Big Data y Machine Learning para economía aplicada

Problem set 2 – Equipo 10

Natalia Buitrago Buitrago

Nicolás David Mocetón Herrera

Juan Felipe Otálora Cedeño

I. Introducción

La pobreza continúa siendo uno de los principales desafíos estructurales en las economías en desarrollo, al reflejar desigualdades persistentes en el acceso a la educación, el empleo y los activos productivos. En el caso colombiano, la Encuesta de Medición de Pobreza Monetaria y Desigualdad (2018) del Departamento Administrativo Nacional de Estadística (DANE) proporciona información detallada que permite identificar las condiciones socioeconómicas de los hogares y cuantificar su situación frente a la línea oficial de pobreza.

El objetivo central de este trabajo es construir un modelo predictivo capaz de clasificar a los hogares como pobres o no pobres a partir de sus características demográficas, educativas, laborales y habitacionales. En particular, el análisis adopta un enfoque de clasificación supervisada, en el cual la variable objetivo `Pobre` se define como una función indicadora que toma el valor uno cuando el ingreso per cápita del hogar se encuentra por debajo de la línea de pobreza y cero en caso contrario.

A través de un conjunto de modelos estadísticos y de aprendizaje automático, como regresión logística, Elastic Net, Random Forest, CART y Naive Bayes, se busca no solo alcanzar un buen desempeño en métricas de precisión y sensibilidad, sino también identificar los determinantes estructurales más relevantes de la pobreza monetaria.

Los resultados muestran que los modelos con regularización, en particular Elastic Net, logran el mejor equilibrio entre poder predictivo e interpretabilidad. Las variables asociadas con educación, empleo, género del jefe de hogar y régimen de salud surgen como los predictores más importantes, confirmando patrones consistentes con la literatura económica. En conjunto, el estudio demuestra el potencial del análisis predictivo como herramienta complementaria para el diseño de políticas públicas orientadas a la reducción de la pobreza.

II. Datos

2.1. Procesamiento de datos

Los datos empleados en este estudio provienen de la Encuesta de Medición de Pobreza Monetaria y Desigualdad (2018) del Departamento Administrativo Nacional de Estadística (DANE). Esta fuente oficial es la base utilizada por el Gobierno colombiano para estimar los indicadores nacionales de pobreza y desigualdad, por lo que su cobertura y calidad la hacen especialmente adecuada para abordar la pregunta predictiva del presente trabajo: clasificar a los hogares colombianos como pobres o no pobres a partir de sus características observables.

La variable dependiente, Pobre, es un indicador binario que toma el valor Yes si el ingreso per cápita del hogar se encuentra por debajo de la línea de pobreza nacional y No en caso contrario.

La construcción de la muestra partió de dos fuentes complementarias: bases de datos a nivel hogar y persona, disponibles tanto para el conjunto de entrenamiento como para el de prueba (train_hogares, train_personas, test_hogares, test_personas). El procesamiento se realizó en R, utilizando paquetes del ecosistema tidyverse, junto con herramientas de modelado y validación como caret, glmnet, MLmetrics y pROC.

El proceso de preparación de los datos incluyó cuatro etapas principales:

Preprocesamiento individual (preprocess_personas): Se generaron indicadores binarios para género, edad y estado laboral. En particular, se definieron las variables mujer, menor_de_edad, adulto_mayor, ocupado, desocupado e inactivo, además de atributos laborales como segundo_empleo, tipo_primer_empleo y recibio_horasextra. Se normalizó la codificación de educación (educLevel) y del régimen de salud (regimen_salud), y se corrigieron valores faltantes según criterios del DANE.

Agregación a nivel hogar (preprocess_personas_agregacion): Se sumó la información individual por hogar, obteniendo totales como número de mujeres, menores, adultos mayores, ocupados e inactivos. Además, se mantuvieron las características del jefe de hogar como género, edad, educación, empleo, tipo de régimen de salud

Procesamiento de la base de hogares (preproces_hogares): Se creó la variable arrienda, que identifica si el hogar paga arriendo, y se mantuvieron identificadores de clase y dominio geográfico.

Integración y estandarización (merge y mutate): Las bases de personas y hogares se integraron mediante el identificador id, generando las bases finales train_ready y test_ready. Se transformaron las variables relevantes en factores con etiquetas descriptivas y se imputaron valores faltantes en Jefe_regimen_salud, asignando por defecto la categoría Subsidiado.

El conjunto de entrenamiento quedó conformado por 164.960 hogares, mientras que el conjunto de prueba incluye 66.168 hogares, manteniendo proporciones muy similares en todas las variables clave. Este diseño garantiza que los modelos puedan generalizar correctamente, evitando sesgos muestrales.

2.2. Análisis descriptivo de los datos

Tabla 1. Estadísticas descriptivas – Conjunto de entrenamiento train

Variable	N	Frecuencias
Pobre	164960	No: 131936, Yes: 33024
arrienda	164960	No: 100507, Yes: 64453
maxEducLevel	164960	Universitario: 80385, Media: 47501, Secundaria: 18111, Primaria: 16492
Jefe_H_mujer	164960	No: 95959, Yes: 69001
Jefe_desocupado	164960	No: 157236, Yes: 7724
Jefe_regimen_salud	164960	Contributivo: 79620, Especial: 76162, Subsidiado: 9178
Jefe_Tipo_primer_empleo	164960	Obr: 87148, Tra: 58084, Obr: 7673, Pat: 6413
Jefe_segundo_empleo	164960	No: 158149, Yes: 6811

El 20% de los hogares (33.024 de 164.960) se clasifican como pobres (Pobre = Yes), 39% de los hogares arrienda su vivienda, lo que refleja una menor acumulación patrimonial y mayor vulnerabilidad económica. Respecto al nivel educativo máximo (maxEducLevel), el 49% de los hogares cuenta con al menos un miembro universitario, mientras que un 22% alcanza solo educación primaria o básica secundaria. Los hogares con jefatura femenina (Jefe_H_mujer = Yes) representan 42% de la muestra, grupo que tiende a exhibir mayor riesgo de pobreza, especialmente en presencia de hijos menores, tan solo el 5% de los jefes de hogar están desempleados, para el régimen de salud el 48% de los hogares hacen parte del régimen contributivo, 46% en regímenes especiales y 6% en el subsidiado. Finalmente, Jefe_Tipo_primer_empleo evidencia trayectorias laborales desiguales: el 53% inició como obrero o empleado particular, el 35% como trabajador independiente, y un 4% como empleador y el 7% de los jefes tienen un segundo empleo.

Tabla 2. Estadísticas descriptivas – Conjunto de prueba test

Variable	N	Frecuencias
arrienda	66168	No: 40967, Yes: 25201
maxEducLevel	66168	Universitario: 31911, Media: 18931, Secundaria: 7446, Primaria: 6844
Jefe_H_mujer	66168	No: 38643, Yes: 27525
Jefe_desocupado	66168	No: 63152, Yes: 3016
Jefe_regimen_salud	66168	Subsidiado: 31404, Contributivo: 30915, Regimen Especial: 3849
Jefe_Tipo_primer_empleo	66168	Obr: 34118, Tra: 23999, Obr: 3088, Pat: 2592
Jefe_segundo_empleo	66168	No: 63414, Yes: 2754

La estructura del conjunto de prueba es casi idéntica a la del entrenamiento: la proporción de hogares que arriendan es 38%, el nivel educativo máximo universitario alcanza 48%, los hogares con jefatura femenina representan 41% y la proporción de desempleo del jefe del hogar es cercana al 4,6%.

Estas similitudes confirman que la partición de la muestra fue aleatoria y que la base de prueba conserva la representatividad poblacional, condición necesaria para evaluar adecuadamente el desempeño predictivo de los modelos.

III. Modelos y Resultados

Esta sección detalla el proceso de modelado predictivo, la calibración de hiperparámetros y la evaluación comparativa de ocho modelos desarrollados para clasificar el estado de pobreza de los hogares. El análisis integra enfoques estadísticos y de aprendizaje automático — Elastic Net, Regresión Logística (Logit), Árboles de Clasificación y Regresión (CART), Bosques Aleatorios (Random Forest) y Naive Bayes —, todos estimados bajo un marco de validación consistente.

Todos los modelos fueron entrenados sobre la misma base de datos depurada, aplicando validación cruzada de cinco pliegues (five-fold cross-validation), utilizando el F1-score como métrica principal de evaluación. Este indicador se prefirió sobre la exactitud (accuracy) debido al desbalance de clases existente, dado que los hogares pobres representaban aproximadamente el 38% del total de la muestra. Los parámetros de cada modelo se optimizaron buscando el mejor equilibrio entre recall (sensibilidad hacia la clase pobre) y precision (exactitud de las predicciones positivas).

3.1 Selección y entrenamiento de modelos

La regresión Elastic Net sirvió como modelo base, combinando las ventajas de la regularización Lasso y Ridge para manejar la multicolinealidad y mejorar la generalización. Se estimaron dos especificaciones, el primer modelo utilizó una grilla amplia con valores de α de 0, 0.5 y 1, y λ entre 10^{-3} y 10^{-1} y el segundo modelo empleó una grilla más refinada con α entre 0.15 y 1.0 y λ entre 10^{-4} y 10^{-2}.

La mejor configuración en el primer modelo (EN_1) fue $\alpha = 0.5$ y $\lambda = 0.0027$, mientras que el segundo (EN_2) alcanzó su mejor desempeño con $\alpha = 0.35$ y $\lambda = 0.0023$. Estos valores reflejan una regularización moderada, evitando una reducción excesiva de los coeficientes y reteniendo predictores relevantes. Los F1 obtenidos fueron 0.63 y 0.64, respectivamente — los más altos entre todos los modelos probados.

Los modelos de Regresión Logística sirvieron como referencia interpretable.
El modelo estándar usó el umbral convencional de 0.5, mientras que una segunda versión optimizada aplicó un umbral de 0.21, derivado de la curva ROC para maximizar el F1-score. Este ajuste incrementó el desempeño del modelo de F1 = 0.58 a 0.61, mejorando significativamente el recall para la clase minoritaria (hogares pobres).

Tanto los modelos Elastic Net como los logísticos incluyeron interacciones como “jefatura femenina × número de menores” y “jefatura femenina × desempleo”, buscando capturar heterogeneidades sociales relevantes. Estas interacciones resultaron particularmente significativas, pues el género y la composición familiar mostraron fuerte influencia en el riesgo de pobreza.

El modelo CART aplicó particionamiento recursivo para identificar fronteras jerárquicas de decisión. Tras ajustar el parámetro de complejidad (cp) en un rango de 0 a 0.005, el valor

óptimo fue cp = 0.0012, que minimizó el sobreajuste preservando la interpretabilidad. Su desempeño (F1 = 0.54) fue inferior en comparación con los modelos regulares o en ensamble, reflejando las limitaciones de los árboles individuales en entornos con alta dimensionalidad.

Se estimaron dos modelos Random Forest (RF). El primero (RF_1) utilizó mtry = 11, min.node.size = 60, y 300 árboles, mientras que el segundo (RF_2) empleó mtry = 9, min.node.size = 55, y 200 árboles para reducir el tiempo de cómputo. Ambos obtuvieron resultados competitivos: F1 = 0.61 para RF_1 y F1 = 0.60 para RF_2.

Finalmente, el modelo Naive Bayes proporcionó una referencia probabilística. Su mejor configuración — Laplace = 1, Kernel = TRUE, Adjust = 1.5 — alcanzó un F1 = 0.52.

3.2 Evaluación y desempeño de los modelos

La Figura 1 muestra los F1-scores de todos los modelos, tanto en entrenamiento como en validación cruzada (out-of-fold). Los resultados confirman que los modelos Elastic Net y Random Forest lograron el desempeño más estable entre pliegues, seguidos de cerca por la Regresión Logística optimizada. El modelo Elastic Net (EN_2) alcanzó el F1 más alto (0.64), seguido de EN_1 (0.63), RF_1 (0.61) y el Logit optimizado (0.61). Modelos como CART y Naive Bayes mostraron mayor dispersión entre los F1 de entrenamiento y validación, sugiriendo un leve sobreajuste o simplicidad estructural.

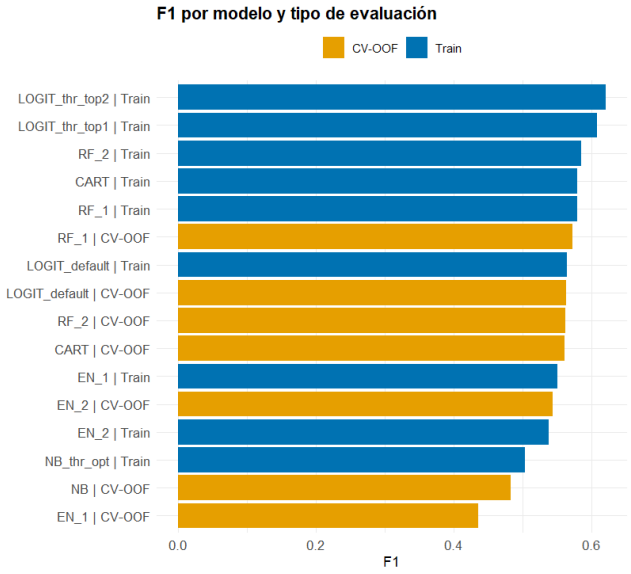

Figura 1. F1-scores por modelo y tipo de evaluación.
(CV-OOF = validación cruzada; Train = desempeño en entrenamiento.)

En general, los F1 de entrenamiento (en azul) superaron a los de validación (en naranja), como era esperable. Sin embargo, la brecha fue moderada en Elastic Net y Random Forest, evidenciando buena capacidad de generalización. CART presentó una caída más pronunciada, señalando ajuste excesivo a los patrones del conjunto de entrenamiento. Cabe destacar que el Logit con umbral optimizado alcanzó resultados prácticamente idénticos al mejor Elastic Net, demostrando que los modelos simples, bien calibrados, pueden igualar el rendimiento de técnicas más complejas.

3.3 Distribución de Probabilidades y Calibración

- ***Elastic Net***

Las **Figuras 2 y 3** presentan los histogramas de probabilidad para EN_1 y EN_2.
Ambos muestran una separación clara entre clases: los hogares pobres (azul) se concentran en probabilidades más altas, mientras que los no pobres (naranja) se agrupan cerca de cero.
No obstante, existe un solapamiento notable en el rango 0.05–0.30, reflejando el continuo socioeconómico entre hogares vulnerables, pero no pobres y hogares oficialmente pobres.

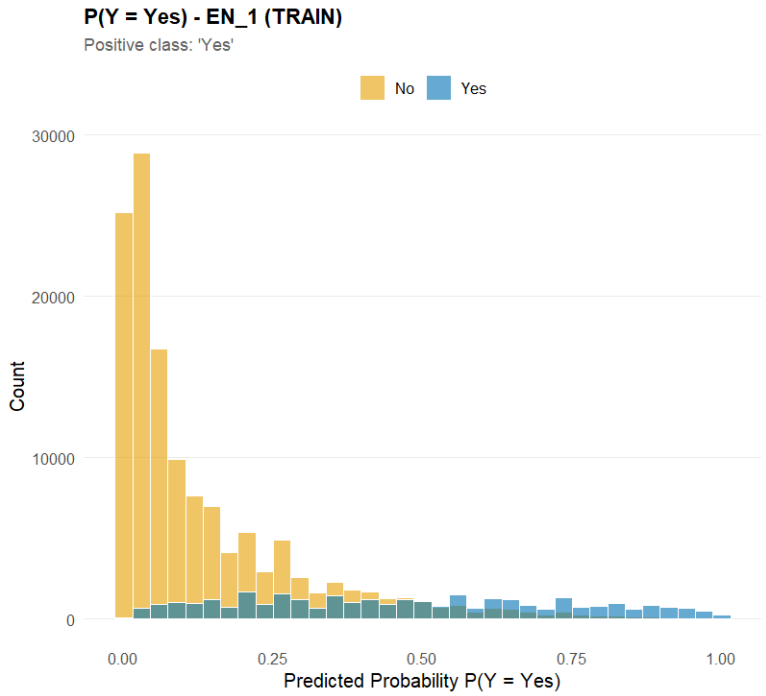

Figura 2. Predicciones de probabilidad para EN_1 (TRAIN).

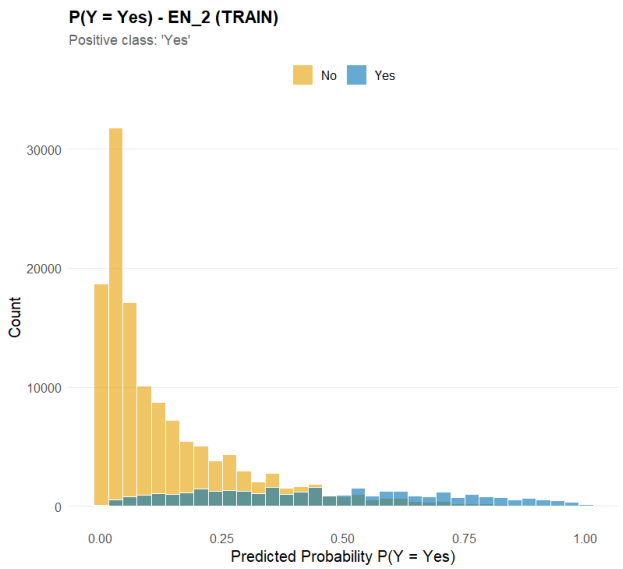

Figura 3. Predicciones de probabilidad para EN_2 (TRAIN).

El modelo EN_2 muestra una separación más nítida y una calibración más suave que EN_1, confirmando que su grilla más fina (menor α y λ) mejoró la discriminación sin perder estabilidad.

- ***Logit***

La Figura 4 muestra la distribución del modelo logístico base. Las probabilidades predichas se concentran cerca de cero para los hogares no pobres, mientras que los hogares pobres se distribuyen en rangos medios.

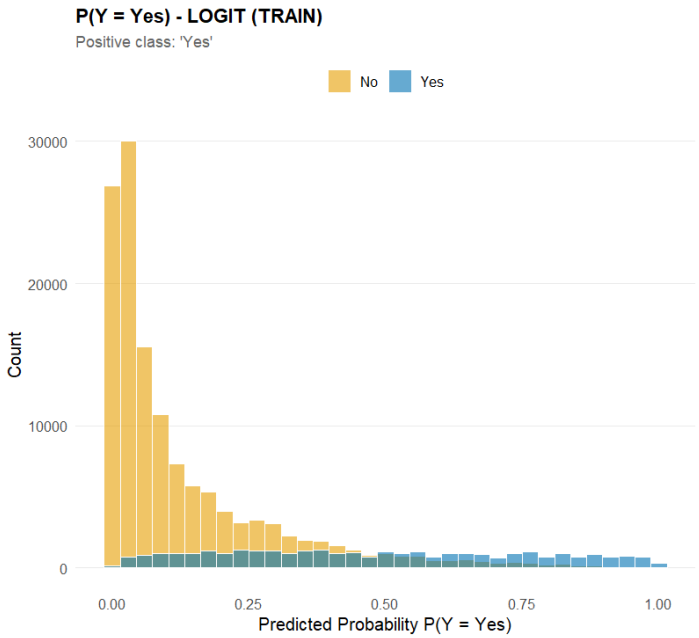

Figura 4. Predicciones de probabilidad para Logit (TRAIN).

Esta forma explica la sensibilidad del modelo al umbral de decisión: el corte de 0.5 tiende a subestimar los casos limítrofes, pero el umbral optimizado de 0.21 captura mejor dichos hogares, elevando el F1 de 0.58 a 0.61.
El modelo logístico sigue siendo un punto de referencia robusto y de alta interpretabilidad para el análisis de políticas.

- ***Random Forest***

La Figura 5 presenta los resultados del Random Forest (RF_1). La distribución de probabilidades muestra diferenciación entre clases, aunque menos marcada que en Elastic Net. El modelo asigna a muchos hogares pobres probabilidades entre 0.2 y 0.6, reflejando su capacidad para capturar relaciones no lineales, aunque con una calibración algo conservadora.

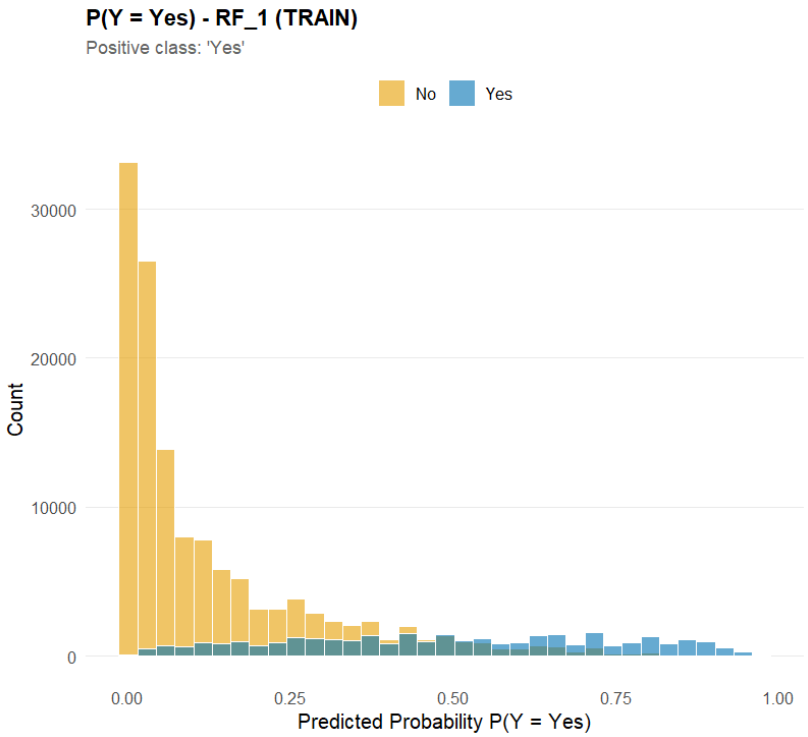

Figura 5. Predicciones de probabilidad para Random Forest (TRAIN).

El rango más amplio de probabilidades indica que el modelo detecta relaciones multidimensionales complejas, a costa de menor interpretabilidad. Aun así, su alto F1 (0.61) demuestra su fiabilidad predictiva y su valor como complemento de los modelos regularizados.

- ***CART***

La Figura 6 muestra la distribución del modelo CART. A diferencia de los métodos en ensamble, CART genera probabilidades discretas que corresponden a las hojas terminales del

árbol. Muchas observaciones se agrupan en probabilidades extremas (cercanas a 0 o 0.75), produciendo una distribución discontinua.

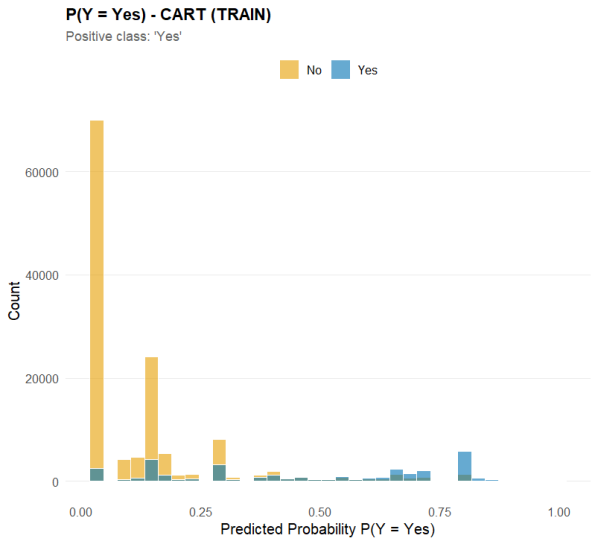

Figura 6. Predicciones de probabilidad para CART (TRAIN).

Este patrón explica su F1 relativamente bajo (0.54): aunque captura tendencias amplias, su estructura por tramos dificulta representar probabilidades intermedias, reduciendo la calibración.

- ***Naive Bayes***

La Figura 7 representa las predicciones del modelo Naive Bayes. Los hogares no pobres se concentran cerca de 0, pero los pobres exhiben una cola larga hacia la derecha con probabilidades moderadas o altas.

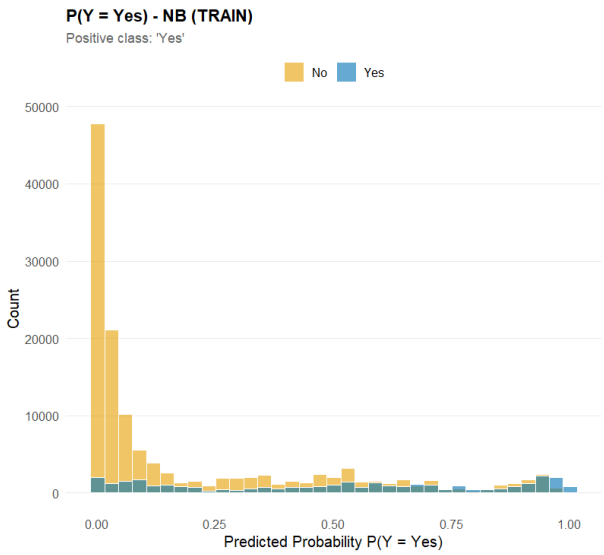

Figura 7. Predicciones de probabilidad para Naive Bayes (TRAIN).

Su forma más suave, respecto a CART, refleja la naturaleza probabilística del modelo y su suavizado mediante kernel density. Sin embargo, el supuesto de independencia entre variables limita su precisión en datos socioeconómicos altamente correlacionados, explicando su menor F1 (0.52).

3.4 Feature importance

El análisis de importancia de variables en los mejores modelos (Elastic Net y Random Forest) muestra un conjunto coherente de determinantes clave de la pobreza. Las variables educativas fueron las más influyentes. Les siguieron las relacionadas con el mercado laboral, como el número de ocupados en el hogar y la condición de desempleo del jefe o jefa. Factores demográficos como el tamaño del hogar, el número de menores y el género de la jefatura también tuvieron peso relevante. Los hogares con jefatura femenina y dependientes pequeños mostraron sistemáticamente mayores probabilidades de pobreza. Las condiciones de vivienda (arriendo, informalidad en la tenencia) y la afiliación al sistema de salud tuvieron efectos moderados pero consistentes.

Los puntajes de importancia por impureza en el Random Forest confirmaron estos resultados, mientras que los coeficientes del Elastic Net aportaron dirección: mayor educación y empleo reducen el riesgo de pobreza, mientras que el desempleo, la mayor carga familiar y la jefatura femenina lo incrementan. La consistencia entre algoritmos respalda la solidez de estos determinantes.

IV. Conclusiones

Los resultados combinados de los modelos estadísticos y de machine learning demuestran que el modelado predictivo puede capturar eficazmente los determinantes estructurales de la pobreza en los hogares colombianos. El Elastic Net alcanzó el desempeño más alto y estable (F1 = 0.64), seguido del Random Forest (F1 = 0.61) y de la Regresión Logística optimizada (F1 = 0.61). El equilibrio entre regularización y flexibilidad fue determinante. Una penalización moderada ($\alpha \approx 0.35$, $\lambda \approx 0.0023$) en Elastic Net controló el sobreajuste sin sacrificar las variables socioeconómicas más relevantes. El Random Forest mejoró el recall y la robustez gracias a su estructura en ensamble, aunque con menor interpretabilidad.
El modelo logístico, en cambio, mostró que un enfoque clásico y transparente puede igualar el rendimiento de métodos más sofisticados cuando se calibra adecuadamente.

Las Figuras 1–7 reflejan este equilibrio entre poder predictivo e interpretabilidad: los modelos con distribuciones de probabilidad más suaves y brechas menores entre entrenamiento y validación (Elastic Net, Logit, Random Forest) fueron los de mejor generalización. Modelos más simples (CART, Naive Bayes) mostraron picos de probabilidad discretos y sobreconfianza, limitando su desempeño. En conclusión, el análisis confirma que la educación, la participación laboral y la composición familiar son los principales predictores de pobreza. Los modelos regularizados como Elastic Net ofrecen el marco más confiable e interpretable para predecir el riesgo de pobreza, apoyando el diseño de políticas públicas y programas sociales basados en evidencia.

Referencias

- **Departamento Administrativo Nacional de Estadística – DANE.** (2019). *Colombia – Medición de Pobreza Monetaria y Desigualdad 2018.* Gobierno Nacional de Colombia. Reporte generado el 12 de julio de 2019. Disponible en: https://microdatos.dane.gov.co/

Github Repository: https://github.com/nbuitrago23/PS2_Equipo_10