# Comparación entre modelos

Catalina Leal Rojas, Lucas Daniel Carrillo Aguirre, Lucas Eduardo Veras Costa

October 21, 2025

# Nuestro mejor modelo: **Random Forest**

- Utilizamos de base un modelo lineal que incluye las 26 variables de la muestra. Inlcuimos variables que describen las condiciones demográficas, laborales, educativas y de vivienda de los hogares colombianos.
- Hiperparámetros:
  - Número de árboles y muestras bootstrap = 1000 (valor fijo)
  - Número de variables seleccionadas aleatoriamente para cada partición = 6 (de grilla [6,8,9,10] con cross-validation de 5 folds y F1 como métrica para optimizar)
  - Número mínimo de observaciones por nodo = 1 (valor fijo)
- Desbalance de clases:
  - Umbral de clasificación = 0.34 (el valor que maximiza el F1 dentro de muestra entre una grilla (0.05,0.95))
  - Downsampling
- **Kaggle score:**0.63

# Top 5 otros modelos

- Random Forest (Bagging) (0.63)
- Logit (0.62)
- Random Forest (diferente técnica para tratar el desbalance de clases) (0.61)
- Lineal (0.61)
- Elastic Net (0.60)

# Bagging (0.63)

- Hiperparámetros:
  - Número de árboles y muestras bootstrap = 500 (valor fijo)
  - Número de variables seleccionadas aleatoriamente para cada partición = 26 (todas las variables)
  - Número mínimo de observaciones por nodo = 50 (de grilla [1,5,15,50] con cross-validation de 5 folds y F1 como métrica para optimizar)
  - *Split Rule* = Hellinger (de grilla [Gini, Hellinger] con cross-validation de 5 folds y F1 como métrica para optimizar)
- Desbalance de clases:
  - Umbral de clasificación = 0.52 (el valor que maximiza el F1 dentro de muestra entre una grilla (0.05,0.95))
  - Downsampling

# Bagging (0.63)

**¿Por qué da el mismo puntaje?:**

- Creemos que se alcanzó una región de estabilidad en el espacio de hiperparámetros (más árboles no mejoran mucho la predicción).
- El número de observaciones en cada nodo podría ser un ejemplo del trade-off sesgo-varianza
- **Buena selección de variables:** las variables con mayor poder predictivo tienden a ser seleccionadas de manera recurrente en las primeras divisiones de los árboles (bosques similares, incluso con distintos hiperparámetros).
- El criterio de división Hellinger (pensado para el desbalance de clases) pudo haber compensado la menor complejidad en términos de número de árboles y condiciones sobre los nodos terminales.

# Logit (0.62)

- Se utilizó un modelo lineal que incluía las 26 variables creadas por el equipo
- Desbalance de clases:
  - Umbral de clasificación = 0.21 (el valor que maximiza el AUC (ROC))

# Logit (0.62)

**¿Por qué da casi el mismo puntaje?:**

- Variables altamente informativas y bien seleccionadas
- Trade-off sesgo-varianza
- En ambos se buscaron soluciones para el desbalance de clases. Parece indicar que solamente cambiar el umbral de clasificación puede mejorar sustancialmente cómo el modelo identifica la clase minoritaria ($F_1$)

# Random Forest: otra técnica para el desbalance (0.61)

- Mismos hiperparámetros que el modelo ganado
- La diferencia radica en que el umbral de clasificación fue elegido con la maximización de la AUC (ROC) igual a 0.21
- En este caso, elegir la misma métrica que evalúa Kaggle, mejora marginalmente el desempeño del modelo.

# Modelo de probabilidad lineal (0.61)

$$Pobre_i = \beta_0 + \mathbf{X}_i'\boldsymbol{\beta} + \beta_1 Nper_i^2 + \beta_2 edad\_head_i^2 + \beta_3(bin\_headWoman_i \times cat\_educHead_i) + \beta_4(edad\_head_i \times cat\_educHead_i) + \beta_5(Nper_i \times t\_dependencia_i) + \varepsilon_i \quad (1)$$

- Desbalance de clases:
  - Umbral de clasificación $= 0.27$ (el valor que maximiza el AUC (ROC))

# Modelo de probabilidad lineal (0.61)

- Para compensar las predicciones fuera de [0,1] y la menor eficiencia del modelo lineal frente a Logit, se incluyeron términos cuadráticos e interacciones para añadir un poco más de complejidad
- También creemos que el cálculo de un umbral de clasificación mejor ajustado a predecir los verdaderos positivos logra equiparar un modelo tan sencillo con técnicas como Random Forest

# Elastic Net (0.60)

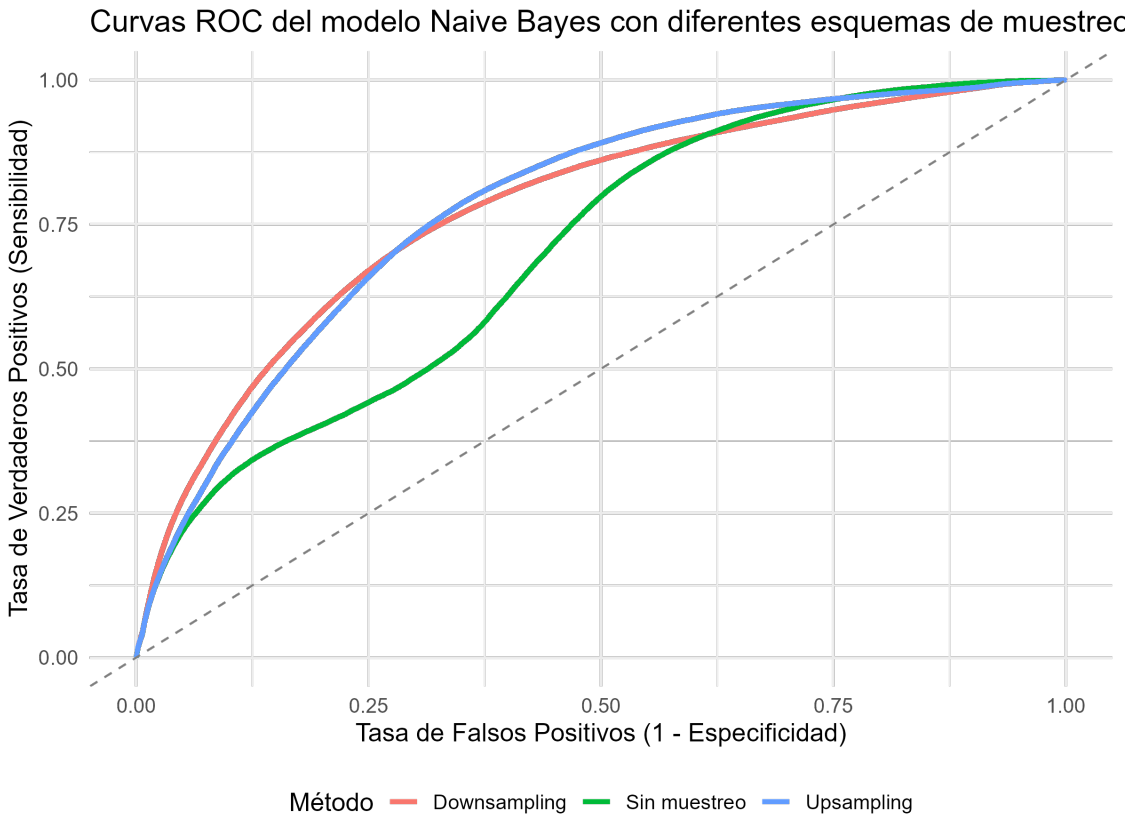
- Hiperparámetros:
  - $\alpha = 0.7$ (grilla de 0 a 1, con saltos de 0.1). Este modelo es 70% Lasso y 30% Ridge
  - $\lambda = 0.001$ (grilla de $10^{-3}$ a $10^{3}$, con 10 valores logarítmicamente espaciados)
- Desbalance de clases:
  - Umbral de clasificación $= 0.2$ (el valor que maximiza el F1 dentro de muestra entre una grilla (0.05,0.95))

# Elastic Net (0.60)

- Aunque combina regularización LASSO y Ridge para manejar multicolinealidad y selección de variables, sigue siendo un modelo lineal en los parámetros. Esto, como ya vimos, tiene un desempeño ligeramente menor que otros modelos como Random Forest
- Nuevamente, los resultados siguen cercanos a otros modelos más complejos gracias al tratamiento para el desbalance de observaciones entre pobres y no pobres.

# Los que quedaron por fuera del top

- Gradient Boosting
- Elastic net. De este último realizamos una prueba para ver los efectos del remuestreo en este modelo en particular:



# Conclusión

- Los modelos presentaron desempeños muy similares, reflejando la alta calidad y relevancia de las variables construidas.
- El *Random Forest* obtuvo el mejor rendimiento ($F_1 = 0.63$) por su capacidad para capturar relaciones no lineales e interacciones complejas.
- Modelos más simples como *Logit* y el modelo lineal lograron resultados comparables, mostrando que la selección de variables fue más determinante que la complejidad del algoritmo.
- El tratamiento del desbalance de clases mediante remuestreo y calibración del umbral con $F_1$ mejoró la identificación de los hogares pobres.
- Una buena preparación de datos y una calibración cuidadosa pueden igualar o superar la complejidad algorítmica en la predicción de pobreza.