María Carolina Gutiérrez Cód. 201922402
Sara Torres Benavides Cód. 201923245
Sofía Quiroga Robles Cód. 201913999
Luis Carlos Rodríguez Cód. 201920638

Problem Set 2: Predicting Poverty

I. Introducción

Los indicadores de desarrollo de Colombia no parecen ser prometedores. En particular, la desigualdad y pobreza del territorio parecen influir negativamente en la distribución del ingreso y el acceso a las oportunidades de los hogares. El dato más reciente de desigualdad, medido a través del índice de Gini para 2021, indica un valor de 0.523 (DANE, 2021), lo cual refleja una alta desigualdad para Colombia, pues el índice sigue muy alejado de tomar valores cercanos a cero para indicar una sociedad menos desigual. Con esto en mente, el gobierno colombiano ha implementado programas y/o ayudas monetarias para ayudar a los más vulnerables y aminorar los efectos de pobreza y desigualdad. Entre estos se encuentran programas como Ingreso Solidario, Familias en Acción, Colombia mayor, Mi Casa Ya, entre otros. Sin embargo, focalizar estas ayudas se ha convertido en un gran desafío para las entidades públicas, pues se percibe en muchos casos manipulación del punto de corte por parte de los hogares ya que varios ocultan información necesaria como por ejemplo los ingresos para recibir ayudas monetarias. Esto ha generado una pérdida de eficiencia y de recursos del gobierno al no poder determinar con certeza qué hogares son los más necesitados.

Con esto en mente, el presente trabajo pretende encontrar un modelo que sea capaz de predecir la pobreza de los hogares. En particular, se aborda el problema bajo dos diferentes enfoques: predecir la pobreza a través de un modelo de clasificación y a través de predecir los ingresos de cada hogar. La segunda forma permite comparar el ingreso con la línea de pobreza monetaria para luego clasificarlo como pobre o no. Ahora bien, para cumplir con el objetivo de la construcción del modelo se utilizó la base de datos del DANE de la Gran Encuesta Integrada de Hogares (GEIH) y la Encuesta Continua de Hogares (ECH). Las bases contienen información a nivel de individuo y de hogares, por lo cual al relacionar el individuo con el hogar es posible crear una medida tanto de ingreso como de clasificación de pobreza del hogar. De esta forma, la metodología empírica logra afrontar el problema de predicción de pobreza. La base también incluye información relevante a nivel de individuo, como el sexo, edad, jerarquía en el hogar, seguridad social y estrato. Por otro lado, a nivel hogar se tiene información sobre la ubicación (i.e área municipal, cabeceras o resto), características físicas del hogar, como número de cuartos, si el lugar habitado es propio o no, el pago por arriendo y número de personas que viven en el hogar. Con los datos previstos es posible abordar una medida cautelosa de pobreza del hogar al combinar las características del hogar con las características individuales.

Entre los resultados principales del trabajo se destaca que los modelos de predicción directa tienen un mayor porcentaje de predicciones correctas. También, los modelos seleccionados fueron los de Random Forest, pues tuvieron el mejor puntaje en accuracy. Más aún, el mejor modelo construido tuvo un porcentaje de acierto del 83.2%, indicando que logró predecir el 83.2% de los hogares en la muestra de testeo que efectivamente eran clasificados como pobres. Este modelo utilizó la metodología de Random Forest que a su vez permite disminuir la variación y la correlación entre los árboles de clasificación, por lo cual es más robusto que los modelos de regresión; sin embargo, es más difícil de interpretar. El resultado parece ser adecuado, pero todavía es posible ajustar nuevas variables y/o parámetros para lograr una mayor capacidad de predicción.

II. Datos

La base contiene información relevante a nivel de individuo, como el sexo, edad, jerarquía en el hogar, seguridad social y estrato. Por otro lado, a nivel hogar se tiene información sobre la ubicación (i.e área municipal, cabeceras o resto), características físicas del hogar, como número de cuartos, si el lugar

habitado es propio o no, el pago por arriendo y número de personas que viven en el hogar. Con los datos previstos es posible abordar una medida cautelosa de pobreza del hogar al combinar las características del hogar con las características individuales.

Las bases anteriores del DANE, se encontraban en formato comprimido csv, por lo cual, usando el comando "saveRDS" se guardan las cuatro bases de datos en formato RDS, para ser cargado al repositorio colaborativo y ser manipulado. Luego de guardar los dataframe, se inicia la limpieza de la base, de forma que en primer lugar se unen variables medidas a nivel individual a las bases medidas a nivel hogar, para los datos de entrenamiento y de prueba. De esta forma, con la llave "id" de cada individuo, se encuentra el total de los ingresos por hogar, el estrato socioeconómico del hogar, así como el promedio del nivel educativo más alto alcanzado por todos los integrantes de la misma familia, el promedio del régimen de seguridad social en salud al cuál pertenece cada individuo del hogar, y finalmente, el promedio de las horas trabajadas a la semana de todos los miembros del hogar. A partir de los datos ponderados anteriores, se utiliza el comando "left_join" para unir estas nuevas bases con las correspondientes de prueba y entrenamiento a nivel hogar, utilizando la variable "id" como llave.

En segundo lugar, para manejar las observaciones en blanco o "Missing Values" (NA), se decide realizar tres enfoques diferentes. De esta forma, se copian tres veces las bases de entrenamiento y prueba de los hogares. En primer lugar, se imputan todas las observaciones faltantes por una constante de valor cero. Sin embargo, bajo esta medida se incurren en cambios en la distribución, además, se puede incurrir en errores de predicción y estimación de los modelos, al sobreestimar o subestimar los parámetros. En segundo lugar, también se plantea la imputación determinística por la media de las observaciones válidas. No obstante, es importante mencionar que bajo esta metodología también se modifica la distribución de las variables, al reducir la varianza dado el incremento artificial en las observaciones. Asimismo, se utiliza este método para tratar las observaciones en blanco debido a que los parámetros de la media y la varianza siguen el valor esperado para cualquier observación al azar de una distribución normal. Finalmente, se imputan los "Missing Values" por la mediana de cada variable dentro de sus observaciones válidas. Siendo este uno de los métodos más utilizados en la práctica, dada su fácil implementación. Asimismo, dadas algunas distribuciones asimétricas, se encuentra que la mediana es la mejor representación para las observaciones faltantes. Sin embargo, como se ha mencionado antes, bajo este método también se puede distorsionar la distribución de las variables y aumentar su varianza.

A continuación, se encuentran las estadísticas descriptivas para las variables utilizadas en la muestra de entrenamiento a nivel hogar, en específico, consisten en las variables de educación, salud y las horas de trabajo para predecir el ingreso. La muestra comprende a 164.960 hogares.

Tabla 1. Estadísticas descriptivas						
Variable	**Media**	**Desv. Est.**	**Mín.**	**Pctl. 25**	**Pctl. 75**	**Máx.**
Número de Cuartos	3.39	1.239	1	3	4	98
Número de dormitorios	1.989	0.898	1	1	3	15
Total de personas en el hogar	3.292	1.775	1	2	4	28
Número de personas en la unidad de gasto	3.28	1.772	1	2	4	28
Pobre (=1 si el hogar es pobre)	0.2	0.4	0	0	0	1
Ingreso total del hogar	2102585.769	2532552.392	0	8e+05	2.518.241,5	858.33.333,3
Estrato del hogar	3.088	1.611	1	2	4	6
Educación del hogar	4.315	1.076	1	3.6	5	9
Salud del hogar	1.968	0.871	1	1	3	9
Horas de trabajo del hogar	45.246	12.199	1	40	49	130

Tabla 1. Estadísticas descriptivas de las varables de predicción, las variables d eresultado (pobreza e ingreso) al igual que algunas características de los hogares y de las viviendas, como el número de cuartos, número total de personas, educación.

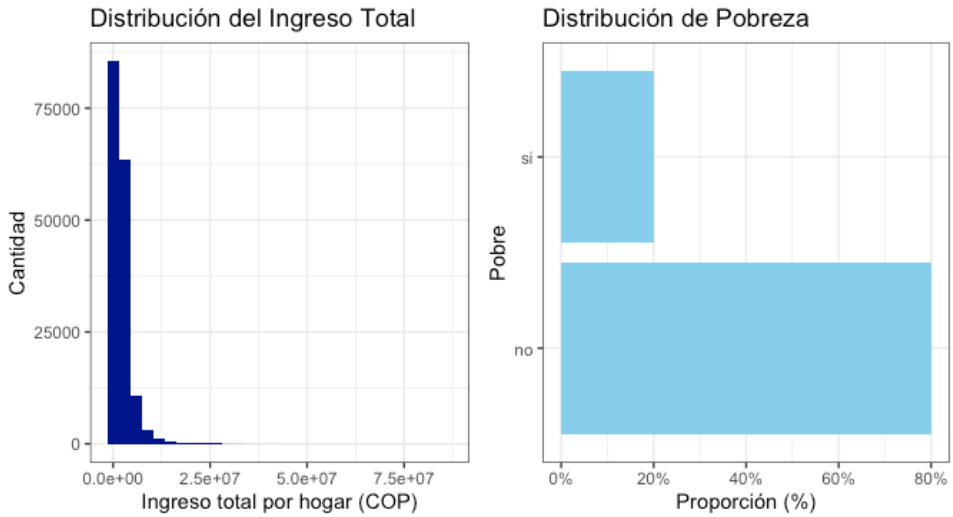
Para comenzar, es importante comenzar el análisis con las características generales de los hogares. En primer lugar, aproximadamente los hogares se comprenden de aproximadamente tres personas, con una desviación estándar de 1 persona más. De hecho, solamente el 25% de los hogares tienen en total 2 personas en promedio y el 75% de los hogares los comprenden aproximadamente cuatro personas. Es importante mencionar que el número máximo de personas viviendo en un mismo hogar son 28 personas. Un número bastante alto y fuera del promedio de los hogares. Esta variable es muy próxima a la variable de número de personas en el hogar como unidad de gasto. Por otro lado, en promedio los 164.960 hogares tienen en su vivienda tres cuartos, el 25% solo tiene a rededor de tres mientras que el 75% de los hogares tienen cuatro cuartos en la vivienda. Nuevamente, hay un dato atípico del número de cuartos máximo, pues 98 cuartos está muy por encima de la media y de la desviación estándar. Es importante mencionar que a pesar de que en promedio los hogares tienen 3 cuartos, en promedio los hogares duermen en únicamente dos cuartos. Esto puede dar indicios de que los hogares colombianos promedio utilizan uno de sus cuartos para trabajar o lo toman como área de descanso. Sin embargo, esto indica una mayor aglomeración del espacio por individuo en el hogar en la categoría de número de cuartos en los que duermen.

En otro orden de ideas, la variable estrato es una variable categórica que indica el estrato de la persona entrevistada. Esta variable toma los valores entre 1-6. El promedio de la variable (3.08) indica que un colombiano promedio se encuentra en estrato 3. Sin embargo, esta variable es altamente dispersa, pues el 75% de la muestra se encuentra entre el estrato 1 y 4 (este cálculo corresponde a sumar y restar dos veces la desviación estándar al promedio de la variable). Adicionalmente, es importante mencionar que, como lo indican los percentiles, únicamente el 25% de la muestra pertenece a los estratos 1 y 2, mientras que el 75% de la muestra se encuentra hasta el estrato 3, donde el restante 25% de la muestra componen los estratos 4, 5 y 6. Esta alta variabilidad da indicios de los desigual que es la población colombiana, donde la mayoría perteneces a la clase media o baja, mientras que los más ricos (clasificados como aquellos que ganan más de 10.000.000 COP) se encuentran en los estratos más altos pero que apenas representan el 25% de la muestra.

Ahora bien, también es importante mirar el comportamiento de las variables de respuesta, como lo son el ingreso y la clasificación directa de pobreza. En cuanto a la variable que clasifica directamente a un hogar como pobre, se identifica que en promedio, de los hogares tomados en consideración, el 20% se clasifican como pobres, con una desviación estándar de 40%. Es una cifra baja al comparar con los datos completos de Colombia, pues para 2018 se estimó que el 27% de los colombianos se encontraban en pobreza monetaria (DANE, 2018). De hecho, el 75% de la muestra no se encuentran en condición de pobreza monetaria. En cuanto a la variable de ingreso total en el hogar, se tiene que en promedio los hogares encuestados tienen en promedio ingresos anuales equivalentes a 2.102.585,769 COP. De nuevo, la tabla 1 pone en evidencia la alta volatilidad de esta variable en la muestra. Por ejemplo, hay hogares que no tienen ingresos (ingreso total equivalente a cero), mientras que el hogar con el máximo de ingresos tiene un monto de 85.833.333,33 COP. También se percibe la alta desigualdad de ingresos dentro de muestra: apenas el 25% de la muestra recibe los mayores ingresos que están entre 8.00.000 y 85.833.333. Mientras que el 75% de los hogares de la muestra perciben ingresos totales entre 0 y 2.518.241.5 COP. De nuevo, estas cifras son altamente preocupantes, pues existe una gran volatilidad de la distribución de los ingresos, que nuevamente se relacionan con el alto índice de desigualdad en Colombia.

La distribución de ambas variables se refleja en la siguiente gráfica:

Gráfica 1: Distribución del ingreso y de clasificación de pobreza por hogar



Gráfica 1: Izquierda: distribución del comportamiento del ingreso total por hogar. Derecha: distribución de la clasificación de pobreza por hogar como porcentaje del total de la muestra.

La gráfica 1 ilustra que la variable de ingreso es altamente dispersa concentrada en bajos niveles monetarios. También, evidencia que la variable está sesgada a la izquierda. Estos resultados son consistentes con la alta desigualdad de ingresos del país. De nuevo, la gráfica muestra que el 20% de los hogares de la muestra son clasificados como pobres, respecto al total de la muestra.

Finalmente, es importante mirar las estadísticas de las variables predictoras del modelo. En primer lugar, se encuentra la educación del hogar. Esta consiste en una variable categórica que indica el nivel máximo de educación alcanzado, como es a nivel hogar consiste en el promedio del nivel educativo máximo alcanzado por integrantes del hogar. Toma el valor de 1 a 7, donde 9 es un indicadora igual a 1 si la persona no responde. En promedio, la muestra tiene un nivel educativo básico de secundaria incompleto (6to-9no), identificado como 4. Es altamente preocupante, pues este resultado indica que un colombiano promedio, tomado de la muestra, no alanza a terminar el colegio. Esto da indicios de la alta probabilidad de incurrir en deserción escolar, pobreza y de no alcanzar los recursos económicos necesarios. En parte, en Colombia ocurre debido a las presiones familiares, donde no se impulsa la

educación sino el trabajo para aportar ingresos y alimentos al hogar. De hecho, alrededor del 75% de los encuestados en 2018 se encontraban con bajos niveles de educación, entre ningún tipo de estudios como apenas acabando el colegio. La varianza en esta variable es relativamente alta, pues con una desviación estándar, una persona puede pasar de un nivel educativo de secundaria básico a uno medio, y de una incompleto a uno de apenas cruzar quinto de primaria. En cuanto a los indicadores de desarrollo y desigualdad, las estadísticas son preocupantes, pues no existe una buena formación de capital humano ni tampoco una buena provisión del servicio por parte del estado. La variable es útil como predictora, pues el nivel educativo es una muestra del nivel de conocimientos y preparación que tienen las personas, donde, según la literatura, cada año de educación se traduce en mayores niveles de ingreso y superación de la trampa de pobreza.

En segundo lugar, la variable de salud del hogar indica si la persona está afiliada o cotiza a un servicio de seguridad social. Es una variable categórica que toma el valor de 1 sí el encuestado está afiliado al régimen contributivo, 2 si está afiliado a un régimen especial y 3 si está subsidiado. El 9 indica que la persona no contestó la pregunta. Esta variable permite tener una idea de la formalidad del trabajo de la persona y en cierta medida del hogar, ya que todos los empleados con contrato formal en Colombia deben cotizar salud. Además, la variable es una buena aproximación para clasificar a un hogar como no pobre, pues en caso de pertenecer a un régimen de salud da evidencia de que se satisfacen las necesidades básicas, sobre todo si pertenece a un régimen contributivo o especial, pues de esta forma un hogar no debería estar clasificado como pobre. Como ilustra la tabla 1, en promedio los hogares encuestados pertenecen bien sea al régimen contributivo o al especial (media de 1.96). Sin embargo, solo el 25% de la muestra está afiliado al régimen contributivo. De nuevo, esta baja cobertura del servicio de salud puede dar evidencia de la alta desigualdad tanto monetaria como de oportunidades de la población.

En tercer lugar, la variable de horas trabajadas es el promedio de las horas de trabajo de los individuos que comprenden un mismo hogar por semana. Como indican las estadísticas descriptivas, las horas promedio de trabajo del hogar son aproximadamente 45.24 por semana, es decir, alrededor de 9 horas por día (de los cinco días de trabajo a la semana). Esta variable también es particularmente variable pues existe una variabilidad de 12 horas más de trabajo en promedio, equivalente a un día más de trabajo. El 25% de la población trabaja 5 horas menos respecto a la media, mientras que el 75% trabaja en promedio 4 horas más que el promedio. Es importante mencionar que el número de horas máximo trabajado es de 130 horas semanales. Es un dato bastante atípico, pues implica que por semana se traban 18 horas (los 7 días de la semana). En particular, esta variable permite capturar la formalidad del trabajo y características del trabajo de los individuos que comprenden el hogar. Puede ayudar a predecir el ingreso, pues entre más horas trabajadas se espera una mayor remuneración. Sin embargo, también puede reflejar la necesidad de trabajo de los más vulnerables, al estar en disposición de aceptar varios trabajos, cada uno de medio tiempo, por ejemplo, para ayudar a los ingresos del hogar. También, para aquellos hogares donde se trabaja lo mínimo (1 hora semanal) se puede estimar un alto nivel de pobreza al no poder suplir con las necesidades básicas de consumo al no recibir ingresos necesarios según las pocas horas trabajadas.

III. Modelos de clasificación

Los modelos de clasificación son un tipo de modelos utilizados en el Machine Learning los cuales tienen la característica de que buscan predecir una cualidad en vez de una cantidad: es decir, que la variable a predecir no es un número, sino una categoría a la cual clasificar un individuo dependiendo el valor de sus predictores. El objetivo de este trabajo es definir si una persona es pobre o no, por lo que se puede hacer por medio de un modelo de clasificación. Donde se puede clasificar a una persona como “Pobre” o “No Pobre”. Ahora el objetivo es utilizar la base de datos para crear un modelo que pueda predecir si una persona es pobre o no. Para lograrlo se utilizaron varios métodos de clasificación, utilizando la métrica “Accuracy” para definir el mejor de ellos.

El primer paso fue escoger las variables del modelo. Las variables escogidas fueron: Educación del hogar, Salud de hogar, horas de trabajo del hogar, personas por cuarto y el gasto de las personas del hogar. Luego de escoger las variables, se utilizaron varios modelos de clasificación para descubrir cual es el mejor de ellos. Los modelos utilizados fueron: 3 Random Forest con diferente cantidad de predictores; 1 Modelo de Gradient Boosting (gbm), 1 modelo de Logit-Lasso usando Up-sampling y 1 modelo de Logit-Lasso usando Down-sampling.

Random Forest

Los modelos de Random Forest son modelos que se caracterizan por su facilidad de uso y adaptabilidad, aunque su interpretabilidad es muy limitada. Estos modelos consisten en hacer varios árboles de decisión utilizando diferentes variables, en diferentes órdenes y con diferentes profundidades. Luego de hacer una gran cantidad de árboles distintos entre sí, se promedia los resultados y se obtiene un modelo de predicción.

A) Modelo Random Forest 3 variables en la media.

Se utilizo un modelo de Random Forest utilizando las variables:

$$Pobre = Educacion + Salud + Horas_trabajo$$

Para este modelo, se realizó un cambio en los missing Values, el cual fue cambiar todos estos por el valor de la media del predictor al cual pertenecían. El Accuracy en muestra del modelo fue de 0.82. El Accuracy en test fue de 0.807.

B) Modelo de Random Forest con 5 variables en la media

Se utilizo un modelo de Random Forest utilizando las variables:

$$Pobre = Educacion + Salud + HorasTrabajo + PersonasGasto + PersonasCuarto$$

Para este modelo, se realizó un cambio en los missing Values, el cual fue cambiar todos estos por el valor de la media del predictor al cual pertenecían. El Accuracy en muestra del modelo fue de 0.842. El Accuracy en test fue de 0.831.

C) Modelo de Random Forest con 5 variables en la mediana

Se utilizo un modelo de Random Forest utilizando las variables:

$$Pobre = Educacion + Salud + HorasTrabajo + PersonasGasto$$

Para este modelo, se realizó un cambio en los missing Values, el cual fue cambiar todos estos por el valor de la mediana del predictor al cual pertenecían. El Accuracy en muestra del modelo fue de 0.849. El Accuracy en test fue de 0.826.

Gradient Boosting

El modelo de Gradient Boosting es un modelo que utiliza arboles de decisión. Este modelo va "aprendiendo" mientras más arboles realiza. La técnica consiste en hacer arboles pequeños y ver cómo funcionan, si el árbol no funciona bien, se cambia en gran parte, pero si funciona bien, se hacen pequeños ajustes. El objetivo del modelo es aprender cuales son las características que mejor funcionan para asi aplicarlas en el modelo final.

D) Modelo de Gradient Boosting

Se utilizo un modelo de Gradient Boosting utilizando las variables:

$$Pobre = Educacion + Salud + HorasTrabajo$$

Para este modelo, se realizó un cambio en los missing Values, el cual fue cambiar todos estos por el valor de la mediana del predictor al cual pertenecían. El Accuracy en muestra del modelo fue de 0.82. El Accuracy en test fue de 0.809.

Logit-Lasso

El modelo utiliza la metodología de Logit, la cual es una metodología de elección binaria (0 y 1) junto al método de máxima verosimilitud para encontrar la probabilidad de que los individuos sean 0 ó 1. Lasso es una forma de regularizar modelos para mejorar sus capacidades a través de una restricción que coloca más peso a las variables, por lo que el error del modelo aumenta y de esta forma acercándose al error que tendría el verdadero modelo al predecir. La metodología Up-sampling es usada cuando las categorías no estan muy balanceadas, teniendo una categoría una proporción mucho mayor a las otras, por lo que se decide crear nuevas observaciones de la categoría menor teniendo en cuenta características de las verdaderas observaciones de la categoría menor. Down-sampling es utilizado de la misma manera para corregir desbalance de clases, pero en este caso se eliminan observaciones de la categoría superior para que se balancee con la categoría menor.

E) Modelo de Logit-Lasso usando Up-sampling.
Se utilizo un modelo de modelo de Logit-Lasso usando Up-sampling utilizando las variables:

$$Pobre = Educacion + Salud + HorasTrabajo$$

Para este modelo, se realizó un cambio en los missing Values, el cual fue cambiar todos estos por el valor de la mediana del predictor al cual pertenecían. El Accuracy en muestra del modelo fue de 0.78. El Accuracy en test fue de 0.714.

E) Modelo de Logit-Lasso usando Down-sampling.
Se utilizo un modelo de modelo de Logit-Lasso usando Down-sampling utilizando las variables:

$$Pobre = Educacion + Salud + HorasTrabajo$$

Para este modelo, se realizó un cambio en los missing Values, el cual fue cambiar todos estos por el valor de la mediana del predictor al cual pertenecían. El Accuracy en muestra del modelo fue de 0.78. El Accuracy en test fue de 0.714.

IV. Modelos de regresión

La segunda alternativa para predecir la pobreza es hacerlo de forma indirecta. Esta alternativa consiste en predecir el ingreso del hogar y usar estas predicciones para clasificar el hogar como pobre o no, comparando la predicción del ingreso con la línea de pobreza monetaria. Para llevar a cabo este proceso se plantearon diferentes modelos de regresión usando como variables las variables en común entre el training set y el test set.

En primer lugar, se creó una variable “valor_arriendo” que es el valor estimado del arriendo en el caso en que el hogar no pague arriendo de su vivienda pero respondió una estimación en la encuesta y el valor real de su arriendo en el caso en que el hogar sí paga arriendo. En segundo lugar, se crearon las variables “arriendo_tamano” indicando el valor del arriendo dividido entre el número de cuartos de la vivienda. Por último, la variable dependiente es el ingreso total de la unidad de gasto del hogar. En todos los modelos planteados, las variables predictoras tuvieron un nivel de significancia del 1%.

$$ingreso\_total_i = \beta_0 + \beta_1 valor\_arriendo_i + \beta_2 tot\_personas_i + u_i (1)$$

$$ingreso\_total_i = \beta_0 + \beta_1 valor\_arriendo_i + \beta_2 tot\_personas_i + \beta_3 valor\_arriendo_i * tot\_personas_i + u_i (2)$$

$$ingreso\_total_i = \beta_0 + \beta_1 valor\_arriendo_i + \beta_2 tot\_personas_i + \beta_3 arriendo\_tamano_i + \beta_4 cuartos\_pc_i\, u_i (3)$$

$$ingreso\_total_i = \beta_0 + \beta_1 valor_arriendo_i * tot_personas_i + \beta_2 arriendo_tamano_i + u_i (4)$$

Para ajustar los modelos de regresión se llevaron a cabo diferentes métodos, entre estos los métodos de oversampling y undersampling. La muestra del training set presenta desbalance de clases, con más del 75% de la muestra clasificada como "No Pobre". Por lo tanto, se llevaron a cabo los dos métodos, uno para aumentar el tamaño de la categoría minoritaria y otro para disminuir el tamaño de la categoría mayoritaria. En términos generales, los resultados de ambos métodos fueron similares. Al predecir la clasificación del hogar, comparando el ingreso predicho con la línea de pobreza monetaria, la medida de accuracy dentro de la muestra para cada uno de los modelos con ambos métodos estuvo entre 0,5 y 0,500003.

Además de las soluciones para el desbalance de clases, se decidió ajustar los modelos sin balanceo de clases y usando técnicas de regularización. Por un lado, se estimaron los modelos sin balanceo de clases y entrenando una regresión lineal estándar. En estas predicciones, la medida de accuracy dentro de la muestra fue mayor que con oversampling y undersampling, manteniéndose entre 0,78 y 0,79. Para los modelos estimados con Ridge y Lasso los resultados fueron similares, la medida de accuracy dentro de la muestra fue de 0,79 en ambos casos.

Al contrastar esta alternativa de predicción con la predicción directa, la predicción directa tiene un mejor desempeño dentro de la muestra. Entre las limitaciones de la predicción indirecta están las variables disponibles, predecir el ingreso del hogar a partir de características de la vivienda implica que se está dejando por fuera las características de las personas que generan ese ingreso. Adicionalmente, el proceso de predecir una categoría a partir de una cantidad predicha puede dificultar el ajuste del modelo y su desempeño por fuera del training set.

Modelo Final

Al momento de realizar las predicciones, los modelos con mejor Accuracy en el test set fueron el de Random Forest con 5 variables en la media y Random Forest con 4 variables en la mediana.

El modelo de Random Forest con 5 variables consiste en la predicción del siguiente modelo:

$$Pobre = Educacion + Salud + HorasTrabajo + PersonasGasto + PersonasCuarto$$

Estas variables tenían una cantidad considerable de observaciones vacías (Missing Values). Para no perder estas observaciones, se optó por calcular la media de las variables a utilizar y de esta forma llenar los missing values con esta medida. Al hacer esto no perdemos observaciones y el modelo estimado puede ser mejor, aunque corre el riesgo de obtener un poco de sesgo.

Las variables escogidas fueron: Educación (Ya que el nivel de educación es un indicador de los recursos de la persona), Salud (una persona con más ingresos tiene acceso a una mejor salud), Ingreso (Mientras mayor ingreso tenga la persona, menor será la probabilidad de que sea pobre). Gasto (una persona con mayores recursos tiende a gastar más dinero), Personas por cuarto (Mientras más personas viven en un mismo cuarto, mayor es la probabilidad de hacinamiento, y este es un indicador de pobreza). (Para mayor análisis de las variables, leer la sección 2 "Datos").

El modelo de Random Forest con 4 variables consiste en la predicción de:

$$Pobre = Educacion + Salud + HorasTrabajo + PersonasGasto$$

A diferencia del modelo de 5 variables, los missing values de este modelo fueron remplazados por la mediana, por lo que los datos usados para hacer la predicción son algo diferentes a los datos del primer modelo.

El algoritmo para realizar la estimación de los modelos fue el de Random Forest. Este método de predicción consiste en crear arboles pequeños usando diferentes variables para luego promediar los resultados de los árboles y así obtener el modelo final. Un árbol se construye eligiendo una variable, luego se observa en que valor de la variable se pueden separar las observaciones en dos regiones. Luego se escoge una segunda variable para que se ubique en alguna de las regiones separadas por la primera variable, de esta forma la nueva variable vuelve a separar la región donde se encuentra en otras dos regiones. Este algoritmo se puede repetir hasta que las variables se terminen, dando de esta forma una respuesta a cualquier conjunto de valores de los predictores.

El Random Forest utilizado en el modelo de 5 variables en la media utiliza un control para cada árbol que crea. Este control consiste en un sistema de Cross Validation con 5 folds. Esto consiste en que antes de que cada árbol sea creado, se dividen los datos en 5 partes iguales. 4 partes se usan para hacer el modelo y la quinta para testearlo. Se repite este procedimiento hasta que se hagan todas las combinaciones de entrenamiento – testeo con los 5 folds. Luego se promedian los resultados para obtener el modelo con la mejor combinación de sesgo-varianza. Al hacer este proceso, el árbol obtenido en cada creación del Random Forest tiene una buena proporción de sesgo-varianza.

El Random Forest utilizado en el modelo, tiene la métrica de que cada árbol que se crea tiene un tamaño de 2, 3 y 5 nodos (Predictores). De esta forma se crean 3 bosques distintos, uno de árboles de 2 nodos, uno de árboles de 3 nodos y uno de árboles de 5 nodos. De esta forma cuando se promedien los bosques, darán resultados distintos, los cuales pueden maximizar alguna métrica de resultado.

Las métricas de resultado son valores que nos muestran que tan bien funciona el modelo dentro y fuera de muestra. Las métricas más utilizadas son: ROC, Sensibilidad, Especificidad, Exactitud (Accuracy) y KAPPA. La métrica de importancia para este modelo es la Accuracy. Esta métrica consiste en el cociente entre los valores verdaderos predichos correctamente más los valores falsos predichos correctamente sobre el total de las observaciones ((TP+TN)/(P+N)). Este valor en resumen mide la cantidad de aciertos que tuvimos al predecir los si una persona era pobre o no.

Para comprar los modelos se utilizó los modelos de 3 variables y de gbm. Pero los resultados de estos modelos tenían un accuracy menor al de los modelos escogidos. El de 3 variables tenía una accuracy de 0.81 en muestra y el gbm un accuracy de 0.82 en muestra. La razón de esto puede ser principalmente por el número de predictores. Al tener más predictores el modelo tiene una mayor capacidad ya que se está evitando los problemas de sesgo de variable omitida. Otra razón de que el modelo gbm no sea tan alto como los otros puede ser por los parámetros escogidos. Este modelo tuvo una limitación de 500 árboles, profundidad de 4 y nodos mínimos de 10. Esto le puede dar una pequeña limitación al modelo y con números diferentes de estos valores podrían generarse diferentes modelos, por lo que alguna podría superar a los modelos escogidos, pero como este método requiere una alta capacidad de procesamiento, por eso se limitó a esos parámetros.

Al comparar el modelo de Random Forest con 5 variables con los modelos de predicción indirecta con regularización por Lasso y Ridge, este modelo tiene un mejor desempeño prediciendo dentro y por fuera de la muestra. Por un lado, el modelo de Random Forest es una forma directa de predecir la pobreza, mientras que los modelos de regresión predicen la pobreza a partir de una predicción del ingreso. Por otro lado, el método de Random Forest permite disminuir la variación en el promedio de los árboles y a partir de los resultados de la medida de accuracy, es una mejor forma de ajustar el modelo que los modelos de regresión para estimar el ingreso.

Por otro lado, al comparar el modelo de 4 variables en la mediana con los modelos de Logit Lasso, tanto en el caso de upsampling como en el de downsampling, el modelo seleccionado tiene un mejor desempeño prediciendo dentro y fuera de la muestra. Las tres especificaciones se hicieron reemplazando los missing values para cada variable por su mediana. Sin embargo, el modelo Logit tiene un menor porcentaje de predicciones correctas, esto es porque los métodos de clasificación son muy diferentes. Mientras que con Logit se estiman los coeficientes al minimizar la función de verosimilitud, el método de random forest puede llegar a ser más robusto al generar varios árboles de clasificación que no están correlacionados entre sí.

V. Conclusión y Recomendaciones

La pobreza es uno de los principales problemas indicadores de desigualdad de los países, por esa razón los gobiernos buscan formas de combatirlas, y una forma de hacerlo es a través de predecir la pobreza teniendo en cuenta las características de los individuos y hogares, para así poder llegar a las personas que necesitan ayuda. Para poder realizar esta estimación, es necesario utilizar modelos de predicción y clasificación. Los modelos que mejor resultados dieron fueron los de Random Forest. Este tipo de modelos son difíciles de interpretar, pero son modelos fáciles de utilizar, ya que solo necitas poner los datos y muestra un resultado con un alto nivel de confianza. El mejor modelo de Random Forest tenía una posibilidad de acierto del 83.2%, esto significa que el modelo es bueno, pero aún se puede mejorar para aumentar la capacidad de predecir correctamente. Otro modelo que dio buenos resultados fue el de gbm, pero tenía una capacidad de predecir correctamente algo menor al Random Forest.

La principal recomendación para predecir la pobreza es tener una buena base de datos de prueba y testeo. Este trabajo tuvo la limitación de que los datos de testeo no eran tan buenos como los de entrenamiento. Si se hubiera tenido una mayor cantidad de variables, se podría hacer un modelo más robusto y con mejor capacidad predictiva. Sin embargo, es importante no sobre ajustar el modelo al incorporar una gran cantidad de variables. Otra limitación que tuvo el trabajo fue la predicción al utilizar los modelos de gbm, ya que al ser modelos que demanda una gran cantidad de poder de procesamiento, se tienen que limitar los parámetros y no se puede correr una gran variedad de modelos de gbm.

VI. Anexos

Enlace al repositorio de GirHub: https://github.com/saratben/Taller2_Repositorio

VII. Referencias

Báez, K. J. (2022). *Estimación de datos faltantes a través de redes neuronales, una comparación con métodos simples y múltiples*. Universidad Santo Tomás. https://repository.usta.edu.co/bitstream/handle/11634/42861/2022kellybaez.pdf?sequence=1&isAllowed=y

Departamento Administrativo Nacional de Estadística, (DANE), (2018). Boletín Técnico, Pobreza Monetaria en Colombia. Recuperado de: https://www.dane.gov.co/files/investigaciones/condiciones_vida/pobreza/2018/bt_pobreza_monetaria_18.pdf

DANE, (2019). Medición de Pobreza Monetaria y Desigualdad 2018. Recuperado de: https://www.kaggle.com/competitions/uniandes-bdml-20231-ps2/data