Universidad de los Andes
Facultad de Economía
Big Data y Machine Learning para Economia Aplicada
PhD. Ignacio Sarmiento-Barbieri

Merit Tejeda: 202210104
Celin Hernández: 202210067

Repositorio Url: https://github.com/chernan77/Taller_3

Problem Set 3: Predicting Poverty

"Wars of nations are fought to change maps. But wars of poverty are fought to map change" M. Ali

I. Introducción

El objetivo principal de este trabajo es realizar predicciones de pobreza, utilizando como fuente de datos la Gran Encuesta Integrada de Hogares – GEIH de 2018, que incluye un conjunto de indicadores socioeconómicos con cobertura urbana y rural de la República de Colombia. Las variables de la GIEH que ayudan a predecir la pobreza se agruparon en diferentes dimensiones basados en el Enfoque de Pobreza Multidimensional; en este sentido, la pobreza está determinado por un conjunto de características estructurales del hogar.

Algunos trabajos como el de Dabús (2020) abordan la realización de predicciones de pobreza siguiendo técnicas de machine leargning para Argentina, encontrando grandes ventajas para modelos de clasificación, considerando que los hogares se definen como pobres y no pobres (1,0), similar a la manera como se está definiendo la variable pobreza en la GEIH de Colombia. Kambuya (2020) utilizando datos de la encuesta de hogares de Tailandia de 2016 aplica modelos de random forest, regresiones Lasso y regresión stepwise y encuentra un trade-off entre el error de exclusión (hogares pobres clasificados como no pobres) y de inclusión (hogares no pobres clasificados como pobres) según el método que se utilice para seleccionar variables. Asimismo, Kshirsagar et al. (2017) muestran a través de una regresión logística regularizada que se puede mejorar la predicción fuera de la muestra para el caso de Zambia. Estos principios son aplicados a los datos de Colombia y se comparan con otros modelos logit sin regularización.

Este trabajo aborda la realización de predicciones sobre pobreza siguiendo dos enfoques, el primero encaró su estimación a partir de pronósticos del ingreso total de los hogares, niveles a partir de los cuales, según los umbrales establecidos de líneas de pobreza se determina si un hogar es pobre, basados en modelos de regresión; el segundo comprendió modelos de clasificación para pronósticos de la variable pobreza, cabe indicar que los resultados mas precisos se obtuvieron prediciendo el ingreso con modelos de regresión, variable (predicha) que fue utilizada como uno de los regresores principales para predecir la pobreza en los modelos de clasificación, en los cuales se destacó una regresión logística estándar.

Los resultados más precisos estiman un 16.1% de hogares pobres en Colombia; en tanto, los datos oficiales presentados por el DANE indican que, en 2022, en el total nacional la pobreza monetaria fue 36.6% y la pobreza monetaria extrema fue 13.9%, por lo cual los pronósticos realizados muestra cierto desface respecto a los datos observados.

II. Datos

i. Proceso de Adquisición los datos

Los resultados presentados en este trabajo se elaboraron en base a datos de la Gran Encuesta Integrada de Hogares – GEIH de 2018, la cual fue armonizada para su comparabilidad en el tiempo por la misión de Empalme de las Series de Empleo, Pobreza y Desigualdad – MESEP, esta información es fundamental para la Medición de Pobreza Monetaria y Desigualdad por parte del Departamento Administrativo Nacional de Estadística (DANE) de Colombia. Esta información comprende datos socioeconómicos para Trece (13) áreas metropolitanas, con cobertura urbana y rural.

Esta encuesta se puede caracterizar como una base de datos de corte transversal, la cual sigue principios y preceptos metodológicos internacionales para la definición y medición de variable como la pobreza, proporcionando los datos básicos para la identificación del conjunto de hogares (o personas) que no satisfacen un grupo específico de necesidades previamente establecidas (condiciones de la vivienda, educación, composición demográfica del hogar, tenencia de activos, etc.). Asimismo, permite establecer umbrales mínimos, asociados a un nivel de ingreso o gasto, por debajo del cual, se considera que una persona no puede satisfacer sus necesidades básicas y por tanto cae en la escala de pobreza.

Los datos fueron obtenidos de https://www.kaggle.com/competitions/uniandes-bdml-202320-ps3/data e incluye datos a nivel de personas y hogares y están separadas en datos para entrenamiento de los modelos a desarrollar y de prueba para hacer las proyecciones de pobreza fuera de muestra.

ii. Proceso de depuración y limpieza de los datos

La base de datos de entrenamiento contiene 543,109 registros a nivel de personas y 178 variables; estas personas son representadas por las 164,960 observaciones en el segmento de hogares; en cambio, la base de datos de pruebas comprende 219,644 registros de personas y 66,168 de hogares. Una de las características inherentes a estas encuestas de hogares es la presencia de valores atípicos y mucha atrición en los datos originales. Para abordar este problema se realizó el proceso de depuración, limpieza y creación de variables utilizando como pivote los indicadores de educación, edad, ingreso, ocupación, etc. del jefe de hogar, de la siguiente manera:

1) Depuración de variables:

- **Horas trabajadas:** Se cuenta con dos (2) variables claves, las horas trabajadas del jefe de hogar y las horas trabajadas de los ocupados en el hogar; se identificó en ambas variables la presencia de valores extremos; en el caso de las horas trabajadas de ocupados, estas se dividieron entre el número de ocupados para obtener las horas trabajadas en promedio de los miembros del hogar; en este sentido se calculó la media y la varianza para ambas variables. La jornada oficial de trabajo comprende 48 horas semanales; en ese sentido, se calculó un límite máximo y mínimo para el total de horas trabajadas, igual a la jornada oficial $\mp$ 2.5 desviaciones estándar, esta es una regla empírica usada en estadísticas, donde los valores por fuera del intervalo de confianza son imputados, lo cual permite eliminar la presencia de valores atípicos carentes de sentido lógico y económico.
- **Categoría Ocupacional:** Sus valores perdidos fueron imputados considerando la variable posición ocupacional, la cual no presentaba datos vacíos, asumiendo que existe una correspondencia entre ambas variables, un ejemplo, es que si la posición ocupacional decía que el jefe de hogar estaba trabajando y la categoría estaba vacía, se imputaba la etiqueta de Obrero o empleado de empresa particular, similar caso se aplicó al relacionar las demás categorías y posiciones ocupacionales.
- **Experiencia de trabajo en empresa:** Esta variable es relevante para el análisis de ingreso y pobreza; sin embargo, presentaba muchos outilers, los cuales se ajustaron obteniendo el promedio y varianza por departamento, lo cual permitió establecer un límite máximo de meses de trabajo, equivalente a la media + 2.5 desviaciones estándar, este umbral se aplicó a los valores por encima del intervalo;

adicionalmente, si el hogar reportaba cero meses de experiencia o dato vacío, pero su categoría ocupacional decía que estaba empleado, sobre estos valores se imputó el promedio de esta variable.

- **Habitaciones por Hogar:** Se calculó la media, varianza, máximo y mínimo por departamento, los valores por encima del máximo se imputaron con el promedio de esta variable.
- **Ingresos por unidad de gasto:** Si el hogar reportaba ingresos 0.00, pero tenía pagos de arriendo, cotizaba a la seguridad social u otros factores que daba indicios que los ingresos nos podían ser nulos, estos valores fueron imputados con el promedio de esta variable a nivel de departamentos.
- **Pago de Arriendos:** Esta variable incluye el alquiler efectivo (el pagado por el hogar) y el alquiler imputado (el que pagaría el hogar si alquilará), para datos atípicos y vacíos se construyó un coeficiente entre el ingreso y el pago de arriendo, si este coeficiente es mayor a uno (1), el hogar no puede pagar en arriendo más allá de su ingreso o es igual a cero, es decir que no paga ni el alquiler imputado, en ambos casos, se imputa el promedio del coeficiente por departamento, luego este coeficiente se multiplica por el ingreso por hogar para obtener el indicador de pago de arriendo depurado.

2) Construcción de variables:

- **Menores de 18 años:** Considerando que, si en un hogar existen muchas personas dependientes económicamente del jefe o jefes de hogar, esta variable incluye la suma de todas aquellas personas miembros del hogar menores de 18 años.
- **Años de educación promedio hijos:** Dado que se segmentó los hijos por 4 escalas de edades, se calculó el promedio de años de escolaridad para cada hogar.
- **Subsidio:** Se construyó esta variable, indicando si el hogar si recibió algún tipo de subsidio por parte del Gobierno.

Las variables categóricas binomiales y multinomiales se convirtieron a factores, lo cual es una práctica común para tratar variables categóricas.

iii. Descripción de los datos:

Las variables que explican la pobreza se determinan en función de un conjunto de características estructurales del hogar: (1) Demográficas, (2) Mercado Laboral, (3) Educación, (4) Vivienda e (5) Ingresos; en base al enfoque multidimensional de la pobreza, las seleccionadas de las bases de datos se presentan en el siguiente cuadro:

Tabla 1: Variables seleccionadas para pronósticos de Pobreza

Dimensiones	Indicadores	Variables
Demográficos	Sexo jefe del Hogar	X_i^1
	Personas por Hogar	X_i^2
	Edad jefe del Hogar	X_i^3
	Edad promedio hijos	X_i^4
	No. menores de 18 años	X_i^5
	Estrato	X_i^6
Mercado Laboral	Tiempo trabajando en la empresa (Meses)	X_i^7
	Horas trabajadas	X_i^8
	Tiene Seguridad Social	X_i^9
	Ocupados Hogar	X_i^{10}
	Categoría Ocupacional jefe Hogar	X_i^{11}
	Posición Ocupacional jefe Hogar	X_i^{12}

Educación	Educ. Jefe Hogar	X_i^{13}
	Grado Escolar aprobado	X_i^{14}
	Años Educ. promedio hijos	X_i^{15}
Vivienda	Habitaciones por Hogar	X_i^{16}
	Dormitorios	X_i^{17}
	Propiedad Vivienda	X_i^{18}
	Pago Alquiler de Vivienda	X_i^{19}
Ingresos	Ingreso por Hogar	X_i^{20}
	Ingreso percápita Hogar	X_i^{21}
	Subsidios	X_i^{22}

En vista de las características de las variables y su coincidencia entre las bases de entrenamiento y prueba, las variables estrato y subsidios fueron excluidas y para efectos de los modelos de regresión y clasificación se elaboraron tres (3) bases de datos:

- Base de datos 1 que incluye 16 variables y 164960 observaciones.
- Base de datos 2 que incluye 18 variables y 141437 observaciones.
- Base de datos 3 que incluye 20 variables y 81435 observaciones.

Para las regresiones utilizadas para predecir el ingreso se consideraron las siguientes variables:

$$Y_i = f\left(X_i^1, X_i^3, X_i^7, X_i^8, X_i^9, X_i^{11}, X_i^{12}, X_i^{13}\right) \quad (1)$$

iv. Análisis descriptivo de los datos

Tabla 2: Estadísticas Variables Seleccionadas

Variables	N	Mean	St. Dev.	Min	Max
Habitaciones por Hogar	164,960	3	1	1	18
Dormitorios	164,960	2	1	1	15
Personas por Hogar	164,960	3	2	1	28
Pago Arriendo	164,960	465,686.5	532,024.7	66.7	27,993,333.0
Ingreso per cápita Hogar	164,960	874,544.5	1,243,141.0	2,083.3	88,833,333.0
Ingreso total Hogar	164,960	2,116,005.0	2,528,109.0	4,000.0	85,833,333.0
Edad jefe de Hogar	164,960	50	16	11	108
Edad Conyugue	88,310	45	15	16	105
Experiencia Empresas	162,393	109	103	0	582
Horas trabajadas jefe Hogar	117,156	47	14	1	78
Horas trabajadas prop. Ocupados	142,737	45	12	16	80
Horas trabajadas ocupados	142,738	78	44	16	640
Menores de 18 años	164,960	1	1	0	10
Educ. promedio hijos	164,960	3	2	0	9
Edad promedio hijos	164,960	13	14	0	90
Estrato	164,960	3	2	1	6

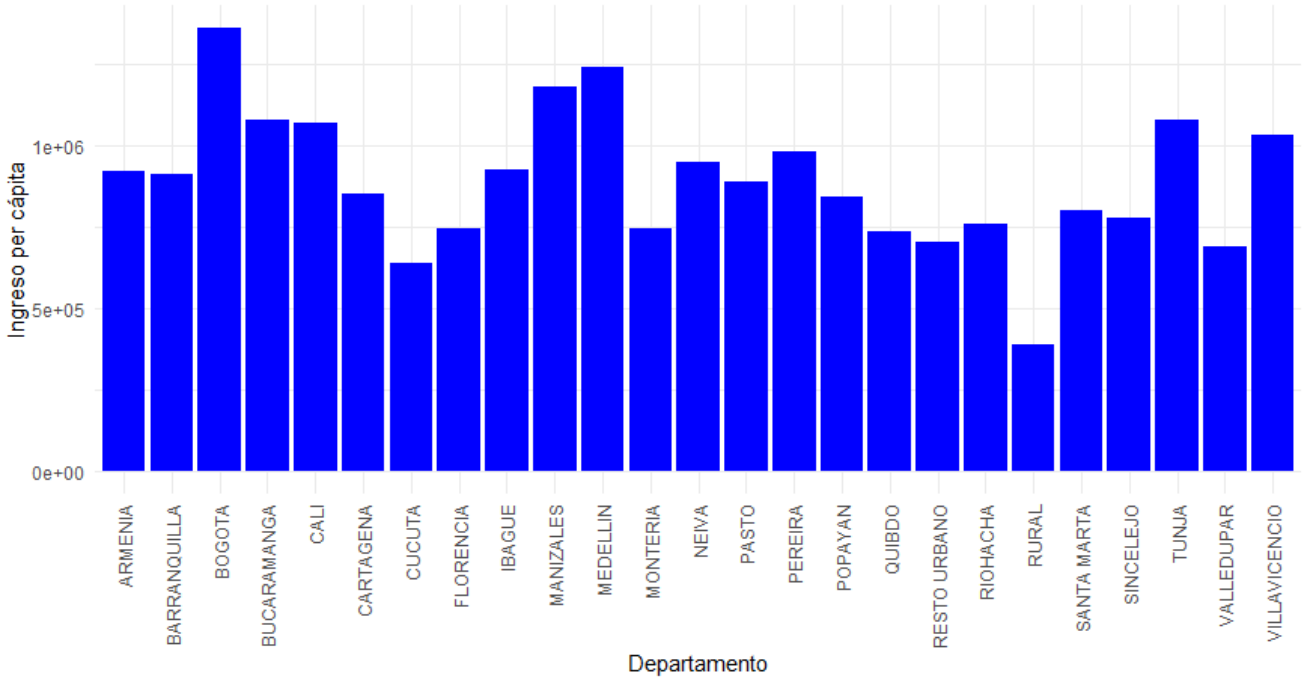
La tabla 2 muestra un conjunto de estadísticas descriptivas de los hogares de Colombia, en promedio los hogares colombianos están constituidos por tres (3) personas; sin embargo, se observa alta variabilidad en esta variable, si estimamos el coeficiente de variabilidad (C.V):

$$C.V = \frac{\text{St. Dev.}}{\text{Mean}} \rightarrow C.V = \frac{2}{3} = 0.667 \quad (2)$$

El C.V de 0.667 indica fuerte heterogeneidad en el tamaño de las familias en Colombia; denotando una fuerte dispersión en esta variable y, por lo tanto, el promedio no es representativo del conjunto de datos. Por otra parte, los hogares en Colombia destinan en promedio un 22% de su ingreso para el pago de arriendo; asimismo, se observa una desviación estándar, incluso mayor al promedio de los datos, explicado por alta disparidad regional en los costos de arriendo, el municipio de Quibdo es que presenta la relación arriendo/ingreso más alto; en cambio, las zonas rurales presentan la relación más baja.

Los hogares de Colombia tienen en su mayoría jefes de hogares masculinos, un 58.2% son hombre y restante 41.8% mujeres. Este jefe de hogar tiene una edad mayor a la de su conyugue, lo cual indica una estructura familiar donde el hombre jefe de hogar es en promedio 5 años mayor que su esposa. Este núcleo familiar tiene en promedio un (1) hijo, cuya edad oscila en torno a los 13 años y ha completado en promedio 3 años de educación. Las familias mas grandes se ubican en los departamentos de Barranquilla y Cartagena. En el mercado laboral, el promedio de horas trabajadas por el jefe de hogar esta de acorde con la jornada oficial de trabajo, los departamentos de Bucaramanga y Cúcuta son los que reportan las jornadas de trabajo más extensas.

Grafica 1: Colombia: Ingreso per cápita promedio por Departamento



A nivel de ingresos, los departamentos de Bogotá, Medellín y Manizales son los que presentan el ingreso per cápita más alto de Colombia; en cambio, como era de esperar, en las zonas rurales se ubican los segmentos poblaciones con ingresos más bajos; sin embargo, cuando Bogotá es de la zonas de más altos ingresos, concentra uno de los mayores porcentajes de pobreza a nivel nacional, lo cual se refleja una alta dispersión del ingreso per cápita a nivel departamental y entre zonas o localidades; es decir, que en el caso de Bogotá, existe una amplia desigualdad en la distribución del ingreso.

III. Modelos y resultados

Con el fin de enfrentar la complejidad de la pobreza a nivel nacional en Colombia, se han desarrollado modelos tanto de clasificación como de regresión. Estos modelos desglosan las predicciones en dos vertientes: por un lado, se emplearon modelos de regresión utilizando dos bases de datos (data1i, data2i). La diferencia entre estos modelos se fundamenta en el número de variables y observaciones, como se describió previamente; por ejemplo, la base data2i incluye variables específicas como la experiencia del jefe de hogar y las horas trabajadas por los ocupados del hogar. Por otro lado, para predecir la incidencia de la pobreza, se utilizaron modelos de clasificación (data1p, data2p, data3p) que incorporan un conjunto más amplio de variables, abordando no solo los ingresos, sino también factores demográficos, y condiciones de vivienda.

MODELOS DE CLASIFICACIÓN

Se desarrollaron modelos de clasificación utilizando diversas técnicas para cada base de datos. Específicamente, se implementaron modelos de regresión logística con regularización (Lasso, Ridge y Elastic Net), árboles de clasificación (CART) y modelos Gradient Boosting. Adicionalmente, se realizaron modelos de clasificación con redes neuronales, los cuales se observan en apéndices. Este proceso metodológico involucró la partición de los datos en conjuntos de entrenamiento y prueba, siendo el 70% destinado al entrenamiento y el 30% a la prueba. La métrica de precisión (Accuracy) se empleó como indicador principal para evaluar el rendimiento de cada modelo.

Resultados Modelos de Regresión Logística, Modelos CART y Gradient Boosting

En este análisis, se estimaron modelos de regresión logística con las bases de datos data1p, data2p y data3p, considerando los enfoques de regularización (Lasso, Ridge, Elastic Net). Se empleó validación cruzada para seleccionar los valores óptimos de lambda en cada método de regularización. La conversión de variables categóricas a dummy fue esencial para la aplicación eficaz de la regularización. Se realizaron pruebas con ingreso total e ingreso per cápita, siendo la variable "Ingreso_Perc_Hogar" más efectiva para predecir la pobreza según la métrica de precisión.

Los resultados, presentados en la Tabla 3, indican que el modelo logístico sin penalización, utilizando data2p, muestra la mayor precisión dentro de muestra (Accuracy: 0.964). Los modelos de regularización se desarrollaron exclusivamente con la base 2 debido a su rendimiento superior. La elección de esta base sobre la base 1 se justifica por la inclusión de variables cruciales para predecir la pobreza: "Exp_Empresa" y "Hrs_Ocupados".

Tabla 3 Modelos de Regresión Logística

Modelos	Metrica	Valor
Logic1	Accuracy	0.955
Logic2	**Accuracy**	**0.964**
Logic3	Accuracy	0.956
Elastic Net	Accuracy	0.962
Ridge	Accuracy	0.875
Lasso	Accuracy	0.962

La metodología CART se fundamentó en la construcción de árboles de decisión. La validación cruzada de esta metodología consistió en dividir repetidamente el conjunto de datos en subconjuntos de entrenamiento y prueba, evaluando el rendimiento en cada iteración y promediando los resultados. Se hizo la búsqueda de hiperparámetros, ajustando factores clave, principalmente el parámetro de complejidad (cp). Se observó que el modelo con la base de datos data2p mostró la máxima precisión, alcanzando un 97.92% en el conjunto de prueba.

Tabla 4 Modelos CART y Gradient Boosting

CART

Modelos	Metrica	Valor
CART1	Accuracy	0.9732663
CART2	**Accuracy**	**0.9791922**
CART3	Accuracy	0.9736353

Gradient Boosting

Modelos	Metrica	Valor
xgboost1	Accuracy	0.796
xgboost2	**Accuracy**	**0.818**
xgboost3	Accuracy	0.798

En el proceso de estimación de la pobreza mediante la metodología de Gradient Boosting con XGBoost, se determinó que el modelo estimado con la base de datos data2p es el que cuenta con la precisión más alta 81.8%.

MODELOS DE REGRESIÓN

Se llevó a cabo un análisis minucioso con el propósito de identificar la variable más precisa y relevante para la predicción de la pobreza[1]. Este estudio concluyó que el ingreso per cápita, derivado del ingreso total y el tamaño del hogar, resultó ser la variable más precisa al emplearse en modelos de clasificación para estimar la pobreza. Este hallazgo fundamentó nuestra decisión de elegir el ingreso per cápita como variable predictora clave en los modelos de regresión. En el proceso, al enfrentarnos a la necesidad de utilizar modelos de regresión para estimar el ingreso y, por ende, posibilitar la estimación de la pobreza, convertimos el ingreso per cápita en ingreso total mediante el tamaño del hogar. Posteriormente, utilizamos la línea de pobreza para determinar la clasificación de pobreza del hogar. Sin embargo, durante la evaluación de los modelos, se observó que este enfoque no era óptimo para la estimación precisa de la pobreza. En contraste, se descubrió que, al combinar modelos de regresión con modelos de clasificación, se lograba una mejora significativa en la precisión de las estimaciones.

Los modelos de regresión desarrollados abarcan desde la Regresión Lineal hasta técnicas más avanzadas como modelos de Regularización (Ridge y Lasso), Arboles y Gradient Boosting con XGBoost. Cada uno de estos modelos fue entrenado y evaluado utilizando métricas como el error absoluto medio (MAE), proporcionando así una evaluación integral y detallada de su rendimiento. La validación de las predicciones se llevó a cabo en dos fases distintas. En la primera, se evaluaron los ejercicios utilizando siempre las dos bases de datos: data1i y data2i. Estas se dividieron en un 70% para entrenamiento y un 30% para pruebas, buscando encontrar el modelo con la mejor predicción del ingreso per cápita dentro de muestra. En la segunda fase, se realizó otro ejercicio que empleó las bases de datos totales (data1i, data2i) como conjunto de entrenamiento, mientras que las bases data1p, data2p y data3p[2] se utilizaron como conjunto de prueba para estimar la precisión de la pobreza.

Resultados de los Modelos Regresión Lineal, Modelos de Regularización, Arboles y Gradient Boosting
Se aplicaron modelos de regresión lineal, regresión Ridge y regresión Lasso con el objetivo de estimar el ingreso per cápita. Los resultados de estos modelos se detallan en la Tabla 5 y Tabla 6, evidenciando que, en todas estas metodologías, aquellos modelos que emplean la base de datos datai2 en la primera fase y, en la segunda fase, utilizan tanto la datai2 como conjunto de entrenamiento y la datap3 como conjunto de prueba, obtienen los mejores resultados para la predicción del ingreso per cápita.

[1] Se ejecutó el análisis inicial utilizando el ingreso total para estimar la pobreza mediante la línea de pobreza; sin embargo, los resultados obtenidos no fueron óptimos. No obstante, al emplear modelos de clasificación, se observó una mejora sustancial en la precisión al utilizar el ingreso per cápita en comparación con el uso del ingreso total.
2 Estas bases se utilizaron porque incluían la base de pobreza.

Tabla 5: Modelos Regresión Lineal

Primera Fase		Segunda Fase	
Modelos	MAE	Modelos	MAE
Modelo1	528573	Modelo1	528047
Modelo2	**506633**	Modelo2	506633
		Modelo3	**427540**

Tabla 6: Modelos de Regularización

RIDGE

Primera Fase		Segunda Fase	
Modelos	MAE	Modelos	MAE
Modelo1	523251	Modelo1	522577
Modelo2	**499480**	Modelo2	499480
		Modelo3	**421507**

LASSO

Primera Fase		Segunda Fase	
Modelos	MAE	Modelos	MAE
Modelo1	528238	Modelo1	527692
Modelo2	**506240**	Modelo2	506240
		Modelo3	**426787**

Tabla 7: Modelos Gradient Boosting y Arboles

Gradient Boosting

Primera Fase		Segunda Fase	
Modelos	MAE	Modelos	MAE
Modelo1	462990	Modelo1	462990
Modelo2	**420050**	Modelo2	420050
		Modelo3	**339397**

Arboles

Primera Fase		Segunda Fase	
Modelos	MAE	Modelos	MAE
Modelo1	462990	Modelo1	462990
Modelo2	**420050**	Modelo2	420050
		Modelo3	**339397**

Por último, se implementaron modelos de Arboles usando el método method = "rpart" y, Gradient Boosting usando la documentación XGBoost. Se encuentra que ambos casos los modelos que utilizan la base de datos datai2 son los que presentan las métricas de error medio absoluto (MAE) más bajos.

MODELOS FINALES

Para abordar la complejidad de predecir la pobreza, exploramos diversas combinaciones de modelos de clasificación y regresión. A pesar de la implementación de modelos más complejos y un extenso proceso de validación cruzada, la combinación de regresión lineal y regresión logística destacó como la más precisa en la competición de Kaggle. Esta combinación, aplicada a la base de datos 2, demostró ser altamente efectiva en la predicción de la pobreza. Las variables clave consideradas en este modelo incluyen Sexo_JHogar, Edad_JHogar, Edad_JHogar2, Pers_por_Hogar, Exp_Empresa, Cat_Ocup_JHogar, Posc_Ocup_JHogar, Educ_JHogar, SS_Jefe, Ingreso_Perc_Hogar. Además, para la predicción de la pobreza, se incorporan Pers_por_Hogar, Menores_18Años, Hrs_Ocupados, Total_Ocup, Educ_prom_Hijos, Hab_por_Hogar, Dormit_Hogar y Pago_Arriendo, todas las cuales fueron detalladas en las secciones previas de este documento.

Entrenamiento de los Modelos y Selección de Hiperparámetros

a) Modelo de regresión lineal

El procedimiento se inició mediante la división de los datos en dos conjuntos, asignando el 70% para entrenamiento y el 30% restante para pruebas. Este paso garantizó conjuntos independientes para ajustar y evaluar los modelos. Durante la selección del mejor modelo de regresión lineal, se incorporó la validación cruzada utilizando el método "k-fold cross-validation". Concretamente, se empleó la función train de la biblioteca caret en R, configurando el parámetro method como "cv" (validación cruzada). Cabe mencionar, que este modelo reflejó un MAE de 506633, pero otros modelos tuvieron mejores predicciones.

b) Modelo de regresión logística

En este caso se utilizó el modelo de regresión logística con los hiperparámetros predeterminados. Se dividió el conjunto de datos en datos de entrenamiento (70%) y datos de prueba (30%) utilizando la función initial_split de rsample. El modelo se entrenó exclusivamente con los datos de entrenamiento y luego se evaluó su rendimiento en los datos de prueba, dando una acuriosidad de 0.964.
El resultado de estos dos modelos (regresión lineal y regresión logística) dio como resultado una precisión de 0.54 siendo la mas alta de las precisiones.

c) Comparación con otros Enfoques

Los pronósticos adicionales estimados dentro de muestra, en términos generales para el concurso de Kaggle, exhibieron una variabilidad en la precisión, situándose mayormente en un rango entre 0.53 y 0.49. Un resultado destacado fue la combinación de modelos que consistió en el siempre eficaz modelo de regresión lineal con la base data2i, junto con el modelo de regresión logística con penalización tipo Ridge, generando una precisión de 0.53 y posicionándose como el segundo modelo más preciso. En tercer lugar, se encontró otro modelo competitivo que empleó regresión lineal y regresión logística utilizando las bases de datos data1i y data1p, alcanzando una precisión de 0.51.
Cabe resaltar que varias de las estimaciones posteriores mostraron una precisión en torno al 0.5, siendo digna de mención la combinación del modelo de regresión árbol con el modelo de regresión logística, ambos utilizando la base de datos 2.

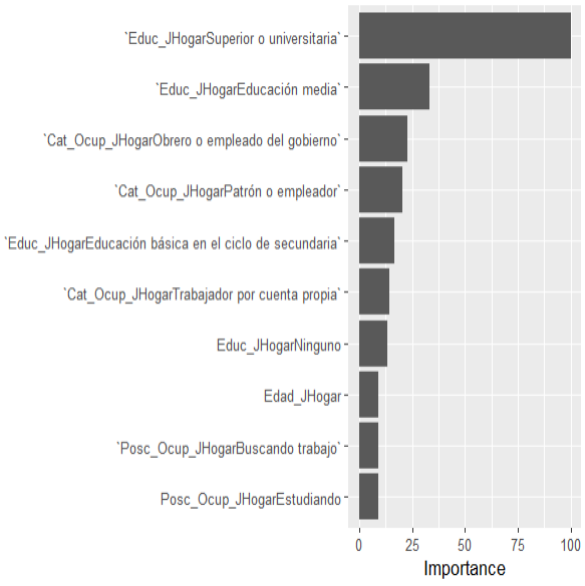
Tabla 8: Resultados Competición en Kaggle Modelos de Regresión y Clasificación

Modelos	Accuracy
Modelo de regresión lineal y logistica base 2	0.54
Modelo de regresión lineal y logistica con penalización	0.53
Modelo de regresión lineal y logistica base1	0.51
Modelo de regresión Arbol y logistica base 2	0.50

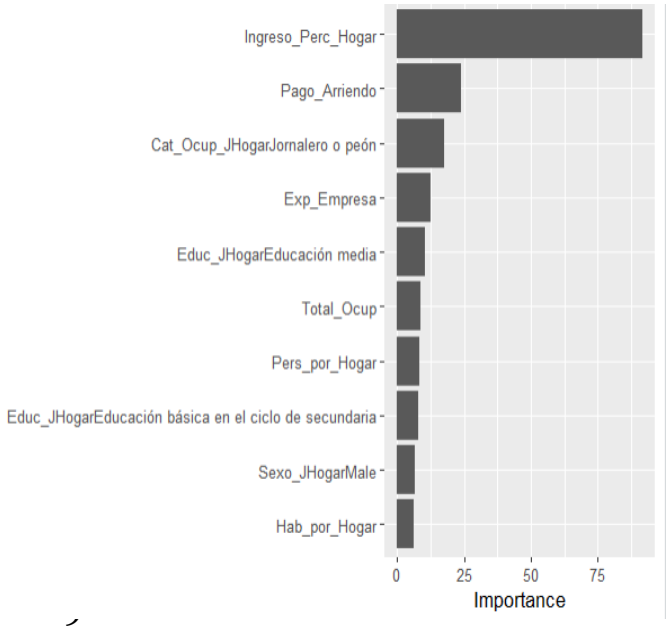
d) Importancia de las Variables en los dos Modelos

Tabla 9: Importancia de las Variables en los Modelos Finales

Modelo de Regresión



Regresión Logística



Como se observan en los resultados de estas tablas se encuentra que las variables que contribuyen más a la predicción del ingreso per cápita es la educación del jefe de hogar. En el caso del modelo clasificación se observa que la variable que tiene la mayor contribución es el ingreso per cápita y el pago de arriendo de la casa.

CONCLUSIONES Y RECOMENDACIONES

- La complejidad de la pobreza en Colombia requiere un enfoque integral que aborde tanto factores económicos como demográficos y de calidad de vida. Se ha destacado la importancia del ingreso per cápita como variable predictora clave para la estimación de la pobreza.

- La integración de modelos de regresión y clasificación se reveló como una estrategia eficaz para mejorar la precisión en la estimación de la pobreza. Al combinar la capacidad de predecir ingresos per cápita con la clasificación de la pobreza, se logró una mejora sustancial en la capacidad general de los modelos para abordar la complejidad de la pobreza a nivel nacional en Colombia.

- A pesar de la diversidad de enfoques utilizados, el modelo que demostró la mejor predicción en el concurso de Kaggle fue sorprendentemente un modelo de regresión lineal con una clasificación logística. Este resultado sugiere que, en ocasiones, la simplicidad y parsimonia de un modelo pueden traducirse en un rendimiento superior. La eficacia de este enfoque resalta la importancia de la adecuada selección y ajuste de modelos, subrayando que la complejidad no siempre garantiza mejores resultados.

- Para mejorar el análisis, se recomienda la inclusión de datos geoespaciales para capturar variaciones regionales en la pobreza. La geolocalización puede ser un factor significativo y agregar un nivel adicional de detalle a la predicción de la pobreza.

- Se sugiere que los esfuerzos futuros se centren en mejorar la precisión del Ingreso Per Cápita en la estimación de esta variable, ya que demostró ser fundamental para predecir la condición de pobreza

IV. Bibliografía

1. Dabús, A. (2020). Pobreza en Argentina: un análisis predictivo utilizando herramientas de Machine Learning. Universidad de San Andrés, Departamento de Economía. DNI: 38.919.616. https://repositorio.udesa.edu.ar/jspui/bitstream/10908/18489/1/%5BP%5D%5BW%5D%20T.M.%20Eco.%20Dabús%2C%20Andrés.pdf

2. Kambuya, P. (2020). Better Model Selection for Poverty Targeting through Machine Learning: A Case Study in Thailand. https://so05.tci-thaijo.org/index.php/TER/article/view/183260/163841

3. Kshirsagar, V., Wieczorek, J., Ramanathan, S., & Wells, R. (2005). Household poverty classification in data-scarce environments: A machine learning approach. *Innovations for Poverty Action*. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://arxiv.org/pdf/1711.06813.pdf

Apéndice

Modelos de Clasificación con Redes Neuronales para la Predicción de Pobreza

En esta sección, se proporciona información detallada sobre la implementación de modelos de clasificación con redes neuronales utilizando TensorFlow y Keras. Estos modelos están diseñados para predecir la variable objetivo "Pobreza". Se sigue un enfoque paso a paso para cada modelo, desde la preparación de datos hasta la evaluación del rendimiento. A continuación, se presenta un resumen por modelo:

Modelo 1: Regresión Lineal Simple

Este modelo utiliza una capa de salida con activación lineal para abordar el problema de clasificación binaria. La función de pérdida es el Error Cuadrático Medio (MSE), y se emplea el optimizador RMSprop para ajustar los pesos del modelo. La elección de la función de activación lineal permite que el modelo realice predicciones en un rango continuo, que luego se pueden ajustar para clasificar entre las clases de pobreza. Capas y Activación: Una capa de salida con activación tanh.

Modelo 2: Red Neuronal con Activación ReLU

En este enfoque, se implementa una capa de salida con la función de activación ReLU, que introduce no linealidades en el modelo. La función de pérdida sigue siendo el MSE, y se utiliza el optimizador RMSprop. La elección de ReLU como función de activación busca capturar relaciones no lineales entre las variables predictoras y la variable objetivo.

Modelo 3: Red Neuronal con Activación Tangente Hiperbólica (tanh)

En este modelo, la capa de salida utiliza la activación tanh, que produce salidas en el rango [-1, 1]. Esta función de activación es útil para problemas de clasificación binaria. Al igual que en los modelos anteriores, la función de pérdida es el MSE, y el optimizador RMSprop ajusta los parámetros de la red.

Modelo 4: Red Neuronal Profunda

Se introduce un modelo más complejo con dos capas ocultas, ambas con activación ReLU, y una capa de salida. La función de pérdida y el optimizador RMSprop se mantienen. Este modelo busca aprender representaciones más complejas de los datos, capturando relaciones no lineales y patrones más sofisticados en las variables predictoras.

Modelo 5: Red Neuronal Más Compleja

Este modelo amplía el Modelo 4 mediante la adición de otra capa oculta. La complejidad adicional tiene como objetivo mejorar la capacidad del modelo para aprender patrones más intrincados. La función de pérdida y el optimizador RMSprop permanecen consistentes con los modelos anteriores.