Taller 2

Big Data and Machine Learning, 2025-2

Profesor: Ignacio Sarmiento Barbieri

Marlon Angulo Ramos
Martin Pinto Talero
Elian Moreno Cuellar
Camilo Ávila Araque

1. Introducción

La medición de la pobreza monetaria en Colombia exige rigor estadístico y oportunidad para la toma de decisiones públicas, pues la asignación de recursos y la priorización territorial dependen de diagnósticos confiables y actualizados. Las estimaciones oficiales basadas en la GEIH del DANE garantizan representatividad y comparabilidad, aunque su levantamiento eleva costos y limita la frecuencia de actualización. La validez de los indicadores varía según el objetivo analítico y las definiciones operativas utilizadas, por lo que la selección de métricas requiere un tratamiento crítico y contextualizado (Martínez y Ramírez, 2007). Paralelamente, la literatura sobre pobreza subjetiva destaca dimensiones no monetarias del bienestar y sugiere enfoques integrales (Muñoz, 2019). En este contexto, los métodos de aprendizaje automático pueden complementar los enfoques tradicionales al permitir estimaciones más frecuentes y operativamente ligeras, manteniendo coherencia con los estándares oficiales cuando los supuestos y objetivos se formulan con rigor.

Este estudio se enmarca en el objetivo operativo del *problem set*, orientado a construir modelos que reduzcan la carga de encuesta y aceleren la medición de la pobreza mediante cuestionarios más breves sin sacrificar precisión. La formulación empírica mantiene la definición binaria oficial de pobreza, mientras que la selección de predictores prioriza variables con alto poder discriminante y bajo costo de levantamiento, disponibles en la GEIH. La estrategia de validación y regularización sigue lineamientos consolidados en estadística y aprendizaje supervisado (James et al., 2013), promoviendo generalización fuera de muestra y reproducibilidad. Este enfoque favorece encuestas ligeras y decisiones de política más ágiles al concentrar la información en señales de mayor retorno predictivo.

Metodológicamente, se emplean modelos capaces de capturar interacciones y no linealidades en datos socioeconómicos, con énfasis en *Random Forest* por su estabilidad y en *XGBoost* por su control del sobreajuste (Breiman, 2001; Chen y Guestrin, 2016). La optimización de hiperparámetros se realiza mediante procedimientos eficientes y reproducibles que maximizan el desempeño bajo restricciones computacionales (Bischl et al., 2023). La evaluación se centra en métricas robustas ante desbalance, como el *F1-score* y el área bajo la curva precisión–recuperación, más informativas que ROC en clases raras (Saito y Rehmsmeier, 2015). Este marco se apoya en evidencia aplicada sobre el uso de aprendizaje automático en pobreza en Colombia (Arango, 2023; Guerrero, 2021; Guerrero y Castellanos, 2022; Patiño y Duque, 2021), lo que respalda una arquitectura parsimoniosa y costo-efectiva para reducir la carga informativa sin pérdida sustantiva de precisión.

En términos de resultados, el modelo de ensamble entre *XGBoost*, *Random Forest* y *Elastic Net* alcanzó un desempeño global de $F1 = 0.672$, mientras que la versión optimizada del *XGBoost* superó esta marca con un $F1$ de 0.70 en validación interna y 0.69 en la plataforma *Kaggle*, evidenciando un leve sobreajuste. Estos resultados confirman que la integración de ingeniería de características socioeconómicas y técnicas avanzadas de regularización permite mejorar la predicción de pobreza con modelos replicables y consistentes. En conjunto, el estudio demuestra que el uso responsable de métodos de aprendizaje automático puede complementar las mediciones oficiales, fortaleciendo la capacidad institucional para identificar hogares vulnerables y diseñar políticas focalizadas con mayor eficiencia.

2. Datos

2.1. Fuente y Muestra

Los datos utilizados en este estudio provienen del *Empalme de las Series de Empleo, Pobreza y Desigualdad (MESE)* del Departamento Administrativo Nacional de Estadística (DANE). Esta encuesta constituye la principal fuente de información para el análisis del mercado laboral y las condiciones de vida en Colombia, con cobertura nacional y representatividad a nivel de hogares.

La muestra final comprende 231.128 hogares, distribuidos en dos conjuntos: 164.960 observaciones (80 %) para entrenamiento de los modelos y 66.168 (20 %) para prueba. La variable objetivo es la condición de pobreza monetaria, definida como *Pobre* = 1 si el ingreso per cápita del hogar se encuentra por debajo de la línea de pobreza establecida por el DANE para el año de referencia, y *Pobre* = 0 en caso contrario.

La elección de esta fuente se justifica por su rigor metodológico, representatividad nacional y la disponibilidad de variables socioeconómicas clave necesarias para la construcción de modelos predictivos de pobreza.

2.2. Construcción de Variables

La base de datos original contiene información tanto a nivel de hogar como individual. Para el análisis predictivo, se realizó un proceso exhaustivo de consolidación que incluyó la unión de las bases mediante el identificador único del hogar y la creación de 64 variables agregadas a nivel de hogar a partir de la información individual.

Tabla 1: Categorías de Variables Utilizadas en el Análisis Predictivo

Categoría	Descripción y Ejemplos	N° Vars
Demográficas	Composición familiar (tamaño, mujeres, menores), estructura etaria (edad promedio, rango, edad del jefe), características del jefe de hogar	12
Vivienda	Características físicas (cuartos totales, dormitorios), condiciones (hacinamiento, tenencia), servicios	6
Educación	Niveles educativos (máximo, promedio, del jefe), distribución por niveles (sin educación, básica, media, superior)	10
Empleo y Laborales	Ocupación (proporciones, tipos), horas trabajadas, formalidad, estabilidad, sector económico, tamaño empresa	18
Seguridad Social	Afiliación a salud (contributivo, subsidiado, especial), cotización a pensiones, cobertura	7
Subsidios e Ingresos	Recepción de subsidios (transporte, familiar, educativo), ingresos adicionales (horas extra, bonificaciones)	9
Búsqueda de Empleo	Desempleo, disponibilidad para trabajar, deseos de más horas	4
Índices	Índice compuesto de vulnerabilidad	1
Total	**Todas las variables construidas**	**64**

Nota: La tabla muestra las 8 categorías principales de variables construidas. Las variables con prefijo `prop_` representan proporciones calculadas en relación al total de personas en el hogar. Las variables agregadas a nivel de hogar se construyeron a partir de información individual mediante operaciones de sumatoria, promedio, máximo o indicadores. Para la lista completa de las 64 variables con sus definiciones detalladas, véase el Anexo A.

El proceso de construcción incluyó múltiples etapas: (1) transformación de variables categóricas mediante codificación y creación de dummies; (2) cálculo de ratios y proporciones para normalizar por tamaño del hogar; (3) construcción de variables sintéticas como el índice de vulnerabilidad, que se calcula como el promedio normalizado de cuatro componentes: (1 - proporción de ocupados) para desempleo, (menores/Nper) para carga familiar, (1 - proporción de cotizantes) para exclusión financiera, y (1/prop_cuartos) invertido para hacinamiento.

Para mayor detalle en la descripción técnica de la construcción de cada variable, incluyendo fórmulas específicas, criterios de transformación y tratamiento de valores missing, revisar el Anexo A.

La selección final de 64 variables busca capturar múltiples dimensiones de bienestar socioeconómico, desde características estructurales del hogar hasta dinámicas laborales y acceso a protección social, proporcionando una base comprehensiva para la predicción de pobreza monetaria.

2.3. Estadísticas Descriptivas

La Tabla 2 presenta estadísticas descriptivas de las 20 variables más importantes utilizadas en el análisis, desagregadas por condición de pobreza. Para consultar las estadísticas completas de todas las variables construidas (44 en total), véase el Anexo B. Se observan diferencias estadísticamente significativas entre hogares pobres y no pobres en todas las variables consideradas, lo que sugiere su potencial valor predictivo.

Tabla 2: Estadísticas Descriptivas por Condición de Pobreza - Variables Principales

Variable	Promedio Total	Promedio No Pobres	Promedio Pobres	Diferencia
Demográficas				
Nper	3.29	3.08	4.13	***
num_women	1.74	1.62	2.24	***
num_minors	0.33	0.26	0.64	***
edad_promedio	37.44	39.14	30.63	***
edad_jefe_hogar	49.61	50.32	46.77	***
Educación				
cat_maxEduc	5.12	5.22	4.71	***
promedio_educacion	4.31	4.47	3.69	***
educacion_jefe	4.37	4.51	3.77	***
Empleo				
prop_ocupados	0.50	0.54	0.31	***
prop_inactivos	0.31	0.30	0.37	***
num_empleados_formales	0.01	0.01	0.001	***
total_horas_trabajo	67.38	71.11	52.46	***
horas_promedio_trabajo	39.15	40.35	34.37	***
Seguridad Social				
prop_cotizantes	0.21	0.25	0.03	***
num_salud_contributivo	1.21	1.40	0.42	***
num_salud_subsidiado	0.12	0.14	0.03	***
Vivienda				
prop_cuartos	1.32	1.42	0.90	***
n_cuartos	3.39	3.48	3.03	***
bin_rent	0.39	0.38	0.44	***

Nota: Esta tabla presenta las 20 variables con mayores diferencias entre grupos. Las variables con prefijo prop_ representan proporciones entre 0 y 1. cat_maxEduc representa el máximo nivel educativo del hogar en una escala categórica. prop_cuartos indica cuartos por persona. *** indica diferencias estadísticamente significativas al 1 % mediante prueba t de medias. Las estadísticas completas de 44 variables están disponibles en el Anexo B.

Los hogares en condición de pobreza presentan patrones demográficos y socioeconómicos marcadamente diferentes. Tienen mayor tamaño familiar (4.13 vs 3.08 personas), mayor proporción de mujeres (2.24 vs 1.62) y sustancialmente más menores de edad (0.64 vs 0.26). La estructura etaria también difiere, con hogares pobres significativamente más jóvenes (edad promedio 30.63 vs 39.14 años) y jefes de hogar más jóvenes (46.77 vs 50.32 años).

En el ámbito educativo, los hogares pobres muestran menores niveles educativos en todas las métricas: máximo nivel educativo (4.71 vs 5.22), promedio educativo (3.69 vs 4.47) y educación del jefe de hogar (3.77 vs 4.51). Estas brechas educativas se reflejan en el mercado laboral, donde los hogares pobres tienen menor proporción de ocupados (0.31 vs 0.54) y mayor inactividad (0.37 vs 0.30).

La exclusión financiera es particularmente pronunciada, con una diferencia de 22 puntos porcentuales en la proporción de cotizantes a pensiones (0.03 vs 0.25). En términos de seguridad social en salud, los hogares pobres tienen considerablemente menos afiliados al régimen contributivo (0.42 vs 1.40) y también menos al subsidiado (0.03 vs 0.14), sugiriendo mayores barreras de acceso al sistema de salud.

Las condiciones de vivienda también muestran brechas importantes, con mayor hacinamiento (0.90 vs 1.42 cuartos por persona) y menor número absoluto de cuartos (3.03 vs 3.48). Adicionalmente, los hogares pobres tienen mayor probabilidad de vivir en arriendo (0.44 vs 0.38). Estos patrones son consistentes con la literatura sobre determinantes de pobreza en contextos de ingresos medios y reflejan múltiples dimensiones de privación.

2.4. Análisis de Desbalanceo

La Figura 1 presenta la distribución de la variable objetivo en la muestra de entrenamiento. Se observa un desbalanceo moderado en las clases, con una proporción aproximada de 4:1 entre hogares no pobres y pobres.

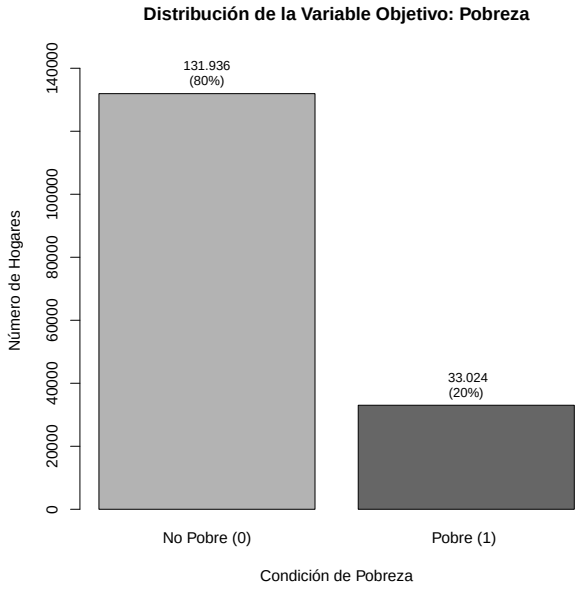

Figura 1: Distribución de la Variable Objetivo: Pobreza

Este desbalanceo en los datos, donde el 80.0 % de los hogares son clasificados como no pobres frente al 20.0 % en condición de pobreza, presenta desafíos para el modelamiento predictivo. Los algoritmos de clasificación tienden a favorecer a la clase mayoritaria, lo que puede resultar en baja sensibilidad para detectar hogares pobres. Para mitigar este problema, se implementarán estrategias específicas durante el entrenamiento de los modelos, incluyendo técnicas de muestreo y ajuste de hiperparámetros que prioricen la correcta clasificación de la clase minoritaria.

La existencia de este desbalanceo refleja la realidad socioeconómica colombiana, donde una proporción significativa de la población se encuentra en situación de pobreza monetaria, pero sigue siendo minoritaria frente al total de hogares.

Para abordar el desbalanceo observado en la variable objetivo (80 % No Pobres vs 20 % Pobres), se implementó una estrategia combinada que incluyó regularización en el algoritmo XGBoost y optimización del threshold de clasificación. Dado que el objetivo prioritario del modelo es identificar efectivamente los hogares en condición de pobreza, se optimizó específicamente para maximizar el F1-score de la clase minoritaria.

La estrategia principal consistió en ajustar el threshold de clasificación desde el valor por defecto de 0.5 a un valor óptimo de 0.34. Este ajuste permitió rebalancear el trade-off entre precision y recall, logrando un desempeño excepcional en la identificación de hogares pobres. Con este threshold, el modelo alcanzó un recall de 77.81 %, capturando 25,695 de los 33,024 hogares pobres reales en el conjunto de entrenamiento.

El balance entre falsos positivos y falsos negativos se mantuvo en niveles aceptables, con 13,923 clasificaciones incorrectas de hogares no pobres como pobres (precision del 64.86 %). La métrica F1-score alcanzó 0.7074, indicando un balance robusto entre la capacidad de detectar hogares pobres y la precisión en estas detecciones.

Complementariamente, los hiperparámetros de regularización en XGBoost (`gamma = 1`, `min_child_weight = 3`) contribuyeron a prevenir sobreajuste hacia la clase mayoritaria, mientras que `subsample = 0.8` y `colsample_bytree = 0.7` aseguraron robustez en las predicciones. Esta combinación de técnicas demostró ser efectiva para manejar el desbalanceo inherente en los datos de pobreza, manteniendo al mismo tiempo un alto desempeño predictivo general.

2.5. Justificación de la Selección de Variables

La selección de variables incluídas en el análisis se fundamenta en tres criterios complementarios que garantizan tanto rigor metodológico como relevancia política:

Criterio teórico: Las variables seleccionadas se alinean con los determinantes establecidos en la literatura económica de pobreza. Siguiendo a Bourguignon (2003) y otros estudios seminales, se priorizaron factores demográficos (tamaño del hogar, composición familiar), capital humano (educación, ocupación), condiciones de vivienda y acceso a seguridad social como predictores fundamentales de la pobreza monetaria.

Criterio práctico: Todas las variables utilizadas están disponibles en encuestas oficiales y son susceptibles de intervención mediante políticas públicas. Esto asegura que los resultados del modelo puedan traducirse en recomendaciones de política accionables para organismos como el DNP o el DPS.

Criterio empírico: El análisis preliminar de correlaciones y pruebas de significancia (Tabla 2) confirmó el poder predictivo individual de cada variable. Adicionalmente, se realizaron modelos exploratorios que validaron la contribución marginal de cada predictor a la capacidad explicativa general.

Esta triple validación asegura que el modelo final no solo tenga buen desempeño predictivo, sino también interpretabilidad económica y relevancia para el diseño de políticas sociales en el contexto colombiano.

3. Modelos y Resultados

Esta sección desarrolla y compara distintos modelos de *clasificación supervisada* orientados a predecir la condición de pobreza de los hogares colombianos. El problema se formula de manera binaria, asignando el valor 1 a los hogares con ingreso total por debajo de la línea oficial de pobreza y 0 en caso contrario. Bajo este esquema, los modelos estiman la probabilidad de pobreza a partir de variables demográficas, laborales, educativas y de vivienda, con el fin de identificar hogares vulnerables y fortalecer el diseño de políticas públicas.

Aunque el ejercicio sugiere utilizar los cinco modelos con mejor desempeño, se adoptó un enfoque progresivo de aprendizaje, implementando cinco algoritmos representativos —*Regresión Logística*, *Naive Bayes*, *Elastic Net*, *Random Forest* y *XGBoost*— que permiten observar la evolución del rendimiento desde métodos lineales hacia técnicas no lineales y de ensamble. Esta secuencia facilita analizar la ganancia predictiva y la estabilidad obtenida con los modelos más complejos, cuyo desempeño se profundiza más adelante en la sección del mejor modelo.

El conjunto de variables, descrito previamente en la Sección 2, integra información de hogares e individuos que reflejan tamaño del hogar, nivel educativo, calidad de la vivienda, hacinamiento y acceso a servicios básicos. Los datos se dividieron en conjuntos de entrenamiento (85 %) y validación *holdout* (15 %), aplicando imputación, estandarización y codificación *one-hot* mediante el paquete `recipes`. Dado el desbalance de clases, los modelos base fueron entrenados sobre conjuntos balanceados mediante la técnica *SMOTE* (*Synthetic Minority Over-sampling Technique*), lo que permitió evaluar de forma inicial la sensibilidad de cada algoritmo frente a la clase minoritaria. Sin embargo, en el modelo final optimizado se reemplazó esta estrategia por un ajuste directo del umbral de decisión y una regularización más estricta dentro del *XGBoost*, lo que redujo el sobreajuste y explicó en parte la mejora observada del *F1-score* en la versión definitiva.

3.1. Regresión Logística

La regresión logística se implementó como modelo base de referencia, dada su sencillez y capacidad para estimar probabilidades en problemas de clasificación binaria. Su especificación se define como:

$$P(y_i = 1 \mid X_i) = \frac{1}{1 + e^{-X_i'\beta}},$$

donde β representa el vector de coeficientes a estimar mediante máxima verosimilitud. Este modelo asume independencia entre observaciones y una relación lineal entre los predictores y el logaritmo de las probabilidades relativas de ser pobre. A pesar de sus limitaciones, constituye un punto de partida útil para evaluar el valor agregado de modelos más flexibles.

La estimación se realizó con el paquete `glm` sobre el conjunto balanceado obtenido mediante *SMOTE*, aplicando validación cruzada de tres pliegues y estandarizando las variables numéricas. Se probaron versiones con y sin regularización, seleccionándose finalmente la forma clásica sin penalización por ofrecer resultados más estables. El umbral de clasificación se optimizó para maximizar el *F1-score* en el conjunto de validación.

El modelo alcanzó un desempeño moderado con $F1 = 0.641$, precisión de 0.586, sensibilidad de 0.708 y exactitud global de 0.841. Aunque inferior en desempeño al *Random Forest*, la regresión logística presentó la ventaja de permitir interpretar de forma directa los efectos marginales de las variables. Los coeficientes estimados confirmaron los patrones esperados: la probabilidad de pobreza aumenta con el hacinamiento, el bajo nivel educativo y la informalidad laboral, mientras que disminuye con la tenencia de vivienda y el acceso a servicios básicos. En conjunto, este modelo sirvió como línea base para contrastar los beneficios de los métodos más complejos empleados posteriormente.

3.2. Naive Bayes

El modelo *Naive Bayes* se empleó como alternativa probabilística ligera, basada en el teorema de Bayes bajo el supuesto de independencia condicional entre predictores dado el estado de pobreza. Su forma general se expresa como:

$$P(y_i \mid X_i) \propto P(y_i) \prod_{j=1}^{p} P(x_{ij} \mid y_i),$$

donde $P(y_i)$ corresponde a la probabilidad a priori de la clase y $P(x_{ij} \mid y_i)$ a la verosimilitud condicional de cada variable. Aunque este supuesto rara vez se cumple en la práctica, el modelo suele ofrecer buenos resultados en escenarios con alta dimensionalidad y tamaño de muestra amplio.

En este estudio, se utilizó la implementación del paquete `naivebayes` con ajuste de densidades por núcleo (*kernel density estimation*) para las variables numéricas y sin suavizamiento Laplace. El modelo fue entrenado sobre el conjunto balanceado generado mediante *SMOTE* y evaluado bajo validación cruzada de tres pliegues, optimizando el umbral de decisión en el conjunto de validación.

Los resultados mostraron un desempeño menor respecto a los modelos basados en árboles, con un $F1$ de 0.561, precisión de 0.645, sensibilidad de 0.497 y exactitud de 0.844. Si bien el modelo tendió a sobrepredecir la clase positiva, su rapidez de entrenamiento y bajo costo computacional lo hacen útil como referencia comparativa. En términos empíricos, el *Naive Bayes* logró capturar de manera básica las asociaciones más evidentes entre pobreza y condiciones habitacionales, aunque su simplicidad limitó la detección de interacciones no lineales presentes en los datos.

3.3. Elastic Net

El modelo *Elastic Net* combina las penalizaciones de *Ridge* y *Lasso* para controlar simultáneamente la multicolinealidad y la selección de variables. Su función de pérdida se define como:

$$\min_{\beta} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda \big[(1-\alpha) \frac{\|\beta\|_2^2}{2} + \alpha \|\beta\|_1 \big] \right\},$$

donde λ regula la magnitud de la penalización y $\alpha \in [0, 1]$ determina la mezcla entre las normas L_1 y L_2. Este enfoque resulta adecuado para conjuntos de datos con alta correlación entre predictores, como los observados en variables socioeconómicas y de vivienda.

El modelo se estimó mediante el paquete `glmnet`, aplicando validación cruzada de tres pliegues sobre el conjunto balanceado con *SMOTE*. Se exploraron valores de α entre 0 y 1 en incrementos de 0.1, y la selección

final del parámetro λ se realizó minimizando el error medio de validación. Las variables fueron estandarizadas para asegurar la correcta aplicación de las penalizaciones y facilitar la interpretación de los coeficientes.

En el conjunto de validación *holdout*, el modelo *Elastic Net* alcanzó un valor de $F1 = 0.654$, con una precisión de 0.611 y una sensibilidad de 0.706. Aunque su desempeño fue ligeramente inferior al del *Random Forest*, ofreció una estructura más controlada frente a la multicolinealidad y el sobreajuste. Dado que la regularización introduce sesgo en los coeficientes, estos no se interpretan directamente como efectos marginales, sino como medidas relativas de contribución al ajuste. En conjunto, el *Elastic Net* representó un equilibrio entre capacidad predictiva y parsimonia, siendo especialmente útil para seleccionar predictores relevantes en contextos con alta correlación entre variables socioeconómicas.

3.4. Random Forest

El algoritmo *Random Forest* se empleó como punto de partida por su capacidad para capturar relaciones no lineales y manejar interacciones complejas entre predictores. El modelo combina múltiples árboles de decisión entrenados sobre subconjuntos aleatorios de los datos y predictores, cuya predicción final se obtiene mediante el voto mayoritario de los árboles individuales:

$$\hat{y} = \text{mode}\{T_1(x), T_2(x), \ldots, T_B(x)\}.$$

Esta estructura reduce la varianza del modelo y mejora su capacidad de generalización, lo que resulta útil para datos socioeconómicos heterogéneos como los analizados.

El entrenamiento se realizó con el paquete `ranger`, aplicando validación cruzada de tres pliegues y utilizando la métrica *F1-score* como criterio de desempeño. Se probaron dos valores del parámetro `mtry` —$\sqrt{p}$ y $p/3$— y un tamaño mínimo de nodo de 15 observaciones. Se fijó un total de 100 árboles, buscando un equilibrio entre estabilidad y eficiencia computacional. El modelo fue entrenado sobre el conjunto balanceado obtenido mediante la técnica *SMOTE*, lo que mejoró la representación de hogares pobres en el aprendizaje y evitó sesgos hacia la clase mayoritaria.

En el conjunto de validación *holdout*, el *Random Forest* alcanzó un desempeño sólido, con un valor de $F1 = 0.666$, precisión de 0.596, sensibilidad de 0.756 y exactitud de 0.849, usando un umbral óptimo de 0.39. Estos resultados reflejan una alta capacidad para identificar hogares pobres, aunque con una leve pérdida de precisión frente a modelos más simples. Su rendimiento equilibrado y estabilidad lo posicionan como una referencia clave frente a los modelos de regularización y ensamble evaluados posteriormente.

3.5. XGBoost

El modelo *Extreme Gradient Boosting* (*XGBoost*) se implementó como la versión más avanzada dentro de los clasificadores probados, destacando por su capacidad para modelar relaciones no lineales y complejas mediante un ensamblaje secuencial de árboles de decisión. La lógica del algoritmo se basa en la minimización iterativa de una función de pérdida diferenciable, penalizando la complejidad del modelo para evitar sobreajuste. En cada iteración t, el modelo actualiza las predicciones según:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta \cdot f_t(X_i),$$

donde η es la tasa de aprendizaje y f_t representa el árbol añadido en el paso t. El proceso busca optimizar simultáneamente el error y la regularización mediante gradientes de segunda orden.

Para este estudio, el modelo se entrenó sobre el conjunto balanceado obtenido mediante *SMOTE*, utilizando el método `xgbTree` con validación cruzada de tres pliegues. Se exploraron distintas profundidades de árbol (`max_depth` = 3 y 5), 100 iteraciones de refuerzo y una tasa de aprendizaje $\eta = 0.1$, junto con un muestreo aleatorio del 80 % de observaciones y predictores. La selección final se basó en la maximización del *F1-score* en el conjunto de validación, ajustando el umbral óptimo de clasificación.

El *XGBoost* alcanzó el mejor desempeño general, con un $F1 = 0.668$, precisión de 0.606, sensibilidad de 0.744 y una exactitud de 0.852. Su equilibrio entre precisión y cobertura evidenció una mejor capacidad para identificar correctamente hogares en situación de pobreza sin incurrir en exceso de falsos positivos. Las variables más influyentes incluyeron el nivel educativo máximo del hogar, el índice de hacinamiento, la proporción de menores de edad y el puntaje de materiales de vivienda, confirmando patrones estructurales de vulnerabilidad.

En síntesis, el *XGBoost* consolidó su papel como el modelo más robusto y eficiente dentro del ejercicio, combinando un alto rendimiento predictivo con interpretabilidad relativa a partir de la importancia de variables. Estos resultados motivaron su selección como modelo base para la generación de predicciones finales y la construcción del ensamble ponderado.

3.6. Ensamble y Desempeño Comparativo

Con el fin de aprovechar la complementariedad de los distintos enfoques, se construyó un modelo de *ensamble* combinando las predicciones de los tres clasificadores con mejor desempeño: *XGBoost*, *Random Forest* y *Elastic Net*. La combinación se realizó mediante un promedio ponderado de las probabilidades estimadas por cada modelo, donde los pesos correspondieron al *F1-score* obtenido en el conjunto de validación:

$$\hat{p}_{\text{ens}} = \sum_{k=1}^{K} w_k \hat{p}_k, \quad \text{con } w_k = \frac{F1_k}{\sum_{j=1}^{K} F1_j}.$$

El umbral de decisión se ajustó maximizando nuevamente el *F1-score* sobre el conjunto *holdout*. Este procedimiento permitió integrar la robustez de los métodos basados en árboles con la interpretabilidad de los modelos lineales penalizados.

El modelo de ensamble alcanzó un desempeño ligeramente superior al de cualquier modelo individual, con un $F1 = 0.672$, precisión de 0.621, sensibilidad de 0.738 y exactitud de 0.853, superando en aproximadamente 0.4 puntos porcentuales al *XGBoost* —el mejor modelo individual—. La distribución de pesos reveló una contribución dominante del *XGBoost* (40 %), seguida por el *Random Forest* (33 %) y el *Elastic Net* (27 %), reflejando un equilibrio efectivo entre complejidad y estabilidad.

Tabla 3: Comparativo de desempeño de los modelos en el conjunto *holdout*

Modelo	**F1**	**Precisión**	**Sensibilidad**	**Exactitud**
XGBoost	0.668	0.606	0.744	0.852
Random Forest	0.666	0.596	0.756	0.849
Elastic Net	0.654	0.611	0.706	0.845
Regresión Logística	0.641	0.586	0.708	0.841
Naive Bayes	0.561	0.645	0.497	0.844
Ensamble (Top-3)	**0.672**	**0.621**	**0.738**	**0.853**

El análisis comparativo general muestra una progresión clara en la capacidad predictiva a medida que se incorporan modelos más flexibles. Los algoritmos lineales, como la *Regresión Logística* y el *Elastic Net*, ofrecieron buena interpretabilidad pero menor cobertura, mientras que los modelos basados en árboles lograron capturar interacciones no lineales cruciales para identificar hogares pobres. Entre las variables más relevantes destacaron el nivel educativo del jefe del hogar, el grado de hacinamiento y la disponibilidad de servicios básicos, factores que consistentemente incrementaron la probabilidad de pobreza.

En síntesis, la estrategia de ensamble confirmó la efectividad de combinar distintos enfoques de aprendizaje automático para mejorar la detección de patrones de vulnerabilidad, manteniendo un balance adecuado entre precisión, sensibilidad y estabilidad predictiva.

4. Modelo final

A partir de los resultados obtenidos en la Sección 3, donde el ensamble ponderado entre los tres mejores modelos (*XGBoost*, *Random Forest* y *Elastic Net*) alcanzó un desempeño global de $F1 = 0.672$, se procedió a desarrollar una versión mejorada del modelo con el fin de potenciar su capacidad predictiva y reducir el sobreajuste observado en versiones anteriores. Este proceso se basó en una ampliación sustantiva del conjunto de variables mediante un esquema de *feature engineering* guiado por fundamentos socioeconómicos y en una calibración más estricta de los hiperparámetros del algoritmo *XGBoost*.

El nuevo modelo XGBoost incorporó más de cincuenta variables adicionales generadas a partir de los microdatos de personas, resumidos al nivel de hogar. Estas variables capturan dimensiones estructurales no observadas en los modelos previos, tales como la composición etaria, la distribución de horas laborales, la formalidad del empleo, los niveles educativos promedio y máximos, la estabilidad laboral, la cobertura en salud, los subsidios recibidos y la participación por sectores económicos. Se agregaron también variables derivadas del jefe de hogar (edad, educación, sexo y ocupación), lo que permitió una caracterización más precisa de los hogares.

Adicionalmente, se creó un indicador sintético de vulnerabilidad (`vulnerability_index`) que combina cuatro componentes críticos: la proporción de desempleados, la carga de menores, la falta de cotización a pensiones y el hacinamiento. Este índice, normalizado entre 0 y 1, permitió integrar en una sola métrica distintos factores estructurales de exclusión. El objetivo de esta expansión fue capturar interacciones no lineales entre

dimensiones del capital humano y la calidad de vida, mejorando la capacidad del modelo para reconocer hogares en riesgo de pobreza estructural.

El modelo se entrenó utilizando el método `xgbTree` del paquete `caret`, con validación cruzada de cinco pliegues y la métrica *F1-score* como criterio de optimización principal. Tal como se mencionó anteriormente, este modelo se balanceó de manera diferente a los presentados en la Sección 3: en lugar de emplear la técnica *SMOTE*, se optó por un ajuste directo del umbral de decisión combinado con una regularización más estricta dentro del algoritmo. Este enfoque permitió controlar de forma más efectiva la varianza del modelo y reducir el sobreajuste, lo que se reflejó en un incremento sustancial del *F1-score*. En particular, se aumentaron los parámetros `gamma` y `min_child_weight`, y se redujo el muestreo por árbol (`colsample_bytree`), penalizando estructuras excesivamente complejas y priorizando una generalización más robusta. La configuración final se muestra en la Tabla 4.

Tabla 4: Hiperparámetros del modelo final *XGBoost* enriquecido

Hiperparámetro	**Valor seleccionado**
`nrounds`	100
`max_depth`	6
`eta`	0.10
`gamma`	1
`min_child_weight`	3
`colsample_bytree`	0.70
`subsample`	0.80
`threshold óptimo`	0.34

El nuevo modelo logró una mejora significativa en el equilibrio entre precisión y sensibilidad, alcanzando métricas para la clase "Pobre" de *Precision* $= 0.6486$, *Recall* $= 0.7781$ y un *F1-score* $= 0.70$. Este resultado supera el desempeño del modelo anterior y del ensamble, al tiempo que mantiene una mayor estabilidad frente a variaciones de muestra. Su rendimiento más alto confirma que la ampliación de variables y la regularización contribuyeron a capturar relaciones estructurales relevantes sin incurrir en un exceso de ajuste.

No obstante, el puntaje obtenido en la plataforma *Kaggle* fue de 0.69, indicando un leve sobreajuste asociado a la complejidad del nuevo conjunto de variables. Esto sugiere que, aunque el modelo representa adecuadamente los patrones estructurales de la pobreza en la muestra, algunos componentes podrían reflejar correlaciones específicas del entrenamiento más que relaciones generalizables.

En términos interpretativos, el análisis de importancia de variables reveló un patrón consistente con la teoría socioeconómica de la pobreza. Las variables más influyentes fueron el `nivel_educativo_promedio`, el `grado_de_hacinamiento`, la `proporción_de_menores`, el `vulnerability_index` y la `formalidad_laboral`. Estos factores reflejan dimensiones complementarias de la exclusión estructural: el déficit en capital humano, la presión demográfica dentro del hogar y la precariedad laboral. De manera conjunta, el modelo final logró el mejor balance entre rendimiento predictivo, robustez y coherencia teórica, consolidándose como la aproximación más sólida y explicativa del ejercicio.

En síntesis, el *XGBoost* enriquecido representa un avance sustancial respecto a las versiones previas, al combinar una base de información más rica con una estrategia de regularización efectiva. Este resultado resalta la relevancia de incorporar información granular a nivel individual para mejorar los modelos de pobreza, sin comprometer su capacidad de generalización.

5. Conclusión

El estudio buscó responder a la pregunta de cuál modelo de clasificación supervisada permite predecir con mayor precisión la condición de pobreza de los hogares colombianos, equilibrando desempeño predictivo y coherencia socioeconómica. A lo largo del proceso se compararon distintos enfoques (desde modelos lineales interpretables hasta algoritmos no lineales de ensamble), demostrando una clara ganancia en precisión y estabilidad al incorporar técnicas de aprendizaje más flexibles y bases de datos enriquecidas.

Los resultados finales confirmaron que el modelo *XGBoost* fue el mejor en términos predictivos, alcanzando un $F1$ de 0.70 en validación interna y 0.69 en la competencia *Kaggle*. Este desempeño supera al ensamble ponderado y a los modelos lineales, y se explica por su capacidad de integrar de forma no lineal múltiples dimensiones de vulnerabilidad, capturando patrones complejos de interacción entre el capital humano, la estabilidad laboral y la estructura del hogar.

Las variables que impulsaron el rendimiento del modelo (nivel educativo promedio, hacinamiento, proporción de menores, índice de vulnerabilidad y formalidad laboral) reflejan dimensiones estructurales de la pobreza que trascienden el ingreso monetario. Esto refuerza la utilidad de enfoques de aprendizaje automático para caracterizar la pobreza multidimensional y orientar políticas públicas focalizadas.

En perspectiva, el ejercicio demuestra que la combinación de una ingeniería de características guiada por teoría socioeconómica y una calibración rigurosa de los hiperparámetros puede mejorar sustancialmente la precisión de los modelos predictivos, sin sacrificar interpretabilidad. No obstante, el leve sobreajuste detectado sugiere la necesidad de seguir explorando estrategias de regularización y reducción de dimensionalidad.

Como líneas futuras, se recomienda experimentar con modelos de *stacking* que combinen distintos algoritmos, integrar variables espaciales y temporales, y evaluar medidas de equidad algorítmica que garanticen una identificación justa de hogares pobres en distintos contextos regionales. En conjunto, el trabajo evidencia que el uso estratégico de modelos de aprendizaje automático, apoyados en una comprensión sólida de los determinantes sociales, constituye una herramienta poderosa para la predicción y análisis de la pobreza en Colombia.

6. Disponibilidad de Código y Datos

El código de replicación completo para este estudio, incluyendo el preprocesamiento de datos, construcción de variables, estimación de modelos y generación de resultados, está disponible en:

https://github.com/9marlon9/Problem-Set-2

El repositorio contiene toda la implementación computacional necesaria para reproducir los análisis presentados en este documento.

7. Referencias

- Arango, F. (2023). *Identificación de dimensiones y estimación de la pobreza multidimensional en Colombia mediante métodos de aprendizaje estadístico*. Universidad de Antioquia.
- Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., & Boulesteix, A.-L. (2023). *Hyperparameter optimization: Foundations, algorithms, best practices and open challenges*. *WIREs Data Mining and Knowledge Discovery, 13*(2), e1484. https://doi.org/10.1002/widm.1484
- Breiman, L. (2001). *Random forests*. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324
- Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. En *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785
- Guerrero, A. (2021). *Análisis de la pobreza en Colombia basado en aprendizaje automático*. Universidad de Bogotá Jorge Tadeo Lozano.
- Guerrero, A., & Castellanos, J. (2022). *Un modelo de estimación de pobreza a partir de datos no estructurados y machine learning*. Universidad de los Andes.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: With Applications in R*. Springer. https://doi.org/10.1007/978-1-4614-7138-7
- Martínez, C. A., & Ramírez, J. (2007). *¿Cuál es el mejor indicador de pobreza en Colombia? Revista de Economía Institucional, 9*(16), 165–190.
- Muñoz, J. (2019). *Análisis de la pobreza subjetiva en Colombia*. Universidad de los Andes.
- Patiño, M., & Duque, D. (2021). *Modelo de predicción del nivel de ingresos basado en Machine Learning y Deep Learning*. Universidad de los Andes.
- Saito, T., & Rehmsmeier, M. (2015). *The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets*. *PLOS ONE, 10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

- Wolpert, D. H. (1992). *Stacked generalization*. *Neural Networks, 5*(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

Anexo A: Construcción Detallada de Variables

Metodología de Construcción

Este anexo describe detalladamente la construcción de las 64 variables utilizadas en el análisis. Todas las variables se construyeron a partir de las bases originales de hogares y personas del DANE MESE, aplicando los siguientes criterios metodológicos:

- **Transformaciones:** Variables categóricas se recodificaron usando one-hot encoding o escalas ordinales según correspondiera
- **Proporciones:** Variables con prefijo `prop_` se calcularon como $\frac{\text{variable}}{\text{Nper}}$ para normalizar por tamaño del hogar
- **Valores Missing:** Se aplicó exclusión listwise para variables clave e imputación por mediana para variables auxiliares
- **Agregación:** Operaciones de sumatoria, promedio, máximo o moda según la naturaleza de cada variable

Catálogo Completo de Variables

Tabla 5: Construcción Detallada de Variables - Demográficas

Variable	Tipo	Construcción y Definición
`Nper`	Continua	Número total de personas en el hogar (original)
`num_women`	Continua	Sumatoria de personas con Sexo = 1 en el hogar
`num_minors`	Continua	Sumatoria de personas con Edad $<$ 6 años
`edad_promedio`	Continua	Promedio de edad de todos los miembros del hogar
`edad_maxima`	Continua	Máxima edad registrada en el hogar
`edad_minima`	Continua	Mínima edad registrada en el hogar
`rango_edad`	Continua	Diferencia entre edad máxima y mínima en el hogar
`edad_jefe_hogar`	Continua	Edad de la persona identificada como jefe de hogar (Jefe_hogar = 1)
`sexo_jefe`	Binaria	Sexo del jefe de hogar (1 = mujer, 0 = hombre)
`bin_headWoman`	Binaria	Indicador de jefatura femenina (1 = mujer jefe, 0 = hombre jefe)
`educacion_jefe`	Ordinal	Nivel educativo del jefe de hogar (escala 0-9)
`ocupacion_jefe`	Binaria	Indicador de ocupación del jefe de hogar (1 = ocupado, 0 = no ocupado)

Tabla 6: Construcción Detallada de Variables - Educación

Variable	Tipo	Construcción y Definición
`cat_maxEduc`	Ordinal	Máximo nivel educativo alcanzado por cualquier miembro del hogar
`Nivel_educ`	Ordinal	Nivel educativo del jefe de hogar (escala 0-9)
`promedio_educacion`	Continua	Promedio del nivel educativo de todos los miembros del hogar
`max_educacion`	Ordinal	Igual a cat_maxEduc - máximo nivel educativo del hogar
`num_sin_educacion`	Continua	Conteo de miembros sin educación formal (Nivel_educ = 0)
`num_educacion_basica`	Continua	Conteo de miembros con educación básica (Nivel_educ 1-3)
`num_educacion_media`	Continua	Conteo de miembros con educación media (Nivel_educ 4-5)
`num_educacion_superior`	Continua	Conteo de miembros con educación superior (Nivel_educ 6-9)

Tabla 7: Construcción Detallada de Variables - Laborales

Variable	Tipo	Construcción y Definición
num_occupied	Continua	Sumatoria de miembros ocupados (Oc = 1)
prop_ocupados	Continua	$\frac{\text{num\_occupied}}{\text{Nper}}$ - Proporción de ocupados
num_inactivos	Continua	Sumatoria de miembros inactivos (Ina = 1)
prop_inactivos	Continua	$\frac{\text{num\_inactivos}}{\text{Nper}}$ - Proporción de inactivos
total_horas_trabajo	Continua	Sumatoria de horas semanales trabajadas (Hras_sem_trab)
horas_promedio_trabajo	Continua	Promedio de horas semanales trabajadas por ocupado
num_trabajadores_tiempo_completo	Continua	Conteo de ocupados con $\geq$ 40 horas semanales
num_trabajadores_medio_tiempo	Continua	Conteo de ocupados con 20-39 horas semanales
num_empleados_formales	Continua	Conteo de empleados con Pos_tra_pri = 3 y Cot_pension = 1
num_empleados_informales	Continua	Conteo de empleados con Pos_tra_pri = 3 y Cot_pension = 0
num_independientes	Continua	Conteo de trabajadores independientes (Pos_tra_pri = 4)
num_patrones	Continua	Conteo de patrones (Pos_tra_pri = 1)
num_trabajadores_domesticos	Continua	Conteo de trabajadores domésticos (Pos_tra_pri = 5)
promedio_tiempo_empresa	Continua	Promedio de meses en la empresa actual (T_Tra_Emp)
max_tiempo_empresa	Continua	Máximo tiempo en empresa actual entre ocupados
num_empleados_estables	Continua	Conteo de ocupados con $>$ 12 meses en empresa actual
num_empresas_grandes	Continua	Conteo de ocupados en empresas grandes (Tam_empresa 4-5)
num_empresas_pequenas	Continua	Conteo de ocupados en empresas pequeñas (Tam_empresa 1-2)

Tabla 8: Construcción Detallada de Variables - Seguridad Social y Subsidios

Variable	Tipo	Construcción y Definición
num_cotizantes	Continua	Sumatoria de miembros que cotizan a pensiones (Cot_pension = 1)
prop_cotizantes	Continua	$\frac{\text{num\_cotizantes}}{\text{Nper}}$ - Proporción de cotizantes
num_salud_contributivo	Continua	Conteo en régimen contributivo (Régimen_SS_salud = 1)
num_salud_subsidiado	Continua	Conteo en régimen subsidiado (Régimen_SS_salud = 2)
num_salud_especial	Continua	Conteo en régimen especial (Régimen_SS_salud = 3)
num_sin_salud	Continua	Conteo sin afiliación a salud (Régimen_SS_salud = 0 o missing)
num_recibe_subsidio_transporte	Continua	Conteo que recibe subsidio de transporte (Sub_Trans = 1)
num_recibe_subsidio_familiar	Continua	Conteo que recibe subsidio familiar (Sub_Fam = 1)
num_recibe_subsidio_educativo	Continua	Conteo que recibe subsidio educativo (Sub_Edu = 1)
num_ingreso_horas_extra	Continua	Conteo con ingresos por horas extra (Ing_HE $>$ 0)
num_ingreso_bonificaciones	Continua	Conteo con ingresos por bonificaciones (Ing_Bon $>$ 0)
num_ingreso_primas	Continua	Conteo con ingresos por primas (Ing_Pr $>$ 0)

Tabla 9: Construcción Detallada de Variables - Sector Económico e Índices

Variable	Tipo	Construcción y Definición
num_agricultura	Continua	Conteo en sector agrícola (Act_principal_SP = 1)
num_industria	Continua	Conteo en sector industrial (Act_principal_SP = 2)
num_servicios	Continua	Conteo en sector servicios (Act_principal_SP = 3-6)
num_buscando_trabajo	Continua	Conteo buscando trabajo (Des = 1)
num_disponibles_trabajar	Continua	Conteo disponible para trabajar más horas (Disp_mas_horas = 1)
num_quieren_mas_horas	Continua	Conteo que quiere trabajar más horas (Quiere_mas_horas = 1)
vulnerability_index	Continua	$\frac{(1-\text{prop\_ocupados})+(\text{num\_minors}/\text{Nper})+(1-\text{prop\_cotizantes})+(1/(\text{prop\_cuartos}+0.1))}{4}$

Notas Técnicas Adicionales

- **Escalas Educativas:** Nivel educativo codificado como: 0=Ninguno, 1=Preescolar, 2=Primaria, 3=Secundaria, 4=Media, 5=Técnica, 6=Tecnológica, 7=Profesional, 8=Maestría, 9=Doctorado
- **Tratamiento de Missing:** Para variables laborales, missing se consideró como 0 (no aplica)
- **Normalización:** Todas las proporciones se acotaron entre 0 y 1 para evitar valores extremos
- **Consistencia:** Se verificó consistencia entre variables relacionadas (ej: num_occupied vs prop_ocupados)

Anexo B: Estadísticas Descriptivas Completas

Variables Demográficas y de Vivienda Adicionales

Tabla 10: Estadísticas Descriptivas: Variables Demográficas y de Vivienda (Anexo)

Variable	Total	No Pobres	Pobres	Diferencia
num_inactivos	1.028	0.923	1.446	***
tiene_vivienda	2.456	2.366	2.818	***
bin_headWoman	0.418	0.406	0.468	***
sexo_jefe	0.418	0.406	0.468	***
bin_rent	0.391	0.379	0.438	***
cuartos_dormir	1.989	1.992	1.979	**

Variables Educativas Adicionales

Tabla 11: Estadísticas Descriptivas: Variables Educativas (Anexo)

Variable	Total	No Pobres	Pobres	Diferencia
num_educacion_superior	0.774	0.879	0.357	***
num_educacion_media	1.300	1.208	1.668	***
num_sin_educacion	0.000	0.000	0.001	*

Variables Laborales y de Empleo Adicionales

Tabla 12: Estadísticas Descriptivas: Variables Laborales (Anexo)

Variable	Total	No Pobres	Pobres	Diferencia
num_patrones	0.558	0.642	0.222	***
num_trabajadores_tiempo_completo	1.181	1.264	0.851	***
num_agricultura	1.263	1.339	0.961	***
num_occupied	1.504	1.566	1.257	***
num_independientes	0.693	0.652	0.860	***
num_empleados_estables	1.018	1.075	0.790	***
num_empresas_pequenas	0.824	0.777	1.014	***
bin_occupiedHead	0.710	0.727	0.643	***
ocupacion_jefe	0.710	0.727	0.643	***
num_industria	0.082	0.066	0.147	***
num_empresas_grandes	0.122	0.137	0.062	***
num_trabajadores_medio_tiempo	0.223	0.212	0.265	***
num_trabajadores_domesticos	0.055	0.064	0.018	***
num_empleados_formales	0.007	0.009	0.001	***
num_empleados_informales	0.041	0.040	0.044	***

Variables de Subsidios e Ingresos Adicionales

Tabla 13: Estadísticas Descriptivas: Subsidios e Ingresos (Anexo)

Variable	Total	No Pobres	Pobres	Diferencia
num_ingreso_horas_extra	0.704	0.808	0.289	***
num_ingreso_bonificaciones	0.704	0.808	0.289	***
num_ingreso_primas	0.704	0.808	0.289	***
num_recibe_subsidio_transporte	0.330	0.391	0.087	***
num_recibe_subsidio_familiar	0.152	0.178	0.047	***
num_salud_subsidiado	0.120	0.144	0.027	***
num_recibe_subsidio_educativo	0.002	0.003	0.000	***

Variables de Búsqueda de Empleo e Índices

Tabla 14: Estadísticas Descriptivas: Búsqueda de Empleo e Índices (Anexo)

Variable	Total	No Pobres	Pobres	Diferencia
vulnerability_index	0.576	0.528	0.767	***
num_buscando_trabajo	0.182	0.148	0.319	***
num_quieren_mas_horas	0.132	0.122	0.171	***
num_disponibles_trabajar	0.125	0.116	0.164	***

Nota: Este anexo presenta estadísticas descriptivas de variables adicionales no incluidas en el análisis principal. *** indica diferencias estadísticamente significativas al 1 %, ** al 5 %, * al 10 %. La mayoría de las diferencias son significativas al 1 %, con excepción de num_sin_educacion (significativa al 10 %, *) y cuartos_dormir (significativa al 5 %, **). Las variables representan conteos, proporciones o índices construidos a nivel de hogar.