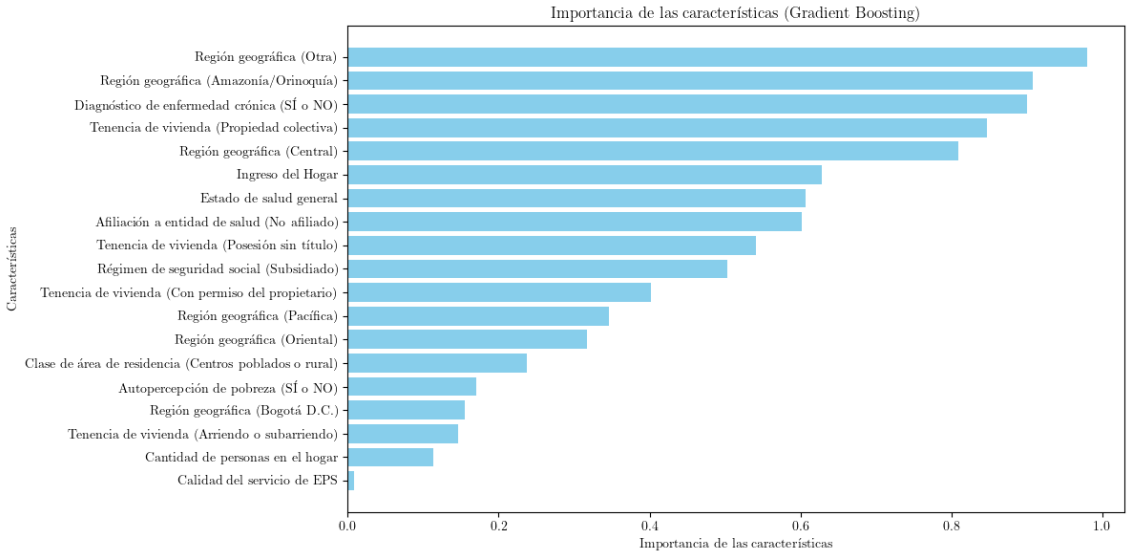
Figura 8.1: Curva ROC del Modelo Gradient Boosting



Importancia de las Variables

La Figura 8.2 ilustra las principales variables que contribuyen a la predicción, siendo `I_HOGAR_ANUAL`, `P1930_2` y `P6127` las más relevantes.

Figura 8.2: Importancia de las Características según el Modelo Gradient Boosting



8.3.4. Interpretación de los Resultados

El modelo Gradient Boosting demostró un desempeño equilibrado en la clasificación binaria. Sin embargo, las tasas de falsos positivos (27.8 %) y falsos negativos (24.9 %) reflejan la necesidad de optimizar aún más las estrategias de clasificación, especialmente para las clases minoritarias.

El análisis de importancia de variables resaltó la relación entre el ingreso anual del hogar (`I_HOGAR_ANUAL`) y el gasto en salud, alineándose con estudios previos. Además, factores socioeconómicos y de salud, como `P1930_2` y `P6127`, desempeñaron un papel significativo en las predicciones.

8.3.5. Conclusión

El modelo Gradient Boosting es una herramienta valiosa para predecir el gasto de bolsillo en salud, aunque sus limitaciones sugieren explorar métodos más avanzados o ajustes en la estrategia de resampling. Esto incluye el uso de métricas personalizadas durante el entrenamiento para priorizar la clase de interés.

8.4. Modelo de Clasificación: Árbol de Decisión

En esta sección se presenta el desarrollo y los resultados obtenidos al implementar un modelo de clasificación basado en árboles de decisión. Este modelo destaca por su simplicidad interpretativa y su capacidad para manejar relaciones no lineales entre variables, lo que lo convierte en una herramienta útil para analizar el gasto de bolsillo en salud.

8.4.1. Metodología

El modelo se ajustó utilizando el conjunto de datos procesado, donde la variable dependiente binaria `GASTO_CATEGORICO` fue el objetivo. Para optimizar los hiperparámetros, se utilizó una búsqueda en cuadrícula (*GridSearchCV*) con validación cruzada de 5 pliegues, maximizando el desempeño en términos de *ROC AUC*. Los hiperparámetros evaluados fueron:

- `max_depth`: Profundidad máxima del árbol.
- `min_samples_split`: Mínimo de muestras necesarias para dividir un nodo.
- `min_samples_leaf`: Mínimo de muestras necesarias en un nodo hoja.

8.4.2. Resultados

Los hiperparámetros óptimos encontrados se presentan en el Cuadro 8.3. Estos parámetros se utilizaron para evaluar el modelo en un conjunto de prueba independiente.

Cuadro 8.3: Hiperparámetros Óptimos para el Modelo de Árboles de Decisión

Hiperparámetro	**Valor Óptimo**
`max_depth`	5
`min_samples_split`	5
`min_samples_leaf`	2

La evaluación del modelo optimizado se presenta en el Cuadro 8.4. El modelo alcanzó una precisión global del 67 %, con un *ROC AUC* de 0.71.

Cuadro 8.4: Evaluación del Modelo de Árboles de Decisión

Clase	**Precisión**	**Recall**	**F1-Score**	**Soporte**
Sin gasto	0.66	0.70	0.68	13,350
Con gasto	0.68	0.64	0.66	13,349
Exactitud (Accuracy)	0.67 (26,699 observaciones)			
Promedio Macro	0.67	0.67	0.67	
Promedio Ponderado	0.67	0.67	0.67	

Matriz de Confusión

La matriz de confusión del modelo optimizado, que incluye valores absolutos y porcentuales, se presenta en el Cuadro 8.5.

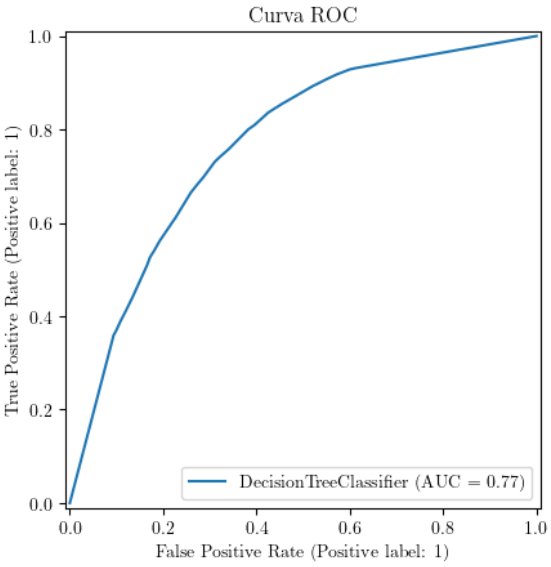
Cuadro 8.5: Matriz de Confusión del Modelo Optimizado

	Clase 0 (Predicho)	**Clase 1 (Predicho)**
Clase 0 (Real)	9,298 (69.6 %)	4,052 (30.4 %)
Clase 1 (Real)	4,808 (36.0 %)	8,541 (64.0 %)

Curva ROC y Análisis de Importancia de Características

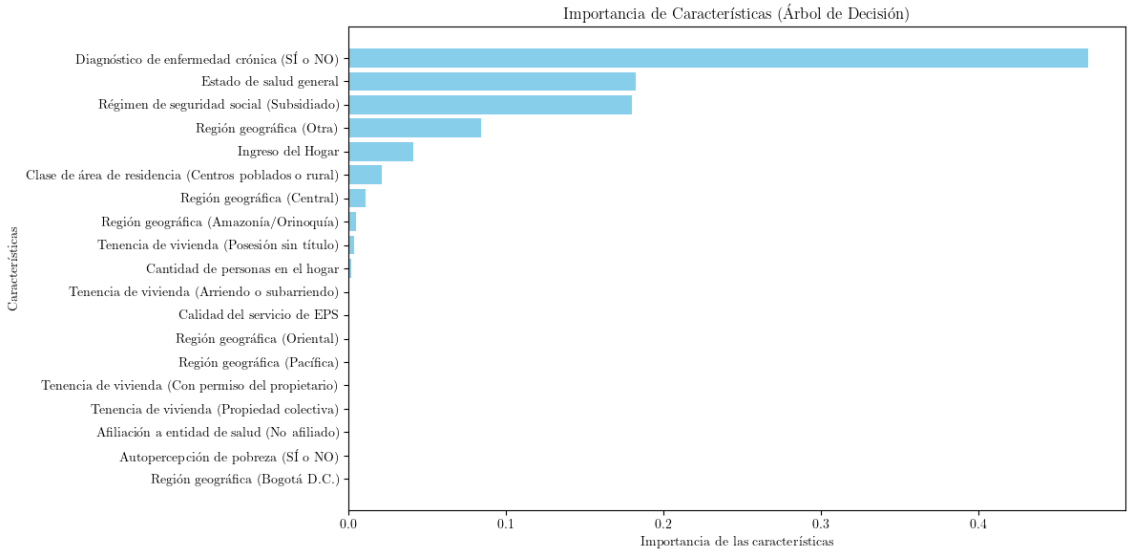
El modelo mostró un *ROC AUC* de 0.71, indicando una capacidad razonable para diferenciar entre las clases. La Figura 8.3 muestra la curva *ROC* obtenida.

Figura 8.3: Curva ROC para el Modelo de Árboles de Decisión



En cuanto a la importancia de las características, la Figura 8.4 destaca las variables más influyentes en el modelo. Las variables `I_HOGAR_ANUAL` (ingreso anual del hogar) y `P1930_2` (tenencia de vivienda) fueron las más relevantes, lo que es consistente con hallazgos previos en la literatura.

Figura 8.4: Importancia de Características en el Modelo de Árboles de Decisión



8.4.3. Interpretación de los Resultados

El modelo de árboles de decisión ofreció un desempeño aceptable, con métricas de clasificación equilibradas entre ambas clases. La relación entre el ingreso del hogar y el gasto de bolsillo en salud, reflejada en la importancia de las variables, refuerza la comprensión de factores determinantes en este contexto.

A pesar de su simplicidad interpretativa, el modelo mostró limitaciones en la captura de relaciones más complejas entre las variables, lo que podría mejorarse con modelos más avanzados o ajustes adicionales en los hiperparámetros.

8.4.4. Conclusión

El modelo de árboles de decisión proporciona una base sólida para explorar la clasificación del gasto de bolsillo en salud, aunque su desempeño puede mejorarse mediante enfoques más sofisticados o el uso de estrategias que capturen mejor las interacciones no lineales.

8.5. Modelo de Clasificación: Random Forest

8.5.1. Contexto y Motivación

El modelo *Random Forest* fue seleccionado debido a su capacidad para manejar grandes conjuntos de datos con características heterogéneas y su robustez frente al sobreajuste, al combinar múltiples árboles de decisión mediante el método de *bagging*. Este enfoque resulta particularmente útil en problemas con clases desbalanceadas, como la clasificación del gasto de bolsillo en salud.

8.5.2. Metodología

Para ajustar los hiperparámetros del modelo, se utilizó una búsqueda en cuadrícula (*GridSearchCV*) con validación cruzada de 5 pliegues, maximizando el desempeño en términos de *ROC AUC*. Los hiperparámetros evaluados fueron:

- `n_estimators`: Número de estimadores.
- `max_depth`: Profundidad máxima del árbol.
- `min_samples_leaf`: Número mínimo de muestras en un nodo hoja.
- `min_samples_split`: Número mínimo de muestras necesarias para dividir un nodo.

Los mejores hiperparámetros obtenidos se presentan en el Cuadro 8.6.

Cuadro 8.6: Hiperparámetros Óptimos del Modelo Random Forest

Hiperparámetro	**Valor Óptimo**
`n_estimators`	300
`max_depth`	15
`min_samples_leaf`	5
`min_samples_split`	10

8.5.3. Resultados del Modelo

El modelo optimizado fue evaluado en un conjunto de prueba independiente. Las métricas de evaluación incluyen precisión, *recall*, F_1-score y soporte, como se detalla en el Cuadro 8.7.

Cuadro 8.7: Evaluación del Modelo Random Forest

Clase	**Precisión**	**Recall**	**F1-Score**	**Soporte**
Sin gasto	0.77	0.76	0.76	13,350
Con gasto	0.76	0.78	0.77	13,349
Exactitud (Accuracy)	0.77 (26,699 observaciones)			
Promedio Macro	0.77	0.77	0.77	
Promedio Ponderado	0.77	0.77	0.77	

Matriz de Confusión

La matriz de confusión se presenta en el Cuadro 8.8, mostrando un buen equilibrio en la clasificación de ambas clases.

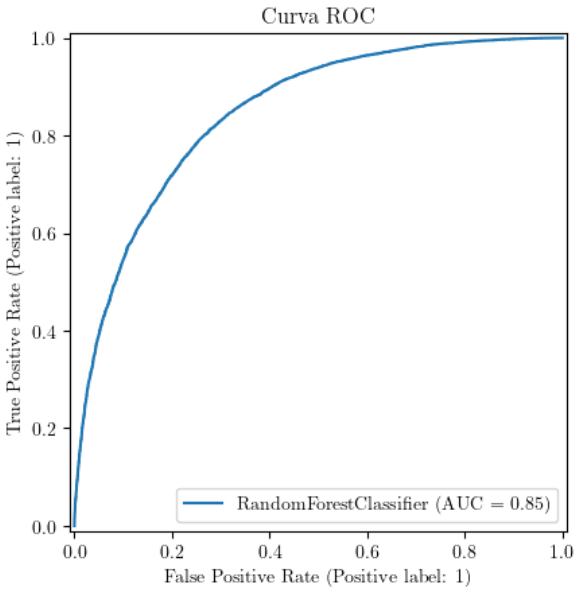
Cuadro 8.8: Matriz de Confusión del Modelo Random Forest

	Clase 0 (Predicho)	**Clase 1 (Predicho)**
Clase 0 (Real)	10,095 (75.6 %)	3,255 (24.4 %)
Clase 1 (Real)	2,985 (22.4 %)	10,364 (77.6 %)

Curva ROC y Análisis de Importancia de Características

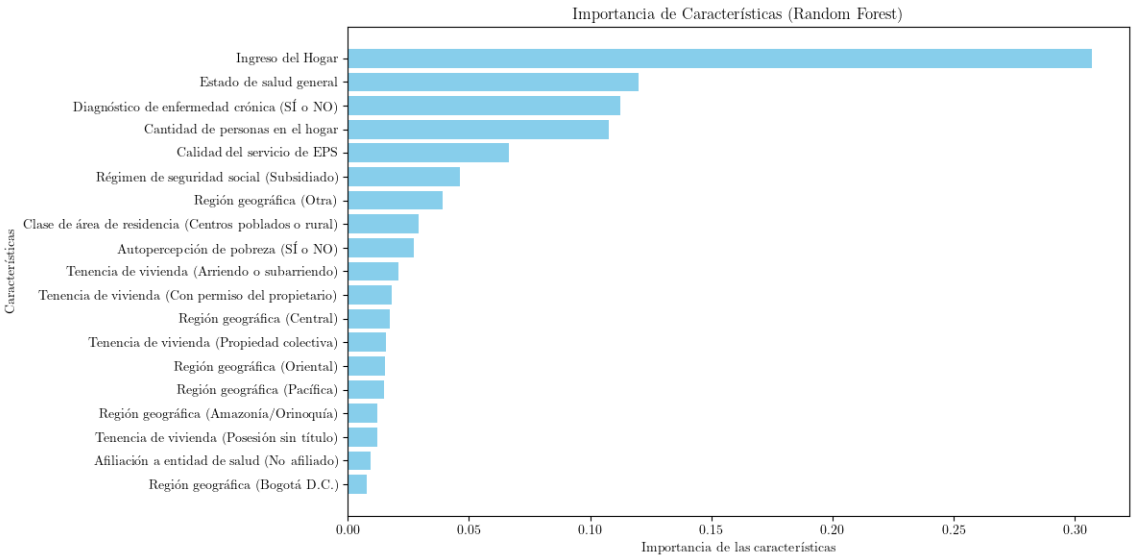
El modelo alcanzó un área bajo la curva ($ROC\ AUC$) de 0.85, lo que indica una buena capacidad discriminativa entre las clases. La Figura 8.5 presenta la curva ROC correspondiente.

Figura 8.5: Curva ROC para el Modelo Random Forest



La importancia de las características en el modelo se ilustra en la Figura 8.6. Las variables `I_HOGAR_ANUAL` (ingreso anual del hogar), `P6127` (estado general de salud) y `CANT_PERSONAS_HOGAR` (cantidad de personas en el hogar) fueron las más influyentes en las predicciones.

Figura 8.6: Importancia de las Características en el Modelo Random Forest



8.5.4. Interpretación de los Resultados

El modelo *Random Forest* mostró un desempeño sólido, con métricas equilibradas entre ambas clases y un área bajo la curva (*ROC AUC*) destacable. La importancia de las características resalta la influencia de factores económicos y sociales en el gasto de bolsillo en salud.

Sin embargo, se observan desafíos relacionados con la predicción de la clase minoritaria. Estrategias como el balanceo de datos o enfoques híbridos podrían mejorar aún más el desempeño.

8.5.5. Conclusión

El modelo *Random Forest* es una herramienta robusta para la clasificación del gasto de bolsillo en salud. Su capacidad para manejar datos heterogéneos y evitar el sobreajuste lo posiciona como una de las opciones más eficaces dentro de esta investigación.

8.6. Modelo de Clasificación: Regresión Logística

8.6.1. Contexto y Motivación

La regresión logística es un modelo estadístico ampliamente utilizado para problemas de clasificación binaria. Su simplicidad, interpretabilidad y eficiencia computacional lo convierten en una herramienta ideal para establecer líneas base en problemas de clasificación. En este estudio, se implementó un modelo de regresión logística para predecir el gasto de bolsillo en salud de los hogares, aplicando técnicas de sobremuestreo para balancear las clases y mejorar el desempeño en la clase minoritaria.

8.6.2. Metodología

El modelo fue optimizado mediante una búsqueda de hiperparámetros utilizando *GridSearchCV*, evaluando configuraciones clave para la regularización y el optimizador. Los hiperparámetros evaluados fueron:

- `penalty`: Regularización l_2 (Ridge).
- `C`: Inverso de la fuerza de regularización, configurado en 1.
- `solver`: Optimizador `saga`, adecuado para conjuntos de datos grandes y regularización l_1 o l_2.

El mejor conjunto de hiperparámetros encontrados se presenta en el Cuadro 8.9.

Cuadro 8.9: Hiperparámetros Óptimos del Modelo de Regresión Logística

Hiperparámetro	**Valor Óptimo**
`penalty`	l_2
`C`	1
`solver`	`saga`

8.6.3. Resultados del Modelo

El modelo optimizado se evaluó en un conjunto de prueba independiente, y sus métricas principales se presentan en el Cuadro 8.10. Las métricas incluyen precisión, *recall*, F_1-score y el área bajo la curva ROC (*ROC AUC*).

Cuadro 8.10: Evaluación del Modelo de Regresión Logística

Clase	**Precisión**	**Recall**	**F1-Score**	**Soporte**
Sin gasto	0.66	0.69	0.68	9,200
Con gasto	0.68	0.65	0.66	8,694
Exactitud (Accuracy)	0.67 (26,699 observaciones)			
Promedio Macro	0.67	0.67	0.67	
Promedio Ponderado	0.67	0.67	0.67	

Matriz de Confusión

La matriz de confusión, que resume los aciertos y errores del modelo, se presenta en el Cuadro 8.11, con valores absolutos y porcentuales.

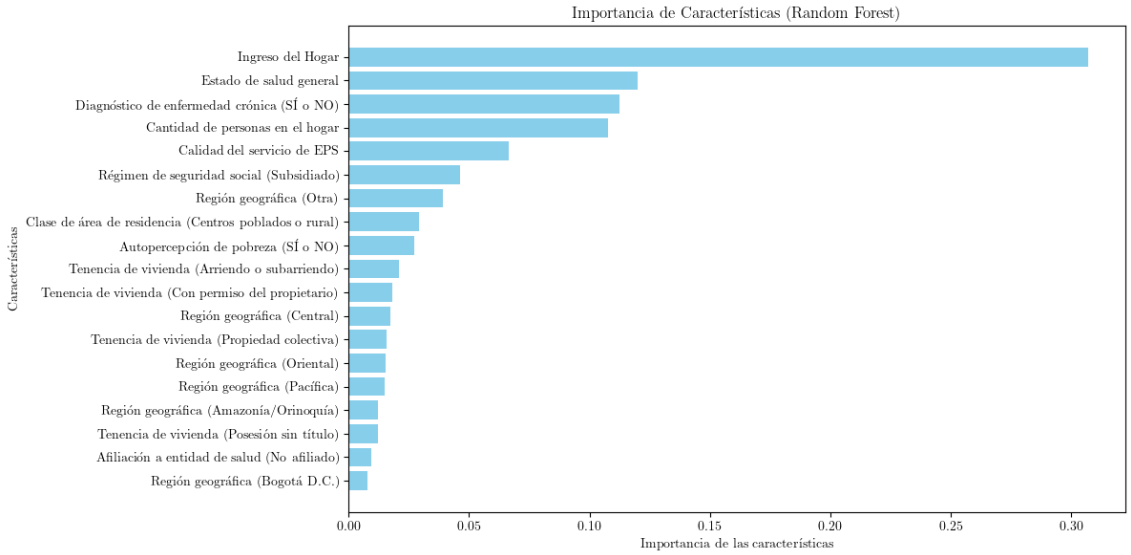
Cuadro 8.11: Matriz de Confusión del Modelo de Regresión Logística

	Clase 0 (Predicho)	**Clase 1 (Predicho)**
Clase 0 (Real)	9,200 (68.9 %)	4,150 (31.1 %)
Clase 1 (Real)	4,655 (34.9 %)	8,694 (65.1 %)

Curva ROC y Análisis de Importancia de Características

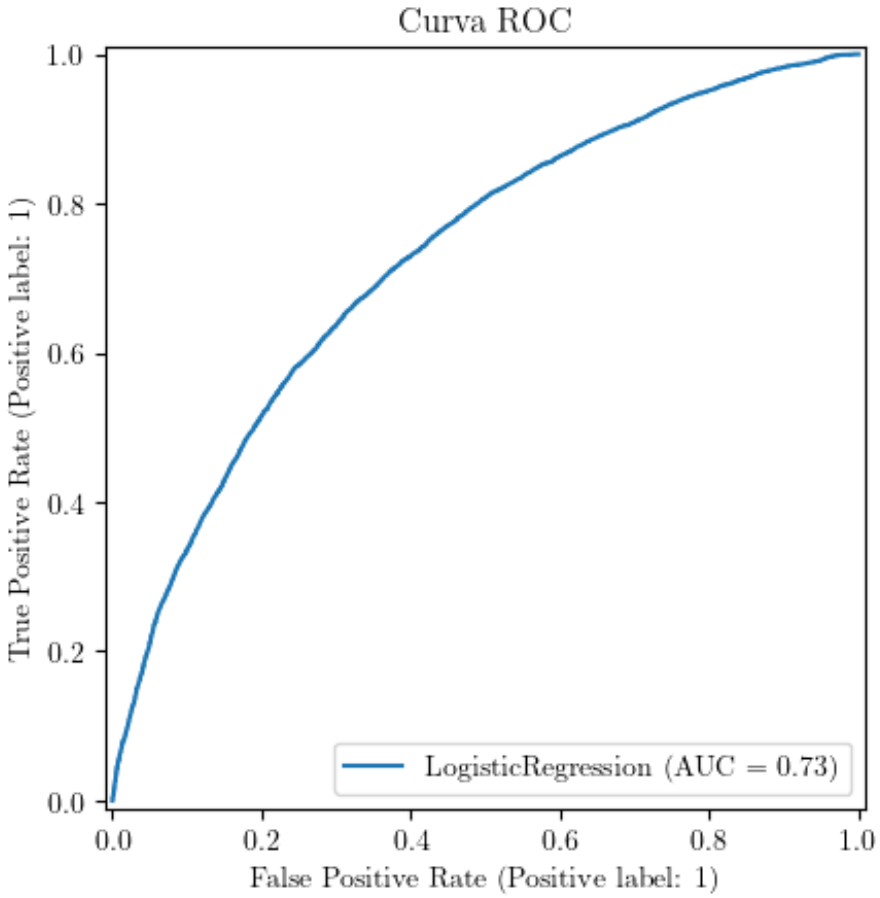
El modelo alcanzó un área bajo la curva (*ROC AUC*) de 0.73, lo que indica un desempeño aceptable en la discriminación de clases. La Figura 8.7 muestra la curva *ROC* correspondiente.

Figura 8.7: Curva ROC del Modelo de Regresión Logística



El análisis de los coeficientes del modelo permitió identificar las características más relevantes, destacando `REGION_9`, `REGION_6` y `P1930_2`. Estas variables mostraron un impacto significativo en la probabilidad de que un hogar incurra en un gasto de bolsillo en salud. La Figura 8.8 ilustra la importancia de las características.

Figura 8.8: Importancia de las Características en el Modelo de Regresión Logística



8.6.4. Interpretación de los Resultados

El modelo de regresión logística mostró un desempeño moderado, con métricas equilibradas entre ambas clases. Aunque su capacidad predictiva es inferior a la de modelos más complejos como *Random Forest*, su simplicidad e interpretabilidad lo convierten en una herramienta útil, especialmente en contextos donde la transparencia del modelo es prioritaria.

8.6.5. Conclusión

La regresión logística sigue siendo una herramienta valiosa para problemas de clasificación binaria, particularmente como línea base en estudios más complejos. Aunque su desempeño en términos de precisión y *ROC AUC* es limitado en comparación con otros modelos, su capacidad para identificar relaciones directas entre variables y su facilidad de implementación la hacen ideal para análisis exploratorios y aplicaciones donde la interpretabilidad es clave.

8.7. Discusión y Comparación de Modelos de Clasificación

La validación cruzada permitió evaluar el desempeño de los modelos de clasificación en términos de dos métricas clave: *ROC AUC*, que mide la capacidad del modelo para distinguir entre clases, y *F1-Score*, que captura el equilibrio entre precisión y sensibilidad. En esta sección, se discuten los resultados obtenidos y se realiza una comparación detallada entre los modelos estimados.

8.7.1. Resultados Generales

EL Cuadro 8.12 presenta un resumen comparativo de los modelos evaluados. Se destacan las métricas *ROC AUC* y *F1-Score*, incluyendo sus medias y desviaciones estándar obtenidas durante la validación cruzada. Cabe destacar que, para abordar el desbalance significativo en la distribución de la variable dependiente, se aplicó la técnica de sobremuestreo SMOTE (*Synthetic Minority Oversampling Technique*) a los datos de entrenamiento de todos los modelos.

Cuadro 8.12: Resumen Comparativo de los Modelos de Clasificación

Modelo	**Métrica**	**Media**	**Desviación Estándar**
Gradient Boosting	*ROC AUC*	0.7210	0.0063
	F1-Score	0.1388	0.0100
Árbol de Decisión	*ROC AUC*	0.7016	0.0078
	F1-Score	0.1282	0.0200
Random Forest	*ROC AUC*	0.7299	0.0058
	F1-Score	0.0543	0.0019
Regresión Logística	*ROC AUC*	0.7189	0.0062
	F1-Score	0.1050	0.0077

Además del desempeño en métricas predictivas, se evaluaron los tiempos de entrenamiento y predicción de cada modelo en el entorno de Google Colab. Los resultados se presentan en el Cuadro 8.13.

Cuadro 8.13: Tiempos de Cómputo de los Modelos de Clasificación

Modelo	**Tiempo de Entre.**	**Tiempo de Predic.**
Gradient Boosting	$\sim$ 165 minutos	$\sim$ 0,005 segundos
Árbol de Decisión	$\sim$ 6 minutos	$\sim$ 0,0005 segundos
Random Forest	$\sim$ 87 minutos	$\sim$ 0,007 segundos
Regresión Logística	$\sim$ 4 minutos	$\sim$ 0,001 segundos

Los resultados muestran que los modelos de ensamblaje, como *Gradient Boosting* y *Random Forest*, requieren tiempos de entrenamiento significativamente

mayores en comparación con modelos más simples como *Árbol de Decisión* y *Regresión Logística*. Sin embargo, el tiempo de predicción por observación sigue siendo bajo en todos los modelos, lo que sugiere que pueden ser aplicados en entornos productivos sin grandes costos computacionales.

Todos los experimentos se realizaron en el entorno de Google Colab, aprovechando los recursos computacionales disponibles en la nube. El detalle del hardware utilizado se presenta a continuación:

Entorno de Ejecución en Google Colab

- **Procesador:** Intel(R) Xeon(R) CPU @ 2.20GHz (2 núcleos virtuales)
- **Arquitectura del CPU:** x86_64, compatible con 32-bit y 64-bit
- **Memoria RAM:** 12 GB (aproximadamente 11 GB disponibles)
- **Sistema Operativo:** Ubuntu 22.04.4 LTS (Jammy Jellyfish)
- **Almacenamiento:** 108 GB de disco, con 79 GB disponibles
- **Virtualización:** Basado en KVM
- **GPU:** No usado

Nota: Todos los cálculos se realizaron utilizando la CPU del entorno de Google Colab.

La diferencia en los tiempos de entrenamiento y predicción resalta la importancia de considerar la eficiencia computacional al seleccionar un modelo. Si bien *Gradient Boosting* y *Random Forest* ofrecen mejor desempeño predictivo, su costo computacional es considerablemente mayor. En aplicaciones donde la interpretabilidad y la velocidad son más relevantes, modelos como *Árbol de Decisión* o *Regresión Logística* pueden ser alternativas viables, especialmente en entornos con recursos limitados.

8.7.2. Análisis de Resultados

Gradient Boosting

El modelo de *Gradient Boosting* destacó como uno de los mejores en términos de discriminación entre clases, con un *ROC AUC* de $0{,}7210 \pm 0{,}0063$. Aunque su *F1-Score* ($0{,}1388 \pm 0{,}0100$) sigue siendo bajo, la aplicación de SMOTE mejoró la identificación de la clase minoritaria en comparación con datos no balanceados, evidenciando la utilidad de esta técnica en este tipo de problemas.

Árbol de Decisión

El Árbol de Decisión mostró un desempeño moderado, con un *ROC AUC* de $0{,}7016 \pm 0{,}0078$ y un *F1-Score* de $0{,}1282 \pm 0{,}0200$. Si bien su simplicidad lo hace interpretable, la técnica de SMOTE ayudó a mejorar ligeramente su capacidad para identificar casos de la clase minoritaria, aunque sigue siendo limitada en su habilidad para capturar relaciones complejas.

Random Forest

El modelo *Random Forest* obtuvo el mejor *ROC AUC* (0,7299 ± 0,0058), lo que refleja su alta capacidad para distinguir entre clases. Sin embargo, su *F1-Score* (0,0543±0,0019) permaneció bajo, incluso con la integración de SMOTE, sugiriendo que este modelo tiende a priorizar la clase mayoritaria, lo cual afecta negativamente la identificación precisa de la clase minoritaria.

Regresión Logística

La Regresión Logística mostró un desempeño competitivo, con un *ROC AUC* de 0,7189 ± 0,0062 y un *F1-Score* de 0,1050 ± 0,0077. Aunque no supera a los modelos más complejos, el uso de SMOTE permitió equilibrar parcialmente su capacidad predictiva y su simplicidad e interpretabilidad siguen siendo ventajas importantes en contextos donde estas características son prioritarias.

8.7.3. Discusión Comparativa

El uso de SMOTE fue fundamental para abordar el problema del desbalance de clases en todos los modelos. Los resultados indican que los modelos más avanzados, como *Random Forest* y *Gradient Boosting*, son los más adecuados para tareas de clasificación binaria cuando se prioriza la capacidad discriminativa (*ROC AUC*). No obstante, los valores relativamente bajos de *F1-Score* sugieren la necesidad de ajustes adicionales para mejorar la identificación de la clase minoritaria.

Por otro lado, los modelos más simples, como Árbol de Decisión y Regresión Logística, ofrecen una base interpretativa sólida, especialmente útil en contextos donde se requiere explicar las predicciones de manera clara. Sin embargo, su desempeño en métricas globales es inferior, lo que los hace menos adecuados si se prioriza la precisión predictiva.

8.7.4. Implicaciones para la Selección del Modelo

La incorporación de SMOTE permitió equilibrar las clases en los datos de entrenamiento, mejorando el desempeño general en todos los modelos. Esto evidencia la importancia de abordar el desbalance de clases en problemas como este. Para aplicaciones donde la interpretabilidad es crítica, la Regresión Logística sigue siendo una opción viable. En cambio, para tareas donde la capacidad predictiva es prioritaria, modelos como *Random Forest* y *Gradient Boosting* son recomendables, pero con ajustes adicionales para optimizar la sensibilidad hacia la clase minoritaria.

Capítulo 9

Conclusiones

La presente tesis abordó la problemática del gasto de bolsillo en salud de los hogares colombianos, utilizando técnicas de aprendizaje automático para desarrollar modelos de predicción que permitan comprender y mitigar los factores asociados a este fenómeno. A continuación, se presentan las principales conclusiones del trabajo realizado:

El análisis exploratorio inicial reveló una distribución altamente asimétrica en la variable objetivo, con un 85 % de observaciones con valores nulos en el gasto de bolsillo. Este hallazgo motivó la transformación de la variable dependiente en una categórica binaria, representando la ocurrencia o ausencia de gasto de bolsillo. Esta decisión, fundamentada en la literatura [38, 37, 41], permitió mejorar la estabilidad de los modelos y abordar el desbalance inherente en los datos.

En este sentido, la aplicación de la técnica SMOTE fue crucial para balancear las clases y garantizar que los modelos aprendieran de manera efectiva las características de la clase minoritaria. Este enfoque resultó particularmente útil en un problema con una marcada asimetría de clases, al incrementar la sensibilidad y mejorar la capacidad predictiva de los modelos, especialmente en términos de *F1-Score* y *ROC AUC*.

En cuanto al desempeño comparativo de los modelos, se desarrollaron y evaluaron cuatro modelos de clasificación: Gradient Boosting, Random Forest, Árbol de Decisión y Regresión Logística. Los resultados de la validación cruzada mostraron que:

- El modelo *Random Forest* obtuvo el mejor desempeño global en términos de *ROC AUC* ($0{,}7299 \pm 0{,}0058$), indicando una alta capacidad para distinguir entre clases.
- *Gradient Boosting* también mostró un buen desempeño ($ROCAUC = 0{,}7210 \pm 0{,}0063$), siendo una alternativa robusta y flexible para problemas de clasificación en datos desbalanceados.
- Aunque los modelos de *Árbol de Decisión* y *Regresión Logística* presentaron resultados menos destacados, su simplicidad y facilidad de interpretación los convierten en herramientas útiles para contextos específicos.

La comparación entre modelos evidenció que las técnicas avanzadas de ensamblaje, como Gradient Boosting y Random Forest, son preferibles en problemas complejos como la predicción del gasto de bolsillo, siempre que se acompañen de estrategias adecuadas de balanceo de clases.

Por su parte, el análisis de importancia de características destacó al **diagnóstico de enfermedades crónicas** como la variable más relevante en todos los modelos evaluados. Otras variables clave incluyeron el **ingreso del hogar, estado de salud y región geográfica**, lo que resalta la influencia de la heterogeneidad socioeconómica en el gasto en salud. Estos resultados son consistentes con estudios previos [37, 41] y subrayan la necesidad de considerar la heterogeneidad socioeconómica en el diseño de políticas públicas.

A la luz de los objetivos propuestos, este trabajo logró determinar las variables críticas del GBS, identificar los modelos de aprendizaje automático más adecuados, desarrollar un protocolo de entrenamiento y validación y, por último, establecer un conjunto de métricas para evaluar la precisión y rendimiento. La identificación de estas variables clave y la selección de modelos óptimos permiten fortalecer el análisis del gasto de bolsillo en salud y ofrecer herramientas más precisas para la toma de decisiones en políticas públicas y planificación del sistema de salud.

Asimismo, la implementación de modelos de clasificación en este estudio cumple con el objetivo central de predecir el gasto de bolsillo en salud. Aunque la predicción tradicionalmente se asocia con modelos de regresión para estimar valores continuos, la clasificación permite predecir la ocurrencia del gasto de bolsillo, proporcionando información relevante para la identificación de grupos poblacionales vulnerables. Esta aproximación no solo mejora la interpretabilidad de los resultados, sino que también facilita el diseño de estrategias de intervención dirigidas a reducir la carga financiera de los hogares y mejorar la equidad en el acceso a los servicios de salud.

Este trabajo contribuye al campo de la economía de la salud y el aprendizaje automático, demostrando la viabilidad de aplicar técnicas avanzadas de clasificación en el análisis de fenómenos socioeconómicos complejos. Sin embargo, aún quedan áreas por explorar. Se sugieren las siguientes líneas de investigación futura:

- Incorporar modelos más avanzados, como redes neuronales profundas y técnicas basadas en *transformers*, para evaluar si mejoran el desempeño predictivo en comparación con los modelos tradicionales.
- Hacer un análisis comparativos de la diferentes técnicas de sobremuestreo.
- Estudiar la integración de métricas personalizadas que penalicen de manera diferenciada los errores en la predicción de la clase minoritaria.

En conclusión, esta tesis resalta el potencial del aprendizaje automático para abordar problemas complejos en el sistema de salud, como el gasto de bolsillo en hogares colombianos. Los resultados obtenidos no solo aportan evidencia empírica para la toma de decisiones, sino que también abren nuevas oportunidades para el desarrollo de herramientas predictivas y políticas públicas que promuevan la equidad y la sostenibilidad en el acceso a los servicios de salud.

Bibliografía

[1] M. Santa María, F. García, S. Rozo, and M. J. Uribe, *Un diagnóstico general del sector salud en Colombia: evolución, contexto y principales retos de un sistema en transformación*, 2011. [Online]. Available: https://www.repository.fedesarrollo.org.co/bitstream/handle/11445/65/LIB_2011_Efectos%20de%20la%20ley%20100%20en%20salud_Completo.pdf?sequence=2&isAllowed=y

[2] L. Alba-Meocera *et al.*, "Aspectos financieros y fiscales del sistema de salud en colombia," *Ensayos sobre Política Económica*, no. 106, October 2023. [Online]. Available: https://doi.org/10.32468/espe106

[3] N. Maldonado, V. Soto, and R. Guerrero, *Gasto de Bolsillo en Salud en Colombia.* PROESA - Centro de Estudios en Protección Social y Economía de la Salud, Universidad Icesi, August 2022. [Online]. Available: https://www.icesi.edu.co/proesa/images/mesa/mesa-gasto-de-bolsillo-en-colombia.pdf

[4] "Colombia's universal health insurance system," *Health Affairs*, vol. 28, no. 3, pp. 853–863, 2009. [Online]. Available: https://openknowledge.worldbank.org/handle/10986/28634

[5] O. para la Cooperación y el Desarrollo Económicos (OCDE), *Health at a Glance 2023*, 2023. [Online]. Available: https://www.oecd.org/health/health-at-a-glance/

[6] "Inequality in healthcare use among older people in colombia," *International Journal for Equity in Health*, 2020. [Online]. Available: https://equityhealthj.biomedcentral.com/articles/10.1186/s12939-020-01262-5

[7] J. Xu *et al.*, "Household catastrophic health expenditure: A global analysis," *The Lancet Global Health*, vol. 3, no. 4, pp. e234–e245, 2015.

[8] E. K. Arah *et al.*, "Understanding the role of health systems in population health and its determinants," *Annual Review of Public Health*, vol. 27, pp. 7–26, 2006.

[9] P. Gertler and J. Gruber, "Insuring consumption against illness," *The American Economic Review*, vol. 87, no. 1, pp. 51–70, 1997.

[10] A. Bitran and U. Giedion, "Waivers and exemptions for health services in developing countries," *Social Science & Medicine*, vol. 54, no. 10, pp. 1465–1477, 2002.

[11] J. Murray and B. Chen, "Health system performance assessment," *Bulletin of the World Health Organization*, vol. 78, no. 6, pp. 717–731, 2000.

[12] J. S. Skinner *et al.*, "The evolving role of chronic conditions in medicare spending," *Health Affairs*, vol. 25, no. 5, pp. 1405–1414, 2006.

[13] S. S. S. Andaleeb, "Service quality perceptions and patient satisfaction: A study of hospitals in a developing country," *Social Science & Medicine*, vol. 52, no. 9, pp. 1359–1370, 2001.

[14] J. D. Grigsby and P. J. Thorndyke, "Geographic access to healthcare for rural medicare beneficiaries: An analysis of physician location and service accessibility," *Medical Care Research and Review*, vol. 42, no. 2, pp. 115–139, 1985.

[15] K. E. Hansen and J. T. Gfroerer, "Rural barriers to access of health services: An analysis of rural versus urban medicaid beneficiaries in the united states," *Journal of Rural Health*, vol. 21, no. 3, pp. 183–189, 2005.

[16] E. S. Fisher and H. G. Welch, "Housing insecurity and health: A review of the literature," *Social Science & Medicine*, vol. 65, no. 4, pp. 571–576, 2007.

[17] M. E. Kruk *et al.*, "Financial protection in health: The role of poverty, income inequality, and health financing," *Health Policy and Planning*, vol. 24, no. 2, pp. 89–99, 2009.

[18] C. M. Bishop, *Pattern Recognition and Machine Learning.* Springer, 2006.

[19] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York: Springer, 2009.

[20] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning with Applications in R.* Springer, 2013.

[21] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[22] A. F. Zuur, E. N. Ieno, N. J. Walker, A. A. Saveliev, and G. M. Smith, *Mixed Effects Models and Extensions in Ecology with R.* New York, NY, USA: Springer, 2009.

[23] A. G. Ospina, H. M. Jaramillo, and J. G. Giraldo, "Determinantes del gasto de bolsillo y gasto catastrófico en la región central de colombia (2008)," *Revista Gestión y Región*, 2011. [Online]. Available: https://revistas.ucp.edu.co/index.php/gestionyregion/article/download/858/850

[24] B. Mihaylova, A. McKeever, and J. P. Newhouse, "Review of statistical methods for analysing healthcare resources and costs," *Health Econ.*, vol. 20, pp. 1029–1046, 2011.

[25] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[26] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[27] J. H. Friedman, "Multivariate adaptive regression splines," *The Annals of Statistics*, vol. 19, no. 1, pp. 1–67, 1991.

[28] Friedman, "Stochastic gradient boosting," *Computational Statistics & Data Analysis*, vol. 38, no. 4, pp. 367–378, 1999.

[29] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Wadsworth International Group, 1984.

[30] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis*. John Wiley & Sons, 2012.

[31] F. Muremyi *et al.*, "Prediction of out-of-pocket health expenditures in rwanda using machine learning techniques," *PAMJ - Clinical Medicine*, vol. 37, no. 357, 2021.

[32] "Multivariable prediction models for health care spending using machine learning: a protocol of a systematic review," *Diagnostic and Prognostic Research*, 2021. [Online]. Available: https://diagnprognres.biomedcentral.com

[33] "Machine-learning-based prediction models for high-need high-cost patients using nationwide clinical and claims data," *npj Digital Medicine*, 2020. [Online]. Available: https://www.nature.com

[34] "Supervised learning methods for predicting healthcare costs: Systematic literature review and empirical evaluation," *PubMed*, 2021. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov

[35] "The application of machine learning to predict high-cost patients: A performance-comparison of different models using healthcare claims data," *PLOS ONE*, 2021. [Online]. Available: https://journals.plos.org

[36] K. Xu, D. B. Evans, K. Kawabata, R. Zeramdini, J. Klavus, and C. J. Murray, "Household catastrophic health expenditure: A global analysis," *The Lancet Global Health*, vol. 3, no. 4, pp. e234–e245, 2015.

[37] A. Hernández-Vásquez, "Gasto de bolsillo en salud en adultos mayores peruanos: análisis de la encuesta nacional de hogares sobre condiciones de vida y pobreza 2017," *Revista Peruana de Medicina Experimental y Salud Pública*, vol. 35, no. 3, pp. 390–397, 2018. [Online]. Available: https://www.scielosp.org/article/ssm/content/raw/?resource_ssm_path=/media/assets/rpmesp/v35n3/1726-4642-rpmesp-35-03-390.pdf

[38] J. R. Jorrat, M. M. Fernández, and E. H. Marconi, "Utilización y gasto en servicios de salud de los individuos en argentina en

2005: comparaciones internacionales de diferenciales socio-económicos en salud," *Salud colectiva*, vol. 4, no. 1, pp. 57–76, 2008. [Online]. Available: https://www.scielosp.org/pdf/scol/2008.v4n1/57-76/es

[39] M. Suhrcke, L. M. Gutacker, and N. Rice, "Economic considerations in the expansion of universal health coverage," *The European Journal of Health Economics*, vol. 20, pp. 605–619, 2019.

[40] Y.-K. Tu, "The use of generalized linear mixed models for analyzing binary data," *American Journal of Epidemiology*, vol. 156, no. 2, pp. 111–118, 2002.

[41] A. G. Ospina, H. M. Jaramillo, and J. G. Giraldo, "Determinantes del gasto de bolsillo y gasto catastrófico en la región central de colombia (2008)," *Revista Gestión y Región*, 2011. [Online]. Available: https://revistas.ucp.edu.co/index.php/gestionyregion/article/download/858/850

[42] M. F. L. Urbano, "Riesgo de familias peruanas en incurrir en gasto catastrófico en salud," *Revista del Centro de Estudios en Población y Desarrollo Humano*, 2010. [Online]. Available: http://biblioteca.ccp.ucr.ac.cr/bitstream/handle/123456789/1513/Riesgo%20de%20familias%20peruanas%20en%20incurrir%20en%20Gasto%20Catastr%C3%B3fico%20en%20Salud.pdf?sequence=1

[43] J. C. Rodríguez, E. G. Caballero, M. A. Esquivel, and N. Peralta, "Análisis de gastos de bolsillo de salud excesivos por quintiles de ingresos en paraguay," *Novapolis*, 2021. [Online]. Available: http://pyglobal.com/ojs/index.php/novapolis/article/download/138/143

[44] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[45] P. Domingos, "A few useful things to know about machine learning," *Communications of the ACM*, vol. 55, no. 10, pp. 78–87, 2012.

[46] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

Capítulo 10

Anexos

Cuadro 10.1: Resumen descriptivo de variables cuantitativas.

Variable	Media	Desv. Est.	Mínimo	Mediana	Máximo
CANT_PERSONAS_HOGAR	2.78	1.50	1	3	24
I_HOGAR_ANUAL	25,124,400	44,779,750	0	15,600,000	3,226,860,000
GASTO_TOTAL_ANUAL	211,169.8	1,313,746	0	0	95,400,000

Cuadro 10.2: Distribución de la variable Afiliación a Seguridad Social en Salud.

Categoría	Frecuencia	Porcentaje (%)
1: Sí	82,772	96.18
2: No	3,031	3.52
9: No sabe/no informa	260	0.30

Cuadro 10.3: Distribución de la variable Régimen de Seguridad Social en Salud.

Categoría	Frecuencia	Porcentaje (%)
3: Subsidiado (EPS-S)	57,223	66.49
1: Contributivo (EPS)	23,488	27.29
9: No sabe/no informa	190	0.22
2: Especial	1,871	2.17
NaN: Faltante	3,291	3.82

Cuadro 10.4: Distribución de la variable Estado General de Salud.

Categoría	Frecuencia	Porcentaje (%)
2: Bueno	59,341	68.95
3: Regular	16,015	18.61
1: Muy bueno	9,678	11.25
4: Malo	1,029	1.20

Cuadro 10.5: Distribución de la variable Diagnóstico de Enfermedades Crónicas.

Categoría	**Frecuencia**	**Porcentaje (%)**
2: No	70,742	82.20
1: Sí	15,321	17.80

Cuadro 10.6: Distribución de la variable Calidad del Servicio de la EPS.

Categoría	**Frecuencia**	**Porcentaje (%)**
2: Buena	58,035	67.43
3: Regular	14,203	16.50
1: Muy buena	6,006	6.98
4: Mala	2,553	2.97
9: No sabe	1,785	2.07
NaN: Faltante	3,481	4.04

Cuadro 10.7: Distribución por región.

Región	**Frecuencia**	**Porcentaje (%)**
1: Caribe	18,566	21.57
2: Pacífica	16,269	18.90
3: Oriental	16,017	18.61
4: Central	8,417	9.78
5: Bogotá	1,786	2.08
6: Orinoquía-Amazonía	3,874	4.50
7: San Andrés	3,134	3.64
8: Exterior	876	1.02
9: No informa	17,124	19.90

Cuadro 10.8: Distribución de la variable CLASE (Tipo de Localidad).

Categoría	**Frecuencia**	**Porcentaje (%)**
1: Cabecera	43,434	50.47
2: Rural	42,629	49.53

Cuadro 10.9: Distribución de la variable Tenencia de Vivienda.

Categoría	**Frecuencia**	**Porcentaje (%)**
1: Propia (pagada)	33,446	38.86
3: Arriendo	24,754	28.76
4: Usufructo	16,786	19.50
6: Propiedad colectiva	4,718	5.48
5: Ocupación sin título	4,701	5.46
2: Propia (pagando)	1,658	1.93

Cuadro 10.10: Distribución de la variable Autopercepción de Pobreza.

Categoría	**Frecuencia**	**Porcentaje (%)**
1: Sí	50,927	59.17
2: No	35,136	40.83

Cuadro 10.11: Resultados del Modelo de Regresión Lineal (OLS)

max width=

Variable	**Coef.**	**Error estándar**	**t**	**P>|t|**	**[0.025]**	**[0.975]**
const	3.921e+05	3.14e+04	12.474	0.000	3.3e+05	4.54e+05
CANT_PERSONAS_HOGAR	-2.01e+04	3446.289	-5.833	0.000	-2.69e+04	-1.33e+04
I_HOGAR_ANUAL	0.0018	0.000	14.962	0.000	0.0018	0.002
P6100_2.0	1.762e+05	3.57e+04	4.943	0.000	1.06e+05	2.46e+05
P6100_3.0	-1.438e+05	1.32e+04	-10.856	0.000	-1.7e+05	-1.18e+05
P6100_9.0	-1.218e+05	1.09e+05	-1.113	0.266	-3.36e+05	9.27e+04
P6127_2	-544.158	1.7e+04	-0.032	0.974	-3.39e+04	3.28e+04
P6127_3	2.035e+05	2.1e+04	9.696	0.000	1.62e+05	2.45e+05
P6127_4	5.637e+05	5e+04	11.274	0.000	4.66e+05	6.62e+05
P1930_2	-2.234e+05	1.48e+04	-15.142	0.000	-2.52e+05	-1.94e+05
P6181_2.0	2682.719	2.1e+04	0.127	0.899	-3.86e+04	4.39e+04
P6181_3.0	1.416e+05	2.4e+04	5.890	0.000	9.45e+04	1.89e+05
P6181_4.0	2.614e+05	3.54e+04	7.387	0.000	1.92e+05	3.31e+05
P6181_9.0	1.213e+05	4.04e+04	3.006	0.003	4.22e+04	2e+05
CLASE_2	-9688.363	1.15e+04	-0.842	0.400	-3.22e+04	1.29e+04
P5095_2	1.203e+05	3.8e+04	3.168	0.002	4.59e+04	1.95e+05
P5095_3	-2.981e+04	1.33e+04	-2.247	0.025	-5.58e+04	-3807.465
P5095_4	-4.529e+04	1.43e+04	-3.159	0.002	-7.34e+04	-1.72e+04
P5095_5	-1.415e+04	2.34e+04	-0.604	0.546	-6.01e+04	3.18e+04
P5095_6	-3.688e+04	2.4e+04	-1.536	0.125	-8.4e+04	1.02e+04
P6090_2	8.039e+04	2.81e+04	2.866	0.004	2.54e+04	1.35e+05
P6090_9	3.5e+04	9.28e+04	0.377	0.706	-1.47e+05	2.17e+05
REGION_2	1.819e+04	1.66e+04	1.097	0.273	-1.43e+04	5.07e+04
REGION_3	6.236e+04	1.64e+04	3.807	0.000	3.03e+04	9.45e+04
REGION_4	4.069e+04	1.99e+04	2.050	0.040	1780.332	7.96e+04
REGION_5	3.91e+05	3.8e+04	10.292	0.000	3.17e+05	4.65e+05
REGION_6	-1.456e+04	2.64e+04	-0.551	0.582	-6.64e+04	3.73e+04
REGION_7	6.835e+04	2.91e+04	2.346	0.019	1.12e+04	1.25e+05
REGION_8	-1.967e+05	5.11e+04	-3.850	0.000	-2.97e+05	-9.66e+04
REGION_9	-2.425e+04	1.64e+04	-1.482	0.138	-5.63e+04	7818.164
P5230_2	7.434e+04	1.15e+04	6.487	0.000	5.19e+04	9.68e+04