Análisis de datos de beneficiarios Más Familias en Acción Colombia

Cristian Camilo Jimenez Mancera

Facultad de Ingeniería

Universidad Nacional de Colombia

Bogotá DC

ccjimenezm@unal.edu.co

# Introducción

“Familias en Acción es un programa a cargo de Prosperidad Social que les ofrece la posibilidad a las familias con niños, niñas y adolescentes en condiciones de pobreza y pobreza extrema, recibir un incentivo económico condicionado a la asistencia escolar y a atenciones de salud de los niños, niñas y adolescentes.” [3]

Como parte de la iniciativa del Gobierno Nacional de datos abiertos, el Departamento Administrativo para la Prosperidad Social ofrece conjuntos de datos anonimizados acerca de los beneficiarios de este programa. La correcta interpretación de estos datos puede suministrar información de alto interés sobre las necesidades de las poblaciones vulnerables en Colombia. Para los motivos de este estudio, se cuenta con un conjunto emitido en el año 2018.

# Objetivos

1. Mediante el uso de técnicas de minería de datos, identificar patrones en beneficiarios del programa para crear agrupamientos (clustering) de los mismos de acuerdo con las similitudes entre sí.
2. A partir de los agrupamientos generados, poder clasificar nuevos beneficiarios con base en sus relaciones de cercanía.

# Descripción del conjunto de datos

El conjunto de datos que tenemos corresponde a una base de beneficiarios del programa de Más Familias en Acción de Colombia, determinando características generales de los beneficiarios, valor y tipo de incentivo recibido, el dataset cuenta con 3,98 millones de registros y con 22 columnas, un peso aproximado de 572 Mb.

Adicionalmente, se incluye un conjunto de datos del Índice de Desarrollo Humano (IDH) por departamento de Colombia a partir de la medición hecha en 2019 por las Naciones Unidas [4]. Este conjunto fue creado manualmente obteniendo un total de 33 registros. Uno por cada departamento y uno para Bogotá D.C. Consistiendo en 2 columnas, departamento e IDH, con lo cual se cruza con el departamento de los beneficiarios en el conjunto original para resultar en un IDH por cada fila de entre 33 posibles valores (sin considerar repetidos).

# Características de los datos

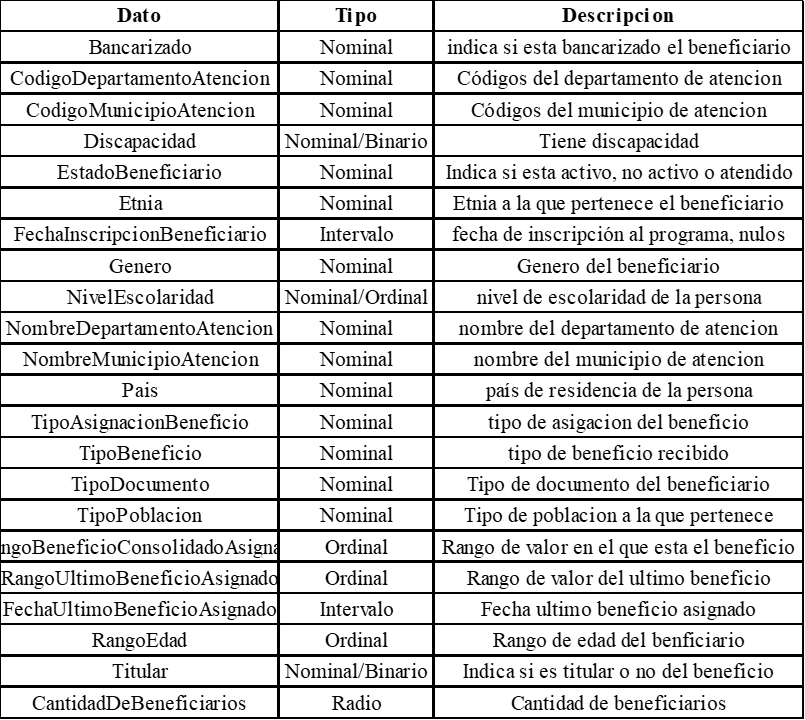


Table 1 Descripción de las características de las columnas del conjunto de datos inicial Más Familias en Acción de Colombia

# Herramientas para el análisis exploratorio

Para la realización del ejercicio de análisis, se utilizaron DeepNote y Google Colaboratory como plataformas colaborativas de código a partir de Jupyter Notebooks basados en lenguaje Python. De igual manera, se incluyeron las librerías Numpy, Pandas, scipy, matplotlib, prince, sklearn y Seaborn, entre otras, para operaciones matemáticas, manipulación y graficado de conjuntos de datos.

# Preprocesamiento inicial

Para la preparación del dataset, tuvimos que utilizar Python para generar una muestra representativa y ponderada de los datos utilizando el campo de CantidadDeBeneficiarios. Realizamos una agrupación de todos los registros sumados en el campo CantidadDeBeneficiarios para luego obtener una muestra de 100000 registros que nos permite trabajar fluidamente en Google Colaboratory, esto elimina también las columnas con valores NULL en este campo.

# Análisis exploratorio

Una vez se hace la corrección, se procede a hacer una revisión de las variables restantes, que son en su mayoría discretas en un rango de valores relativamente pequeño. Se encuentra que existen múltiples valores en algunas columnas que pueden ser considerados como uno solo para motivos de simplificación de este análisis. La discrepancia puede ser atribuida a la forma de recolección de la información. Se reducen las categorías en las columnas País y Etnia, de 6 a 2 y 10 a 7 respectivamente. El resultado del cambio se observa en la figura 2. Para País se agruparon en ‘Colombia’ todos los valores menos ‘ND’ ya que son formas diferentes de referirse al mismo país, y se unieron para Etnia en ‘AFROCOLOMBIANO’ los 3 valores que contienen la palabra, así como se unió en ‘ROM’ los 2 que tienen la palabra.

Tras la agrupación de valores redundantes, se evidencia que la columna País cuenta con 2 opciones, ‘ND’ y ´Colombia´, donde ‘ND’ significa No Definido. En la columna TipoAsignacionBeneficio se observa la misma situación con ‘ND’ y ‘MONETARIO’. Estas columnas por tanto no aportan información significativa al ser binarias inconclusas y se excluyen del dataset. Adicionalmente, se identifica que NombreDepartamentoAtencion y CodigoDepartamentoAtencion hacen referencia al mismo dato por lo que se unen en la nueva columna DepartamentoAtencion para solo manipular una columna. Lo mismo ocurre con NombreMunicipioAtencion y CodigoMunicipioAtencion quedando en MunicipioAtencion.

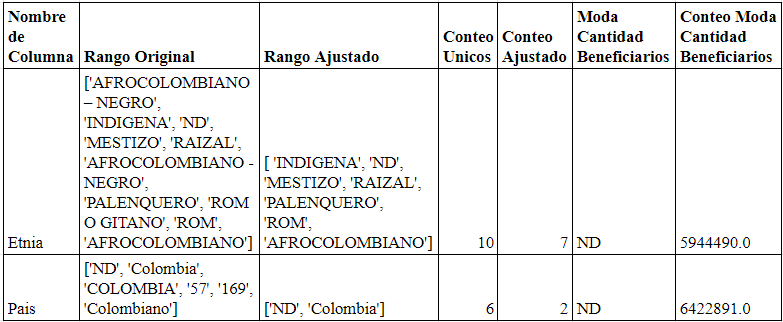


Table 2 Reducción de rangos y medidas de tendencia de las columnas.

Seguimos con el preprocesamiento de datos revisando las columnas con nulos, para ver como completar estas columnas, a las columnas Bancaraizado Genero y TipoPoblacion, les completamos los registros nulos, con los valores de su Moda correspondiente, ya que tenían una proporción de nulos baja, menos de un 20%.

Proseguimos con columnas que tenían una proporción alta de valores faltantes, Etnia y Discapacidad, En etnia los valores nulos, en este caso ND (no definido) eran cerca de un 90%, y solo un poco menos del 10% de los registros tiene alguna etnia, lo cual concuerda con el panorama nacional de acuerdo a las cifras del DANE reportadas en 2018, cerca de un 10% de la población colombiana se identifica como NARP (Negro-Afrocolombiano-Raizal-Palenquero) o indígena en el país[[1]](#footnote-1); la mayoría de la población no pertenece a ninguna etnia, así que los nulos los completamos con el valor NINGUNA, y la Discapacidad con el valor de NO, lo mas seguro es que fueran valores que se omiten a la hora de recolectar los datos. Los cambios en este campo se ven reflejados en la tabla 3:

Tabla

Descripción generada automáticamente

Table 3 Etnia vs cantidad de beneficiarios

Finalmente, para el campo de NivelEscolaridad, se realizó una llenado de los nulos revisando la moda por dataset filtrado de manera que para beneficiarios de rango de edad 6-17, fueron completados con el nivel de escolaridad PRIMARIA, el resto de los registros se le aplicado la media total de los registros, que corresponde a SECUNDARIA.

Tabla

Descripción generada automáticamente con confianza baja

Table 4 NivelEscolaridad vs cantidadDeBeneficiarios

Como se ve en la tabla 4, y luego de realizar los procesos antes mencionados encontramos que los beneficiarios, están en un nivel de escolaridad entre SECUNDARIA Y PRIMARIA, en menor proporción TRANSICION, llama la atención encontrar personas así sea una proporción muy pequeña, que tienen estudios técnicos y tecnólogos, educación superior e igual ser beneficiarios del programa.

Comenzamos nuestra exploración grafica Figure 1 y 2:

Imagen que contiene Forma

Descripción generada automáticamente

Figure 1Cantidad de beneficiarios por fecha de inscripción

Imagen que contiene Forma

Descripción generada automáticamente

Figure 2 Cantidad de beneficiarios por fecha de ultima asignación

Las dos figuras nos muestran el importante sesgo que hay en los datos entre la inscripción, que en su mayoría se dio en los primeros años del programa, vs la última asignación obtenida, esto nos da entender que hay varias personas que seguramente han tomado el beneficio mas de una vez a lo largo del tiempo, siguen siendo beneficiarios desde la primera vez que estuvieron en el programa.

Tabla

Descripción generada automáticamente

Figure 3 TipoBeneficio por rangoedad

La figura 3 nos muestra que existe una coherencia en los registros, para el rango de edad 00-05 no hay prácticamente registros de tipos de beneficios en escuela o colegio, este grupo son bebés o infantes pequeños, por lo que su beneficio en su mayoría es nutrición menor. Los dos grupos poblaciones más grandes de esta tabla son los rangos de 06-17, niños y jóvenes, 18-29 adultos jóvenes. Sobre todo, en el rango de 06-17 el beneficio de Educación Primaria es el que predomina sobre los demás, seguido por educación secundaria y Nutrición Menor, se mantiene la coherencia. Una proporción importante de los registros está en rango 18-29, inicios de adultez donde el beneficio de educación secundaria es el más importante, eso muestra que esta población a pesar de pasar los 18 años, siguen estudiando secundarias seguramente repitentes o validando.

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Figure 4 Distribución de tipos de beneficio otorgados según tipo de documento.

A pesar de contar rangos amplios de edades, se puede tener una idea de edades más exactas en etapas fundamentales como lo son la infancia mediante el tipo de documento del beneficiario. Esto debido a que se conoce las edades exactas en las que un documento se cambia, como lo es el Registro Civil RC (0-7 años), Tarjeta de Identidad TI (7-18 años), y Cédula de Ciudadanía CC (>18 años). Considerando lo anterior y teniendo en cuenta la figura 4, se puede esperar con cierta seguridad una distribución poblacional sesgada a la izquierda en el rango de edad 18-29, ya que la mayoría de los beneficiarios todavía están inscritos con su TI. Lo mismo ocurre con el rango 6-17, ya que alrededor de la mitad todavía usa RC. De tal manera, los porcentajes de distribución de tipos de beneficios según edades de la figura 3 tienen más sentido, considerando que una gran parte de los beneficiarios pueden estar en el límite inferior de su categoría por lo que los niveles de educación menores son acordes a su edad.

Gráfico, Gráfico en cascada

Descripción generada automáticamente

Figure 5 Distribución de tipo de documento de beneficiarios según rango de edad.

La figura 5 muestra una correlación esperable de los tipos de identificación de los beneficiarios y su edad, mostrando coherencia en los datos, donde rangos de 06-17 y 17-29 jóvenes y adultos jóvenes tienen en su mayoría TI, el rango de 00-05 básicamente bebés, con registro civil, que comparten también con el rango 06-17, mostrando que en esta población no hay un afán por actualizar su documentación o puede estar sesgada a niños que apenas están cumpliendo la edad para obtener la TI. Y como era de esperar los rangos de edad 18 a 65 o más en su mayoría ya cuenta con su CC, el valor de CE, no tiene mayor relevancia en la generalidad de los datos.

Tabla

Descripción generada automáticamente con confianza baja

Figure 6 Distribución de tipos de beneficio otorgados según tipo de población

La siguiente figura 6, muestra la relación entre tipo de población y el beneficio recibido, encontramos una mayoría de la población en el programa de SISBEN del gobierno, en este caso la mayoría están en los programas de educación secundaria y primaria, estos programas son los más comunes en todos los tipo de población con casi el 70%, y se evidencia una problemática del país con más de un 25% de beneficiarios desplazados, el otro programa del gobierno UNIDOS tiene un cerca del 16% de la población, con el programa de educación secundaria.

Gráfico de embudo

Descripción generada automáticamente

Figure 7 Distribución de tipos de beneficio otorgados según género

Acá podemos ver que la mayoría de los beneficiarios son mujeres con casi un 60% de los datos, dejando un restante casi 40% para hombres, las mujeres tienen una mayor proporción de beneficios de Nutrición, posiblemente relacionada con madres cabeza de familia.

Interfaz de usuario gráfica

Descripción generada automáticamente con confianza media

Figure 8 Distribución de tipos de beneficio otorgados según rango total del valor del beneficio

En la figura 8 vemos en consolidado de rango de beneficios consolidado asignado con el tipo de beneficio, mostrando que una gran mayoría de la población más de un 80% recibe el rango de Beneficios de 0-15000000, esto corresponde a recibir aunque sea uno de estos apoyos del gobierno el apoyo básico del gobierno, seguramente los otros rangos de beneficios implican haber recibido más de uno de estos beneficios, a los largo de la historia del programa, que son pocos casos, en comparación.

Gráfico

Descripción generada automáticamente

Figure 9 Distribución de tipos de población según rango de edades.

De acuerdo con la figura 9, se evidencia una distribución normal con sesgo a la izquierda en las edades de los beneficiarios, no se evidencia una relación entre el tipo de población y la edad a la que pertenece el beneficiario, la distribución se manifiesta por igual en todos. Esto está acorde con la pirámide poblacional general de Colombia [2].

# MCA y Variables finales

Luego de estos análisis utilizamos la herramienta de MCA (Multiple Correspondence Analysis), la herramienta utilizada para ver relaciones entre variables categóricas para determinar las variables que vamos a utilizar de ahora en adelante en nuestros modelos. Cabe aclarar que en este punto por razones de rendimiento de los modelos comenzamos a trabajar con una muestra aun menor de los registros, 10000 registros igual ponderados.

Escala de tiempo

Descripción generada automáticamente

Figure 10 MCA realizado sobre un set de variables

Como vemos en la figura 10, con este set de variables, se evidencia lo que la exploración de datos previamente nos había mostrado, los beneficios de Nutrición y Nutrición menor están relacionados, a su vez con el nivel de educación primaria, vemos una relación que no revisamos antes entre el rango de edad 18-29 t estados de beneficiario NoActivo, el tipo de población indígena con la etnia indígena. Afrocolombiano y raizal muy relacionado con educción secundaria y este benéfico.

Interfaz de usuario gráfica, Aplicación, Tabla

Descripción generada automáticamente con confianza media

Table 5 Listado de columnas resultantes del preprocesamiento y numeración de valores únicos.

Para algunos ejercicios posteriores también aplicamos One hot encoding, para convertir en numéricas varias de las variables que tenemos, y de esta manera aplicarles técnicas de clusterización basadas en distancias.

# Agrupación

* 1. K Modes:

Aplicamos un método llamado Kmodes que permite hacer clúster sobre variables categóricas. Inicialmente realizamos el análisis con el método de codo para determinar la cantidad de clústeres a realizar:

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Figure 11 Grafico elbow para Kmodes

Dado que este modelo es el optimo para utilizarse en variables categóricas, lo tomamos como modelos base para nuestros siguientes análisis, en pocas palabras los clústeres generados van a ser el referente para los siguientes clústeres y modelos.

* 1. K Means:

Realizamos la aplicación de Kmeans para nuestro conjunto de datos numérico obtenido del One hot encoding,, inicialmente aplicamos el método del codo para determinar la cantidad óptima de clústeres a generar.

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Figure 12 Grafico de elbow para Kmeans

Como vemos en la figura 12 la línea tiende a suavizarse luego de 3 clústeres, por lo que aplicamos este K al conjunto de datos, además ya con las dos validaciones de elbow (Kmenas y Kmodes), dejamos ya estos 3 clústeres para el siguiente modelo de cluster.

* 1. GaussianMixture:

Finalmente aplicamos un modelo de GaussianMixture colocando de parámetro 3 clústeres para que los 3 modelos tengan la misma cantidad de clústeres, evitando que tengan mal rendimiento por que tiene un clúster de mas alguno o de menos.

# Asociación

En el anexo 1 se encuentran los resultados de las reglas de asociación obtenidas por modelo y por clúster:

En términos generales el Cluster\_0 se determina sobre todo por el rango consolidado de ingresos, el cluster\_2, por el nivel de escolaridad SECUNDARIA, y el cluester\_1 tienen el restante sin una regla determinante definida, y con mas registros.

# Comparación resultados Clústeres:

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Figure 13 PCA\_KMODES

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Figure 14 PCA\_KMEANS

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Figure 15 PCA\_GAUSSIAN\_MIXTURE

:



Table 6 Resumen Consolidado

Como podemos ver en las figuras 13,14 y 15 a pesar de que el modelo de Kmodes, utiliza Modas en vez de distancias, obtiene unos resultados bastante similares a modelos netamente cuantitativos como lo son Kmeans y GaussianMixture. y esta similitud también se ve en la comparación de sus medias, logrando muy buenos valores para accuracy, recall, precisión y F1\_score.

Y como era de esperarse al ser datos categóricos, que se muestran siempre espaciados entre sus valores al aplicar one hot encoding, sus valores de distancias internas si son bastante malos.

# Modelos de clasificación

Finalmente tomando los cluesters generados por el Kmodes, vamos a realizar una validación por medio de 3 modelos de clasificación, Random forest, SVM y Multinomial Naive Bayes.

* 1. Random Forest:

Realizamos un random forest, iterando su profundidad de 1 a 10 y 10 iteraciones de árboles de decisión y proporción de test de 0.3, para ver cuál era el resultado mas adecuado para revisar, encontramos que entre profundidad 4 -6 se obtienen buenos resultados, se puede graficar con facilidad, y aunque mejora entre mas profundidad, no es una mejora sustancial, y si complejiza y vuelve muy grande el árbol resultante, dejamos el valor de profundidad de 4, aplica CrossFold validation, para revisar si existía algún desbalanceo en los datos, pero no se encontró nada relevante, que afectara en gran medida el desempeño del modelo.

* 1. SVM

Realizamos un Support Vector Machine; con los siguientes parámetros: kernel='poly', degree=3, C=1, para realizar la comparación con nuestro anterior modelo, para estos también realizamos una variación entre la proporción train/test, con test de 0.1,0.2 y 0.3 para ver sus resultados. El mejor fue con 0.3

* 1. Multinomial Naive Bayes.

Finalmente aplicamos un modelo Naive bayes, con sus parámetros default, para estos también realizamos una variación entre la proporción train/test, con test de 0.1,0.2 y 0.3 para ver sus resultados. El mejor fue con 0.3

* 1. Resultados Consolidados:



Table 7 Resumen modelos clasificación

De los resultados obtenidos, vemos que el que mejor rendimiento tiene es el SVM con un rendimiento casi perfecto en todos los scores, el random forest también consigue unos valores altos si le dejamos la variable de profundidad abierta, lo cual no es tan recomendable.

# Conclusiones

La composición del conjunto en su mayoría por atributos de tipo categórico con pocos valores posibles reduce las necesidades de reducción de dimensiones y preprocesamiento, pero así mismo dificulta y limita las herramientas de exploración para extraer información útil. No se encuentran suficientes relaciones diferenciales de alto peso entre atributos con los procesos llevados a cabo, solo unas relaciones con los atributos que mas aparecen en el dataset.

Encontramos solo unas reglas relacionadas a el rango consolidado del beneficio, al género, y al nivel de educación como las variables que mas sirven para segmentar los datos.

##### Referencias

1. Departamento Administrativo para la Prosperidad Social, 2022. Beneficiarios Más Familias en Acción. [online] Datos.gov.co. Available at: <https://www.datos.gov.co/Inclusi-n-Social-y-Reconciliaci-n/Beneficiarios-M-s-Familias-en-Acci-n/xfif-myr2> [Accessed 29 August 2022].
2. datosmacro.com. 2021. Colombia - Piramide de población 2021. [online] Available at: <https://datosmacro.expansion.com/demografia/estructura-poblacion/colombia> [Accessed 29 August 2022].
3. Prosperidad Social. 2022. Familias en Acción - Transferencias Condicionadas de Prosperidad Social. [online] Available at: <https://prosperidadsocial.gov.co/sgpp/transferencias/familias-en-accion/> [Accessed 14 September 2022].
4. United Nations, 2022. Human Development Index. [online] Hdr.undp.org. Available at: <https://hdr.undp.org/data-center/human-development-index#/indicies/HDI> [Accessed 14 September 2022].

Anexo 1. Reglas de asociación por Modelo y CLuster:

KMODES:

A priori:

**Cluster\_0:**

{Etnia\_NINGUNA, NivelEscolaridad\_SECUNDARIA, RangoEdad\_18-29, Titular\_NO} -> {RangoBeneficioConsolidadoAsignado\_0 - 1.500.000} (conf: 1.000, supp: 0.644, lift: 1.085, conv: 78162622.708)

{Etnia\_NINGUNA, RangoEdad\_18-29, TipoBeneficio\_EDUCACIÓN SECUNDARIA, Titular\_NO} -> {RangoBeneficioConsolidadoAsignado\_0 - 1.500.000} (conf: 1.000, supp: 0.637, lift: 1.085, conv: 78162622.708)

{NivelEscolaridad\_SECUNDARIA, RangoEdad\_18-29, TipoBeneficio\_EDUCACIÓN SECUNDARIA, Titular\_NO} -> {RangoBeneficioConsolidadoAsignado\_0 - 1.500.000} (conf: 1.000, supp: 0.714, lift: 1.085, conv: 78162622.708)

**Cluster\_1:**

{NivelEscolaridad\_PRIMARIA, RangoEdad\_06-17} -> {RangoBeneficioConsolidadoAsignado\_0 - 1.500.000, Titular\_NO} (conf: 1.000, supp: 0.959, lift: 1.000, conv: 0.000)

{NivelEscolaridad\_PRIMARIA, RangoEdad\_06-17} -> {Titular\_NO} (conf: 1.000, supp: 0.959, lift: 1.000, conv: 0.000)

{NivelEscolaridad\_PRIMARIA, RangoBeneficioConsolidadoAsignado\_0 - 1.500.000} -> {Titular\_NO} (conf: 1.000, supp: 0.968, lift: 1.000, conv: 0.000)

**Cluster\_2:**

{EstadoBeneficiario\_ACTIVO, Etnia\_NINGUNA, Titular\_SI} -> {NivelEscolaridad\_SECUNDARIA} (conf: 1.000, supp: 0.801, lift: 1.000, conv: 0.000)

{EstadoBeneficiario\_ACTIVO, Genero\_Mujer, Titular\_SI} -> {NivelEscolaridad\_SECUNDARIA} (conf: 1.000, supp: 0.838, lift: 1.000, conv: 0.000)

{EstadoBeneficiario\_ACTIVO, Etnia\_NINGUNA} -> {NivelEscolaridad\_SECUNDARIA} (conf: 1.000, supp: 0.801, lift: 1.000, conv: 0.000)

FP\_growth:

**Cluster\_0**

('Etnia\_NINGUNA', 'RangoEdad\_18-29', 'TipoBeneficio\_EDUCACIÓN SECUNDARIA', 'Titular\_NO')

('Etnia\_NINGUNA', 'NivelEscolaridad\_SECUNDARIA', 'RangoEdad\_18-29', 'TipoBeneficio\_EDUCACIÓN SECUNDARIA', 'Titular\_NO')

('Etnia\_NINGUNA', 'NivelEscolaridad\_SECUNDARIA', 'RangoEdad\_18-29', 'Titular\_NO')

('RangoEdad\_18-29', 'TipoBeneficio\_EDUCACIÓN SECUNDARIA', 'Titular\_NO')

('NivelEscolaridad\_SECUNDARIA', 'RangoEdad\_18-29', 'TipoBeneficio\_EDUCACIÓN SECUNDARIA', 'Titular\_NO')

('NivelEscolaridad\_SECUNDARIA', 'RangoEdad\_18-29', 'Titular\_NO')

**Cluster\_1:**

('EstadoBeneficiario\_ACTIVO', 'Etnia\_NINGUNA', 'NivelEscolaridad\_PRIMARIA', 'RangoBeneficioConsolidadoAsignado\_0 - 1.500.000', 'RangoEdad\_06-17')

('EstadoBeneficiario\_ACTIVO', 'Etnia\_NINGUNA', 'NivelEscolaridad\_PRIMARIA', 'RangoEdad\_06-17', 'Titular\_NO')

('EstadoBeneficiario\_ACTIVO', 'NivelEscolaridad\_PRIMARIA', 'RangoBeneficioConsolidadoAsignado\_0 - 1.500.000', 'RangoEdad\_06-17')

**Cluster\_2:**

('EstadoBeneficiario\_ACTIVO', 'Genero\_Mujer')

('EstadoBeneficiario\_ACTIVO', 'Genero\_Mujer', 'Titular\_SI')

('EstadoBeneficiario\_ACTIVO', 'Titular\_SI')

KMEANS

A priori:

**Cluster\_0:**

{Etnia\_NINGUNA, RangoEdad\_18-29, TipoBeneficio\_EDUCACIÓN SECUNDARIA, Titular\_NO} -> {RangoBeneficioConsolidadoAsignado\_0 - 1.500.000} (conf: 1.000, supp: 0.738, lift: 1.000, conv: 214592.275)

{NivelEscolaridad\_SECUNDARIA, RangoEdad\_18-29, TipoBeneficio\_EDUCACIÓN SECUNDARIA, Titular\_NO} -> {RangoBeneficioConsolidadoAsignado\_0 - 1.500.000} (conf: 1.000, supp: 0.827, lift: 1.000, conv: 214592.275)

{Etnia\_NINGUNA, NivelEscolaridad\_SECUNDARIA, RangoEdad\_18-29, TipoBeneficio\_EDUCACIÓN SECUNDARIA, Titular\_NO} -> {RangoBeneficioConsolidadoAsignado\_0 - 1.500.000} (conf: 1.000, supp: 0.737, lift: 1.000, conv: 214592.275)

**Cluster\_1**

{RangoEdad\_06-17} -> {RangoBeneficioConsolidadoAsignado\_0 - 1.500.000, Titular\_NO} (conf: 1.000, supp: 0.989, lift: 1.000, conv: 0.000)

{NivelEscolaridad\_PRIMARIA, RangoEdad\_06-17} -> {RangoBeneficioConsolidadoAsignado\_0 - 1.500.000, Titular\_NO} (conf: 1.000, supp: 0.931, lift: 1.000, conv: 0.000)

{NivelEscolaridad\_PRIMARIA} -> {RangoBeneficioConsolidadoAsignado\_0 - 1.500.000, Titular\_NO} (conf: 1.000, supp: 0.942, lift: 1.000, conv: 0.000)

**Cluster\_2**

{Etnia\_NINGUNA, Titular\_SI} -> {NivelEscolaridad\_SECUNDARIA} (conf: 1.000, supp: 0.911, lift: 1.000, conv: 0.000)

{Genero\_Mujer, TipoBeneficio\_EDUCACIÓN SECUNDARIA} -> {NivelEscolaridad\_SECUNDARIA} (conf: 1.000, supp: 0.802, lift: 1.000, conv: 0.000)

{Genero\_Mujer, Titular\_SI} -> {NivelEscolaridad\_SECUNDARIA} (conf: 1.000, supp: 0.925, lift: 1.000, conv: 0.000)

FP\_growth

**Cluster\_0**

('Genero\_Mujer', 'TipoBeneficio\_EDUCACIÓN SECUNDARIA', 'Titular\_NO')

('Genero\_Mujer', 'NivelEscolaridad\_SECUNDARIA', 'TipoBeneficio\_EDUCACIÓN SECUNDARIA', 'Titular\_NO')

('Genero\_Mujer', 'NivelEscolaridad\_SECUNDARIA', 'TipoBeneficio\_EDUCACIÓN SECUNDARIA')

**Clúster\_1**

('EstadoBeneficiario\_ACTIVO', 'RangoBeneficioConsolidadoAsignado\_0 - 1.500.000')

('EstadoBeneficiario\_ACTIVO', 'Titular\_NO')

('EstadoBeneficiario\_ACTIVO', 'RangoBeneficioConsolidadoAsignado\_0 - 1.500.000', 'RangoEdad\_06-17')

**Clúster\_2**

**(**'Genero\_Mujer', 'NivelEscolaridad\_SECUNDARIA', 'Titular\_SI')

('Genero\_Mujer', 'TipoBeneficio\_EDUCACIÓN SECUNDARIA', 'Titular\_SI')

('EstadoBeneficiario\_ACTIVO', 'Etnia\_NINGUNA')

GAUSSIANMIXTURE:

**Cluster\_0:**

{NivelEscolaridad\_PRIMARIA, RangoEdad\_06-17, Titular\_NO} -> {RangoBeneficioConsolidadoAsignado\_0 - 1.500.000} (conf: 1.000, supp: 0.933, lift: 1.000, conv: 0.000)

{NivelEscolaridad\_PRIMARIA, RangoBeneficioConsolidadoAsignado\_0 - 1.500.000, RangoEdad\_06-17} -> {Titular\_NO} (conf: 1.000, supp: 0.933, lift: 1.000, conv: 0.000)

{NivelEscolaridad\_PRIMARIA, RangoEdad\_06-17} -> {RangoBeneficioConsolidadoAsignado\_0 - 1.500.000, Titular\_NO} (conf: 1.000, supp: 0.933, lift: 1.000, conv: 0.000)

**Cluster\_1**

{NivelEscolaridad\_SECUNDARIA, TipoBeneficio\_EDUCACIÓN SECUNDARIA, Titular\_NO} -> {RangoBeneficioConsolidadoAsignado\_0 - 1.500.000} (conf: 1.000, supp: 0.983, lift: 1.000, conv: 0.000)

{NivelEscolaridad\_SECUNDARIA, RangoBeneficioConsolidadoAsignado\_0 - 1.500.000, TipoBeneficio\_EDUCACIÓN SECUNDARIA} -> {Titular\_NO} (conf: 1.000, supp: 0.983, lift: 1.000, conv: 0.000)

{NivelEscolaridad\_SECUNDARIA, TipoBeneficio\_EDUCACIÓN SECUNDARIA} -> {RangoBeneficioConsolidadoAsignado\_0 - 1.500.000, Titular\_NO} (conf: 1.000, supp: 0.983, lift: 1.000, conv: 0.000)

**Cluster\_2**

{EstadoBeneficiario\_ACTIVO, Genero\_Mujer, TipoBeneficio\_EDUCACIÓN SECUNDARIA, Titular\_SI} -> {NivelEscolaridad\_SECUNDARIA} (conf: 1.000, supp: 0.731, lift: 1.000, conv: 0.000)

{Etnia\_NINGUNA, Genero\_Mujer, TipoBeneficio\_EDUCACIÓN SECUNDARIA, Titular\_SI} -> {NivelEscolaridad\_SECUNDARIA} (conf: 1.000, supp: 0.723, lift: 1.000, conv: 0.000)

{EstadoBeneficiario\_ACTIVO, Etnia\_NINGUNA, Genero\_Mujer, TipoBeneficio\_EDUCACIÓN SECUNDARIA, Titular\_SI} -> {NivelEscolaridad\_SECUNDARIA} (conf: 1.000, supp: 0.656, lift: 1.000, conv: 0.000)

FP\_growth

**Cluster\_0**

('EstadoBeneficiario\_ACTIVO', 'RangoBeneficioConsolidadoAsignado\_0 - 1.500.000')

('EstadoBeneficiario\_ACTIVO', 'Titular\_NO')

('EstadoBeneficiario\_ACTIVO', 'RangoBeneficioConsolidadoAsignado\_0 - 1.500.000', 'RangoEdad\_06-17')

**Cluster\_1**

('Etnia\_NINGUNA', 'NivelEscolaridad\_SECUNDARIA', 'RangoBeneficioConsolidadoAsignado\_0 - 1.500.000')

('Etnia\_NINGUNA', 'NivelEscolaridad\_SECUNDARIA', 'Titular\_NO')

('Etnia\_NINGUNA', 'RangoBeneficioConsolidadoAsignado\_0 - 1.500.000')

**Cluster\_2**

('TipoBeneficio\_EDUCACIÓN SECUNDARIA', 'Titular\_SI')

('EstadoBeneficiario\_ACTIVO', 'Etnia\_NINGUNA')

('EstadoBeneficiario\_ACTIVO', 'Etnia\_NINGUNA', 'Titular\_SI')

ANEXO 2 resultado Random Forest: tree.jpg

1. Dane, 2018 <https://www.dane.gov.co/files/investigaciones/boletines/grupos-etnicos/presentacion-grupos-etnicos-2019.pdf> [↑](#footnote-ref-1)