Mejor Modelo: Random Forest

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

Big Data y Machine Learning — PS2

20 de octubre de 2025

Descripción General del Mejor Modelo

Random Forest

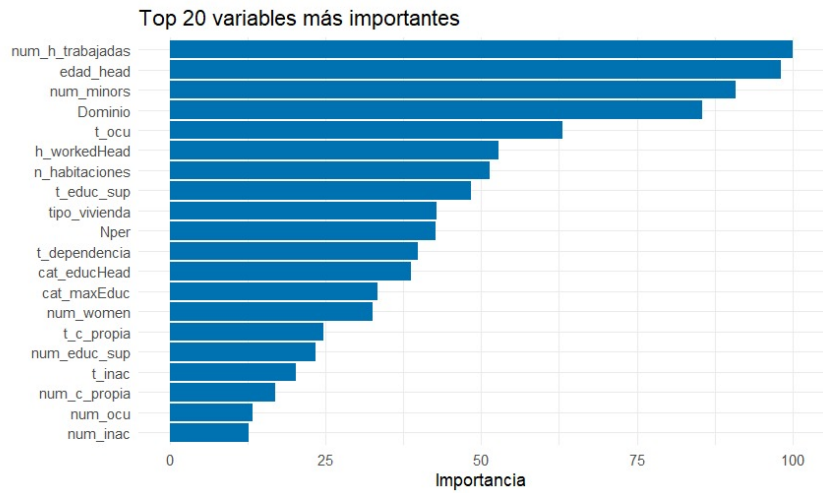
Versión 1: Bagging (Hellinger)	Versión 2: Selección de variables (Gini)
`num.trees` = 500	`num.trees` = 1000
`mtry` = 26 (todas las variables)	`mtry` = 6
`node.size` = 50	`node.size` = 1
`split.rule` = Hellinger	`split.rule` = Gini
Umbral $\tau^* = 0{,}522$ (max F1 en CV)	Umbral $\tau^* = 0{,}340$ (max F1 en CV)

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

Versión 1: Bagging

- **Idea**: ensamble con `mtry = 26` (todas), `num.trees = 500`; **Hellinger** favorece la clase minoritaria.
- **Tuning**: 5-fold CV con métrica **F1**; barrido de $\tau \in [0{,}05, 0{,}95] \Rightarrow \tau^* = 0{,}522$.
- **Diagnóstico**: menor varianza, buena sensibilidad en hogares pobres; *node.size*=50 evita sobreajuste.
- **Score**: **F1 = 0.63**.

Importancia (permutación)

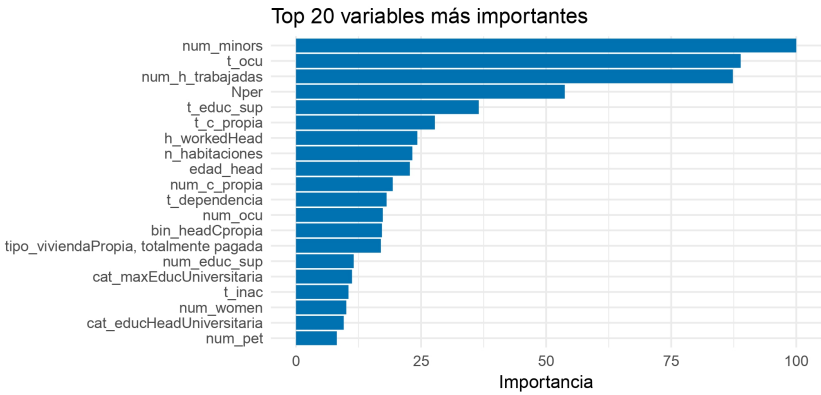


Top señales: horas trabajadas totales, edad del/la jefe, # menores, tasa de ocupación, tipo de vivienda.

Versión 2: Selección de Variables

- **Idea**: submuestreo de variables por división (`mtry = 6`) para reducir correlación entre árboles; **Gini**.
- **Tuning**: 5-fold CV + barrido de umbral $\Rightarrow$ $\tau^* = 0{,}340$ para maximizar **F1**.
- **Diagnóstico**: `node.size`=1 permite árboles profundos; `num.trees`=1000 estabiliza el ensamble.
- **Score**: **F1 = 0.63** (empate técnico).

Importancia (Gini)



Patrón similar al Bagging: mercado laboral, demografía y vivienda dominan la predicción.

Comparación con otros modelos

- **Métrica única**: **F1** (alineada con Kaggle).
- **Hallazgo**: RF domina por su capacidad de capturar no linealidades e interacciones, su robustez ante colinealidad y la calibración del umbral τ^*.

Modelo	F1	Notas
RF (Bagging, Hellinger)	**0.63**	mtry=26, trees=500, node=50, $\tau^* = 0{,}522$
RF (Gini, mtry=6)	**0.63**	trees=1000, node=1, $\tau^* = 0{,}340$
RF (Umbral ROC)	0.62	Misma familia; umbral definido por curva ROC, sin ajuste de F1 específico
Logit	0.62	Lineal en log-odds; mejora con calibración de τ^* pero no capta interacciones complejas
Lineal (MPL)	0.61	Fronteras lineales; desempeño estable y simple
Elastic Net	0.60	Regulariza coeficientes; penaliza variables redundantes pero pierde no linealidad
Gradient Boosting	0.57	Árboles poco profundos; sin optimización de pesos de clase ni early stopping
Naive Bayes (Down)	0.51	Supone independencia; mejora con remuestreo pero persiste subajuste

Aprendizajes y Recomendaciones

- **Principales aprendizajes metodológicos**
  - El desempeño superior del **Random Forest** se explica por su capacidad para capturar interacciones y no linealidades, y por su **robustez frente a colinealidad y ruido**.
  - La combinación de **criterio Hellinger** y **calibración explícita del umbral** (τ^*) resultó determinante para optimizar la métrica de **F1** en contextos desbalanceados.
  - Las variables con mayor poder predictivo se concentran en la **inserción laboral del hogar**, la **estructura demográfica** y las **condiciones habitacionales**.
- **Implicaciones para política pública y focalización**
  - La calibración de τ^* permite una mejor identificación de hogares vulnerables y, por tanto, una focalización más precisa de la política.
  - Es necesario incorporar esquemas de **validación fuera de muestra por Dominio** (territorio) y monitoreo de posibles **sesgos de exclusión**.
  - Estrategias **cost-sensitive** o el uso de **class weights** son recomendables para penalizar los falsos negativos en poblaciones vulnerables.
- **Líneas de extensión técnica**
  - Ampliar el **grid de hiperparámetros** y explorar variantes de boosting como *XGBoost*.