# **Diplomatura en análisis de datos aplicados al desarrollo de políticas públicas**

****

**Título del tema de trabajo Final**: Programa Cambio Rural - SAGyP

**Integrantes** Constanza Guerrini, Patricia Perrone, Marianela Pi

**Objetivos Generales**

El objetivo general de esta investigación es caracterizar a los beneficiarios/as del Programa Cambio Rural.

Cambio Rural es una política pública que busca, a través de la asistencia técnica, promover y facilitar la intensificación y reconversión productiva, como un medio para mejorar la situación productiva y socioeconómica de los pequeños y medianos productores rurales y propender al desarrollo agroindustrial en todo el territorio nacional, impulsando el aprendizaje grupal.

**Objetivos Específicos**

Obtener un *dataset* limpio, determinar las variables a utilizar para la caracterización,

Análisis estadísticos

**Antecedentes:**

La situación actual es la reestructuración del Programa Federal de Reconversión Productiva que tuvo lugar en septiembre de 2017 y que fue relanzado como Programa Cambio Rural (CR) a través de la Resolución E 249/2017, junto con un nuevo Manual Operativo. En ese mismo documento se crea el Registro de Integrantes de Grupos Cambio Rural y se desarrolla un Sistema de Gestión para sistematizar la información de este Registro. La decisión de reestructuración se tomó debido a que, a pesar de los esfuerzos y recursos invertidos, los resultados obtenidos no fueron suficientes para posicionar a la pequeña y mediana empresa rural en los niveles óptimos y necesarios de eficiencia productiva que les permitieran enfrentar exitosamente las fluctuaciones económicas y climáticas.

Como antecedente, se tomó la Resolución N° 227 de fecha 4 de mayo de 1993 de la Secretaría de Agricultura, Ganadería y Pesca del entonces Ministerio de Economía y Obras y Servicios Públicos, que creó el Programa Federal de Reconversión Productiva con el propósito de promover y facilitar la intensificación y reconversión productiva de la pequeña y mediana empresa rural. La creación del programa se llevó a cabo en el contexto histórico de los años 90 y ante la crisis económica reinante, y se solicitó propuestas al sector, una de las cuales, elaborada por el INTA, se convirtió en el Programa Federal de Reconversión Productiva. En su origen, el Programa otorgó gran importancia al trabajo coordinado con las Provincias y las entidades del sector, incluidas las intermedias, para posibilitar y potenciar la asistencia técnica, el acceso al crédito y el intercambio tecnológico necesario para una mayor eficiencia y diversificación productiva que, junto al esfuerzo asociativo, generaran economías competitivas.

Políticas similares

https://agriculture.ec.europa.eu/common-agricultural-policy/rural-development\_es

**Actividades y metodología:**

1. Ver las variables relevadas
2. Selección de las variables a trabajar
3. Limpieza de datos
4. Homogeneización de datos
5. Generación dataset
6. Análisis estadístico de las variables

**Factibilidad:**

Es factible trabajar con variables categóricas en un análisis de clustering, pero se requiere una técnica específica para manejarlas adecuadamente. Por ejemplo, se podría utilizar una técnica de codificación de variables categóricas, como la codificación "one-hot" (también conocida como "dummy variables"), que consiste en crear una variable binaria para cada posible valor de la variable categórica original.

Sin embargo, trabajar con demasiadas variables categóricas puede ser problemático, ya que esto puede generar una gran cantidad de variables binarias en el análisis, lo que puede afectar la calidad del modelo. Además, trabajar con variables categóricas puede requerir un mayor conocimiento experto y experiencia en la selección de técnicas de clustering adecuadas y en la interpretación de los resultados que todavia no tenemos y requiere más de dos meses.

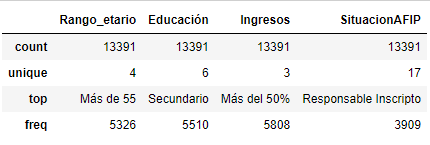
En resumen, es factible trabajar con variables categóricas en un análisis de clustering, pero es importante considerar cuidadosamente el número y la calidad de las variables, así como la selección de técnicas de clustering apropiadas y la interpretación adecuada de los resultados.

Resultados (3 páginas máximo)

Se realizó un análisis de la tabla y se identificaron varios problemas en las variables de edad y superficie. Se intentó resolver el problema de los datos de la superficie utilizando una transformación logarítmica, pero se encontró que no se podían convertir los valores "NAN" a 0 debido a que este valor es válido, ya que hay productores que no tienen ninguna hectárea. En la edad utilizamos el rango etario mayores a 18 años y menores de 99.

Debido a esto se decidió trabajar con las siguientes variables categóricas, rango etario, educación, ingresos y situación AFIP.

**Resumen estadístico de las cuatro variables**

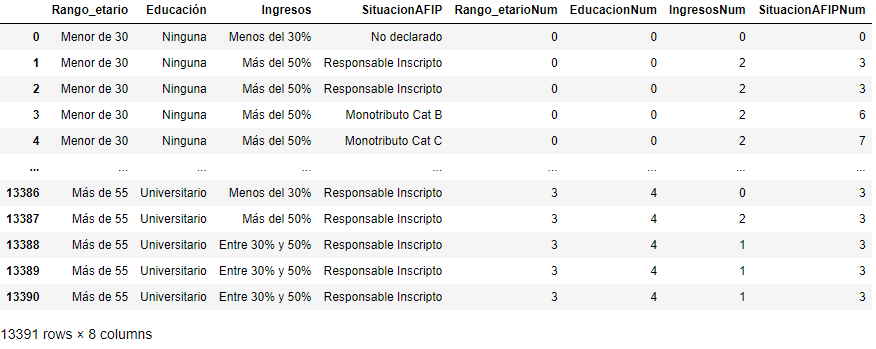


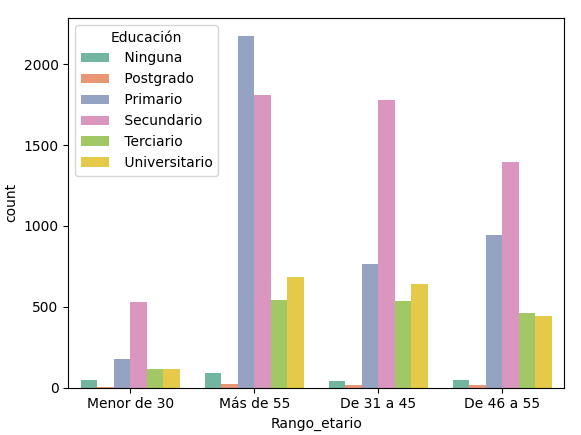
El valor "top" en cada variable muestra la categoría más frecuente en esa variable, lo que indica que hay 5,326 registros con la categoría "Más de 55" en la variable "Rango\_etario", 5,510 registros con la categoría "Secundario" en la variable "Educación", 5,808 registros con la categoría "Más del 50%" en la variable "Ingresos" y 3,909 registros con la categoría "Responsable Inscripto" en la variable "SituacionAFIP".

**Construcción del Modelo**

Nuestro objetivo en este cuaderno es solo demostrar el algoritmo K-modes, y omitiremos en esta ocasión el EDA y nos pasaremos directamente a la construcción del modelo

Contrucción del modelo-Convertimos las variables categóricas a numéricas



En este gráfico, se puede observar que en el rango etario de más de 55 años, hay una mayor proporción de individuos que solo han completado la educación primaria.

**Comenzamos a utilizar los métodos para clusterizar**

Usando K-Modes con Inicialización "CAO" y K-Modes con Inicialización "Huang"

K-Modes con Inicialización "Huang"

Run 3, iteration: 1/100, moves: 3953, cost: 25707.0

Init: initializing centroids

Init: initializing clusters

Starting iterations...

Run 4, iteration: 1/100, moves: 4323, cost: 24331.0

Run 4, iteration: 2/100, moves: 500, cost: 24331.0

Init: initializing centroids

Init: initializing clusters

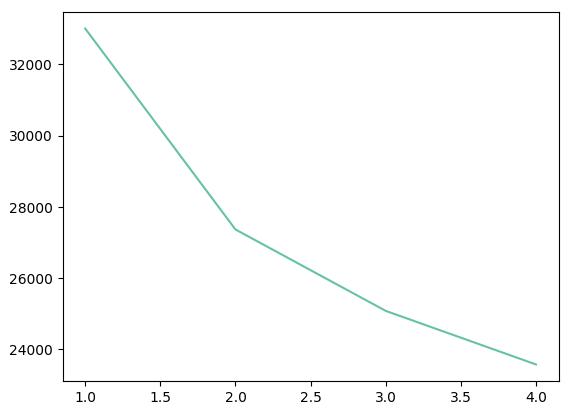
Starting iterations...

Run 5, iteration: 1/100, moves: 2880, cost: 24710.0

Run 5, iteration: 2/100, moves: 1089, cost: 24710.0

Best run was number 4

En cada "Run" se registra el número de iteración actual, el número de movimientos realizados, el costo actual y el número de la ejecución. Se puede observar que en la ejecución número 4 se logró el menor costo, lo que indica que esa es la mejor solución encontrada por el algoritmo.



### Características de los clusters

count 1797

unique 4

top De 46 a 55

freq 1385

Name: Rango\_etario, dtype: object

count 1797

unique 6

top Secundario

freq 1058

Name: Educación, dtype: object

count 1797

unique 3

top Entre 30% y 50%

freq 1128

Name: Ingresos, dtype: object

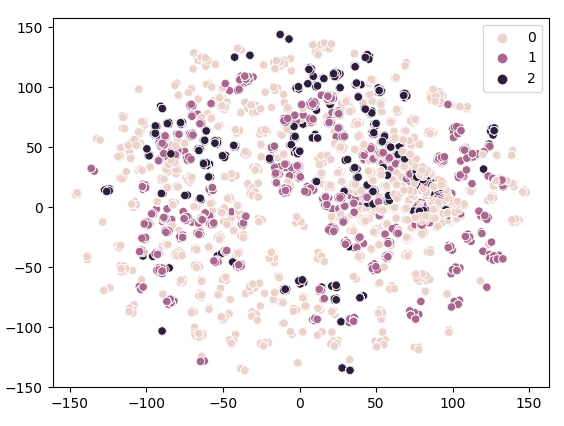
count 1797

unique 17

top Monotributo Cat A

freq 688

Name: SituacionAFIP, dtype: object



**Material adicional:**

*Material adicional relevante, desarrollos programas, visualizaciones, documentos generales, y extensiones que permitan evaluar el trabajo*.

**Referencias:**

<https://biblioguias.uma.es/citasybibliografia/ejemplosAPA>

<https://www.kaggle.com/code/halflingwizard/clustering-categorical-data-using-gower-distance>

<https://www.youtube.com/watch?v=S5cL5MAFon8>

https://www.youtube.com/watch?v=o4bn2ZEGr4g