
**UNIVERSIDAD NACIONAL DE COLOMBIA,**
**SEDE MEDELLÍN**

# INFORME TÉCNICO TRABAJO No. 1
# TÉCNICAS EN APRENDIZAJE ESTADÍSTICO

**INTEGRANTES**

**Carolina Osorio Grajales - 1007338277**
**Mateo Restrepo Higuita - 1035236413**
**Julian Esteban Cadavid - 1000193246**
**Nicolás Laniado Valencia - 1040758375**
**Alejandro Ortiz Mejía - 1152710288**

**CONTENIDO:**

## I. Introducción y definición del usuario.

El presente trabajo se realiza basado en la encuesta de calidad de vida publicada por el DANE en el año 2019. Esta encuesta se encuentra dirigida a hogares y a partir de la información obtenida de ellos se calculan indicadores para la medición de varios 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. La Encuesta de Calidad de Vida (ECV) es una investigación que el DANE realiza con el objeto de recoger información sobre diferentes aspectos y dimensiones del bienestar y las condiciones de vida de los hogares, incluyendo temas como: el acceso a bienes y servicios públicos, privados o comunales, salud, educación, atención integral de niños y niñas menores de 5 años, entre otros. (DIMPE,2019).

Para darle solución al problema se define al jefe del hogar como la fuente principal de información con respecto a las características que tiene, valga la redundancia, el hogar. Así el número de hijos hará referencia a los hijos que tenga dicho jefe. Las variables explicativas a usar serán escogidas minuciosamente en términos de correlaciones con la variable respuesta de tal forma que permita mejorar la efectividad del modelo sin caer en sobre ajuste. Finalmente se creará una aplicación web con el modelo predictivo que tendrá como usuario final a cualquier persona del territorio nacional colombiano que desee identificar el número de hijos con base en ciertas variables que describen el hogar; además, es importante aclarar que el usuario previamente debe leer este reporte técnico. Un ejemplo de un usuario podría ser un trabajador comunitario que quiera establecer cuántos hijos hay en ciertos hogares donde realiza su trabajo social, así esta persona puede ingresar a la aplicación y con un número reducido de variables, definir el total de hijos por hogar y con esto establecer estrategias efectivas.

## II. Creación de la variable respuesta.

### II.a. Definición del hogar.

De acuerdo con la estructura de la base de datos, en una vivienda puede haber varios hogares, por tanto, es preciso distinguir con un identificador a los hogares que habitan una vivienda. Para ello se procedió así:

- Dado que cada vivienda se identifica con la variable ***Directorio***, y cada hogar de dicha vivienda se identifica con la variable ***Secuencia_p***, se procedió a hacer las respectivas agrupaciones, de tal manera que un hogar sería identificado por el ***Directorio*** de la vivienda al cual pertenece, un guión y después la ***Secuencia_p*** del respectivo hogar.

Así la base de datos resultante tendrá la siguiente forma:

| Id_Hogar | Hijos | P1_DEPARTAMENTO | CLASE | P1070 | P4005 | P4015 | P4567 | P8520S1 |
|---|---|---|---|---|---|---|---|---|
| 7120001-1 | 1 | 13 | 1 | 2 | 1 | 4 | 3 | 1 |
| 7120002-1 | 5 | 13 | 1 | 1 | 1 | 4 | 3 | 1 |
| 7120005-1 | 0 | 13 | 1 | 1 | 1 | 4 | 3 | 1 |
| 7120006-1 | 0 | 13 | 1 | 2 | 1 | 4 | 3 | 1 |
| 7120007-1 | 0 | 13 | 1 | 2 | 1 | 4 | 3 | 1 |
| 7120008-1 | 0 | 13 | 1 | 2 | 1 | 6 | 3 | 1 |
| 7120009-1 | 1 | 13 | 1 | 2 | 1 | 4 | 1 | 1 |
| 7120010-1 | 3 | 13 | 1 | 2 | 1 | 4 | 4 | 1 |
| 7120013-1 | 0 | 13 | 1 | 1 | 1 | 4 | 3 | 1 |
| 7120016-1 | 2 | 13 | 1 | 2 | 1 | 4 | 3 | 1 |
| 7120018-1 | 2 | 13 | 1 | 2 | 1 | 4 | 4 | 1 |

Tabla 1. El id del hogar está compuesto por el directorio y la secuencia_p.

### II.b. Definición de la variable respuesta.

Para la creación de la variable respuesta se define la cantidad de hijos del hogar como el número de hijos que tiene el jefe del hogar.

Debido a que la variable respuesta no estaba explícita en los datos, se procedió al cálculo de ésta de la siguiente manera: se tomó la variable ***P6051***, que corresponde al parentesco del encuestado con el jefe del hogar, y a partir de esto se identificaron 2 posibles casos:

1. En el hogar hay uno (1) o más hijos: este caso ocurre cuando entre las respuestas a la pregunta ***P6051*** para cada hogar hay una o más ocurrencias de la respuesta número 3, la cual corresponde a que el parentesco del encuestado con el jefe de hogar es hijo(a) o hijastro(a). Por lo tanto, el número de hijos de ese hogar es igual al número de ocurrencias de la respuesta número 3.

2. En el hogar no hay hijos: ocurre cuando entre las respuestas a la pregunta ***P6051*** no está la número 3.

Basado en lo anteriormente expuesto la variable respuesta corresponderá a:

$$\#\ de\ hijos\ =\ \#\ de\ hijos\ del\ jefe\ del\ hogar$$

**II.c. Presentación de ejemplos.**

Tomando como ejemplo la tabla 1.

| Id_Hogar | Hijos | P1_DEPARTAMENTO | CLASE | P1070 | P4005 | P4015 | P4567 | P8520S1 |
|---|---|---|---|---|---|---|---|---|
| 7120001-1 | 1 | 13 | 1 | 2 | 1 | 4 | 3 | 1 |
| 7120002-1 | 5 | 13 | 1 | 1 | 1 | 4 | 3 | 1 |
| 7120005-1 | 0 | 13 | 1 | 1 | 1 | 4 | 3 | 1 |

Tabla 1.1. Recorte de tabla 1

Para el caso del tercer hogar, el jefe del hogar no tiene ningún hijo, es por ello que se le asigna un 0, en el caso del segundo hogar, el jefe tiene 5 hijos, por tanto se le asigna 5, es decir, que en las encuestas correspondientes a su hogar cinco personas marcaron ser hijos o hijastros del jefe de hogar.

## III. Descripción de las variables explicativas.

**III.a. Definición del procedimiento usado.**

La selección de las variables explicativas se hizo por medio de un proceso de varios pasos así:

1. Se revisó cuáles variables tenían definición en la documentación suministrada por el DANE, con el fin de descartar aquellas que carecían de este significado.

Al hacer una revisión de cada variable dentro de las diferentes tablas proporcionadas por el DANE en la encuesta de calidad de vida, se encontró que algunas variables presentes eran creación del DANE formuladas a partir de ciertos indicadores preestablecidos y en otros casos no se encontró descripción alguna que permitiera identificar la pregunta en cuestión. Basado en la descripción del usuario final que hará uso de la aplicación se decide que las variables explicativas usadas deben tener sentido y permitir una respuesta fácil por parte del jefe del hogar, a su vez que estas tengan interpretación y tengan sentido para explicar el número de hijos del hogar.

2. Se hace imputación de datos, pues para los modelos que se propondrán no pueden existir valores nulos.

Esta imputación de datos se hace creando una nueva categoría ‘000’ en variables categóricas y usando la media de los datos para el caso donde las variables son de tipo numérico.

3. Se calcularon las correlaciones de las variables predictoras con la variable de salida, con el fin de encontrar una relación significativa entre estas.

Se hace un barrido inicial de la base de datos y se calculan dos correlaciones: Pearson y Spearman. Con el fin de encontrar dependencia lineal o no lineal. Se define un conjunto de cincuenta variables iniciales que cumplen con dos requisitos: Correlación significativa e interpretables. Se propone un primer modelo lineal con resultados no significativos, debido a que, aunque las correlaciones son diferentes de cero, estas no son suficientes para aumentar un factor de determinación R2 altamente significativo y problemas graves en términos de residuales; es por ello que no se define este como un modelo final.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.8738726  0.0285824  65.560  < 2e-16 ***
P6040       -0.0168016  0.0003257 -51.593  < 2e-16 ***
P60833      -0.1121225  0.0100461 -11.161  < 2e-16 ***
P60813      -0.0219930  0.0096138  -2.288 0.022160 *
P55022       0.7080978  0.0251226  28.186  < 2e-16 ***
P55023       0.4957333  0.0286092  17.328  < 2e-16 ***
P55024       0.3401865  0.0261483  13.010  < 2e-16 ***
P55025      -0.3112726  0.0265052 -11.744  < 2e-16 ***
P55026       0.6031599  0.0258867  23.300  < 2e-16 ***
P60802      -0.5378345  0.1767964  -3.042 0.002350 **
P60803      -0.4977710  0.0633811  -7.854 4.09e-15 ***
P60804      -0.5946442  0.1542526  -3.855 0.000116 ***
P60805      -0.2957180  0.0168716 -17.528  < 2e-16 ***
P60806      -0.3951160  0.0127488 -30.992  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.087 on 86786 degrees of freedom
Multiple R-squared:  0.153,     Adjusted R-squared:  0.1529
F-statistic:  1206 on 13 and 86786 DF,  p-value: < 2.2e-16
```

Figura 1. Modelo de regresión lineal múltiple salida de R

A pesar de que todas las variables resultan significativas, el resultado solo explica 15% de la variabilidad de los hijos del hogar. Se hace un análisis de residuales y no se encuentra validez en los supuestos de normalidad en los errores y varianza constante. Es por ello que se descarta esta alternativa.

4. Con base en los resultados del punto anterior, se realizó una discusión entre los miembros del equipo para determinar, según el contexto del problema, cuáles variables eran mejores candidatas para la predicción del número de hijos.

Debido a los deficientes resultados del primer modelo lineal realizado, se busca hacer una reducción de dimensionalidad poniendo en contexto con el equipo aquellas variables con correlaciones altas y significativas en el primer modelo para, en conjunto, leer detenidamente cuales tenían sentido para explicar el número de hijos.

5. Posteriormente, se ajustó un modelo con la metodología *XGBOOST* buscando hacer una clasificación para múltiples clases, con el fin de encontrar las variables más

significativas de las ya preseleccionadas en los pasos previos. Esto nos dio como resultado 40 variables significativas, y a partir de estas se realizó una nueva discusión, en la cual finalmente se escogieron las 18 variables predictoras definitivas.

**III.b. Presentación de las variables escogidas.**

Es importante aclarar que para las variables predictoras escogidas las respuestas tomadas fueron las del jefe del hogar.

Luego del exhaustivo proceso de selección de variables realizado, se escogió el siguiente conjunto de variables:

- **P6390S2: ¿A qué actividad se dedica principalmente la empresa o negocio en la que ... realiza su trabajo?**

*La respuesta a esta pregunta es un código que se debe de validar con el DANE, por medio del código CIIU que se puede verificar en el siguiente link de la cámara de comercio de bogotá donde se puede buscar la actividad económica.*
*https://linea.ccb.org.co/descripcionciiu/*

- **P1_DEPARTAMENTO: Departamento de nacimiento.**

*La respuesta a esta pregunta es un código que se debe de validar con el DANE por medio del siguiente enlace:*
*https://fopep.gov.co/wp-content/uploads/2019/02/Tabla-C%C3%B3digos-Dane.pdf*

- **P8587S1: Grado o año aprobado.**

*Los grados posibles de cursar en cada nivel educativo corresponden con los grados que conforman el nivel. Cinco (5) representa Básica primaria, nueve (9) representa básica secundaria y once (11) o doce (12) representa educación media.*

- **P415: ¿Cuántas horas a la semana trabaja normalmente ….. en ese trabajo ?**

*Se utilizan para la medición de horas de trabajo, información que hace posible caracterizar las condiciones de empleo diferenciando entre empleo de tiempo completo y de tiempo parcial. Para diligenciar esta pregunta se indaga por el número de horas que la persona trabaja normalmente a la semana. Se aclara que no se debe tener en cuenta las horas que trabajó de más o de menos por cualquier motivo en la semana, sino aquellas horas que trabaja normalmente en su empleo principal.*

- **P6750: ¿cuál fue la ganancia neta o los honorarios netos de... En esa actividad, negocio, profesión o finca, el mes pasado ?**

*En esta pregunta su respuesta es la Ganancia neta Valor ($ COP) un número positivo.*

- **P756S2: Municipio.**

*La respuesta a esta pregunta es un código que se debe de validar con el DANE.*

- **P8587: ¿cuál es el nivel educativo más alto alcanzado por ... Y el último año o grado aprobado en este nivel?**

1 Ninguno.
2 Preescolar.
3 Básica Primaria (1º - 5º).
4 Básica secundaria (6º - 9º).
5 Media (10º - 13º).
6 Técnico sin título .
7 Técnico con título .
8 Tecnológico sin título.
9 Tecnológico con título.
10 Universitario sin título.
11 Universitario con título.
12 Postgrado sin título.
13 Postgrado con título.

- **P8520S1A1: Estrato para tarifa.**

*¿Con cuáles de los siguientes servicios públicos, privados o comunales cuenta la vivienda?*
*1. Energía eléctrica*
*Estrato para tarifa:*
*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.*

- **P1896: En general, qué tan satisfecho(a) se siente ... con su ingreso actualmente?**

*En general, qué tan satisfecho(a) se siente ... con su ingreso actualmente? 10 Totalmente satisfecho(a) 9 8 7 6 5 4 3 2 1 0 Totalmente insatisfecho(a).*

- **P1897: En general, qué tan satisfecho(a) se siente ... con su salud actualmente?**

*En general, qué tan satisfecho(a) se siente ... con su salud actualmente? 10 Totalmente satisfecho(a) 9 8 7 6 5 4 3 2 1 0 Totalmente insatisfecho(a).*

- **P1899: En general, qué tan satisfecho(a) se siente ... con su trabajo/actividad actualmente?**

*En general, qué tan satisfecho(a) se siente ... con su trabajo/actividad actualmente? 10 Totalmente satisfecho(a) 9 8 7 6 5 4 3 2 1 0 Totalmente insatisfecho(a).*

- **P1927: ¿En cuál escalón diría usted que se encuentra parado(a) en este momento?**

*Imagine una escalera con escalones numerados de 0 a 10, donde 0 es el escalón más bajo y 10 el escalón más alto. El más alto representa la mejor vida que usted podría tener y el más bajo, la peor.*

- **P4015: Material predominante de los pisos.**

*Material predominante de los pisos 1. Alfombra o tapete de pared a pared 2. Madera pulida y lacada, parqué 3. Mármol 4. Baldosa, vinilo, tableta, ladrillo, laminado 5. Madera burda, tabla, tablón, otro vegetal 6. Cemento, gravilla 7. Tierra, arena.*

- **P4567: ¿Cuál es el material predominante del techo o cubierta?**

*1. Plancha de concreto, cemento u hormigón 2. Tejas de barro 3. Teja de asbesto - cemento 4. Teja metálica o lámina de zinc 5. Teja plástica 6. Paja, palma u otros vegetales 7. Material de desecho (tela, cartón,latas, plástico, otros) .*

- **P5502: Actualmente…**

*Actualmente…: 1. No está casado(a) y vive en pareja hace menos de dos años 2. No está casado(a) y vive en pareja hace dos años o más 3. Está viudo(a) 4. Está separado(a) o divorciado(a) 5. Está soltero(a) 6. Está casado(a).*

- **P6087: ¿Cuál es o fue el nivel de educación más alto alcanzado por el padre de…..?**

*Cuál es o fue el nivel de educación más alto alcanzado por el padre de …? 1. Algunos años de primaria 2. Toda la primaria 3. Algunos años de secundaria 4. Toda la secundaria 5. Uno o más años de técnica o tecnológica 6. Técnica o tecnológica completa 7. Uno o más años de universidad 8. Universitaria completa 9. Ninguno 10. No sabe.*

- **P6088: ¿cuál es o fue el nivel de educación más alto alcanzado por la madre de.....?**

*Cuál es o fue el nivel de educación más alto alcanzado por la madre de …? 1. Algunos años de primaria 2. Toda la primaria 3. Algunos años de secundaria 4. Toda la secundaria 5. Uno o más años de técnica o tecnológica 6. Técnica o tecnológica completa 7. Uno o más años de universidad 8. Universitaria completa 9. Ninguno 10. No sabe.*

- **P6040: ¿cuántos años cumplidos tiene...?**

*Valor numérico de la edad.*

**III.c. Transformación de los datos.**

- **Variable Respuesta: Hijos.**

```
     Hijos
Min.   : 0.000
1st Qu.: 0.000
Median : 1.000
Mean   : 1.098
3rd Qu.: 2.000
Max.   :11.000
```

El 75% de los jefes de hogar tienen menos de dos hijos. Esto quiere decir que es muy extraño encontrar personas que tengan más de este número de hijos.

```
     Hijos
0       :34437
1       :24440
2       :17792
3       : 6886
4       : 2162
5       :  709
(Other):  374
```

Al hacer este mismo análisis se puede identificar que solo poco más de mil jefes de hogar tienen más de cinco hijos, es decir, menos del 5% de la población evaluada.

- **Reemplazo de guiones por espacios vacíos y renombre de variables.**

Para las siguientes variables se hace el cambio de '-' por espacios, con el fin de tomarlos como valores nulos:

- *P6390S2*
- *P8587S1*
- *P415*
- *P6750*
- *P756S2*
- *P8587*
- *P8520S1A1*
- *P1_DEPARTAMENTO*
- *P1896*
- *P1897*
- *P1899*
- *P1927*
- *P4015*
- *P4567*
- *P5502*
- *P6087*
- *P6088*
- *P6040*

Adicionalmente se hace un renombre de las variables con el fin de que sea más fácil de trabajar con ellas al momento de crear el modelo.

El procedimiento es realizado en el software Qlik, mediante el siguiente código:

```
IF(P6390S2='-','',P6390S2) AS ACTIVIDAD_ECONOMICA,
IF(P8587S1= '-','',P8587S1) AS GRADO_APROBADO,
P415 AS HORAS_SEM_TRABAJA,
P6750 AS GANANCIA_NETA,
IF(P756S2= '-','',P756S2) AS MUNICIPIO_NACIMIENTO,
IF(P8587= '-','',P8587) AS NIVEL_EDUCATIVO ,
IF(P8520S1A1= '-','',P8520S1A1) AS ESTRATO_TARIFA,
IF(P1_DEPARTAMENTO= '-','',P1_DEPARTAMENTO) AS DEPARTAMENTO,
IF(P1896= '-','',P1896) AS SACTISFACCION_INGRESO,
IF(P1897= '-','',P1897) AS SACTISFACCION_SALUD,
IF(P1899= '-','',P1899) AS SACTISFACCION_TRABAJO,
IF(P1927= '-','',P1927) AS ESCALON_VIDA,
IF(P4015= '-','',P4015) AS MATERIAL_PISOS,
IF(P4567= '-','',P4567) AS MATERIAL_TECHO,
IF(P5502= '-','',P5502) AS ESTADO_CIVIL,
IF(P6087= '-','',P6087) AS NIVEL_EDUCACION_PADRE,
IF(P6088= '-','',P6088) AS NIVEL_EDUCACION_MADRE,
IF(P6040= '-','',P6040) AS EDAD


FROM [lib://Datos/Resultado.csv](txt, utf8, embedded labels, delimiter is ';', msq)
Where P756S2<>'-' AND len(P6087)>0 AND len(P6088)>0 AND Len(P6390S2)>0 AND Len(P8587S1)>0;
Store Resultados1 into lib://Datos/Resultado1.csv(txt,delimiter is ';');
```

**III.d. Análisis descriptivo de las variables explicativas.**

```
   ï..HIJOS      ACTIVIDAD_ECONOMICA GRADO_APROBADO  HORAS_SEM_TRABAJA
 Min.   :0.000   141    : 4263      5      :14362   Min.   :  1.00
 1st Qu.:0.000   161    : 3668      11     :12704   1st Qu.: 35.00
 Median :1.000   123    : 3054      2      : 7362   Median : 48.00
 Mean   :1.178   4111   : 2782      3      : 6905   Mean   : 43.81
 3rd Qu.:2.000   113    : 2706      4      : 3572   3rd Qu.: 50.00
 Max.   :5.000   4921   : 2285      1      : 3137   Max.   :120.00
                 (Other):39924      (Other):10640
```

Como se dijo anteriormente, el 75% de la población tiene menos de 2 hijos. La actividad económica que más se repite en el conjunto de datos es la 141. La gran mayoría de jefes de hogar aprobaron solo hasta quinto grado, el porcentaje que sigue son aquellos jefes que llegaron hasta undécimo grado. En promedio se tiene que los jefes de hogar trabajan 43 horas, y el 75% menos de 50 horas a la semana.

```
GANANCIA_NETA       MUNICIPIO_NACIMIENTO NIVEL_EDUCATIVO ESTRATO_TARIFA
Min.   :         0   11001  :  963       3      :24659   1      :28220
1st Qu.:    200000   8001   :  365       5      :13892   2      :15333
Median :    500000   76001  :  317       4      : 8705   3      : 4855
Mean   :    734658   50001  :  294       11     : 4170   0      : 2484
3rd Qu.:    800000   5001   :  267       7      : 3570   4      : 1419
Max.   :100000000   (Other):22445       13     : 1409   (Other): 3629
NA's   :23580       NA's   :34031       (Other): 2277   NA's   : 2742
```

En promedio los jefes de hogar ganan $734.658 algo que es muy preocupante ya que para el salario mínimo del 2019, este valor es inferior. El municipio que más se repite es el 11001,

con un nivel educativo de 1 que hace referencia a ningún tipo de educación. Estrato para tarifa 1 y 2 corresponden al nivel bajo y medio-bajo respectivamente.

```
 DEPARTAMENTO   SACTISFACCION_INGRESO SACTISFACCION_SALUD
18      : 2409   8      :11971        8      :14880
66      : 2311   7      :10298        10     :13937
41      : 2175   6      : 7636        9      : 9750
5       : 2162   5      : 7409        7      : 8594
25      : 2115   10     : 6952        6      : 4765
52      : 2100   9      : 5165        5      : 3940
(Other):45410   (Other): 9251        (Other): 2816
SACTISFACCION_TRABAJO  ESCALON_VIDA   MATERIAL_PISOS MATERIAL_TECHO
8      :14673         8      :15046   1:   72        1:10121
10     :12085         7      :11691   2:  473        2: 3549
7      : 9977         10     :11154   3:   98        3:16534
9      : 7749         9      : 6920   4:24906        4:26951
6      : 5933         6      : 6492   5: 4288        5:  202
5      : 4348         5      : 4848   6:23398        6: 1199
(Other): 3917         (Other): 2531   7: 5447        7:  126
```

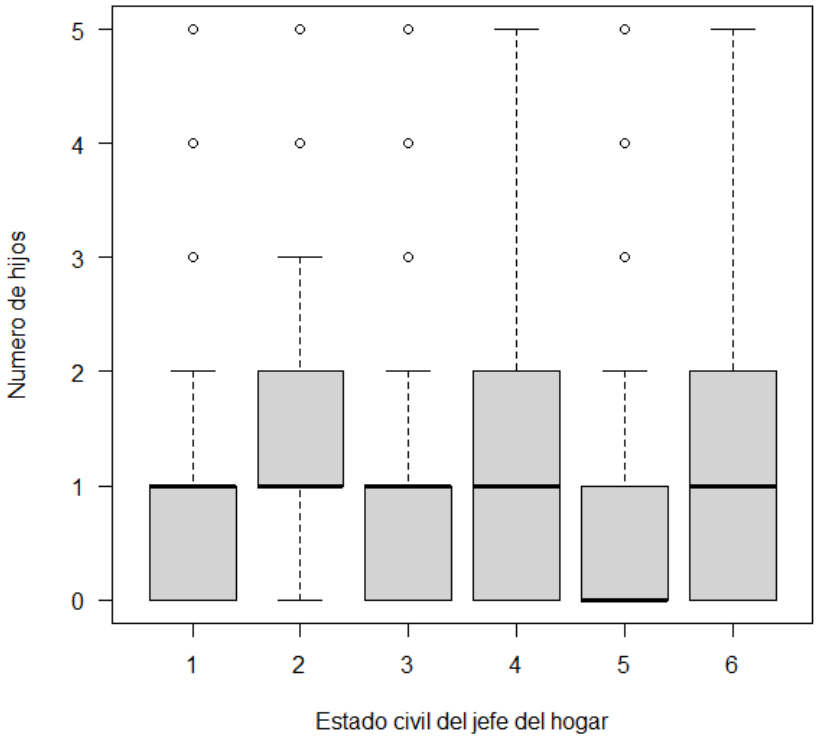
El departamento parece ser muy equitativo, es decir que para cada departamento se sacó aproximadamente la misma cantidad de población para realizar el estudio. Tanto para satisfacción de ingresos, salud, trabajo y escalón de vida, la mayoría de los jefes de hogar se mostraron con resultados positivos entre 8,9,10. Esto quiere decir que en general se encuentran satisfechos con alguno de estas cuatro dimensiones. El material de piso que más predomina es la alfombra así mismo como el techo de concreto u hormigón.

```
ESTADO_CIVIL NIVEL_EDUCACION_PADRE NIVEL_EDUCACION_MADRE      EDAD
1: 1705      1      :19707         1      :20698         Min.   :15.00
2:23532      10     :11786         9      :11519         1st Qu.:35.00
3: 1984      9      :11289         2      : 9790         Median :45.00
4: 9739      2      : 8698         10     : 8716         Mean   :45.27
5: 7314      4      : 3252         4      : 3737         3rd Qu.:55.00
6:14408      3      : 2066         3      : 2562         Max.   :98.00
             (Other): 1884         (Other): 1660
```

Del anterior podemos concluir que la gran mayoría de jefes de hogar se encuentran no casados pero viven en pareja hace dos años o más o se encuentran separados. El nivel de educación de los padres y las madres de los jefes de hogar de esta población a lo sumo llegaron al grado 1. La edad de los jefes de hogar en promedio es de 36 años, y el 75% de la población tiene 55 años o menos.

**IV. Relación entre variables explicativas y la variable respuesta.**

Al realizar un análisis detallado de cada variable categórica con respecto a la variable respuesta número de hijos del hogar no se encuentra ninguna variable que muestre diferencias significativas en el boxplot salvo el estado civil del jefe del hogar, para el cual se obtiene:




La mediana para el estado civil 5 es inferior a todas las demás. Dicho estado civil corresponde a estar soltero. Es decir que si un jefe de hogar está soltero a lo sumo tiene 1 hijo, y esto es válido para el 75% de la población. Así mismo para un jefe de hogar que se encuentren no casados pero viven con su pareja hace más de dos años es quienes tienen en promedio más hijos.

## V. Definición del Modelo.

### V.a. Modelo escogido.

Luego de realizar algunas pruebas pilotos con enfoques de regresión lineal, hacer sus comparaciones de métricas y verificaciones de supuestos se decide implementar un modelo ***Gradient Boosting*** el cual está formado por un conjunto de árboles de decisión individuales, entrenados de forma secuencial, de forma que cada nuevo árbol trata de mejorar los errores de los árboles anteriores. La predicción de una nueva observación se obtiene agregando las predicciones de todos los árboles individuales que forman el modelo.

Un modelo Gradient Boosting Trees está formado por un conjunto (ensemble) de árboles de decisión individuales, entrenados de forma secuencial. Cada nuevo árbol emplea información del árbol anterior para aprender de sus errores, mejorando iteración a iteración. En cada árbol individual, las observaciones se van distribuyendo por bifurcaciones (nodos) generando la estructura del árbol hasta alcanzar un nodo terminal. La predicción de una nueva observación se obtiene agregando las predicciones de todos los árboles individuales

que forman el modelo. Para entender cómo funcionan los modelos Gradient Boosting Trees es necesario conocer primero los conceptos de ensemble y boosting.

El algoritmo es implementado en python debido a que tiene una función de regularización mejor que la de R, es decir que a la hora de determinar la probabilidad de escoger una clasificación u otra este presenta mejor desempeño.

```
datos_train, datos_test = train_test_split(datos, test_size=.20, random_state=1)
```

```
datos_train_mat = xgb.DMatrix(datos_train.drop("HIJOS", 1), label=datos_train["HIJOS"])
datos_test_mat = xgb.DMatrix(datos_test.drop("HIJOS", 1), label=datos_test["HIJOS"])
```

```
parametros = {
    'max_depth': 7,
    'learning_rate':0.01,
    'objective': 'multi:softmax',  # error evaluation for multiclass training,
    'num_class': 6
}
```

```
rondas =2000
```

```
evaluacion = [(datos_test_mat, "eval"), (datos_train_mat, "train")]
```

```
modelo = xgb.train(parametros, datos_test_mat, rondas, evaluacion)
```

```
In [338]: prediccion = modelo.predict(datos_test_mat)
```

```
In [339]: print(classification_report(datos_test["HIJOS"], prediccion))
```

```
              precision    recall  f1-score   support

           0       0.89      0.92      0.90      4310
           1       0.87      0.88      0.87      3258
           2       0.82      0.87      0.84      2632
           3       0.98      0.71      0.82      1039
           4       1.00      0.77      0.87       329
           5       0.99      0.92      0.95       169

    accuracy                           0.88     11737
   macro avg       0.92      0.84      0.88     11737
weighted avg       0.88      0.88      0.87     11737
```

```
In [340]: cm = confusion_matrix(datos_test["HIJOS"], prediccion)
          cm
```

```
Out[340]: array([[3976,  158,  172,    4,    0,    0],
                 [ 264, 2857,  130,    6,    0,    1],
                 [ 185,  148, 2294,    4,    0,    1],
                 [  51,   91,  157,  740,    0,    0],
                 [  12,   21,   42,    2,  252,    0],
                 [   1,    4,    9,    0,    0,  155]], dtype=int64)
```

Una de las ventajas de este algoritmo es que permite la entrada de datos nulos, adicional a esto permite la exportación en formato binario para ser usado por ejemplo en R. Se define como el número de nodos siete debido a que si es menor a esto la varianza se hace pequeña pero el sesgo crece, es por eso que se define un punto de equilibrio de sesgo y varianza y se definió a partir de ensayo y error, para llegar a este modelo final se hicieron varios con el que se probaba efectividad del ajuste. Se define el learning rate a un valor de 0.01 este es una media que ayuda a que el modelo no se deje influenciar por el paso anterior (o árbol anterior) es por ello que se garantiza que el modelo aprenda verificando siempre que no se cayera en un problema de sobre parametrización.

**V.b. Medidas de efectividad.**

Para evaluar la efectividad y precisión del modelo se presenta la siguiente tabla

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.89 | 0.92 | 0.90 | 4310 |
| 1 | 0.87 | 0.88 | 0.87 | 3258 |
| 2 | 0.82 | 0.87 | 0.84 | 2632 |
| 3 | 0.98 | 0.71 | 0.82 | 1039 |
| 4 | 1.00 | 0.77 | 0.87 | 329 |
| 5 | 0.99 | 0.92 | 0.95 | 169 |
| accuracy | | | 0.88 | 11737 |
| macro avg | 0.92 | 0.84 | 0.88 | 11737 |
| weighted avg | 0.88 | 0.88 | 0.87 | 11737 |

En la tabla anterior se encuentran especificados los porcentajes de datos clasificados correctamente se puede identificar que se tiene una tasa de acierto del 88% con muy buenas precisiones en cada una de las posibles respuestas de cantidad de hijos.

## VI. Creación de aplicación web.

El modelo fue desarrollado en python sin embargo la aplicación web se realizó con la herramienta Shiny de Rstudio.

Por medio del siguiente código se exporta el modelo en formato binario con el fin de que pueda ser leído por R y hacer uso de él.

```
: modelo.save_model('modelo1.bin')
```

Luego, por medio de la siguiente función se lee en la aplicación que será creada

```
> datos <- read.csv('Resultado1.csv',sep = ';', encoding = 'UTF-8')
>
> modelopython<-xgb.load('modelo1.bin')
```

Finalmente, el código que crea la aplicación se expone a continuación:

```
ui <-(
   navbarPage(
      title="TAE",
      theme=shinytheme("spacelab"),
      inverse=TRUE,
      position = "static-top",
      tabPanel("Clasificacion Hijos",

   fluidPage(
   # titlePanel("---------------------------------------------------------------------------- "),
   flowLayout(
      numericInput("S1",
                   label = h4("Actividad Economica"),
                  value=1, min = min(datos$ACTIVIDAD_ECONOMICA),
```

```
                max=max(datos$ACTIVIDAD_ECONOMICA)),
      numericInput("S2",
                label = h4("Grado Aprobado"),
                value=1, min = min(datos$GRADO_APROBADO),
                max=max(datos$GRADO_APROBADO)),
      numericInput("S3",
                label = h4("Horas Trabaja Semana"),
                value=1, min = min(datos$HORAS_SEM_TRABAJA),
max=max(datos$HORAS_SEM_TRABAJA)),
      numericInput("S4",
                label = h4("Ganancia Neta"),
                value=1, min = min(datos$GANANCIA_NETA),
                max=max(datos$GANANCIA_NETA)),
      numericInput("S5",
                label = h4("Municipio Nacimiento"),
                value = 1, min = min(datos$MUNICIPIO_NACIMIENTO),
                max=max(datos$MUNICIPIO_NACIMIENTO)),
      numericInput("S6",
                label = h4("Nivel Educativo"),
                value=1, min = min(datos$NIVEL_EDUCATIVO),
                max=max(datos$NIVEL_EDUCATIVO)),
      numericInput("S7",
                label = h4("Estrato Tarifa"),
                value=1, min = min(datos$ESTRATO_TARIFA),
                max=max(datos$ESTRATO_TARIFA)),
      numericInput("S8",
                label = h4("Departamento"),
                value=1, min = min(datos$DEPARTAMENTO),
                max=max(datos$DEPARTAMENTO)),
      numericInput("S9",
                label = h4("Satisfaccion Ingreso"),
                value=1, min = min(datos$SACTISFACCION_INGRESO),
                max=max(datos$SACTISFACCION_INGRESO)),
      numericInput("S10",
                label = h4("Satisfaccion Salud"),
                value=1, min = min(datos$SACTISFACCION_SALUD),
                max=max(datos$SACTISFACCION_SALUD)),
      numericInput("S11",
                label = h4("Satisfaccion Trabajo"),
                value=1, min = min(datos$SACTISFACCION_TRABAJO),
                max=max(datos$SACTISFACCION_TRABAJO)),
      numericInput("S12",
                label = h4("Escalon Vida"),
                value=1, min = min(datos$ESCALON_VIDA),
                max=max(datos$ESCALON_VIDA)),
      numericInput("S13",
                label = h4("Material Pisos"),
                value=1, min = min(datos$MATERIAL_PISOS),
                max=max(datos$MATERIAL_PISOS)),
      numericInput("S14",
                label = h4("Material Techo"),
                value=1, min = min(datos$MATERIAL_TECHO),
                max=max(datos$MATERIAL_TECHO)),
      numericInput("S15",
                label = h4("Estado Civil"),
```

```
                value=1, min = min(datos$ESTADO_CIVIL),
                max=max(datos$ESTADO_CIVIL)),
        numericInput("S16",
                label = h4("Nivel Educativo Padre"),
                value=1, min = min(datos$NIVEL_EDUCACION_PADRE),
                max=max(datos$NIVEL_EDUCACION_PADRE)),
        numericInput("S17",
                label = h4("Nivel Educativo Madre"),
                value=1, min = min(datos$NIVEL_EDUCACION_MADRE),
                max=max(datos$NIVEL_EDUCACION_MADRE)),
        numericInput("S18",
                label = h4("Edad"),
                value=1, min = min(datos$EDAD),
                max=max(datos$EDAD)),
        verbatimTextOutput("text",placeholder = F),
        verbatimTextOutput(""),
        verbatimTextOutput(""),
        verbatimTextOutput(""),
        verbatimTextOutput(""),
        verbatimTextOutput(""),
        img(src="1.png", height="150", width="250")

    )
)
)))

server <- function(input, output,session) {
    output$text <- renderText({
        {
            Data1 = reactive({
                df<-data.frame(
                        ACTIVIDAD_ECONOMICA = input$S1,
                        GRADO_APROBADO = input$S2,
                        HORAS_SEM_TRABAJA = input$S3,
                        GANANCIA_NETA = input$S4,
                        MUNICIPIO_NACIMIENTO = input$S5,
                        NIVEL_EDUCATIVO = input$S6,
                        ESTRATO_TARIFA = input$S7,
                        DEPARTAMENTO = input$S8,
                        SACTISFACCION_INGRESO = input$S9,
                        SACTISFACCION_SALUD = input$S10,
                        SACTISFACCION_TRABAJO = input$S11,
                        ESCALON_VIDA = input$S12,
                        MATERIAL_PISOS = input$S13,
                        MATERIAL_TECHO = input$S14,
                        ESTADO_CIVIL = input$S15,
                        NIVEL_EDUCACION_PADRE = input$S16,
                        NIVEL_EDUCACION_MADRE = input$S17,
                        EDAD = input$S18)
            })

            prediccion <- predict(modelopython, newdata = as.matrix(Data1()))
            #print(prediccion)
            paste(c("Hijos Estimados:" ,prediccion))
```

```
    }
  })
}

# Run the application
shinyApp(ui = ui, server = server)
```

En la siguiente página web se podrá encontrar un formulario mediante el cual el usuario podrá ingresar los datos con los cuales el sistema hará la predicción de cuántos hijos tendrá.

https://tae-2021-1.shinyapps.io/Clasificador2/

## VII. Referencias


1. *COLOMBIA - Encuesta Nacional de Calidad de Vida - ECV 2019. (2019,Noviembre). MIMPE. Bogotá, Colombia. Departamento Nacional de Estadística.*
2. *Amat Rodrigo, J. (2020, octubre). Gradient Boosting con Python. Ciencia de datos.https://www.cienciadedatos.net/documentos/py09_gradient_boosting_python.html*