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.

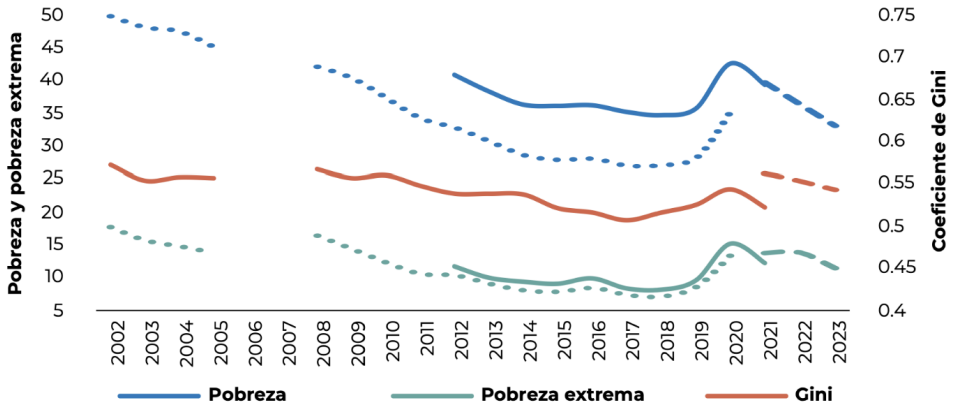
En este punto, la medición de la pobreza adquiere especial relevancia y su análisis, con el paso del tiempo, ha incluido inferencia, teoría, estadística, abstracción, idealización, predicción, cálculo y construcción de instrumentos Tal2016. Sin embargo, aún falta teoría y una estructura definida que permita a los investigadores concluir que determinada asignación numérica representa adecuadamente la característica de la pobreza que se pretende medir HuffmanNajera2024.

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 de DANE y la misión ."Empalme de las Series de Empleo, Pobreza y Desigualdad (MESE)". Estos datos son adecuados para resolver el problema de predicción de pobreza por varias razones. Primero, contienen información detallada a nivel de hogar e individuo, lo que permite capturar la heterogeneidad socioeconómica de la población colombiana. Segundo, incluyen la variable objetivo de interés (Pobre) que está correctamente definida según el criterio oficial del DANE: un hogar es clasificado como pobre si su ingreso per cápita es menor a la línea de pobreza establecida. Tercero, contienen variables sobre composición demográfica, características laborales, educativas y de vivienda que son teóricamente relevantes para explicar la pobreza.

Para este análisis, disponemos de cuatro conjuntos de datos divididos en entrenamiento y prueba a nivel de hogar e individuo: `train_hogares.csv`, `train_personas.csv`, `test_hogares.csv` y `test_personas.csv`. Esta división permite realizar predicciones

fuera de muestra, lo que representa un desafío adicional ya que algunas 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.

Verificamos que la variable objetivo *Pobre* esté correctamente definida siguiendo la metodología del DANE, mediante la comparación con cálculos propios:

$$\text{Pobre} = I(\text{Ingpcug} < \text{Lp}) \tag{1}$$

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: el 80 % de los hogares se clasifican como no pobres, mientras que el 20 % son considerados pobres, lo que requirió estrategias específicas durante el modelamiento.

Incluir tabla: Distribución de la Variable Objetivo (Pobreza)

2.2. Construcción de la muestra

El proceso de construcción de la muestra involucró varios pasos clave para garantizar datos limpios y adecuados para el modelamiento:

2.2.1. Análisis de la base train_hogares y comparación con test_hogares

Se inició el trabajo con una revisión detallada de las variables disponibles en la base de entrenamiento de hogares (`train_hogares`), comparándolas con las de la base de prueba (`test_hogares`). Este paso fue crucial para identificar variables que, al estar ausentes en la prueba, debían eliminarse del análisis o de la base consolidada. En particular, se eliminaron las siguientes variables porque contenían información directa del ingreso o porque solo existían en la base de entrenamiento:

`Ingtotug, Ingtotugarr, Ingpcug, Indigente, Npobres, Nindigentes`

Estas variables, aunque relevantes, no estaban disponibles para predicción en la base de prueba, por lo tanto, su inclusión habría implicado una filtración de información inadecuada. Además, se eliminaron variables que no aportan al análisis predictivo como:

`Directorio, Secuencia_p, Mes, P5130, Fex_c, Fex_dpto`

Las variables `P5100` y `P5140` fueron tratadas como gastos que pueden legítimamente tomar valor cero, por lo cual los valores faltantes se imputaron con ceros.

2.2.2. Análisis de la base train_personas y comparación con test_personas

Se realizó una limpieza similar en las bases de individuos. Se eliminaron variables que no están disponibles en ambas bases o que no describen características estables del hogar. En este grupo se encuentran identificadores (`Directorio, Secuencia_p`), variables temporales (`Mes`), ponderadores (`Fex_c, Fex_dpto`) y una lista extensa de variables relacionadas con ingresos individuales, que no están disponibles en la base de prueba.

Además, se excluyeron variables que no aportaban valor predictivo directo a nivel hogar, como `Orden`. Para las variables categóricas binarias que solo tomaban valores 1 y NA (como `Pet, Oc, Des, Ina`), los valores faltantes fueron transformados en ceros, con el fin de estandarizar su uso en los conteos que alimentan la base de hogares.

2.2.3. Construcción de nuevas variables para consolidar la base de hogares

Utilizamos la variable `id` como clave para unir las bases de datos de hogares e individuos. Sin embargo, esto requirió un paso previo de agregación para transformar la información a nivel de individuos en variables a nivel de hogar:

- **Conteos:** Se contó el número de miembros que cumplían ciertas condiciones (por ejemplo, número de ocupados, hombres, mujeres, menores de edad, personas en edad de trabajar, desocupados e inactivos).
- **Proporciones:** Para variables categóricas como afiliación a seguridad social (`P6090`) o deseo de más horas (`P6240`), se calcularon proporciones sobre el total de personas en edad de trabajar (`Pet = 1`).
- **Estadísticas demográficas:** Se calcularon edad promedio, máximo nivel educativo en el hogar, promedio de horas trabajadas usando variables como `P6040`, `P6210s1` y `P6800`.
- **Indicadores de seguridad social:** Proporción de afiliados, cotización a pensión.
- **Características laborales:** Proporciones por tipo de ocupación, tamaño de empresa, actividades adicionales.
- **Características específicas del jefe de hogar:** Edad, sexo, nivel educativo y situación laboral.

Esta agregación nos permitió generar 37 nuevas variables derivadas que capturan la estructura y características socioeconómicas del hogar, enriqueciendo significativamente el conjunto de datos. El enfoque utilizado garantiza que las variables incorporadas describan características estructurales del hogar y puedan ser calculadas en la base de prueba, utilizando únicamente la información disponible en ambas bases.

2.3. Limpieza de datos y tratamiento de valores faltantes

El análisis de valores faltantes reveló patrones importantes:

- La mayoría de variables tienen menos del 15 % de valores faltantes, siendo procesables con técnicas de imputación.
- Identificamos variables con alta proporción de valores faltantes (>33 %) como `jefe_tiempo_traba` y varias proporciones de características específicas que fueron candidatas a eliminación.
- Las variables relacionadas con ingresos complementarios presentaban patrones de valores faltantes no aleatorios, siendo más frecuentes en hogares no pobres.

Como parte de la limpieza, se eliminaron variables con más del 33 % de valores faltantes en la base consolidada. Esta decisión se basó en las buenas prácticas revisadas en clase y en los tutoriales de preprocesamiento de datos.

Para las variables restantes con valores faltantes, se adoptaron las siguientes estrategias:

- **Imputación con la mediana:** Para variables numéricas como `promedio_horas_trab`, `prop_cotiza_pension`, `prop_actividad_adicional` y `prop_desea_mas_horas`, se imputó la mediana dentro de cada base (train o test).
- **Imputación con la moda:** Variables categóricas como `max_nivel_educativo` y discretas como `max_años_educ` fueron imputadas con su moda, debido a su naturaleza y su importancia como predictores.
- **Para variables como `prop_afiliados_ss` y `prop_busca_trabajo`:** También se utilizó la mediana para la imputación.

Estas decisiones fueron implementadas en el script `03_data_cleaning_Fast.R`.

2.3.1. Manejo de variables categóricas

Un análisis especial se realizó para la variable `Oficio`, que presentaba más de 80 categorías diferentes. Mediante un análisis de asociación con la variable objetivo, agrupamos los oficios en tres categorías según su relación con la pobreza:

- **Grupo 1:** Oficios con baja tasa de pobreza (promedio 17 %)
- **Grupo 2:** Oficios con tasa media de pobreza (promedio 37 %)
- **Grupo 3:** Oficios con alta tasa de pobreza (promedio 59 %)

Esta agrupación simplificó el modelamiento y mejoró la interpretabilidad manteniendo la relevancia predictiva.

2.4. Análisis descriptivo

2.4.1. Distribución de la pobreza

La variable objetivo presenta un desbalance notable: el 79.9 % de los hogares se clasifican como no pobres y el 20.1 % como pobres, con un ratio de desbalance de aproximadamente 4:1.

Incluir tabla: Distribución de la Variable Objetivo (Pobreza)

2.4.2. Características por estado de pobreza

El análisis reveló diferencias significativas entre hogares pobres y no pobres:

Incluir figura: Distribución de horas trabajadas por estado de pobreza

Los hogares pobres tienden a reportar menos horas trabajadas en promedio, lo que se relaciona con condiciones laborales más precarias y menor estabilidad en el empleo.

Encontramos variables altamente correlacionadas con la pobreza, destacando:

- **Negativas (menor valor asociado a mayor pobreza):** Proporción de cotizantes a pensión (-0.51), proporción de trabajadores en empresas grandes (-0.47), y máximo nivel educativo (-0.39).
- **Positivas (mayor valor asociado a mayor pobreza):** Proporción de oficios del grupo 3 (0.43), número de menores (0.37) y proporción de inactivos (0.31).

Incluir figura: Variables con mayor correlación con la pobreza

Este análisis de correlaciones ayudó a identificar las variables más relevantes para la predicción, permitiéndonos crear un conjunto parsimonioso y predictivo.

2.4.3. Importancia de la composición familiar y laboral

El análisis chi-cuadrado para variables categóricas reveló asociaciones significativas entre pobreza y diversas características:

- La presencia de jefes de hogar ocupados reduce significativamente la probabilidad de pobreza (V de Cramér = 0.23)
- Los hogares con mayor proporción de miembros en oficios del grupo 3 (alta tasa de pobreza) tienen mayor probabilidad de ser pobres (V de Cramér = 0.31)
- El nivel educativo del jefe de hogar muestra una fuerte asociación con la pobreza (V de Cramér = 0.27)

Estos hallazgos confirman la importancia de variables relacionadas con capital humano, estructura familiar y características laborales para predecir la pobreza.

2.5. Justificación de la selección de variables

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

1. **Relevancia predictiva:** Utilizamos correlaciones con la variable objetivo y pruebas chi-cuadrado para identificar las variables más predictivas.
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 mínimo de variables que maximizara el poder predictivo.

Eliminamos variables con más del 33 % de valores faltantes y aquellas con alta redundancia (correlacionadas entre sí a más de 0.7). Para variables con correlación alta, mantuvimos la que presentaba mayor asociación con la variable objetivo.

En el caso de variables categóricas como la ocupación, optamos por transformaciones que preservaran su poder predictivo al tiempo que simplificaban el modelamiento (agrupación en 3 categorías).

Las variables finales incluyeron:

- **Características del hogar:** Número de miembros, relación de dependencia, proporción de ocupados.
- **Capital humano:** Nivel educativo máximo y del jefe de hogar, proporción de afiliados a seguridad social.
- **Características laborales:** Distribución por grupos de ocupación, horas trabajadas, tamaño de empresa.
- **Características del jefe de hogar:** Sexo, edad, ocupación.
- **Vivienda y ubicación:** Características de la vivienda, departamento.

El flujo de trabajo seguido en esta etapa responde a varios objetivos:

- **Evitar fuga de información:** Se eliminaron variables disponibles solo en el entrenamiento.
- **Garantizar la coherencia entre bases:** Se preservaron únicamente las variables comunes entre train y test.

- **Convertir la información individual en predictores agregados:** Esto permite maximizar el uso de la información disponible sin violar las restricciones impuestas por la estructura del problema.
- **Preparar una base limpia y funcional para modelado:** Las bases `train_cleaned.csv` y `test_cleaned.csv` contienen las variables seleccionadas e imputadas, listas para el entrenamiento de modelos.

Este conjunto final de variables balanceó el poder predictivo, la disponibilidad en los datos de prueba y la parsimonia, permitiéndonos construir modelos robustos y generalizables para la predicción de pobreza.

2.6. Análisis descriptivo

Exploración de la variabilidad de los datos con tablas y gráficos.

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, "pobreçomo 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 "pobreçuando 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:

$$P\left(y=1|X\right)=\frac{1}{1+e^{-(\beta_0+\beta_{1+\ldots+\beta_k})}}$$

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:

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

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, el esta metodologia 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:

$$P\left(Y|X_1, X_2, \ldots, X_k\right) = P(Y) \prod P(X_J|Y)$$

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 **??** 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 afortunamente 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 **??** 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 **??** 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 importe es que este fue el único modelo que se estimo con todas las variables, el resto fueron estimados siguiendo otra especificación (Pobre ~ hacinamiento + r_gast+ hacinamiento_c + TGP +educ_cab +trabajando +CotizaPension + Subsidiado + Pequena + Subsidios + CotizaPension + Pensionado + Ingresos_AlquilerPensiones + OtrosIngresos) esto debido a que no todas las variables aportaban poder explicativo.

La Tabla **??** 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 cp. 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 deespecificación.

La Tabla ?? 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 alpha y lambda. 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 alpha (que varió de 0.10 a 1.00) y lambda (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 ?? del Anexo 8 presenta los resultados del modelo ajustado por la metodología de Naive Bayes que fue ajustado con los hiperparámetros alpha y lambda. 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 ?? 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 **??** del Anexo 10 se muestra la matriz de confusión para el primer modelo CART.

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

La Tabla **??** 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 **??** del Anexo 13.

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

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

3.4. Importancia de Variables

[Esta sección queda pendiente para incluir un análisis de la importancia relativa de las variables en los modelos]

4. Conclusión

Resumen de los hallazgos principales y posibles mejoras futuras.

Cuadro 2. 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
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

Cuadro 4. 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
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

Anexos

A. Estadísticas descriptivas de variables continuas

B. Estadísticas descriptivas de variables discretas

C. Variables utilizadas en los modelos

D. Estimaciones mediante metodología LOGIT

E. Estimación 1 metodología CARTs

F. Estimación 2 metodología CARTs

G. Estimaciones por Elastic Net

H. Estimaciones por Naive Bayes

I. Estimaciones por GBM

J. Matriz de confusión - Modelo CART 1

K. Matriz de confusión - Modelo CART 2

L. Matriz de confusión - Modelo Elastic Net

M. Matriz de confusión - Modelo Logit

N. Matriz de confusión - Modelo Naive Bayes

Ñ. Matriz de confusión - Modelo GBM

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.

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.

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.