Taller 2 - Predicción de Pobreza en Colombia

Harold Stiven Acuña

José David Cuervo
José David Dávila
César Augusto Alfaro

31 de marzo de 2025

Resumen

Este documento presenta el análisis de datos y la implementación de modelos de clasificación para la predicción de la pobreza en Colombia.

Palabras clave: pobreza, clasificación, aprendizaje automático
Clasificación JEL: J31, C53, J16

Repositorio GitHub:
https://github.com/alfarocesar/BDML_Predicting_Poverty_Equipo8

1. Introducción

Breve descripción del problema, antecedentes y motivación del estudio.

2. Datos

2.1. Descripción de la base de datos

Explicación de la estructura de los datos y variables clave.

2.2. Procesamiento y limpieza de datos

Metodologías aplicadas para la adecuación de los datos.

2.3. Análisis descriptivo

Exploración de la variabilidad de los datos con tablas y gráficos.

3. Modelos y Resultados

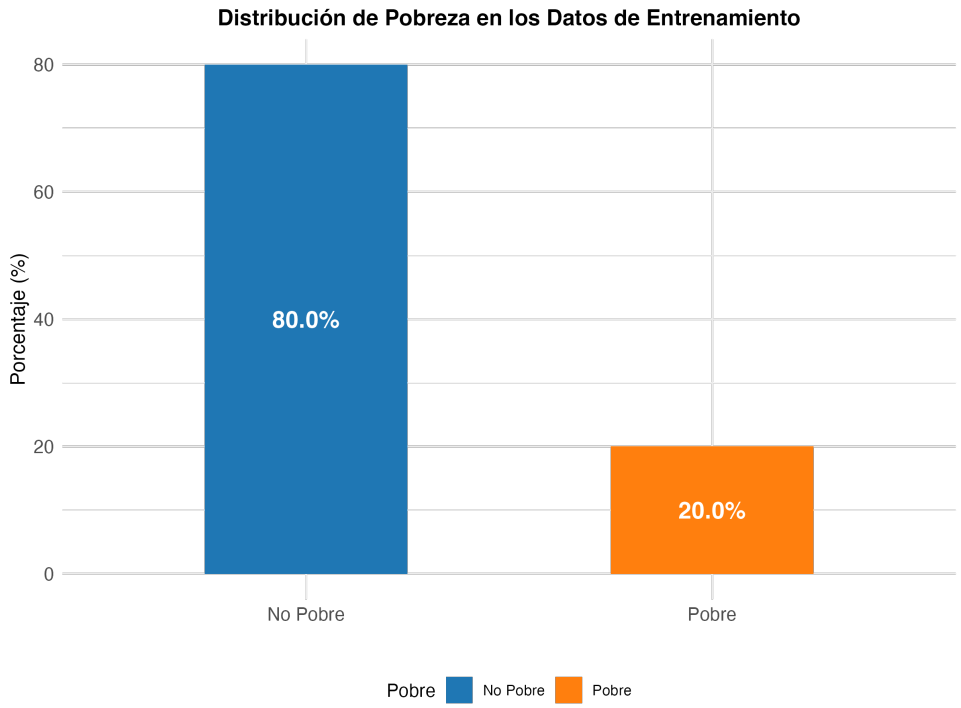
3.1. Metodología

Descripción de los algoritmos utilizados: Regresión Logística, Random Forest, etc.

3.2. Entrenamiento y validación de modelos

Explicación del proceso de entrenamiento y evaluación.

Figura 1. Distribución de la pobreza



Cuadro 1. Estadísticas Descriptivas de Variables Numéricas

Statistic	N	Mean	St. Dev.	Min	Max
P5000	164,960	3.39	1.24	1	98
P5010	164,960	1.99	0.90	1	15
P5090	164,960	2.46	1.26	1	6
P5130	100,507	499,840.80	4,163,131.00	98	600,000,000
P5140	64,453	437,911.80	1,447,543.00	20	300,000,000
Nper	164,960	3.29	1.77	1	28
Npersug	164,960	3.28	1.77	1	28
Ingtotugarr	164,960	2,307,865.00	2,628,933.00	0.00	88,833,333.00
Ingpcug	164,960	870,639.30	1,244,350.00	0.00	88,833,333.00
Lp	164,960	271,522.30	33,656.89	167,222.50	303,816.70
Pobre	164,960	0.20	0.40	0	1
n_miembros	164,960	3.29	1.77	1	28
n_mujeres	164,960	1.74	1.18	0	14
n_menores	164,960	0.98	1.16	0	15
n_ocupados	164,960	1.50	1.03	0	14
edad_promedio	164,960	37.44	16.88	5.67	102.00
jefe_mujer	164,960	0.42	0.49	0	1
jefe_edad	164,960	49.61	16.39	11	108

Cuadro 2. Relación entre Variables Categóricas y Tasa de Pobreza

variable	categorias_analizadas	variabilidad_tasas	min_tasa	max_tasa	rango_tasas
Dominio	25	59.01	8.30	32.17	23.87

Cuadro 3. Resumen de Variables Categóricas

variable	categorias_unicas	categoria_mas_frecuente	frecuencia_max_cat	porcentaje_max_c
Dominio	25	RESTO URBANO	17049	10.

3.3. Comparación de modelos

Tabla comparativa con métricas de desempeño.

3.4. Importancia de variables

Análisis de las características más relevantes en la predicción.

4. Conclusión

Resumen de los hallazgos principales y posibles mejoras futuras.

Cuadro 4. Correlación de Variables Numéricas con Pobreza

variable	correlacion_con_pobreza
n_menores	0.36
Ingtotugarr	-0.30
Ingpcug	-0.28
Npersug	0.24
Nper	0.24
n_miembros	0.24
n_mujeres	0.21
edad_promedio	-0.20
P5000	-0.14
P5090	0.14
n_ocupados	-0.12
Lp	-0.09
jefe_edad	-0.09
jefe_mujer	0.05
P5140	-0.04

Cuadro 5. Variables con Valores Faltantes

variable	n_missing	pct_missing
P5140	100507.00	60.93
P5130	64453.00	39.07

Cuadro 6. Resumen de Variables Numéricas

variable	media	sd	cv_pct	min	q1	mediana	q
P5000	3.39	1.24	36.56	1.00	3.00	3.00	4.0
P5010	1.99	0.90	45.15	1.00	1.00	2.00	3.0
P5090	2.46	1.26	51.37	1.00	1.00	3.00	3.0
P5130	499840.83	4163131.05	832.89	98.00	200000.00	350000.00	500000.0
P5140	437911.80	1447543.24	330.56	20.00	250000.00	380000.00	500000.0
Nper	3.29	1.77	53.91	1.00	2.00	3.00	4.0
Npersug	3.28	1.77	54.05	1.00	2.00	3.00	4.0
Ingtotugarr	2307864.63	2628933.20	113.91	0.00	900000.00	1581242.00	2785322.0
Ingpcug	870639.26	1244349.74	142.92	0.00	300000.00	543568.48	987367.4
Lp	271522.31	33656.89	12.40	167222.48	275594.03	279944.53	285649.5
Pobre	0.20	0.40	199.88	0.00	0.00	0.00	0.0
n_miembros	3.29	1.77	53.91	1.00	2.00	3.00	4.0
n_mujeres	1.74	1.18	67.78	0.00	1.00	2.00	2.0
n_menores	0.98	1.16	118.28	0.00	0.00	1.00	2.0
n_ocupados	1.50	1.03	68.27	0.00	1.00	1.00	2.0
edad_promedio	37.44	16.88	45.08	5.67	24.00	33.50	48.2
jefe_mujer	0.42	0.49	117.93	0.00	0.00	0.00	1.0
jefe_edad	49.61	16.39	33.04	11.00	37.00	49.00	61.0

Cuadro 7. Análisis de Outliers (Método IQR)

variable	n_total	n_outliers	pct_outliers	lower_bound	upper_bound	min_value	
P5000	164960	17716	10.74	1.50	5.50	1.00	
P5010	164960	70	0.04	-2.00	6.00	1.00	
P5090	164960	0	0.00	-2.00	6.00	1.00	
P5130	100507	7507	7.47	-250000.00	950000.00	98.00	6
P5140	64453	3216	4.99	-125000.00	875000.00	20.00	3
Nper	164960	3929	2.38	-1.00	7.00	1.00	
Npersug	164960	3901	2.36	-1.00	7.00	1.00	
Ingtotugarr	164960	11692	7.09	-1927983.00	5613305.00	0.00	
Ingpcug	164960	13199	8.00	-731051.13	2018418.54	0.00	
Lp	164960	19787	12.00	260510.81	300732.73	167222.48	
n_miembros	164960	3929	2.38	-1.00	7.00	1.00	
n_mujeres	164960	12168	7.38	-0.50	3.50	0.00	
n_menores	164960	687	0.42	-3.00	5.00	0.00	
n_ocupados	164960	6579	3.99	-0.50	3.50	0.00	
edad_promedio	164960	761	0.46	-12.30	84.50	5.67	
jefe_mujer	164960	0	0.00	-1.50	2.50	0.00	
jefe_edad	164960	52	0.03	1.00	97.00	11.00	

Cuadro 8. Distribución de la Variable Objetivo (Pobreza)

Pobre	n	proportion
0	131936	79.98
1	33024	20.02

Cuadro 9. Resumen Estadístico del Conjunto de Entrenamiento

skim_type	skim_variable	n_missing	complete_rate	character.min	character.max	charact
character	id	0	1.00	24	24	
character	Dominio	0	1.00	4	13	
numeric	P5000	0	1.00			
numeric	P5010	0	1.00			
numeric	P5090	0	1.00			
numeric	P5130	64453	0.61			
numeric	P5140	100507	0.39			
numeric	Nper	0	1.00			
numeric	Npersug	0	1.00			
numeric	Ingtotugarr	0	1.00			
numeric	Ingpcug	0	1.00			
numeric	Lp	0	1.00			
numeric	Pobre	0	1.00			
numeric	n_miembros	0	1.00			
numeric	n_mujeres	0	1.00			
numeric	n_menores	0	1.00			
numeric	n_ocupados	0	1.00			
numeric	edad_promedio	0	1.00			
numeric	jefe_mujer	0	1.00			
numeric	jefe_edad	0	1.00			

Cuadro 10. Variables Candidatas a Eliminar

variable	motivo
id	Variable de identificación
P5140	Más del 30 % de valores faltantes
P5130	Más del 30 % de valores faltantes

Cuadro 11. Variables Categóricas Candidatas para One-Hot Encoding

variable	categorias_unicas

Cuadro 12. Clasificación de Variables

variable	tipo	es_id
id	character	TRUE
Dominio	character	FALSE
P5000	integer	FALSE
P5010	integer	FALSE
P5090	integer	FALSE
P5130	numeric	FALSE
P5140	numeric	FALSE
Nper	integer	FALSE
Npersug	integer	FALSE
Ingtotugarr	numeric	FALSE
Ingpcug	numeric	FALSE
Lp	numeric	FALSE
Pobre	integer	FALSE
n_miembros	integer	FALSE
n_mujeres	integer	FALSE
n_menores	integer	FALSE
n_ocupados	integer	FALSE
edad_promedio	numeric	FALSE
jefe_mujer	integer	FALSE
jefe_edad	integer	FALSE

Cuadro 13. Variables con Mayor Porcentaje de Outliers

variable	pct_outliers	min_value	max_value
Lp	12.00	167222.48	303816.69
P5000	10.74	1.00	98.00
Ingpcug	8.00	0.00	88833333.33
P5130	7.47	98.00	600000000.00
n_mujeres	7.38	0.00	14.00
Ingtotugarr	7.09	0.00	88833333.33

Cuadro 14. Variables Más Correlacionadas con Pobreza

variable	correlacion_con_pobreza
n_menores	0.36
Ingtotugarr	-0.30
Ingpcug	-0.28
Npersug	0.24
Nper	0.24
n_miembros	0.24
n_mujeres	0.21
edad_promedio	-0.20
P5000	-0.14
P5090	0.14

Cuadro 15. Variables Redundantes y Propuestas de Eliminación

par_correlacionado	correlacion	propuesta_eliminar	propuesta_mantener	justificacion
Nper y n_miembros	1.00	Nper	n_miembros	Mayor corre
Nper y Npersug	1.00	Nper	Npersug	Mayor corre
Npersug y n_miembros	1.00	n_miembros	Npersug	Mayor corre
Nper y n_mujeres	0.78	n_mujeres	Nper	Mayor corre
n_miembros y n_mujeres	0.78	n_mujeres	n_miembros	Mayor corre
Ingtotugarr y Ingpcug	0.78	Ingpcug	Ingtotugarr	Mayor corre
edad_promedio y jefe_edad	0.78	jefe_edad	edad_promedio	Mayor corre
Npersug y n_mujeres	0.78	n_mujeres	Npersug	Mayor corre
Npersug y n_menores	0.76	Npersug	n_menores	Mayor corre
Nper y n_menores	0.76	Nper	n_menores	Mayor corre
n_miembros y n_menores	0.76	n_miembros	n_menores	Mayor corre