Problem Set 3:
Predicting Poverty

Integrantes: Juan Diego Duarte[1]; Erick Julian Villabon[2]

1. Introducción

La problemática de la pobreza en América Latina constituye un desafío persistente, especialmente en el contexto de países en vía de desarrollo, como es el caso de países Latinoamericanos. Según datos recientes de la CEPAL, la región enfrenta una tasa de pobreza del 29%, con un 11.2% de la población viviendo en condiciones de pobreza extrema, proyectándose un aumento a 29.1% y 11.4% para el año 2023. Este fenómeno complejo se atribuye a factores como la desigualdad económica, la limitada accesibilidad a oportunidades educativas y al sistema de salud, así como una oferta laboral insuficiente. Como respuesta a esta problemática el Banco Mundial tiene una iniciativa de abordar la medición de la pobreza a través de una competencia "Pover-T tests: Precting Poverty", la cual tiene el objetivo de construir modelos más eficientes que permitan predecir la pobreza de manera acertada.

Para le caso específico de Colombia, predecir la pobreza adquiere una relevancia debido a la meta del Gobierno en cumplir con los Objetivos de Desarrollo Sostenible, en los cuales se especifica la erradicación de la pobreza extrema. Según cifras del DANE, para el 2022 se observó una disminución de la pobreza monetaria, alcanzando el 36,6% de la población, y la pobreza extrema se situó en un 13,8%, como parte de las estrategias para reducir la pobreza monetaria, se implementaron medidas que aumentaron el número de beneficiarios en programas sociales, sin embargo, las políticas implementadas para este objetivo tienen limitaciones debido a la información incompleta de los hogares en el país. En este contexto, competencias como "Pover-T tests" ofrece oportunidades de explorar y mejorar herramientas que permitan abordar de manera eficiente este tipo de problemáticas al acceso y costo de la información, cabe destacar la importancia de la precisión en la predicción para maximizar la eficacia de las intervenciones y políticas dirigidas a combatir este fenómeno.

De esta forma, los datos derivados de la Misión de Empalme de las Series de Empleo, Pobreza y Desigualdad (MESEP), recopilados por el DANE, se erigen como una herramienta esencial en la evaluación de la pobreza monetaria en Colombia. Esta iniciativa introduce ajustes significativos en la definición de la línea de pobreza y en la construcción del ingreso familiar, permitiendo así la comparación a nivel regional al adoptar métodos de medición en línea con otros países de la región. Por lo que, este estudio utiliza datos provenientes del MESEP a nivel hogar e individual que después de realizar el tratamiento completamos datos a nivel hogar para lograr el objetivo de predecir la pobreza de los hogares. Este estudio se centra en predecir la pobreza de los hogares a través de diferentes metodologías para evaluar cual compone la mejor predicción.

El siguiente trabajo de desarrolla de la siguiente manera: la sección 2 proporciona información sobre la recopilación, tratamiento y descripción de los datos. Posteriormente, la sección 3 presenta

[1] Código: 202011999
[2] Código: 201815677

las especificaciones y los modelos utilizados en los ejercicios de predicción. Por último, la sección 4 concluye.

2. Datos

Los datos provienen de la encuesta "Medición de Pobreza Monetaria y Desigualdad 2018", construida a partir de la GEIH de 2018, fundamental para las estimaciones del Índice de Pobreza Multidimensional (IPM) y la clasificación de hogares en pobres y no pobres según la línea de pobreza colombiana establecida en 2018. Utilizada por el DANE y como insumo clave para la MESEP. Los datos correspondientes parten de las bases de entramiento y de prueba, divididas a nivel hogar y personas, por lo que contamos con cuatro bases las cuales requirieren un emparejamiento a nivel de hogar para obtener dos bases de datos (entrenamiento y prueba) a nivel hogar. Este estudio se centra en la unidad de observación del hogar para analizar características propias de cada hogar.

Debido a la importancia de tener variables que cuenten con la totalidad de los valores y no perder observaciones importantes al momento de realizar predicciones, se realizó una serie de pasos para completar la información de cada hogar e incluir variables que consideramos esenciales para calcular los ingresos de los hogares. Dado que las bases de datos a nivel individual poseen un mayor número de variables y observaciones que los datos a nivel de hogares, se realizó una unión utilizando la llave de hogar como referencia. En otras palabras, en la base de datos de personas se identificaron los miembros de los hogares a través de una variable clave común, facilitando la combinación de información. Este proceso permitió, en primer lugar, obtener los datos faltantes a nivel de hogar y, en segundo lugar, adquirir más variables que potencialmente explican los ingresos del hogar. Esto, a su vez, mejora la precisión en la predicción de ingresos y contribuye a una clasificación más certera.

La muestra cuenta con un total de 231,128 observaciones, de las cuales 66,168 hogares se asignaron a la muestra de prueba. Con el objetivo de no perder observaciones, se llevaron a cabo imputaciones de valores faltantes utilizando datos a nivel de personas. La imputación se basó en el promedio de las personas que integran cada hogar. Para las variables categóricas con datos faltantes, se asignó el valor cero, ya que la comparación posterior tomará el valor de uno si cumple con la característica de la categoría y cero en caso contrario. Dada la considerable cantidad de valores perdidos en la variable de ingreso total, se optó por utilizar la variable ingreso total de la unidad de gasto, la cual representa la suma de los ingresos dentro del hogar y no presenta valores faltantes.

La Tabla 1 muestra la descripción de las variables del estudio, donde se discrimina la muestra de testeo y entrenamiento, y se realiza la separación de pobres y no pobre. En la muestra de entrenamiento se cuenta con 164.960 hogares, mientras que en la muestra de prueba se cuenta con 66.168 hogares. Entre las cuales encontramos características que no tienen mayor diferencia entre los hogares. Podemos encontrar que el promedio de edad en ambas poblaciones se encuentra en promedio en 37 años, las mujeres componen más del 50% de las personas dentro del hogar, y menos del 10% de las personas del hogar son menores de edad. En términos del mercado laborar la experiencia promedio de ambas muestras están alrededor de 60 meses, con respecto a los niveles de educación, podemos observar que tan solo 5% cuenta con educación primaria, el 25% con educación superior y el 11% de la muestra corresponde a estudiantes.

Tabla 1. Descripción de variables

	Muestra de prueba		**Muestra de entrenamiento**	
	Promedio	Mediana	Promedio	Mediana
Edad	37,46	33,50	37,44	33,50
Mujer	0,66	0,50	0,52	0,50
Número de cuartos por persona	0,68	0,66	0,68	0,66
Menores de edad	0,93	1,00	0,91	1,00
Experiencia actual	60,35	24,36	59,41	24,00
Educación				
Primaria	0,06	0,00	0,05	0,00
Secundaria	0,16	0,00	0,16	0,00
Media	0,22	0,00	0,22	0,07
Superior	0,25	0,00	0,26	0,00
Estudiante	0,11	0,00	0,11	0,00
Observaciones	66.168		164.960	

3. Modelos y Resultados

En esta sección, abordaremos el desarrollo de modelos y resultados obtenidos en el ejercicio de predicción de la pobreza de los hogares en Colombia. Dada la naturaleza compleja de los datos y la necesidad de evitar el sobreajuste a la base de entrenamiento, se optó por una estrategia donde en lugar de depender únicamente de una única división de los datos, se divide la base de entrenamiento original en dos conjuntos distintos, el 75% se convierte en una nueva base de entrenamiento, mientras que el 25% restante forma nuestra subbase de prueba. Ambas submuestras se derivan de nuestra base original de entrenamiento.

Esta metodología se adoptó para mitigar el riesgo de sobreajuste, permitiendo una validación más robusta de los modelos. Los modelos mejor ajustados se cargaron en la plataforma Kaggle y se sometieron a pruebas adicionales utilizando la base de prueba original. A continuación, exploraremos los modelos utilizados, las especificaciones adoptadas y, finalmente, los resultados obtenidos en términos de precisión en la predicción de la pobreza de los hogares.

3.1 Modelos

En esta etapa del análisis, nos sumergimos en un paso intermedio crucial para predecir la pobreza: la estimación de los ingresos. Abordar la predicción de los ingresos totales por hogar se presenta como un enfoque estratégico que proporciona información fundamental para realizar estimaciones subsiguientes de la pobreza. En lugar de depender directamente de datos específicos de pobreza, estos modelos se centran en la capacidad predictiva de los ingresos como precursor esencial en la comprensión y abordaje de la problemática de la pobreza.

Para esto, tenemos diferentes modelos y métodos, los cuales deben probar su capacidad predictiva de los ingresos totales por hogar. A continuación, se describirán los modelos que fueron propuestos en esta sección

- Modelo 1:
  Regresión lineal

$$Ingreso\ total = \beta(X) + \varepsilon_i$$

Donde X incluye una serie de variables explicativas como la edad, edad al cuadrado, si es mujer o hombre, si es estudiante o no, los niveles de escolaridad antes mencionados, la experiencia laboral con la que cuenta, cuantas personas por cuarto tiene el hogar, el número de menores de edad que están en el hogar, la ciudad, si la vivienda posee algún tipo de hipoteca, valor del arriendo, el tipo de vivienda, si es propia o en arriendo.

- Modelo 2:
  Lasso es un método de regresión que penaliza la magnitud absoluta de los coeficientes, forzando algunos de ellos a cero. Esto facilita la selección de variables relevantes y contribuye a evitar el sobreajuste.

- Modelo 3:
  Ridge utiliza una penalización de la magnitud cuadrada de los coeficientes, lo que ayuda a mitigar la multicolinealidad. Este modelo ayuda cuando varias variables explicativas están fuertemente correlacionadas.

- Modelo 4:
  Elastic net, este modelo combina la penalización de los dos modelos anteriormente explicados, es particularmente útil cuando hay multicolinealidad en los datos, permitiendo la selección y ponderación automática de variables importantes.

- Modelo 5:
  Boosting trees es un método de aprendizaje automatizado que se basa en la construcción de secuencial de árboles de decisión, donde cada nuevo árbol se enfoca en corregir errores de los modelos anteriores, este método iterativo permite que el modelo se adapte de manera más precisa y eficaz a los datos, mejorando su capacidad de generalización. El ajuste de hiperparámetros específicos del Boosting trees como la tasa de aprendizaje y el número de árboles base, se seleccionaron a medida que iba mejorando la precisión de las predicciones.

- Modelo 6:
  Random forest, en este caso se constituyen múltiples árboles de decisión en paralelo, utilizando subconjuntos aleatorios de los datos de entrenamiento como reemplazo y realizando la selección de características aleatorias en cada nodo de decisión.

Cada uno de estos modelos representa un enfoque que nos ayudará a abordar la predicción de ingresos, y su evaluación detallada proporcionará información valiosa sobre su capacidad predictiva en el contexto de la pobreza en hogares colombianos. En la tabla 2 se podrán observar los resultados obtenidos al aplicar estos modelos.

Tabla 2. Criterios de evaluación de los modelos.

	RMSE	R^2	MAE
Regresión Lineal	0,454	0,097	0,206
Ridge	0,454	0,097	0,206
Lasso	0,447	0,099	0,200
Elastic Net	0,453	0,097	0,206
Boosting	0,372	0,254	0,138

Como se puede observar el modelo con los mejores resultados es el modelo de Boosting ya que incluye el menor valor de MAE, RMSE y el mayor R^2 en comparación con los otros modelos que se utilizaron para este ejercicio de predicción.

3.2 Pobreza

Para llevar a cabo la predicción de si un hogar es pobre o no, nos adentramos en la estimación de modelos de clasificación binaria. En este contexto, la variable "pobre" toma el valor de uno para indicar que el hogar es pobre y cero en caso contrario. La construcción de estos modelos implica la utilización de un vector de predictores X, que comprende variables fundamentadas en la teoría económica y que reflejan la composición del hogar en diversas características.

Este vector de predictores abarca una serie de variables cuidadosamente seleccionadas, las cuales, según la teoría económica, desempeñan un papel crucial en determinar la condición de pobreza de un hogar. Entre estas variables se incluyen factores como la estructura demográfica del hogar, niveles educativos, ingresos, y otras características socioeconómicas que proporcionan información valiosa para la clasificación binaria.

A continuación, nos enfocaremos en la estimación de modelos de clasificación binaria, explorando el rendimiento y la capacidad predictiva de estos modelos en la identificación de hogares pobres. Cada modelo se ajustará a la complejidad de la tarea, utilizando el vector de predictores X para comprender de manera precisa cuales hogares son pobres y no pobres. Para este ejercicio se utilizaron dos enfoques, en el primero se realizó la predicción estimando el ingreso y posteriormente pobreza y en el segundo se estimo directamente la pobreza.

- Modelo 1: Logit

$$\Pr\left(Pobre_i = 1|X\right) = I(Ingreso < Linea\ de\ pobreza)$$

  En este modelo utilizamos la regresión logística para estimar la probabilidad de que un hogar sea clasificado como pobre Pr $(Pobre_i = 1)$ basándose en si su ingreso es inferior a la línea de pobreza. El vector de predictores X abarca una serie de variables explicativas como la edad, edad al cuadrado, si es mujer o hombre, si es estudiante o no, los niveles de escolaridad antes mencionados, la experiencia laboral con la que cuenta, cuantas personas por cuarto tiene el hogar, el número de menores de edad que están en el hogar, la ciudad, si la vivienda posee algún tipo de hipoteca, valor del arriendo, el tipo de vivienda, si es propia o en arriendo.

- Modelo 2: CART
  Este enfoque emplea árboles de decisión para realizar la clasificación binaria. Estos modelos dividen iterativamente los datos en nodos basándose en variables predictoras, generando reglas de decisión jerárquicas para clasificar. En este contexto de la clasificación de pobreza, el árbol divide el conjunto de datos en nodos donde cada uno representa una decisión basada en una característica particular y las hojas del árbol representan las clases finales de si el hogar es pobre o no pobre.

- Modelo 3: Random Fores
  Utiliza múltiples árboles de decisión en paralelo, utilizando subconjuntos aleatorios de datos y selección aleatoria de características en cada nodo. Este enfoque mejora la generalización reduciendo el sobreajuste.

- Modelo 4: Boosting

Es un método de aprendizaje automatizado que se basa en la construcción secuencial de árboles de decisión. Cada nuevo árbol se enfoca en corregir errores de los modelos anteriores, mejorando la capacidad de generalización del modelo. Boosting es efectivo para mejorar el rendimiento y la precisión de la clasificación.

- Modelo 5: Genative model
  Los modelos generativos se centran en la estimación de la distribución de probabilidad de los datos. En este caso de la clasificación de pobreza, este modelo busca entender y modelar la distribución de las variables observadas para predecir la probabilidad de que un hogar sea pobres o no pobres.

- Modelo 6: Neural Networks
  Las redes neuronales se emplean para clasificar hogares como pobres o no pobres, a través del aprendizaje de patrones complejos y no lineales a partir de un conjunto de variables explicativas. Donde este modelo cuenta con capas neuronales de entrada que contiene nodos que representan las variables de entrada; capas ocultas que contienen nodos adicionales que procesan y transforman la información y las capas de salida que en este caso es la probabilidad de que un hogar sea clasificado como pobre o no pobre.

Cada uno de estos modelos representa un enfoque único para abordar la clasificación de pobreza de los hogares colombianos, para decidir sobre estos modelos vamos se evaluarán el rendimiento y su capacidad de predecir la condición de pobreza. La tabla 3 resume el resultado de los modelos utilizados y los mejores resultados en Kaggle.

Tabla 3. Criterios de evaluación de los modelos

	Accuracy	Recall	Kaggle
Logit	0,852	0,455	0,56
CART	0,846	0,432	0,56
LDA	0,839	0,394	0,50
AdaBoost	-	-	0,57
Neural Net	0,893	0.619	0,60

Tras analizar los resultados, se ha determinado que el modelo de redes neuronales destaca como el mejor en términos de la puntuación en Kaggle, obteniendo un valor de 0,60, que al tiempo tiene el mayor valor del accuracy con 0.89. Esto sugiere que, según la calificación de Kaggle utilizada para la competencia, el modelo de redes neuronales basado en un conjunto de variables fundamentadas en teoría económica es la elección preferida para predecir la pobreza en hogares colombianos.

Durante el entrenamiento, la arquitectura del modelo se compone de capas densas con funciones de activación, seguidas de capas de dropout para regularización, lo que contribuye a la prevención del sobreajuste. La capa de salida utiliza la función de activación sigmoid, apropiada para problemas de clasificación binaria, adicionalmente, el modelo se ajustó a los datos de entrenamiento a lo largo de 200 épocas, con un tamaño de lote de 80. Estas especificaciones detalladas delinean la configuración y el proceso de entrenamiento del modelo de red neuronal, proporcionando un marco robusto para la predicción precisa de la pobreza en hogares colombianos.

Es relevante señalar que, si bien el modelo de redes neuronales lidera en Kaggle, la diferencia en la evaluación de modelos no es significativamente grande entre los métodos empleados. Cabe

destacar que la elección del modelo óptimo debe considerar no solo la métrica de evaluación, sino también otros factores como la interpretabilidad del modelo, la complejidad computacional y la capacidad de generalización a nuevos datos.

4. Conclusiones

Este estudio se enfoco en probar diferentes métodos de predicción y clasificación de los hogares en condición de pobreza en Colombia con el objetivo de evaluar cual metodología y método es más eficiente para esta medición.

Tras explorar modelos que van desde técnicas tradicionales como la regresión lineal hasta enfoques más avanzadas como las redes neuronales, se ha identificado el modelo logístico Logit como el más efectivo en términos del puntaje en la competencia em kaggle el cual corresponde a 0.57. Sin embargo, es importante aclarar que la diferencia en la evaluación de los modelos no es significativamente amplia, lo que implica que pueden ser igual de eficientes prediciendo la pobreza y solo cambia en la complejidad de aplicación de los mismos.

GitHub:

https://github.com/Erick-Villabon/Problem_Set_3

Referencias

CEPAL. (2023). Panorama Social de América Latina y el Caribe 2023: la inclusión laboral como eje central para el desarrollo social inclusivo.

DANE. (2023). Publicación de pobreza monetaria extrema y pobreza monetaria, Declaración Comité de Expertos en Pobreza.