Taller 2.

Predicción de la pobreza

Integrantes: Isabella Mendez Pedraza. Cód.: 201814239
Manuela Ojeda Ojeda. Cód.: 201814476
Juan Sebastian Tellez Melo. Cód.: 201513710
Andres Mauricio Palacio Lugo. Cód.: 201618843

Link del repositorio: https://github.com/AndresMPL/Repositorio_PS2.git

1. Introducción.

Llegar a entender la pobreza es un reto que ha llamado la atención de diferentes entidades e investigadores, buscando así incrementar la efectividad de diferentes iniciativas que tienen como objetivo combatir la pobreza, con el fin de orientar de manera óptima las diferentes intervenciones y políticas que enfocadas en su reducción y maximizando el impacto con el costo más bajo posible.

Con la meta de cumplir el Objetivo de Desarrollo Sostenible sobre "Poner fin a la pobreza en todas sus formas en todas partes", se han buscado diferentes métodos para la predicción de la pobreza llegando a utilizar herramientas de Inteligencia Artificial. Anteriormente, los investigadores se limitaban a usar datos oficiales, sin embargo, ahora se pueden utilizar cualquier tipo de datos lo que brinda grandes oportunidades y muchos más desafíos por desarrollar. Llegando a desarrollarse durante el 2016 y marzo de 2022 la publicación de 22 artículos relacionados a la predicción de la pobreza mediante la aplicación de métodos de IA (Usmanova et al., 2022)

En la medición de pobreza no hay una definición única y existen enfoque monetarios y no monetarios. El primer enfoque, considera que las personas son pobres cuando no tienen suficiente dinero para mantener su sustento. Sin embargo, actualmente existe una discusión frente a que la pobreza comprende la falta de oportunidades, educación, atención médica, entre otros. Actualmente, los investigadores concuerdan con que la pobreza es un fenómeno multidimensional que no puede explicarse solo por el dinero (Usmanova, Aziza et al, 2022).

Siguiendo la metodología del Banco Mundial: Pover-T Tests: Predicting Poverty, en este documento se busca desarrollar una metodología para predecir la pobreza en Colombia. Para esto se utilizaron datos a nivel de hogares y personas, provenientes de la Gran Encuesta Integrada de Hogares (GEIH) y del empalme de las Series de Empleo y Pobreza (MESEP).

En primer lugar, se aborda un problema de clasificación para predecir si el hogar es pobre o no pobre dadas sus características observables. Posteriormente, se aborda un problema de predicción de ingresos, en donde se busca predecir los ingresos del hogar para clasificar cada hogar por encima o por debajo de la línea de pobreza.

Para abordar el problema de clasificación se utilizaron modelos Logit, LDA y Árboles de decisión con ajustes de regularización Lasso, Ridge y Elastic Net y se abordó el problema de balance de clases con metodologías de remuestreo. Para el caso de predicción de ingresos se comparó el RMSE de un modelo de Regresión Lineal Simple con regularización con modelos de Arboles, Ramdom Forest y Boosting, haciendo tuning de hiperparámetros para maximizar la capacidad predictiva. Se buscó maximizar la capacidad predictiva de los modelos usando como métrica principal el Accuracy. Sin embargo, no se dejó de lado medidas como Sensitivity y el F1-Score como indicadores de rendimiento de los modelos calculados.

De manera general, se observó que los modelos de clasificación de mejor capacidad predictiva fueron los desarrollados con Logit con regularización Elastic Net y remuestreo Upsampling. Y, en el caso de los modelos de predicción de ingreso, las regresiones lineales con regularización Ridge tuvieron el menor RMSE y generaron la mejor clasificación de las muestras de testeo y evaluación. Las principales variables predictivas fueron el número de cuartos, la relación entre número de ocupados y el núumero de personas en el hogar, y el nivel educativo del jefe del hogar.

2. Datos.

Para este ejercicio utilizamos los datos a nivel de hogar y personas provenientes del DANE, del Empalme de las Series de Empleo, Pobreza y Desigualdad (MESE). Estas bases de datos brindan información sobre características de los hogares y de las condiciones en que se habita la vivienda. Adicionalmente, a nivel de personas, ofrecen información sobre los integrantes de estos hogares.

La base de hogares contiene una variable llamada *Pobre* que identifica los hogares en condiciones de pobreza y cuenta con información sobre el límite de ingresos por debajo del cual un hogar es considerado en pobreza (línea de pobreza). La base de personas contiene la variable *Ingtot* que corresponde al ingreso total por persona que resulta de sumar cada una de las fuentes de ingresos tanto observadas como imputadas.

Para las predicciones a través de clasificación se utilizó información a nivel hogar y como controles se utilizaron características del jefe del hogar, del tipo de vivienda y de los ocupantes del hogar, entre otras, que serán detalladas más adelante.

Creamos variables de interés como *Num_menores_edad*, que es una variable que cuenta la cantidad de menores de 14 años en el hogar y *Num_adulto_mayor* que es una variable que cuenta la cantidad de personas en el hogar mayores de 65 años.

De igual manera, se hicieron variables dummys de aquellas variables que indican si los individuos son desempleados, inactivos, ocupados o si están dentro de la población en edad de trabajar.

A partir de la base de personas se tomó información sobre los jefes de hogar como edad, género, nivel educativo y variables relacionadas al nivel del empleo de los individuos pueden llegar a afectar directamente el ingreso de los hogares o determinar si un hogar es pobre o no pobre y con esto se unieron las dos bases de datos hogares y personas.

Adicionalmente, fue creada la variable *Numper_por_dor* que indica el número de personas por cuartos totales en el hogar. Se creó la variable *Hacinamiento* que toma el valor de 1 si el número de personas por cuartos totales en el hogar es mayor a 3, es decir, que si en promedio más de 3 personas se quedan por cuarto en un hogar esta variable toma el valor de 1 y se creó la variable *Ocupados_por_perhog* en el cual si el número de ocupados en el hogar es mayor a 0 se calcula el número de personas en la unidad de gasto sobre el número de ocupados en el hogar.

Posteriormente, se realizó un tratamiento de los datos con valores no disponibles, manteniendo únicamente las variables que tienen un porcentaje de dichos valores menor o igual a 50%.

Estadísticas descriptivas

Luego de hacer una limpieza de datos, se obtienen las estadísticas descriptivas para las principales variables. En primer lugar, se evidencia en la tabla 1 que el ingreso tiene un mínimo de 0, un máximo de 88,833,333 y una media de 870,639.3, lo que indica que tiene una alta dispersión. Por otro lado, la edad del jefe de hogar tiene un mínimo de 11 años, un máximo de 108 y una media de 49.6 años. El número de cuartos en cada hogar tiene una media de 1.9, un mínimo de 1 y un máximo de 15. Y finalmente, el número de personas en el hogar tiene un máximo de 28, un mínimo de 1 y una media de 3.2.

La figura 1, indica la distribución de la pobreza por hogares. Es decir, indica que cantidad de hogares clasifica como pobre y como no pobre. El 0 indica no pobre y el 1 indica pobre, es decir, hay muchos más hogares no pobres.

Nuestra variable de interés toma el valor de 0 si no es pobre y el 1 si es pobre. El principal problema de esto es que es inicialmente arbitraria y está desbalanceada, lo podemos evidenciar en la figura 2 en donde se tiene que el 80% de los hogares es no pobre y el 20% es

pobre. Cabe resaltar que hay un interés en predecir si una persona es pobre, por lo que se espera reducir este tipo de error.

En la figura 3, podemos ver que el número de personas en el hogar está correlacionada con el ingreso de forma negativa y estadísticamente significativa, esto sucede para hombre y mujeres. Por otro lado, la el número de cuartos y el número de personas se correlacionan de forma positiva como es de esperarse y esto es estadísticamente significativo, de nuevo, ocurre tanto para hombres como para mujeres. Finalmente, el ingreso está correlacionado de manera negativa con el número de cuartos.

La figura 4 muestra la distribución del ingreso por hacinamiento, los ingresos son más altos cuando no hay hacinamiento en el hogar, además, indica que aproximadamente un 95% de los hogares no viven en hacinamiento. La figura 5, hace evidente que entre más alto sea el nivel de educación del jefe de hogar, los ingresos pueden ser a su vez más altos, y la mayoría de jefes de hogar tienen como máximo nivel educativo la primaria básica. La figura 6 indica que los ingresos son un poco más altos cuando el jefe de hogar es hombre y que más del 50% de los jefes de hogar son hombres. La figura 7 nos muestra que el ingreso es más alto cuando la clase del hogar es urbana en vez de rural y aproximadamente un 80% de los hogares son de clase urbana. Y finalmente la figura 8 evidencia que, el nivel de ingresos de los hogares es más alto cuando la vivienda es propia pero no paga, a su vez, indica que la mayoría de los hogares viven en arriendo, seguido por vivienda propia paga.

3. Modelos y resultados.

3.1 Modelos de clasificación

El ejercicio de clasificación de los hogares como "Pobre" (1) y "No Pobre" (0) se realizó a partir de las bases de datos "train_hogares" y "train_personas", dispuestas en la plataforma Kaggle para el desarrollo de este análisis y como parte de la competencia por generar un modelo de clasificación.

El alistamiento de las bases de datos consideró el hecho de que las matrices de datos para el entrenamiento de los modelos contenían diferentes variables respecto a las matrices de datos de la prueba final ("test_hogares.csv" y "test_personas"), por este motivo fue necesario restringir el planteamiento de los modelos a las variables de las matrices de prueba. Adicionalmente, de la base de datos de entrenamiento *personas* se utilizaron variables como la edad, género, situación laboral y nivel de educación del jefe de hogar, se calculó el número de menores de edad y adultos mayores de cada hogar y se estimó el número total de ocupantes en cada hogar, y estas variables fueron implementadas en la base de datos de *hogares*.

El proceso de alistamiento de la información incluyó la conversión de variables como factores, según se requirió, la verificación de variables con valores no disponibles y la generación de dummys; todo esto se encuentra en el script denominado "1_Cleaning", en el repositorio de Git Hub mencionado al inicio de este documento.

El ejercicio buscó la clasificación de los hogares en las clases "Pobre" y "No Pobre", con 1 y 0, respectivamente. Al iniciar el análisis se identificó que la participación de cada clase en la muestra total de datos fue de 80% en el caso de la clase "No Pobre" y 20% la clase "Pobre", lo que evidenció un desbalance moderado de estas clases.

Para el planteamiento de los modelos, teniendo en cuenta este desbalance de clases, se consideraron modelos de tipo Logit, LDA (Linear Discriminant Analysis) y Árboles de decisión; en el caso de los modelos Logit se utilizaron los métodos de regularización de Rigde, Lasso y Elastic Net, y en todos los casos se implementaron métodos para balancear las clases de la muestra, mediante Upsampling, Downsamplig y Oversampling (ROSE). Ahora bien, por capacidad computacional, en las simulaciones se utilizó una grilla de 100 lambdas y en la validación cruzada un k = 5. Así mismo, debe señalarse que los modelos fueron entrenados con la métrica de Accuracy, toda vez que el puntaje en la competencia de Kaggle sería obtenido mediante esta medida, no obstante, en cada caso se calcularon igualmente las medidas de Sensitivity y el F1 Score, con el fin de estimar el rendimiento de los modelos bajo diferentes criterios.

Ahora bien, el enfoque de cada modelo para intentar clasificar los hogares consistió en tener como variable de respuesta la clasificación ya otorgada por el DANE que se encuentra en el archivo de hogares, denominada "Pobre", en la que los hogares pobres toman el valor de 1 y los hogares no pobres toman el valor de 0 y utilizar como predictoras las siguientes variables: número de cuartos de cada hogar y número de cuartos en los que duermen personas, características del jefe del hogar como edad, nivel educativo y situación laboral, características de la vivienda y su ubicación y características de los integrantes del hogar.

Finalmente, cada uno de los modelos se estimó a partir de una base de datos de entrenamiento, que consistió en el 70% de los datos totales, y se evaluó en dos muestras diferentes; una muestra de Test, que correspondía al 20% de los datos iniciales y una muestra de Evaluación que correspondía al 10%, todo esto con el fin de evaluar el rendimiento y la precisión de predicción de cada modelo fuera de muestra, comparando con la variable existente "pobre" en la base de datos de hogares, correspondiente a la clasificación real.

Este proceso, así como los pasos de estandarización, escalado y aleatorización de las submuestras de entrenamiento, test y evaluación, se encuentra descrito en el script denominado "2_Classification" y permitió estimar así un total de 24 modelos, cada uno

evaluado en dos muestras diferentes a las de entrenamiento. Estos modelos y sus medidas de rendimiento en las muestras de Test se encuentran detallados en la Tabla 1.

Así, en la tabla 1 se observa que los modelos con los resultados más altos en la exactitud de la predicción (Accuracy), es decir, el porcentaje de casos que el modelo clasificó correctamente, son aquellos en los que no se realizó ningún tipo de balanceo de la muestra, no obstante, la ausencia de este tratamiento podría generar errores de clasificación al aumentar el número de datos de la muestra de evaluación, toda vez que el desbalance de observaciones hace que el modelo aprenda poco de la clase minoritaria en la muestra de entrenamiento y resulte fácil clasificar la clase mayoritaria y con eso incrementar el acierto del modelo.

Por lo anterior, cada modelo se evaluó balanceando las muestras mediante el incremento de datos de la clase minoritaria (Up-sampling), reduciendo el número de datos de la clase mayoritaria (Down-sampling) o mediante combinación de estas dos técnicas (ROSE - Over-sampling).

3.2 Modelos de regresión

Para el caso de predicción de ingresos se evaluaron modelos de regresión lineal, Ridge, Lasso, Elastic Net y árboles. Se utilizó como variable de interés el logaritmo de la variable *Ingtotugarr* que es el ingreso total de la unidad de gasto con imputación de arriendo a propietarios y usufructuarios. Para elegir el mejor modelo para nuestras predicciones comparamos el RMSE en los diferentes casos. A partir de estos ingresos y por medio de la línea de pobreza se buscó predecir la clasificación de pobreza de los hogares.

Dividimos nuestra base en 70% de la muestra para train, 15% para test y 15% para evaluación y realizamos la estandarización correspondiente.

1. Regresión lineal

$$min_\beta\, E(\beta) = \sum_{i=1}^{n} (Log\ ing\ - \beta_0 - \beta_1 num\ cuartos\ - \beta_2 num\ cuartos\ dormir - \beta_3 num\ personas - \beta_4 ocupados\ por\ hog\ - \beta_5 rural\ - \beta_6 sexo - \beta_7 nivel\ educ\ superior\ - \beta_8 jefe\ de\ hogar\ desocupado\ )^2$$

Estimamos este modelo de regresión lineal a través del método (lm).
Posteriormente corremos otro modelo de regresión lineal incluyendo más controles y algunas interacciones.

$$Log\ ing = \beta_0 + \beta_1 num\ cuartos + \beta_2 num\ cuartos\ dormir + \beta_3 num\ personas + \beta_4 edad + \beta_5 edad^2 + \beta_6 num\ menores\ de\ edad + \beta_7 num\ adulto\ mayor + \beta_8 numper\ por\ dor + \cdots + \beta_i\ rural * edu\ secundaria + \varepsilon_i$$

2. Ridge

$$min_\beta\ E(\beta) = \sum_{i=1}^{n} (Log\ ing - \beta_0 - \beta_1 num\ cuartos - \beta_2 num\ cuartos\ dormir - \beta_3 num\ personas - \beta_4 ocupados\ por\ hog - \beta_5\ rural - \beta_6 sexo - \beta_7\ nivel\ educ\ superior - \beta_8\ jefe\ de\ hogar\ desocupado)^2 + \lambda \sum_{j=1}^{p} (\beta_j)$$

Para estimar el modelo de Ridge usamos el paquete gmlnet con un $\alpha = 0$

3. Lasso

$$min_\beta\ E(\beta) = \sum_{i=1}^{n} (Log\ ing - \beta_0 - \beta_1 num\ cuartos - \beta_2 num\ cuartos\ dormir - \beta_3 num\ personas - \beta_4 ocupados\ por\ hog - \beta_5\ rural - \beta_6 sexo - \beta_7\ nivel\ educ\ superior - \beta_8\ jefe\ de\ hogar\ desocupado)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Para estimar el modelo de Lasso usamos el paquete gmlnet con un $\alpha = 1$

4. Elastic Net

$$min_\beta\ E(\beta) = \sum_{i=1}^{n} (Log\ ing - \beta_0 - \beta_1 num\ cuartos - \beta_2 num\ cuartos\ dormir - \beta_3 num\ personas - \beta_4 ocupados\ por\ hog - \beta_5\ rural - \beta_6 sexo - \beta_7\ nivel\ educ\ superior - \beta_8\ jefe\ de\ hogar\ desocupado\ )^2 + \lambda \left(\alpha \sum_{j=1}^{p} |\beta_j| + \frac{(1-\alpha)}{2} \sum_{j=1}^{p} (\beta_j)^2 \right)$$

Si $\lambda = 1$ es Lasso
Si $\lambda = 0$ es Ridge

3.3 Modelo final

A partir del proceso descrito en los puntos 3.1 y 3.2 del presente documento, los modelos de clasificación seleccionados fueron los que se detallan en la tabla que sigue:

Modelo	Muestreo	Muestra de evaluación	Sensitivity	Accuracy	F1 Score
Logit	—	Test	0.3886	0.8420	0.5022
Logit - Ridge	Oversamplig (ROSE)	Test	0.7049	0.7756	0.5571
Logit - EN	Upsampling	Test	0.7924	0.6337	0.8502
LDA	—	Test	0.3937	0.8400	0.4963
LDA	Downsampling	Test	0.7211	0.7804	0.5679
LDA	Oversamplig (ROSE)	Test	0.7073	0.7813	0.5642

Estos modelos fueron seleccionados teniendo como criterio principal la medida de Accuracy, es decir, el número de casos en los que el modelo acertó en su predicción y fueron entrenados como se describió en el numeral 3.1 con el siguiente planteamiento:

$$Pobre = \beta_0 + \beta_p X_p + \varepsilon$$

La propuesta en este documento es que la clasificación de Pobreza de los hogares se puede realizar a partir de las 16 características o variables de control para cada X_p que se describen en el Anexo 1 del apéndice.

A partir de las variables permitidas en las bases de datos del Problem Set 2, consideramos que estas características que fueron seleccionadas permiten clasificar si un hogar es pobre o no.

Ahora bien, la estrategia de submuestreo, tal como se describió en el numeral 3.1, consistió en dividir la muestra de manera aleatoria en entrenamiento (70%), test (20%) y evaluación (10%), entrenar el modelo en la primera muestra y evaluarlo en las dos siguientes. Adicionalmente, se utilizaron métodos de regularización y técnicas de balanceo de muestra, dado que la clase "Pobre" presentaba una participación apenas del 20% en el total de los datos, lo que haría que el modelo aprendiera poco de las características de esta clase y realizara clasificaciones al aumentar el número de datos evaluados. Finalmente, con las evaluaciones de las predicciones fuera de la muestra de entrenamiento, identificamos los mejores modelos para cargar en Kaggle.

De manera general, se encontró que los resultados evaluados en las muestras de Test mediante Up-sampling tuvieron un promedio de exactitud de clasificación (Accuracy) de 0.6825, mediante Down-sampling un promedio de exactitud de 0.7714 y mediante Over-sampling de 0.7583.

No obstante, si el objetivo es generar un modelo que sea capaz de identificar y predecir correctamente la clase de un hogar, es recomendable fijarse en las medidas de Sensitivity (Recall) y en el F1-Score, este último mide la precisión y la exhaustividad del modelo. En tal sentido, los modelos recomendados serían aquellos que se balanceron incrementando los datos de la clase minoritaria, mediante Logit, Logit con Ridge, Logit con Elastic Net y LDA.

Al finalizar la competencia, se identificó que el modelo con mayor puntaje, de aquellos que cargamos, fue el que se realizó mediante Logit con Elastic Net y balanceando la muestra mediante Up-sampling, logrando un puntaje de 0.8153.

Debe señalarse que al estimar los modelos Logit mediante Ridge, Lasso y Elastic Net, variables como el número de ocupados del hogar y el número de menores de edad se mantuvieron, guardando relación con la probabilidad de que un hogar puedas ser pobre o no.

4. Conclusiones y recomendaciones.

Se concluye que los modelos de clasificación que mostraron mejor rendimiento dentro y fuera de muestra fueron los modelos Logit con regularización Elastic Net y remuestreo Up-Sampling. Mientras que los modelos de regresión con mejor desempeño prediciendo ingresos para posteriormente clasificar los hogares entre pobre y no pobre, fueron las regresiones lineales con regularización Lasso.

Adicionalmente, desde los modelos de Árboles se encontró que la variable más importante para predecir pobreza son el número de cuartos de la vivienda, relacionado con el hacinamiento; la relación entre número de ocupados en el hogar y el número total de personas en la unidad de gasto; el nivel educativo del jefe del hogar, siendo la educación superior la más significativa en diferencia de ingresos. Finalmente, se identificó que la variable de sexo del jefe del hogar no apareció como una variable contundente al momento de predecir pobreza.

5. Bibliografía.

Usmanova, Aziza et al. "Utilities of Artificial Intelligence in Poverty Prediction: A Review." Sustainability (Basel, Switzerland) 14.21 (2022): 14238–. Web.

Usmanova, A., Aziz, A., Rakhmonov, D., Osamy,W. Utilities of Artificial Intelligence in Poverty Prediction: A Review . Sustainability 2022, 14, 14238. https://doi.org/10.3390/su142114238

6. Apéndice de Anexos e Imágenes.

Anexo 1. Variables de control utilizadas para estimar los modelos de clasificación:

- Número de cuartos en el hogar
- Número de cuartos del hogar en los cuales duermen personas
- Número de personas en el hogar
- Edad del jefe del hogar
- Número de menores de edad en el hogar
- Número de adultos mayores en el hogar
- Número de personas por dormitorio
- Número de personas ocupadas en el hogar
- Clase de ubicación de la vivienda (2 factores)
- Tipo de vivienda (6 factores, en el modelo se crearon dummys)
- Tipo de vivienda (7 factores, en el modelo se crearon dummys)
- Género del jefe de hogar (2 factores)
- Nivel educativo del jefe del hogar
- El jefe del hogar está desocupado o no
- El jefe del hogar está inactivo o no
- Se considera hacinamiento en el hogar o no

Tabla 1. Estadísticas descriptivas.

Statistic	N	Mean	St. Dev.	Min	Max
Número de cuartos	164,960	1.989	0.898	1	15
Número de personas	164,960	3.292	1.775	1	28
Ingreso	164,960	870,639.300	1,244,350.000	0.000	88,833,333.000
Edad jefe hogar	164,960	49.612	16.390	11	108
Jefe de hogar hombre	164,960	0.582	0.493	0	1
Urbano	164,960	0.906	0.292	0	1
Vivienda própia y paga	164,960	0.378	0.485	0	1
Vivienda en arriendo	164,960	0.391	0.488	0	1
Educación jefe hogar media	164,960	0.261	0.439	0	1

Figura 1. Distribución de la Clasificación de Pobreza por Hogares.

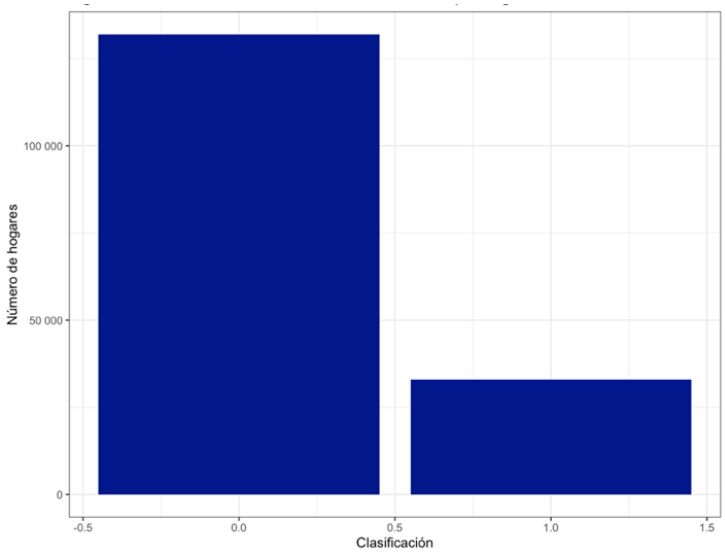

Figura 2. Pie chart con porcentaje clasificación de pobreza. Pobre (1), No Pobre (0)

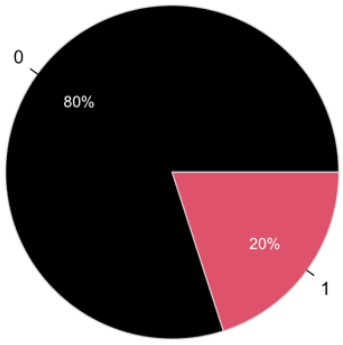

Figura 3. Estadísticas de variables numéricas.

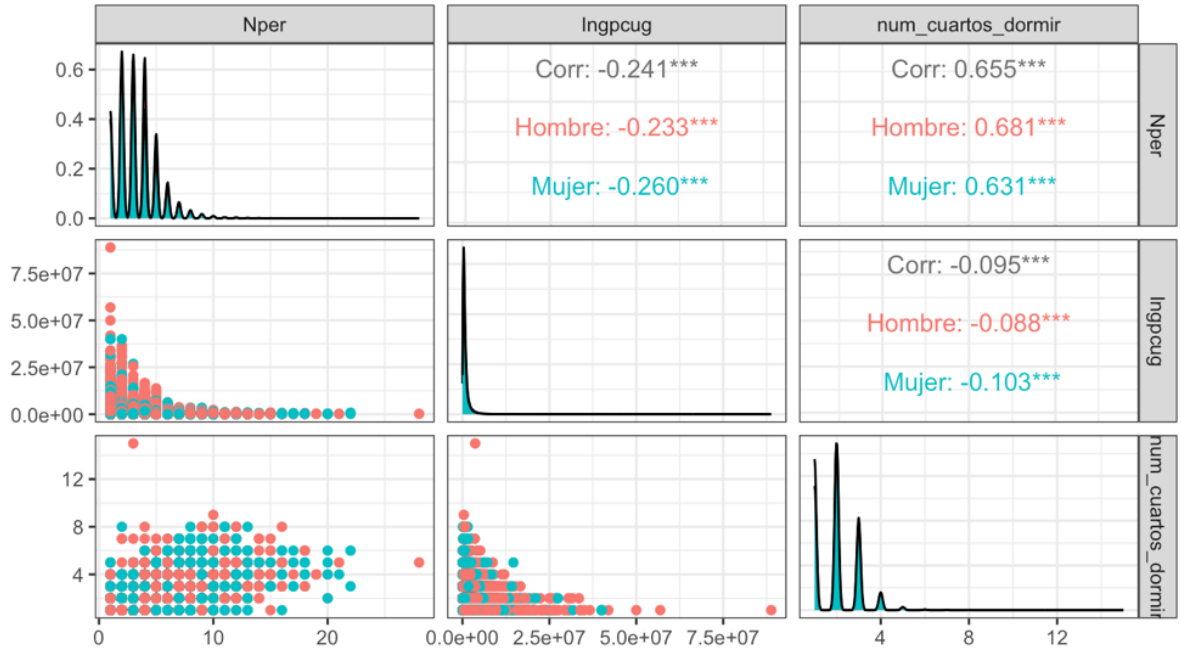

Figura 4. Distribución de ingreso y hacinamiento.

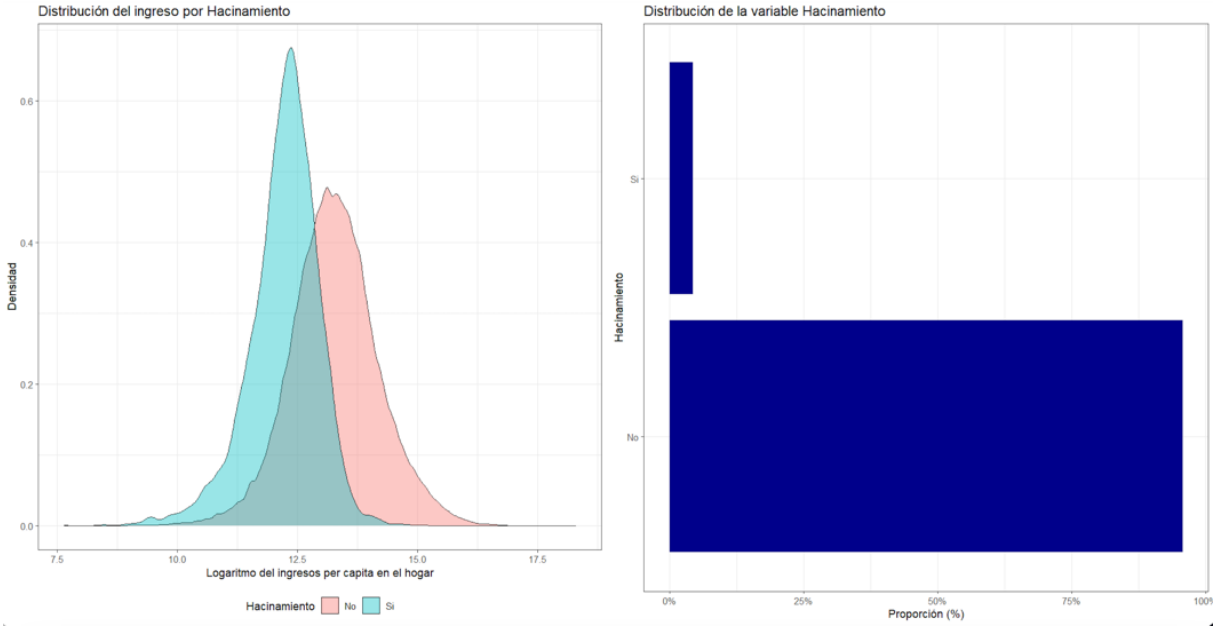

Figura 5. Distribución de ingreso y educación.

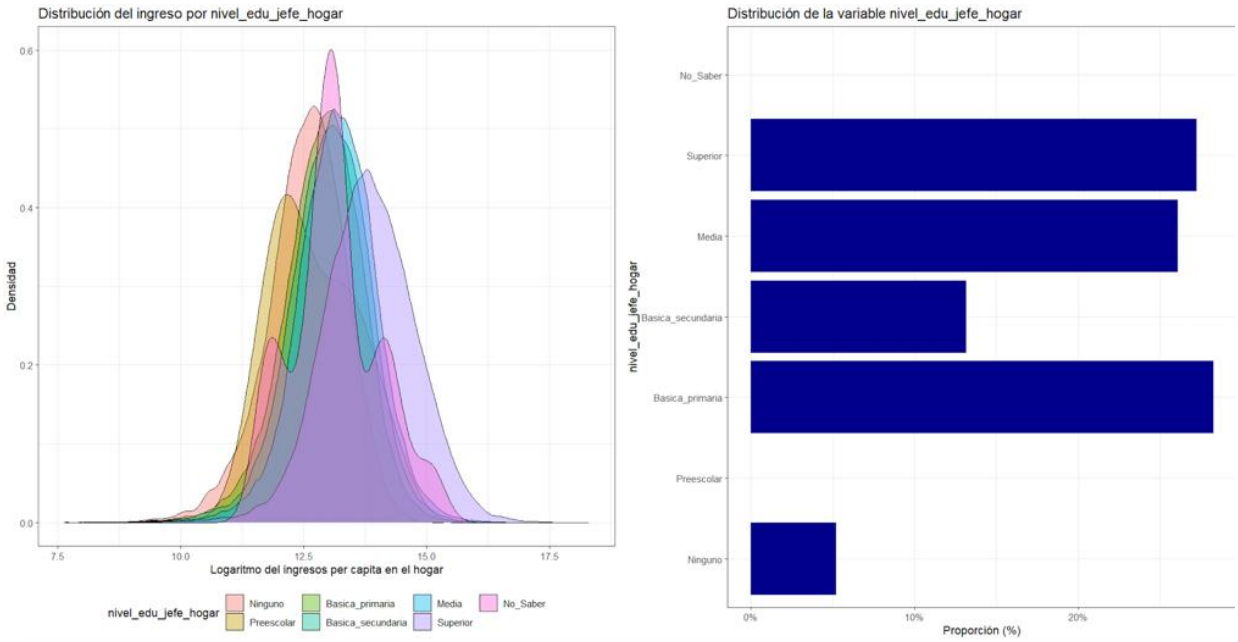

Figura 6. Distribución de ingreso y género del jefe del hogar.

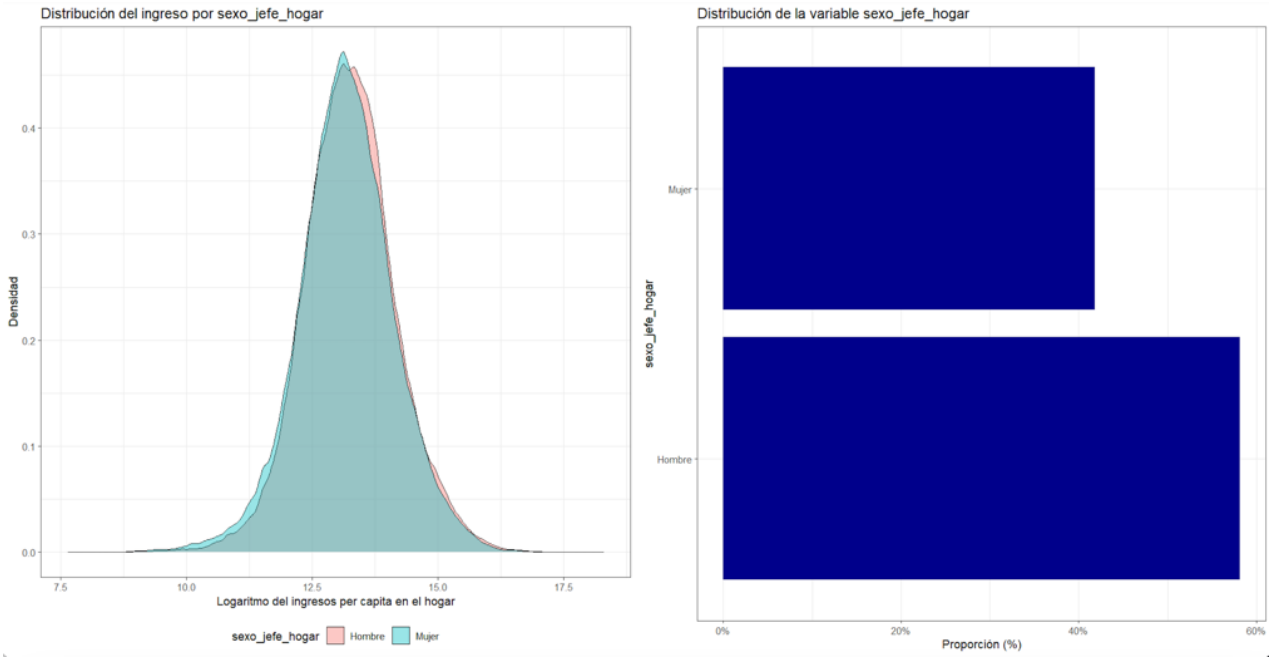

Figura 7. Distribución de ingreso y clase.

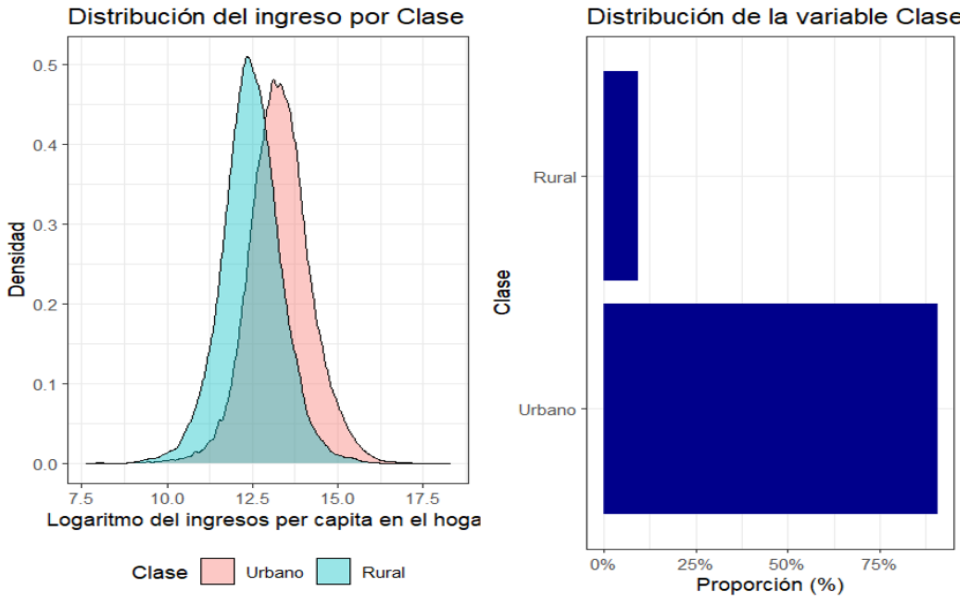

Figura 8. Distribución de ingreso y vivienda.

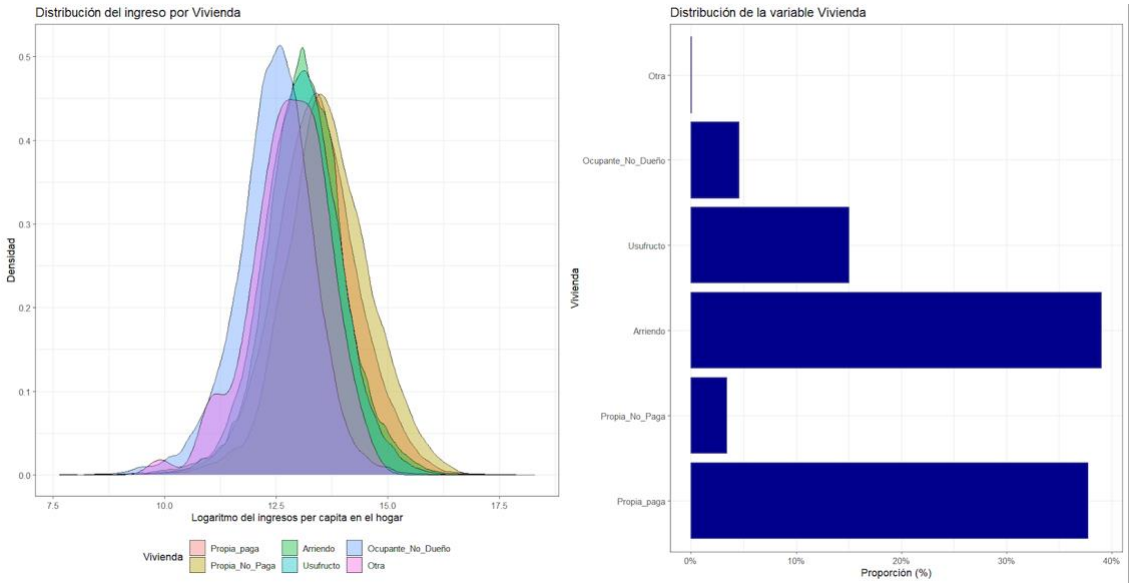

Tabla 1. Comparación del rendimiento de los modelos para clasificación de hogares.

Modelo	Muestreo	Muestra de evaluación	Sensitivity	Accuracy	F1 Score
Logit	—	Test	0.3886	0.8420	0.5022
Logit	Upsampling	Test	0.7835	0.6266	0.8476
Logit	Downsampling	Test	0.7368	0.7741	0.5663
Logit	Oversamplig (ROSE)	Test	0.7206	0.7752	0.5620
Logit - Lasso	—	Test	0.0083	0.8014	0.0165
Logit - Lasso	Upsampling	Test	0.9999	0.7997	0.8895
Logit - Lasso	Downsampling	Test	0.6620	0.7538	0.5184
Logit - Lasso	Oversamplig (ROSE)	Test	0.6283	0.7598	0.5115
Logit - Ridge	—	Test	0.2324	0.8309	0.3549
Logit - Ridge	Upsampling	Test	0.7933	0.6345	0.8506
Logit - Ridge	Downsampling	Test	0.7135	0.7772	0.5619
Logit - Ridge	Oversamplig (ROSE)	Test	0.7049	0.7756	0.5571
Logit - EN	—	Test	0.2324	0.8309	0.3549
Logit - EN	Upsampling	Test	0.7924	0.6337	0.8502
Logit - EN	Downsampling	Test	0.7111	0.7776	0.5614
LDA	—	Test	0.3937	0.8400	0.4963
LDA	Upsampling	Test	0.7943	0.6353	0.8523
LDA	Downsampling	Test	0.7211	0.7804	0.5679
LDA	Oversamplig (ROSE)	Test	0.7073	0.7813	0.5642
Árbol de decisión	—	Test	0.2050	0.8255	0.3199
Árbol de decisión	Upsampling	Test	0.6280	0.7655	0.5174
Árbol de decisión	Downsampling	Test	0.6278	0.7655	0.5173
Árbol de decisión	Oversamplig (ROSE)	Test	0.7000	0.7000	0.4830