Problem Set 2

William Aguirre, Andrés Arévalo & John Londoño

2024-04-14

1. Introducción

La pobreza es uno de los temas centrales de investigación en la economía del bienestar y en las agendas políticas de todos los gobiernos, especialmente en los países en vías de desarrollo. Sin embargo, no hay un consenso en su definición y en su forma de ser medida. A pesar de esto, se puede entender que ser pobre es no disponer de los recursos para obtener los medios mínimos de subsistencia, una definición general. El Banco Mundial define la pobreza como incapacidades para llevar una vida plena, desagregándolo en buena alimentación, calidad de la vivienda, buena salud, buena educación y tener un trabajo digno. También tenemos la definición de uno de los economistas que más influencia tiene en el estudio de la pobreza, Amartya Sen. Con la teoría de las capacidades, este autor dice que la pobreza no es simplemente la falta de ingreso, sino la falta de capacidades básicas y la libertad para obtenerlo, definición muy parecida a la propuesta del Banco Mundial. En resumen, los avances del estudio teórico de la pobreza han reposado en que es un fenómeno multidimensional y no solamente carencia de ingresos.

De la misma forma que las definiciones, la forma de medir la pobreza es variada, donde se tiene en cuenta la multidimensionalidad de la pobreza que tanto se menciona en la etapa de definición. En primer lugar, se tiene el NBI (Necesidades Básicas Insatisfechas) que señala la insuficiencia que tiene un hogar en una de las siguientes necesidades básicas: Vivienda con Materiales adecuados, servicios públicos de acueducto y alcantarillado, nivel bajo de hacinamiento, bajo grado de dependencia. También se encuentra el método de la línea de pobreza, que calcula el costo de una canasta básica de bienes y servicios, luego calcula los ingresos del hogar y se compara con esta línea, si no alcanza este mínimo entonces se considera un hogar pobre. Por último, también se trabaja con el Índice de Condiciones de vida, que comprende variables que miden la calidad de la vivienda, el capital humano actual y potencial, el acceso a la calidad de los servicios y las condiciones del hogar. En este trabajo se va a trabajar con el indicador de línea de pobreza.

El objetivo de este artículo es encontrar los determinantes económicos y sociales que más inciden en clasificar a una persona como pobre o no pobre a través de la medida de línea de pobreza. Para esto se utilizan los datos del DANE que vienen de la Misión de Empalme de las Series de Empleo, Pobreza y Desigualdad. En esta base de datos se encuentra a nivel de hogar y de personas y recoge toda la información sobre las condiciones de empleo de las personas, además de todas las características socioeconómicas generales y que son importantes para determinar la pobreza de un hogar. Por tanto, con estas variables se puede estimar en términos de probabilidades las características de la población menos favorecida.

El estudio de los determinantes de la pobreza es importante porque muestra los temas en los que es necesario focalizar las políticas públicas. En Colombia se han realizado varios estudios que buscan encontrar los determinantes de la pobreza con datos de tipo transversal. Nunes & Ramírez (2002) utilizan un modelo de respuesta cualitativa para encontrar las características socioeconómicas de los hogares más pobres a través de la estimación de probabilidades. Estos autores encuentran que el nivel educativo, la cantidad de los miembros del hogar, la tasa de ocupación y la ubicación del hogar son variables que afectan la probabilidad de ser o no hogares pobres.

En la misma dirección, Chaves (2010) encuentra con un modelo logit que la educación del jefe de hogar, la posesión de activos, el tamaño del hogar, el tipo de vivienda y el género del jefe de hogar son determinantes de que este se encuentre o no en situación de pobreza. De esta forma, la mayoría de los estudios de los determinantes de la pobreza tienen una forma similar de trabajar, se establecen las variables de control, se crea la variable de respuesta binaria a través de la línea de pobreza y se procede con la estimación de los parámetros con modelos probit o logit (Nunes & Ramírez, 2002), (Nunes, Ramírez, Cuesta, 2005), (Chaves, 2010), (Marrugo et al., 2015).

Un factor común que utilizan todos los estudios es trabajar y estimar la pobreza de un hogar a través de las condiciones del jefe de hogar, esto debido a que es el responsable sobre todo económico de todos los integrantes. Por esta razón, es importante estudiar la posición del jefe de hogar en el mercado laboral, ya que muestra la capacidad de los ingresos y calidad de vida. Investigar sobre su posición en el mercado y si trabaja en condiciones de informalidad puede ser un determinante de la condición de pobreza. Esto debido a que trabajar en condiciones de informalidad, por sus características precarias y bajos salarios, puede explicar la pobreza en un hogar (Torres et al., 2022). También la edad del jefe de hogar puede servir como determinante de la pobreza porque, desde la teoría, se considera que hay edades donde es tiene más incidencia el desempleo como en la población joven y más adulta (Klose, 2012).

2. Datos

Seleccion de los datos

La seleccion de las variables se hacen tomando como referencia los estudios analizados para este proceso y todo se trabaja a nivel de hogar como lo recomienda Chaves(2010), (Torres, et al,2022), (Klose,2012)

$p6050$: Si es jefe de hogar; $p620$: Sexo del jefe de hogar; $p6100$: Tipo de regimen; $p6210$: Educacion; $p6240$: Empleado; $p6430$: $Posicionlaboraldeljefedehogar$; $Personasporhabitacion$; $Tipodevivienda$: propia o no; $Edaddeljefedehogar$; $Zona$: Rural o Urbana; $Tipodetrabajo$

Limpieza de los datos.

Todo el analisis exploratorio de datos y la modelación se hace para las personas que han reportado que son jefes de hogar, por esto la limpieza mas importante es crear la variable jefe, que establece que si es jefe de hogar tiene el valor de 1 y 0 en los otros caso, por esto al momento de traer las variables del nivel de personas al de hogar se hace tomando como referencia si es jefe de hogar,

```
train_personas$jefe<-ifelse(test = train_personas$P6050==1,1,0)
test_personas$jefe<-ifelse(test = test_personas$P6050==1,1,0)
```

Para automatizar el proceso de traer las caracteristicas que pueden explicar la pobreza monetaria de un hogar se crea la función crear_, que tiene el siguiente funcionamiento. Primero se guarda la función segun seal nombre de la variable a traer, esta funcion recibe como parametro la base de datos sobre la cual se va a construir la variable, el primer filtro que se aplica es que en la base de datos de personas debe estar filtrada solo para los jefes de hogar, del dataframe se toma la variable de interes junto con el identificador unico de personas que crea la base de datos y por ultimo se crea el join con la base de datos de hogares por medio de este identificador. La función llamada traer variable garantiza que no se repitan los id y que tenenmos registros unicos en toda la muestra.

```
crear_sexo_jefe<-function(df){ aux<-df %>% filter(jefe==1)
  aux2<-data.frame(sexo_jefe=aux$P6020,id=aux$id)
  df<-left_join(df,aux2,by="id")
  return(df)}
train_personas<-crear_sexo_jefe(train_personas)
train_hogares<-traer_variable(train_hogares,train_personas,"sexo_jefe")
train_hogares$sexo_jefe<-as.factor(train_hogares$sexo_jefe)
test_personas<-crear_sexo_jefe(test_personas)
test_hogares<-traer_variable(test_hogares,test_personas,"sexo_jefe")
test_hogares$sexo_jefe<-as.factor(test_hogares$sexo_jefe)
```

Este proceso se hace para cada una de las variables , en este dosumento solo se presenta el caso para el sexo del jefe del hogar, en el script Creación de variables de interes esta el proceso para cada una de las variables.

Este proceso se hace para las siguientes variables:

Tabla de Variables

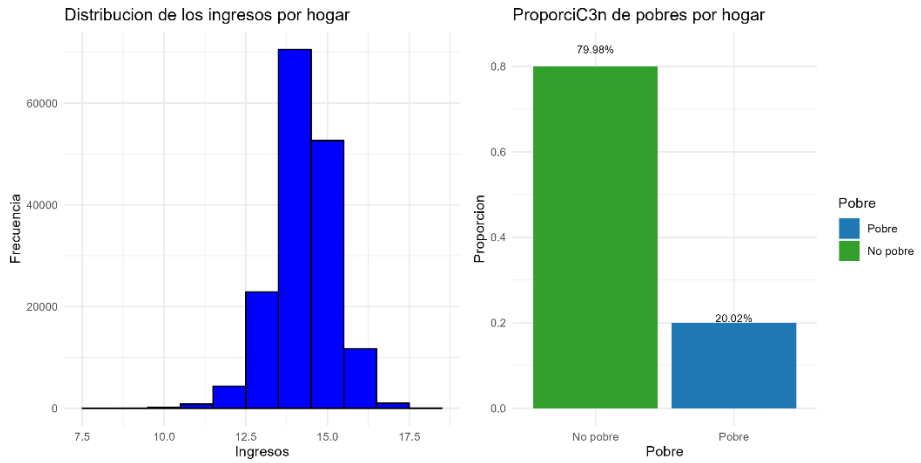
Descripción de las variables y sus hipótesis

Variable	Transformacion	Hipotesis
Sexo_jefe	Se trabaja como una variable dummy 1 cuando es hombre y cero cuando es mujer	Las mujeres son mas vulnerables que los hombres
Regimen_subsidiado	1 cuando el jefe de hogar esta afiliado en el regimen subsidiado, 0 en otro caso	Las personas afiliadas al regimen subsidiado son mas vulnerables
Educacion_jefe	Ultimo grado aprobado por el jefe de hogar	Entre mas alto el grado de educacion, menos es la probabilidad de ser pobre
Ocupacion_jefe	Si el jefe de hogar manifiesta que la semana anterior trabajo, 0 y 1 para cualquier otra respuesta	Los jefes de hogar que se encuentren desempleados hacen que el hogar sea mas vulnerable a estar en condicion de pobreza
Posicion_Jefe	Se trabajan todas las categorias de la variable sin ninguna transformacion	La actividad que desarrolla un jefe de hogar ayuda a determinar si su hogar puede estar en condicion de pobreza por el diferencial de ingresos en cada actividad
Personas_habitacion	Se divide el total de personas por la cantidad de habitaciones que reporto el jefe de hogar	Entre mas alto sea el número de personas por hogar mas vulnerable es a estar en condición de pobreza
Edad del jefe del hogar	Se toma la definición de informalidad del DANE si el establecimiento tiene menos de 10 trabajadores se considera como informal	Por la teoria del ciclo vital hay intervalos donde se es mas vulnerable a estar en la pobreza.
Informalidad_jefe	Se trabaja tal cual se reporta en la encuesta	Las personas en condición de informalidad son mas vulnerables a estar en pobreza

Analisis exploratorio de los datos

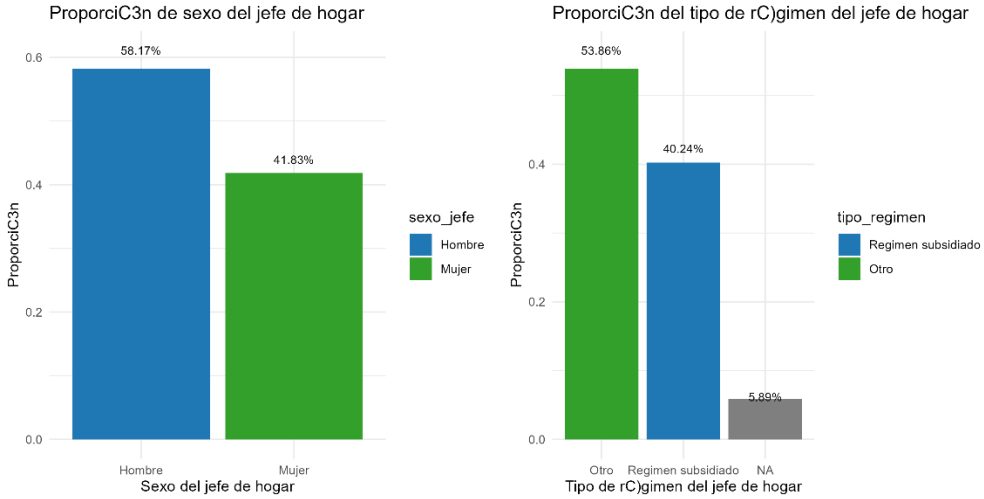
El analisis exploratorio se datos se divide en dos partes: Analisis univariado y analisis multivariado de datos. En el analisis univariado el proposito es analizar como se distribuyen de forma individual las variables que vamos a utilizar en el proceso de modelacion. Luego con el analisis multivariado nuestro objetivo es ver como se relacionan las variable independientes con el ingreso de los hogares.

Analisis univariado



Estas son las dos variables objetivo que se van a explicar en este informe. En la primera gráfica se presentan los ingresos agregados del hogar, que se utilizan para determinar si un hogar es pobre utilizando la línea de pobreza, mientras que la variable 'Pobre' está categorizada como 0 cuando el hogar no es pobre y 1 cuando lo es. Los datos muestran que aproximadamente el 80% de los hogares no están en condición de pobreza, lo que implica un desafío en términos del proceso de modelado debido a que hay un claro desbalance de clases.

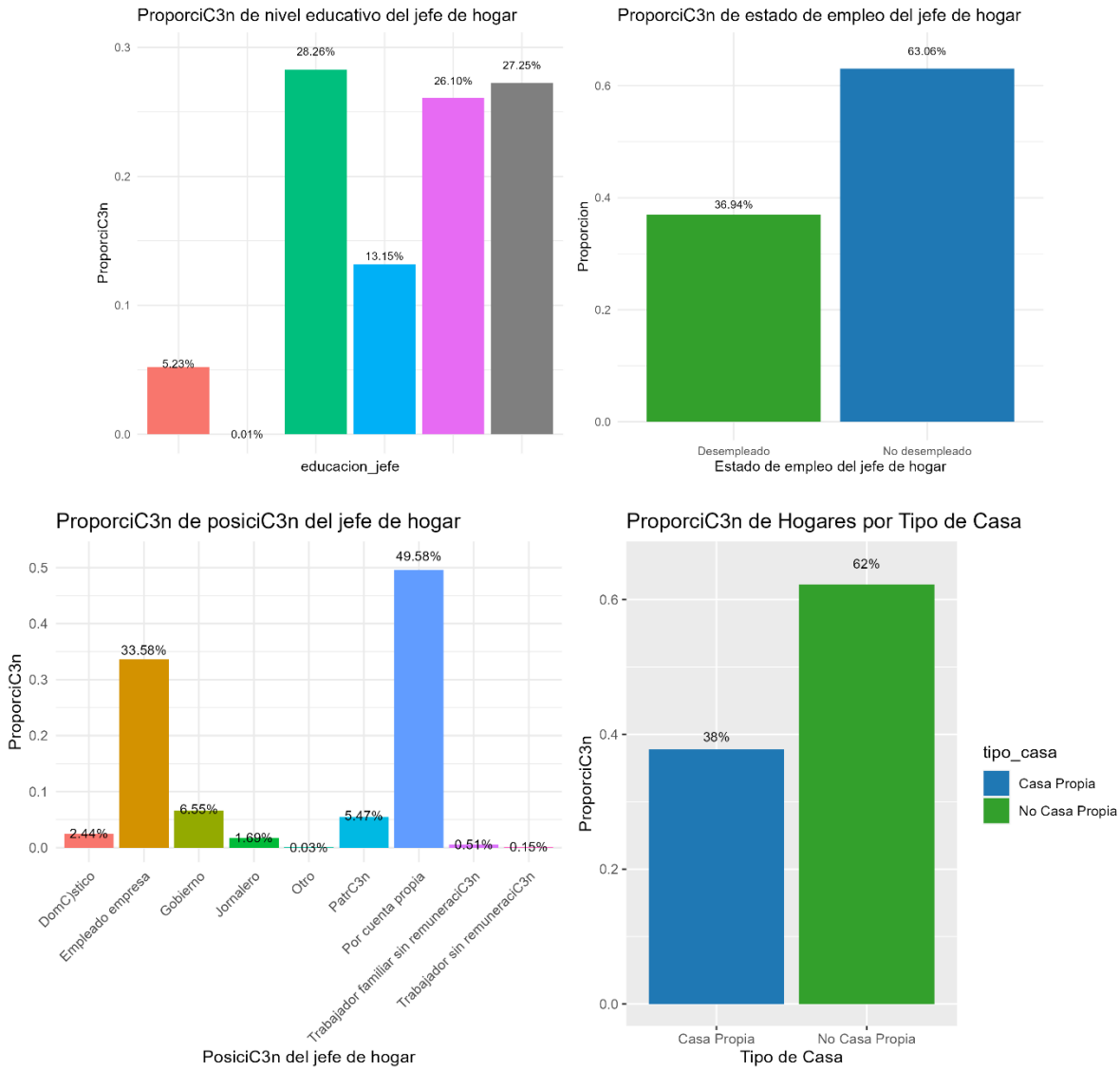
Variables explicativas



En este caso, observamos que la representación de jefes de hogar masculinos es ligeramente mayor que la de las mujeres que encabezan el hogar. Según la literatura económica, cuando la mujer es la jefa de hogar, existe un mayor riesgo de pobreza en el hogar. Esta situación puede atribuirse a diversos factores, como la brecha salarial de género y la responsabilidad desproporcionada de cuidado y trabajo doméstico no remunerado.

Respecto a los jefes de hogar afiliados al régimen subsidiado, representan aproximadamente el 42% del total. Esta variable es crucial de analizar, ya que el acceso a este mecanismo por parte de los colombianos para obtener servicios cuando no tienen capacidad de pago, sirve como una medida indirecta de la condición de pobreza de un hogar.

Para revisar la educación de los jefes de hogar, primero se filtran aquellos que no saben o no responden, y se calcula la proporción con respecto al total de personas que reportaron su nivel educativo. En este análisis, observamos que la mayoría de los jefes de hogar tienen al menos aprobada la educación básica secundaria. Esta variable es una de las más importantes para predecir el nivel de ingresos de un hogar. Como se mencionó en la introducción de este informe, a medida que el jefe de hogar obtiene un mayor nivel de educación formal, la probabilidad de ser pobre disminuye, ya que una mayor educación se traduce en mayores niveles de ingresos. Sin embargo, es preocupante la gran cantidad de personas que no reportan su último nivel educativo alcanzado.

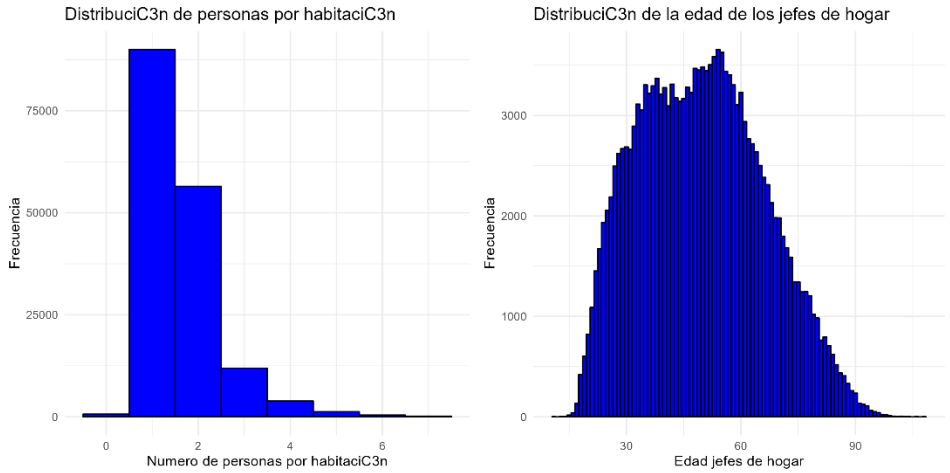


El estado de empleo es fundamental para determinar si un hogar está en condición de pobreza. Esto se explica porque si el jefe de hogar está desempleado, significa que el hogar no está percibiendo ingresos por parte de la persona encargada del sustento familiar, lo cual aumenta la probabilidad de ser pobre. En este caso, observamos que aproximadamente el 37% de los jefes de hogar no están empleados, una cifra considerable que puede explicar la situación del hogar.

La posición del jefe de hogar en el empleo es importante para determinar si el hogar corre riesgo de estar en situación de pobreza. Se espera que los jefes de hogar que trabajen sin remuneración tengan un mayor riesgo de pobreza. Sin embargo, en nuestro análisis no se observa que esta sea la situación predominante, ya que al combinar estas dos categorías apenas llegan al 1%.

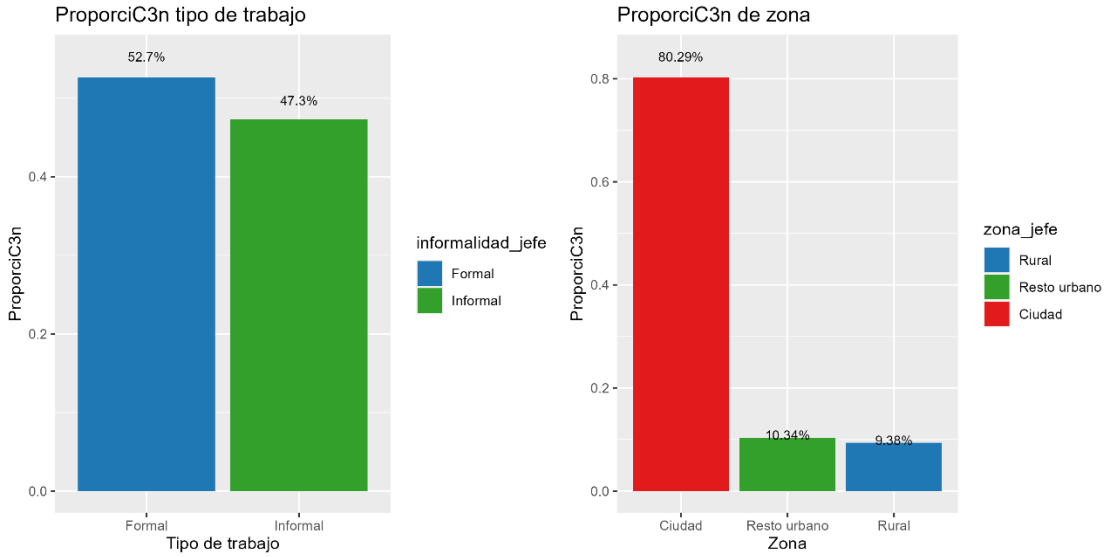
Tener casa propia, en principio, puede ayudar a explicar la situación de pobreza del hogar. Esto se debe a que contar con una vivienda propia permite disponer de una mayor cantidad de ingresos que pueden destinarse a cubrir otras necesidades. La propiedad de una casa se considera una inversión y un seguro en momentos de dificultades económicas. Se supone que las personas que no tienen casa propia podrían ser más vulnerables relativamente a ser pobres. En nuestra muestra, observamos que la mayoría de los jefes de hogar no cuentan con vivienda propia y deben pagar arriendo por su alojamiento. Es fundamental profundizar en esta información para comprender mejor la dinámica de la pobreza en los hogares.

El número de personas por habitación también se considera importante para predecir la pobreza, ya que se espera que a mayor número de personas por habitación, exista un hacinamiento que aumente el riesgo de pobreza. En nuestro análisis, la mayoría de los datos se concentran en 2 personas por habitación, lo cual es una situación normal. Es fundamental incorporar la edad del jefe del hogar en nuestro análisis, ya que está intrínsecamente ligada a la teoría del ciclo de vida en economía, la cual postula que las personas atraviesan diferentes etapas en su vida, cada una con características económicas específicas que pueden influir en su vulnerabilidad a la pobreza.



En el contexto colombiano, esta teoría cobra especial relevancia, ya que tanto los jóvenes como los ancianos son considerados grupos de población vulnerables. Por un lado, los jóvenes tienden a enfrentar mayores dificultades económicas debido a su entrada reciente al mercado laboral y su menor experiencia, lo que puede traducirse en ingresos más bajos y una mayor vulnerabilidad financiera. Este fenómeno se ve exacerbado por las altas tasas de desempleo que suelen afectar a este grupo demográfico. Por lo tanto, es plausible suponer que los jefes de hogar más jóvenes enfrenten desafíos adicionales para alcanzar la estabilidad financiera, lo que podría aumentar su probabilidad de experimentar situaciones de pobreza.

Por otro lado, la teoría del ciclo de vida también reconoce que los ancianos, especialmente después de su retiro, pueden enfrentar dificultades económicas debido a ingresos fijos y limitados, como pensiones o jubilaciones, que pueden no ser suficientes para cubrir todas sus necesidades. A pesar de haber contribuido al mercado laboral durante años, las limitaciones en los ingresos de jubilación pueden dejar a los ancianos en una situación económica precaria, lo que aumenta la probabilidad de que los hogares encabezados por personas mayores estén en situación de pobreza. Si bien en nuestra muestra de datos observamos una distribución uniforme de las edades de los jefes de hogar, la teoría del ciclo de vida nos permite anticipar que los hogares liderados por personas más jóvenes y más mayores podrían enfrentar mayores dificultades económicas, cada uno por razones asociadas a su etapa de vida.



La distinción entre el tipo de empleo a través de la informalidad laboral es un factor crucial para determinar las condiciones de pobreza de un hogar. El trabajo informal se caracteriza por salarios bajos y condiciones laborales precarias, lo que puede aumentar significativamente el riesgo de pobreza para quienes se desempeñan en este sector. En nuestros datos, observamos que aproximadamente el 47% de las personas trabajan en condiciones de informalidad, una cifra que concuerda con los datos nacionales y subraya la magnitud del desafío que representa la informalidad laboral en términos de pobreza.

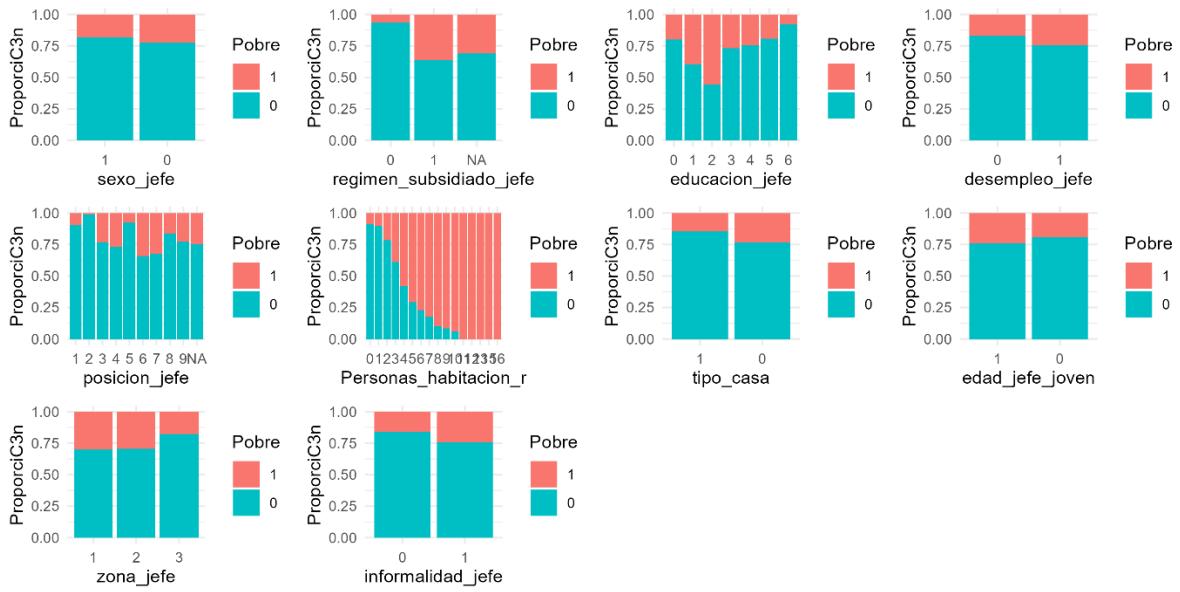
Además, la ubicación geográfica del hogar también desempeña un papel crucial, ya que las viviendas ubicadas en áreas más alejadas del centro económico suelen considerarse más vulnerables. Esta disparidad se explica en parte por la teoría del centro-periferia, que destaca las diferencias en el desarrollo económico entre las áreas urbanas centrales y las regiones periféricas. En el contexto urbano, el centro se caracteriza por un mayor desarrollo económico e industrial, mientras que las áreas rurales tienden a tener un menor desarrollo, con actividades económicas centradas en la agricultura y la extracción de recursos, que generalmente se asocian con salarios más bajos. Aunque en nuestra muestra hay una representación limitada del área rural, este análisis nos brinda una

aproximación útil para evaluar la probabilidad de pobreza en función de la ubicación del hogar y su relación con la teoría del centro-periferia.

En la gráfica anterior, se observan las variables seleccionadas en relación con el logaritmo de los ingresos del hogar. Este enfoque se utiliza para normalizar los datos y facilitar la comparación entre las variables de interés. En términos generales, se puede apreciar que las personas afiliadas al régimen subsidiado tienen ingresos ligeramente inferiores en comparación con aquellos que son cotizantes. Del mismo modo, se evidencia que las personas desempleadas tienden a tener ingresos menores en el hogar. En cuanto al sexo del jefe de hogar, las diferencias en los ingresos no son tan marcadas. Sin embargo, se observan diferencias significativas en los ingresos agregados según el nivel educativo del jefe de hogar y su posición laboral.

Además, se aprecia que a medida que aumenta el número de personas por habitación, la media de ingresos del hogar tiende a disminuir, lo cual coincide con lo reportado en la literatura. Respecto a las variables tipo_casa y edad_jefe_joven, los resultados son coherentes con lo esperado, ya que los jefes de hogar que poseen vivienda propia y son adultos suelen tener mayores ingresos, lo que reduce la probabilidad de estar en situación de pobreza.

El tema de la informalidad laboral también muestra diferencias significativas en los ingresos de las personas. En general, aquellos que trabajan en condiciones de informalidad tienen ingresos más bajos, lo que aumenta su vulnerabilidad económica y, por ende, la probabilidad de estar en situación de pobreza.



Cuando se analiza en relación con la variable de pobreza, se observa un patrón muy similar al analizar los ingresos del hogar. Sin embargo, se destaca que la tenencia de vivienda propia marca una diferencia más significativa que cuando se analiza únicamente por medio de los ingresos. Esto sugiere que la posesión de vivienda propia puede ser un factor más determinante en la reducción de la probabilidad de pobreza, independientemente de los niveles de ingresos del hogar.

3. Modelos y resultados

3.1. Modelos de clasificación

En esta sección se muestran los resultados de los modelos más relevantes encontrados para clasificar como pobres o no pobres a los hogares. La variable objetivo a predecir es *Pobre* y como predictores se utilizaron las variables que se encuentran en Predictores.

Capturamos la data que se encuentra en el repositorio con las diferentes variables que hemos traido a partir de la información de personas y las transformaciones de interés. Para ver mayor detalle del procesamiento de la data se puede consultar el script Creación de variables de interés.R.

Predicciones con Logit.

Inicialmeante se entrena un modelo con todos los predictores que no poseen valores perdidos en los set de entrenamiento y prueba.Inicialmente se hizo un entrenamiento con la data sin Bogotá para utilizar la variable Dominio, sin embargo, al no ser significativa se retiró la variable y trabajamos con el set de entrenamiento completo. Tambien se retiraron las variables que tienen multicolinealidad, estos predictores se pueden consultar en Predictores Logit.

```
logit<-glm(as.formula(paste("Pobre~",paste(predictores_logit, collapse = " + "))),data = train_hogares)
```

Los detalles de los coeficientes, niveles de significancia y métricas de ajuste se pueden observar en summary logit. Se evalua el modelo dentro de muestra:

```
train_hogares$prob_logit<-predict(logit,newdata = train_hogares)
train_hogares$logit_predict<-ifelse(train_hogares$prob_logit>0.5,1,0)

train_hogares$logit_predict<-as.factor(train_hogares$logit_predict)
cf_logit<-confusionMatrix(data = train_hogares$logit_predict,
reference = train_hogares$Pobre, positive = "1",mode="prec_recall")
cf_logit$byClass["F1"]
##                                                                                  F1
## 0.2889066
```

Los detalles de la salida de la matriz de confusion se encuentran en matriz de confusión modelo Logit. Se realiza la predicción fuera de muestra y se genera el submission para Kaggle utilizando la sintaxis que se encuentra en Submission Logit. También se genera un modelo con las interacciones entre las variables predictores para interacciones. Evaluamos el modelo dentro de muestra.

```
logit2<-glm(as.formula(paste("Pobre~",paste(interacciones, collapse = " * "))),
data = train_hogares)

train_hogares$prob_logit2<-predict(logit2,newdata = train_hogares)
train_hogares$logit_predict2<-ifelse(train_hogares$prob_logit2>0.5,1,0)
summary(train_hogares$prob_logit2)

##                 Min.    1st    Qu.        Median                Mean    3rd    Qu.                 Max.
## 0.02374 0.13061 0.17777 0.20019 0.27360 0.43697
```

La probabilidad maxima que predice la interacción entre estas variables es inferior a 0.5 por lo que no se tomó en cuenta para hacer predicciones fuera de muestra.

Predicciones con árbol de clasificación

Para entrenar el árbol se utilizó la siguiente sintaxis.

```
arbol<-rpart(formula = as.formula(paste("pobre_texto~",paste(predictores, collapse = " + "))),
data=train_hogares, method = "class",parms = list(split="Gini"))
prp(arbol,box.col = "gray")

train_hogares$predic_arbol<-predict(arbol,newdata =train_hogares, type = "class")
cf_arbol<-confusionMatrix(data = train_hogares$predic_arbol, reference = train_hogares$pobre_texto,
positive="Pobre", mode = "prec_recall")
cf_arbol$byClass["F1"]

##                                                                                  F1
## 0.3313086
```

Se evaluan los resultados dentro de muestra, el detalle de la matriz de confusión se encuentra en matriz de confusión árbol de clasificación. Se realiza la predicción fuera de muestra y se genera el submission para Kaggle. La sintaxis utilizada para generar el submission se encuentra en Submission Árbol.

```
test_hogares$predic_arbol<-predict(arbol,newdata =test_hogares,type = "class")
table(test_hogares$predic_arbol)
## No_Pobre    Pobre
##    61469     4699
```

Predicciones con Random Forest

El algoritmo de Random Forest no nos permite utilizar variables con valores perdidos por lo que se excluyen de los predictores algunas variables que no aplican para toda la muestra. Los predictores utilizados para entrenar el modelo se pueden consultar en Predictores Random Forest. Se realiza un entrenamiento inicial de un modelo de Ramdom Forest utilizando la siguiente sintaxis.

```
RF<- ranger(formula = as.formula(paste("pobre_texto~",paste(predictores, collapse = " + "))), data =
train_hogares,
            num.trees= 500, ## Numero de arboles a estimar
            mtry= 4,    # N. var aleatoriamente seleccionadas en cada partición.
            min.node.size  = 1, ## Numero minimo de observaciones en un nodo
            importance="impurity")
## [1] "OBB prediction error"

## [1] 0.1691319
```

Para definir la cantidad optima de variables seleccionada por cada partición se prueba aumentar la cantidad de arboles para identificar que hay una reducción importante del OOB predictor error.

```
RF1000<- ranger(formula = as.formula(paste("pobre_texto~",paste(predictores, collapse = " + "))),data =
train_hogares, num.trees= 1000, ## Aumentamos de 500 a 1000
            mtry= 4, min.node.size  = 1,
            importance="impurity")

## OOB prediction error:             16.95 %
```

Se observa que no hay reducción importante del OBB predictor error. Para continuar de afinar hiperparametros del modelo aumentamos la cantidad de minima de observaciones por nodo.

```
RF_NODE100<- ranger(formula = as.formula(paste("pobre_texto~",paste(predictores, collapse = " + "))),data =
train_hogares, num.trees= 500, mtry= 4,
 min.node.size  = 100, #Aumentamos de 1 a 100
importance="impurity")

## OOB prediction error:             16.92 %
```

Vemos que tampoco existe un cambio importante en el OBB predictor error. Finalmente probamos aumentar la cantidad de variables por árbol.

```
RF_iVAR<- ranger(formula = as.formula(paste("pobre_texto~",paste(predictores, collapse = " + "))), data =
train_hogares, num.trees= 500, mtry= i, #i=c(5,6,7)
          min.node.size  = 1, importance="impurity")
## [1] "OBB predictor error"
## [1] 0.1703322
## [1] 0.1721266
## [1] 0.1749333
```

Observamos que después de cinco variables por árbol el OBB predictor error deja de reducirse. Por lo que se toma la decisión de trabajar con el primer árbol. Podemos observar también la importancia que tienen las variables en el arbol para identificar que algunas cobran mayor relevancia en comparación de un arbol sencillo. Analizamos las predicciones realizadas dentro de muestra. El detalle de la matriz de confusión se encuentra en matriz de confusión de modelo Ramdom Forest.

```
cf_rf<-confusionMatrix(data = RF$predictions, reference = train_hogares$pobre_texto, positive = "Pobre",

mode = "prec_recall")
cf_rf$byClass["F1"]
   F1
## 0.4414414
```

Predecimos fuera de muestra y generamos el submission para Kaggle. Podemos consultar la sintaxis con la que se dió formato a las predicción en formato de submission con Random Forest.

3.2 Modelos de regresión del ingreso

En esta sección se muestran los resultados de los modelos más relevantes encontrados para predecir el ingreso con el objetivo de apartir de estas predicciones poder clasificar como pobres o no pobres a los hogares, tomando el umbral de pobreza como límite de clasificación entre los dos grupos de hogares.

Predicciones con regresión lineal

Para realizar la predicción del ingreso de los hogares utilizamos como variable dependiente el ingreso total de la unidad de gasto con impotación de arriendo a propietarios y usufructuario (*Ingtotugarr*). Las variables predictoras se pueden consultar en predictores para regresión líneal. Para correr el modelo se utilizó la siguiente sintaxis:

```
modelo_lm<-lm(as.formula(aste("Ingtotug~",paste(predictores,      collapse      =      "      +      "))),
  data =train_hogares_sin_bogota)
```

Los detalles de la salida de regresión se pueden consultar en summary modelo de regresión lineal. Realizamos las predicciones en el conjunto de entrenamiento para poder clasificar, la tabla derivada de la matriz de confusión se puede consultar completa en matriz de confusión de modelo de regresión lineal.

```
train_hogares_sin_bogota$modelo_lm_ingreso<-predict(object = modelo_lm, newdata = train_hogares_sin_bogota)
#Clasificamos los hogares en pobres o no según el ingreso predicho
train_hogares_sin_bogota$modelo_lm_predict<-
ifelse(train_hogares_sin_bogota$modelo_lm_ingreso<train_hogares_sin_bogota$Lp,1,0)

##        F1
## 0.1371305
```

Luego realizamos la predicción en el conjunto de prueba y exportamos con el formato adecuado para realizar submitir en Keaggle. La sintaxis utilizada para esto se encuentra en formato para submission con regresión lineal. Posteriormente se realizan diferentes combinaciones de variables predictivas para encontrar aquellas que tengan una mejor capacidad de predicción.

```
sub1<-test_hogares %>% select(id,modelo_lm_predict)
sub1<-sub1 %>% rename(pobre=modelo_lm_predict)
write_csv(x = sub1,"C:/Users/HP-Laptop/Documents/GitHub/Curso-Big-Data/Taller
2/2.Entregables/Submission1.csv",)
```

Predicciones con Regularización

Para realizar estos modelos utilizamos solo las variables que no poseen NA en train ni en test, estos se pueden consultar en predictores regularización. Para aplicar regularización tenemos que trabajar con nuestros datos en formato matriz, la sintaxis utilizada también se puede observar en transformación en matriz para regularización.

Ridge, Lasso y Elastic Net

Para entrenar los modelos fue necesario afinar los parámetros lambda utilizados para la predicción. En este caso, el parámetro lambda mínimo fue encontrado mediante cross validation utilizando la siguiente sintaxis:

```
cv_ridge <- cv.glmnet(x = X, y = Y,alpha =i)
# Para Ridge i=1, Lasso i=0 Lasso, Elastic Net i=0.75.
coef(cv_ridge, s = "lambda.min")
test_hogares$predict_ingreso_rige<-predict(cv_ridge, newx = X_test, s = "lambda.min")
```

Elastic Net con Cartet

Para identificar los hiperparametros optimos de Elastic Net con Caret se utilizó la siguiente sintaxis.

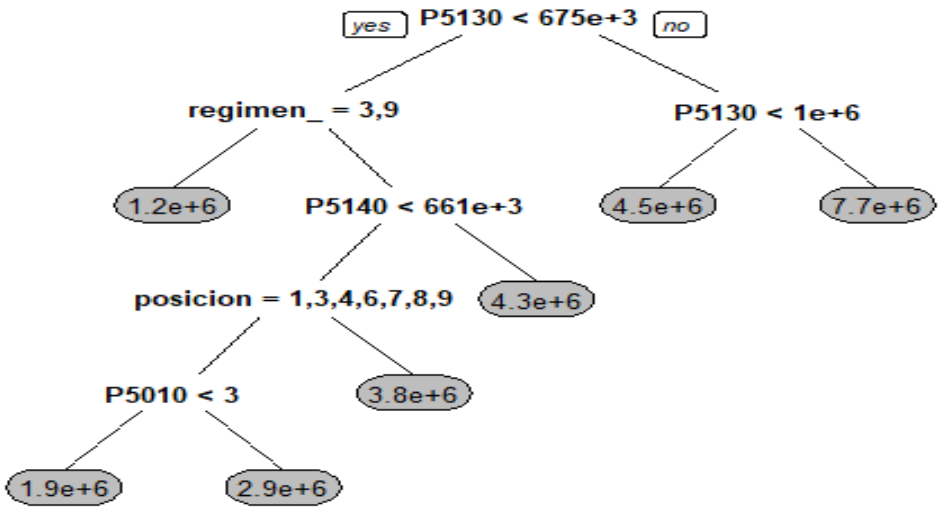
```
tc_10 <- trainControl(method = "cv", number = 10)

en_caret <- train(x=X,y=Y,method = "glmnet",trControl = tc_10, tuneLength=100)
```

Los resultados de los modelos con Regularización no fueron tenidos en cuenta para realizar submissions ya que la cantidad de pobres predicha con estos fue muy pequeña en comparación con el resto de algoritmos.

Predicciones con árbol de regresión

Para predecir el ingreso contamos con mas flexibilidad al momento de utilizar variables con NA, estas variables se pueden consultar en predictores para árbol de regresión . Para realizar la predicción con árboles de decisión se realizó la siguiente sintaxis

```
arbol_regresion<-rpart(formula = as.formula(paste("Ingtotugarr~",paste(predictores, collapse = " + "))),
data=train_hogares_sin_bogota, parms = list(split="Gini"))
prp(arbol_regresion,box.palette = "gray")
```



Luego de entrenar el modelo realizamos la predicción en train y evaluamos su resultado dentro de muestra. La totalidad de las métricas derivadas de la matriz de confusión se encuentran en matriz de confusión árbol de regresión.

```
train_hogares_sin_bogota$predic_arbol_ingreso<-predict(arbol_regresion,
```

```
 newdata =train_hogares_sin_bogota)

train_hogares_sin_bogota$predict_arbol<-
ifelse(train_hogares_sin_bogota$predic_arbol_ingreso<train_hogares_sin_bogota$Lp*train_hogares_sin_bogota$Npe
rsug,1,0)

train_hogares_sin_bogota$predict_arbol<-as.factor(train_hogares_sin_bogota$predict_arbol)

cf_arbol_reg<-confusionMatrix(data =train_hogares_sin_bogota$predict_arbol ,
reference=train_hogares_sin_bogota$Pobre, positive="1", mode = "prec_recall")
cf_arbol_reg$byClass["F1"]
##                                                                                                        F1
## 0.3388104
```

Luego realizamos la predicción fuera de muestra:

```
test_hogares$predic_arbol_ingreso<-predict(arbol_regresion, newdata =test_hogares)

test_hogares$predict_arbol_regresion<-ifelse(test_hogares$predic_arbol_ingreso<
    test_hogares$Lp*test_hogares$Npersug,1,0)
table(test_hogares$predict_arbol_regresion)
##     0     1
## 59622  6546
```

3.3 Selección del mejor modelo

Para la selección del mejor modelo se compararon los resultados de las submussions subidas a Kaggle y se trabajó en el perfeccionamiento del modelo que mejor puntaje F1 obtuvo en la competencia, en este caso fue el Random Forest con 0.45. La estrategia utilizada para mejorar la predicción fuera de muestra de este modelo fue realizar un balanceo del conjunto de entrenamiento utilizando smote. El modelo entrenado con un dataset balanceado obtuvo un mejor rendimiento fuera de muestra, logrando un puntaje 0.52. Este modelo se puso a competir con otros algortmos mas robustos como Xgboots, sin embargo, no se evidenció un aumento del puntaje fuera de muestra con estos modelos.

4. Conclusión Final

El proyecto finalizó con un enfoque robusto en la aplicación de modelos avanzados de aprendizaje automático para predecir la pobreza en hogares colombianos. Iniciamos con una regresión Logit, que sirvió como base para entender las dinámicas y las variables más influyentes en la pobreza. A partir de ahí, el proceso evolucionó hacia técnicas más sofisticadas, incluyendo árboles de decisión y Random Forest. En la fase final del proyecto, implementamos un modelo de Random Forest utilizando la técnica SMOTE (Synthetic Minority Over-sampling Technique) para abordar el desbalance de clases entre hogares pobres y no pobres en nuestro conjunto de datos. Este método permitió generar observaciones sintéticas de la clase minoritaria, proporcionando un entrenamiento más equilibrado y efectivo para nuestro modelo. El modelo ajustado alcanzó una puntuación F1 de 0.8111, indicando un excelente equilibrio entre la precisión y la sensibilidad, lo que refleja una alta capacidad para identificar correctamente tanto a los hogares pobres como a los no pobres.

Este modelo fue publicado en la competición de Kaggle, donde obtuvimos un score de 0.52. Este resultado puede interpretarse como una moderada efectividad del modelo en la plataforma de Kaggle, considerando las complejidades y las variaciones inherentes en los datos de test reales y competitivos. La puntuación F1 alta sugiere que el modelo es bastante robusto en términos de manejar el desequilibrio de clases y efectivo en la clasificación binaria en un escenario controlado, mientras que el score de Kaggle nos proporciona una perspectiva realista de cómo el modelo podría performar en escenarios externos y con datos que podrían diferir ligeramente de nuestro conjunto de entrenamiento.

Se puede concluir que, el uso de Random Forest con SMOTE demostró ser la estrategia más eficaz entre los métodos evaluados, destacándose en la capacidad de manejar grandes volúmenes de datos y capturar interacciones complejas entre las variables sin necesidad de ajustes manuales extensos. Esto subraya la importancia de técnicas de muestreo adecuadas y la selección de modelos en el análisis de datos socioeconómicos.

Link al repositorio de GitHub: Curso-Big-Data/Taller 2 at main · Esteban7777/Curso-Big-Data (github.com)

Referencias

- Sánchez Torres, R. M., Manzano Murillo, L. D., & Maturana Cifuentes, L. A. (2022). Informalidad laboral, pobreza monetaria y multidimensional en Bogotá y el Área Metropolitana. Problemas del Desarrollo. Revista Latinoamericana de Economía, 53(208), 31-56. https://doi.org/10.22201/iiec.20078951e.2022.208.69754

- Marí-Klose, P., & Marí-Klose, M. (2012). Edad, vulnerabilidad económica y Estado de bienestar: La protección social contra la pobreza de niños y personas mayores. Panorama Social, (15), 107-117.

- Núñez, J., & Ramírez, J. C. (2002). Determinantes de la pobreza en Colombia. Años recientes. Naciones Unidas, CEPAL - Serie Estudios y Perspectivas. Web.

- Hernández, M. C., Cingolani, J., & Chaves, M. (2015). Espacios con edades: El barrio y la pobreza desde los niños y los jóvenes. Web.

- Del Risco Serje, K. P., & Martelo Amaya, J. E. (2015). Determinantes de la pobreza en la región Caribe Colombiana. Universidad de Cartagena, Facultad de Economía, Programa de Economía. Cartagena de Indias, D.T. y C.