

Taller 2 - Predicción de Pobreza en Colombia

Harold Stiven Acuña

José David Cuervo

José David Dávila

César Augusto Alfaro

13 de abril de 2025

Resumen

Este documento presenta el análisis de datos y la implementación de modelos de clasificación para la predicción de la pobreza en Colombia.

Palabras clave: pobreza, clasificación, aprendizaje automático

Clasificación JEL: J31, C53, J16

Repositorio GitHub:

https://github.com/alfarocesar/BDML_Predicting_Poverty_Equipo8

1. Introducción

La migración del campo a la ciudad, producto de la Primera y Segunda Revolución Industrial, trajo consigo un aumento significativo de la población, así como conflictos de orden político, económico y social, en igual medida para todos los países que decidieron sumarse a dicho proceso de transformación emergente, como Gran Bretaña, Francia, Alemania, Estados Unidos, entre otros. Es así como la pobreza, como concepto abstracto y simbólico, se introduce en la discusión del mundo científico, mismo que se encontraba en auge, no solo para las ciencias exactas que derivan de las leyes de la física, sino también en ciencias cuyo objeto de estudio resultaba más abstracto y retador, aún en nuestros tiempos: la sociedad.

La pobreza, entonces, generó la inminente necesidad de ser considerada, sobre todo, para encontrar formas de medición acertadas. La razón de ser de ello es muy sencilla: a medida que la sociedad avanzaba, el discurso de los derechos se hizo más fuerte; la dignidad, el mínimo vital, el contrato social, entre otros, se hicieron vigentes y, en consecuencia, la pobreza se convirtió en objeto de estudio, con el propósito de determinar su existencia, evolución y distribución.

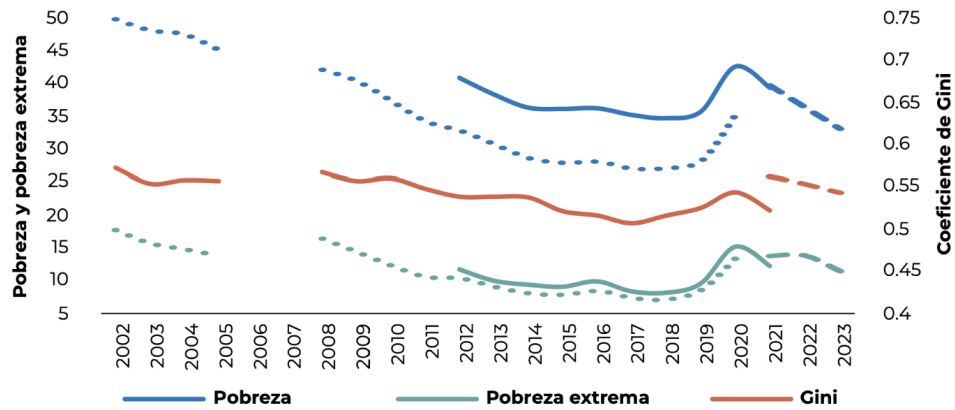
En este punto, la medición de la pobreza adquiere especial relevancia y su análisis, con el paso del tiempo, ha incluido inferencia, teoría, estadística, abstracción, idealización, predicción, cálculo y construcción de instrumentos [Tal \(2016\)](#). Sin embargo, aún falta teoría y una estructura definida que permita a los investigadores concluir que determinada asignación numérica representa adecuadamente la característica de la pobreza que se pretende medir [Huffman and Nájera \(2024\)](#).

Amartya Sen señala que la pobreza se mide no solo frente a la falta de ingresos y recursos, sino también en la falta de capacidades que restringen las opciones y oportunidades de las personas. Así las cosas, sobrevienen dos dificultades hasta el momento: la falta de una estructura para medir la pobreza y la determinación de las características que representan dicha pobreza.

Una medición precisa, eficiente y oportuna de la pobreza es clave para diseñar políticas públicas que sean no solo efectivas, sino también costo-eficientes. Sin embargo, los métodos tradicionales de medición —basados en encuestas extensas— implican altos costos y tiempos prolongados de recolección, procesamiento y análisis. En respuesta a esta limitación, el uso de técnicas de *machine learning* ofrece una alternativa prometedora: construir modelos que permitan predecir la condición de pobreza de los hogares utilizando un número reducido de variables y, en consecuencia, realizar evaluaciones más rápidas y baratas.

Este proyecto busca desarrollar modelos de clasificación binaria que permitan identificar si un hogar colombiano se encuentra en condición de pobreza, utilizando microdatos del DANE y la Misión MESE, a nivel de hogar e individuo. La Figura 1 ilustra la evolución reciente de la pobreza en el país, motivando así la necesidad de mejorar las herramientas de diagnóstico.

Figura 1. Evolución reciente de la pobreza monetaria en Colombia.



Fuente: DANE, GEIH 2002-2023.

Nota: Las líneas de pobreza y pobreza extrema se actualizaron en 2019 y hay una interrupción metodológica a partir de 2021.

A lo largo del documento se evalúan diferentes algoritmos de clasificación (como regresión logística, árboles de decisión, random forest, entre otros) y se presenta una comparación sistemática de su desempeño. El modelo con mejor rendimiento logró una puntuación F1 de **[0.5686]**, utilizando solo **[número]** variables, lo que representa un avance en términos de precisión y simplicidad. Este modelo fue seleccionado como la base para las predicciones entregadas en Kaggle.

Además de identificar el mejor algoritmo, se discute la importancia relativa de las variables predictoras, destacando aquellas con mayor capacidad explicativa. Estos hallazgos permiten no solo mejorar la focalización de políticas sociales, sino también abrir camino para sistemas de monitoreo más ágiles y adaptativos.

2. Datos

2.1. Adecuación de los datos

Los datos utilizados en este estudio provienen del DANE y la misión “Empalme de las Series de Empleo, Pobreza y Desigualdad (MESE)”. Estos datos son idóneos para resolver el problema de predicción de pobreza por varias razones fundamentales. Primero, contienen información detallada a nivel de hogar e individuo, permitiendo capturar la heterogeneidad socioeconómica de los diferentes grupos poblacionales. Segundo, incluyen la variable objetivo (Pobre) correctamente definida según el criterio oficial: un hogar es clasificado como pobre si su ingreso per cápita es menor a la línea de pobreza establecida. Finalmente, contienen un conjunto completo de variables sobre composición demográfica, características laborales, educativas y de vivienda que la literatura ha identificado como teóricamente relevantes para explicar la pobreza.

Para este análisis, disponemos de cuatro conjuntos de datos: `train_hogares.csv` (164,960 observaciones y 23 variables), `train_personas.csv` (543,109 observaciones y 135 varia-

bles), `test_hogares.csv` (66,168 observaciones y 16 variables) y `test_personas.csv` (150,367 observaciones y 63 variables). Esta configuración permite realizar predicciones fuera de muestra, lo que representa un desafío adicional ya que varias variables presentes en los datos de entrenamiento están ausentes en los de prueba, simulando un escenario real donde no toda la información está disponible al momento de realizar predicciones.

Como verificación preliminar, comprobamos que la variable objetivo *Pobre* estuviera correctamente definida siguiendo la metodología del DANE mediante dos métodos diferentes:

$$\text{Pobre_calculado} = I(\text{Ingpcug} < \text{Lp}) \quad (1)$$

$$\text{Pobre_calculado2} = I(\text{Ingto tugarr} < \text{Lp} \times \text{Npersug}) \quad (2)$$

Al realizar esta validación, obtuvimos una coincidencia del 100 % con la variable original, confirmando su correcta definición. La distribución de la variable objetivo reveló un desbalance importante: aproximadamente el 80 % de los hogares se clasifican como no pobres (131,936 hogares), mientras que el 20 % son considerados pobres (33,024 hogares), con un ratio de desbalance de aproximadamente 4:1.

2.2. Construcción de la muestra

El proceso de construcción de la muestra involucró varios pasos clave:

2.2.1. Análisis y limpieza de las bases de datos

Iniciamos con un examen detallado de las variables disponibles en las bases de entrenamiento y prueba, identificando que entre las bases de hogares existen 16 variables comunes, mientras que entre las bases de personas hay 63 variables comunes. Este análisis fue crucial para determinar qué variables podíamos utilizar en nuestros modelos predictivos.

Se identificaron y eliminaron variables que no aportaban valor predictivo o que no estaban disponibles en los conjuntos de prueba:

1. Variables exclusivas de entrenamiento que indicaban directamente el resultado: `Ingto tug`, `Ingto tugarr`, `Ingpcug`, `Indigente`, `Npobres`, `Nindigentes`.
2. Variables de identificación o ponderación que no aportaban valor predictivo: `Directorio`, `Secuencia_p`, `Mes`, `P5130`, `Fex_c`, `Fex_dpto`.
3. En la base de personas, se eliminaron 72 variables que no estaban presentes en los datos de prueba, principalmente relacionadas con montos de ingresos.

Para las variables P5100 (valor del arriendo) y P5140 (pago por servicios), las cuales presentaban valores faltantes, realizamos una imputación específica reemplazando los NA con 0, entendiendo que estos representan gastos que pueden legítimamente no existir.

2.2.2. Transformación de variables de personas

La base de personas requirió un procesamiento especial:

1. **Variables binarias:** Para variables como `Pet`, `0c`, `Des`, `Ina` que solo toman valores 1 y NA, transformamos los NA a 0 para estandarizar su uso en los cálculos agregados.
2. **Agrupación de la variable Oficio:** Analizamos la relación entre las diferentes categorías de oficios y la pobreza. Mediante un análisis de clustering (k-means con $k=3$), agrupamos los 83 códigos originales en tres categorías según su asociación con la pobreza:
 - **Grupo 1:** 45 ocupaciones con baja tasa de pobreza (media de 5.3 %)
 - **Grupo 2:** 28 ocupaciones con tasa media de pobreza (media de 19.4 %)
 - **Grupo 3:** 10 ocupaciones con alta tasa de pobreza (media de 37.5 %)
3. **Variables calculadas adicionales:** Creamos variables como `tiene_arriendo`, `tiene_otros_ingresos` y `otro_bene_año` para capturar información económica relevante.

2.2.3. Agregación de datos a nivel de hogar

Para poder construir un modelo a nivel de hogar utilizando la información de individuos, realizamos varias agregaciones:

1. **Conteos básicos y promedios:** Calculamos el número de miembros por hogar, número de hombres, mujeres, menores, personas en edad de trabajar, ocupados, desocupados e inactivos. También calculamos la edad promedio, máxima edad y máximo nivel educativo en el hogar.
2. **Características del jefe de hogar:** Extrajimos variables específicas del jefe de hogar (identificado con `P6050 = 1`), como sexo, edad, nivel educativo, años de educación y situación laboral.
3. **Proporciones de variables categóricas:** Para 31 variables categóricas relevantes, calculamos proporciones dentro de cada hogar (por ejemplo, proporción de miembros que cotizan a pensión o proporción con actividad adicional).

Tras este procesamiento, logramos construir conjuntos de datos unificados con 79 variables para entrenamiento y 78 para prueba (la diferencia está en la variable objetivo `Pobre` que solo existe en entrenamiento), manteniendo la estructura de un registro por hogar.

2.3. Limpieza de datos y tratamiento de valores faltantes

El análisis de valores faltantes reveló patrones importantes en nuestros datos. Mediante el uso de las funciones `skim` y `vis_miss`, identificamos que:

1. El conjunto de entrenamiento unido presentaba valores faltantes en 36 variables, con proporciones que variaban desde el 10.9 % (para `prop_cotiza_pension`) hasta el 54.5 % (para `prop_P7110` y `prop_P7120`).
2. Se observó un patrón diferente de valores faltantes entre el conjunto de entrenamiento y prueba. En el conjunto de prueba, la proporción de faltantes era significativamente mayor para todas las variables, llegando a superar el 50 % en muchos casos (por ejemplo, `prop_P7510s1` a `prop_P7510s7` con 55.7 % de faltantes).
3. El análisis de correlación de valores faltantes reveló que variables relacionadas con características similares (como distintos tipos de ingresos o subsidios) presentaban patrones similares de valores faltantes, lo que sugiere un origen estructural en la ausencia de datos.

Para abordar estos valores faltantes, tomamos las siguientes decisiones:

1. **Variables con alta proporción de faltantes (>30 %):** Evaluamos cuidadosamente su relevancia. Decidimos mantener algunas de estas variables a pesar de su alta proporción de faltantes debido a su potencial poder predictivo, aplicando métodos adecuados de imputación.
2. **Imputación para variables numéricas:** Para variables como `promedio_horas_trab`, `prop_cotiza_pension` y `prop_actividad_adicional`, utilizamos la imputación por la mediana para mantener la robustez frente a outliers.
3. **Imputación para variables categóricas:** Para variables como `max_nivel_educativo` y `max_años_educ`, aplicamos imputación por la moda, dada su naturaleza discreta.

2.4. Análisis descriptivo

2.4.1. Distribución de variables numéricas

El análisis descriptivo de las variables numéricas revela aspectos importantes de las características socioeconómicas de los hogares estudiados. Las estadísticas descriptivas de las variables continuas (como se muestra en el Anexo 1) y discretas (Anexo 2) proporcionan una visión detallada de la distribución de estas variables.

Entre los hallazgos más relevantes, observamos que:

- El promedio de miembros por hogar es de 3.29 personas, con un máximo de 28 personas en algunos casos extremos.

- La edad promedio en los hogares es de 37.44 años, mientras que la edad media de los jefes de hogar es considerablemente más alta: 49.61 años.
- El promedio de horas trabajadas entre las personas ocupadas es de 45.25 horas semanales, con una importante variabilidad (desviación estándar de 13.11).
- Solo el 31 % de las personas en edad de trabajar cotiza a pensión, lo que refleja una importante informalidad laboral.
- La proporción de ocupaciones por grupos muestra que el 51 % de los trabajadores se encuentran en ocupaciones del grupo 2 (tasa media de pobreza), mientras que solo el 6 % están en ocupaciones del grupo 3 (alta tasa de pobreza).

2.4.2. Análisis de outliers

Mediante la aplicación del método IQR (rango intercuartílico), identificamos variables con alta proporción de valores atípicos. Variables como `prop_P7120`, `prop_P7150` y `prop_P7510s3` presentaron más del 23 % de outliers, mientras que variables fundamentales como `promedio_horas_trab` (12.4 % de outliers) y `prop_afiliados_ss` (12.32 % de outliers) también mostraron una cantidad significativa de valores extremos. Estos hallazgos fueron considerados durante el procesamiento y modelado para evitar que valores extremos afectaran indebidamente nuestras predicciones.

2.4.3. Relación con la variable objetivo

El análisis de la relación entre las variables predictoras y la pobreza reveló asociaciones significativas:

1. **Variables numéricas:** Calculamos la correlación de cada variable numérica con la variable objetivo, encontrando que variables como `prop_cotiza_pension` (-0.51), `prop_oficio_grupo3` (0.43) y `max_nivel_educativo` (-0.39) presentaban las correlaciones más fuertes.
2. **Variables categóricas:** El análisis chi-cuadrado reveló asociaciones significativas entre la pobreza y variables como `Dominio` (V de Cramér = 0.20), `P5090` (tipo de ocupación de la vivienda, V de Cramér = 0.18) y `max_nivel_educativo` (V de Cramér = 0.24). Estas asociaciones confirmaron la importancia de incluir variables relacionadas con ubicación geográfica, vivienda y capital humano en nuestros modelos.

2.5. Selección final de variables

La selección final de variables para nuestros modelos se basó en tres criterios principales:

1. **Relevancia predictiva:** Utilizamos las correlaciones con la variable objetivo y los resultados de las pruebas chi-cuadrado para identificar las variables con mayor poder predictivo.
2. **Disponibilidad en datos de prueba:** Garantizamos que todas las variables seleccionadas estuvieran disponibles tanto en el conjunto de entrenamiento como en el de prueba.
3. **Parsimonia:** Buscamos un conjunto reducido de variables que maximizara el poder predictivo y minimizara la complejidad.

Eliminamos variables con más del 33% de valores faltantes, a menos que mostraran una fuerte asociación con la variable objetivo. También evaluamos la redundancia entre variables, eliminando aquellas que presentaban correlaciones muy altas entre sí (>0.7), manteniendo la que mostraba mayor asociación con la pobreza.

El conjunto final de variables seleccionadas incluyó características demográficas del hogar (número y composición de miembros), variables de capital humano (nivel educativo y afiliación a seguridad social), características laborales (distribución por grupos de ocupación, horas trabajadas) y características específicas del jefe de hogar. Esta selección equilibrada nos permitió construir modelos robustos y parsimoniosos para la predicción de pobreza.

3. Modelos y Resultados

3.1. Modelo de Selección y Entrenamiento

3.1.1. Metodología

3.1.1.1. 1. Regresión Logística Esta primera metodología estima la probabilidad de que una observación pertenezca a una clase (por ejemplo, "pobrequando en este caso) en función de una combinación lineal de las variables independientes. Esta combinación se transforma usando la función de distribución de probabilidad logística, lo que garantiza que los valores estimados estén entre 0 y 1. Aunque no tiene muchos hiperparámetros, en este trabajo se utilizó la metodología de validación cruzada para ajustar el umbral de clasificación (por ejemplo, elegir si se clasifica como "pobrequando la probabilidad es mayor a 0.5 o a otro valor), además de evaluar su capacidad predictiva comparado con modelos más complejos. La forma funcional básico de dicho modelo es:

$$P(y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 + \dots + \beta_k)}}$$

3.1.1.2. 2. Elastic Net Este modelo es una variante penalizada de la regresión logística que incluye regularización mediante una combinación de Lasso (penalización L1)

y Ridge (penalización L2). Es útil cuando hay muchas variables correlacionadas o cuando se busca reducir el número de variables relevantes. Los hiperparámetros clave son lambda (intensidad de la penalización) y alpha (mezcla entre L1 y L2). Ambos se seleccionaron con validación cruzada, evaluando el rendimiento del modelo en diferentes combinaciones para elegir la más adecuada. La forma funcional de este modelo es:

$$LOSS = -\log(\beta) + \rho(\alpha \sum |\beta_j| + (1-\alpha) \sum \beta_j^2)$$

3.1.1.3. 3. Árboles de Clasificación (CART) Este método construye un árbol de decisión dividiendo el conjunto de datos en subconjuntos más homogéneos según los valores de las variables. En cada nodo, esta metodología selecciona la variable y el punto de corte que maximizan la separación entre clases. Aunque es muy fácil de interpretar, un solo árbol puede sobreajustarse a los datos si es muy profundo o si se permite dividir con pocos datos. Por eso, se utilizó validación cruzada para determinar la profundidad óptima del árbol, el número mínimo de observaciones por nodo y el parámetro de complejidad que regula la poda del árbol.

3.1.1.4. 4. GBM (Gradient Boosting Machines) GBM (Gradient Boosting Machines) es una técnica avanzada de Boosting utilizada para mejorar la precisión de los modelos predictivos, especialmente en problemas de clasificación y regresión. A diferencia de los métodos tradicionales de Boosting, GBM utiliza el gradiente descendente para minimizar la función de error, lo que permite que el modelo ajuste las predicciones de manera más eficiente.

3.1.1.5. 5. Naive Bayes Naive Bayes es un clasificador probabilístico basado en el teorema de Bayes, que asume independencia condicional entre las variables predictoras. Aunque esta suposición es bastante fuerte, el modelo suele funcionar sorprendentemente bien en la práctica, especialmente cuando las variables tienen efectos independientes. No requiere un ajuste intensivo de hiperparámetros, pero aun así se utilizó validación cruzada para evaluar su desempeño y compararlo con los otros modelos. Para calcular dichas probabilidades de pertenecer a una clase u otra se utiliza la siguiente formula:

$$P(Y|X_1, X_2, \dots, X_k) = P(Y) \prod P(X_j|Y)$$

Donde esto es posible dado el supuesto de independencias de las variables.

3.1.2. Variables utilizadas

A continuación se presentan las principales variables utilizadas en los modelos, como se puede observar en la Tabla 3 del Anexo 3.

3.2. Resultados

A continuación se presentan los principales resultados con las metodologías anteriormente explicas, cabe resaltar que al ser una muestra de la población presenta un desbalance entre clases, pero afortunadamente los mismos microdatos ayudan a resolver esto mediante un factor de expansión que no es mas que pesos muestrales que ayudan a corregir este problema.

Los resultados de la Tabla 4 del Anexo 4 corresponden a las estimaciones del modelo Logit, los resultados muestran que variables como el hacinamiento, pertenecer al régimen subsidiado de salud y trabajar en empresas pequeñas están asociadas positivamente con la probabilidad de ser pobre, mientras que contar con educación, estar trabajando, cotizar a pensión y ser pensionado reducen significativamente dicha probabilidad. Por ejemplo, cotizar a pensión o tener mayor nivel educativo parece ser un factor protector frente a la pobreza, mientras que vivir en condiciones de hacinamiento o depender de subsidios refleja situaciones de vulnerabilidad estos resultados hay que interpretarlos de manera cuidadosa ya que no necesariamente representan una relación causal. Algunos coeficientes presentan magnitudes muy elevadas (como los de TGP o CotizaPension), lo cual puede estar relacionado con problemas de escalamiento o multicolinealidad, especialmente dado el tamaño de la muestra. A pesar de ello, la mayoría de las variables son altamente significativas y la reducción en la varianza residual sugiere que el modelo tiene un buen ajuste en comparación con uno sin predictores.

Los resultados de la Tabla 5 del Anexo 5 corresponden a la metodología CARTs, resultados muestran que, en promedio, el modelo alcanzó una precisión del 81.5 % con una desviación estándar baja (0.17 %), lo que sugiere que el modelo es bastante estable entre las diferentes particiones de la muestra. El coeficiente Kappa, que mide la concordancia ajustada por azar, tuvo un valor medio de 0.47, lo cual indica un nivel de acierto moderado entre las predicciones del modelo y la realidad. A pesar de que los valores máximos de precisión alcanzaron hasta 83.4 % y el Kappa llegó a 0.54, los valores mínimos fueron cercanos al 81.2 % y 0.46, lo que sugiere que el modelo no presenta grandes variaciones entre distintas configuraciones que dicha metodología contempló. Estos resultados indican que el árbol de decisión logró capturar patrones relevantes para predecir la pobreza con un buen nivel de precisión.. Una aclaración importante es que este fue el único modelo que se estimó con todas las variables, el resto fueron estimados siguiendo otra especificación (Pobre $\hat{\text{hacinamiento}}$ + r_{gast} + hacinamiento_c + TGP + educ_{cab} + trabajando + CotizaPension + Subsidiado + Pequena + Subsidios + CotizaPension + Pensionado + $\text{Ingresos_AlquilerPensiones}$ + OtrosIngresos) esto debido a que no todas las variables aportaban poder explicativo.

La Tabla 6 del Anexo 6 representa otra estimación bajo metodología CARTS pero con otra especificación, también ajustado mediante validación cruzada, presenta resultados consistentes en cuanto a precisión y estabilidad. El valor promedio de precisión (Accuracy) fue de 81.2 %, con una desviación estándar baja (0.24 %), lo que indica que el modelo mantiene un rendimiento estable al aplicarse sobre distintas particiones de la muestra. El coeficiente de Kappa promedio fue de 0.46, lo cual representa un acuerdo moderado entre las predicciones del modelo y la clasificación real al igual que la especificación anterior. Aunque los valores máximos de precisión y Kappa alcanzaron 82.5 % y 0.50

respectivamente, los mínimos estuvieron en torno al 80.9 % y 0.45, lo que muestra que el rendimiento es bastante homogéneo a lo largo de los diferentes valores del hiperparámetro α . En general, estos resultados refuerzan que el árbol de decisión logra capturar patrones relevantes para identificar condiciones de pobreza, aunque no presento grandes diferencias ante el cambio de especificación.

La Tabla 7 del Anexo 7 por su parte utilizó la metodología de Elastic Net. Los resultados muestran una variabilidad en los valores de precisión (Accuracy) y Kappa en función de los hiperparámetros α y λ . En promedio, la precisión del modelo fue de 80.45 %, con una desviación estándar de 0.12 %, lo que sugiere que el rendimiento del modelo es relativamente consistente a través de las diferentes particiones de la muestra. A lo largo del rango de α (que varió de 0.10 a 1.00) y λ (desde 0.00017 hasta 0.13491), la precisión mostró poca variación, manteniéndose cercana al 81 % en los percentiles 1, 2 y 3, lo que indica un buen ajuste del modelo. La tendencia en la precisión y el Kappa sugiere que la regularización aplicada por Elastic Net ayudó a controlar el sobreajuste sin perder mucho rendimiento, lo cual es positivo para problemas de alta dimensionalidad o multicolinealidad. En resumen, el modelo Elastic Net ofrece un buen balance entre precisión y estabilidad, con una capacidad moderada de clasificación, que es acorde a la magnitud de las variables involucradas.

La Tabla 8 del Anexo 8 presenta los resultados del modelo ajustado por la metodología de Naive Bayes que fue ajustado con los hiperparámetros α y λ . Los resultados indican que, en promedio, el modelo alcanzó una precisión (Accuracy) de 80.45 %, con una desviación estándar de 0.12 %, lo que indica una precisión bastante estable entre las particiones de la muestra. El coeficiente Kappa promedio fue de 0.38, la precisión se mantuvo relativamente constante, alcanzando un máximo de 81.31 % y un Kappa de 0.45. Los valores de precisión en el primer y tercer cuartil se mantienen alrededor de 80.7 % y 81.3 %, lo que sugiere que el modelo es robusto y no presenta grandes variaciones al ajustar los hiperparámetros. En cuanto a la desviación estándar de precisión, los valores son bajos, lo que refleja la estabilidad del modelo a través de las particiones de la muestra. En resumen, el modelo Naive Bayes parece ser efectivo para este conjunto de datos, logrando una buena precisión con una moderada capacidad de discriminación, y su rendimiento se mantiene estable a pesar de la variación en los hiperparámetros.

La Tabla 9 del Anexo 9 muestra las estimaciones del modelo mediante boosting, con los hiperparámetros shrinkage , interaction.depth , minobsinnode , y n.trees . Los resultados muestran una variabilidad en la precisión (Accuracy) y el coeficiente Kappa en función de los valores de los hiperparámetros. La precisión promedio fue de 78.86 %, con una desviación estándar baja (0.11 %). Al analizar los percentiles, la precisión alcanzó un máximo de 82.33 %, con un Kappa de 0.49 en el cuartil superior, lo que muestra que con valores de shrinkage y interaction.depth más altos, el modelo logra una mayor capacidad discriminativa. Por otro lado, los valores mínimos de precisión y Kappa fueron bastante bajos (74.86 % y 0.00, respectivamente), lo que refleja que con configuraciones subóptimas de los hiperparámetros, el modelo no logró un buen rendimiento. La mediana de precisión fue de 79.72 %, lo que indica que el modelo con una configuración intermedia de hiperparámetros proporciona un rendimiento razonablemente bueno y estable. Además, la desviación estándar de la precisión fue muy baja (cerca de 0.0004), lo que refuerza la estabilidad del modelo. En resumen, el modelo ajustado por boosting parece ser bastante efectivo, logrando buenos niveles de precisión y Kappa, especialmente con configuraciones

más altas en los hiperparámetros shrinkage e interaction.depth, lo cual podría indicar que el modelo está aprovechando mejor las interacciones no lineales entre las variables.

3.3. Matrices de Confusión

A continuación, se presentan las matrices de confusión para los diferentes modelos evaluados:

En la Tabla 10 del Anexo 10 se muestra la matriz de confusión para el primer modelo CART.

En la Tabla 11 del Anexo 11 se presenta la matriz de confusión para el segundo modelo CART.

La Tabla 12 del Anexo 12 contiene la matriz de confusión para el modelo Elastic Net.

La matriz de confusión para el modelo de regresión logística se puede observar en la Tabla 13 del Anexo 13.

Los resultados de la matriz de confusión para el modelo Naive Bayes se encuentran en la Tabla 14 del Anexo 14.

Finalmente, la matriz de confusión para el modelo GBM se presenta en la Tabla 15 del Anexo 15.

4. Conclusión

Este estudio buscó estimar la pobreza utilizando técnicas de Big Data y algoritmos de aprendizaje automático aplicados con datos de “Empalme de las Series de Empleo, Pobreza y Desigualdad – MESE”. A pesar de las limitaciones inherentes al problema, se logró un desempeño razonable, alcanzando un F1-score de 0.56. El modelo con mejor rendimiento fue XGBoost, gracias a su capacidad para capturar relaciones no lineales y su potencia en términos de regularización y ajuste fino de parámetros [Huang et al. \(2025\)](#). Uno de los principales aprendizajes de este ejercicio fue la estrategia utilizada para la imputación de datos faltantes. Se optó por imputar con la media en variables numéricas y con la moda en variables categóricas, una técnica sencilla de implementar, eficiente en términos computacionales y útil para mantener el tamaño original del conjunto de datos. Además, esta estrategia no altera la escala de las variables, lo que facilita la aplicación de diversos algoritmos de Machine Learning sin necesidad de ajustes adicionales.

El resultado obtenido también evidencia oportunidades de mejora significativas. Para incrementar la precisión del modelo y lograr métricas más robustas, es fundamental avanzar en la selección de variables más relevantes y explorar con mayor profundidad la sintonización de hiperparámetros. Como han señalado [Karmaker et al. \(2025\)](#), la identificación precisa de los determinantes socioeconómicos es clave para fortalecer la capacidad predictiva de los modelos de pobreza, especialmente en contextos donde los recursos y las

políticas deben dirigirse de manera efectiva.

En conclusión, aunque los resultados obtenidos son alentadores, el desarrollo de modelos más precisos y accionables para la predicción de la pobreza requiere una combinación de mejoras técnicas, refinamiento metodológico.

Anexos

A. Estadísticas descriptivas de variables continuas

Cuadro 1. Estadísticas descriptivas de variables continuas (personas)

Variable	% Missing	Mínimo	Máximo	Media	Mediana	Desv. Estándar	Coef. Vari.
P6040	0.0000	0.00E+00	1.10E+02	34	31	22	0
P6210s1	0.1757	0.00E+00	9.90E+01	6	5	4	0
P6426	0.5432	0.00E+00	9.48E+02	86	36	114	1
P6500	0.7862	0.00E+00	5.00E+07	1,068,740	781,550	1,280,535	1
P6510s1	0.7864	0.00E+00	4.00E+06	6,218	0	49,300	7
P6545s1	0.7864	0.00E+00	1.50E+07	4,286	0	97,143	22
P6580s1	0.7864	0.00E+00	4.72E+07	9,169	0	180,938	19
P6585s1a1	0.7865	0.00E+00	5.00E+06	2,149	0	30,520	14
P6585s2a1	0.7869	0.00E+00	8.00E+06	30,517	0	62,719	2
P6585s3a1	0.7864	0.00E+00	2.85E+07	9,257	0	87,498	9
P6585s4a1	0.7862	0.00E+00	1.30E+07	1,497	0	71,032	47
P6590s1	0.7862	0.00E+00	8.00E+06	23,920	0	83,961	3
P6600s1	0.7862	0.00E+00	8.00E+06	7,174	0	63,245	8
P6610s1	0.7862	0.00E+00	4.00E+06	4,689	0	44,050	9
P6620s1	0.7862	0.00E+00	6.30E+06	1,261	0	32,886	26
P6630s1a1	0.7862	0.00E+00	3.00E+08	570,510	80,000	1,400,285	2
P6630s2a1	0.7862	0.00E+00	9.00E+07	254,978	0	1,017,974	3
P6630s3a1	0.7862	0.00E+00	3.50E+08	144,078	0	1,214,662	8
P6630s4a1	0.7862	0.00E+00	1.30E+08	43,273	0	882,883	20
P6630s6a1	0.7862	0.00E+00	2.80E+07	2,238	0	152,204	68
P6750	0.7735	0.00E+00	1.00E+08	692,936	490,000	1,266,078	1
P6760	0.7724	1.00E+00	1.20E+01	1	1	1	0
P550	0.9734	0.00E+00	3.60E+08	4,348,796	3,200,000	6,963,873	1
P6800	0.5432	1.00E+00	1.30E+02	45	48	16	0
P7045	0.9775	0.00E+00	9.90E+01	13	12	9	0
P7070	0.5461	0.00E+00	4.80E+07	17,166	0	237,135	13
P7422s1	0.9509	0.00E+00	1.40E+07	114,021	0	399,057	3
P7472s1	0.6879	0.00E+00	6.00E+06	5,505	0	83,442	15
P7500s1a1	0.1757	0.00E+00	5.00E+07	25,354	0	251,015	9
P7500s2a1	0.1757	0.00E+00	3.18E+08	75,032	0	780,024	10
P7500s3a1	0.1757	0.00E+00	2.40E+07	2,569	0	91,565	35
P7510s1a1	0.1757	0.00E+00	4.80E+08	297,441	0	2,132,103	7
P7510s2a1	0.1757	0.00E+00	2.50E+08	37,796	0	931,776	24
P7510s3a1	0.1757	0.00E+00	3.00E+08	57,891	0	661,984	11
P7510s5a1	0.1757	0.00E+00	4.00E+08	23,669	0	1,461,142	61
P7510s6a1	0.1757	0.00E+00	3.60E+08	20,127	0	759,002	37
P7510s7a1	0.1757	0.00E+00	7.00E+08	73,670	0	2,296,475	31

Continúa en la siguiente página

(Continuación de la Tabla 1)

Variable	% Missing	Mínimo	Máximo	Media	Mediana	Desv. Estándar	Coef. Vari.
Impa	0.5586	0.00E+00	7.00E+07	936,228	760,000	1,301,617	1
Isa	0.5461	0.00E+00	4.80E+07	17,166	0	237,135	13
Ie	0.7862	0.00E+00	9.00E+06	36,546	0	128,336	3
Imdi	0.9879	0.00E+00	1.40E+07	604,396	450,000	716,699	1
Iof1	0.1757	0.00E+00	3.33E+07	1,972	0	121,762	61
Iof2	0.1757	0.00E+00	3.60E+07	73,363	0	443,466	6
Iof3h	0.1757	0.00E+00	4.00E+07	30,447	0	215,024	7
Iof3i	0.1757	0.00E+00	1.10E+06	4,234	0	20,985	4
Iof6	0.1757	0.00E+00	5.00E+07	25,354	0	251,015	9
Impaes	0.9479	6.67E+03	5.20E+07	1,265,281	850,000	1,711,962	1
Isaes	0.9978	6.00E+03	4.80E+07	608,732	250,000	1,656,914	2
Iees	0.9948	2.00E+03	8.00E+06	246,371	180,000	284,032	1
Imdies	0.9994	7.00E+03	6.00E+06	717,752	600,000	746,657	1
Iofles	0.9990	8.30E+01	2.50E+07	643,318	166,667	1,795,105	2
Iof2es	0.9927	7.04E+05	1.80E+07	1,882,954	1,375,000	1,556,259	0
Iof3hes	0.9857	1.67E+03	3.00E+07	357,441	200,000	649,108	1
Iof3ies	0.9983	2.50E+03	8.00E+05	70,841	65,000	63,244	0
Iof6es	0.9975	3.00E+04	3.00E+07	902,848	550,000	1,392,386	1
Ingtotob	0.1757	0.00E+00	8.58E+07	664,523	280,000	1,267,809	1
Ingtotes	0.9206	8.30E+01	5.50E+07	1,145,262	780,000	1,693,550	1
Ingtot	0.1757	0.00E+00	8.58E+07	774,792	435,000	1,373,891	1

B. Estadísticas descriptivas de variables discretas

Cuadro 2. Estadísticas descriptivas de variables discretas (personas)

Variable	% Missing	Clases	Moda	Frec. Moda	% Pobres en moda	% Pobres en clase más
Estrato1	0.0000	6	2	0.2597	0.2736	0
P6020	0.0000	2	2	0.5286	0.5408	0
P6050	0.0000	9	3	0.3534	0.4211	0
P6090	0.1757	9	1	0.9324	0.6614	0
P6100	0.2315	9	1	0.4772	0.1015	0
P6210	0.0419	9	3	0.2593	0.3210	0
P6240	0.1757	6	1	0.4656	0.2325	0
Oficio	0.5432	100	47	0.0500	0.0000	0
P6430	0.5432	9	4	0.4611	0.2080	0
P6510	0.7862	9	2	0.9348	0.0679	0
P6510s2	0.9863	2	2	0.7364	0.0015	0
P6545	0.7862	9	2	0.9876	0.0698	0
P6545s2	0.9975	2	2	0.6842	0.0001	0
P6580	0.7862	9	2	0.9665	0.0693	0

Continúa en la siguiente p

Cuadro 2 – Continuación

Variable	% Missing	Clases	Moda	Frec. Moda	% Pobres en moda	% Pobres en clase más
P6580s2	0.9931	2	2	0.7984	0.0004	0
P6585s1	0.7862	9	2	0.9745	0.0693	0
P6585s1a2	0.9948	2	2	0.6892	0.0004	0
P6585s2	0.7862	9	2	0.5273	0.0486	0
P6585s2a2	0.8997	2	2	0.7822	0.0168	0
P6585s3	0.7862	9	2	0.7830	0.0584	0
P6585s3a2	0.9539	2	2	0.9823	0.0112	0
P6585s4	0.7862	9	2	0.9965	0.0698	0
P6585s4a2	0.9993	2	2	0.9644	0.0001	0
P6590	0.7862	9	2	0.8581	0.0549	0
P6600	0.7862	9	2	0.9654	0.0683	0
P6610	0.7862	9	2	0.9711	0.0689	0
P6620	0.7862	9	2	0.9914	0.0697	0
P6630s1	0.7862	2	1	0.6019	0.0194	0
P6630s2	0.7862	2	2	0.7826	0.0633	0
P6630s3	0.7862	2	2	0.8253	0.0648	0
P6630s4	0.7862	2	2	0.9733	0.0694	0
P6630s6	0.7862	2	2	0.9982	0.0698	0
P6870	0.5432	9	1	0.3869	0.1766	0
P6920	0.5451	3	2	0.5969	0.2697	0
P7040	0.5432	2	2	0.9506	0.2896	0
P7050	0.9775	9	4	0.7356	0.0111	0
P7090	0.5432	2	2	0.9125	0.2627	0
P7110	0.9600	2	2	0.5580	0.0206	0
P7120	0.9600	2	1	0.9524	0.0397	0
P7140s1	0.8907	2	1	0.6036	0.0583	0
P7140s2	0.8907	2	1	0.9745	0.1121	0
P7150	0.8907	2	2	0.5910	0.0648	0
P7160	0.8907	9	1	0.9154	0.1064	0
P7310	0.9446	2	2	0.8859	0.0685	0
P7350	0.9509	9	1	0.6266	0.0387	0
P7422	0.9509	2	2	0.8262	0.0609	0
P7472	0.6879	2	2	0.9886	0.3468	0
P7495	0.1757	2	2	0.9083	0.7191	0
P7500s1	0.9244	9	2	0.6013	0.0065	0
P7500s2	0.9244	9	1	0.6293	0.0032	0
P7500s3	0.9244	9	2	0.9410	0.0083	0
P7505	0.1757	2	2	0.7902	0.5633	0
P7510s1	0.8271	9	2	0.5413	0.0842	0
P7510s2	0.8271	9	2	0.9619	0.1639	0
P7510s3	0.8271	9	2	0.6916	0.0757	0
P7510s5	0.8271	9	2	0.9817	0.1668	0
P7510s6	0.8271	9	2	0.8324	0.1638	0
P7510s7	0.8271	9	2	0.8646	0.1595	0

Continúa en la siguiente p

Cuadro 2 – Continuación

Variable	% Missing	Clases	Moda	Frec. Moda	% Pobres en moda	% Pobres en clase más
Pet	0.1757	1	1	1.0000	0.7308	0
Oc	0.5432	1	1	1.0000	0.3040	0
Des	0.9446	1	1	1.0000	0.0771	0
Ina	0.6879	1	1	1.0000	0.3497	0
Cclasnr2	0.9499	1	1	1.0000	0.0324	0
Cclasnr3	0.9978	1	1	1.0000	0.0008	0
Cclasnr4	0.9948	1	1	1.0000	0.0021	0
Cclasnr5	0.9994	1	1	1.0000	0.0005	0
Cclasnr6	0.9990	1	1	1.0000	0.0003	0
Cclasnr7	0.9927	1	1	1.0000	0.0005	0
Cclasnr8	0.9848	1	1	1.0000	0.0163	0
Cclasnr11	0.9975	1	1	1.0000	0.0004	0

Cuadro 3. Variables utilizadas en los modelos

Variable	Descripción
id	Identificador único del hogar o individuo
Fex_c	Factor de expansión del hogar o persona (para extrapolar a la población total)
Pobre	Dummy (1/0) que indica si el hogar es pobre según algún umbral oficial
dummy_urbano	Dummy que indica si el hogar está ubicado en zona urbana
hacinamiento	Número de personas por cuarto (medida de hacinamiento)
r_gast	Relación entre ingresos y gastos del hogar (o gasto per cápita)
hacinamiento_c	Dummy que indica si hay hacinamiento crítico (por ejemplo, >3 personas/cuarto)
gen_cab	Género del jefe del hogar
cab_formal	Dummy que indica si el jefe del hogar tiene un empleo formal
TGP	Tasa Global de Participación (proporción de la población económicamente activa)
tasa_desempleo	Tasa de desempleo del hogar o área donde reside
educ_cab	Nivel educativo del jefe del hogar
trabajando	Dummy que indica si el individuo está trabajando actualmente
CotizaPension	Dummy que indica si el individuo cotiza a pensión
Oc	Dummy que indica si el individuo está ocupado
Des	Dummy que indica si el individuo está desempleado
Ina	Dummy que indica si el individuo es inactivo
NoSeguro	Dummy que indica si el hogar o persona no tiene ningún tipo de seguro de salud
Subsidiado	Dummy que indica si el hogar está en régimen subsidiado de salud
BasicaSecundaria	Dummy que indica si el nivel educativo máximo es básica o secundaria
Media	Dummy que indica si el nivel educativo máximo es media (secundaria completa)
Independiente	Dummy que indica si el jefe del hogar trabaja por cuenta propia
Microempresa	Dummy que indica si trabaja en una microempresa
Pequena	Dummy que indica si trabaja en una pequeña empresa
Subsidios	Monto total de subsidios recibidos por el hogar
Pensionado	Dummy que indica si la persona es pensionada
Ingresos_AlquilerPension	Ingresos por alquileres y/o pensiones (no laborales)
OtrosIngresos	Otros ingresos no clasificados (remesas, loterías, etc.)
AyudasEco	Ayudas económicas recibidas de programas sociales, ONG, etc.

Cuadro 4. Estimaciones mediante metodología LOGIT

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	$6,493 \times 10^0$	$2,97 \times 10^4$	0.218	0.827
hacinamiento	$8,826 \times 10^{-1}$	$7,67 \times 10^{-1}$	*	$< 2 \times 10^{-16}$ ***
r_gast	$3,159 \times 10^0$	$3,34 \times 10^1$	94.711	$< 2 \times 10^{-16}$ ***
hacinamiento_c	$-4,225 \times 10^{-1}$	$2,27 \times 10^0$	-186.250	$< 2 \times 10^{-16}$ ***
cab_formal	$6,981 \times 10^{-1}$	$2,23 \times 10^0$	312.970	$< 2 \times 10^{-16}$ ***
TGP	$9,342 \times 10^{-1}$	$1,72 \times 10^0$	542.255	$< 2 \times 10^{-16}$ ***
educ_cab	$-1,105 \times 10^{-1}$	$3,70 \times 10^{-1}$	-298.519	$< 2 \times 10^{-16}$ ***
trabajando	$-1,444 \times 10^{-1}$	$5,78 \times 10^{-1}$	-250.066	$< 2 \times 10^{-16}$ ***
CotizaPension	$-1,311 \times 10^0$	$1,65 \times 10^0$	-794.844	$< 2 \times 10^{-16}$ ***
Subsidiado	$9,825 \times 10^{-1}$	$1,18 \times 10^0$	834.897	$< 2 \times 10^{-16}$ ***
Pequena	$2,233 \times 10^{-1}$	$2,40 \times 10^0$	93.114	$< 2 \times 10^{-16}$ ***
Subsidios	$-2,566 \times 10^{-1}$	$4,07 \times 10^{-1}$	-630.917	$< 2 \times 10^{-16}$ ***
Pensionado	$-2,415 \times 10^0$	$1,01 \times 10^1$	-238.875	$< 2 \times 10^{-16}$ ***
Ingresos_AlquilerPensiones	$-1,207 \times 10^1$	$2,97 \times 10^4$	-0.406	0.685

Signif. codes: *** 0.001 ** 0.01 * 0.05 (Dispersion parameter for binomial family taken to be 1)

Null deviance: 41518665 on 543108 degrees of freedom

Residual deviance: 28680402 on 543095 degrees of freedom

AIC: 28681433

Number of Fisher Scoring iterations: 7

Cuadro 5. Estimación 1 metodología CARTs

cp	Accuracy	Kappa	AccuracySD	KappaSD
Min. :0.000100	Min. :0.8126	Min. :0.4624	Min. :0.0009727	Min. :0.004039
1st Qu.:0.002475	1st Qu.:0.8126	1st Qu.:0.4624	1st Qu.:0.0014666	1st Qu.:0.008747
Median :0.004850	Median :0.8127	Median :0.4664	Median :0.0019802	Median :0.010789
Mean :0.004850	Mean :0.8150	Mean :0.4745	Mean :0.0017894	Mean :0.009487
3rd Qu.:0.007225	3rd Qu.:0.8154	3rd Qu.:0.4771	3rd Qu.:0.0020633	3rd Qu.:0.010789
Max. :0.009600	Max. :0.8342	Max. :0.5431	Max. :0.0020633	Max. :0.012497

Cuadro 6. Estimación 2 metodología CARTs

cp	Accuracy	Kappa	AccuracySD	KappaSD
Min. :0.000100	Min. :0.8097	Min. :0.4482	Min. :0.001521	Min. :0.005476
1st Qu.:0.002475	1st Qu.:0.8097	1st Qu.:0.4504	1st Qu.:0.002043	1st Qu.:0.011268
Median :0.004850	Median :0.8100	Median :0.4504	Median :0.002707	Median :0.012063
Mean :0.004850	Mean :0.8121	Mean :0.4585	Mean :0.002460	Mean :0.011135
3rd Qu.:0.007225	3rd Qu.:0.8123	3rd Qu.:0.4570	3rd Qu.:0.002868	3rd Qu.:0.012642
Max. :0.009600	Max. :0.8253	Max. :0.5008	Max. :0.002868	Max. :0.015887

Cuadro 7. Estimaciones por Elastic Net

alpha	lambda	Accuracy	Kappa	AccuracySD	KappaSD
Min. :0.10	Min. :0.0001663	Min. :0.7486	Min. :0.0000	Min. :5.844e-06	Min. :0.0000
1st Qu.:0.30	1st Qu.:0.0008876	1st Qu.:0.8070	1st Qu.:0.3859	1st Qu.:1.030e-03	1st Qu.:0.0000
Median :0.55	Median :0.0047370	Median :0.8127	Median :0.4414	Median :1.169e-03	Median :0.0000
Mean :0.55	Mean :0.0264176	Mean :0.8045	Mean :0.3842	Mean :1.196e-03	Mean :0.0000
3rd Qu.:0.80	3rd Qu.:0.0252798	3rd Qu.:0.8128	3rd Qu.:0.4508	3rd Qu.:1.489e-03	3rd Qu.:0.0000
Max. :1.00	Max. :0.1349104	Max. :0.8131	Max. :0.4510	Max. :1.532e-03	Max. :0.0000

Cuadro 8. Estimaciones por Naive Bayes

alpha	lambda	Accuracy	Kappa	AccuracySD	KappaSD
Min. :0.10	Min. :0.0001663	Min. :0.7486	Min. :0.0000	Min. :5.844e-06	Min. :0.0000
1st Qu.:0.30	1st Qu.:0.0008876	1st Qu.:0.8070	1st Qu.:0.3859	1st Qu.:1.030e-03	1st Qu.:0.0000
Median :0.55	Median :0.0047370	Median :0.8127	Median :0.4414	Median :1.169e-03	Median :0.0000
Mean :0.55	Mean :0.0264176	Mean :0.8045	Mean :0.3842	Mean :1.196e-03	Mean :0.0000
3rd Qu.:0.80	3rd Qu.:0.0252798	3rd Qu.:0.8128	3rd Qu.:0.4508	3rd Qu.:1.489e-03	3rd Qu.:0.0000
Max. :1.00	Max. :0.1349104	Max. :0.8131	Max. :0.4510	Max. :1.532e-03	Max. :0.0000

Cuadro 9. Estimaciones por GBM

shrinkage	interaction.depth	minobsinnode	n.trees	Accuracy	Kappa
Min. :0.01000	Min. :1	Min. :10	Min. : 10	Min. :0.7486	Min. :0.0000
1st Qu.:0.01000	1st Qu.:1	1st Qu.:10	1st Qu.: 30	1st Qu.:0.7486	1st Qu.:0.0000
Median :0.05000	Median :3	Median :10	Median : 55	Median :0.7972	Median :0.0000
Mean :0.05333	Mean :3	Mean :10	Mean : 55	Mean :0.7886	Mean :0.2000
3rd Qu.:0.10000	3rd Qu.:5	3rd Qu.:10	3rd Qu.: 80	3rd Qu.:0.8177	3rd Qu.:0.4000
Max. :0.10000	Max. :5	Max. :10	Max. : 100	Max. :0.8233	Max. :0.4900
AccuracySD			KappaSD		
Min. :4.469e-06			Min. :0.000000		
1st Qu.:4.469e-06			1st Qu.:0.000000		
Median :1.227e-03			Median :0.004886		
Mean :1.055e-03			Mean :0.005150		
3rd Qu.:1.433e-03			3rd Qu.:0.005963		
Max. :4.463e-03			Max. :0.028524		

Cuadro 10. Matriz de confusión - Modelo CART 1

Prediction	Reference	
	0	1
0	84,723	84,844
1	25,457	24,620
Accuracy : 0.4978		
95 % CI : (0.4957, 0.4999)		
No Information Rate : 0.5016		
P-Value [Acc > NIR] : 0.9998		
Kappa : -0.0061		
Mcnemar's Test P-Value : <2e-16		
Sensitivity : 0.7690		
Specificity : 0.2249		
Pos Pred Value : 0.4996		
Neg Pred Value : 0.4916		
Prevalence : 0.5016		
Detection Rate : 0.3857		
Detection Prevalence : 0.7720		
Balanced Accuracy : 0.4969		
'Positive' Class : 0		

Cuadro 11. Matriz de confusión - Modelo CART 2

Prediction	Reference	
	0	1
0	87,279	87,477
1	22,901	21,987
Accuracy : 0.4975		
95 % CI : (0.4954, 0.4996)		
No Information Rate : 0.5016		
P-Value [Acc > NIR] : 1		
Kappa : -0.007		
Mcnemar's Test P-Value : <2e-16		
Sensitivity : 0.7921		
Specificity : 0.2009		
Pos Pred Value : 0.4994		
Neg Pred Value : 0.4898		
Prevalence : 0.5016		
Detection Rate : 0.3974		
Detection Prevalence : 0.7956		
Balanced Accuracy : 0.4965		
'Positive' Class : 0		

Cuadro 12. Matriz de confusión - Modelo Elastic Net

Prediction	Reference	
	0	1
0	89,525	89,563
1	20,655	19,901
Accuracy : 0.4982		
95 % CI : (0.4961, 0.5003)		
No Information Rate : 0.5016		
P-Value [Acc > NIR] : 0.9994		
Kappa : -0.0057		
Mcnemar's Test P-Value : <2e-16		
Sensitivity : 0.8125		
Specificity : 0.1818		
Pos Pred Value : 0.4999		
Neg Pred Value : 0.4907		
Prevalence : 0.5016		
Detection Rate : 0.4076		
Detection Prevalence : 0.8154		
Balanced Accuracy : 0.4972		
'Positive' Class : 0		

Cuadro 13. Matriz de confusión - Modelo Logit

Prediction	Reference	
	0	1
0	87,304	87,485
1	22,876	21,979
Accuracy : 0.4975		
95 % CI : (0.4955, 0.4996)		
No Information Rate : 0.5016		
P-Value [Acc > NIR] : 0.9999		
Kappa : -0.0068		
McNemar's Test P-Value : <2e-16		
Sensitivity : 0.7924		
Specificity : 0.2008		
Pos Pred Value : 0.4995		
Neg Pred Value : 0.4900		
Prevalence : 0.5016		
Detection Rate : 0.3975		
Detection Prevalence : 0.7958		
Balanced Accuracy : 0.4966		
'Positive' Class : 0		

Cuadro 14. Matriz de confusión - Modelo Naive Bayes

Prediction	Reference	
	0	1
0	60,665	60,377
1	49,515	49,087
Accuracy : 0.4997		
95 % CI : (0.4976, 0.5018)		
No Information Rate : 0.5016		
P-Value [Acc > NIR] : 0.9663		
Kappa : -0.0010		
Mcnemar's Test P-Value : <2e-16		
Sensitivity : 0.5506		
Specificity : 0.4484		
Pos Pred Value : 0.5012		
Neg Pred Value : 0.4978		
Prevalence : 0.5016		
Detection Rate : 0.2762		
Detection Prevalence : 0.5511		
Balanced Accuracy : 0.4995		
'Positive' Class : 0		

Cuadro 15. Matriz de confusión - Modelo GBM

Prediction	Reference	
	0	1
0	60,665	60,377
1	49,515	49,087
Accuracy : 0.4997		
95 % CI : (0.4976, 0.5018)		
No Information Rate : 0.5016		
P-Value [Acc > NIR] : 0.9663		
Kappa : -0.0010		
Mcnemar's Test P-Value : <2e-16		
Sensitivity : 0.5506		
Specificity : 0.4484		
Pos Pred Value : 0.5012		
Neg Pred Value : 0.4978		
Prevalence : 0.5016		
Detection Rate : 0.2762		
Detection Prevalence : 0.5511		
Balanced Accuracy : 0.4995		
'Positive' Class : 0		

Referencias

- Anthropic (2023). Conversaciones con claude ai assistant. Utilizado para estructura de documentos LaTeX, y mejoramiento del código en R.
- Banco Mundial (2024). Trayectorias: Prosperidad y reducción de la pobreza en el territorio colombiano. Technical report, Banco Mundial, Washington, D. C., Estados Unidos de América. Publicado el 3 de diciembre de 2024.
- Castillo-Álvarez, G. and Sarmiento-Barbieri, I. (2024a). Data pre-processing: Visualizing and handling missing values (part 1). Notebook proporcionado como material del curso.
- Castillo-Álvarez, G. and Sarmiento-Barbieri, I. (2024b). Data pre-processing: Visualizing and handling missing values (part 2). Notebook proporcionado como material del curso.
- Huang, S., Shao, Y., Hu, L., Bi, J., and Ma, Z. (2025). Unveiling the spatio-temporal patterns and mechanisms of poverty incidence in southwest china through multi-sourced data and interpretable machine learning. *Applied Geography*, 176:103541.
- Huffman, C. and Nájera, H. (2024). What we talk about when we talk about measurement in poverty research. *Problemas del desarrollo*, 55(217):89–102.
- James, G., Witten, D., Hastie, T., and Tibshirani, R. (2021). *An Introduction to Statistical Learning: with Applications in R*. Springer, New York, NY, 2 edition.
- Karmaker, S. C., Rjbongshi, A., Pal, B., Sen, K. K., and Chapman, A. J. (2025). Machine learning-based prediction of energy poverty in bangladesh: Unveiling key socioeconomic drivers for targeted policy actions. *Socio-Economic Planning Sciences*, 99:102213.
- OpenAI (2023). Conversaciones con chatgpt. Utilizado para estructura de documentos LaTeX, y mejoramiento del código en R.
- Sarmiento-Barbieri, I. (2024). Limpieza de datos con tidyverse. Notebook proporcionado como material del curso.
- Sarmiento-Barbieri, I. and Castillo-Álvarez, G. (2024a). Caret para clasificación. Material del curso BDML, Universidad de los Andes. Notebook proporcionado como material del curso.
- Sarmiento-Barbieri, I. and Castillo-Álvarez, G. (2024b). Classification - cuaderno de clase. Notebook proporcionado como material del curso.
- Sarmiento-Barbieri, I. and Castillo-Álvarez, G. (2024c). Uniendo bases y calculando pobreza. Material del curso BDML, Universidad de los Andes. Notebook proporcionado como guía para el taller.
- Tal, E. (2016). How does measuring generate evidence? the problem of observational grounding. *Journal of Physics: Conference Series*, 772(1):012001.