Mejor Modelo: Random Forest

Equipo 03 — Catalina Leal, Lucas D. Carrillo, Lucas E. Veras

Big Data y Machine Learning — PS2

October 20, 2025

Descripción General del Mejor Modelo

Random Forest

Versión 1: Bagging (Hellinger)	Versión 2: Selección de variables (Gini)
`num.trees` = 500	`num.trees` = 1000
`mtry` = 26 (todas las variables)	`mtry` = 6
`node.size` = 50	`node.size` = 1
`split.rule` = Hellinger	`split.rule` = Gini
Umbral $\tau^* = 0.522$ (max F1 en CV)	Umbral $\tau^* = 0.340$ (max F1 en CV)

Kaggle (Public Leaderboard): F1 = 0.63 (ambos modelos)

Fuente de datos

- Los datos provienen de la *Medición de Pobreza Monetaria y Desigualdad 2018*, elaborada por el *Departamento Administrativo Nacional de Estadística (DANE)* en el marco de la misión *Empalme de las Series de Empleo, Pobreza y Desigualdad (MESE)*.
- Esta iniciativa busca unificar y armonizar las series de empleo, pobreza y desigualdad en Colombia, fortaleciendo la consistencia de los indicadores sociales.
- La fuente contiene información detallada sobre ingresos, educación, ocupación, vivienda y composición familiar, tanto a nivel individual como de hogar.
- Los datos se obtuvieron de la plataforma *Kaggle*, donde se desarrolla una competencia para predecir la condición de pobreza.

Procesamiento y limpieza de datos

- Se integraron las bases de hogares y personas mediante el identificador único `id`.
- Se excluyeron las variables con más del 40% de valores faltantes, evitando imputaciones poco confiables.
- La selección final se basó en criterios definidos por el equipo a partir de supuestos teóricos y empíricos sobre los determinantes de la pobreza.
- Las variables de ingreso del conjunto de entrenamiento no se incluyeron, ya que no estaban disponibles en el conjunto de prueba.

Tamaño muestral:
164.959 hogares en el conjunto de entrenamiento y 66.168 en el de prueba (40% del primero).

Variables del mercado laboral

- Se construyeron las tasas de **ocupación** e **inactividad**, definidas como:

$$t_{ocu} = \frac{\text{Ocupados}}{\text{Personas en edad de trabajar}}, \quad t_{inac} = \frac{\text{Inactivos}}{\text{Personas en edad de trabajar}}$$

- Hipótesis: mayor participación laboral $\Rightarrow$ menor probabilidad de pobreza.
- Otras variables laborales incluidas:
  - Horas totales trabajadas por los miembros del hogar.
  - Número de trabajadores por cuenta propia.
  - Dicotómica: jefe(a) de hogar cuenta propia.
- Estas variables reflejan la **resiliencia económica** del hogar ante choques del mercado laboral.

Características socioeconómicas y demográficas

- Variables demográficas y de género:
  - Número de mujeres.
  - Jefe(a) de hogar mujer (binaria).
  - Tasa de dependencia:

$$t_{dep} = \frac{\text{Población } <15 \text{ o } >65}{\text{Población en edad de trabajar}}$$

- Variables educativas:
  - Proporción de personas con educación superior.
  - Nivel educativo máximo alcanzado en el hogar.
- Estas medidas capturan el **capital humano** y su efecto sobre los ingresos y la movilidad social.

Condiciones habitacionales y geográficas

- Variables de vivienda:
  - Número de habitaciones.
  - Tipo de vivienda (propia, arrendada, usufructo, ocupada sin título).
- Variable demográfica de control:
  - Número total de personas en el hogar.
- Variable geográfica:
  - Dominio (área metropolitana o rural).
- Estas variables complementan las dimensiones laboral y educativa, reflejando el **bienestar material** y las **condiciones estructurales del hogar**.

Estadísticas descriptivas

Estadísticas descriptivas para las variables numéricas

Variable	Media		Mediana		Desv. Est.	
	Pob.	No Pob.	Pob.	No Pob.	Pob.	No Pob.
N. Personas	4.13	3.08	4.00	3.00	2.03	1.64
N. Habitaciones	3.03	3.48	3.00	3.00	1.13	1.25
Edad Jefe	46.77	50.32	45.00	50.00	16.24	16.35
H. Trab. Jefe	28.62	34.49	32.00	42.00	25.23	24.67
N. Mujeres	2.24	1.62	2.00	1.00	1.35	1.10
N. Menores	1.55	0.62	1.00	0.00	1.34	0.86
N. Mayores	0.27	0.33	0.00	0.00	0.57	0.61
N. Educ. Sup.	0.36	0.88	0.00	1.00	0.69	1.00
N. Cta. Propia	0.86	0.65	1.00	0.00	0.85	0.80
Horas Trab. Tot.	52.46	71.11	48.00	60.00	43.68	49.59
N. Ocupados	1.26	1.57	1.00	1.00	0.99	1.03
N. PET	3.02	2.64	3.00	2.00	1.53	1.35
N. Inactivos	1.45	0.92	1.00	1.00	1.17	0.96
Tasa Ocup.	0.43	0.62	0.50	0.67	0.31	0.34
Tasa Inac.	0.46	0.33	0.50	0.33	0.31	0.33
Tasa Educ. Sup.	0.13	0.35	0.00	0.25	0.25	0.38
Tasa Dependencia	67.42	37.84	60.00	33.33	52.86	39.45
Tasa Cta. Propia	0.30	0.26	0.25	0.00	0.29	0.33

Estadísticas descriptivas

Table: Promedios de variables categóricas por condición de pobreza

Variable	Pobre	No Pobre
Jefe del Hogar Mujer	0.468	0.406
Jefe del Hogar Ocupado	0.643	0.727
Jefe del Hogar Cuentra Propia	0.473	0.322
Tipo de vivienda		
Propia Totalmente Pagada	0.275	0.403
Propia la Están Pagando	0.017	0.038
En arriendo o subarriendo	0.438	0.379
En usufructo	0.158	0.149
posesión sin título	0.111	0.030
Nivel educativo del jefe del hogar		
Preescolas	0.0002	0.00003
Primaria	0.378	0.259
Medias	0.253	0.263
Secundaria	0.162	0.124
Universitaria	0.103	0.315
Máximo nivel educativo del hogar		
Preescolar	0.001	0.0001
Primaria	0.128	0.093
Media	0.385	0.264
Secundaria	0.201	0.087
Universitaria	0.261	0.544

Conclusiones descriptivas

- Los hogares pobres presentan mayor número de integrantes, más menores y mayores tasas de inactividad.
- Exhiben menor nivel educativo y mayor proporción de trabajo por cuenta propia.
- Los hogares no pobres muestran una inserción laboral más estable y mejores condiciones de vivienda.
- Estas diferencias confirman la coherencia teórica de las variables incluidas en el modelo predictivo.