Problem Set 2[1]

Nicolás Moreno[2]
Mateo Isaza Díaz[3]
María Camila Caraballo [4]
Javier Antonio Amaya [5]

1 Introducción

Para el año 2024, se estima que más de 16 millones de personas en Colombia viven en situación de pobreza, según datos del Banco Mundial. Este panorama representa un retroceso en los avances sociales logrados durante la última década, debido al impacto de la pandemia, que revirtió más de diez años de progreso en la reducción de la pobreza y la pobreza extrema. Aunque se evidencian señales de recuperación, este proceso aún no se ha consolidado Banco Mundial (2024). Autores como Lipton and Ravallion (1995) sostienen que factores como la ubicación geográfica del hogar, la distribución del ingreso y el nivel educativo son determinantes clave para la generación de ingresos y, por tanto, para el riesgo de caer en la pobreza.

En este contexto, las transferencias monetarias de los programas sociales, así como las decisiones de política pública y la asignación de subsidios orientados a mitigar condiciones de vulnerabilidad, se fundamentan en el uso de modelos econométricos. Estos modelos permiten analizar las asociaciones entre diversos factores a nivel del hogar y características individuales de sus integrantes, con el objetivo de determinar la probabilidad de que una persona sea clasificada como pobre Haughton (2009). La correcta focalización de los recursos públicos es crucial, dado que asignar transferencias a personas que no se encuentran en situación de pobreza representa un costo social elevado, especialmente en un escenario marcado por el retroceso en los avances sociales ocasionado por la pandemia Vallejo Zamudio (2021).

Una contribución significativa del aprendizaje automático al estudio de la pobreza fue la competencia organizada por el Banco Mundial, cuyo objetivo fue identificar el algoritmo más preciso para predecir la pobreza en tres países distintos Banco Mundial and DrivenData (2020). Con este enfoque, otras investigaciones han implementado técnicas como los bosques aleatorios, por ejemplo, en el caso de Tailandia en 2016 Pave and Stender (2017), o en Costa Rica en 2019, donde se utilizó una encuesta de hogares para clasificar a la población en cuatro categorías según umbrales específicos: pobreza extrema, pobreza, vulnerabilidad y no pobreza Solís-Salazar (2022).

[1]El repositorio a este proyecto lo puede encontrar en este link
[2]Código: 201615907
[3]Código: 202412526
[4]Código: 201613424
[5]Código: 202214392

Tomando como referencia estas investigaciones y en consideración del contexto actual de Colombia, resulta fundamental abordar el problema de la clasificación de hogares según su situación de pobreza, con el propósito de asignar de manera más eficiente los recursos públicos destinados a satisfacer necesidades básicas. Para este análisis en particular, se emplea la base de datos de la Encuesta de Medición de Pobreza Monetaria y Desigualdad del Departamento Administrativo Nacional de Estadística (DANE) correspondiente al año 2018. Esta base incorpora la metodología más reciente desarrollada por la Misión de Empalme de las Series de Empleo (MESEP), la cual mide la pobreza monetaria mediante la definición de líneas de pobreza e indigencia, a partir del ingreso per cápita disponible en los hogares DANE (2018).

2 Datos

Los datos que tuvimos a nuestra disposición provienen de la Encuesta de Medición de Pobreza Monetaria y Desigualdad del DANE, y se estructuraron en cuatro bases de datos: dos a nivel individual (una para entrenamiento y otra para test) y dos a nivel de hogar, también separadas para estos mismos fines. La principal diferencia entre las bases de entrenamiento y test fue que las primeras incluían la variable Pobre, así como información detallada sobre los ingresos individuales y del hogar. Además, las bases de entrenamiento contenían múltiples variables continuas que no estaban presentes en las de test, como ingresos por arriendos, conceptos extra-laborales, remesas, entre otros.

No obstante, dado que las bases de test incluían variables dicótomas relacionadas con los ingresos, fue posible incorporarlas en las fases de entrenamiento con el fin de construir modelos más robustos y con mayor capacidad predictiva. A nivel individual, seleccionamos la variable de nivel educativo y, con base en esta, asignamos a cada hogar una categoría según el nivel educativo alcanzado por sus miembros: se asignó un valor de 0 si ningún integrante había alcanzado al menos el nivel de bachillerato; un valor de 1 si al menos un miembro había finalizado el bachillerato, pero ninguno contaba con educación superior; y un valor de 2 si al menos un integrante del hogar tenía formación universitaria o superior. Asimismo, se creó una variable que indicaba la condición laboral del hogar: se asignó el valor de 1 si al menos un miembro era parte de la población ocupada y 0 en caso contrario. Estas decisiones metodológicas se basaron en la premisa de que, en general, un mayor nivel educativo y la participación en el mercado laboral están asociados con mayores niveles de ingreso, lo cual puede reducir la probabilidad de que un hogar se encuentre en situación de pobreza.

Asimismo, se consideraron variables asociadas a ingresos laborales adicionales, tales como el pago de horas extras, auxilio de transporte, primas de servicios, de Navidad y de vacaciones, viáticos y cotización a fondos de pensiones. Con base en estas, se construyó una variable dicótoma que toma el valor de 1 si al menos un miembro del hogar reportaba recibir alguno de estos ingresos y 0 en caso contrario. Estos conceptos corresponden a beneficios asociados al empleo formal, y su presencia se asume como un indicador de mayor estabilidad y privilegio económico dentro del hogar. Esta decisión se fundamenta en el hecho de que, para el año 2018, poco más de la mitad de las personas ocupadas en Colombia se encontraban vinculadas formalmente al mercado laboral, lo cual refuerza la idea de que estas prestaciones son distintivas de una mejor posición económica relativa DANE (2018).

Se incluyó una variable que refleja el promedio de antigüedad, en meses, en el empleo

actual de los miembros del hogar que se encontraban ocupados. Esta variable busca capturar la estabilidad laboral dentro del hogar, bajo la premisa de que una mayor permanencia en el puesto de trabajo puede estar asociada con mayor seguridad económica. Adicionalmente, se incorporaron variables dicotómicas basadas en las categorías ocupacionales reportadas por los miembros del hogar. Por ejemplo, si al menos un integrante se desempeñaba como empleado público, al hogar se le asignaba un valor de 1, y 0 en caso contrario. Se aplicó el mismo criterio para otras categorías como trabajadores independientes, empleados domésticos o jornaleros, entre otros. Finalmente, se incluyó una variable continua que representa la proporción de mujeres dentro del hogar, con el objetivo de explorar posibles asociaciones entre la composición de género y las condiciones socioeconómicas.

Con el objetivo de capturar fuentes de ingreso no laborales, se incluyeron variables dicotómicas a nivel de hogar que tomaban el valor de 1 cuando al menos uno de sus integrantes reportaba ingresos provenientes de arriendos, remesas u otras fuentes, y 0 en caso contrario. Posteriormente, tras la construcción y agregación de las variables mencionadas, se procedió a integrar la base de datos resultante con la base original de hogares, utilizando la variable "id" como llave primaria. Esta fusión permitió incorporar información adicional relevante, como el valor de la cuota de amortización para hogares con vivienda financiada, el monto del arriendo pagado por hogares arrendatarios y la estimación del arriendo imputado en el caso de hogares propietarios. Estas variables complementaron los modelos al aportar datos sobre ingresos potenciales o gastos asociados a la tenencia de la vivienda, ofreciendo así una visión más integral de las condiciones económicas de los hogares.

Table 1: Estadísticas descriptivas para variables continuas

Variable	**Media**	**Desv. Est.**	**Mín.**	**Mediana**	**Máx.**
Ingreso hogar	2090895	2512488	0	1400000	85833333
No. de personas	3.292	1.775	1	3	28
No. Habitaciones	3.390	1.239	1	3	98
No. Dormitorios	1.989	0.898	1	2	15
Crédito vivienda	31.38	1141.65	0	0	280000
Valor arriendo	304.54	3258.73	0	100	600000
Arriendo estim.	171.1	929.7	0	0	300000
Antigüedad laboral	128.995	167.849	0	60	2077
Promedio de antigüedad	77.851	100.343	0	36.5	948
Media de edad	37.440	16.876	5.667	33.5	102
Proporción de mujeres	52.575	27.724	0	50	100

Nota. Número de observaciones: 16,542 para todas las variables. La variable dependiente es *Pobre*, dicotómica, que toma valor de 1 para hogares con ingresos promedio por debajo de la línea de pobreza y 0 de lo contrario. Las variables *Crédito vivienda*, *Valor arriendo* y *Arriendo estim.* están expresadas en miles de pesos.

3 Modelos

Esta sección presenta un análisis comparativo de los modelos desarrollados, con énfasis en el modelo que mostró el mejor desempeño predictivo frente a las demás aproximaciones implementadas. Asimismo, se describen los procesos de definición y ajuste de hiperparámetros realizados para cada uno de los algoritmos empleados. El propósito central de cada modelo es construir una herramienta predictiva robusta que permita clasificar correctamente a los hogares por debajo del umbral de pobreza y a aquellos que no lo están.

La intención final es aplicar este modelo a datos externos, no utilizados en el entrenamiento, con el fin de mejorar la identificación de hogares en situación de pobreza y contribuir así a una asignación más eficiente y focalizada de los recursos públicos. Motivados por la competencia organizada por el Banco Mundial para predecir la pobreza, en este ejercicio académico se seleccionó el modelo que obtuvo el mejor desempeño en la plataforma Kaggle, con el objetivo de responder cuál es el modelo más adecuado para esta tarea.

3.1 Selección de modelos y entrenamiento

Modelo XGBoost y GLMNET con balanceo de clases mediante SMOTE

El modelo que logró el mejor F1 Score (0.7237) combinó XGBoost y GLMNET en un ensamble ponderado (70%-30%), optimizado con 3-Fold CV y balanceo de clases mediante SMOTE. Se aplicó feature engineering con nuevas variables como ing_per_capita y renta_vs_educ, y se estandarizaron los datos. Para manejar el desbalance (20% clase "Pobre"), se usó SMOTE (K=5, dup_size=3) junto con scale_pos_weight=4 en XGBoost, mejorando el recall en 29%. Los hiperparámetros se ajustaron con grid search: XGBoost usó un learning rate bajo (n=0.05) y profundidad controlada (max_depth=8), mientras que GLMNET empleó regularización Elastic Net (a=0.2). El threshold óptimo se determinó en 0.45 (frente al default 0.5) para maximizar el F1.

Las transformaciones incluyeron log1p para variables de ingresos y one-hot encoding para categóricas. El código implementó early stopping en XGBoost (50 rondas sin mejora) y validación cruzada estratificada para garantizar representatividad. La selección final del modelo se basó en el F1 Score (no solo AUC), priorizando el equilibrio entre precision y recall. Las predicciones se generaron combinando las probabilidades de ambos modelos y aplicando el threshold óptimo. Esta estrategia demostró que abordar explícitamente el desbalance y optimizar el threshold son clave en problemas con clases desiguales.

El análisis de relevancia revela que reg_cotiz1 (18.7% de Gain) y cotiz_pen1 (12.4%) son las variables más determinantes, destacando el impacto de las contribuciones al sistema de seguridad social en la predicción de pobreza. Le siguen en importancia ind_arriendol (7.9%) y Nper (6.3%), que reflejan el efecto del arriendo y el tamaño del hogar. Variables económicas como renta_per_capita y ing_per_capita mostraron una frecuencia alta en los árboles (Frequency $> 4\%$), confirmando su papel clave en el modelo. Estos resultados validan que los indicadores laborales y de ingresos son críticos, mientras que características demográficas (Nper, max_educ) y de condiciones de vivienda (propiedad5) aportan información complementaria.

Resultados de Validación Cruzada y Métricas:

La optimización mediante 3-Fold Cross-Validation arrojó un F1-Score promedio de 0.7209 (OUT-OF-FOLD), demostrando robustez en datos no vistos. El ensemble final (87% XGBoost + 13% GLMNET) alcanzó un F1-Score de 0.9531 en entrenamiento con SMOTE, pero el umbral óptimo se ajustó a 0.72 (no 0.5) para maximizar el equilibrio entre precision y recall en la clase minoritaria.

3.2 Tuneo de hiperparametros

Modelo XGBoost y GLMNET (Ensemble)

Para la selección de hiperparámetros en XGBoost, se implementó una búsqueda en grilla (grid search) con validación cruzada de 5-Fold, evaluando el F1-Score como métrica principal debido al desbalance de clases (20% "Pobre"). Se probaron combinaciones clave: nrounds (100, 200), max_depth (4, 6, 8), eta (0.01, 0.1, 0.3, 0.05), gamma (0, 1), colsample_bytree (0.6, 0.8, 1.0), min_child_weight (1, 5) y subsample (0.7, 0.8, 1.0), buscando equilibrio entre complejidad y generalización. Los mejores resultados se obtuvieron con max_depth=8, eta=0.05, min_child_weight=5 y subsample=0.7, logrando un F1-Score de 0.7159 en CV. Posteriormente, se refinó el modelo añadiendo scale_pos_weight=4 para priorizar la clase minoritaria y combinándolo con SMOTE y GLMNET, alcanzando el F1-Score final de 0.7237. La estrategia demostró que hiperparámetros como min_child_weight y max_depth son críticos para manejar el desbalance, mientras que eta y subsample optimizaron la convergencia y robustez del modelo.

El modelo GLMNET se configuró con a=0.2 (mezcla de regularización L1/L2) para equilibrar la selección de variables (como en Lasso) y la estabilidad (como en Ridge), evitando overfitting en datos desbalanceados. Se usó validación cruzada (5-Fold) con métrica AUC para optimizar el parámetro de penalización (lambda), priorizando la capacidad discriminativa del modelo. La elección de lambda.min (en lugar de lambda.1se) maximizó el recall para la clase "Pobre", mientras que el bajo valor de alpha preservó variables clave identificadas en el análisis de importancia (ej. reg_cotiz1, ing_per_capita). Esta combinación aseguró robustez en el ensemble final, complementando a XGBoost al manejar relaciones lineales y ruido en los datos.

Modelo de Probabilidad Lineal con Elastic Net

En este ejercicio se implementó un modelo de probabilidad lineal para la clasificación binaria de hogares en situación de pobreza, utilizando como variable dependiente la categoría 'Pobre'. Para evaluar su desempeño fuera de muestra, se dividió el conjunto de datos en un 80% para entrenamiento y un 20% para validación. Esta partición permitió una evaluación interna sólida, empleando métricas como el F1 Score, especialmente relevante en contextos con clases desbalanceadas, ya que captura de manera equilibrada los errores al combinar precisión y sensibilidad.

La selección de hiperparámetros se realizó mediante una búsqueda en grilla con validación cruzada con 5 particiones. Este enfoque permitió ajustar progresivamente los valores de los parámetros a medida que se obtenían resultados. En una primera etapa, se exploró un rango de valores para alpha (entre 0 y 1, en incrementos de 0.01) y para lambda (entre 0 y 100, en incrementos de 10). El modelo que mostró mejor desempeño correspondió a un alpha = 1 y lambda = 0, lo que indica una preferencia por una regularización predominantemente tipo Lasso, aunque conservando el enfoque combinado de Elastic Net. La elección del modelo final se basó en el valor más alto del F1 Score durante el proceso de validación cruzada el cual tuvo un valor de 0,61.

Modelo de regresión lineal con Elastic Net

Para este modelo, se utilizaron datos a nivel individual, definiendo como variable dependiente el ingreso de cada persona. La especificación incluyó el conjunto completo de variables disponibles, tanto aquellas asociadas a las características individuales como las correspondientes a su contexto familiar. El conjunto de datos fue dividido en dos partes: una base de entrenamiento que representó el 80% de los datos y una base de validación con el 20% restante. Con el objetivo de mejorar la capacidad predictiva del modelo, se aplicó un proceso de balanceo sobre la base de entrenamiento, de modo que se igualara la proporción de individuos clasificados como pobres y no pobres.

Posteriormente, se implementó un procedimiento de validación cruzada con $K = 10$ pliegues, explorando iterativamente distintas combinaciones de los hiperparámetros de α y λ. La búsqueda se concentró en un rango amplio de valores para λ (0 a 10.000) y valores cercanos a cero para α, con el fin de buscar la configuración que minimizara el RMSE. El modelo óptimo se obtuvo con un $\alpha = 0,009$ lo que indica una penalización prácticamente equivalente a Ridge, y un $\lambda = 5.800$ lo cual sugiere una fuerte regularización.

Una vez identificado el modelo con el menor RMSE en la etapa de validación cruzada, este se aplicó al conjunto de validación. Los valores de ingreso predichos fueron transformados en una variable categórica binaria a partir del umbral correspondiente a la línea de pobreza. Finalmente, se calculó el F1 Score comparando las variables Pobre (observada) y Pobre (estimada), obteniendo un valor de 0.6732, lo que refleja un desempeño razonable del modelo en términos de balance entre precisión y sensibilidad en la clasificación binaria de la pobreza.

Modelo de regresión logística

En la fase inicial, se compararon distintos enfoques para modelar la pobreza mediante regresión logística: LDA, QDA, KNN y GLM-binomial estándar (sin balanceo de clases). Para validar los modelos, se evaluaron estrategias de k-folds (entre 3 y 15 particiones), observando que los resultados de F1 se estabilizaban con más de 5 folds. Entre estos métodos, el GLM-binomial estándar obtuvo el mejor desempeño, con un F1-score de 0.6120. Posteriormente, se implementó un balanceo de clases mediante Propensity Score Matching (PSM), ajustando la distribución a 50%-50% entre pobres y no pobres. Al reentrenar el GLM-binomial con estos datos balanceados, se alcanzó un F1 ligeramente superior (0.6120), confirmando que el balanceo no mejoró significativamente el modelo. También, se exploró regularización con Elastic Net (GLM-EN), probando una grilla que abarcaba desde configuraciones cercanas a Lasso (alpha=1, lambda entre 1e-04 y 1) hasta Ridge (alpha=0). El mejor resultado (F1=0.6105) se logró con alpha=1 y lambda=1e-04, indicando que un modelo cercano a Lasso puro, pero con mínima penalización, era óptimo. Esto sugiere que la selección automática de variables (propia de Lasso) fue más relevante que la regularización fuerte. Finalmente, para el mejor modelo, agregamos SMOTE como estrategia de balanceo para el modelo del GLM-binomial y el F1 score mejoró a 0.638.

Classification and Regression Trees(CART)

Para la selección de hiperparámetros de los modelos CART, se utilizó una estrategia de búsqueda en grilla. Se probaron diferentes valores del parámetro de complejidad (cp) en rangos específicos para cada modelo. En el primer modelo, se evaluaron valores de cp desde 0 hasta 0.04 en incrementos de 0.01 eligiendo un cp de 0 (F1=0.6148). En el segundo modelo, se probaron valores de cp desde 0 hasta 0.0115 en incrementos de 0.002, siendo el cp de 0 (F1=0.5032). Para el tercer modelo, se exploraron valores de

cp desde 0 hasta 0.0001 en incrementos de 0.00001 (mejor CP de 0.00004[F1=0.6538]), y en el cuarto modelo, se evaluaron valores de cp desde 0 hasta 0.0002 en incrementos de 0.00002 (mejor CP=0.000325[F1=0.6323]). La elección de estos rangos se basó en la necesidad de encontrar un equilibrio entre la complejidad del modelo y su capacidad de generalización. Además de esto, la decisión de ir cerrando el rango de búsqueda de la grilla se basó en iteraciones que buscaban explorar rangos de valores con los mejores F1 de forma progresiva. El mejor modelo de CART usó un CP de 0.00004 usando SMOTE para balancear la muestra(F1=0.6538).

Random Forest

Este modelo tomó datos a nivel de hogar incluyendo aquellos que originalmente venían con la base de datos de hogares y otras variables creadas a partir de la agregación de datos correspondientes a los individuos del mismo hogar. El entrenamiento del modelo se enfocó en calcular los coeficientes de cada variable para la predicción de los hogares pobres y no pobres. El conjunto de datos fue dividido en dos partes: una base de entrenamiento que representó el 80% de los datos y una base de validación con el 20% restante. Con el objetivo de mejorar la capacidad predictiva del modelo, se aplicó un proceso de balanceo sobre la base de entrenamiento, de modo que se igualara la proporción de individuos clasificados como pobres y no pobres.

La selección de hiperparámetros se llevó a cabo primero seleccionando la cantidad de predictores que serían elegidos al azar en cada partición de los árboles. Iniciamos con la raíz cuadrada de la cantidad de predictores ($\sqrt{44} \approx 6$) - como se recomienda en James *et al.* (2020, pp. 343 y 344) - y probamos con valores cercanos, pero terminamos obteniendo mejores resultados con 30 predictores. Posteriormente, pasamos a experimentar fijando diferentes valores para la cantidad mínima de observaciones por nodo terminal con valores de 10, 25 y 50 observaciones. Luego de varias pruebas, concluímos que el mejor desempeño se obtenía con 13 observaciones por nodo terminal. Por último, probamos la cantidad de árboles que el modelo generaba para predecir el resultado, iniciando con 50, luego 100, 500 y 1000. El mejor puntaje F1 (0.6724) lo obtuvimos con el máximo número de árboles.

3.3 Analisis comparativo

Table 2: Resultados de modelos logísticos

Model	**Accuracy**	**Precision**	**Recall**	**F1_Score**
XGBoost ensemble GLMNET SMOTE	0.878	0.667	0.786	0.721
RL_ElasticNet	0.852	0.736	0.606	0.665
CART_F1op_SMOTE	0.863	0.666	0.633	0.649
modelo_mpl	0.865	0.72	0.532	0.612
LR_QDA_5fold	0.700	0.398	0.862	0.535
Random_Forest	0.846	0.670	0.620	0.6724

La tabla comparativa muestra que el modelo XGBoost ensemble GLMNET SMOTE presenta el mejor desempeño general, alcanzando el mayor valor de F1 Score (0.721), gracias a un equilibrio sólido entre precisión (0.667) y recall (0.786). Este resultado sugiere que el uso conjunto de métodos ensemble, regularización y balanceo de clases contribuye significativamente a capturar correctamente la clase positiva (pobreza), sin sacrificar demasiada precisión. Le sigue el modelo RL_ElasticNet, con un F1 Score de 0.665 y una

precisión superior (0.736), aunque con menor recall (0.606), lo que indica una mayor capacidad para evitar falsos positivos, pero menor cobertura de los verdaderos casos de pobreza. Modelos como CART_F1op_SMOTE y Random Forest también muestran desempeños competitivos (F1 $>$ 0.64), mientras que LR_QDA_5fold obtiene el mayor recall (0.862) pero a costa de una baja precisión (0.398), lo que reduce su F1 Score a 0.535. En conjunto, los resultados resaltan la importancia del balance entre precisión y sensibilidad, así como el impacto positivo del SMOTE y la optimización del umbral en los modelos con mejores métricas.

3.4 Relevancia de las variables

El modelo óptimo identificó variables clave que reflejan dimensiones críticas de la pobreza, alineadas con hallazgos de organismos internacionales. Los indicadores laborales como reg_cotiz1 (18.7% de Gain) y cotiz_pen1 (12.4%) demostraron ser los predictores más fuertes, corroborando estudios de la CEPAL que vinculan la informalidad laboral con mayores tasas de pobreza en América Latina CEPAL (2023). Variables económicas como renta_per_capita y ing_per_capita respaldaron esta relación, mostrando alta frecuencia en los árboles (Frequency $>$ 40%), lo que coincide con metodologías del Banco Mundial para medir pobreza monetaria Mundial (2021). Estas métricas, junto con características del hogar (Nper, suma_antiguedad), capturan dinámicas multidimensionales reconocidas en el Índice de Pobreza Multidimensional global OPHI (2020).

La creación de variables sintéticas y el balanceo de clases enriquecieron la capacidad predictiva del modelo. Interacciones como edad_por_ocupado y ind_arriendol (7.9% de Gain) evidenciaron cómo factores combinados —como edad y acceso a vivienda— agravan el riesgo de pobreza, un patrón documentado por ONU-Hábitat en entornos urbanos ONU-Hábitat (2022). El uso de SMOTE para abordar el desbalance de clases (20% pobreza) mejoró el recall en un 29%, asegurando que el modelo no ignorara la clase minoritaria, una práctica respaldada por investigaciones en aprendizaje automático para datos desbalanceados Fernández et al. (2018). Estas decisiones técnicas, junto con la optimización del threshold (0.72), permitieron un equilibrio robusto entre precisión y sensibilidad, esencial para políticas públicas basadas en evidencia.

4 Conclusiones

Este ejercicio surgió de la preocupación por mejorar la identificación de hogares en condición de pobreza mediante el uso de modelos predictivos basados en características observables. Motivados por la competencia organizada por el Banco Mundial, se exploraron diferentes aproximaciones con el objetivo de responder cuál es el mejor modelo para predecir la pobreza. Tras un análisis comparativo, el modelo XGBoost, complementado con regularización mediante GLMNET y balanceo de clases con SMOTE, presentó el mejor desempeño en términos de precisión y capacidad de generalización.

La elección de este modelo se justifica por varias razones. En primer lugar, XGBoost permite capturar relaciones no lineales y complejas entre variables, como las interacciones entre edad y ocupación, que modelos lineales no pueden representar adecuadamente. En segundo lugar, su combinación con GLMNET permitió optimizar relaciones lineales dentro del modelo mediante regularización, reduciendo el riesgo de sobreajuste. Finalmente, el uso de SMOTE para generar muestras sintéticas de la clase minoritaria "Pobre" mejoró significativamente el recall, alcanzando un 75 por ciento frente al 58 por ciento obtenido

sin balanceo, lo cual resulta fundamental para evitar omisiones en contextos de alta sensibilidad como la asignación de recursos sociales.

5 Referencias

Banco Mundial (2024). Trayectorias: prosperidad y reducción de la pobreza en el territorio colombiano. Disponible en: https://www.bancomundial.org/.

Banco Mundial and DrivenData (2020). Poverty mapping challenge: Predicting poverty. https://www.drivendata.org/competitions/50/worldbank-poverty-prediction/page/97/. Consultado el 12 de abril de 2025.

CEPAL (2023). Panorama social de américa latina 2023. https://www.cepal.org. Consultado en abril de 2025.

DANE (2018). Colombia - medición de pobreza monetaria y desigualdad 2018. Technical report, Departamento Administrativo Nacional de Estadística.

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., and Herrera, F. (2018). Learning from imbalanced data sets. *Knowledge and Information Systems*, 55(1):1–51.

Haughton, Jonathan Khandker, S. R. (2009). *Handbook on Poverty and Inequality*. World Bank Publications, Washington, DC.

Lipton, M. and Ravallion, M. (1995). Poverty and policy. In Behrman, J. and Srinivasan, T., editors, *Handbook of Development Economics*, volume 3, pages 2551–2657. Elsevier.

Mundial, B. (2021). Medición de la pobreza: Definición y métodos. https://www.worldbank.org. Consultado en abril de 2025.

ONU-Hábitat (2022). Informe mundial sobre asentamientos humanos 2022: Ciudades y vivienda en la era de la urbanización masiva. Technical report, Programa de las Naciones Unidas para los Asentamientos Humanos. Consultado en abril de 2025.

OPHI (2020). Global multidimensional poverty index 2020. Technical report, University of Oxford. Consultado en abril de 2025.

Pave, T. and Stender, N. (2017). Is random forest a superior methodology for predicting poverty? an empirical assessment. *Poverty & Public Policy*, 9(1):118–133.

Solís-Salazar, María Madrigal-Sanabria, J. (2022). A machine learning proposal to predict poverty. *Tecnología en Marcha*, 35(4):84–94.

Vallejo Zamudio, L. E. (2021). Magnitud e implicaciones de la pobreza en colombia. *Apuntes del Cenes*, 40(72):7–13.

A Apéndice 1 - Diccionario de variables

Esta tabla contiene una descripción corta de todas las variables utilizadas para el mejor modelo implementado:

Variable	**Tipo**	**Descripción**
Pobre	Numérica binaria	Indicador de pobreza (0 = no pobre, 1 = pobre)
Cabecera	Factor (2 niveles)	Ubicación (1 = cabecera, 2 = resto)
Dominio	Factor (25 niveles)	Área geográfica o ciudad de dominio
num_room	Entera	Número de habitaciones
num_bed	Entera	Número de camas
propiedad	Factor (6 niveles)	Tipo de tenencia de la vivienda
pago_amort	Factor (576 niveles)	Valor de amortización
renta_h	Numérica	Renta del hogar
renta_r	Numérica	Renta reportada
Nper	Entera	Número de personas en el hogar
Depto	Factor (24 niveles)	Departamento de residencia
suma_antiguedad	Entera	Suma de antigüedad laboral en el hogar
promedio_antiguedad	Entera	Promedio de antigüedad laboral
tiene_empleado_publico	Factor (2 niveles)	Indicador de empleado público en el hogar
tiene_patron	Factor (2 niveles)	Hogar tiene empleadores
tiene_cuenta_propia	Factor (2 niveles)	Hogar con trabajadores por cuenta propia
tiene_emp_domestico	Factor (2 niveles)	Empleados domésticos en el hogar
tiene_jornalero	Factor (2 niveles)	Hogar con jornaleros
tiene_sin_remuneracion	Factor (2 niveles)	Miembros sin remuneración en el hogar
n_posiciones_lab_distintas	Entera	Número de posiciones laborales distintas
aux_trans	Factor (2 niveles)	Recibe auxilio de transporte
ind_prima	Factor (2 niveles)	Recibe prima
prima_serv	Factor (2 niveles)	Recibe prima de servicios
prima_nav	Factor (2 niveles)	Recibe prima de navidad
prima_vac	Factor (2 niveles)	Recibe prima de vacaciones

ind_viaticos	Factor (2 niveles)	Recibe viáticos
ocupado	Factor (2 niveles)	Persona ocupada
ind_oficio	Factor (2 niveles)	Persona con oficio
ind_arriendo	Factor (2 niveles)	Recibe ingreso por arriendo
pet_trabajo	Factor (19 niveles)	Petición de tipo de trabajo
max_educ	Entera	Nivel educativo máximo alcanzado
hr_extr	Entera	Horas extra trabajadas
otro_trr	Entera	Otro tipo de trabajo relacionado
rem_ext	Entera	Remesas del exterior
reg_cotiz	Factor (2 niveles)	Registrado como cotizante
cotiz_pen	Factor (2 niveles)	Cotiza a pensión
ing_otros	Entera	Otros ingresos
edad_prom	Numérica	Edad promedio en el hogar
perc_fem	Numérica	Porcentaje de mujeres en el hogar
ing_total	Numérica	Ingreso total del hogar
ing_total_log	Numérica	Logaritmo del ingreso total
renta_per_capita	Numérica	Renta per cápita
ocupado_por_fem	Numérica	Ocupados por porcentaje de mujeres
edad_prom_cuadrado	Numérica	Edad promedio al cuadrado
edad_prom_cubico	Numérica	Edad promedio al cubo
ing_per_capita	Numérica	Ingreso per cápita (duplicado de renta_per_capita)
empleo_por_antig	Numérica	Promedio de empleo por antigüedad
renta_vs_educ	Numérica	Relación entre renta y educación
edad_por_educ	Numérica	Relación edad promedio y educación
log_renta_r	Numérica	Logaritmo de la renta reportada
sqrt_renta_r	Numérica	Raíz cuadrada de la renta reportada
antiguedad_por_educ	Numérica	Antigüedad por educación
renta_por_habitacion	Numérica	Renta por habitación
ratio_otros_renta	Numérica	Relación otros ingresos sobre renta
ratio_educ_antig	Numérica	Relación educación vs antigüedad
interaccion_ocupado_renta	Numérica	Interacción entre ocupación y renta

Tabla de estadísticas descriptivas de las principales variables socioeconómicas

variable	categoria	n	porcentaje
Cabecera	1	149488	90.60
propiedad	1	62276	37.80
propiedad	2	5626	3.40
propiedad	3	64453	39.10
propiedad	4	24865	15.10
propiedad	5	7574	4.60
propiedad	6	166	0.10
Pobre	1	33024	20.00
tiene_empleado_publico	1	11608	7.00
tiene_patron	1	8391	5.10
tiene_cuenta_propia	1	84919	51.50
tiene_emp_domestico	1	7677	4.70
tiene_jornalero	1	2867	1.70
tiene_sin_remuneracion	1	7067	4.30
aux_trans	1	44027	26.70
ind_prima	1	1287	0.80
prima_serv	1	55588	33.70
prima_nav	1	21822	13.20
prima_vac	1	17762	10.80
ind_viaticos	1	2987	1.80
pension	1	22981	13.90
ocupado	1	142737	86.50
ind_oficio	1	11406	6.90
ind_arriendo	1	36134	21.90
max_educ	0	37074	22.50
max_educ	1	47501	28.80
max_educ	2	80385	48.70
hr_extr	1	6985	4.20
otro_tr	1	11259	6.80
rem_ext	1	10836	6.60
reg_cotiz	1	102887	62.40
cotiz_pen	1	71487	43.30
ing_otros	1	72415	43.90