Maestría en Economía

Problem Set 2: Predicting Poverty

Bohorquez, Emiliano

Condori Luque, Brayan Alexis

Onofri, Milagros

Profesor: Ignacio Sarmiento Barbieri

Materia: Machine Learning

15 de Diciembre de 2024

1. Repositorio

Realizamos el trabajo en el siguiente repositorio de GitHub: https://github.com/milagrosonofri/Problem_Set_2.

2. Introducción

La predicción de la pobreza es un desafío crítico para la formulación de políticas públicas, ya que evaluaciones precisas de la misma permiten intervenciones focalizadas y una asignación eficiente de recursos.

Este *problem set* está basado en la iniciativa del Banco Mundial, "Pover-T Tests: Predicting Poverty". Esta propuesta subraya la importancia de desarrollar modelos predictivos que permitan diseñar encuestas más ágiles y específicas, logrando así evaluar la efectividad de políticas e intervenciones de manera rápida y económica. Contar con modelos más precisos no solo optimiza la asignación de recursos, sino que también permite ajustar las políticas públicas para maximizar su impacto y eficiencia.

El objetivo principal será predecir la pobreza a nivel de hogares en Colombia, utilizando datos provenientes del Departamento Administrativo Nacional de Estadística (DANE) y la "Misión para el Empalme de las Series de Empleo, Pobreza y Desigualdad" (MESEP). Según DANE, en 2023 la tasa de incidencia de pobreza a nivel nacional fue del 33 % [3].

En este contexto, se buscará construir un modelo predictivo que clasifique a los hogares como pobres o no pobres en función de si su ingreso se encuentra por debajo de una línea de pobreza definida. Este enfoque puede abordarse desde dos perspectivas: como un problema de clasificación, donde se predice la condición de pobreza directamente, o como un problema de predicción de ingresos, en el cual el ingreso estimado se compara con la línea de pobreza para determinar la clasificación final. Este *problem set* explorará ambas estrategias para alcanzar el objetivo planteado.

Para implementar este análisis, se utilizarán diversos algoritmos de aprendizaje supervisado que permiten abordar tanto la simplicidad como la complejidad del problema. En primer lugar, se realizará un análisis de **regresión lineal**, la cual servirá como un modelo de referencia intermedio para explorar correlaciones entre la variable dependiente y las distintas variables independientes consideradas. Además, se emplearán modelos **Elastic Net**, que combinan las penalizaciones de Ridge y Lasso, **CARTs** (árboles de decisión), **Random Forest** y **Regresión Logística** (Logit).

La combinación de estos algoritmos permitirá comparar y contrastar sus capacidades predictivas y explicativas. Este enfoque también facilitará la identificación de los factores más relevantes que determinan la pobreza en los hogares, proporcionando así un análisis robusto para la toma de decisiones basada en evidencia.

3. Antecedentes

El término pobreza hace referencia a carencias o privaciones que limitan el acceso a un nivel mínimo de vida digno. Según autores como Gasparini, Cicowiez y Sosa Escudero [7], la pobreza no solo es uno de los problemas sociales más graves, sino que su eliminación debería ser una prioridad central para cualquier sociedad. Su estudio adquiere especial relevancia debido a las consecuencias que tiene sobre otras dimensiones económicas y sociales.

Entre los efectos más destacados, la pobreza puede condicionar seriamente la acumulación de capital humano y otros factores productivos, afectando negativamente las perspectivas de crecimiento económico.

Además, sus implicancias van más allá del ámbito económico, contribuyendo a problemas de salud pública, inseguridad y, en algunos casos, a la inestabilidad democrática.

Su correcta medición es particularmente relevante para gobiernos y hacedores de política. Su utilidad responde a la importancia de identificar hogares en condición de vulnerabilidad y mejorar el alcance de programas sociales destinados a combatir la pobreza [1]. Una medición incorrecta puede conducir a pérdidas de eficiencia por sobreexclusión (o sobreinclusión) de participantes [11].

Como destaca Sosa Escudero [14], no existe una forma inequívoca de definir o medir la pobreza, ya que las estrategias tradicionales enfrentan importantes limitaciones en términos de cobertura y profundidad. La literatura económica ha explorado ampliamente formas de medir la pobreza para mejorar su eficacia: datos administrativos usando Proxy Mean Tests [8], mapas de pobreza con Machine Learning [2], *community-based* [9], entre otros.

En este contexto, la revolución tecnológica basada en Big Data y algoritmos de Machine Learning ofrece una oportunidad sin precedentes para mejorar las mediciones del bienestar y la pobreza. Estos enfoques permiten aprovechar datos no convencionales para obtener estimaciones más rápidas, precisas y económicas. Según Sosa Escudero, estas herramientas pueden superar las limitaciones de los sistemas tradicionales de encuestas, proporcionando una mayor granularidad y alcance en las estimaciones, especialmente en países en desarrollo donde los recursos son limitados [14].

Este marco conceptual enfatiza la importancia de utilizar herramientas rigurosas y enfoques metodológicos sólidos para comprender la dinámica de la pobreza y su relación con otras variables. Así, el presente análisis busca contribuir a esta agenda, aplicando técnicas modernas de predicción para abordar este problema desde una perspectiva innovadora y práctica.

4. Datos

El conjunto de datos utilizado en este proyecto proviene del Departamento Administrativo Nacional de Estadística (DANE) y de la "Misión para el Empalme de las Series de Empleo, Pobreza y Desigualdad" (MESEP).

La estructura del conjunto de datos incluye cuatro archivos principales, divididos en información a nivel de hogares e individuos, y en subconjuntos de entrenamiento y prueba:

- `train_hogares.csv`: Conjunto de entrenamiento con información a nivel de hogares.
- `train_personas.csv`: Conjunto de entrenamiento con información a nivel individual.
- `test_hogares.csv`: Conjunto de prueba con información a nivel de hogares.
- `test_personas.csv`: Conjunto de prueba con información a nivel individual.

4.1. Limpieza y Transformación

El proceso completo se dividió en dos etapas principales, implementadas en dos scripts de `Python`. En el primer script, `1_working_data`, se realizó la limpieza, transformación y construcción de las variables a partir de los datos originales. En el segundo script, `2_analisis`, se llevó a cabo el análisis exploratorio, la generación de visualizaciones y la implementación de modelos predictivos. Esta separación permitió organizar el flujo de trabajo de manera ordenada y facilitar la reproducibilidad del proyecto.

[1] Por ejemplo, en Colombia, el Sistema de Identificación de Potenciales Beneficiarios de Programas Sociales (SISBEN IV) clasifica a los hogares en cuatro grupos desde los más pobres a los no pobres. La metodología para el cálculo del indicador se encuentra detallado en el Documento CONPES 3877 [4].

- Se identificaron y reemplazaron valores faltantes codificados como `98` y `99` por valores `NaN`.
- Se eliminaron columnas redundantes resultantes de la fusión de tablas (`Dominio`, `Fex_c`, `Fex_dpto`).
- Se imputaron valores faltantes utilizando la mediana. Utilizamos la mediana ya que es una medida menos sensible a los outliers que la media.

4.2. Construcción de Variables

- **Variables Binarias:**
  - `subsidio`: Indicador binario que toma el valor 1 si el hogar recibe algún subsidio alimentario, educativo o de transporte.
  - `ing_capital`: Toma el valor 1 si el hogar recibe ingresos de capital o pensiones.
- **Variables del Jefe de Hogar:** A partir de individuos identificados como jefes de hogar (`P6050 = 1`), se construyeron las siguientes variables:
  - `anios_educ_head`: Años de educación del jefe de hogar.
  - `hs_work_head`: Horas semanales trabajadas por el jefe de hogar.
  - `subsidio_head`: Indicador binario si el jefe de hogar recibe subsidio.
  - `edad_head`: Edad del jefe de hogar.
  - Variables dummies relacionadas con el grupo ocupacional del jefe de hogar (`oficio_new`).
- **Proporciones Relativas:**
  - `mayoria_infancia`: Proporción de individuos menores de 15 años respecto al total de miembros del hogar.
  - `gen_relativo`: Proporción de mujeres respecto al total de miembros del hogar.
- **Condiciones de Vivienda:**
  - `hacinamiento`: Relación entre el número total de personas en el hogar (`Nper`) y el número de cuartos destinados a dormir (`P5010`).
  - `tipo_vivienda`: Se generaron variables dummy para identificar el tipo de vivienda: propia, alquilada, tomada o de otro tipo.

4.3. Fusión y Generación de los Conjuntos Finales

Los datos individuales y de hogares se fusionaron utilizando la clave `id`. Posteriormente, se incorporaron las variables construidas del jefe de hogar, así como las dummies correspondientes a los grupos ocupacionales y tipos de vivienda.

De esta forma, se generaron los conjuntos finales:

- `train_set_final.csv`: Conjunto de entrenamiento.
- `test_set_final.csv`: Conjunto de prueba.

4.4. Análisis descriptivo

4.4.1. Elección de variables

En primer lugar, definimos la pobreza como el porcentaje de personas cuyo ingreso per cápita familiar se encuentra por debajo de una línea de pobreza previamente establecida. Para determinar esto, llevamos a cabo una predicción del ingreso per cápita familiar.

Posteriormente, abordamos el análisis de la pobreza como un problema de clasificación, categorizando a los hogares en dos grupos: hogares pobres (asignados con un valor de 1) y hogares no pobres (asignados con un valor de 0).

En ambos enfoques, regresión y clasificación, utilizamos el mismo conjunto de variables explicativas. Sin embargo, en el caso del análisis de clasificación, excluimos la variable binaria "Dominio".

A continuación, se detallan las variables elegidas. Los gráficos A1 y A2 muestran las matrices de correlación.

- **Proporción de mujeres en el hogar (`gen_relativo`):** Esta variable actúa como una aproximación a la brecha de ingresos dentro del hogar. Se anticipa que a medida que aumenta la proporción de mujeres, el aporte relativo de estas al ingreso total del hogar tiende a ser menor, reflejando posibles disparidades salariales y, en consecuencia, capturando niveles de ingreso más bajos. La diferencia entre los salarios percibidos por hombres y mujeres es un tema ampliamente estudiado en la literatura [6]. De esta forma, se espera una relación negativa con el ingreso per cápita y positiva con la pobreza, confirmada por la matriz de correlación, aunque con valores muy moderados.

  Índice de hacinamiento (`hacinamiento`): Esta característica de condiciones de vivienda es incluida como una variable relevante en la medición de la pobreza multidimensional[2] Se espera que a mayor hacinamiento, incrementa la probabilidad de que el hogar sea pobre y se reduzca el nivel de ingreso per cápita. La matriz de correlaciones muestra esta relación.

- **Cantidad de miembros del hogar (`Nper`):** La fecundidad puede afectar a los hogares pobres en la medida que reduce el ingreso per cápita familiar (calculado como el ingreso total dividido por la cantidad de miembros). En América Latina, se ha documentado la existencia de una brecha entre el número de hijos de los percentiles más bajos y los más altos de la distribución [1]. Se espera que la relación sea negativa con el ingreso per cápita y positiva con la pobreza. La matriz de correlaciones muestra esta relación.

- **Proporción de miembros del hogar menores a 15 años (`mayoria_infancia`):** La cantidad de miembros que no están en edad de trabajar y, por tanto, no generan ingresos para el hogar, influye en la determinación del nivel de ingresos per cápita del hogar [1]. Se espera que su relación con el ingreso per cápita y con la pobreza, sea negativa y positiva, respectivamente. La matriz de correlación muestra esta relación, con valores significativos.

- **Nivel educativo del jefe de hogar** (`anios_educ_head`): La educación es uno de los determinantes más importantes y ampliamente estudiados en la literatura sobre ingresos. Se espera que exista una relación positiva y significativa entre el nivel educativo y el salario (componente importante del ingreso per cápita familiar), ya que mayores años de escolaridad incrementan las habilidades y productividad del individuo, traduciéndose en mayores ingresos. Esta relación está formalmente modelada en la Ecuación de Mincer [12], un modelo clásico en la economía laboral que estima los

[2] Colombia la utiliza como parte de la dimensión Çondiciones de Vivienda y Servicios Públicos"detallado en el Documento CONPES 160 [5].

retornos a la educación. Respecto a su relación con pobreza, se espera que la misma sea positiva. Efectivamente, ambas correlaciones se comprueban en la matriz.

- **Edad del jefe de hogar** (`edad_head`): La edad es un indicador de la experiencia laboral acumulada, la cual tiene un impacto significativo en los ingresos. Se espera una relación positiva entre la edad y el salario en las etapas iniciales de la vida laboral, ya que el aumento de la experiencia suele asociarse con mayores ingresos. Sin embargo, más allá de cierto punto, esta tendencia podría estabilizarse o incluso revertirse, reflejando un posible estancamiento o disminución en los ingresos con el envejecimiento. La ecuación de Mincer también tiene en cuenta la experiencia laboral. Respecto a su relación con pobreza, debido al mismo proceso descrito previamente, se espera una relación negativa, aunque no estrictamente lineal. Ambas relaciones se comprueban en la matriz de correlación.

- **Ingresos de capital (`ing_capital`):** Los retornos del capital han sido ampliamente estudiados como uno de los determinantes de la desigualdad, bajo un enfoque de capitalistas y trabajadores. Típicamente, los tenedores de capital se encuentran en los percentiles altos de la distribución [13], por tanto, se espera una correlación positiva con los ingresos y negativa con la pobreza. La matriz de correlación muestra claramente esta asociación.

- **Subsidio** (`subsidio_head`): Los hogares que reciben subsidios son, en general, hogares vulnerables. De esta forma, el subsidio puede actuar como un complemento del ingreso, aumentando el ingreso per cápita. Además, en hogares de bajos ingresos pueden ser determinantes para superar la línea de pobreza. La relación ambigua, si bien se espera un efecto riqueza que incremente el ingreso per cápita familiar y, en consecuencia, la pobreza se reduzca, la matriz de correlación muestra una relación nula con el ingreso per cápita y negativa con la pobreza.

- **Variables asociadas a la vivienda** (`P5130 y P5140`): reflejan el costo estimado de un arriendo de la vivienda (para quienes no pagan arriendo) y el valor del arriendo (para quienes si lo hacen). El alquiler es componente significativo en el gasto del hogar. Un arriendo más alto (real o estimado) puede estar asociado a mayores ingresos, dado que hogares con mayores recursos tienden a acceder a mejores viviendas. Sin embargo, un costo elevado de arriendo también podría reducir el ingreso disponible para otros gastos, afectando negativamente el ingreso per cápita familiar. La matriz de correlaciones muestra una correlación positiva débil con los ingresos per cápita y negativa con pobreza.

- **Variables asociadas a la propiedad de la vivienda** (`P5090`): afectan el ingreso per cápita familiar al determinar los gastos asociados a la vivienda. Al mismo tiempo, permiten determinar el grado de vulnerabilidad del hogar. Los hogares propietarios pueden disponer de mayores ingresos si no pagan alquiler, mientras que los hogares arrendatarios ven reducido su ingreso disponible por los costos del mismo. En contraste, viviendas en usufructo o posesión sin título pueden implicar menores gastos, aunque a menudo están asociadas a condiciones de mayor precariedad.

- **Dummies de Oficio** (`oficio_new`): Es lógico pensar que existirá una amplia variabilidad en los ingresos entre sectores de ocupación. Asimismo, se ha estudiado ampliamente en la literatura las brechas de ingresos entre trabajadores calificados y no calificados, así como el lugar que ocupan en la distribución de ingresos. Típicamente, se espera que los hogares con el jefe de hogar ocupado en empleos de mayor calificación tengan un ingreso per cápita mayor. Por tanto, dependiendo de la categoría de oficio, la correlación con la variable de ingreso per cápita y pobreza puede ser positiva o negativa.

- **Dummies regionales** (`Dominio`): Indican la ubicación geográfica de los hogares, pueden afectar significativamente el ingreso per cápita familiar debido a las variaciones en el costo de vida, las oportunidades económicas y las políticas locales. Las correlaciones con la variable de ingreso puede

ser positiva o negativa, dependiendo de la región.

4.5. Estadísticas descriptivas

La tabla A1 muestra las estadísticas descriptivas para las variables incluidas en los modelos de clasificación y predicción.

Respecto a las características del hogar, observamos que la proporción de mujeres en el hogar tiene una media de 0,53, lo que indica que, en promedio, los hogares mantienen un equilibrio en la proporción de género. El tamaño promedio del hogar es 3,29 personas, donde la proporción promedio de menores de 15 años en el hogar es de 19 %, lo que sugiere que una proporción importante de personas se encuentran en relación de dependencia económica en el hogar.

Las variables relacionadas con la vivienda, nos brindan detalles relevantes para la caracterización del ingreso per cápita y la pobreza. El valor del arriendo mensual tiene como media 437.911 pesos colombianos, con una desviación estándar de 1.447.543 y valores mínimos de 20 y 300 millones, sugieren una alta desigualdad en los costos de vivienda. Es preciso resaltar que el arriendo mensual promedio representa más del 50 % del ingreso per cápita promedio de la unidad de gasto (870,639 pesos). Además, el 41 % de los hogares son propietarios de sus viviendas, lo que implica que más de la mitad de hogares recurre al alquiler o, en su defecto, a habitar viviendas de manera informal (viviendas tomadas, 5 %).

Con un promedio de 6.1 años de educación, los jefes de hogar alcanzan un nivel equivalente a la primaria o secundaria incompleta. Este nivel educativo afecta directamente las oportunidades laborales, limitando el acceso a trabajos formales y mejor remunerados. Además, el promedio de horas trabajadas por semana alcanza 46,91 con una desviación estándar de 15,32 horas. Asimismo, resulta alarmante el valor máximo de horas trabajadas de 130, lo cual sugiere que existe una dispersión importante en los datos.

Por otro lado, la tabla incluye una descomposición de ocupación donde la categoría de oficios administrativos y comerciales es la más frecuente (17 %), seguida por servicios personales y comunitarios (14 %). Por otro lado, las categorías típicamente asociadas a ingresos altos son las de menor frecuencia como Arte, Cultura y Medios (1 %) y Ciencias Biológicas y de la Salud (1 %).

5. Modelos y Resultados

5.1. Modelos

5.1.1. Regresión Lineal

A través de una regresión lineal podemos analizar la relación entre una variable dependiente Y y múltiples variables predictoras $X_1, X_2, \ldots, X_p$. En el contexto de este problema, la regresión lineal se utiliza para modelar los ingresos per cápita del hogar (Y) en función de diversas características socioeconómicas y del hogar (X), así como *dummies* por región. Contamos con 49 variables predictoras, las mismas que se detallan en la tabla A1.

La ecuación general del modelo es la siguiente:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon,$$

donde:

- Y: Es la variable dependiente.
- β_0: Es el término de intercepto, que representa el valor esperado de Y cuando todas las variables X_j son iguales a 0.
- β_j: Son los coeficientes de regresión asociados con las variables predictoras X_j, que miden el cambio promedio en Y por un incremento unitario en X_j, manteniendo constantes las demás variables.
- ϵ: Es el término de error, que captura la variabilidad en Y no explicada por las variables X.

Para estimar los parámetros, se elige $\beta_0, \beta_1, \ldots, \beta_p$ con el objetivo de minimizar la suma de residuos al cuadrado (RSS):

$$\min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p)^2$$

Los valores $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ que minimizan esta expresión son las estimaciones de mínimos cuadrados de los coeficientes de regresión.

Este enfoque se basa en minimizar el error "dentro de la muestra". De esta forma, aunque la regresión lineal es una herramienta útil para identificar correlaciones y comprender el impacto promedio de las variables predictoras, su capacidad predictiva fuera de la muestra puede ser limitada.

En este contexto, este modelo servirá como punto de referencia inicial para nuestro análisis, proporcionando una base para evaluar la necesidad de modelos más complejos. En particular, motiva el uso de técnicas de regularización como Elastic Net, que optimizan la capacidad predictiva al penalizar la magnitud de los coeficientes y reducir el riesgo de sobreajuste.

5.1.2. Elastic Net

Elastic Net es un modelo de regularización que combina las propiedades de Ridge y Lasso, proporcionando un equilibrio entre ambos. La función objetivo de Elastic Net incorpora una penalización que incluye tanto la suma de los valores absolutos de los coeficientes como el cuadrado de los coeficientes, lo que permite manejar problemas como la correlación entre variables y realizar selección de predictores simultáneamente. Su función de optimización se define como:

$$\min_{\beta} EN(\beta) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \left(\alpha \sum_{j=1}^{p} |\beta_j| + \frac{(1-\alpha)}{2} \sum_{j=1}^{p} \beta_j^2 \right),$$

donde:

- λ es el parámetro que controla el grado de penalización global.
- α define el balance entre Lasso y Ridge, donde $\alpha = 1$ equivale a Lasso y $\alpha = 0$ a Ridge.

De acuerdo con Zou y Hastie [15], Elastic Net supera a Lasso al incorporar un efecto de agrupamiento, lo que significa que predictores altamente correlacionados tienden a ser seleccionados juntos o excluidos simultáneamente. Esto contrasta con Lasso, que tiende a seleccionar solo uno de los predictores correlacionados.

5.1.3. Árboles de Decisión

Árboles de Decisión es un método para regresión y clasificación que utiliza los predictores para dividir el espacio en función de reglas de decisión. En otras palabras, cada nodo se forma en función de una nueva decisión, basado en una predicción que toma como argumento el valor de una variable. Primero se analizarán los árboles para regresión y posteriormente los de clasificación.

Árboles de Regresión

Para regresar una variable dependiente continua, esta metodología se basa en segmentar el espacio de entrada en regiones disjuntas y ajusta un valor basado en la media de la distribución de los datos para predecir cada región.

Su principal ventaja, respecto al modelo lineal clásico, es que no requiere linealidad, es decir, la relación entre las variables predictoras y la variable predicha puede ser no lineal. Además, captura interacciones entre variables sin necesidad de especificación. Por último, cuenta con una estructura de interpretación accesible, mostrando cómo las reglas de decisión determinan los resultados.

La división en nodos se realiza a través de criterios que minimicen el error. Para regresión se utiliza la suma de residuos cuadráticos:

$$SCR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Con y_i siendo los valores observados de la variable dependiente, y siendo $\hat{y}_i$ los valores predichos del modelo.

Por último, para evitar el sobreajuste, es común limitar la profundidad del árbol, como así también el número mínimo de muestras necesarias y el mínimo de muestras que debe tener una hoja posterior a dividir el nodo.

Árboles de Clasificación

Un Árbol de Clasificación se utiliza para predecir variables binarias, dividiendo los datos en regiones según criterios que maximizan la pureza de las clases en cada nodo.

Dado su naturaleza no paramétrica, no requiere supuestos sobre la distribución de los datos. Además, también es de fácil interpretabilidad, puesto que las reglas de decisión se desarrollan en función de la variable binaria que busca predecirse.

Los dos criterios de división más usados son:

- **Índice de Gini:** indica el nivel de pureza de las clases en función de la heterogeneidad de las mismas.

$$G = \sum_{i=1}^{K} p_k 1 - p_k)$$

  Donde p_k es la proporción de observaciones de la clase k en un nodo. Un valor cercano a cero indica un mayor nivel de pureza, mientras que lo contrario aplica para valores cercanos a 1.

- **Índice de Entropía:** indica el nivel de incertidumbre en el nodo.

$$\mathrm{E} = -\sum_{k=1}^{K} p_k \log(p_k)$$

  Donde p_k es la proporción de observaciones de la clase k en un nodo. Un valor cercano a cero indica, nuevamente, un mayor nivel de pureza, mientras que para valores cercanos a 1 indica una distribución uniforme de las clases.

Adicionalmente, la clasificación puede evaluarse mediante la Curva ROC, una representación del rendimiento del modelo para clasificar. La misma se obtiene tras graficar la tasa de verdaderos positivos contra la tasa de faltos positivos.

5.1.4. Random Forest

Random Forest es una versión sencilla de *bagging* que toma un subconjunto de los predictores y se basa en la construcción de múltiples árboles de decisión utilizando muestras aleatorias de los datos de entrenamiento. Cada vez que se considera una división en un árbol, se selecciona un subconjunto aleatorio de predictores entre todos los disponibles (si hay p predictores, en cada partición se usan $m < p$). El número de árboles es un hiperparámetro a encontrar.

Esta metodología para clasificación toma los mismos criterios que los árboles utilizados para clasificar. El funcionamiento se centra en realizar múltiples remuestreos para reemplazar el conjunto original y así entrenar a cada árbol de decisión. Posteriormente, se realiza una selección aleatoria de características posibles para introducir heterogeneidad y minimizar la correlación entre árboles. Por último, y dado que nos referimos a clasificación, las clases tomadas para la predicción son aquellas que fueron elegidas por más árboles.

Los hiperparámetros importantes son:

- El número de árboles, el cual marca un trade-off entre precisión y costo computacional.
- La máxima profundidad de cada árbol para evitar sobreajustes.
- Número de características, donde una menor cantidad incrementa la heterogeneidad entre los árboles.
- El tamaño mínimo de hojas para evitar árboles específicos.

5.1.5. Regresión Logística

La regresión logística es un modelo estadístico usado para predecir una variable categórica binaria en función del conjunto de predictores elegidos. Para modelar la probabilidad de pertenencia a uno de los grupos se utiliza la función logística:

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)}}$$

A diferencia del modelo lineal, utiliza Maxima verosimilitud, para medir la probabilidad de ajuste de los datos al modelo. Además, utiliza como criterios de evaluación la precisión de la estimación y la Curva ROC (mencionada más arriba). También se debe seleccionar un umbral, que por defecto es $0,5$ para clasificar si una observación pertenece a un grupo (igual a 1) o a otro (igual a 0).

5.2. Resultados

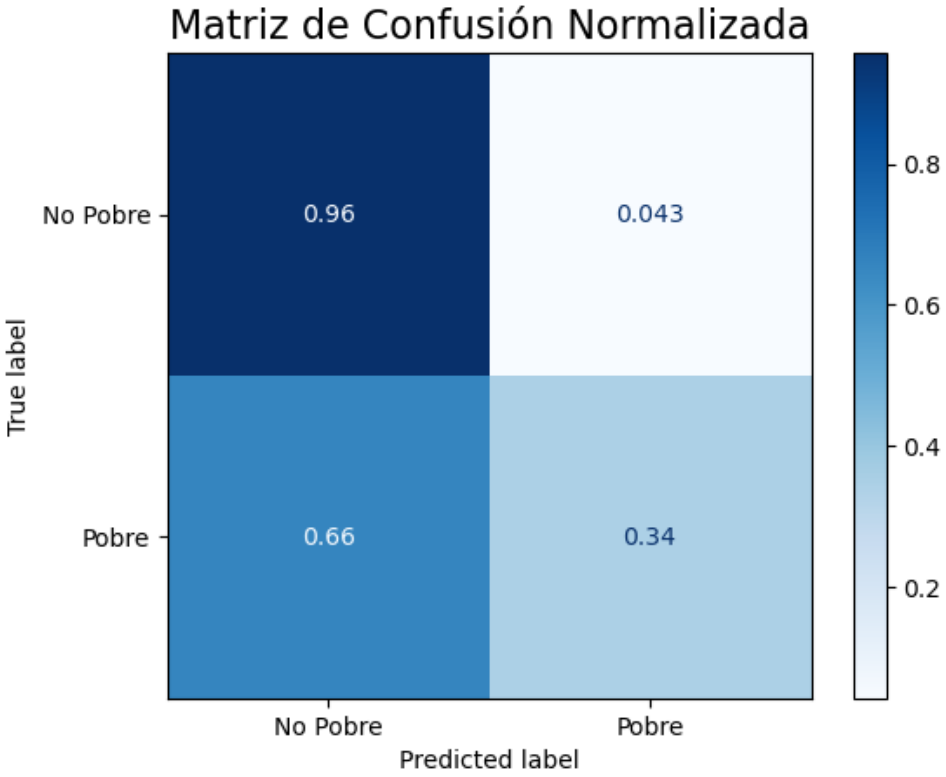
5.2.1. Regresión

Para el problema de clasificación usando la predicción de ingresos, realizamos la predicción mediante (i) regresión lineal y (ii) ElasticNet.

Para el modelo de regresión lineal, los coeficientes asumimos una relación lineal entre las variables. El modelo elegido es el polinomio de grado 2 para los 49 predictores que minimiza el RMSE y tiene un R cuadrado de 0.42. No es de extrañar que el poder predictivo del modelo sea bajo. Sin embargo, el análisis de los coeficientes permite identificar correlaciones clave y relaciones subyacentes entre las variables, lo cual constituye un punto de partida valioso para la construcción de modelos más sofisticados.

El gráfico 1 corresponde a la matriz de confusión para el modelo de regresión lineal. En total, la tasa de verdaderos positivos es de 22 %, frente a una identificación correcta del 98 % para los no pobres.

Figura 1: Matriz de Confusión - Regresión lineal



Asimismo, se utilizó el modelo ElasticNet para ajustar los datos, implementando una búsqueda de hiperparámetros con validación cruzada (5 folds) para optimizar la combinación de regularización. Se probaron diferentes valores para el hiperparámetro $\lambda \in [0{,}1, 0{,}5, 0{,}7, 0{,}9, 1{,}0]$ y para el parámetro de penalización αin $[0{,}1, 0{,}5, 1{,}0, 5{,}0, 10{,}0, 20{,}0]$. Este enfoque permitió encontrar la configuración que minimiza el error en el conjunto de entrenamiento.

5.2.2. Clasificación

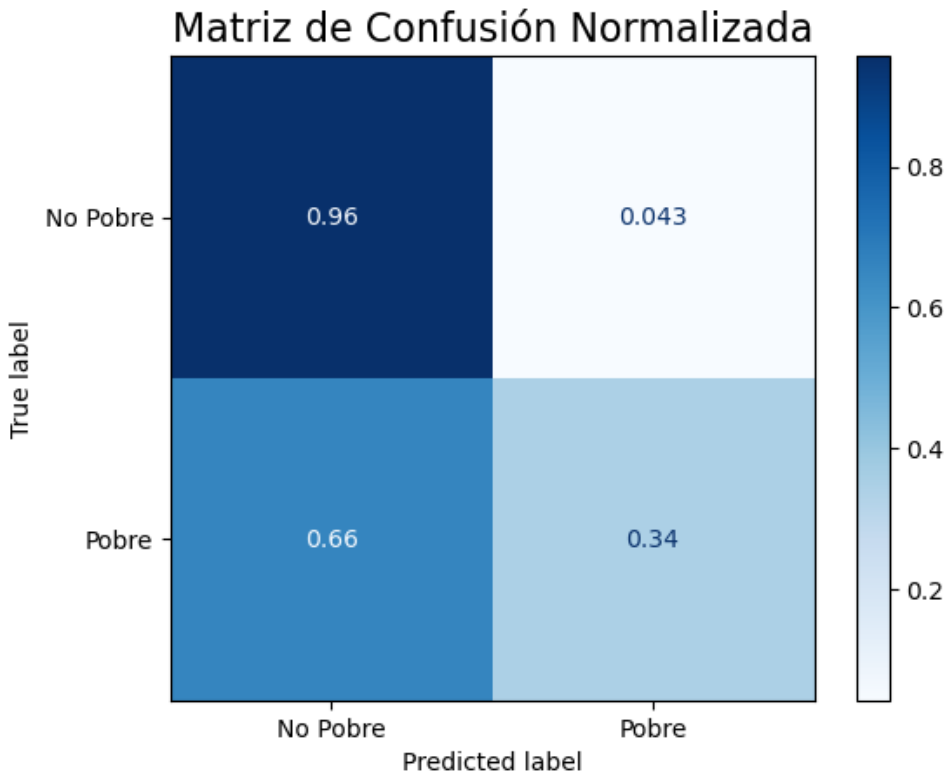
En la siguiente sección se busca predecir la clasificación de un hogar sobre si es pobre o no, es decir, bajo una serie de características relevantes, encontrar la proporción de hogares que están en dicha situación.

Los modelos implementados son: (i) Árboles de Decisión para clasificación; (ii) Random Forests para clasificación; y (iii) Regresión Logística.

Para **Árboles de Clasificación**, en primer lugar se utilizó como variables predictoras la imputación del costo por alquilar estimado para propietarios, el monto abonado por alquilar la propiedad, la proporción de menores de 15 años en el hogar, el nivel de hacinamiento del mismo, el número de personas en la unidad de gasto junto con el total de convivientes en total, la proporción de mujeres en el hogar, Una variable indicadora si la vivienda es propia, otra si es alquilada y otra si es ocupada sin título. Para este modelo, la tasa de verdaderos positivos fue del 29 %, mientras que la precisión en identificar a los que no son pobres fue del 96 %. La precisión global fue del 83 %.

Para mejorar la performance, se incorporaron variables relacionadas con el jefe/a del hogar. Estas son: la edad, el oficio (variable binaria por oficio registrado en la encuesta), la cantidad de años de educación, la cantidad de horas trabajadas en la última semana y si cuenta con ingresos de capital o rentas. Aquí, la tasa de verdaderos positivos mejoró, subiendo a un 34 %, aunque sin modificación en la identificación de los no pobres (96 %). La precisión global del modelo es de 84 %, no evidenciando mejoras significativas en este punto. El cuadro 1 resume la relevancia (en porcentaje) de cada predictor que haya superado el 5 %. Por otra parte, el gráfico 2 es la matriz de confusión correspondiente.

Figura 2: Matriz de Confusión - Árbol de Decisión



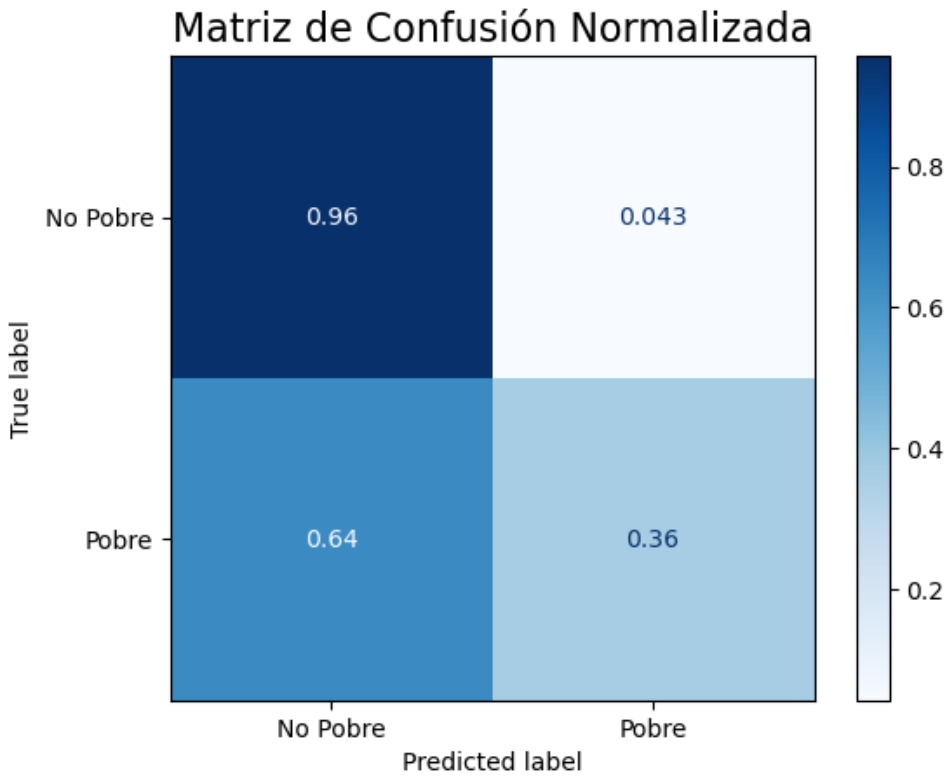
Cuadro 1: Importancia de los predictores

Rango	Predictor	Importancia
3	mayoria_infancia	0.31
8	imput_vivienda	0.19
7	subsidio_head	0.12
14	hacinamiento	0.08
1	monto_alquiler	0.08
6	ing_capital	0.06

Siguiendo la misma lógica que el modelo anterior, para **Random Forests** primero se procedió sin contar con las variables vinculadas al jefe de hogar. Aquí, la tasa de verdaderos fue de 29 %, mientras que la precisión en la identificación de los no pobres fue de 97 %. La precisión del global fue de 83 %.

Incorporando las variables del jefe de hogar, el modelo mejora la tasa de falsos positivos en 7 puntos (36 %), mientras que reduce un punto la identificación de los no pobres (96 %). La precisión global fue de 84 %, no demostrando ganancias importantes en este punto. Para más información, véase el cuadro 2 con la importancia de los predictores y el gráfico 3 con la matriz de confusión.

Figura 3: Matriz de Confusión - Random Forest

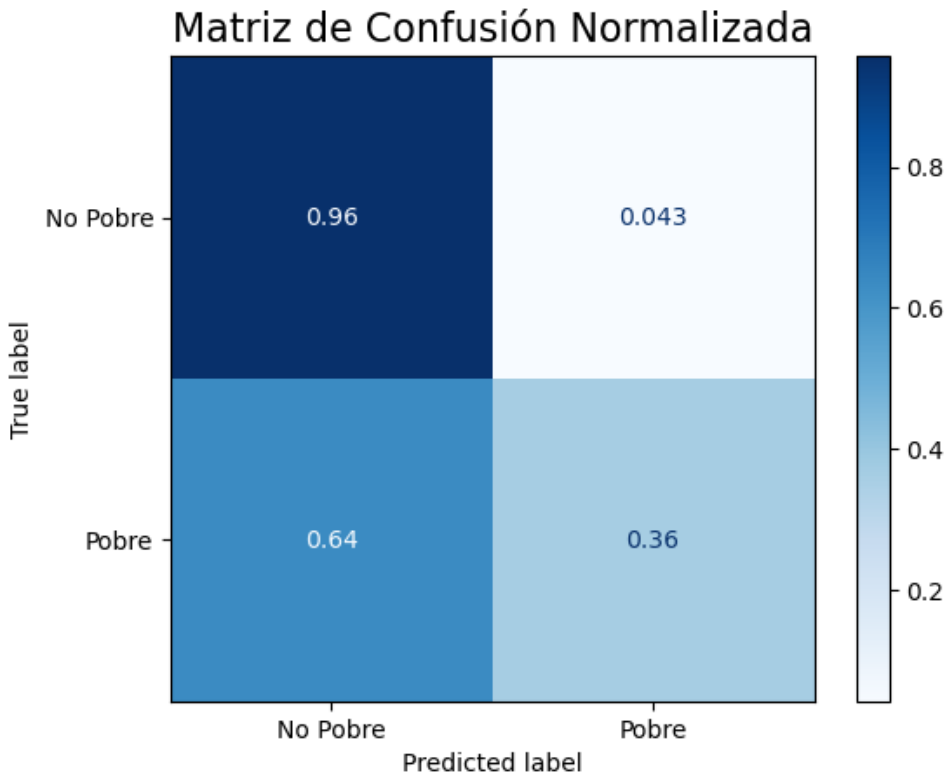


Cuadro 2: Importancia de los predictores

Rango	Predictor	Importancia
3	mayoria_infancia	0.32
14	hacinamiento	0.14
8	imput_vivienda	0.14
7	subsidio_head	0.11
1	monto_alquiler	0.09

Por último, para **Regresión Logística** la tasa de verdaderos positivos con el primer conjunto de variables fue de 26 %, frente a 97 % de identificación de no pobres, contando con un *accuracy* global de 83 %. Incorporando los predictores del jefe/a de hogar, la tasa de verdaderos positivos asciende a 37 % y la identificación de no pobres decrece a 95 %. La precisión global fue de 84 %. El cuadro 3 y el gráfico 4 resumen la tabla de relevancia de predictores y la matriz de confusión, respectivamente.

Figura 4: Matriz de Confusión - Logistic Regression

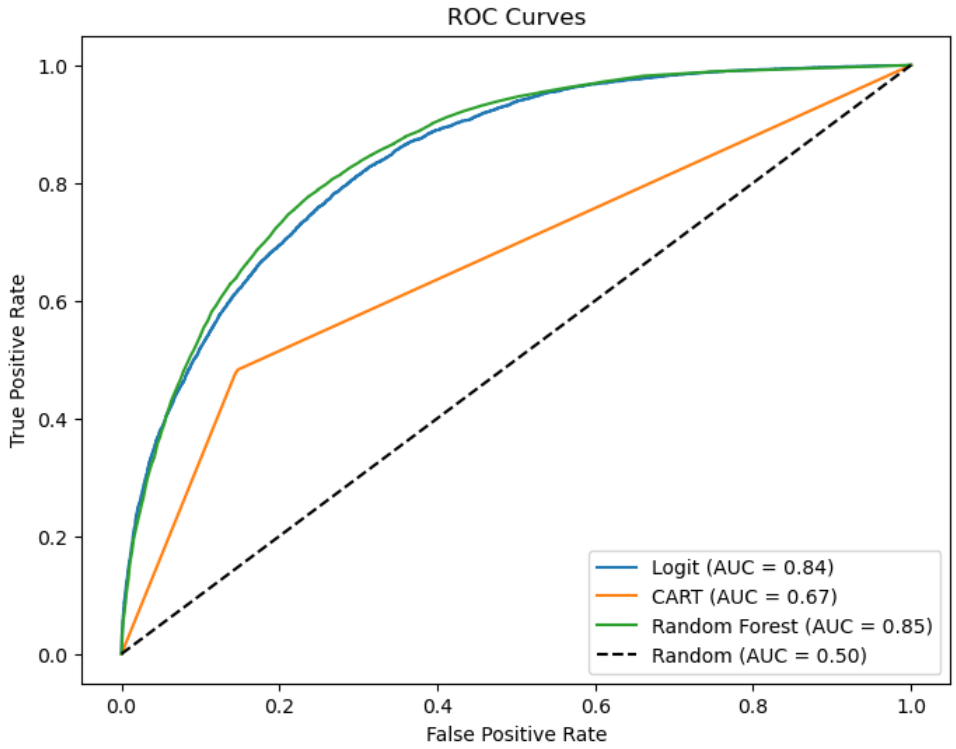


Cuadro 3: Importancia de los predictores

Predictor	**Importancia Relativa (%)**
imput_vivienda	0.20
monto_alquiler	0.18
Npersug	0.10
Nper	0.09
ing_capital	0.05

Por último, el gráfico 5 representa la Curva ROC para medir el rendimiento entre modelos. Más cercanos al extremo superior izquierdo, sería una mejor performance predictiva. Como se observa en la imagen, el área más grande corresponde al modelo de Random Forest, lo cual condice con sus ventajas para una predicción más eficiente.

Figura 5: Curvas ROC de los modelos



6. Conclusión y Recomendaciones

El presente trabajo ha explorado la predicción de la pobreza en hogares colombianos a partir de técnicas modernas de aprendizaje automático y análisis estadístico. Se implementaron diversos modelos predictivos que, además de evaluar la capacidad de predicción de los ingresos y la clasificación de los hogares como pobres o no pobres, permitieron identificar los factores más relevantes que inciden en la pobreza. Esta aproximación ofrece una herramienta valiosa para la formulación de políticas públicas focalizadas y la asignación eficiente de recursos.

Respecto a los modelos de regresión, las raíces de los errores cuadráticos medios son significativas, indicando una proporción mayor de error en la predicción respecto a la media de ingreso per cápita familiar del conjunto de prueba. Por otro lado, los modelos de clasificación identificaban correctamente a hogares en situación de pobreza entre un 25 % y un 37 %, siendo el modelo Random Forests el de mejor performance predictiva según lo visto en la Curva ROC. En paralelo, el modelo de Regresión Logística tiene la tasa de verdaderos positivos más alta (37 %). Sin embargo, estas predicciones producen una subestimación de la pobreza, por lo que se recomienda como insumo de política pública la incorporación de variables no contempladas aquí, como, por ejemplo, si la madre del hogar trabaja o realiza tareas domésticas, si los niños/as están escolarizados o cuentan con sobreedad y/o están en situación de abandono, y, principalmente, si la fuente principal de ingresos proviene de un trabajo en situación de formalidad.

Reconocemos las limitaciones del presente análisis, pero este constituye un primer paso hacia la construcción de herramientas más robustas y precisas para la medición de la pobreza. De esta forma y como se mencionó anteriormente, incorporar nuevas variables contribuiría a mejorar las predicciones y a proporcionar una base más sólida para el diseño e implementación de políticas públicas más efectivas.

Referencias

Leonardo y Marchionni Mariana Badaracco, Nicolás y Gasparini. Distributive implications of fertility changes in latin america. Technical Report 206, CEDLAS, Universidad Nacional de La Plata, January 2017. URL http://www.cedlas.econo.unlp.edu.ar.

Heath y Segovia Sandra Corral, Paul y Henderson. Mapeo de la pobreza en la era del aprendizaje automático. *Journal of Development Economics*, 172:103377, 2025.

Departamento Administrativo Nacional de Estadística (DANE). Presentación de resultados: Pobreza monetaria 2023, December 2023. URL https://www.dane.gov.co/files/operaciones/PM/pres-PM-2023.pdf. Accedido el 15 de diciembre de 2024.

Documento CONPES 3877. Declaración de importancia estratégica del sistema de identificación de potenciales beneficiarios (sisben iv), December 2016.

Documento CONPES Social 160. Metodologías oficiales y arreglos institucionales para la medición de la pobreza en colombia, May 2012.

Mariana Gasparini, Leonardo y Marchionni, editor. *Bridging Gender Gaps? The Rise and Deceleration of Female Labor Force Participation in Latin America*. CEDLAS, Universidad Nacional de La Plata, La Plata, Argentina, 2015. URL http://www.cedlas.econo.unlp.edu.ar. Análisis sobre la participación laboral femenina y las disparidades de género en América Latina.

Martín y Sosa Escudero Walter Gasparini, Leonardo y Cicowiez. *Pobreza y Desigualdad en América Latina: Conceptos, herramientas y aplicaciones*. CEDLAS, Universidad Nacional de La Plata, La Plata, Argentina, 2010.

Judy L. Grosh, Margaret y Baker. Pruebas de medios proxy para enfocar programas sociales. *Living Standards Measurement Study Working Paper*, 118:1–49, 1995.

Stefan y Pacere Noraogo A. Hillebrecht, Michael y Klonner. La dinámica del enfoque de la pobreza. *Journal of Development Economics*, 161:103033, 2023.

Daniela y Hastie Trevor y Tibshirani Robert James, Gareth y Witten. *An Introduction to Statistical Learning*. Springer Texts in Statistics. Springer, New York, 1st edition, 2013. ISBN 978-1-4614-7138-7. URL https://www.statlearning.com/.

Stephen Kidd. Exclusión social y acceso a esquemas de protección social. *Journal of Development Effectiveness*, 9(2):212–244, 2017.

Jacob Mincer. *Schooling, Experience, and Earnings*, volume 2 of *Human Behavior and Social Institutions*. National Bureau of Economic Research, New York, 1974. Trabajo seminal sobre la relación entre educación, experiencia e ingresos.

Thomas Piketty. *Capital in the Twenty-First Century*. Harvard University Press, Cambridge, Massachusetts, 2014. ISBN 978-0674430006. Trabajo seminal que analiza la desigualdad de riqueza e ingresos en la era moderna.

Walter Sosa Escudero. Big data y algoritmos para la medición de la pobreza y el desarrollo. CEDLAS Working Papers 319, CEDLAS, Universidad Nacional de La Plata, October 2023. URL https://www.cedlas.econo.unlp.edu.ar/wp-content/uploads/2023/10/doc_cedlas319.pdf.

Trevor Zou, Hui y Hastie. Regularización y selección de variables mediante la red elástica. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

A. Anexos

Cuadro A1: Estadísticas descriptivas

Variable	Media	Desvío Estándar	Observaciones	Mínimo	Máximo
Proporción de mujeres en el hogar	0.53	0.28	164960	0.00	1.00
Valor del arriendo (mensual)	437911.80	1447543.24	64453	20.00	300000000.00
Cantidad de miembros del hogar	3.29	1.77	164960	1.00	28.00
Proporción de miembros menores de 15 años	0.19	0.22	164960	0.00	1.00
Años de educación del jefe de hogar	6.10	3.58	164939	0.00	15.00
Edad del jefe de hogar	49.61	16.39	164960	11.00	108.00
Percibe ingreso de Capital (=1)	0.13	-	164960	-	-
Percibe subsidio (=1)	0.16	-	164960	-	-
Arriendo estimado	553046.74	4375785.97	90836	2000.00	600000000.00
Cantidad de personas en la unidad de gasto	3.28	1.77	164960	1.00	28.00
Vivienda alquilada (=1)	0.54	-	164960	-	-
Vivienda propia (=1)	0.41	-	164960	-	-
Vivienda tomada (=1)	0.05	-	164960	-	-
Horas trabajadas por el jefe de hogar (semana)	46.91	15.32	117156	1.00	130.00
Oficio: Administración y Gestión (=1)	0.02	-	164960	-	-
Oficio: Agricultura, Pesca y Explotación de Recursos Naturales (=1)	0.07	-	164960	-	-
Oficio: Arte, Cultura y Medios (=1)	0.01	-	164960	-	-
Oficio: Ciencias Biológicas y de la Salud (=1)	0.01	-	164960	-	-
Oficio: Ciencias Sociales y Humanidades (=1)	0.05	-	164960	-	-
Oficio: Ciencias, Ingeniería y Tecnología (=1)	0.02	-	164960	-	-
Oficio: Industria, Manufactura y Construcción (=1)	0.13	-	164960	-	-
Oficio: Oficios Administrativos y Comerciales (=1)	0.17	-	164960	-	-
Oficio: Otros Oficios y Trabajos No Clasificados (=1)	0.00	-	164960	-	-
Oficio: Servicios Personales y Comunitarios (=1)	0.14	-	164960	-	-
Oficio: Transporte y Logística (=1)	0.08	-	164960	-	-
Región: Barranquilla (=1)	0.04	-	164960	-	-
Región: Bogotá (=1)	0.06	-	164960	-	-
Región: Bucaramanga (=1)	0.03	-	164960	-	-
Región: Cali (=1)	0.04	-	164960	-	-
Región: Cartagena (=1)	0.03	-	164960	-	-
Región: Cúcuta (=1)	0.03	-	164960	-	-
Región: Florencia (=1)	0.03	-	164960	-	-
Región: Ibagué (=1)	0.03	-	164960	-	-
Región: Manizales (=1)	0.04	-	164960	-	-
Región: Medellín (=1)	0.05	-	164960	-	-
Región: Montería (=1)	0.03	-	164960	-	-
Región: Neiva (=1)	0.03	-	164960	-	-
Región: Pasto (=1)	0.03	-	164960	-	-
Región: Pereira (=1)	0.03	-	164960	-	-
Región: Popayán (=1)	0.04	-	164960	-	-
Región: Quibdo (=1)	0.02	-	164960	-	-
Región: Resto Urbano (=1)	0.10	-	164960	-	-
Región: Riohacha (=1)	0.03	-	164960	-	-
Región: Rural (=1)	0.09	-	164960	-	-
Región: Santa Marta (=1)	0.04	-	164960	-	-
Región: Sincelejo (=1)	0.03	-	164960	-	-
Región: Tunja (=1)	0.03	-	164960	-	-
Región: Valledupar (=1)	0.03	-	164960	-	-
Región: Villavicencio (=1)	0.03	-	164960	-	-
Ingreso per cápita de la unidad de gasto	870639.26	1244349.74	164960	0.00	300000000.00

Figura A1: Matriz de Correlación - Predicción

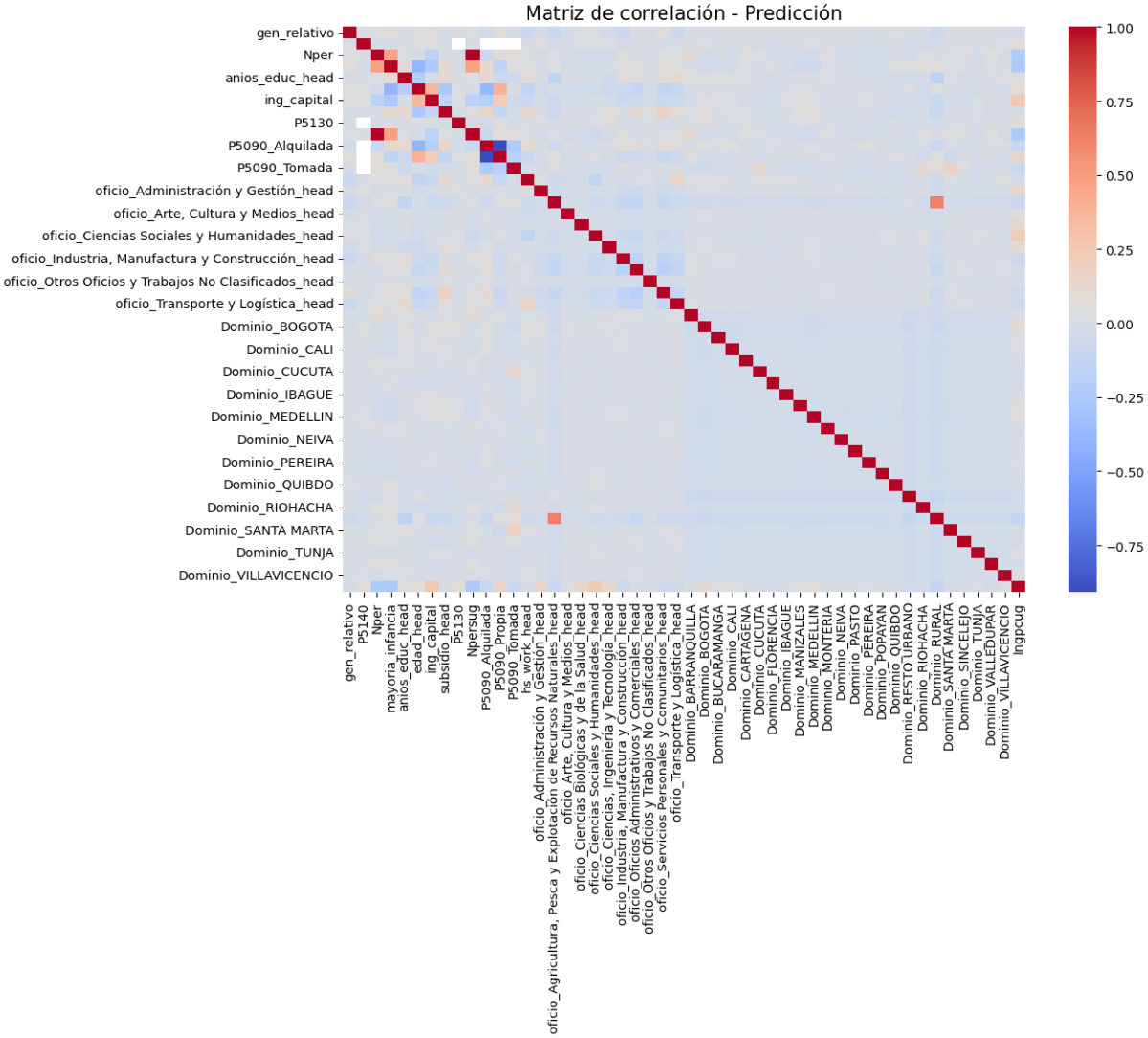


Figura A2: Matriz de Correlación - Clasificación