Profesor: Ignacio Sarmiento

Camila Beltrán - 202213110
Julian Pulido - 201814833
Alejandra Guevara - 202324570
Juan P. Rodríguez - 201716771

Problem Set 2

Big Data y Machine Learning para Economía Aplicada, 2024-I

Fecha: **14 Abril 2024**.

Universidad de los Andes

Índice

Índice de cuadros

Índice de figuras

1. Introducción

El Banco Mundial tiene como meta acabar la pobreza en 2030, luego de que en 2023 se estimó hay cerca de 700 millones de personas en situación de pobreza extrema. En los últimos años se ha diseñado múltiples enfoques para reducir la pobreza. Por ejemplo, considerando que la mitad de la población pobreza extrema se encuentra en África subsahariana, el organismo internacional ha venido trabajando en los países de esta región, y en su acompañamiento propuso que el enfoque en la prestación de bienes básicos y resolución de conflictos armados tienen un resultado a corto plazo. La educación y el ajuste a la medicina moderna tienen efectos en generaciones posteriores, siempre y cuando este desarrollo esté acompañado de una reducción de la desigualdad.

Ahora bien, para cumplir con esta tarea e impulsar políticas públicas para mitigar los efectos de la pobreza, es fundamental tener mediciones precisas de la misma a lo largo de todos los países, especialmente aquellos que tienen mayores índices de concentración de pobreza. No obstante, medir la pobreza es difícil, requiere mucho tiempo y es costoso. De ahí que el Banco Mundial haya diseñado nuevos mecanismos de recopilación de datos (además de las encuestas a hogares), como datos satelitales y de navegación en internet de los individuos. Este tipo de datos masivos permiten utilizar nuevas herramientas que puedan predecir la pobreza, y de esta manera, observar de manera anticipada qué estrategias son más efectivas contra la reducción de este fenómeno, focalizando de mejor manera las acciones y políticas.

Al hacer una revisión de la literatura sobre predicción de pobreza para países latinoamericanos, se encontró un gran número de artículos con diferentes metodologías y datos, se destaca el estudio de Sosa y Cornejo (2022) para la CEPAL, donde se buscaba predecir la tasa de pobreza agregada para varios países de América Latina a través de un enfoque "micro – macro", que combina información agregada (macro) con datos de encuestas de hogares (micro). Para esto, se utilizaron: "1) microdatos de ingresos para distintos periodos y países, 2) la tasa de variación del ingreso medio, para cada país y período, 3) las tasas de variación del coeficiente de Gini, para cada país y periodo"(Sosa & Cornejo, 2022, p. 8). En la presentación de resultados sobre la evaluación desagregada de las proyecciones de 2019, es importante resaltar que el mejor *score F1* que obtuvieron fue $0{,}878$ para Honduras en la agregación en mediana, los demás países obtuvieron *scores* que oscilaban entre $0{,}3$ y $0{,}7$.

Por otra parte, en el trabajo de Muñetón-Santa y Manrique-Ruiz (2023) se estimó el índice de pobreza multidimensional para Medellín, Colombia, utilizando datos espaciales a nivel de manzana empleando cinco algoritmos de aprendizaje automático, incluidos *XGBoost*, *Lightboost* y *Random Forest*, de acuerdo a los resultados que obtuvieron, el mejor modelo predictivo fue el Modelo Lineal General (GLM), la regresión lineal arrojó un MAE de 0,622, mientras que los siguientes mejores de acuerdo al rendimiento MAE fueron *Random Forest* (0.07504), *XGBoost* (0.07804). Esto implica que la distribución espacial de la estimación de la pobreza multidimensional está altamente correlacionada con los valores reales de la distribución. Por otra parte, el trabajo de Sabogal, García-Bedoya y Granados (2021) "analiza la pobreza en Colombia utilizando herramientas de aprendizaje automático supervisado a partir de los datos de Hogares, Personas y Vivienda del DANE para el periodo 2016 a 2019. Se examina la percepción de factores que influyen en la pobreza teniendo en cuenta las especificidades estructurales que conforman la medición de la pobreza, como la salud, el trabajo y la educación"(p. 2). Se utilizaron 5 diferentes algoritmos, donde destaca el rendimiento de *XGBoost_Classifier* el cual obtuvo un *score F1* de $0{,}96$ en la predicción a nivel de hogares. En este estudio se destaca la importancia de variables predictoras como el desempleo, la salud, el nivel educativo, y los materiales de construcción de las viviendas.

Desde finales de los años 80´s, en Colombia se ha desarrollado distintas medidas de pobreza. Actualmente, mediante una medida indirecta se calcula la pobreza monetaria; medida que utilizaremos para el análisis predictivo en este ejercicio. A pesar de que la tarea de predecir si un individuo es pobre o no se puede ver como un problema de clasificación, también se puede concluir este resultado mediante la predicción del ingreso que tenga un individuo o, en su defecto, el ingreso per cápita de un hogar.

El Dane calcula la pobreza monetaria como la capacidad de un hogar o persona para adquirir una canasta básica de alimentos y bienes básicos. Respecto a esto, se establece dos líneas de pobreza o umbrales. En 2018, la línea de pobreza extrema se encontraba en $121.449 pesos, mientras que la pobreza monetaria estaba en $275.594 pesos. El segundo umbral será la medida de interés para este estudio. Esta medida toma en cuenta alimentos básicos clasificados por la Organización de las Naciones Unidas para la Alimentación y la agricultura (FAO), más otros bienes básicos de subsistencia. Este umbral, por su puesto, no define una medida de bienestar o de clase media en Colombia. Los datos que se utilizaron para la construcción del presente documento provienen del DANE y la misión del "Empalme de las Series de Empleo, Pobreza y Desigualdad - MESE". Los datos contienen cuatro

conjuntos divididos en 'entrenamiento' y 'prueba' a nivel doméstico e individual. Cómo se verá en las secciones 3 y 4, se encontró que el mejor modelo de predicción de una variable binaria es el método de *XGBoosting*, mientras que el modelo lineal fue el más acertado a la hora de predecir el ingreso total del hogar, que posteriormente permitiría la clasificación del hogar de acuerdo con el umbral de pobreza.

Con el objetivo de que los resultados sean replicables, el trabajo cuenta con un repositorio en GitHub. El repositorio cuenta con cuatro carpetas. La primer carpeta, cuenta con los resultados predictivos que se subieron a la competencia Kaggle y las bases de datos con las cuales se realizaron las predicciones. En la segunda, se encuentra el documento final. La tercera carpeta contiene el código de R, dividido en siete documentos scripts; en caso de correr el código, se recomienda usar el *00_main_script*, el cual es la base del código. Finalmente, la cuarta carpeta contiene los gráficos y tablas usadas en este documento.

2. Datos

2.1. Manejo de datos

Los datos utilizados hacen parte de la Misión de Empalme de las Series de Empleo, Pobreza y Desigualdad (MESE) del DANE, en donde se busca medir la pobreza de los hogares a través de sus fuentes de ingreso; además, almacena información a nivel de personas y hogar. La base de entrenamiento de los hogares cuenta con 23 variables y 164.960 observaciones, mientras que la base de personas cuenta con 135 variables y con 543.109 observaciones. Respecto a las bases de datos de testeo, la de hogares tiene 16 variables y 66.168 observaciones, mientras que la de personas 63 variables y 219.644 observaciones. El reto de la predicción se encuentra en que las bases de testeo no cuentan con información sobre el ingreso del hogar.

Considerando que el objetivo de la predicción es determinar si un hogar es pobre o no, se utilizan las variables a nivel hogar y se crean algunas variables por hogar a partir de la base de personas. Lo anterior es posible dado que las dos bases contienen la variable "id", que identifica el hogar al cual pertenece cada persona. En general, las variables seleccionadas buscan describir patrones que inciden en la pobreza en el hogar, tales como: número de niños, niveles de educación, habitaciones por persona, ubicación, entre otros. Además, se busca caracterizar la fuente de ingresos del hogar asociada con el jefe de hogar, a través de su ocupación, oficio y mecanismos que indican si tiene un trabajo informal.

La tabla 1 presenta las variables utilizadas para la construcción de los modelos de predicción. Para la base de datos de personas, se construyen variables a nivel hogar basadas en características de las personas que conforman el hogar y del jefe de hogar.

Tabla 1: Descripción de variables

Fuente de datos	**Variable**	**Descripción**
Hogares	$Nper$	Número de personas.
	$Cabecera$	Dummy si se encuentra en área urbana.
	$Dpto$	Departamento donde está ubicado el hogar.
	Fex_dpto	Factor de expansión del departamento.
	$Dominio$	Clasifica ciudades principales y agrupa las demás en 'resto urbano' y 'resto rural'.
	$Ncuartos$	Número de cuartos.
	$Ocup\_vivienda$	Ocupación la vivienda (propia, en arriendo, entre otros).
	$Ln(Cuota)$	Logaritmo natural de la cuota de amortización (hogares que están pagando la vivienda).
	$Ln(Arriendo)$	Logaritmo natural del valor del arriendo de la vivienda.
	$PerXCuarto$	Número de cuartos por persona (no. personas / no. cuartos).
	$Nmujeres$	Número de mujeres.
	$Menores\_5$	Menores de 5 años.

Continúa en la siguiente página

Tabla 1 – Continuación de la tabla 1

Fuente de datos	Variable	Descripción
	$Menores\_6\_11$	Menores entre 6 y 11 años.
	$Menores\_12\_17$	Menores entre 12 y 17 años.
	$Nmenores\_ocup$	Número de menores de 18 años que se encuentran ocupados.
	$P5010$	Número de cuartos en que duermen las personas del hogar
	Max_educ	Máximo nivel de educación.
	$Ninc\_trab$	Número de personas que estaban incapacitadas para trabajar la semana anterior a la encuesta.
	$Jefe\_mujer$	Dummy si el jefe de hogar es mujer.
	$Jefe\_ocupacion$	Ocupación del jefe de hogar.
	$Jefe\_oficio$	Oficio del jefe de hogar.
	$Jefe\_ocup$	Dummy si el jefe de hogar se encuentra ocupado.
	$Jefe\_educ$	Nivel de educación del jefe de hogar.
	$Jefe\_SS$	Dummy si el jefe de hogar se encuentra afiliado al sistema de seguridad social en salud.
	$Jefe\_RSS$	Dummy si el jefe de hogar pertenece al régimen subsidiado en salud.
	$Jefe\_rec\_viv$	Dummy si el jefe de hogar recibe vivienda como parte de su salario.
	$Jefe\_rec\_ali$	Dummy si el jefe de hogar recibe alimentación como parte de su salario.
	$Jefe\_ult\_exp$	Tiempo de experiencia del último trabajo del jefe de hogar.
	$Jefe\_Ntrab$	Número de trabajadores de la empresa en la que trabaja el jefe de hogar.
	$Jefe\_Strab$	Dummy si el jefe de hogar tiene un segundo trabajo.
	$Jefe\_primas$	Dummy si el jefe de hogar recibe primas.
	$Jefe\_bonif$	Dummy si el jefe de hogar recibe bonificaciones.
	$Jefe\_subsidio$	Dummy si el jefe de hogar recibe subsidios.
	$Jefe\_pension$	Dummy si el jefe de hogar cotiza pensión.

Una vez las variables de la base de personas se agregan por hogar, se adicionan estas variables a la base de hogares de entrenamiento y testeo. Para el entrenamiento, se utiliza la variable "$Pobre$", que asigna 1 a los hogares pobres y 0 en caso contrario, para los modelos de clasificación. En cuanto a los modelos de predicción del ingreso y clasificación de hogares pobres según la línea de pobreza, se utilizaron las variables: "$Ingtotug$"(Ingreso total de la unidad de gasto antes de imputación de arriendo a propietarios y usufructuarios), "$Ingtotugarr$"(Ingreso de la unidad de gasto con imputación de arriendo a propietarios y usufructuarios), "$Ingpcug$"(Ingreso per cápita de la unidad de gasto con imputación de arriendo a propietarios y usufructuarios). En múltiples modelos de predicción indirecta se utilizó como variable de ingreso el logaritmo de "$Ingtotug$", es decir, "$Ln(IngHogar)$" y el logaritmo de "$Ingtotugarr$", es decir, "$Ln(Ing\_Hogar\_Imp\_Arriendo)$".

2.2. Imputación de datos

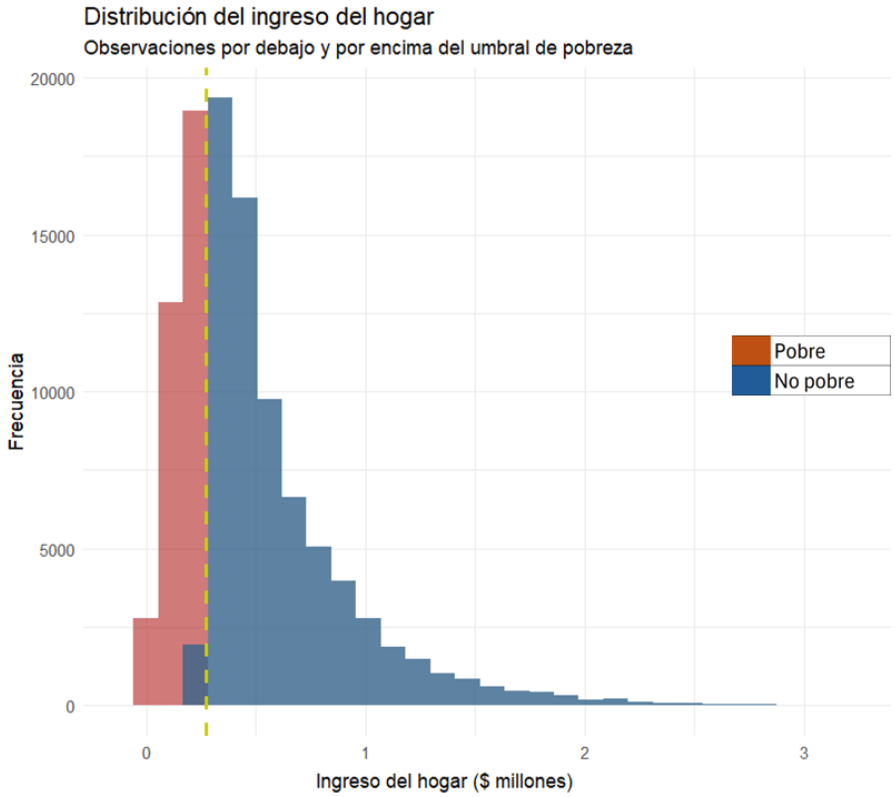
El valor del arriendo del hogar es una variable importante en la predicción de pobreza dado que limita el ingreso disponible del hogar y, además, puede utilizarse como una aproximación de estrato o tipo de vivienda. Sin embargo, este valor solo está disponible para los hogares que viven en arriendo o subarriendo, quienes representan el 39 % y el 38 % de los hogares del entrenamiento y de prueba, respectivamente. Para imputar los datos faltantes de los otros tipos de ocupación de vivienda (propia, usufructo, entre otros) se aplica la misma metodología del DANE, utilizando la pregunta de la encuesta: si tuviera que pagar arriendo por esta vivienda, ¿cuánto estima que tendría que pagar mensualmente? Esto da como resultado que la variable arriendo no tenga datos faltantes para ninguna de las dos bases.

Por otro lado, se utilizan indicadoras de formalidad o tipo de trabajo para el jefe de hogar, lo cual limita el análisis a los jefes de hogar que se encuentran ocupados. Para las variables relacionadas al trabajo (ocupación, oficio, tamaño de la empresa, entre otros), se reemplazan los valores faltantes con 0. Para las variables numéricas y dicótomas asigna el menor valor (0), para las categóricas como oficio u ocupación crea un nuevo nivel para las personas desocupadas.

2.3. Estadísticas descriptivas

Dentro del gran conjunto de variables, la más relevante para el presente estudio resulta ser el ingreso total del hogar. Luego de comparar el ingreso con el umbral de pobreza definido por el DANE (de acuerdo a su ubicación geográfica) que define un mínimo de subsistencia, se puede clasificar a un hogar como pobre o no pobre.

Figura 1: Histograma del ingreso de los hogares clasificados por pobreza



En nuestra muestra, agregada por hogares, se puede observar la distribución del ingreso acotada hasta dos millones de pesos; esto con el fin de eliminar parte de la cola derecha de la distribución que representa los hogares con mayores ingresos, pero que son una minoría. De esta manera, se puede visualizar de forma más clara la partición en las distribuciones entre los hogares clasificados como pobres.

No obstante, para no perder la vista general del ingreso del hogar se observa que la distribución completa tiene un sesgo positivo (o a la derecha) debido a que la media excede en gran medida a la mediana. Esto ocurre porque hay *outliers* de grandes ingresos. Lo anterior se puede inferir ya que el percentil 75 se ubica en $ 2.5 millones, mientras que el máximo se ubica en $85 millones de pesos.

La frecuencia se representa en el eje de las ordenadas de la gráfica 1. Para el caso de los hogares clasificados como pobres se representan en color rojo –el 20 % de la muestra–, mientras que los hogares no pobres se pueden visualizar en la parte azul –80 % de los hogares del histograma–. La línea puntada de color amarillo muestra el promedio nacional del umbral de pobreza monetaria.

Al observar la distribución del ingreso y, por tanto, la distribución de pobreza se observa un desbalance. Por esta razón, la partición de datos entre entrenamiento y testeo se hará manteniendo la misma proporción de pobreza. En el caso hipotético de que una partición aleatoria no conserve las mismas proporciones, el modelo podría generar un resultado distinto al esperado en los datos de testeo.

3. Modelos de clasificación

3.1. Resumen

En esta sección se presentan los modelos utilizados en la competencia que predicen la pobreza del hogar a partir de modelos de clasificación utilizando como variable dependiente si un hogar es pobre (1) o no (0) para el entrenamiento. Para la construcción de los modelos se utilizaron las siguientes variables:

Grupo 1: Todas las variables detalladas en la sección de datos. Además, se crean dummys para variables categóricas de $Dominio$,$Dpto$,$Ocup\_vivienda$,Max_educ,$Jefe\_educ$,$Jefe\_oficio$,$Jefe\_ocup$. Para poder hacer la predicción, se dejan únicamente las categorías que se comparten en las bases de prueba y testeo. En total, suman 187 predictores. La totalidad de variables predictoras en este grupo no necesariamente coincide con las usadas en los modelos de predicción indirectos; aunque sí se comparte variables clave que describen la vivienda y las características de la jefe del hogar.

Dado que la base de datos se encuentra desbalanceada, con una participación del 20 % de los hogares pobres, se realizan técnicas de remuestreo híbrido como *Synthetic Minority Over-sampling Technique (SMOTE)* para rebalancear la muestra de entrenamiento y mejorar la predicción, lo cual permite que la proporción de pobres aumente a 43 %. Por otra parte, se consideraron distintos métodos y subgrupos de variables, sin embargo, los algoritmos que utilizan *SMOTE* y todo el conjunto de variables son los que obtuvieron mejores resultados de predicción.

La tabla 2 presenta un resumen de los modelos con mayor puntaje de F1 que otorga la competencia de Kaggle:

Tabla 2: Modelos de clasificación

No. Modelo	Algoritmo	Variables	Remuestreo	Puntaje Kaggle
1	Linear Discriminant Analysis	Grupo 1	SMOTE	0,66
2	Elastic Net			0,66
3	Logit			0,66
4	Random Forest			0,63
5	XGBoosting			0,67

3.2. Modelos lineales

En la primera predicción se usó *Linear Discriminant Analysis (LDA)*. La motivación para usar este método es que un modelo Logit puede sufrir cuando las clases son asimétricas; en este caso la variable *Pobre* tiene una menor proporción, pero a su vez la variable base (Ingreso del hogar) también tiene una distribución distinta cuando se divide por umbral de pobreza. Por lo tanto, separar las clases por medio de LDA conlleva a un modelo más robusto. El segundo modelo es una predicción lineal utilizando Elastic Net, este algoritmo permite ajustar los coeficientes de las variables para evitar el sobreajuste y, en algunos casos, lleva a cero los coeficientes de las variables que generan mucha varianza. Para este algoritmo se utilizó *cross-validation* de 5 *folds* para identificar los hiper-parámetros que optimizaban el *Accuracy* (α =0,2 y $\lambda = 0{,}01$). Como resultado, los dos modelos de enfoque lineal para predecir la clasificación de hogares pobres fueron exitosos en el 66 % de los casos de la competencia.

3.3. Modelos no lineales

Los modelos de *Random Forest* permiten escoger de manera aleatoria una muestra de predictores, lo que permite que las estimaciones sean más robustas al reducir la correlación entre los árboles. Para este algoritmo se realizó *cross-validation* de 5 *folds* y el mejor resultado de los hiper - parámetros fue un número mínimo de observaciones en el nodo terminal de 50 observaciones, 30 variables que se escogen de forma aleatoria y se utilizó *gini* como regla de división. Por otro lado, se implementó un modelo de probabilidad *Logit*, este enfoque mejora algunas limitaciones de los modelos lineales, en particular asigna probabilidades entre 0 y 1 de pertenecer a una clase (pobre y no pobre), asumiendo que la distribución de la variable independiente sigue una función logística. Su

estimación se realiza a partir de máxima verosimilitud, lo cual permite que sea un modelo más flexible. Como resultado, se obtuvo un puntaje de $0{,}66$ a través de *cross-validation* con 5 *folds*.

El *Boosting* es una técnica de aprendizaje automático en la que se construyen una serie de modelos de predicción débiles, generalmente árboles de decisión simples, de manera iterativa. En cada iteración, se da más peso a los casos que fueron mal clasificados en las iteraciones anteriores, de modo que los modelos subsiguientes se centren en corregir los errores cometidos por los modelos anteriores. Para el modelo de clasificación para predecir la pobreza se utilizó el algoritmo *XGBoost*, el cual se basa en el principio de *Boosting*, sin embargo, se destaca por características clave como el uso de un enfoque de optimización de gradiente, lo que significa que ajusta los modelos de manera incremental buscando minimizar una función de pérdida mediante la optimización de los gradientes de esta función. Esto permite que los modelos se ajusten de manera más precisa a los datos y mejoren gradualmente su rendimiento. Además, implementa técnicas de regularización para controlar la complejidad del modelo y prevenir el sobreajuste, lo que también permite una mayor eficiencia en el manejo de conjuntos de datos grandes y complejos.

Este último algoritmo reflejó el mejor puntaje con un F1 de $0{,}67$ en la competencia, a través de *cross-validation* de 5 *folds* se obtuvo que el mejor modelo utilizó 500 modelos para entrenar el algoritmo. Los hiper-parámetros que escogió el algoritmo fueron: una profundidad máxima de cada árbol de 4, un número de observaciones en los nodos terminales de mínimo 50, una tasa de aprendizaje *eta* de $0{,}25$, una pérdida mínima de división *gamma* de 0, el 33 % de variables predictoras escogidas de manera aleatoria para cada árbol (*colsample*) y el 40 % como proporción de muestras utilizadas para entrenar cada árbol en el proceso de *boosting* (*subsample*).

4. Modelos de clasificación indirecta

En la anterior sección se proyectó mediante un método de clasificación si un hogar era pobre o no. Tal como se vio en la sección 2.3, el ingreso es la base de esta clasificación. En este apartado se busca predecir la pobreza mediante un método indirecto en el cual la variable dependiente será continua y representará el ingreso total del hogar en su transformación logarítmica.

4.1. Resumen

La tabla 3 muestra la comparación entre los tres modelos que predijeron el ingreso del hogar para luego clasificarlo como pobre o no pobre. De acuerdo con la clasificación de Kaggle, el modelo lineal obtuvo el mejor puntaje.

Tabla 3: Comparación entre los modelos de clasificación indirecta

No. Modelo	Algoritmo	Puntaje Kaggle
1	Lineal	0.63
2	Random Forest	0.20
3	XGBoosting 1	0.41
4	XGBoosting 2	0.61

4.2. Modelo líneal

El primer modelo resulta ser un modelo simple que cuenta con una mezcla de 5 predictores que reúne características sobre la vivienda y detalles del mercado laboral al cual pertenece la jefa del hogar. Estas variables son: el número de personas por cuarto del hogar, el número de ocupados, un dummy sobre si la jefe del hogar es mujer, si se cotiza a pensiones y si se encuentra afiliado al régimen subsidiado de salud.

$$\begin{aligned} Ln(IngHogar) = \alpha + \beta_1 PerXCuarto + \beta_2 Nocupados + \beta_3 HeadMujer \\ + \beta_4 HeadCotpension + \beta_5 HeadRecsubsidio + \epsilon \end{aligned}$$

Al ser un modelo simple, para hacer la revisión de variables se usó el método *Best Subset Selection*. El algoritmo concluye que se deberían usar las cinco variables, en vista de que arroja el menor indicador de error sobre la

media de coeficiente de variación ($0{,}577$) sobre la submuestra de entrenamiento. Además, se aplicó el método de regulación *Elastic Net* y se concluye que los mejores parámetros son: $\alpha = 0{,}05$ y $\lambda = 0{,}1$. Los coeficientes de cada variable se presentan en la tabla 5, en el apéndice A. El modelo (1) se estima por MCO, mientras que al segundo se le aplica el método de regulación de manera que los coeficientes son menores, en especial la variable dummy que define el género del jefe de hogar.

Finalmente se configuró un modelo saturado con 144 variables (gran parte de ellas son categóricas, lo que resulta en la generación de nuevos coeficientes para $k-1$ categorías). Este modelo se realizo con una regularización por medio del método de *Elastic Net* y obtuvo el mayor el puntaje *F1*. Lo anterior se debe a que los modelos lineales parecen comportarse de manera más efectiva al predecir sobre variables continuas.

4.3. Modelos no lineales

4.3.1. *XGBoost*

El segundo modelo con mejor poder predictivo fue en el cual se utilizó el algoritmo de *XGBoost*, el cual combina características de los *Random Forest*, pues lo que busca es construir los arboles secuencialmente, aprendiendo poco a poco de los errores anteriores. Este modelo es una implementación específica y muy eficiente de la técnica de *boosting*, dado que utiliza un algoritmo de árbol de decisión optimizado y paralelizado, lo que lo hace extremadamente rápido y escalable. Además, *XGBoost* ofrece una serie de características adicionales, como regularización, manejo de valores faltantes y funciones de importancia de características, que lo hacen aún más potente y versátil.

Debido a que el objetivo era hacer una predicción indirecta de la pobreza a través de la predicción del ingreso, el modelo utilizó como variable dependiente el logaritmo del ingreso total del hogar con imputación de arriendo (*Ln(Ingtotugarr)*), como variables predictoras se tomaron 28 variables detalladas en la sección de datos, incluyendo algunas de estas variables elevadas al cuadrado:

$$\begin{aligned} Ln(Ing\_Hogar\_Imp\_Arriendo) = \alpha &+ \beta_1 Dominio + \beta_2 Depto + \beta_3 P5010 + \beta_4 P5010^2 \\ &+ \beta_5 Npersug + \beta_6 Npersug^2 + \beta_7 Nmujeres^2 \\ &+ \beta_8 PerXCuarto^2 + \cdots + \beta_{28} Jefe\_Pension + \epsilon \end{aligned}$$

Además de las variables mostradas en la ecuación anterior, específicamente se utilizaron las siguientes variables: $Ncuartos$, $Nper$, $Menores\_5$, $Menores\_6\_11$, $Menores\_12\_17$, $Nocupados$, $Nincapacitados$, $Nmenores\_ocup$, $Jefe\_mujer$, $Jefe\_SS$, $Jefe\_subsidio$, $Jefe\_rec\_viv$, Max_educ, $Jefe\_Strab$, $Ln(Cuota)$, $Jefe\_oficio$, $Ln(Arriendo)$, $Ocup\_vivienda$, $Cabecera$.

Una vez se tenía la predicción del logaritmo del ingreso con imputación de arriendo, se comparó con la línea de pobreza multiplicada por el número de personas por unidad de gasto para determinar los hogares marcados como "$Pobre$", con este modelo se obtuvo un puntaje de $0{,}41$ en Kaggle. Adicionalmente, se realizó un segundo modelo de *XGBoosting* que utilizó las mismas métricas de variables e hiper-parámetros del modelo no. 5 de la tabla 2, con la distinción que no se crean dummys para variables categóricas sino que se mantienen en factores, lo cuál ayudó a mejorar la predicción y acertar en un 61 % de las veces, siendo este algoritmo el segundo mejor resultado de medición por ingresos.

4.3.2. Random Forest

Por último, el modelo que obtuvo menor poder predictivo frente a la muestra de prueba en este análisis fue el realizado con *Random Forest*. La variable objetivo fue el *Ingreso total de la unidad de gasto con arriendo imputado a propietarios y usufructuarios* y las variables predictoras utilizadas fueron el número de cuartos por persona, si el jefe de hogar es mujer, clasificación entre ciudades principales, resto urbano y resto rural, ocupación del jefe de hogar, nivel de educación del jefe de hogar, si el jefe de hogar se encuentra ocupado, si el jefe de hogar recibe subsidios, oficio del jefe de hogar, si el jefe de hogar tiene un segundo trabajo.

$$
\begin{aligned}
Ingtotugarr = \beta_0 &+ \beta_1 PerXCuarto + \beta_2 Jefe\_Mujer + \beta_3 Dominio \\
&+ \beta_4 Jefe\_Ocupacion + \beta_5 Jefe\_Educ + \beta_6 Jefe\_Ocup \\
&+ \beta_7 Jefe\_Subsidio + \beta_8 Jefe\_Oficio + \beta_9 Jefe\_Strab
\end{aligned}
$$

En el proceso de optimización del modelo mediante *Random Forest*, se realizó una búsqueda exhaustiva de hiper-parámetros. Esto incluyó la exploración de diferentes valores que determinen el número de variables predictoras consideradas en cada división del árbol, con opciones de 1, 2 y 3. Asimismo, se evaluaron distintos criterios de división de nodos, especificados con opciones como “variance”, “extratrees” y “gini”. Además, se ajustó el parámetro que establece el tamaño mínimo de un nodo terminal en el árbol, probando valores de 1, 3 y 5. El objetivo fue encontrar la combinación óptima de estos hiper-parámetros para maximizar el rendimiento del modelo, medido en términos de la métrica F1 en el conjunto de datos de prueba. Este proceso permitió que se ajustara mejor a las características específicas de los datos y mejorara su capacidad predictiva. Como resultado, se obtuvo un valor de F1 de $0{,}2$ en el conjunto de datos de prueba. Este resultado sugiere que el modelo tiene una capacidad limitada para predecir con precisión la cantidad total de ingresos, lo que indica la necesidad de explorar otras técnicas de modelado o recopilar más datos para mejorar el rendimiento del modelo.

5. Modelos finales

Esta sección detalla la construcción de los modelos que generaron mejores resultados de predicción a partir del puntaje F1 de la competencia de *Kaggle*. Adicionalmente, se analizan sus ventajas con respecto a los otros modelos evaluados y la interpretación de los predictores utilizados.

5.1. Remuestreo

Entre los retos de la predicción de la pobreza se encuentra que una clase minoritaria de la población se clasifica como pobre, según la muestra de entrenamiento alrededor del 20 % de los hogares eran pobres. Este desequilibrio entre clases puede hacer que las estimaciones estén sesgadas hacia los hogares que no son pobres, dado que representan la mayoría de observaciones. Para nuestro ejercicio se utilizó la técnica de remuestreo de SMOTE, la cual permite crear muestras sintéticas que son combinaciones lineales de la clase que está sub-representada, en este caso los hogares pobres. Adicionalmente, esta técnica resulta útil para evitar un sobreajuste que generaría al aplicar otras técnicas que repliquen observaciones, en lugar de generar muestras sintéticas.

El modelo final de clasificación implementa SMOTE para su entrenamiento, el cual se explicará en más detalle en la siguiente subsección. Por otro lado, los modelos de clasificación indirecta, es decir, los que predicen ingresos, no incluyen está técnica dado que la medición es distinta y se cambia el enfoque a una predicción que refleja la distribución de ingresos de los hogares.

Las variables seleccionadas en estas predicciones abarcan diferentes aspectos económicos y sociales que influyen en la situación de pobreza de los hogares. En primer lugar, consideramos características del hogar, como el tamaño y la composición familiar, así como la infraestructura y condiciones de vida, como indicadores relevantes de la capacidad económica de los hogares. Además, las características individuales, como la demografía y el nivel educativo, así como la participación en el mercado laboral y el acceso a la seguridad social y beneficios laborales, son factores importantes a tener en cuenta. También se consideran aspectos financieros, como los gastos relacionados con la vivienda, y la ocupación y experiencia laboral del jefe de hogar. Estas variables fueron seleccionadas cuidadosamente por su potencial para ofrecer información relevante sobre la situación económica de los hogares y su relación con la pobreza.

5.2. Modelos

El modelo de clasificación que tuvo el mejor resultado fue un modelo de *Boosting*, en específico, utilizando el enfoque de *XGBoosting* (modelo no. 5 de la tabla 2). Para su construcción se consideraron resultados de modelos

más simples y que arrojaban buenos resultados de predicción. En ese sentido, se utilizaron todas las variables del "Grupo 1" que se especificaron en la sección de modelos, y se implementaron algunas grillas para obtener el mejor resultado:

Tabla 4: Parámetros del modelo

Parámetros	**Valor**
Número de árboles	500
Número mínimo de observaciones en nodos terminales	50
Profundidad del árbol	4
Tasa de aprendizaje (*eta*)	0,01, 0,25, 0,5
Pérdida mínima de división (*gamma*)	0
Proporción de variables	0,33 y 0,66
Proporción de la muestra	0,4

Los valores fijos como el número de árboles y las observaciones de los nodos terminales se determinaron a partir de los resultados de algunos modelos como *Random Forest*, para reducir el número de iteraciones y la carga computacional. Como resultado los valores que escoge el algoritmo es una tasa de aprendizaje intermedia (0,25) y la proporción de las variables más baja (33 %). Esta muestra tiene 187 predictores, lo que implica que los árboles pueden llegar a ser muy distintos, el árbol con mayor importancia del modelo final que pondera los 500 árboles es el árbol número 1 (la secuencia inicia en 0), el cual utiliza como variable inicial la dummy que se habilita cuando el máximo nivel de educación en el hogar es universitario. En el apéndice B se encuentra el diagrama de los árboles 1 y 499 que refleja la variabilidad entre ellos.

Por el contrario, en los modelos de predicción de la pobreza de manera indirecta, los modelos no lineales como las diferentes variaciones de los árboles no son los más acertados. Como se observa en la tabla 3 uno de en algunos modelos de árboles como *XGBoosting* y *Random Forest* predice peor que la distribución aleatoria binomial. Esto contrasta con los resultados encontrados en los modelos de clasificación. Para el caso de la predicción de la pobreza con la muestra nacional se encuentra que el modelo lineal saturado (144 variables predictoras) es el mejor modelo, estableciendo un puntaje *F1* de $0{,}63$ ($0{,}03$ puntos menos que el mejor modelo de clasificación).

Cabe resaltar que en el entrenamiento de este modelo se hizo con una transformación logarítmica y, por lo tanto, la predicción se le aplicó un exponente natural (i.e. e^x). Además, fue importante aplicar un método de regularización, disminuyendo el impacto de los distintos coeficientes.

Al comparar el *performance* predictivo del modelo de clasificación que empleó el algoritmo *XGBoost* con el resto de modelos de la Tabla 2 encontramos que solo aumentó la precisión en un $0{,}01$ respecto a los modelos con algoritmos de *Linear Discriminant Analysis*, *Elastic Net* y *Logit*. Sim embargo, sí existe una diferencia signiticativa respecto al modelo 4 que empleó el algoritmo de *Random Forest*. Por otra parte, al comparar el *performance* predictivo del modelo 1 de predicción indirecta de pobreza que empleó la regresión lineal con los modelos de la Tabla 3, encontramos que es mucho mejor que la mayoría de otros algoritmos, el más cercano es el de *XGBoost*, con una diferencia solo de $0{,}02$.

6. Conclusiones

En el presente trabajo se buscó, como objetivo principal, la construcción de un modelo preciso predictivo de la pobreza de los hogares en Colombia a partir de los datos de la Misión de Empalme de las Series de Empleo, Pobreza y Desigualdad (MESE) del DANE, en la búsqueda de este modelo se emplearon dos enfoques: 1) una predicción directa de clasificación, donde se predecía directamente 1 ($Pobre$) o 0 ($No\ pobre$) a partir de variables que desde un punto de vista económico fueran determinantes de la pobreza, 2) una predicción indirecta de pobreza a partir del ingreso, donde se utilizaron variables determinantes del ingreso de los hogares para luego comparar esta predicción con la línea de pobreza y así clasificar si un hogar era pobre o no. Para esto se entrenaron más de 22 modelos que emplearon los algoritmos de regresión lineal, *Elastic Net*, *CARTs*, *Random Forest*, y *Boosting*.

De acuerdo a los resultados encontrados, el mejor modelo en términos de precisión de predicción, que obtuvo un puntaje *F1* de $0{,}67$ fue el quinto modelo de predicción directa de la Tabla 2, el cual empleó el algoritmo de *XGBoost*. En este modelo se utilizaron 187 variables predictivas que contenían información del número de personas en los hogares, en dónde se encontraban ubicados estos hogares, el número de habitaciones, el número de menores en los hogares, el nivel de educación y características del jefe de hogar como por ejemplo si era mujer, el tipo de ocupación, si estaba afiliado a seguridad social, etc. Estas variables fueron fundamentales para entender la pobreza desde características multidimensionales de los integrantes de los hogares, pues más allá de la pobreza monetaria, estas variables logran capturar las precarias condiciones en que viven estos hogares, donde podemos encontrar casos en los que los menores de edad se ven en la obligación de trabajar para poder sobrevivir, o donde por ejemplo el hecho de que el jefe de hogar sea mujer con bajo nivel de estudios alcanzados implica una brecha salarial que determina que un hogar sea pobre.

Se encontró también, que usualmente los modelos de clasificación directa tenían un mejor *performance* que aquellos modelos con el segundo enfoque de medición indirecta, casi sin diferenciar el tipo de algoritmo que se usara, por ejemplo, el modelo 4 de la Tabla 2 obtuvo un puntaje *F1* de $0{,}63$, el más bajo de los mejores de clasificación, este fue el mismo puntaje que obtuvo el mejor modelo de medición indirecta (modelo 1 de la Tabla 3). Una característica común de estos "mejores modelos" es que entre más variables tuviera en cuenta para la predicción, más preciso era el modelo. El que menor puntaje obtuvo fue el modelo 2 de la Tabla 3, el cual solo consideraba 8 variables predictoras. Otra de las razones por las que el mejor modelo obtuvo un mayor puntaje *F1* es debido a que se utilizó el algoritmo de *XGBoost*, pues su enfoque de optimización de gradiente permite que los modelos se ajusten de manera más precisa a los datos y mejoren gradualmente su rendimiento, además implementa técnicas de regularización para controlar la complejidad del modelo y prevenir el sobreajuste.

Para mejorar los resultados obtenidos, es esencial tener en cuenta la limitación computacional con el fin de comparar distintos modelos con otros grupos de variables, distintos números de árboles y especificación de parámetros, dado que estos modelos más complejos requieren mucho más tiempo y poder computacional. Por otra parte, teniendo en cuenta la literatura revisada, otra forma de potencializar estos resultados sería adicionando otro tipo de datos como tipo de vivienda, materiales en que está construida, alcantarillado, enfermedades de los ocupados y no ocupados del hogar, gasto en salud (esto nos permitiría predecir la pobreza multidimensional), además de datos espaciales a nivel de manzana.

Referencias

Muñetón-Santa, Guberney y Luis Manrique-Ruiz (2023). «Predicting Multidimensional Poverty with Machine Learning Algorithms: An Open Data Source Approach Using Spatial Data». En: *Social Sciences* 12.5, pág. 296. URL: https://doi.org/10.3390/socsci12050296.

Sabogal, Hermes, Olmer García-Bedoya y Oscar Granados (2021). «Un análisis de la pobreza en Colombia basado en aprendizaje automático». En: URL: http://hdl.handle.net/20.500.12010/22282.

Sosa, Walter y Magdalena Cornejo (2022). «Predicciones agregadas de pobreza con información a escala micro y macro: Evaluación, diagnóstico y propuestas». En: *CEPAL*. URL: https://hdl.handle.net/11362/48018.

7. Apéndice A

Tabla 5: Comparación de los coeficientes de un modelo lineal y uno equivalente con un método de regulación

	Variable dependiente:	
	Ln (Ingreso Hogar)	
	(1 - MCO)	(2 - Regulación)
Personas x Cuarto	−0.254	−0.207
	(0.003)	
N° Ocupados	0.365	0.319
	(0.002)	
Mujer	0.013	0.000
	(0.004)	
H (Cotiz. Pension)	0.529	0.488
	(0.005)	
H (Reg. Subsidiado)	−0.491	−0.453
	(0.008)	
Constant	13.853	13.886
	(0.005)	
Observations	131,400	131,400
R^2	0.308	
R^2 Ajustado	0.308	
Residual Std. Error (df = 131394)	0.760	

8. Apéndice B

Figura 2: Árbol 1. Clasificación con XGBoosting

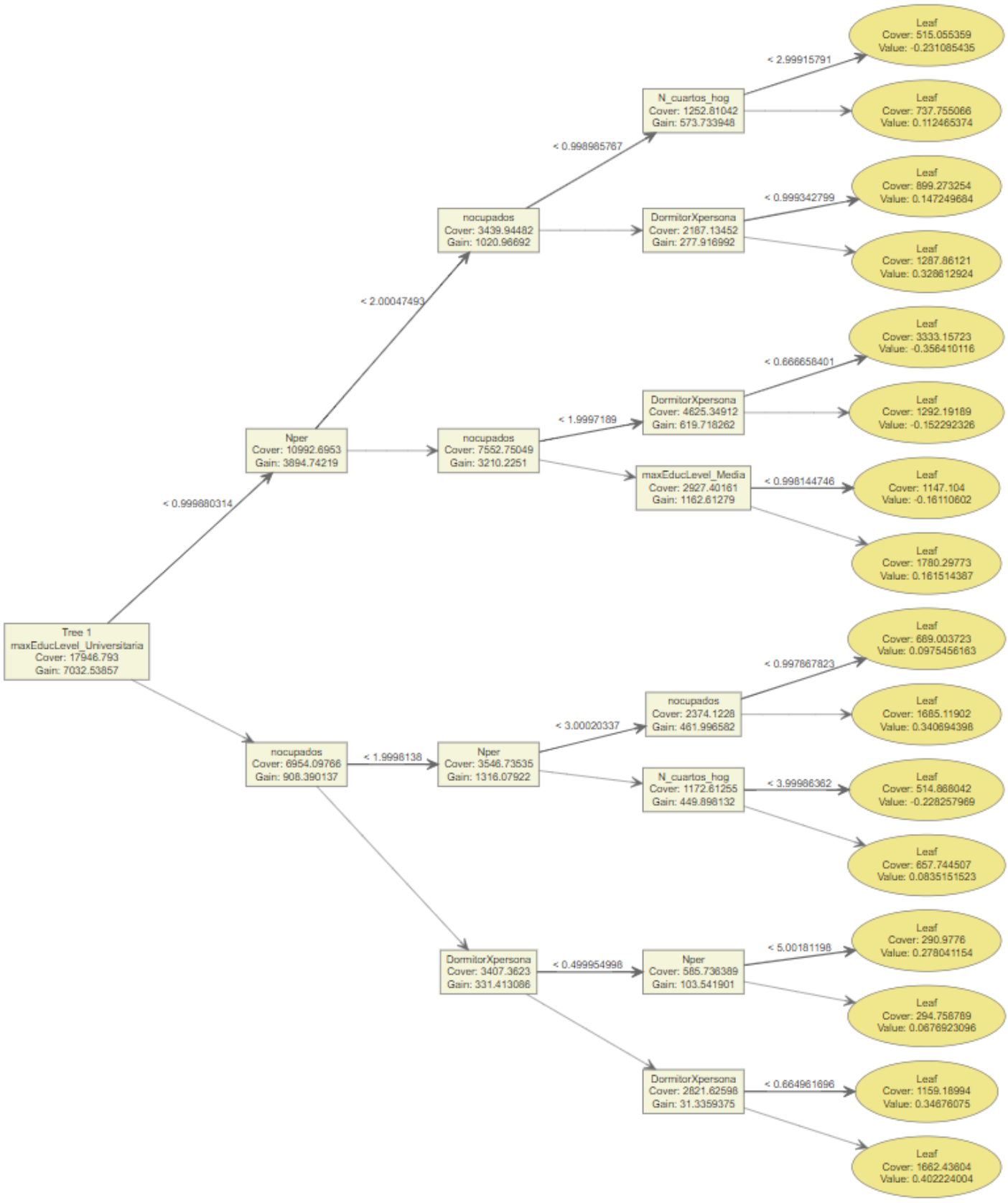


Figura 3: Árbol 499. Clasificación con XGBoosting