UNLP - Machine Learning - 2024

Problem Set 2: Predicting Poverty

Victoria Mascaró

1. Introducción

La pobreza es un fenómeno complejo que puede medirse de diferentes formas, dependiendo del contexto y los objetivos del análisis. Los métodos para medirla se han diversificado a lo largo del tiempo, y se pueden agrupar en dos enfoques generales propuestos por Sen (1979): el **método directo** y el **método indirecto**.

El **método directo** consiste en identificar a los hogares o individuos que no satisfacen un conjunto específico de necesidades previamente definidas, tales como condiciones de vivienda, acceso a la educación, composición demográfica del hogar, y tenencia de activos, entre otras. Ejemplos de este enfoque incluyen el Índice de Necesidades Básicas Insatisfechas (NBI) desarrollado por la CEPAL (Comisión Económica para América Latina y el Caribe) y el Índice de Pobreza Multidimensional (IPM), que capturan la pobreza desde una perspectiva multidimensional.

Por otro lado, el **método indirecto** se basa en la determinación de un umbral mínimo de ingresos o gasto, debajo del cual se considera que una persona no puede satisfacer sus necesidades básicas. Este método está asociado a la medición de pobreza monetaria a través de la definición de líneas de pobreza e indigencia. La línea de indigencia corresponde al valor de una canasta básica de alimentos; mientras que la línea de pobreza corresponde al valor de una canasta que incluye alimentos y además otros bienes básicos. Esta medición consiste en determinar cuántos hogares carecen de ingresos suficientes para comprar dichas canastas.

En países desarrollados, se suele emplear la pobreza relativa, que compara el ingreso de un hogar con el promedio de la sociedad. Sin embargo, en contextos de países en Desarrollo, donde los niveles de pobreza suelen ser estructurales, las aproximaciones más relevantes son las dos primeras.

El objetivo principal de este trabajo es predecir el nivel de pobreza en Colombia, utilizando para ello, modelos predictivos basados en datos económicos cuantitativos.

Este análisis adopta la perspectiva del Banco Mundial, que mide la pobreza como una función del ingreso, estableciendo una línea de pobreza para clasificar a los hogares como pobres o no pobres:

$$Pobre = I(Inc < Pl)$$

donde $I()$ es una función indicadora que toma el valor de 1 si el ingreso del hogar está por debajo de un umbral de pobreza específico Pl.

Esto sugiere dos enfoques principales para predecir la pobreza:

1. **Como un problema de clasificación**: Predecir valores de 0 (no pobre) y 1 (pobre).
2. **Como un problema de predicción de ingresos**: Con el ingreso predicho, usar la línea de pobreza para clasificar los hogares.

En este trabajo, se explorarán ambos enfoques.

Los datos utilizados provienen del Departamento Administrativo Nacional de Estadística (DANE) de Colombia y de la misión para el "Empalme de las Series de Empleo, Pobreza y Desigualdad - MESE". Estos datos están organizados en cuatro conjuntos: entrenamiento y prueba, tanto a nivel de hogar como a nivel individual. La disponibilidad de datos en dos niveles permite una mayor flexibilidad analítica, ya que la información individual puede ser procesada y transformada en variables agregadas que enriquecen las predicciones a nivel de hogar.

Para abordar el problema de clasificación, se utilizaron datos a nivel de hogar, con modelos como Logistic Regression y Gradient Boosting Classifier. Por otro lado, para abordar el problema desde la perspectiva de predicción de ingresos, se utilizaron datos individuales. Estos fueron combinados con los datos de hogares a través de la variable identificadora de hogares (id). Posteriormente, los ingresos individuales se sumaron para calcular el ingreso total del hogar y determinar si este supera la línea de pobreza definida en términos monetarios. Modelos como Linear Regression y Elastic Net resultaron adecuados para esta aproximación.

Como resultado de este trabajo, se encontraron diferencias significativas en el desempeño de los modelos al abordar el problema desde ambas perspectivas. Los modelos de clasificación mostraron un rendimiento sólido al identificar y clasificar hogares pobres, mientras que los modelos de predicción de ingresos permitieron un análisis más detallado de los factores que contribuyen a superar la línea de pobreza, pero obtuvieron un desempeño poco satisfactorio a la hora de clasificar dichos hogares. Estos hallazgos resaltan la importancia de seleccionar el enfoque y el modelo adecuados según los objetivos específicos de la política pública y el contexto del análisis. Adicionalmente, estos enfoques pueden ser utilizados de manera conjunta para analizar diferentes perspectivas, y continuar desarrollando búsquedas futuras con la implementación de otros modelos.

En resumen, este artículo presenta un análisis integral de la pobreza desde la perspectiva económica del Banco Mundial, explorando dos enfoques predictivos. Los resultados y hallazgos ofrecen información valiosa tanto para académicos como para formuladores de políticas interesados en diseñar estrategias más efectivas para combatir la pobreza.

2. Datos

El dataset de hogares es adecuado para abordar la tarea predictiva de identificar condiciones de pobreza, ya que contiene variables clave que permiten inferir el nivel socioeconómico y las condiciones de vida de los hogares. La variable objetivo principal es **"Pobre"**, que clasifica a los hogares en condiciones de pobreza. Esta variable es binaria, con un valor de 1 indicando que el hogar está en situación de pobreza, y un valor de 0 que indica que no lo está.

El proceso de construcción de la muestra implicó varias etapas de limpieza y transformación de datos. Se eliminaron columnas irrelevantes y se manejaron valores faltantes de manera responsable. Se descartaron variables como **"Dominio"**, que traía inconvenientes, y se utilizó **"Dpto"** (departamento administrativo) como variable geográfica relevante.

El tratamiento de valores faltantes se realizó en las variables monetarias de hipoteca y alquiler, como **P5100** (cuota mensual de amortización de crédito de vivienda), **P5130** (estimación de arriendo mensual) y **P5140** (valor del arriendo mensual). Estas variables fueron manejadas cuidadosamente. Los valores **98** y **99** indican que el informante no sabe o no quiere informar el monto correspondiente, y se imputaron de acuerdo con la documentación de microdatos del DANE, que establece estos valores como predeterminados para datos faltantes. Se utilizó el método de imputación con la mediana para las columnas **P5130**, **P5100**, y **P5140**.

Por el otro lado, las variables provenientes del dataset de individuos, son de caracter sociodemográficas, como **edad**, **sexo**, y **educación**, por lo que complementan la información del hogar, permitiendo crear un perfil más completo de los hogares en análisis.

En la literatura económica relacionada con la educación y economía laboral, se reconoce que el nivel de ingresos está influido por una variedad de factores individuales y culturales. Las interacciones entre estas variables permiten capturar dinámicas complejas que no se observan directamente al analizar cada factor por separado. Este enfoque es particularmente útil para estudiar desigualdades económicas y comprender cómo diferentes características sociodemográficas afectan los ingresos.

Entre las **interacciones típicas** destacadas en la literatura, se han aplicado las siguientes en este estudio:

1. **Edad y Edad al Cuadrado (P6040 y P6040^2):**
   Esta interacción permite modelar la relación no lineal entre edad e ingresos. La literatura económica sugiere que los ingresos tienden a aumentar con la edad, reflejando mayor experiencia laboral y productividad, pero decrecen a medida que las personas se acercan a la jubilación, asimilando una Curva de Laffer. Incorporar la edad al cuadrado captura esta dinámica de forma precisa.

2. **Género × Nivel Educativo (P6020 × P6210):**
   Diferencias en los ingresos entre hombres y mujeres pueden variar según el nivel educativo alcanzado. Este cruce examina si el retorno a la educación es homogéneo entre géneros o si existen desigualdades significativas. Esto es relevante para identificar brechas salariales asociadas a diferencias educativas y de género.
3. **Zona Urbana/Rural × Nivel Educativo (Clase × P6210):**
   Las diferencias geográficas en el retorno a la educación son un tema recurrente en los estudios de pobreza y desigualdad. Las personas en zonas urbanas generalmente tienen acceso a mejores oportunidades económicas y educativas, lo que puede resultar en mayores ingresos en comparación con zonas rurales. Esta interacción busca capturar estas disparidades espaciales.

Estas interacciones no solo reflejan relaciones bien documentadas en la literatura, sino que también enriquecen el modelo predictive de ingresos.

El análisis descriptivo realizado sobre el dataset de hogares revela información importante sobre la distribución de las variables clave. A continuación se presentan las estadísticas descriptivas para algunas de las variables más relevantes.

Variable	Count	Mean	Std	Min	25%	50%	75%	Max
P5100 (cuota de crédito de vivienda)	5.626	919.943	6.115.976	98	300	500	893875	280.000.000
P5130 (alquiler estimado)	100.507	499.841	4.163.131	98	200	350	500	600.000.000
P5140 (alquiler pagado)	64.453	437.912	1.447.543	20	250	380	500	300.000.000

La variable **Pobre** muestra una distribución desequilibrada, con 20% de hogares que no superan la línea de pobreza. Las variables monetarias (**P5100**, **P5130**, **P5140**) presentan una amplia variabilidad, lo que indica que los hogares tienen diferentes capacidades económicas para afrontar pagos de hipoteca o alquiler.

P5100 (Cuota de crédito de vivienda): La media de la cuota mensual de crédito de vivienda es de 919,943, pero con una desviación estándar considerable (6,115,976), lo que indica una gran variabilidad en los montos. Los valores extremos alcanzan hasta 280,000,000, aunque el valor mínimo es 98, lo que corresponde a los casos en que no se informa o no se sabe el valor de la cuota.

P5130 (Alquiler estimado): El valor promedio del alquiler estimado que los hogares consideran pagar es de 499,841. La desviación estándar es de 4,163,131, lo que indica que los valores de alquiler varían significativamente. El valor mínimo de 98 indica que, en algunos casos, no se pudo obtener una estimación del alquiler.

P5140 (Alquiler pagado): El monto promedio del alquiler mensual pagado es de 437,912, con una desviación estándar de 1,447,543. El valor mínimo de 20 sugiere que algunos hogares informaron un monto muy bajo o que hubo inconsistencias en la declaración del alquiler.

Las variables **P5130**, **P5100**, y **P5140** tenian como valores de default los valores **98** y **99;** los cuales se reemplazaron con valores imputados basados en el análisis de la distribución de las variables.

Además, el análisis de las variables sociodemográficas muestra que, en general, los hogares con menores niveles de educación y mayor proporción de adultos mayores tienden a estar más representados en las clases de pobreza.

Pobre	Hogares No Pobres	Hogares Pobres
Proporción de PET (población en edad de trabajar)	70%	58%
Proporción con educación superior (P6210 > 5)	29%	9%
Proporción con educación hasta secundaria o menor (P6210 ≤ 5)	68%	85%

La población en edad de trabajar (15 a 64 años) constituye un indicador clave para evaluar el potencial productivo de los hogares. En los hogares no pobres, la proporción de PET asciende al **70%**, mientras que en los hogares pobres esta cifra se reduce al **58%**.

Esta diferencia pone de manifiesto que los hogares no pobres cuentan con una mayor proporción de personas en condiciones de participar activamente en el mercado laboral. Por el contrario, en los hogares pobres, la menor proporción de PET puede estar asociada a una mayor carga de dependencia, ya sea de adultos mayores o de niños en edad escolar, lo que limita la capacidad económica de estos hogares. Este resultado destaca la relevancia de políticas que faciliten la incorporación laboral de personas en edad de trabajar, especialmente en contextos de pobreza.

Para obtener un análisis más completo y preciso de los ingresos, es crucial integrar datos a nivel individual con datos de hogares. Este enfoque permite capturar los efectos combinados de las características individuales y las decisiones colectivas del hogar, proporcionando una visión más rica de las dinámicas económicas y sociales. A nivel práctico, esta integración mejora la capacidad predictiva de los modelos, favorece el diseño de políticas públicas más focalizadas y ofrece una comprensión más profunda de las desigualdades económicas.

Se analizaron de forma extensive diversas hipótesis para abordar los valores faltantes en el dataset. Se exploraron varias técnicas de imputación, tales como la imputación por la media, la mediana y métodos más avanzados basados en algoritmos de machine learning, como K-Nearest Neighbors. Además, se evaluó la distribución de los datos para asegurar una imputación adecuada y precisa. Para obtener un detalle completo de este análisis, se puede consultar la correspondiente notebook de Jupyter.

3. Modelos y Resultados

Esta sección describe los modelos implementados para predecir la pobreza utilizando diferentes estrategias y niveles de análisis. Los enfoques adoptados incluyen tanto problemas de clasificación como de predicción de ingresos, siguiendo la metodología del Banco Mundial, que define la pobreza como una función del ingreso según la ecuación:

$$Pobre = I(Inc < Pl)$$

Aquí, $I(\quad)$ es una función indicadora que toma el valor de 1 si el ingreso del hogar está por debajo de un umbral de pobreza específico (PlPlPl). Este marco permite explorar dos aproximaciones principales:

1. **Problema de Clasificación:** Predecir valores binarios de pobreza (0 para no pobres y 1 para pobres).
2. **Problema de Predicción de Ingresos:** Predecir el ingreso del hogar y utilizar la línea de pobreza (PlPlPl) para clasificar posteriormente.

Además, se incluyó un tercer enfoque: combinar los datasets de hogares e individuos para entrenar un nuevo modelo de clasificación, lo que no resultó en una mejora significative.

Los modelos entrenados incluyeron Regresión Logística, Gradient Boosting Classifier, Regresión Lineal y Elastic Net, y nuevamente un XGBoost con datos de nivel de hogar e individuos. Se empleó Grid Search para seleccionar los mejores hiperparámetros de cada modelo posible.

Regresión Lineal (Linear Regression):

El modelo de **Regresión Lineal** se utiliza para predecir un valor continuo, en este caso, el ingreso del hogar. Aunque es un modelo de regresión y no de clasificación, se utilizó para predecir ingresos y luego clasificar los hogares como pobres o no pobres en función de un umbral (línea de pobreza). En este contexto, la regresión lineal no es ideal para la clasificación, pero proporciona una buena base de comparación para otros modelos más complejos. Al igual que la regresión logística, este modelo también fue utilizado como **baseline** en el análisis. Los parámetros en la regresión lineal incluyen los **coeficientes de las variables independientes** y el **intercepto** que determinan la relación entre las variables predictoras y la variable de interés (ingreso). Este modelo no presenta hiperparámetros a optimizar. Este modelo mostró un rendimiento inferior en comparación con los modelos de clasificación.

Elastic Net:

El **Elastic Net** es un modelo de regresión que combina las penalizaciones de **Lasso** (L1) y **Ridge** (L2). El parámetro **alpha** controla la fuerza de la penalización, mientras que el parámetro **l1_ratio** determina el balance entre las penalizaciones L1 y L2. Cuando **l1_ratio = 1**, el modelo se comporta como Lasso, y cuando **l1_ratio = 0**, como Ridge. En este trabajo, los mejores hiperparámetros fueron **alpha = 0.1** y **l1_ratio = 0.8**, lo que permitió una buena regularización, simplificando el modelo mientras mantenía la capacidad de capturar

relaciones complejas entre las variables. A pesar de su flexibilidad, el Elastic Net no logró superar los modelos de clasificación en términos de rendimiento predictivo.

Regresión Logística (Logistic Regression):

La **Regresión Logística** es un modelo de clasificación binaria utilizado para predecir la probabilidad de que una observación pertenezca a una de dos clases, en este caso, "pobre" o "no pobre". Este modelo es particularmente adecuado cuando la variable dependiente es binaria. Los parámetros del modelo incluyen el **coeficiente de regularización** (penalización) que controla la complejidad del modelo para evitar sobreajustes (overfitting). En su versión estándar, no se requiere la optimización de hiperparámetros adicionales, por lo que este modelo sirvió como un **baseline** para la comparación con otros modelos más complejos. Los resultados obtenidos con este modelo fueron utilizados como una referencia inicial para evaluar la mejora con los modelos más avanzados.

Gradient Boosting Classifier:

El **Gradient Boosting Classifier** es un modelo basado en árboles de decisión que utiliza un enfoque de **ensemble learning**, donde varios modelos débiles (árboles de decisión) se entrenan secuencialmente, cada uno corrigiendo los errores del anterior. Los hiperparámetros clave incluyen **n_estimators** (el número de árboles en el modelo), **learning_rate** (la tasa de aprendizaje que controla la contribución de cada árbol a la predicción final) y **max_depth** (la profundidad máxima de cada árbol). Los mejores hiperparámetros encontrados fueron **learning_rate = 0.1**, **max_depth = 5**, y **n_estimators = 200**, lo que ayudó a mejorar la precisión del modelo respecto a los modelos más simples como la regresión logística. Este modelo fue el de mejor desempeño en la tarea de clasificación, demostrando una mejora en la capacidad predictiva en comparación con el baseline. Este modelo fue optimizado usando una búsqueda de cuadrícula para encontrar los mejores hiperparámetros. Este modelo se utilizó para predecir la clase de pobreza (pobre o no pobre), y su rendimiento fue superior a los otros modelos entrenados en este análisis.

3.2 Evaluación de Resultados

Los modelos fueron evaluados en función de su capacidad predictiva, comparando las métricas obtenidas en los conjuntos de entrenamiento y prueba. Al observar los resultados, se encontró que el **Gradient Boosting Classifier** fue el modelo con el mejor desempeño en la tarea de clasificación, mejorando ligeramente de 0.50 a 0.55 en comparación con el modelo de **Regresión Logística**. Esta mejora, aunque modesta, sugiere que el Gradient Boosting tiene un mejor poder predictivo para este tipo de problemas.

En contraste, los modelos de **Elastic Net** y **Regresión Lineal** obtuvieron resultados significativamente más bajos, con un puntaje de 0.33, lo que demuestra su inferioridad frente a los modelos de clasificación. La diferencia de rendimiento entre estos modelos y los de clasificación puede explicarse por su naturaleza de regresión, que no está optimizada para manejar tareas de clasificación binaria, como es el caso de la pobreza (pobre o no pobre).

Aunque los modelos de clasificación superaron a los modelos de regresión en términos de precisión, aún existen áreas de mejora. Específicamente, la eliminación de variables durante el preprocesamiento podría haber afectado negativamente la capacidad predictiva. De igual manera, el **desbalance de clases** en los datos podría haber influido en el rendimiento general de los modelos. En futuros trabajos, se recomienda explorar técnicas de submuestreo más avanzadas para abordar estos desbalances, como **SMOTE (Synthetic Minority Oversampling Technique)** o el uso de **ensembles balanceados**.

Para los modelos de clasificación, se optimizaron los hiperparámetros con el objetivo de maximizar el **F1-score**, una métrica que balancea la precisión (precision) y el recall, lo cual es crucial cuando se enfrentan desbalances en las clases. El **F1-score** se define como la media armónica de la precisión y el recall, lo que lo convierte en una medida adecuada para evaluar el rendimiento de un modelo en escenarios donde las clases son desbalanceadas. La **precisión** (precision) se refiere a la proporción de verdaderos positivos (casos correctamente clasificados como positivos) entre todos los casos que el modelo predijo como positivos, mientras que el **recall** (también conocido como sensibilidad) es la proporción de verdaderos positivos entre todos los casos que realmente son positivos. Dado que la pobreza es un fenómeno relativamente infrecuente en muchas poblaciones, el desbalance entre las clases "pobre" y "no pobre" puede afectar negativamente el rendimiento del modelo si no se optimizan adecuadamente estas métricas. Por ello, el F1-score fue elegido como la métrica principal para la optimización, ya que proporciona un equilibrio entre la capacidad del modelo para identificar correctamente a los hogares pobres (recall) y la precisión de esas predicciones, minimizando el impacto de las predicciones erróneas.

El **XGBoost** implementado como un tercer enfoque tras combinar los datasets de hogares e individuos, no logró mejorar el desempeño en comparación con los modelos entrenados de manera independiente. Este resultado sugiere que la combinación de datos a nivel de hogar e individuo puede introducir ruido y dificultar la convergencia del modelo. Sin embargo, es importante señalar que aún existen áreas de mejora, como la incorporación de variables eliminadas previamente o el ajuste más preciso de los hiperparámetros. Por lo tanto, aunque la combinación de datasets no haya proporcionado una mejora inmediata en el rendimiento, todavía es un área que merece ser explorada y optimizada en futuros enfoques.

Una posible área de mejora en este análisis sería la exploración de **Árboles de Decisión (CART)** y **Random Forest**. Estos algoritmos son conocidos por su capacidad para manejar interacciones no lineales y relaciones complejas entre variables, lo que podría mejorar la predicción de la pobreza, especialmente cuando se trata de datos heterogéneos a nivel de hogar e individuo. Los árboles de decisión son fáciles de interpretar y podrían ser útiles para capturar reglas de decisión claras, mientras que el Random Forest, al ser un modelo de ensamble, podría mejorar la estabilidad y precisión del modelo al combinar múltiples árboles, reduciendo así el riesgo de sobreajuste.

El análisis de **Variable Importance** es crucial en la construcción de modelos predictivos, ya que permite identificar qué variables tienen un mayor impacto en la predicción de la variable

objetivo, en este caso, la pobreza. Este análisis no solo ayuda a mejorar la interpretación y comprensión del modelo, sino que también permite una selección más eficiente de características, optimizando el modelo y reduciendo su complejidad. Al conocer las variables más influyentes, se pueden realizar ajustes en el modelo, como la eliminación de características irrelevantes o la inclusión de variables adicionales que podrían mejorar la precisión. Esta es un área de mejora importante para trabajos futuros y se recomienda realizar un análisis más exhaustivo de la **Variable Importance** al combinar los datasets a nivel de individuo y hogar, lo cual podría revelar nuevas interacciones y relaciones entre las variables que actualmente podrían estar siendo pasadas por alto, mejorando así la capacidad predictiva del modelo.

4. Conclusiones y Recomendaciones

El análisis de pobreza en Colombia a través de modelos predictivos ha revelado diferencias en el desempeño de los enfoques de clasificación y predicción de ingresos. Los modelos de clasificación, como *Logistic Regression* y *Gradient Boosting Classifier*, demostraron un buen desempeño en la identificación de hogares pobres. En contraste, los modelos de predicción de ingresos, como *Linear Regression* y *Elastic Net*, mostraron limitaciones en la clasificación de pobreza, aunque fueron útiles para analizar los factores determinantes del ingreso.

Se recomienda analizar nuevamente el enfoque de trabajar con los dos datasets de hogares e individuos juntos, combinándolos a nivel de individuo y hogar. Esto permitirá realizar un análisis de *Variable Importance*, lo cual podría revelar nuevas interacciones y relaciones entre las variables que actualmente podrían estar siendo pasadas por alto. Además, se sugiere analizar modelos como Árboles de Decisión (CART) y Random Forest, que son útiles para la identificación de patrones complejos en los datos. En cuanto a los desbalances en el dataset, se recomienda explorar técnicas de submuestreo más avanzadas para abordar estos desbalances, mejorando así la precisión y la generalización de los modelos.