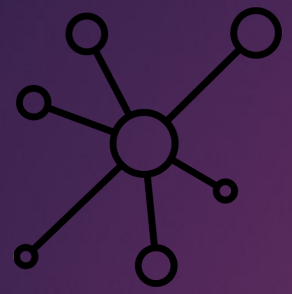


Proyecto Minería de Datos

CASEN

Integrantes:

Nicolás Gatica Cristóbal Igor Nelson Moreno Emilio Roa Thomas Waldura

Profesor(a):

Felipe Bravo Bárbara Poblete

Mejorar hito 1: Nuestras Preguntas

Una observación del hito 1 fue que la mayor parte de nuestras preguntas se pueden responder trivialmente mediante exploración de datos y cálculos de relaciones. Para resolver eso decidimos plantear nuevas preguntas las cuales son:

Problemas encontrados en el Hito 1

Después de explorar nuestros datos de nuevo nos planteamos lo siguiente:

1. ¿Es posible establecer una relación entre la distancia de clusters y la desigualdad económica entre la distintas comunas?
2. ¿Es posible predecir comunas en base a los materiales de la construcción de sus viviendas, qué materiales caracterizan a las comunas?
3. ¿Es posible a través de los clusters formados en los datasets de salud, educación, vivienda y bancarios o a través de los atributos presentes en ellos encontrar reglas de asociación que permitan establecer un rango o situación socioeconómica a la cual se pertenece?, ¿Cuáles serían estas reglas de asociación?

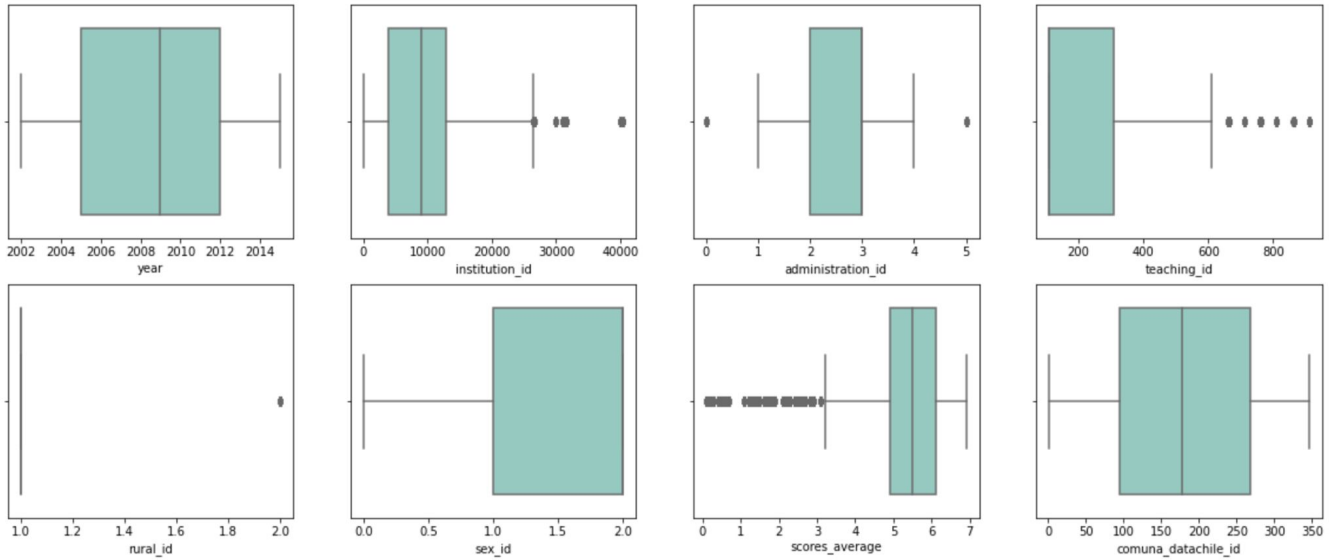
Mejorar hito 1: Nuevo dataset - Educación

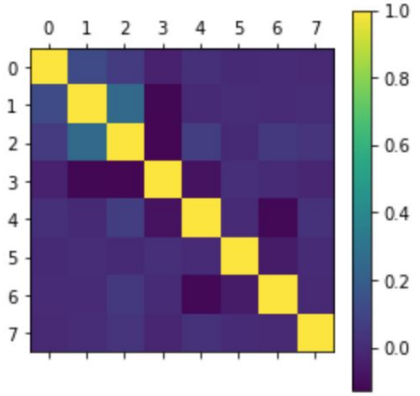
Para responder correctamente a nuestras nuevas preguntas decidimos agregar un nuevo dataset que viene de la misma fuente que los otros datos, es decir, datachile.io, y que presenta notas para cada institución escolar del país de cada comuna entre 2002 y 2015. A continuación algunas visualizaciones de su exploración:

Análisis general

El dataframe tiene 8 columnas de tipos int64 y float64, 37.369.703 filas con 0,00% de valores faltantes. Como se ve abajo en los boxplots hay algunos outliers.

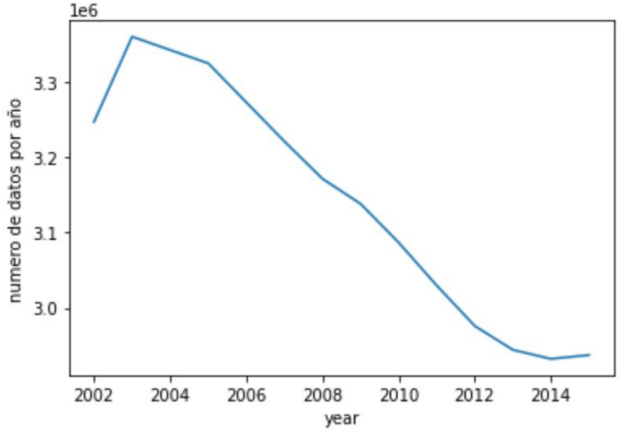
Análisis estadístico de cada columna





Correlación

Hay poca correlación entre los atributos. Lo que es bastante lógico dada la estructura del dataframe



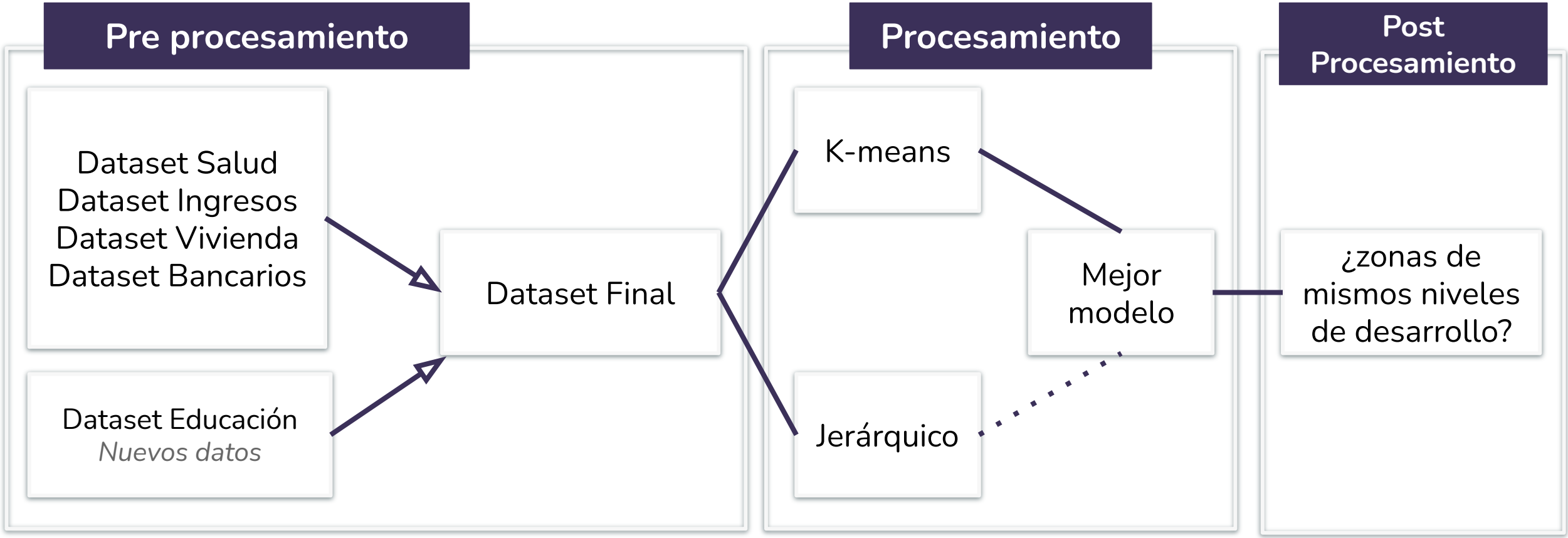
Análisis temporal

El número de datos a lo largo del tiempo no es homogénea.

Hito 2: Propuesta experimental pregunta 1

Para responder a la pregunta : **¿Es posible establecer una relación entre la distancia de clusters y la desigualdad económica entre la distintas comunas?**

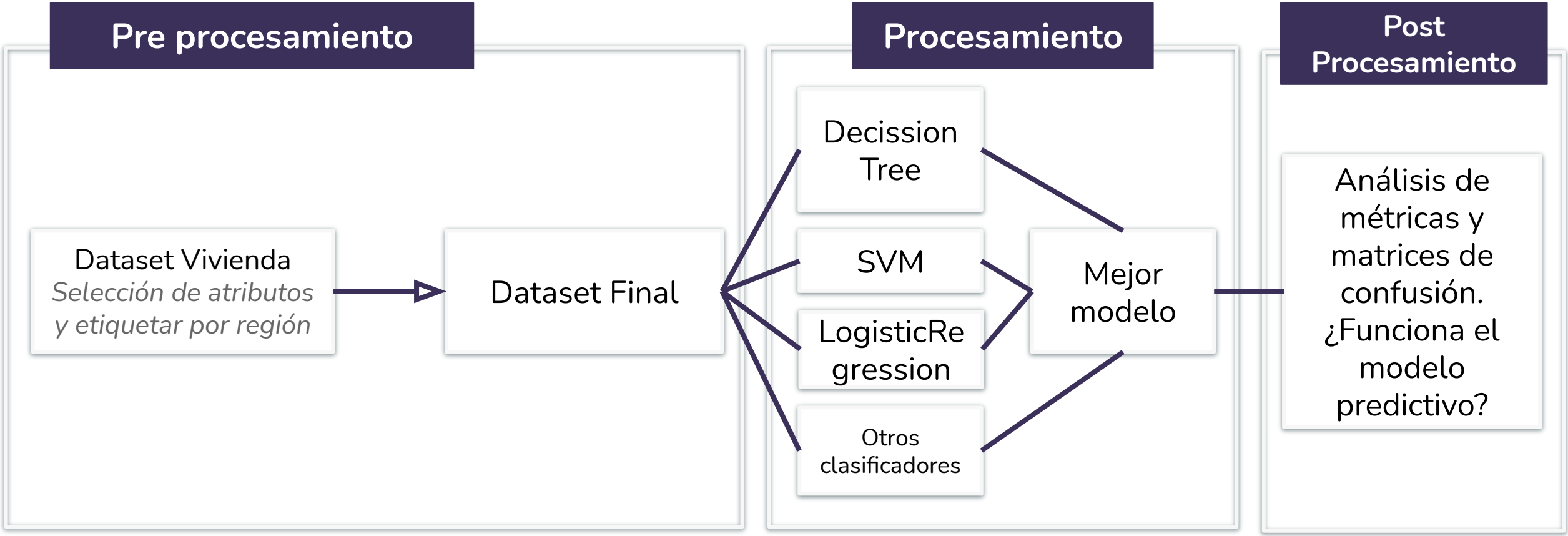
Este es la propuesta experimental:



Hito 2: Propuesta experimental pregunta 2

Para responder a la pregunta : **¿Es posible predecir comunas en base a los materiales de la construcción de sus viviendas, qué materiales caracterizan a las comunas?**

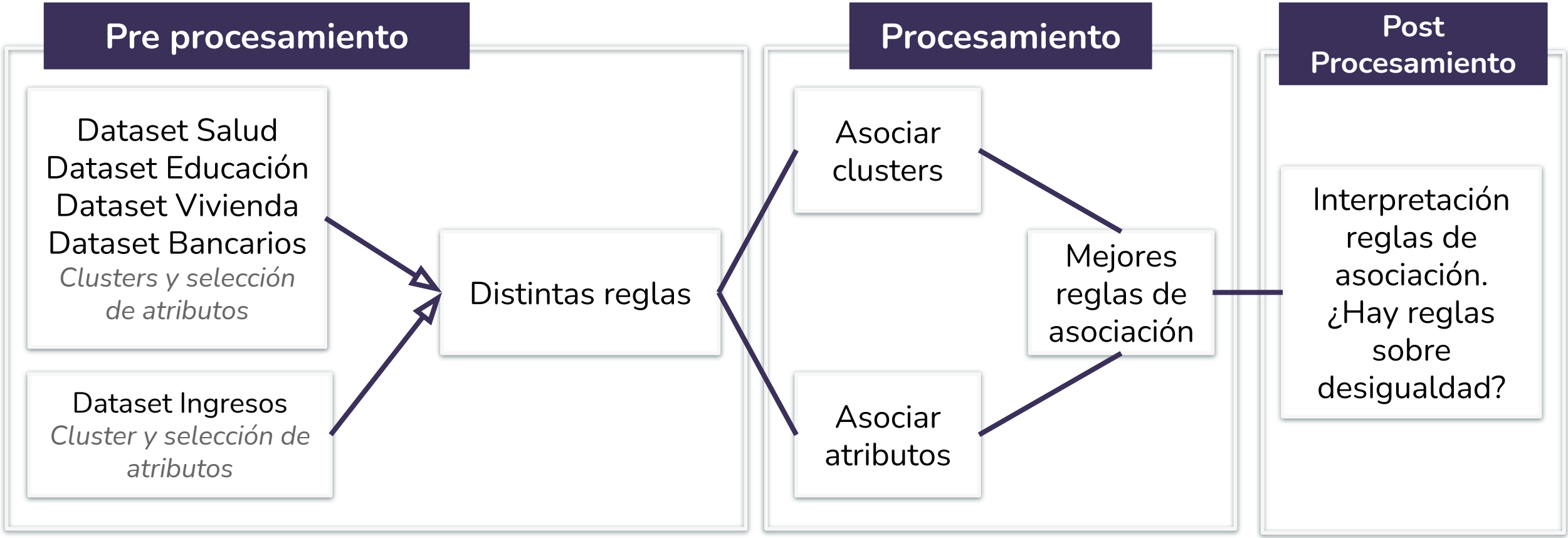
Este es la propuesta experimental:



Hito 2: Propuesta experimental pregunta 3

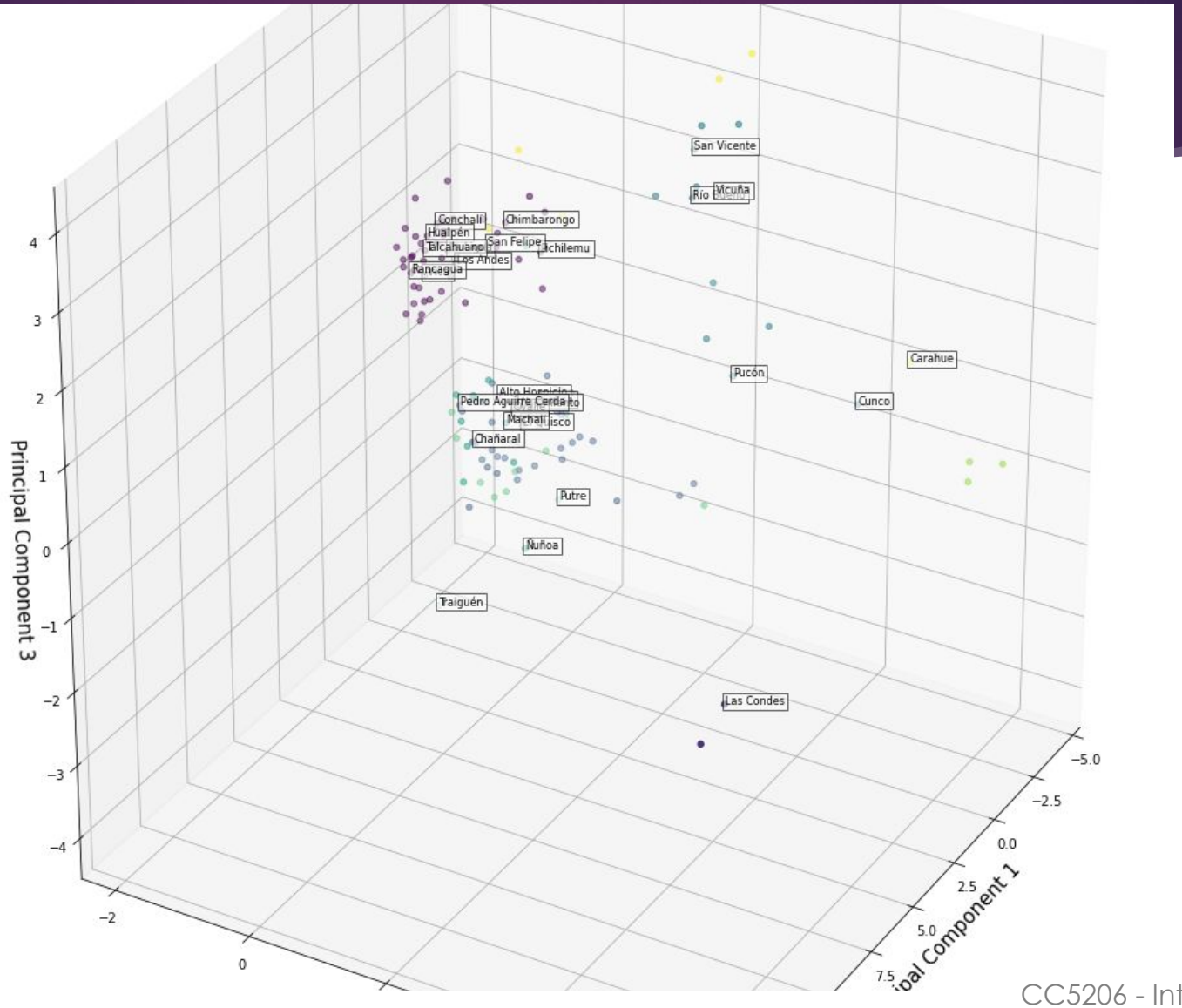
Para responder a la pregunta : **¿Es posible a través de los clusters formados en los datasets de salud, educación, vivienda y bancarios o a través de los atributos presentes en ellos encontrar reglas de asociación que permitan establecer un rango o situación socioeconómica a la cual se pertenece?, ¿Cuáles serían estas reglas de asociación?**

Este es la propuesta experimental:



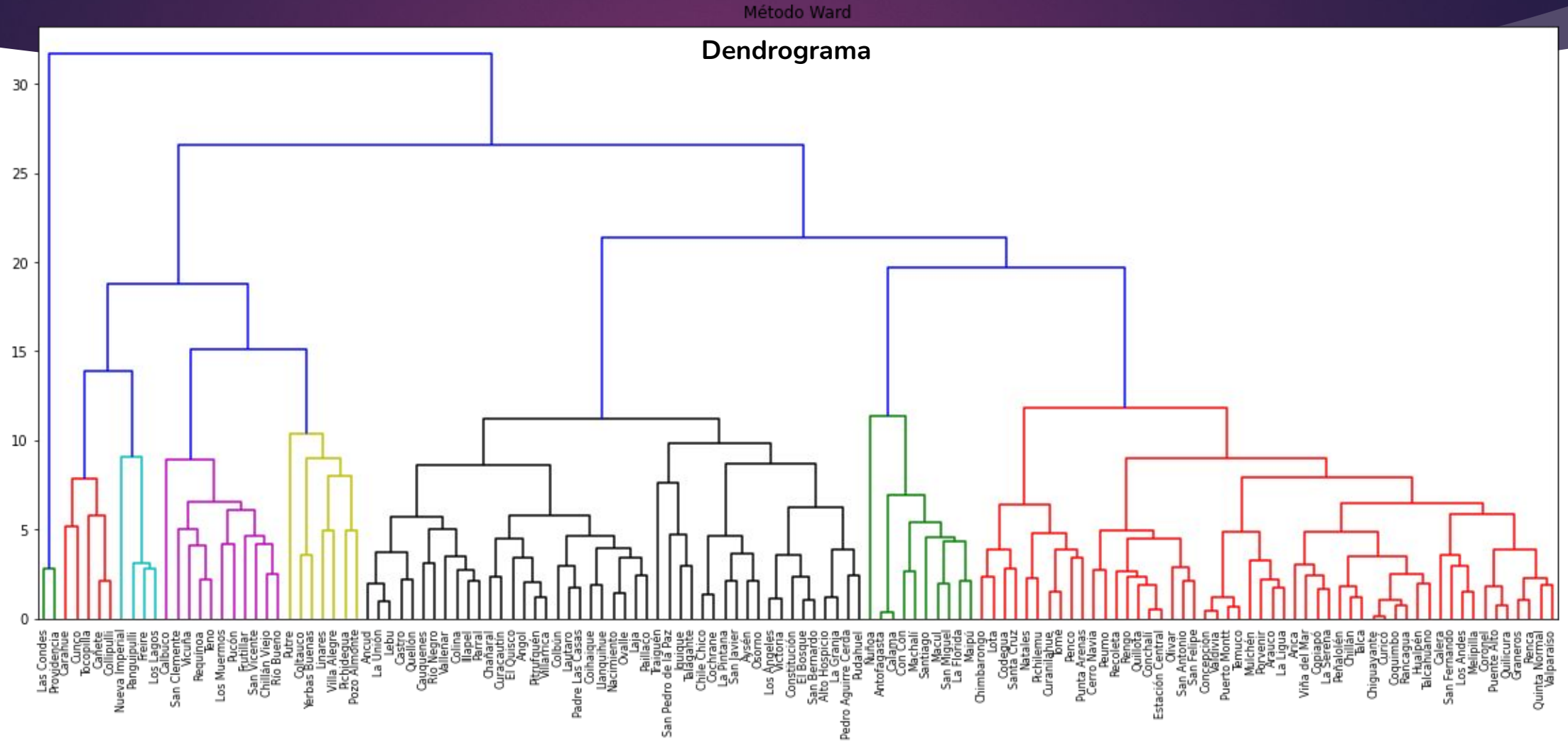
Resultados Preliminares pregunta 1

Resultados de hacer K-Means sobre todos los datasets y ploteando en un PCA de 3 dimensiones



Resultados Preliminares pregunta 1

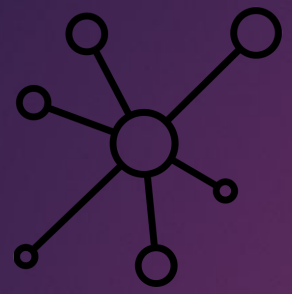
Resultado de aplicar un dendrograma sobre los atributos más importantes en cuanto a desigualdad de todos los datasets



Conclusiones

¿Qué encontramos tras la clusterización?

- 8 Clusters (3 Grandes grupos)
- Diferencias entre distancias de clusters
- Evidencia de la gran brecha socioeconómica del país
- Influencia que tiene la diferencia de clases no solo sobre Ingresos y Datos Bancarios sino también sobre Educación, Vivienda, Salud.



Proyecto Minería de Datos

MUCHAS GRACIAS

Repositorio del proyecto, la **Base de datos** y **Análisis**:
[Github](https://github.com/Bruno-Moreno/fondaMinera)