

Trabajo 1 - Técnicas en Aprendizaje Estadístico

Predicción del número de hijos de los hogares colombianos

April 4, 2021

Integrantes

- Santiago Franco Valencia
- Isabela Lujan Jaramillo
- Ana María Sánchez Henao
- Daniel Alexander Naranjo Ríos
- Stephany Michell Lobo Laguado

Objetivo:

El objetivo de este trabajo es predecir la cantidad de hijos en un hogar colombiano a partir de características específicas, partiendo de la Encuesta Nacional de Calidad de Vida - ECV 2019, publicada por el Departamento Administrativo Nacional de Estadística DANE.

Introducción:

Las encuestas dirigidas a los distintos hogares constituyen una de las principales fuentes de datos socioeconómicos con las que cuentan los países. A partir de la información obtenida de ellas, es posible calcular indicadores para la medición de variados aspectos económicos y sociales; además, facilitan el conocimiento y explicación de los determinantes o factores causales del comportamiento de dichos aspectos, lo cual es de gran importancia para el diseño, monitoreo y medición de resultados de las políticas públicas.” (DANE, 2018)

Por ello, estas encuestas son herramientas muy usadas actualmente por el gobierno, aprovechando la alta capacidad que tienen para recoger grandes cantidades de información rápidamente, sobre todo cuando dicha información proviene de la percepción que tiene el pueblo o las condiciones en las que cada ciudadano vive. Para ello, se crearon instituciones como el el Departamento Administrativo Nacional de Estadística DANE, que se encarga de planear, recopilar, procesar, analizar y difundir las estadísticas oficiales referentes a información de interés sobre varias situaciones actuales del país y su población.

Contextualización de la encuesta:

Las encuestas a hogares normalmente se centran en temas específicos como, por ejemplo, la Encuesta de Mercado Laboral (Gran Encuesta Integrada de Hogares) que se aplica en forma regular y continua durante todo el año. Otro ejemplo es la Encuesta Nacional de Presupuestos de los Hogares que se aplica cada diez años. Si bien, estas encuestas indagan sobre algunos aspectos que permiten hacer análisis particulares del bienestar, no brindan información que posibilite conocer íntegramente las diferentes variables que determinan las condiciones de vida del hogar en todas sus dimensiones.” (DANE, 2018)

Estas encuestas incluso abarcan temas como la calidad de vida de los colombianos, como lo hace la Encuesta Nacional de Calidad de Vida - ECV, la cual es realizada por el DANE cada año aproximadamente; la información más reciente publicada por el DANE sobre esta encuesta fue la base de donde se recopiló, depuró y analizó información para el desarrollo de este documento.

Problemas encontrados:

Inicialmente, se debían recopilar los datos, presentes en la página del DANE, que correspondían a información sobre la Encuesta Nacional de Calidad de Vida - ECV 2019.

El problema inicial, se encontró al momento de realizar una exploración general de los datos por la forma en la que estos se encontraban distribuidos; la Encuesta Nacional de Calidad de Vida - ECV, fue una encuesta realizada por medio de muestreo trietápico, considerando municipios del país, diferentes sectores del municipio y segmentos de cada sector como unidades de muestreo. Adicionalmente, se quería obtener información sobre distintos temas tales como salud, educación, situación económica y otras características para cada hogar dentro de la muestra, por lo que la información suministrada se encontraba dividida de acuerdo a los distintos agentes del hogar que fueron encuestados; se encontraron datos correspondientes a la vivienda de la cual hacía parte el hogar, datos para cada uno de estos hogares, y además, datos sobre cada una de las personas que hacían parte de los hogares dentro de la muestra.

Al tener la información dividida de esta manera, se debía buscar alguna forma en que la información sobre la cuál se hiciera un análisis, se encontrara en términos de cada hogar dentro de la muestra, que era la unidad de estudio que se reguería.

Para obtener una predicción sobre el número de hijos en un hogar, se requería el uso de alguna metodología de modelos de regresión, en los cuales se requieren ajustar modelos, por medio del método de validación cruzada. Para esto se necesitaba darle al modelo un conjunto de variables que dieran información sobre el hogar, las cuales actuarían como predictores de la variable respuesta, que sería el número de hijos en cada uno de los hogares del conjunto de muestra.

Debido a que los datos proporcionados por la encuesta tenían información tanto para la vivienda, el hogar y las personas el hogar, se necesitaba definir un conjunto de variables predictoras a partir de estos datos que dieran información sobre cada uno de los hogares considerados en la muestra.

Otro problema con los datos era que, para utilizar validación cruzada, se requería que el conjunto de datos también tuviera a la variable respuesta dentro de él, es decir, se necesitaba una variable que diera cuenta del número de hijos presentes en cada uno de los hogares del conjunto de muestra.

Esta variable no estaba explícitamente definida dentro de los datos proporcionados en la encuesta por lo que debía ser inferida por medio de otra información.

Variables a considerar:

La información recolectada por la Encuesta Nacional de Calidad de Vida - ECV 2019 se encuentra conformada por 14 bases de datos en total. Para la selección de variables se realizó un estudio riguroso de las variables a considerar teniendo en cuenta características específicas del hogar y del jefe de hogar.

La variable inicial a considerar en el conjunto de datos fue el numero de hijos para cada hogar, la cual, como se dijo anteriormente, no estaba definida de manera explícita, por lo tuvo que construirse a partir de otra información que sí era suministrada.

Debido a que las bases de datos tenían información sobre viviendas, hogares y personas, se debían considerar los datos dieran información sobre los hogares dentro de la muestra. Se tomo inicialmente el conjunto de datos de “Servicios del hogar”, el cual contenía 93993 registros y 57 variables con información sobre cómo eran adquiridos en el hogar los servicios básicos de agua, luz y electricidad, así como diversa información asociada a estos servicios (tipo de servicio, clasificación de desechos, etc).

Con estos datos, ya se sabía entonces que el número de hogares a considerar era de 93993, y con estos datos como base, se añadieron algunas variables que se consideraron podían llegar a tener relación con el número de hijos.

Estas variables añadidas fueron tomadas de los conjuntos de datos " Datos de Vivienda", "Características y composición del hogar", "Uso de energéticos del hogar" y "Fuerza de trabajo", tomando para cada hogar, es decir, el registro correspondiente a cada jefe del hogar, la siguiente información:

1. **Región** (1 Caribe, 2 Oriental, 3 Central, 4 Pacífica(sin valle), 5 Bogotá, 6 Antioquia, 7 Valle del cauca, 8 San Andrés, 9 Orinoquía - amazonía).
2. **Tipo de vivienda** (1 Casa, 2 Apartamento, 3 Cuarto(s), 4 Vivienda tradicional indigena, 5 Otro (carpa, contenedor, vagón, embarcación, cueva, refugio natural, etc)).
3. **Energía eléctrica** (0 - no, 1 - si).
4. **estrato de recibo de electricidad** (1 Bajo - Bajo, 2 Bajo, 3 Medio - Bajo, 4 Medio, 5 Medio - Alto, 6 Alto, 8 Planta eléctrica, 9 No conoce el estrato o no cuenta con recibo de pago, 0 Recibos sin estrato o el servicio es pirata).
5. **acueducto** (0 - no, 1 - si).
6. **alcantarillado** (0 - no, 1 - si).
7. **vivienda_ocupada** *¿La vivienda ocupada por este hogar es?* (1 Propia - totalmente pagada, 2. Propia - la están pagando, 3 En arriendo o subarriendo, 4 Con permiso del propietario - sin pago alguno (usufructuario), 5 Posesión sin título (ocupante de hecho), 6 - Propiedad colectiva).
8. **arriendo** *¿cuánto pagan mensualmente por arriendo?*
9. **num_computadores** *Número de computadores en el hogar.*
10. **fisico_t** *el trabajo del jefe del hogar exige mucho esfuerzo físico* (0 - no, 1 - si).
11. **intelectual_t** *el trabajo del jefe del hogar exige mucho esfuerzo intelectual* (0 - no, 1 - si).
12. **conyuges** (0 No tiene, 1 Si vive en el hogar, 2 No vive en el hogar)
13. **padre_hogar** *padre de los hijos vive en el hogar* (1 si, 2 no, 3 fallecido).
14. **educacion_padre** *educación del padre* (1 primaria incompleta, 2 primaria completa, 3 secundaria incompleta, 4 secundaria completa, 5 técnica incompleta, 6 técnica completa, 7 universidad incompleta, 8 universitaria completa, 9 Ninguno, 10 No sabe).
15. **madre_hogar** *madre de los hijos vive en el hogar* (1 si, 2 no, 3 fallecida).
16. **educacion_madre** *educación de la madre* (1 primaria incompleta, 2 primaria completa, 3 secundaria incompleta, 4 secundaria completa, 5 técnica incompleta, 6 técnica completa, 7 universidad incompleta, 8 universitaria completa, 9 Ninguno, 10 No sabe).
17. **cultura** (1 Indígena, 2 Gitano (a) (Rom), 3 Raizal del archipiélago de San Andrés, Providencia y Santa Catalina, 4 Palenquero (a) de San Basilio, 5 Negro (a), mulato (a) (afrodescendiente), afro-colombiano(a), 6 Ninguno de los anteriores).
18. **es_campesino** (1 Si, 2 No, 3 No informa).
19. **condicion_vida** *¿En cuál escalón diría usted que se encuentra parado(a) en este momento?* (10 Mejor vida, 9, 8, 7, 6, 5, 4, 3, 2, 1, 0 Peor vida).

Nota: La unión de los datos se hizo siguiendo el documento guía suministrado en la página de la encuesta, en el que se daba información sobre las variables llave que unían las distintas tablas de viviendas, hogares y personas.

¿Cómo se encontró la variable respuesta *y*?

De la base de datos “Características y composición del hogar” se tomó la variable “P6051”, en la que se le preguntó a cada uno de los miembros del hogar “¿cuál es la relación directa con el jefe(a) de hogar?”.

Las personas que eran hijos del jefe del hogar, tenían un valor de 3 en la columna de esta variable; se filtraron los datos de las personas que eran hijos, es decir, los valores de 3 en la variable “P6051” para todos los hogares, y se realizó un conteo de las observaciones que se repetían para cada hogar, a partir de un índice único creado para distinguir cada hogar; este índice único se creó pegando los valores de las variables que funcionaban como llave para una las bases de datos que contenían información de los hogares.

Este conteo se guardó en otra variable, que correspondía al número de hijos, pero como estos datos iban a estar para cada una de las personas del hogar por estar en una base de datos de personas, se encontraban repetidos. Para proyectar esta información en términos de los hogares, se unió esta variable, identificada con el índice único creado para cada hogar, con la tabla de “Servicios del hogar”, la cual contenía información para cada uno de los hogares y no de las personas.

Al tener la variable de hijos dentro de la tabla de “Servicios del hogar”, simplemente se unió a la base de datos construida con las demás variables anteriormente definidas.

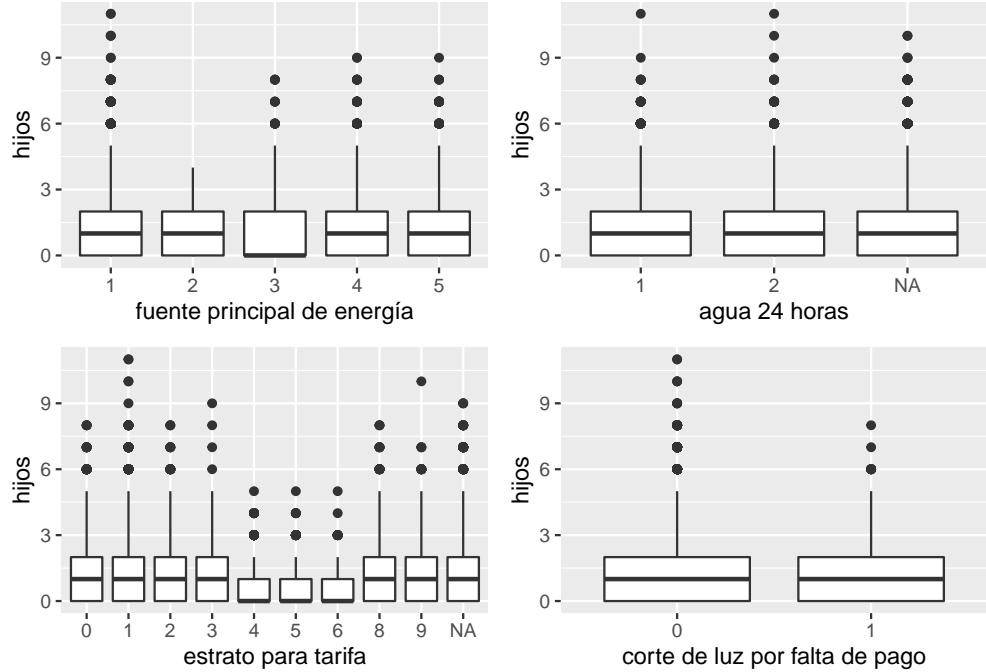
Análisis descriptivo :

El primer filtro para descartar algunas de las variables consiste en graficar boxplots para variables categóricas y gráficos de dispersión para las variables numéricas. A continuación se presentan algunos gráficos usados para descartar variables no relacionadas con el número de hijos del hogar:

Variables categóricas:

Se encontraron muchas variables que no son significativas a la hora de graficarlas ya que su media permanecía constante al número de hijos y se pudo comprobar que no aportaban mucha información, un ejemplo de estos gráficos se encuentra a continuación:

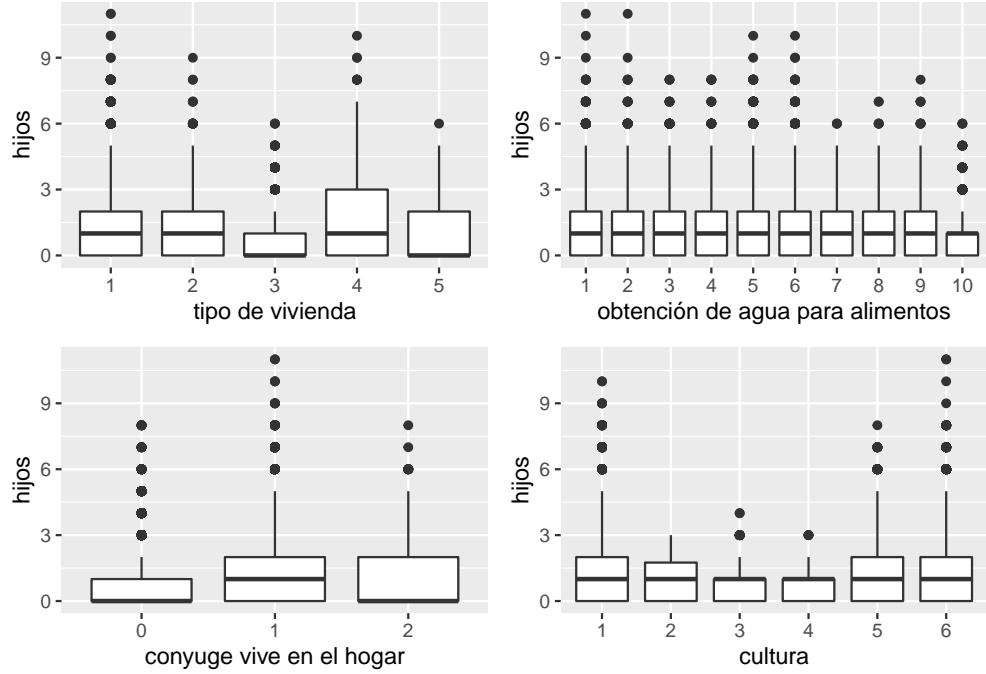
```
##### gráficos juntos #####
fuente principal de energia
a <- ggplot(aes(y = hijos, x = ilum_ppal), data = datos) + geom_boxplot() + xlab("fuente principal de energia")
# agua 24 horas
b <- ggplot(aes(y = hijos, x = agua_24h), data = datos) + geom_boxplot() + xlab("agua 24 horas")
# estrato
c <- ggplot(aes(y = hijos, x = estrato), data = datos) + geom_boxplot() + xlab("estrato para tarifa")
# corte luz
d <- ggplot(aes(y = hijos, x = corte_energia_fp), data = datos) + geom_boxplot() +
    xlab("corte de luz por falta de pago")
grid.arrange(a, b, c, d)
```



En estos gráficos se observa que para variables como fuente principal de energía, si la vivienda tenía agua las 24 horas del día, estrato por tarifa y algún corte de luz en los últimos 30 días, no aportan mucha información para el número de hijos, contrario a lo que se pensaba, pues son variables directamente relacionadas con la calidad de vida.

Por otro lado, también se encontraron variables que aunque su variación no es mucha, se consideró prudente dejarlas para la implementación del modelo, pues se llegó a la conclusión que eran importantes a la hora de evaluar el número de hijos. Alguna de estas variables son:

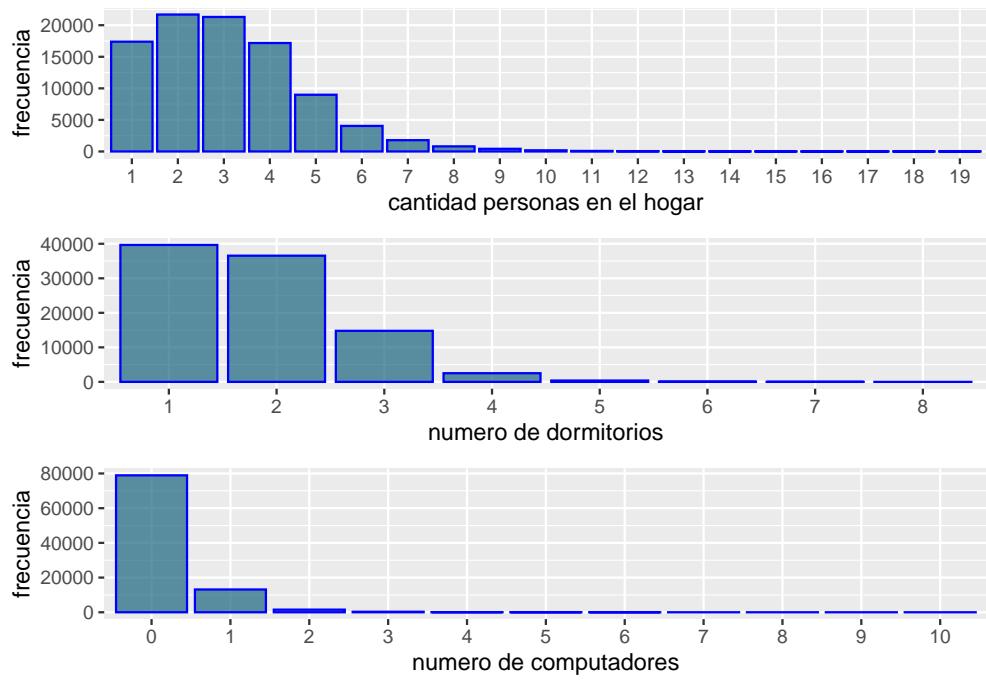
```
# boxplot tipo de vivienda
e <- ggplot(aes(y= hijos, x =tipo_vivienda), data=datos) + geom_boxplot() + xlab("tipo de vivienda")
# obtención agua
f <- ggplot(aes(y= hijos, x = obtencion_agua_alimento), data=datos) + geom_boxplot() + xlab("obtención")
# conyugues
g <- ggplot(aes(y= hijos, x = conyuges), data=datos) + geom_boxplot() +
  xlab("conyuge vive en el hogar")
# cultura
h <- ggplot(aes(y= hijos, x =cultura), data=datos) + geom_boxplot()
grid.arrange(e,f,g,h)
```



Cabe recalcar que, aunque se piense que son necesarias, por medio de métodos de selección de variables es que se decidió finalmente qué tan significativas eran algunas variables o si efectivamente se debían retirar para hacer el ajuste del modelo.

Variables numéricas:

```
# cantidad personas
i <- ggplot(datos, aes(x=as.factor(CANT_PERSONAS_HOGAR) )) +
  geom_bar(color="blue", fill=rgb(0.1,0.4,0.5,0.7) ) +
  xlab("cantidad personas en el hogar") + ylab("frecuencia")
# número dormitorios
j <- ggplot(datos, aes(x=as.factor(num_dormitorios) )) +
  geom_bar(color="blue", fill=rgb(0.1,0.4,0.5,0.7) ) +
  xlab("numero de dormitorios") + ylab("frecuencia")
# número computadores
k <- ggplot(datos, aes(x=as.factor(num_computadores) )) +
  geom_bar(color="blue", fill=rgb(0.1,0.4,0.5,0.7) )+
  xlab("numero de computadores") + ylab("frecuencia")
grid.arrange(i,j,k)
```



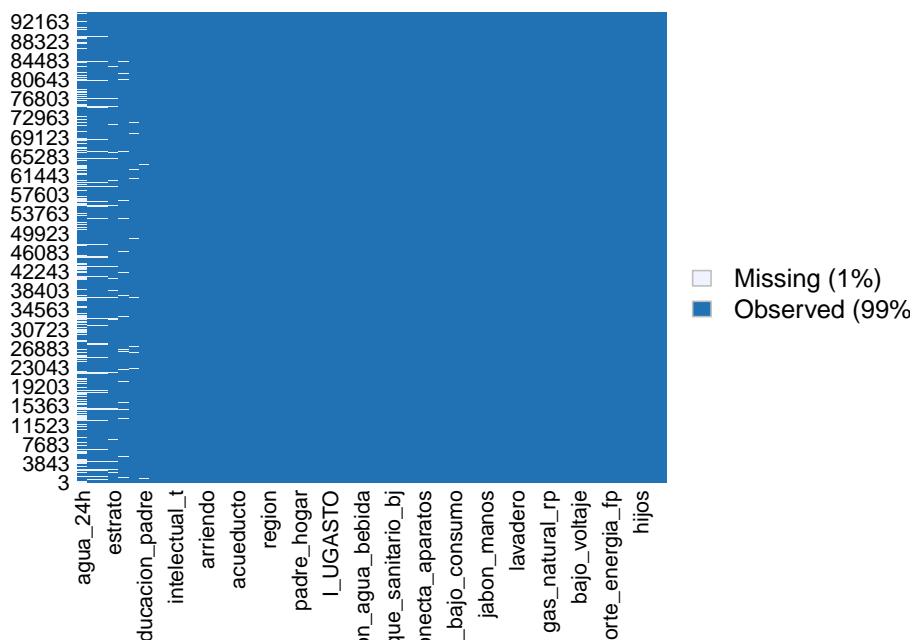
La mayoría de hogares están conformados por entre dos y tres personas; además, se encontró que el número de cuartos más recurrente es uno, y al preguntar por cantidad de computadores en el hogar, los gráficos indican que la mayoría respondieron entre no tiene o solo tiene uno en el hogar.

Valores faltantes:

Con ayuda del siguiente gráfico se detectaron los valores faltantes que se encontraban en la base de datos:

```
missmap(datos, main = "Valores faltantes vs. Observados")
```

Valores faltantes vs. Observados



Se puede observar que no hay gran presencia de valores faltantes en relación con la cantidad total de datos; además, otras variables en las que se encontraron valores faltantes fueron:

- ubicación del servicio sanitario (dentro o fuera del hogar)
- educación de la madre
- condición de vida
- tipo energía
- estrato
- es campesino

La manera de resolver los datos faltantes consistió en hacer un reemplazo en los valores faltantes con la moda correspondiente de cada columna (por tratarse de variables categóricas en su mayoría) ya que como se observa en el gráfico, los valores faltantes corresponden solo al 1% de los datos en total.

Selección de variables:

Aunque el análisis descriptivo dio una idea intuitiva de cuáles variables podrían ser seleccionadas en el modelo, se decidió que era importante requerir de una prueba específica para la selección de variables.

El método empleado para la selección de variables fue por medio de la **selección univariada**, el cual se basa en la prueba estadística chi-cuadrado para hacer una selección que devuelve las 10 variables más relacionadas con la variable respuesta. Información y guía

En este caso, las variables que tuvieron una relación fuerte con los hijos fueron:

Variable	Puntaje
I_UGASTO	2095865000
I_HOGAR	2067919000
arriendo	137660600
CANT_PERSONAS_HOGAR	58415.41

Variable	Puntaje
num_dormitorios	9660.812
tipo_serv_sanitario	2351.219
conyuges	2028.149
num_cuartos	14161.81
obtencion_agua_alimento	1343.717

Finalmente, estas fueron las 10 variables seleccionadas para la modelación.

Nota: este método fue empleado en Python.

La definición explícita de las variables utilizadas en la modelación se encuentra a continuación:

Variable respuesta:

hijos Se refiere al número de hijos del jefe de cada hogar.

Variables predictoras:

1. **I_HOGAR** ingreso mensual del hogar.
2. **I_UGASTO** gasto mensual del hogar.
3. **arriendo** arriendo mensual que se paga en el hogar.
4. **CANT_PERSONAS_HOGAR** cantidad de personas en el hogar.
5. **num_dormitorios** número de habitaciones en las que duermen las personas del hogar.
6. **num_cuartos** número de habitaciones del hogar.
7. **tipo_serv_sanitario** 1 Inodoro conectado a alcantarillado, 2 Inodoro conectado a pozo séptico, 3 Inodoro sin conexión, 4 Letrina, 5 Inodoro con descarga directa a fuentes de agua (bajamar), 6 No tiene servicio sanitario.
8. **conyuges** 0 No tiene, 1 Si vive en el hogar, 2 No vive en el hogar.
9. **obtencion_agua_alimento** 1 Acueducto público, 2 Acueducto comunal o veredal, 3 Pozo con bomba, 4 Pozo sin bomba-aljibe-jagüey-barreno, 5 Agua lluvia, 6 Río-quebrada-manantial-nacimiento, 7 Pila pública, 8 Carro tanque, 9 Aguatero, 10 Agua embotellada o en bolsa.

Materiales

Se utilizó R y R studio para el preprocesamiento de los datos, así como también para análisis descriptivos, modelación y escritura del reporte por medio de la herramienta R Markdown.

Adicionalmente, se usó del lenguaje de programación Python por medio de Google Collab para realizar la selección de variables.

Métodos

El siguiente paso era entrenar un modelo que pudiera predecir la cantidad de hijos por hogar. Inicialmente se usó la técnica de **validación cruzada**, la cual consiste en realizar una partición del total de los datos en subconjuntos que se utilizan más adelante para entrenar y evaluar el modelo predictivo.

Un 75% de los datos fue utilizado para realizar el entrenamiento del modelo, y el 25% restante se utilizó para evaluar el comportamiento de este.

```

# base_nueva <- datos %>%
#   select(I_UGASTO, I_HOGAR, arriendo, CANT_PERSONAS_HOGAR,
#         num_dormitorios, tipo_serv_sanitario, conyuges, num_cuartos, obtencion_agua_alimento, hijos)
#
# train.set = base_nueva[index,]
# test.set = base_nueva[-index,]

```

Debido a que la variable respuesta, que es el número de hijos de un hogar, se trata de una variable que toma valores discretos, se decidió abordar la problemática por medio de un método de clasificación.

Inicialmente se pensó en utilizar el método de K-Nearest Neighbors (kNN), pero este fue descartado ya que se encontró que, cuando hay presencia de variables categóricas dentro de los predictores, este método puede conducir a resultados erróneos debido a la dificultad de medir “distancias”. link

Los modelos considerados fueron los siguientes:

Modelo 1: Árbol de decisión

Se consideró el método de árboles de decisión, ya que es uno de los algoritmos de Machine Learning más populares, debido a que puede ser fácilmente visible y entendible la forma en la que este método trabaja.

Un **árbol de decisión** tiene una estructura similar a un diagrama de flujo, donde un nodo interno representa una variable o atributo, la rama representa una regla de decisión y cada nodo u hoja representa el resultado. El algoritmo básico es el siguiente:

1. Selecciona el mejor atributo utilizando una medida de selección (heurísticamente).
2. El atributo se vuelve un nodo de decisión que divide el conjunto de datos en subconjuntos más pequeños.
3. Se construye el árbol recursivamente para cada atributo hasta que una de las siguientes condiciones coincida:
 - Todas las variables pertenecen al mismo valor de atributo.
 - Ya no quedan más atributos.
 - No hay más casos.

Además de la facilidad de interpretación, otra ventaja es que al considerar este modelo no se necesitaba hacer normalización de las variables. Más información.

```

# library(rpart)
# library(rpart.plot)
#
# # prueba árboles de decisión
# fit <- rpart(formula=as.factor(hijos)~I_UGASTO + I_HOGAR + arriendo + CANT_PERSONAS_HOGAR+
#               num_dormitorios+tipo_serv_sanitario+conyuges+num_cuartos+obtencion_agua_alimento, data=
# summary(fit)
# rpart.plot(fit)

# # accuracy
# accuracy(test.set$hijos, round(prediccion))

```

Al correr este modelo con las variables previamente seleccionadas para el ajuste, se obtuvo un accuracy del 38.204%.

Modelo 2: Bosque Aleatorio

Los Bosques Aleatorios o Random Forest, es un algoritmo de aprendizaje supervisado que puede utilizarse tanto para la clasificación como para la regresión. También es un algoritmo flexible y fácil de usar.

Un bosque está compuesto de árboles, y mientras más árboles tenga, más robusto será el bosque.

Los Bosques Aleatorios crean árboles de decisión a partir de muestras de datos seleccionados al azar, obtienen predicciones de cada árbol y seleccionan la mejor solución mediante una votación. También proporcionan un indicador bastante bueno de la importancia de la variable. Este algoritmo funciona de la siguiente manera:

1. Construir un árbol de decisión para cada muestra y obtener un resultado de predicción de cada árbol de decisión.
2. Realizar una votación por cada resultado previsto.
3. Seleccionar el resultado de la predicción con más votos como predicción final.

Más información.

```
# # prueba random forest
# modelo2 <- randomForest(as.factor(hijos)~I_UGASTO + I_HOGAR + arriendo + CANT_PERSONAS_HOGAR+
#                           num_dormitorios+tipo_serv_sanitario+conyuges+num_cuartos+obtencion_agua_alimen
# data=train.set)
# print(modelo2) # view results
# importance(modelo2) # importance of each predictor
#
# # predicción y accuracy
# prediccion1 = predict(fit1, test.set)
# accuracy(test.set$hijos, prediccion1)
```

Se obtuvo un accuracy de 76.97578%.

Resultados

Accuracy para el modelo 1: 38.20204%

Accuracy para el modelo 2: 76.97578%

Entre ambos modelos de clasificación implementados, el mejor modelo fue el modelo 2, por lo que se decidió trabajar sobre este la implementación de una aplicación Shiny.

Definición de la app Shiny y el usuario

Con el modelo 2, trabajado con el método de Random Forest, se implementó la app Shiny “Ramilia”, que sirve para obtener predicciones sobre el número de hijos en un hogar, ingresando la siguiente información: gasto mensual del hogar en pesos colombianos, ingreso mensual del hogar en pesos colombianos, arriendo mensual del hogar en pesos colombianos, número total de habitaciones en el hogar, número total de dormitorios en el hogar, tipo de servicio sanitario en el hogar, si el conyuge del jefe del hogar vive en el hogar y cómo se obtiene el agua para la preparación de alimentos en el hogar.

La aplicación se encuentra dirigida a cualquier persona o entidad que desee tomar decisiones acorde a la composición de un hogar en Colombia, entre estas pueden estar:

- Entidades financieras
- Entes políticos

- Entidades de salud
- Sector educacional
- Aseguradoras
- Entre otros

Conclusiones

Finalmente, a pesar de que se decidió implementar el modelo de bosque aleatorio, se debe tener presente que la interpretación para este método no es tan sencilla como la del método de árboles de decisión.

Se pensó también en implementar el método de boosteo, pero finalmente no se usó debido a que no se tenía mucha claridad sobre este.

Bibliografía, referencias y guías

- K vecinos más cercanos.
- Árboles de decisión.
- Bosques Aleatorios
- Métodos para selección de variables.