Machine Learning
Trabajo Práctico 2

Nicolás Abbate, Rocío Bisang, Lucio Garay Méndez

December 18, 2023

Los códigos utilizados para computar los resultados de este trabajo están disponibles en el repositorio `https://github.com/Queeno11/ml_problem_set_2`.

Introducción

La identificación de la pobreza representa un desafío a nivel global, y su comprensión precisa es esencial para el diseño de políticas efectivas, correctamente focalizadas y orientadas al desarrollo sostenible de los países. Estimar la pobreza de manera precisa es sin embargo una tarea costosa y compleja, por lo que el desarrollo de métodos que permitan predecir la misma cobra particular relevancia, en particular en países en desarrollo, como Colombia, donde según las últimas mediciones, la pobreza monetaria alcanzó en 2022 al $36,6\%$ de la población.

Si bien la selección de los modelos fue arbitraria, nos basamos parcialmente en los trabajos ganadores de T-test Poverty del Banco Mundial. Las estrategias ganadores tienen en común (1) el uso de modelos de boosting, en particular XGboost y LightGBM, junto a redes neuronales relativamente simples, (2) el agrupamienti (bagging) de modelos para reducir la varianza de las predicciones, y (3) el uso de un gran numero de variables para los modelos. Utilizando estos criterios, buscamos utilizar algunos de estos modelos tanto para clasificación como para regresión. Asimismo, decidimos ir por un enfoque generalista y agnóstico de no eliminar información potencialmente relevante para los modelos, por lo que mantuvimos la mayor cantidad posible de variables en los distintos datasets. Además de esto, buscanos resolver el problema de los valores faltantes generando variables binarias para todas las columnas que identificaran si, para ese hogar, esa variable tenía valores faltantes; al tiempo que se rellenan los valores restantes con ceros. Finalmente, todas las variables (menos el ingreso en el caso de regresión) se estandarizaron según su media y su varianza, de forma diferenciada para los diferentes conjuntos. Con estos puntos en mente implementamos diez modelos diferentes, obteniendo un F1 mayor con el modelo XGBoost de clasificación de pobraza.

Datos

Para la estimación de la pobreza a nivel hogar se utilizan los datasets provistos en clase, tanto la encuesta a nivel personas como a nivel de hogares, que surgen de los microdatos

de la "Medición de Pobreza Monetaria y Desigualdad 2018" efectuada por el Departamento Administrativo Nacional de Estadística (DANE). Con el objetivo de estimar de forma mas precisa la pobreza a nivel hogares, se incorporan de manera progresiva al análisis variables construidas a partir de la muestra individual.

Al momento de seleccionar las variables predictoras se especificó que cumplan con la condición de que estuvieron tanto en el train set como en el test set. Esto se realiza en un contexto donde el conjunto de entrenamiento tenía más variables que el conjunto de prueba y esto generaría problemas a la hora de predecir la variable de interés, pobreza, con menos variables de las que fue entrenada. Por fuera de las variables en común, se seleccionó la variable Pobreza e Ingreso del hogar, que ambas serían necesarias para el enfoque de estimación directo e indirecto, respectivamente.

Otro procedimiento realizado fue eliminar las variables que se repetían en el set de hogares y en el set de individuos para no doble contabilizar variables.

Además, se estandarizaron las columnas numéricas con media igual a cero. Las columnas categóricas se codificaron como tipos de datos categóricos. En lo referido a la completitud de los datos, en ciertas variables se observó la presencia de missing data, por lo que la manera de tratar con este asunto fue crear dummies que indiquen aquel caso en el que un dato de una variable, ya sea del hogar o de algun individuo, estuviese faltando para ver si esta dummy que se construye correlaciona con la variable de interés y para limpiar la base de valores faltantes. Todos estos procedimientos son beneficiosos para los algoritmos de optimización que serán luego utilizados en los modelos de predicción.

En un primer dataset, A, se le añade a cada hogar además de las variables propias del hogar, todas las variables observables individuales del jefe de hogar.

El segundo dataset, B, consta de todas las variables que constituyen el dataset A y se le suman como variables todas las características observables de primeros 10 integrantes del hogar adicionales al jefe de hogar. Para los modelos más simples, es esperable que esta estrategia resulte en estimaciones con varianza alta. En modelos que filtran la importancia de variables, como CART, XGBoost o LigthGBM, ese problema es menor.

Por último, el tercer dataset además de contar con todas las variables del dataset A se les agregaron ocho variables construidas a partir de combinaciones entre las variables de la encuesta de hogares y de individuos. Estas son:

- El ratio de hijos en edad escolar que se encuentran actualmente estudiando dentro del hogar: incluye individuos entre 10 y 18 años cuya actividad principal es el estudio.
- El ratio de personas buscando trabajo dentro del hogar.
- La cuota de amortización per cápita del hogar, entendida como aquello que efectivamente pagan sobre la cantidad de integrantes del hogar.
- La cuota de amortización per cápita estimada del hogar, entendida como aquello que estiman que pagarían en el caso de pagar arriendo sobre la cantidad de integrantes del hogar.

- El ratio de personas que viven en el hogar y la cantidad de dormitorios de la misma, como medida de hacinamiento.
- Una variable dummy que indica si hay una persona mayor (mayores de 65 años) viviendo dentro del hogar, condicional a que el jefe de hogar sea menor a 65 años y que no sea el cónyuge del jefe de hogar.
- Una variable dummy que identifica los hogares donde el jefe de hogar es menor a 25 años de edad y hay al menos un niño en el hogar, con el fin de identificar el fenómeno de padres jóvenes.
- La proporción de niños en el hogar, con el objetivo de ver la proporción de población pasiva menor de edad en el hogar en relación a la activa, dejando de lado las personas mayores cuya información es traída por otra variable construida.

Análisis de los datos

Con el objetivo de comprender mejor la información, realizamos un análisis exploratorio de los datos. En primer lugar calculamos algunas estadísticas descriptivas básicas de la distribución de la pobreza e indigencia a nivel hogares 1. Además incorporamos una visualización gráfica de las distintas distribuciones de ingresos, medidas en logaritmo, por región, con sus respectivas líneas de pobreza e indigencia 2.

Región	% Pobreza	% Indigencia	Ingreso Promedio	Desvío
MEDELLIN	11,1	2,5	2.753.886	3.106.650
BARRANQUILLA	15,6	1,9	2.590.081	2.596.069
BOGOTA	9,8	2,4	3.180.430	4.230.223
CARTAGENA	19,1	2,6	2.277.562	2.513.515
TUNJA	13,7	2,6	2.407.702	2.416.819
MANIZALES	8,3	1,7	2.598.130	2.803.279
FLORENCIA	26,8	5,6	1.687.100	1.995.882
POPAYAN	24,2	7,4	1.957.092	2.109.959
VALLEDUPAR	26,7	6,0	1.893.891	1.991.576
MONTERIA	21,1	2,8	1.999.533	2.143.865
QUIBDO	32,2	10,7	1.707.237	1.951.103
NEIVA	17,8	3,5	2.203.768	2.513.953
RIOHACHA	32,2	9,9	1.947.249	2.139.301
SANTA MARTA	22,8	4,5	2.051.998	2.036.126
VILLAVICENCIO	13,1	3,4	2.328.211	2.942.517
PASTO	20,2	3,2	2.119.366	2.309.842
CUCUTA	27,8	5,6	1.718.115	1.931.105
ARMENIA	15,7	3,0	2.001.036	2.086.660
PEREIRA	8,7	1,2	2.185.964	2.170.691
BUCARAMANGA	11,8	2,0	2.635.917	2.862.747
SINCELEJO	21,6	2,4	2.114.157	2.259.001
IBAGUE	13,4	2,4	2.132.576	2.497.939
CALI	11,7	3,1	2.446.537	2.696.386
RESTO URBANO	29,5	6,3	1.641.149	1.935.903
RURAL	30,0	12,4	949.033	1.138.949

Figure 1: Estadísticas básicas por región

Por otra parte, una de las variables más relevantes en el presente análisis fue el número de personas en el hogar, que es de las que más correlaciona con la dummy indicativa de pobreza. En tal sentido, la Figura 3 muestra que a medida que aumenta la cantidad de integrantes del hogar, aumenta en promedio la proporción de hogares pobres.

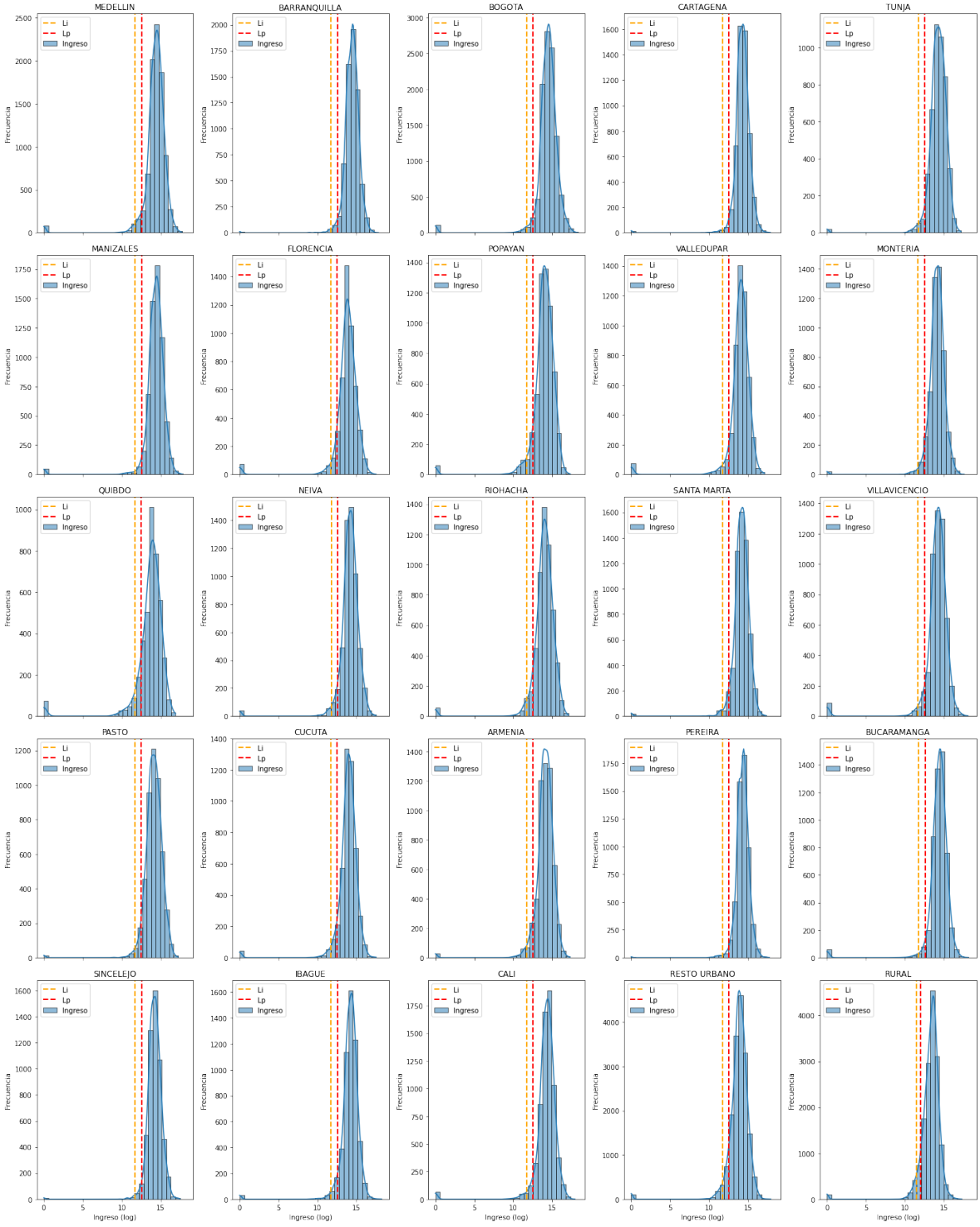


Figure 2: Distribución del ingreso en log

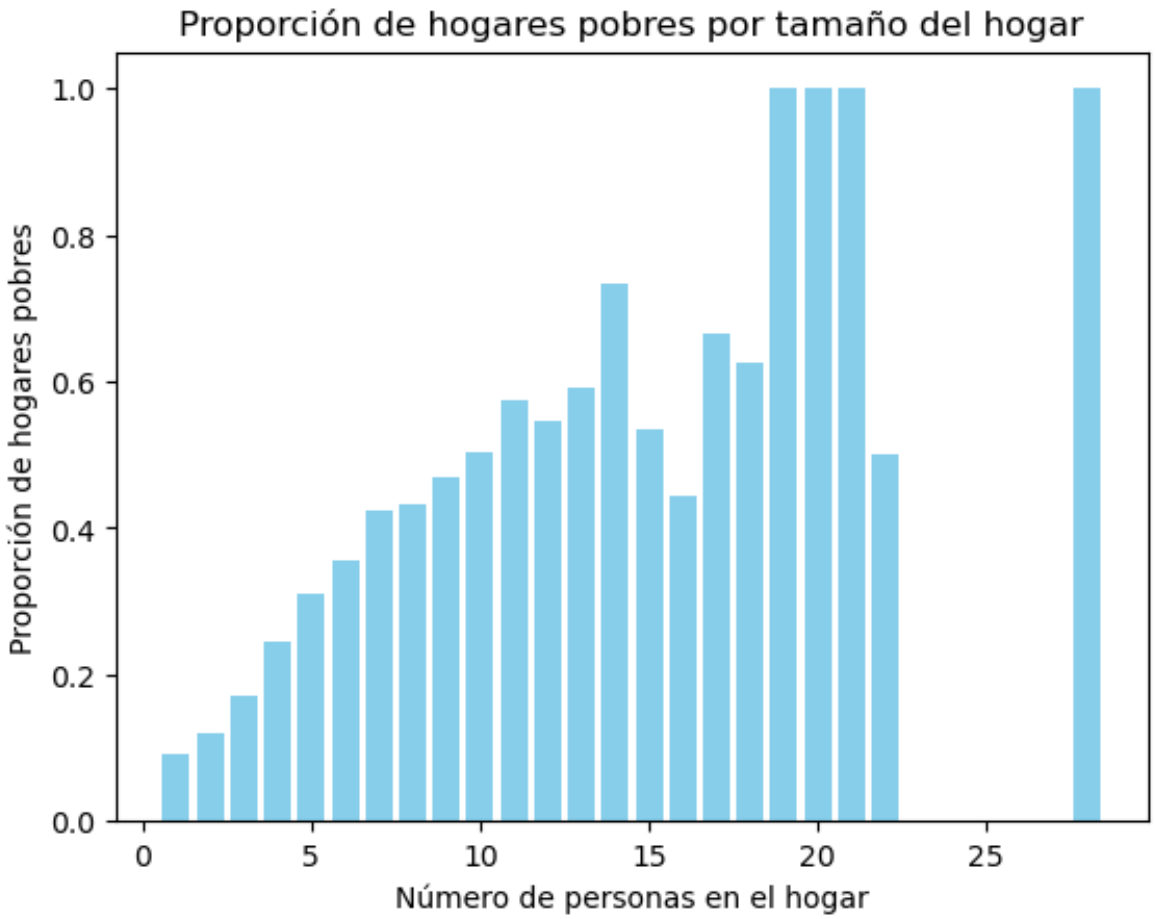


Figure 3: Proporción de hogares pobres según cantidad de integrantes

Tal como se mencionó anteriormente, los datasets utilizados cuentan con valores faltantes que en el caso de las variables a nivel hogar pueden observarse en la Figura 4. En este caso, está toda la información disponible para casi todas las variables a excepción del Dominio, en referencia a la ubicación geográfica y por otra parte, las variables que refieren al pago por arriendo o el estimado, se complementan es decir, entre ambas columnas se tiene la información para el total de la muestra.

Por su parte, el problema de los datos faltantes es relevante para las características observables a nivel individual, donde se agrava a medida que se profundiza en el número de integrante del hogar. En la Figura 5 para el caso del jefe de hogar y en la Figura 6 para el segundo integrante entrevistado, en el caso que exista, se muestran los valores faltantes. Ambos resultados justifican la adición de las dummies indicativas si para cierta característica observable, el individuo computa un missing value.

Además, se realizó un análisis de correlación entre las variables que constituyen la base de datos a nivel hogar con la dummy indicativa de pobreza y los resultados muestran que la

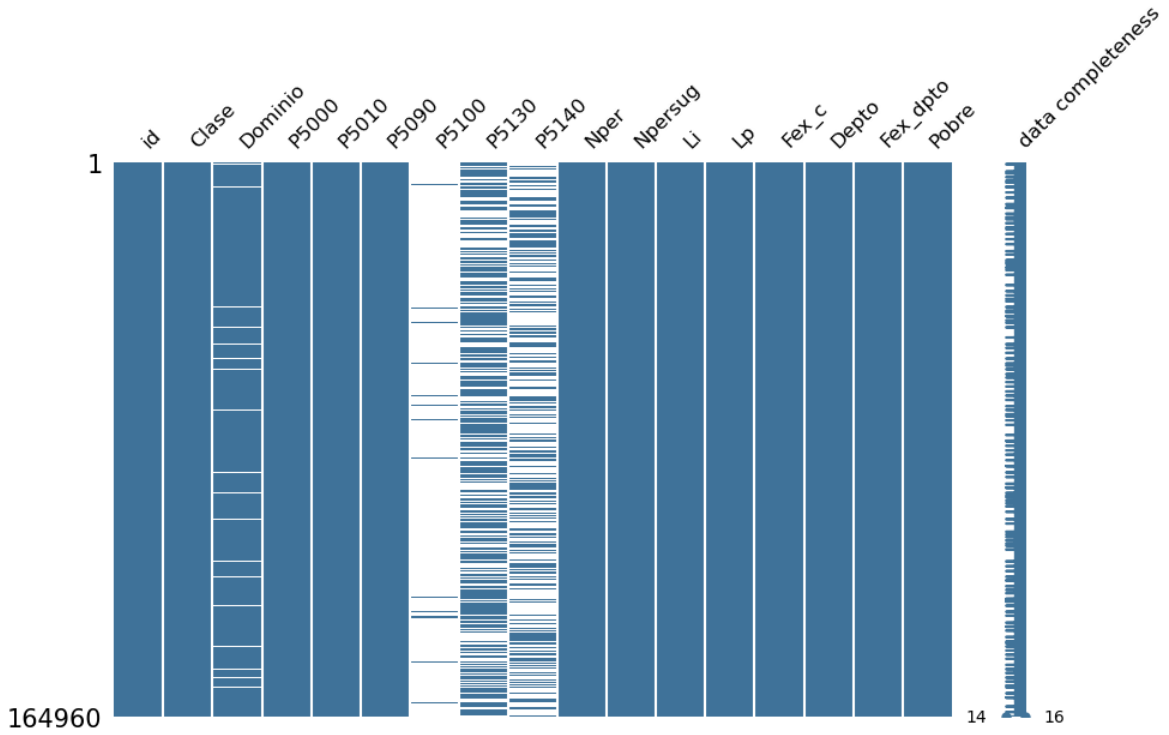


Figure 4: Data completeness: Encuestas hogares

correlación más fuerte es con el número de personas en el hogar (Nper), seguido de si son propietarios o arrendatarios del hogar (P5090) y de la cantidad de ambientes que posee el hogar (P5000). El resto de los valores pueden observarse en la Figura 7

Por último, se analizó la correlación de las ocho variables construidas con la variable de interés. Hay resultados auspiciosos, si se los compara con los coeficientes de correlación del párrafo anterior. La variable que muestra la proporción de niños en edad escolar que se encuentran actualmente estudiante, la variable que muestra la proporción de niños en el hogar sobre el total de integrantes y la variable que computa el ratio entre miembros del hogar y ambientes del mismo muestran un coeficiente de correlación mayor al 0,3, por encima de todos los coeficientes obtenidos previamente. Por su parte, la variable que indica la presencia de padres jóvenes en el hogar y la variable que muestra la proporción de individuos que buscan trabajo dentro de la unidad muestran un coeficiente entre 0,05 y 0,1, más parecido a los coeficientes resaltados en el párrafo anterior.

Predicción de la pobreza

Dado que la pobreza es una identificación binaria basada en la relación entre el ingreso y una línea de pobreza (que puede variar entre hogares), el problema de estimación puede ser resuelto tanto utilizando un modelo de predicción del ingreso, que luego se compara con la línea de pobreza de esos hogares, como también con un modelo de clasificación, identificando

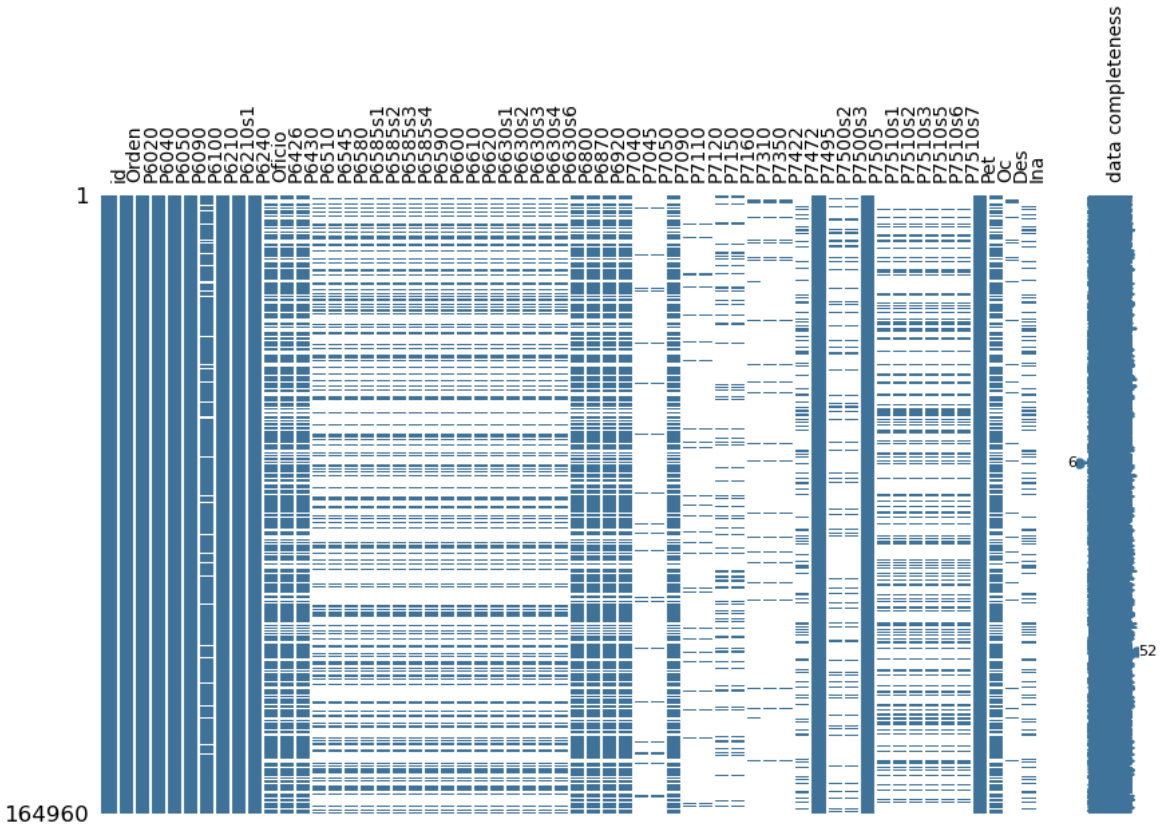


Figure 5: Data completeness: Encuestas invididuos - Jefe de hogar

si un hogar es pobre o no.

Para los modelos de regresión utilizamos (a) un modelo lineal, (b) un modelo RIDGE, (c) un modelo de Redes Neuronales, (d) un modelo de LightGBM y (e) un bagging de los modelos previos. La optimización se realizó minimizando el error cuadrático medio. Además, la selección del modelo se realizo dividiendo el dataset entre validación y entrenamiento utilizando un algoritmo de partición KFold con 5 particiones, de forma tal de garantizar una correcta evaluación de la performance del modelo. No se realizaron técnicas de resampling para resolver el desbalance de clases ya que la predicción del modelo era el ingreso per capita del hogar.

Para los modelos de clasificación utilizamos (a) un modelo CART, (b) un modelo XGBoost, (c) un modelo de LightGBM, (d) un modelo de Redes Neuronales y (e) un bagging de los previos. La optimización se realizó minimizando la función de entropía cruzada (log loss). Para los modelos de clasificación, utilizamos un algoritmo de KFold estratificado para corregir el desbalance de clases entre hogares pobres y no pobres, ya que la categoría de pobre era aproximadamente el 20% de las observaciones. Se consideró en las predicciones a hogares pobres aquellos cuya probabilidad predicha sea mayor a al 50%.

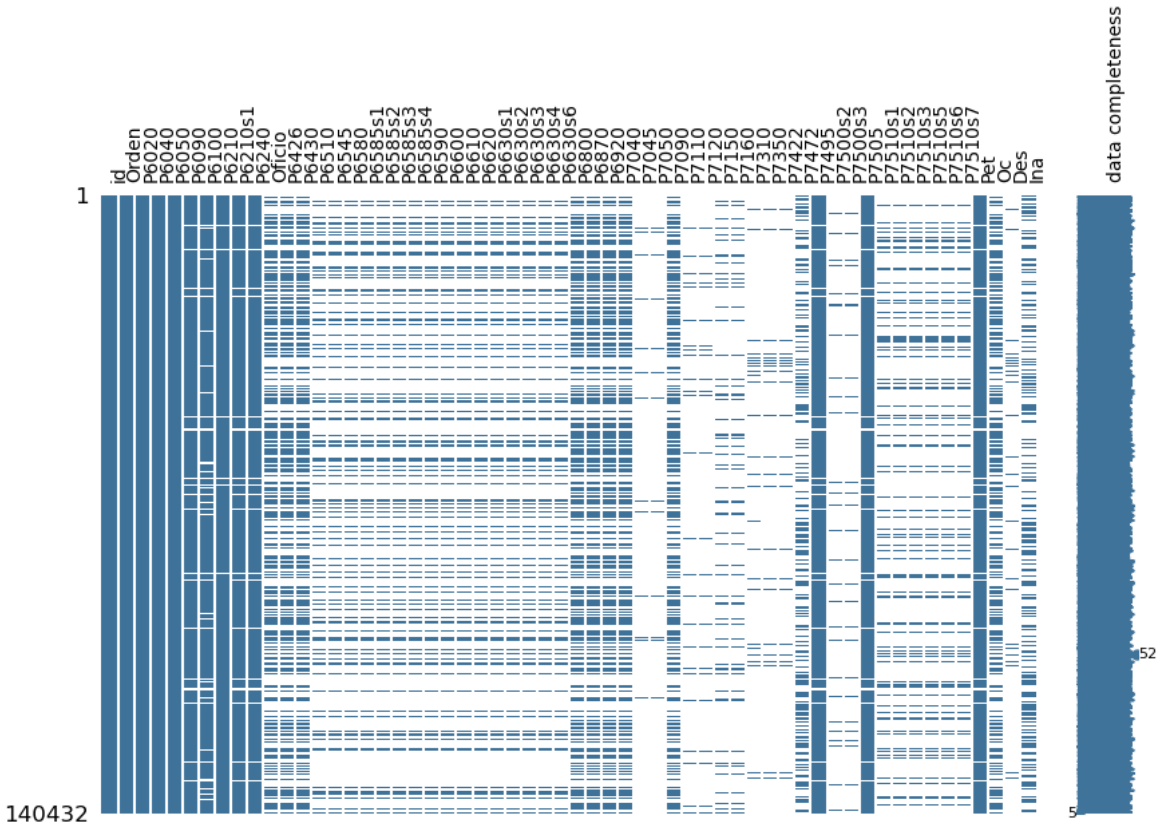


Figure 6: Data completeness: Encuestas invididuos - Jefe de hogar

Las predicciones enviadas a Kaggle son la moda de las predicciones de los 5 modelos entrenados en cada uno de los KFolds. Es importante destacar que en algunos de los modelos enviados, las métricas obtenidas en el conjunto de prueba difieren significativamente de las métricas obtenidas en el conjunto de validación. Entendemos que esto se debe a un error no identificado en el código que no logramos solucionar en el tiempo disponible.

La figura 9 muestra la performance promedio en los conjuntos de validacion de cada uno de los modelos enviados. Como puede observarse, en general los modelos de clasificación generan una mejor performance que los de regresión. Es posible que esto se deba a que, para predecir pobreza a través del ingreso, es necesario tener una importante presicion ya que variaciones pequeñas en el ingreso cercano a la linea generan cambios discretos en la clasificación.

Presentamos en la figura 10 los resultados particulares de los dos modelos de mayor rendimiento que obtuvimos: XGBoost para clasificación y LightGBM para clasificación. En el primer caso, el modelo predice un 17.5% de pobres (frente a un verdadero 20%). En base a esta matriz, parece ser que el modelo identifica con bastante certeza los hogares no pobres (ya que solo el 5% de los hogares no pobres está mal identificado como pobre), mientras que es más complejo para el modelo identificar a hogares realmente pobres (el 31.3% son identificados incorrectamente como no pobres). En el segundo caso se observa prácticamente el mismo patrón: una leve

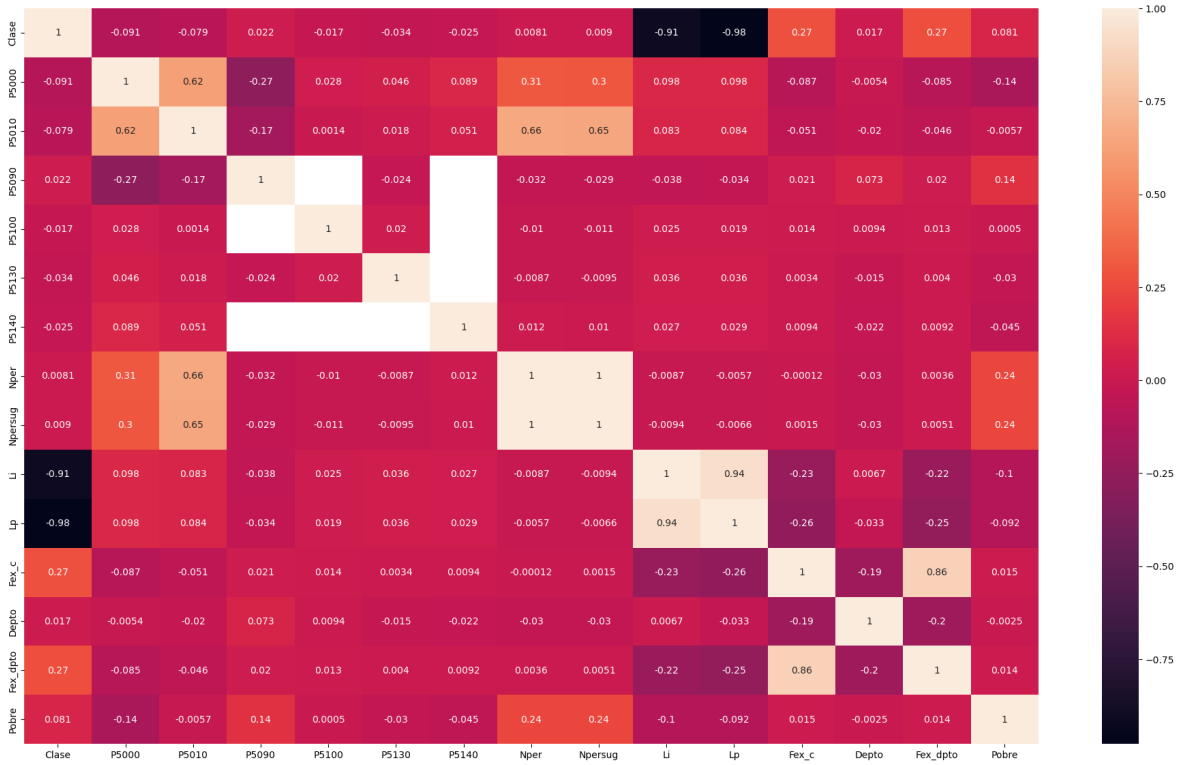


Figure 7: Correlación entre variables: Encuesta hogares

dificultad para identificar efectivamente a los hogares Pobres. Sin embargo, en ambos casos la performance del modelo es muy buena, ya que aproximadamente el 90% de los hogares son clasificados correctamente. Finalmente, en los dos modelos se buscó ajustar manualmente el hiperparámetro de learning rate (o *eta* en XGBoost), intentando con variaciones de múltiplos de 10 a los valores propuestos inicialmente. Los resultados presentados se corresponden con los parámetros que mejor resultado dieron en términos de la métrica F1.

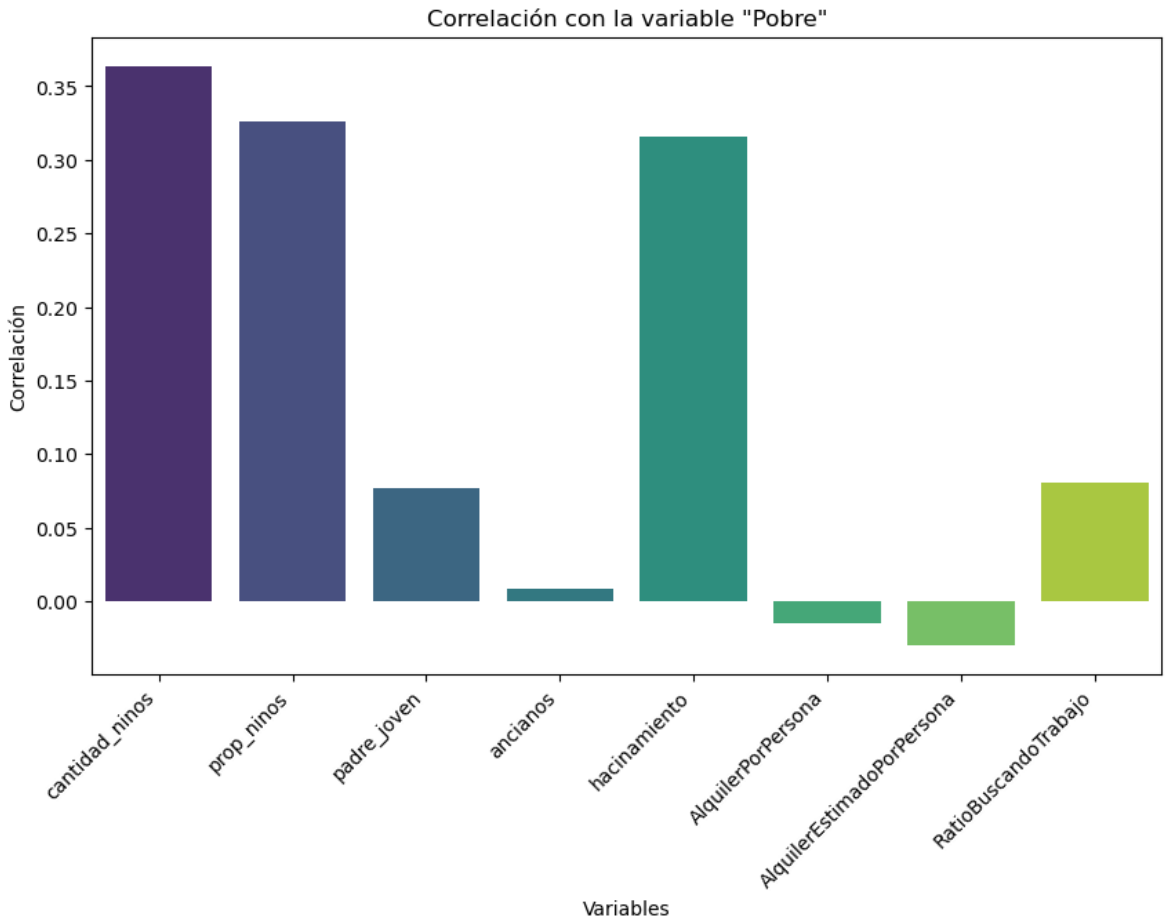


Figure 8: Correlación de variables construidas con la variable de interés

Figure 9: Performance de los diferentes modelos sobre el conjunto de validación (F1)

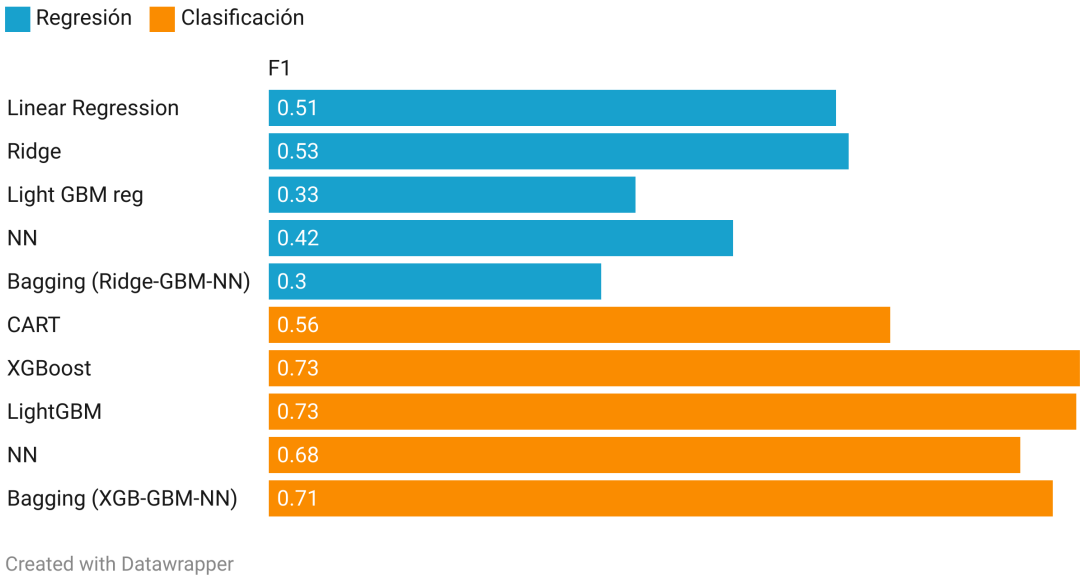
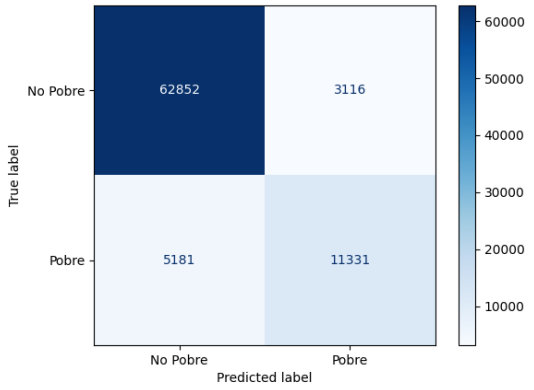
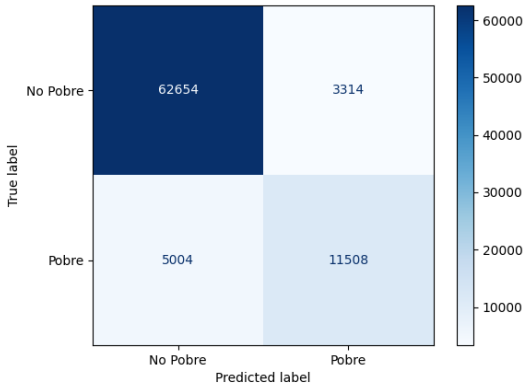


Figure 10: Matriz de confusión para modelos de clasificación seleccionados



(a) LightGBM



(b) XGBoost