Andrés Chavarro

Juan David Vélez

Ricardo Silva (20182178)

Federico Camacho (202011648)

Taller 3 Big data & Machine Learning

Introducción

La pobreza es uno de los temas culmines de la economía, pues una es una de sus preguntas, más viejas, y más difíciles de contestar. Para ser más específicos uno de los grandes santos griales de la economía es dar a la respuesta de ¿qué factores hacen que un país sea pobre, ergo que sus ciudadanos y hogares lo sean? Con la era digital la conquista de esta pregunta se ha vuelto más intensa, dadas las modernas metodologías que existen para, precisamente, conocer que factores afectan la pobreza, y que pesos tienen estos. Con el creciente uso de herramientas como el "machine learning" y el "deep learning" predecir la pobreza se ha vuelto un ejercicio popular, pero esto no significa que se hayan generado metodologías homogéneas, sino más bien una gran cantidad de soluciones a estos ejercicios.

Para conocer un poco de estas metodologías se realizaremos una pequeña revisión de literatura, para no tan solo estar conscientes de los enfoques actuales, sino también tener recursos para conducir nuestro análisis. Uno de estos trabajos que tiene una popularidad importante es el de "Is Random Forest a Supirior Methodology for predicting poberty" de los autores Pave, T & Stender, N; en este se trabaja la pobreza desde el consumo, comparando el funcionamiento de dos enfoques, uno es econométrico ("Multiple Imputation") y el otro son los "Random Forest". La construcción se dio con 500 árboles con un mínimo de 4 observaciones por hoja. Evidentemente, para el estudio se dividió la muestra en dos, dónde una parte se usó para el entrenamiento del modelo, mientras la otra para testear la efectividad de este. El estudio termino concluyendo que la precisión fue más alta usando los "Random Forest". Otro estudio interesante es el de "Machine Learning Approach for Bottom 40 Percent Households (B40) Poverty Classification", donde quieren predecir los hogares que se encuentran sobre el 40% más pobre en Malasia. Los autores utilizan tres métodos, "Naive Bayes", arboles de decisión y "k-Nearest Neighbors". Todos los clasificadores fueron optimizados con parámetros diferentes, pero todos con 10 *folds* de validación cruzada, todos los modelos fueron ejecutados dos veces, la primera con 16 variables, para luego reducir la selección a las 8 mejores. Para comparar los modelos se utilizó tanto la precisión de los modelos como el estadístico de Kappa, dando como resultado los árboles de decisión. Como última parte de la literatura revisamos el estudio de "Does Artificial Intelligence Prevail in Poverty Measurement?" , dónde se realiza un metaanálisis de las tendencias sobre la estimación de la pobreza donde los métodos más populares terminan siendo los "Random Forest", LASSO, y OLS. Ya con el conocimiento de la literatura podemos guiar nuestro estudio hacia las metodologías que parecen tener mejor impacto.

Respecto al trabajo que se realizó para este taller, se menciona que el ejercicio fue inspirado debido a una competencia dirigida por el Banco Mundial donde se buscaba predecir la pobreza. Y en particular, el presente trabajo fue el de predecir la pobreza en Colombia. Dicha pobreza ha de ser predicha a nivel de hogares, tal es el objetivo. Y toda la información con la que se

contaba venía desde niveles de personas y hogares. Todo siendo suministrado por la entidad gubernamental del DANE (Departamento Administrativo Nacional de Estadística). Por otro lado, con el fin de hacer que el ejercicio fuese de mayor dificultad, las bases de datos para hacer testeo no contenían todas las mismas variables que aquellas con las que se entrenaban los modelos. Por lo cual, era necesario crear y jugar con la información disponible para poder realizar predicciones.

De acuerdo con el propio DANE, a un hogar se le considera pobre cuando sus ingresos per cápita con imputación de arriendo a propietarios y usufructos es menor que la línea de pobreza. Por lo cual, el problema se podía abordar desde dos frentes diferentes. Generar modelos que predijeran los niveles de ingreso de los hogares/personas o predecir si un hogar es pobre o no. En pocas palabras, problemas de regresión y clasificación. Siendo la métrica principal el f-score. En nuestro caso, el mejor modelo terminó siendo por medio del enfoque de clasificación. Se enviaron un total de 11 modelos en la competencia y en Kaggle, el mejor modelo terminó con un resultado de 0.62 y fue por medio de la metodología de Random Forest. Este f-score de 0.62 contrasta que se acanzao por medio de un random forest de clasificacion contrasta con el mejor modelo en la metodologia de regresion que alcanzo un f-score de 0,32 en Kaggle.

A nivel de conclusiones podemos decir que el manejo de datos es un trabajo retador. Esto debido a que las bases de testeo no contenían toda la información que las de entrenamiento y porque hay que tener un alto grado de entendimiento y creatividad a la hora de generar variables que puedan ser de relevancia a la hora de predecir algo tan delicado como lo puede ser la pobreza de un hogar. Es necesario tener un alto grado de entendimiento para evitar cometer posibles equivocaciones que puedan llevar a una mal distribución de recursos dentro de lo que puede llegar a ser una política pública. Toda la información y trabajo lo pueden encontrar en el siguiente enlace a github: BDML_PS3.

Construcción de variables y limpieza de datos

La construcción de variables presentó un desafío significativo debido a la discrepancia en las observaciones entre la base de entrenamiento y la base de prueba. A pesar de lograr un modelo exitoso en la base de entrenamiento, alcanzando una precisión cercana al 95% mediante la inclusión de variables como el estrato y el ingreso promedio por miembro del hogar, no fue posible evaluar este modelo en la base de prueba debido a la ausencia de las variables de ingreso por hogar y estrato en la base de datos test_hogares o test_personas.

La labor de construcción de variables se basó en la extracción de información de la base de datos de personas para agrupar características individuales por hogar, extrapolándolas posteriormente a nivel de hogar. Un ejemplo de este enfoque es la creación de una variable inicialmente a nivel personal, como la condición de desempleo de una persona. Esta variable se exportó al nivel de hogar utilizando una Dummy que indicaba la presencia de al menos una persona desempleada y en edad de trabajar en ese hogar (valor 1) o la ausencia de dicha situación (valor 0). Se aplicó una lógica similar para la variable relacionada de pensión, donde había una Dummy para el hogar que contara con una persona pensionada frente a aquellos hogares que carecían de ello, asimismo sucedió con la variable dummy de si el jefe de hogar contaba con educación superior (valor 1) en comparación a los hogares que no cumplen la condición. La utilización de la función mínima (min) fue crucial para asignar el valor de 1 a la Dummy del hogar y pensión, independientemente de la cantidad de personas desempleadas o

beneficiarias de pensiones en el hogar con que al menos una cumpliera la condición ya se activaba la Dummy.

Otra estrategia creativa consistió en derivar variables de nivel personal a nivel de hogar, como calcular un promedio de los años de educación de los miembros del hogar y agregarlo a la base de datos de hogares. Aunque este enfoque podría introducir ciertos errores de interpretación, como la dilución del promedio debido a la presencia de hogares con mayor cantidad de niños, se reconoce que esta dinámica puede influir en la pobreza, pues al haber mas niños hay mayores gastos y sujetos dependientes económicamente. De igual manera, Se exploró un efecto similar con la variable de edad promedio de los miembros del hogar, con la hipótesis de que hogares con una mayor edad podrían tener una menor incidencia de pobreza debido a una menor cantidad de miembros dependientes ya la presencia de más miembros potencialmente generadores de ingresos (PET).

Finalmente, se construyeron variables adicionales utilizando las observaciones disponibles en la base de datos de hogares, como una Dummy que indicaba si la vivienda era propia y ya se había pagado, versus si estaba en alquiler o si la vivienda era propia pero aún no se había pagado en su totalidad. También se incluyó el número de habitaciones del hogar como indicador de calidad, aunque se reconoció su vínculo con la cantidad de personas en el hogar. Además, la elección del departamento como variable proporciona una dimensión geográfica que podría influir en la predicción de la pobreza.

En el proceso de limpieza de la base de datos, nos enfrentamos a desafíos significativos, principalmente relacionados con la presencia de valores faltantes en varias variables. Algunas de estas variables presentaban más del 50% de valores ausentes, llegando incluso a alcanzar porcentajes alarmantes, como el 79% y 96%. Optamos por excluir estas variables de cualquier modelo, ya sea de clasificación o regresión, dado que la imputación de valores para ellas podría introducir un sesgo considerable, comprometiendo su poder predictivo. Ejemplos de tales variables incluyen si el hogar recibía remesas, si recibía subsidios gubernamentales o si los miembros del hogar trabajaban horas extras. Aunque podrían haber sido relevantes para predecir la pobreza en circunstancias diferentes, su falta de observaciones adecuadas las descartó de nuestro análisis.

Por otro lado, abordamos variables con porcentajes de valores faltantes inferiores al 35%, con la esperanza de mantenerlos en torno a variables con únicamente 20% de nulos y trabajamos estos datos haciendo imputaciones. En casos de variables discretas, como las dummy de desempleo, propiedad de la casa y pensiones, optamos por reemplazar los valores nulos con la mediana. Esta estrategia fue efectiva, ya que la probabilidad de que las observaciones faltantes representaran la situación opuesta era bastante alta, especialmente en variables con bajos porcentajes de incidencia, como el 5% de la muestra que tenía pensiones. En situaciones donde las variables eran continuas, como la edad promedio y la educación promedio, remplazamos los valores nulos con la media. En donde claramente era posible incluir observaciones con decimales. Es crucial destacar que la eliminación de observaciones nulas fue esencial para el correcto funcionamiento de los modelos, y las variables limpias se utilizaron tanto en los modelos de clasificación como en el modelo de regresión, garantizando la coherencia y la validez de los resultados obtenidos.

Estadísticas descriptivas

Dado que se trabajó con 4 bases de datos distintas y con variables extra por el lado de las bases destinadas al entrenamiento, a continuación, se pueden observar las siguientes tablas de estadísticas descriptivas de las bases de testeo y entrenamiento de los hogares.

Tabla 1 (Estadísticas descriptivas)

Estadisticas Descriptivas Train-Hogares

Statistic	N	Mean	St. Dev.	Min	Max
Clase	164,960	1.1	0.3	1	2
Cuartos	164,960	3.4	1.2	1	98
Cuartos_dormir	164,960	2.0	0.9	1	15
Amort_mes	5,626	919,943.2	6,115,976.0	98	280,000,000
Est_amort_mes	100,507	499,840.8	4,163,131.0	98	600,000,000
Arriendo_mes	64,453	437,911.8	1,447,543.0	20	300,000,000
Personas_hogar	164,960	3.3	1.8	1	28
Personas_gasto	164,960	3.3	1.8	1	28
Ingtotug	164,960	2,090,895.0	2,512,488.0	0.0	85,833,333.0
Ingtotugarr	164,960	2,307,865.0	2,628,933.0	0.0	88,833,333.0
Ingpcug	164,960	870,639.3	1,244,350.0	0.0	88,833,333.0
Lp	164,960	271,522.3	33,656.9	167,222.5	303,816.7
Indigente	164,960	0.05	0.2	0	1
Npobres	164,960	0.8	1.9	0	28
Nindigentes	164,960	0.2	1.0	0	17
Depto	164,960	37.1	23.9	5	76
train	164,960	1.0	0.0	1	1
edad_max	164,960	49.6	16.4	11	108
Promedio_edad	164,960	37.4	16.9	5.7	102.0
tiempo_edu	164,960	16.8	11.1	0	230
educ_superior	164,960	0.3	0.4	0	1
educ_media	164,960	0.3	0.4	0	1
regimen_salud_cont	155,237	0.5	0.5	0	1
horas_trabajo_total	164,960	67.4	49.0	0	640
pension	117,155	0.4	0.5	0	1
Pago_arr_pen	164,959	0.2	0.4	0	1
Pago_otros	164,959	0.3	0.5	0	1
desempleo	164,959	0.02	0.2	0	1
hogar_propio_arr	164,960	0.8	0.4	0	1
hogar_usu_otro	164,960	0.2	0.4	0	1

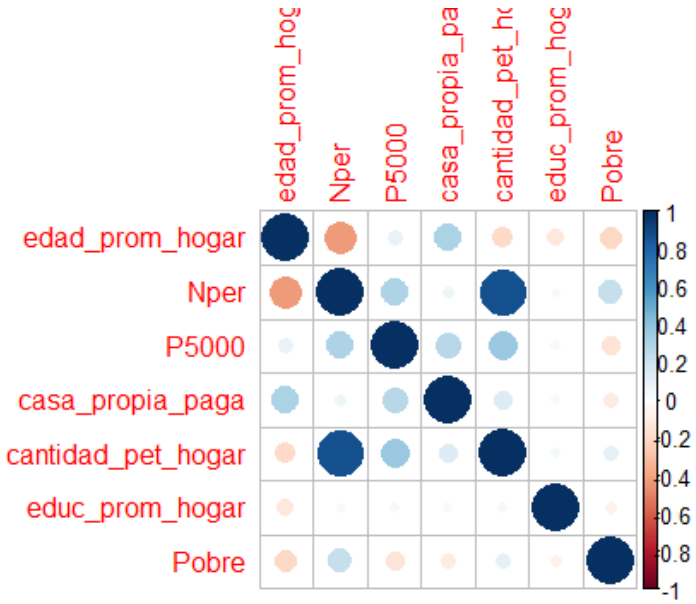
Estadisticas Descriptivas Test-Hogares

Statistic	N	Mean	St. Dev.	Min	Max
Clase	66,168	1.1	0.3	1	2
Cuartos	66,168	3.4	1.2	1	18
Cuartos_dormir	66,168	2.0	0.9	1	10
Amort_mes	2,138	716,047.9	2,094,332.0	98	90,000,000
Est_amort_mes	40,967	492,701.1	5,967,990.0	98	800,000,000
Arriendo_mes	25,201	421,770.9	1,016,769.0	98	150,000,000
Personas_hogar	66,168	3.3	1.8	1	21
Personas_gasto	66,168	3.3	1.8	1	21
Lp	66,168	270,984.5	34,849.5	167,222.5	303,816.7
Depto	66,168	39.0	23.6	5	76
train	66,168	0.0	0.0	0	0
edad_max	66,168	49.7	16.4	11	101
Promedio_edad	66,168	37.5	16.9	6.3	100.0
tiempo_edu	66,168	16.8	11.3	0	223
educ_superior	66,168	0.3	0.4	0	1
educ_media	66,168	0.3	0.4	0	1
regimen_salud_cont	62,331	0.5	0.5	0	1
horas_trabajo_total	66,168	67.4	48.9	0	544
pension	47,022	0.4	0.5	0	1
Pago_arr_pen	66,167	0.2	0.4	0	1
Pago_otros	66,167	0.3	0.4	0	1
desempleo	66,167	0.02	0.2	0	1
hogar_propio_arr	66,168	0.8	0.4	0	1
hogar_usu_otro	66,168	0.2	0.4	0	1

Las bases de entrenamiento de personas y hogares contaban con un total de 543,109 y 164960 observaciones respectivamente. Mientras que las de testeo de personas y hogares contaban con 219,644 y 66168 observaciones respectivamente.

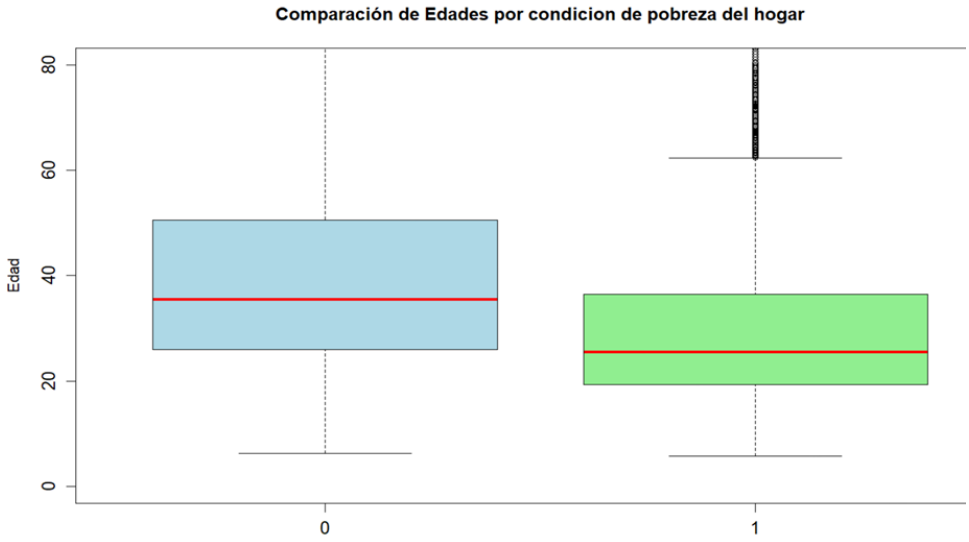
Ya luego de mencionar los cambios y los datos que dejamos después de hacer limpieza, podemos pasar a los análisis descriptivos, precisamente de estas variables. Comenzamos con el gráfico número uno, el cuál muestra la correlación entre las variables, esto lo hacemos por dos razones, la primera revisar que no exista correlación demasiado fuerte entre las variables explicativas y que efectivamente exista correlación entre estas últimas y la dependiente

Gráfico 1 (correlación)



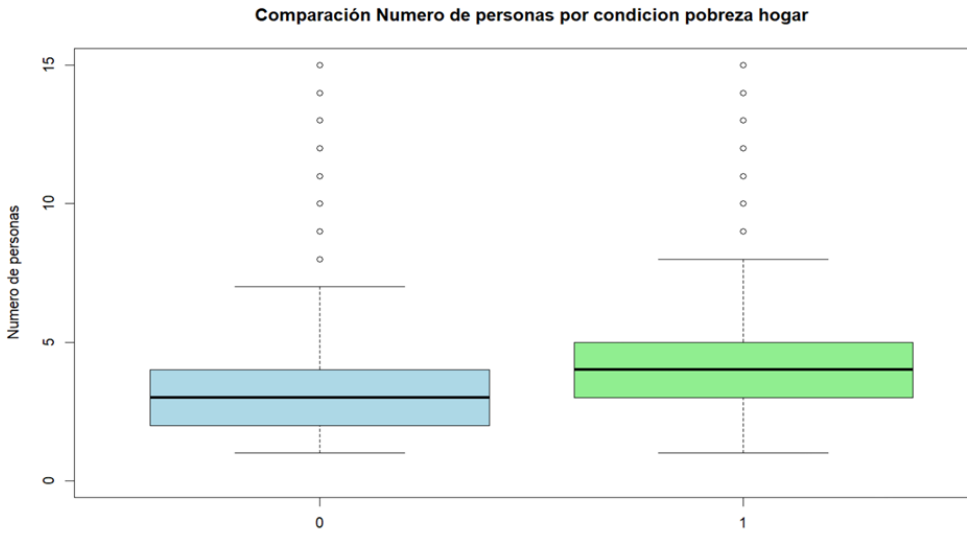
Siguiendo la hipótesis de cuando se hizo la construcción de variables que exploran estadísticas descriptivas en edad y cantidad de personas sobre pobreza.

Gráfico 2



La representación gráfica de caja y bigotes que muestra la edad promedio del hogar segmentado por su condición de pobreza confirma la noción de que hogares con una edad promedio más joven tienden a ser más propensos a ser catalogados como hogares pobres. Observamos que la edad promedio de los hogares pobres tiene una mediana de 23 años, en comparación con los 36 años de los hogares no pobres. Esta brecha de casi 15 años subraya cómo la presencia de miembros de mayor edad y una menor proporción de niños puede influir en la situación de pobreza de un hogar.

Gráfico 3



Este mismo patrón se explora a través de la Gráfica 3, que representa la distribución de personas en el hogar, diferenciando entre hogares pobres y no pobres. Resulta notable la comparación entre las dos cajas, donde la mediana es más elevada en los hogares pobres, y a sy vez también lo son cuartil inferior y el superior. Este hallazgo fortalece la relación positiva entre la cantidad de personas en un hogar y su tendencia a la pobreza.

Podemos vincular estos resultados con la teoría de Barten (1964), que postula que hogares con mayor cantidad de personas destinan una proporción significativa de sus recursos al gasto en alimentos. Esto se relaciona directamente con la pobreza, ya que, al igual que el coeficiente de Orshansky indica, una señal de situación precaria es el alto porcentaje de ingresos dedicado a alimentos.

No obstante, existe cierta divergencia teórica en cuanto a cómo las escalas de equivalencia podrían afectar la percepción de la pobreza del hogar en relación con la cantidad de niños. Algunas teorías sugieren que hogares más grandes pueden generar economías a escala y, en ocasiones, estar en mejores condiciones que hogares más pequeños (Lasso, 2002). Además, las escalas de equivalencia plantean que el gasto en un hijo no es equiparable al de un adulto, lo que podría incidir en la definición de pobreza del hogar. En nuestra base de datos, los hallazgos respaldan la teoría de Barten, indicando que hogares más jóvenes y con mayor cantidad de personas tienen una mayor propensión a la pobreza, evidenciada por menores ingresos, mayor dedicación de recursos a alimentos y una mayor carga de personas dependientes sobre el jefe de hogar.

Ahora bien, dado que la educación del jefe del hogar puede ser de gran interés debido a su relación que puede llegar a tener con el nivel de ingresos de un hogar y de la estabilidad laboral, se generó una variable dicótoma cuando este contaba con una educación superior (universitaria) y otra cuando tenía educación media.

Modelos

Clasificación

En el ámbito de los modelos de clasificación, es importante establecer que lo que se realiza es determinar si un hogar es pobre o no a partir de una cantidad de características o variables observables. En este caso se llevó a cabo un proceso de entrenamiento en la base de datos de hogares. Esta base de datos de entrenamiento se dividió en un 75-80% para entrenamiento y un 20-25% para prueba. Posteriormente, se aplicó la validación cruzada con el objetivo de determinar los hiperparámetros óptimos para diversos modelos, incluyendo Ridge, Lasso, Elastic Net, Random Forest y Boosting.

Durante este proceso, se emplearon métricas cruciales como la puntuación F1 y la precisión (accuracy) para seleccionar los modelos más eficientes. El modelo seleccionado se evaluó posteriormente en una base de datos de entrenamiento ciega, que no contenía la variable dummy indicando si la persona era pobre o no. Previamente, se construyó una matriz de confusión en la base de entrenamiento de hogares para comprender cuántos casos de pobreza fueron correctamente predichos y cuántos no.

La predicción resultante se sometió a la plataforma Kaggle, donde las variables construidas y las observaciones del hogar demostraron tener un poder predictivo significativo en la clasificación de hogares como pobres o no pobres. Este modelo final se destacó por su rendimiento superior en las métricas evaluadas y en la matriz de confusión durante el entrenamiento.

No obstante, el mejor modelo terminó siendo uno de Random Forest. Arrojando un valor de 0.62 en Kaggle. A diferencia de los demás, este contó con las siguientes variables y siguiente ecuación:

$$
\begin{aligned}
Pobre_1 = {} & \beta_0 + \beta_1\, Cuartos\ + \beta_2\, Cuartos.dormir + \beta_3\, Est.amort.mes \\
& + \beta_4 Personas.hogar + \beta_5\, Depto\ + \beta_6\, edad.max + \beta_7 tiempo.edu \\
& + \beta_8\, educ.superior + \beta_9\, educ.media + \beta_{10} regimen.salud.cont \\
& + \beta_{11} horas.trabajo.total + \beta_{12} pension\ + \beta_{13}\, Pago.arr.pen \\
& + \beta_{14}\, Pago.otros + \beta_{15} desempleo + \epsilon_i
\end{aligned}
$$

Dicho modelo de Random Forest, a partir de una grilla para calcular sus mejores hiper-parámetros terminó con los siguientes valores: Número de variables a considerar (mtry) de 4. Cantidad de árboles siendo 139 (trees). Profundidad del árbol de 9 (min_n).

Dado que el estado actual de la vivienda donde se considera si el hogar tiene su espacio como propio, arrendado, en usufructo, sin papeleo, entre otros estados, la veíamos como importante, decidimos correr un segundo modelo Random Forest idéntico al anterior, pero con esta variable:

$$Pobre_i = \beta_0 + \beta_1 Cuartos + \beta_2 Cuartos.dormir + \beta_3 EstadoVivienda + \beta_4 Est.amort.mes + \beta_5 Personas.hogar + \beta_6 Depto + \beta_7 edad.max + \beta_8 tiempo.educ + \beta_9 educ.superiro + \beta_{10} educ.media + \beta_{11} regimen.salud.total + \beta_{12} horas.hora.total + \beta_{13} pension + \beta_{14} Pago.arr.pen + \beta_{15} Pago.otros + \beta_{16} desempleo + \epsilon_i$$

De dicho segundo modelo Random Forest obtuvimos un resultado de Kaggle de 0.61. Sus hiper-parámetros resultaron siendo: Número de variables a considerar (mtry) de 3. Cantidad de árboles siendo 183 (trees). Profundidad del árbol de 3 (min_n). Esto solo nos lleva a pensar que entre estos modelos y los anteriores, puede haber una sobre especificación y por tanto un ajuste a los datos inadecuado.

Regresión

En el caso del problema a partir de una perspectiva de predicción de ingresos y su comparación con la línea de pobreza, aquí la clave es la de determinar si el nivel de los ingresos del hogar, tomando información de las bases a nivel personas, llevan a que se considere al hogar como pobre o no. En la implementación de los modelos de regresión, se siguió un proceso análogo. Se procedió a dividir la base de entrenamiento en conjuntos de entrenamiento y prueba, manteniendo las mismas proporciones que en los modelos de clasificación. Asimismo, se aplicó la validación cruzada como parte integral del proceso. Nuevamente se aplicaron modelos de ridge, Lasso, elastic net y random forest.

Sin embargo, surgió una complejidad adicional en este contexto. Dado que la base de datos de prueba (test_hogares) no contenía información sobre el ingreso del hogar, se adoptó un enfoque sustitutivo. Se utilizó la línea de pobreza como un ingreso mínimo que el hogar debe alcanzar para no ser clasificado como pobre. En este sentido, la metodología para construir la variable de pobreza se basó en que las variables observables fueran predictivas de la línea de pobreza.

La variable dependiente se estableció como la línea de pobreza. Si la predicción arrojaba un valor superior al de la línea de pobreza, el hogar se clasificaba como no pobre, ya que superaba el umbral definido por el Dane. Por el contrario, si la predicción de la línea de pobreza era inferior a la establecida en la base de datos, el hogar se consideraba pobre. La métrica utilizada en regresión para predecir la línea de pobreza fue el error cuadrático medio (mean square error).

Este proceso permitió construir la variable dummy de pobreza con valores de 1 y 0. Luego, se comparó con la base de datos de entrenamiento para construir una matriz de confusión. Esta matriz proporcionó información sobre la precisión de la clasificación de hogares como pobres o no pobres en la base de datos de entrenamiento. Asimismo, se recalculó la precisión (accuracy) y el F1 para evaluar la calidad del modelo en la tarea de clasificación de pobreza.

Para los modelos que se hicieron de regresión el que mejor logro resultados fue tambien un modelo de random forest.

Para este modelo la especificación era la siguiente:

$$
\begin{aligned}
Linea.de.pobreza_i &= \beta_0 + \beta_1 edad.prom.hogar_i + \beta_2 número.de.personas_i \\
&+ \beta_3 número.de.cuartos_i + \beta_4 casa.propia.paga_i \\
&+ \beta_5 educación.promedio.hogar_i + \beta_6 desempleados_i \\
&+ \beta_7 pensión.hogar_i + \beta_8 Depto_i + \epsilon_i
\end{aligned}
$$

Se implementó una búsqueda de hiper-parámetros utilizando un rango específico. Se buscó que cada hoja del árbol tuviera entre 1 y 30 observaciones, la cantidad de árboles variara entre 15 y 50, y que, al realizar la partición de los árboles, se tuvieran en cuenta 4 variables de la especificación. La optimización se llevó a cabo mediante validación cruzada, seleccionando el modelo que proporcionara el error cuadrático medio (RMSE) más bajo. El modelo final de Random Forest eligió 49 árboles, acercándose al límite superior de la búsqueda, y estableció un número mínimo de 24 observaciones por nodo terminal.

Este modelo se evaluó en la base de datos de prueba, prediciendo el valor de la línea de pobreza. Al aplicar la transformación a variables dummy de 1 y 0 para clasificar la pobreza según si la predicción estaba por encima o por debajo de la línea de pobreza, se calculó el accuracy y el F1 Score del modelo. El accuracy resultó en un 36%, mientras que el F1 Score apenas alcanzó el 30%. Estos resultados eran esperados, ya que, al observar la matriz de confusión, se asignaron muchos valores de pobreza incorrectamente, clasificando hogares como pobres cuando no lo eran. Este error afectó aproximadamente 20,000 observaciones, de las 42,000 que representan el 25% de la muestra en la base de datos de entrenamiento (train_hogares), que contiene alrededor de 162,000 observaciones a nivel de hogar. El modelo optimizado, al ser subido a kaggle, obtuvo un resultado en la métrica F1 del 32%, incluso mejor que el resultado en la base de entrenamiento.

Por otro lado, un modelo que destacó al exhibir un resultado sorprendentemente alto fue el segundo mejor modelo de regresión, construido mediante Elastic Net. Este modelo buscó la combinación óptima de lambda y alpha para regularizar la regresión lineal, específicamente en la predicción del nivel de pobreza del hogar, que interpretamos como el ingreso mínimo esperado. Utilizamos las mismas variables que en la especificación anterior y ajustamos el modelo para mejorar la métrica F1. Exploramos un rango de valores de lambda entre 20 y -20, utilizando 100 números diferentes para regularizar la predicción de la línea de pobreza. Durante la selección del mejor modelo, notamos que se eligió un número decimal extremadamente cercano a 0 (1e-20), indicando que se requería muy poca regularización.

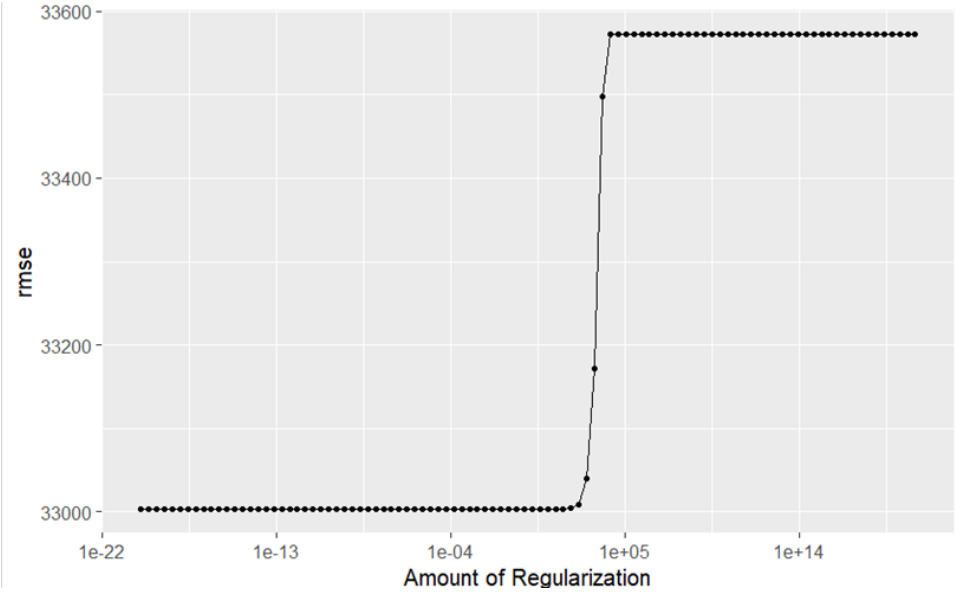
Al observar la Gráfica 4, notamos que a medida que aumentaba la penalización combinada elegida, el error cuadrático medio crecía, sugiriendo predicciones cada vez menos precisas. Esta metodología no demostró ser la más efectiva. Al realizar la transformación explicada previamente para construir la variable dummy de pobreza y al construir la matriz de confusión en la base de prueba sintética, nos enfrentamos a una situación similar a la de Random Forest, donde muchos hogares fueron incorrectamente clasificados como pobres, llegando a casi

24,000 hogares mal ubicados y aproximadamente 6,200 hogares pobres correctamente identificados. La precisión del modelo fue del 32%, y el F1 Score fue del 34%.

Sin embargo, al extrapolar estos resultados para la prueba real, el rendimiento del F1 Score empeoró, sorprendiendo ya que en la muestra de entrenamiento había sido más alto que con Random Forest. El resultado en Kaggle fue del 29%.

Selección de penalización en Elastic net para predecir la cercanía del ingreso a la línea de pobreza respecto al mean squared error

Gráfico (4)



De hecho, observamos la misma tendencia al intentar hacer predicciones con Lasso, donde el RMSE también empeoraba a medida que se incrementaba la penalización. Además, no conseguimos obtener una predicción en la base de kaggle con un F1 que mejorara; este se empeoro en un 23%. Sin embargo, a través de Lasso, pudimos identificar que la variable de pensión y el número de personas eran altamente relevantes, ya que sus coeficientes no se acercaban tanto a cero con el aumento de la penalización. Esto contrasta con, por ejemplo, la variable del departamento al que pertenecía el hogar, que posiblemente estaba introduciendo ruido en la especificación o se estaba manejando incorrectamente internamente.

Conclusiones

No sobra resaltar que nuestros resultados no solo tienen coherencia interna-validez interna- sino que también creemos que tenemos validez externa, por lo que resulta ser un ejercicio robusto. Como ya se mencionó en los resultados, nuestros modelos mejor comportados, en el sentido que mejores predicciones hicieron fueron “Random Forest” esto en concluyente con lo que se menciona en Sohnesen, T.P ., & Stender, N (2017) pues precisamente nuestros mejores modelos de predicción fueron los que utilizaron esta metodología. Esto mismo es coherente con las tendencias que se están dando para pronosticar la pobreza, como los muestra Isnin, R. Et al. (2020), pues desde el 2007 el “random forest” ha tendio la mayor popularidad. Con estos datos anteriores confirmamos la robustez del ejercicio en un sentido de validez externa, ya que precisamente sigue los métodos más efectivos, y con lo que se está publicando en la actualidad.

Asimismo, el trabajo fue funcional en utilizar variables relevantes en la literatura de pobreza para introdúcelas en modelos predictivos, variables como si había un desempleado, la cantidad de cuartos en el hogar, la cantidad de personas en el hogar, la educación que alcanzaban los miembros del hogar, entre otras variables, fueron útiles para predecir y tenían signos congruentes tanto en los modelos de clasificación por medio de regresiones logísticas tanto en los otros que fueron utilizados en este trabajo. Los modelos de Ridge y Lasso nos ayudaron para encontrar la importancia de estas variables dada que en la regylarizacion eran los últimos en acercar su coeficiente a 0, pero estos mismos modelos al usarlos para predecir la pobreza del hogar tanto en regresión como en clasificación no fueron muy eficientes y dejaron mucho que desear. Hubiese sido interesante haber explorado modelos de redes neuronales y relaciones no lineales en las variables. Cabe reiterar que todo se puede encontrar en el repositorio siguiente: BDML_PS3.

Estos ejercicios son claves para el futuro puesto que determinar si un hogar es pobre o no puede terminar siendo extremadamente costoso dada la cantidad de gente que puede llegar a tener un país, por complicaciones a nivel de movilidad o por la geografía del mismo. Por lo cual, abordar dicha clasificación por medio de machine-learning pueden ser de gran utilidad y de bajo costo.

Referencias

- Barten, A. (1964). "Family Composition, Prices, and Expenditure Patterns." En: Economic Analysis for National Economic Planning, editado por G. Hart, Milis, y J. Whithaker, Londres: Butterworths.
- Sohnesen, T. P., & Stender, N. (2017). Is random forest a superior methodology for predicting poverty? An empirical assessment. *Poverty & Public Policy*, *9*(1), 118-133.
- Sani, N. S., Rahman, M. A., Bakar, A. A., Sahran, S., & Sarim, H. M. (2018). Machine learning approach for bottom 40 percent households (B40) poverty classification. *Int. J. Adv. Sci. Eng. Inf. Technol*, *8*(4-2), 1698.}
- Lasso,F (2002). Pobreza y economias de escala en hogares. Planeacion y desarrollo.
- Isnin, R., Bakar, A. A., & Sani, N. S. (2020, April). Does Artificial Intelligence Prevail in Poverty Measurement?. In *Journal of Physics: Conference Series* (Vol. 1529, No. 4, p. 042082). IOP Publishing.