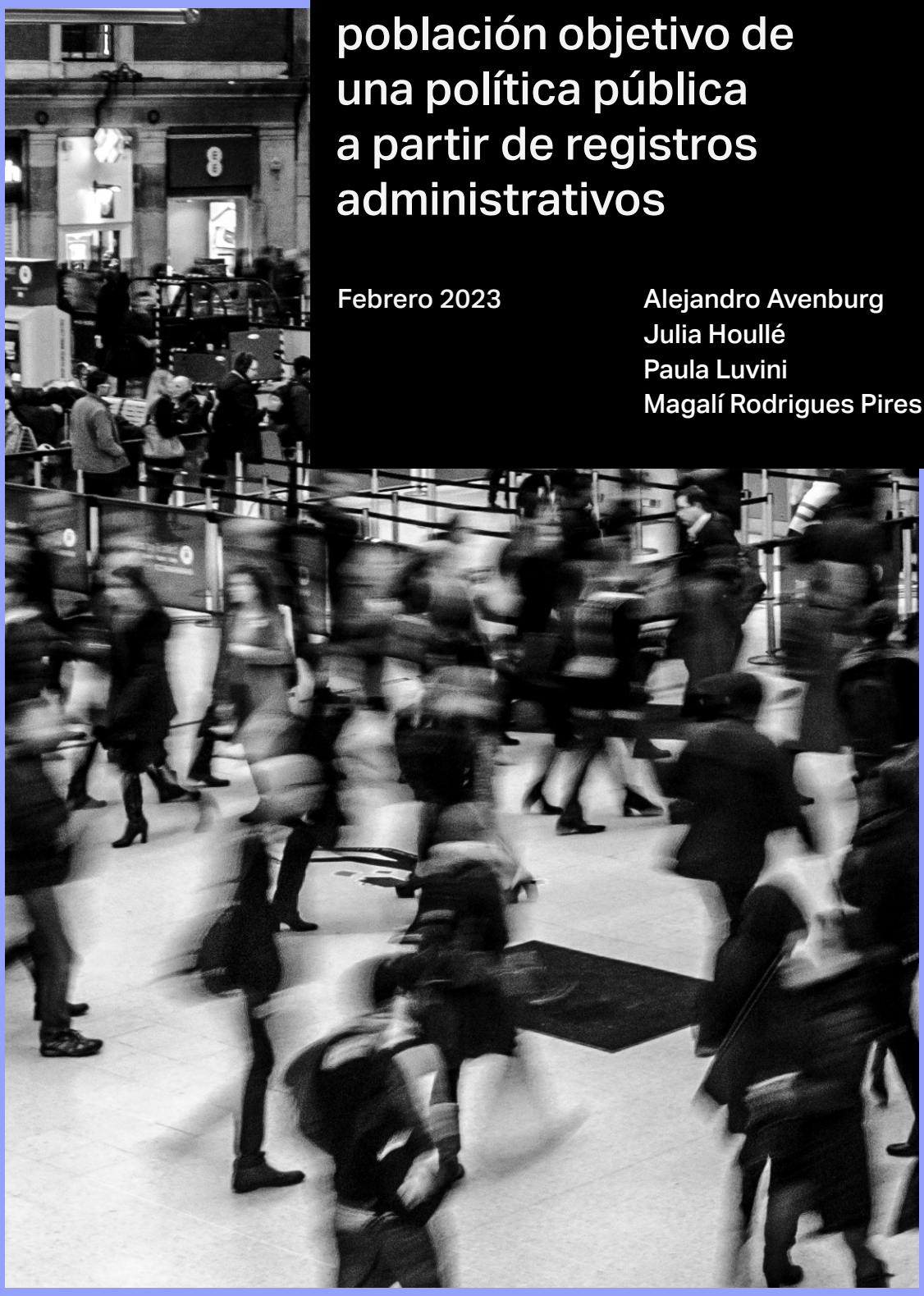
fundar

Ministerio de Cultura
Argentina

Área de
Datos

# Guía práctica para caracterizar a la población objetivo de una política pública a partir de registros administrativos

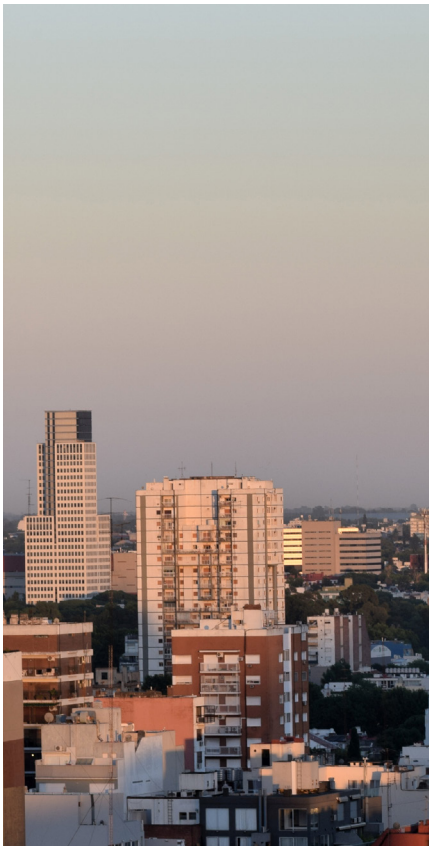


Febrero 2023

Alejandro Avenburg
Julia Houllé
Paula Luvini
Magalí Rodrigues Pires

Área de
Datos

Febrero 2023



# Guía práctica para caracterizar a la población objetivo de una política pública a partir de registros administrativos

El uso de registros administrativos es una herramienta muy frecuente en la Administración Pública: permite a los/as ciudadanos/as registrarse en programas, solicitar subsidios, y puede facilitar la interacción con el Estado. Esta disponibilización de datos, acompañada por el desarrollo tecnológico y de algoritmos de aprendizaje automático, permite realizar análisis novedosos para dotar de información y herramientas a los/las decisores/as de políticas públicas. La presente guía busca orientar este tipo de proyectos de manera general: ¿cómo generar una herramienta para conocer el público objetivo de un ministerio a partir de registros administrativos?

Para ello, se describe la experiencia colaborativa entre el Área de Datos de Fundar y la Dirección de Planificación y Seguimiento de Gestión del Ministerio de Cultura de Nación, donde se buscó identificar herramientas útiles para caracterizar a la población objetivo de las políticas culturales.

Área de Datos

Guía práctica para caracterizar a la población objetivo de una política pública a partir de registros administrativos

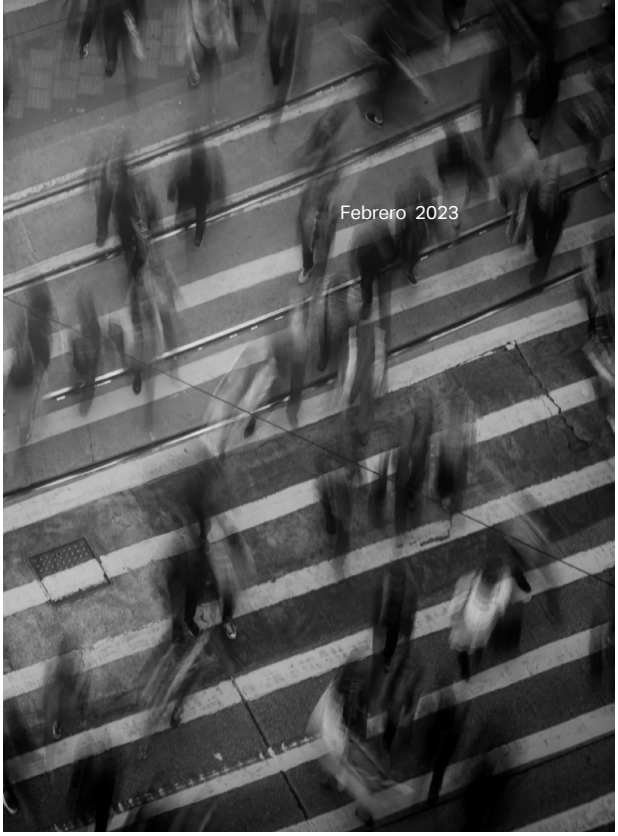
Febrero 2023

# Población objetivo en políticas públicas

¿Cómo podemos conocer mejor este público objetivo[1]? Un recurso muy útil para este fin es hacer análisis de subgrupos, es decir identificar características comunes entre diversos subgrupos más homogéneos dentro de un universo amplio y diverso. A través de este análisis, podemos identificar mejor cuáles son las características, necesidades, beneficios estatales que han obtenido (o no) grupos con distintas características.

En el caso de la población objetivo del Ministerio de Cultura de Nación, existen perfiles tan diversos como el de **escenógrafos/as, artesanos/as, músicos/as, iluminadores/as, actores/actrices, docentes de disciplinas artísticas,** entre otros, así como un amplio espectro de destinatarios/as en términos geográficos, de situación laboral, etarios. Por eso, la segmentación en diversos perfiles puede ayudar a dar mayor precisión a los objetivos de cualquier política pública para el sector.

1 Toda política pública tiene una población o público objetivo. La definición de tal público estará mediada por una definición del ámbito de alcance de las políticas públicas de cualquier agencia gubernamental.



Febrero 2023

# Experiencia colaborativa con el Ministerio de Cultura

Desde la Dirección de Planificación y Seguimiento de Gestión del Ministerio de Cultura de Nación se produce información para el diseño de programas y la toma de decisiones estratégicas, como así también para aportar al análisis y el seguimiento de las políticas implementadas. Con diferentes líneas de trabajo se busca conocer y caracterizar a su población objetivo y aquella con la que dialoga. En este contexto, Fundar junto con la Dirección llevan a cabo un trabajo colaborativo con el objetivo de dimensionar y segmentar a las personas que se encuentran inscriptas en el Registro Federal de Cultura (en una versión actualizada a noviembre de 2021), en tanto pueden considerarse población objetivo para diseñar políticas específicas[2].

El alcance del Registro Federal de Cultura está orientado hacia los/as trabajadores/as de la cultura más que al público consumidor de actividades culturales. Al constituirse como registro administrativo para acceder a las convocatorias del Ministerio de Cultura de Nación, se entiende que constituye un recorte específico de los/as trabajadores del ámbito cultural que por diversas razones se inscribieron (las características de este registro se describen luego en este informe).

# Metodología

La manera de responder a la pregunta inicial de esta Guía fue mediante un análisis de *clustering*. La utilización de esta metodología permitió dividir a la población en 6 grupos con características distintas y que indican la presencia de diferentes perfiles. Para ello, consideramos tanto la actividad que desarrollaban como su lugar de residencia, relación laboral e ingresos culturales, entre otros. A continuación vamos a detallar la metodología utilizada en este estudio, que es fácilmente replicable a otros casos en los que se desee hacer un análisis de perfiles.

2 Para ello, se compartió una versión previamente anonimizada de Registro Federal de Cultura. Para una guía del procedimiento de anonimización de datos ver Luvini, Paula (2022). Guía práctica para la protección de datos. Buenos Aires: Fundar.

Área de Datos

Guía práctica para caracterizar a la población objetivo de una política pública a partir de registros administrativos

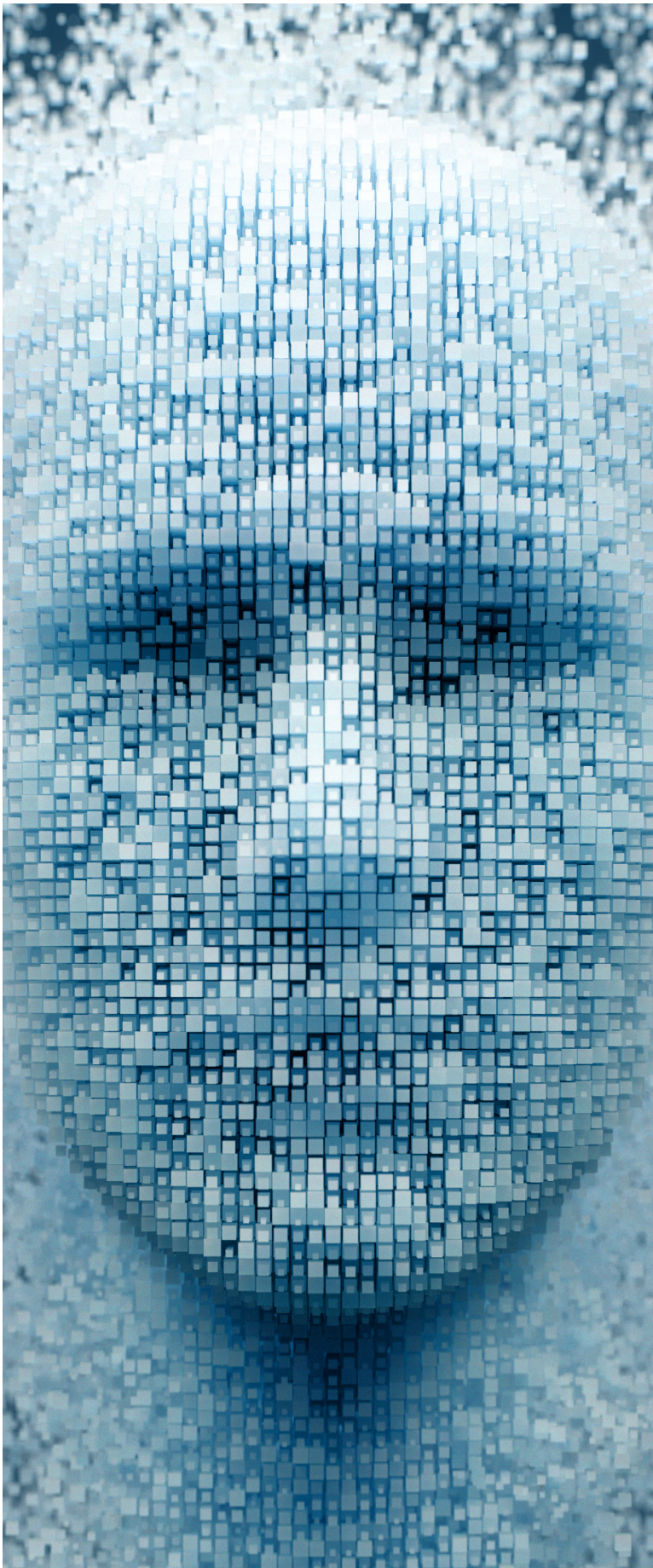
Febrero 2023

# *Machine learning* para políticas públicas

El **aprendizaje no supervisado** utiliza algoritmos de *machine learning* para agrupar datos. El objetivo de estos algoritmos es descubrir patrones en los datos para agruparlos sin la necesidad de definir más que algunos parámetros correspondientes a cada modelo.

Llevar a cabo un *clustering* implica dividir nuestros datos en grupos —los llamamos precisamente *clusters*— que están conformados por objetos que son más similares a sus pares de *cluster* que a los de los demás grupos, dada una medida elegida con anterioridad. La conformación de estos grupos va a depender de las variables que elijamos como *inputs* de nuestro modelo. Si incluimos variables como la edad, el género, trabajo o localización de individuos, el algoritmo las leerá y buscará patrones transversales a estas variables.

Si incluimos variables como la edad, el género, trabajo o localización de individuos, el algoritmo las leerá y buscará patrones transversales a estas variables



Área de Datos

Guía práctica para caracterizar a la población objetivo de una política pública a partir de registros administrativos

Febrero 2023

Paso 1 ⟶

# Elegir el modelo. Algoritmos de particiones

Dentro del aprendizaje no supervisado hay muchos tipos de metodologías que se pueden utilizar: algoritmos de particiones, mixturas, jerárquicos, de densidad, espectrales, entre otros. En la mencionada colaboración hemos utilizado algoritmos de partición, los cuales **dado un conjunto de observaciones en un espacio van a determinar una partición de los datos en K grupos, de manera tal que las observaciones dentro de cada partición son similares entre sí y diferentes a las de otras particiones.**

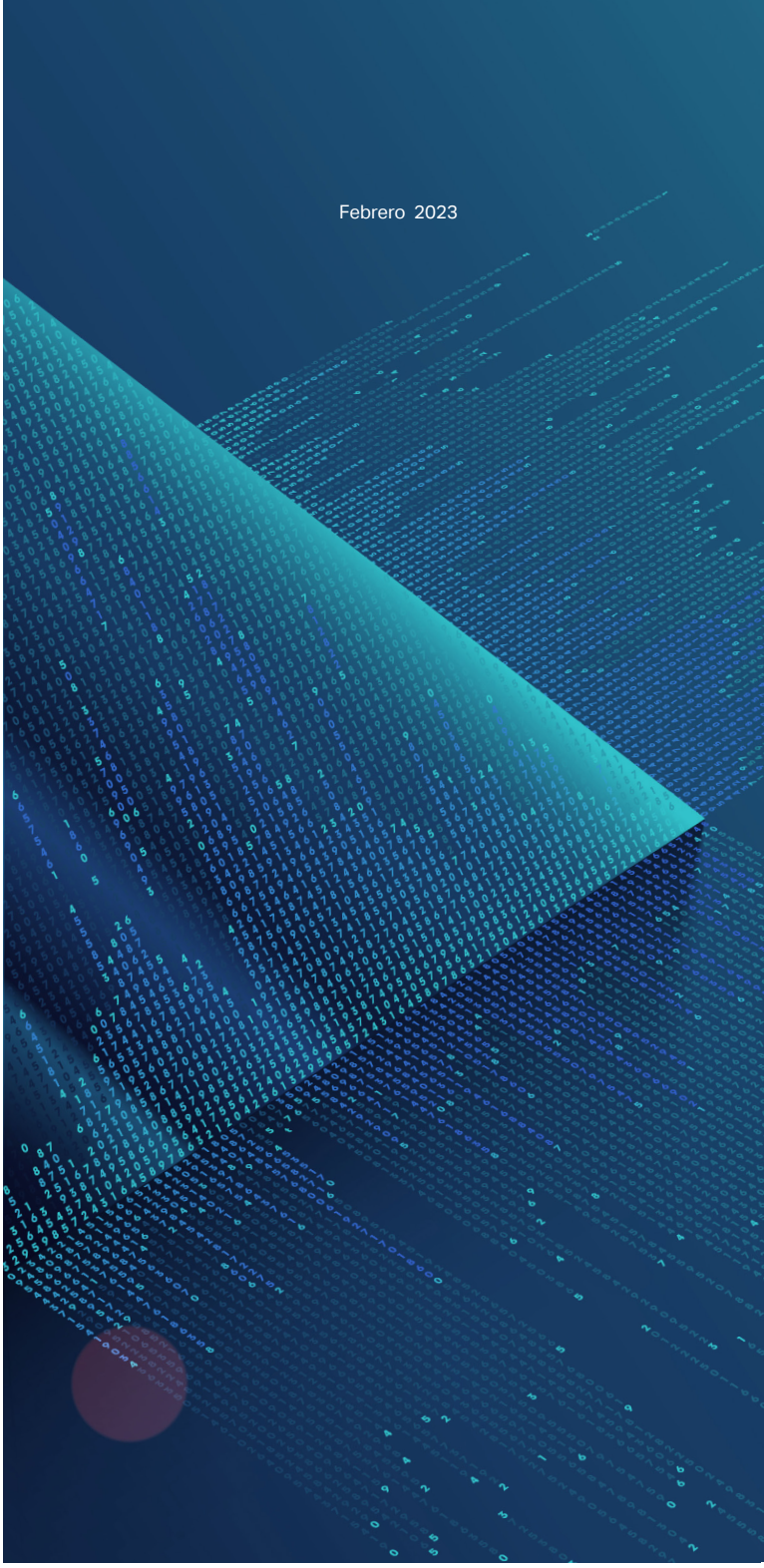
De esto último se desprende que necesitamos un **criterio para evaluar qué significa que dos observaciones sean similares entre sí.** Un criterio posible es aquel usado en K Medias: **distancias**. De esta manera, observaciones que se encuentren cerca unas de otras van a considerarse similares, mientras que aquellas que estén más alejadas en el espacio van a considerarse distintas. Este criterio de distancias va a encontrar, entonces, las particiones que minimicen la dispersión dentro de los grupos y maximice la dispersión entre grupos.

K Medias es uno de los algoritmos partitivos de aprendizaje no supervisado más usados para hallar grupos dentro de datos continuos y descubrir patrones en ellos. El número de grupos o k que encontrará es definido de antemano e indicado al algoritmo. Con ello, K Medias identificará k centroides en los datos: los centros geométricos del *cluster.* A partir de los centroides clasificará los demás puntos dentro de los grupos correspondientes. Esto lo hará colocando a cada punto en el *cluster* más cercano, es decir minimizando la distancia de los puntos con los centroides en una serie de iteraciones hasta encontrar la mejor configuración. Para realizar el agrupamiento, K Medias considera como métrica de distancia entre puntos a la distancia Euclídea al cuadrado.

Área de
Datos

Guía práctica para caracterizar a la población objetivo de una política pública a partir de registros administrativos

Febrero 2023



Paso 2 →

# ¿Qué pasa cuando tenemos datos categóricos?

En los registros administrativos, encuestas y formularios, es muy común encontrar que contamos con datos categóricos y no numéricos. Datos como la provincia de residencia, el tipo de relación laboral, el género o la ocupación.

En el caso de contar con variables categóricas, utilizar la distancia Euclídea ya no es posible sin binarizar de alguna manera las variables. Otra opción es utilizar un algoritmo de partición que tome como función de evaluación alguna específica para este tipo de datos: K Modas. Este reemplaza las medias de los *clusters* con modas, utiliza la frecuencia de las variables para la asignación a los *clusters* y como función de costos toma una medida de disimilaridad. La moda es el valor que tiene mayor frecuencia absoluta en una categoría determinada.

Los pasos del algoritmo son los siguientes:

1. → Se seleccionan aleatoriamente las k modas iniciales que van a ser los centroides de los *cluster*.

2. → Se adjudica cada punto a un *cluster* utilizando la función de disimilaridad y luego se actualiza su moda (centroide).

3. → Se vuelven a evaluar los puntos según la función de disimilaridad y los nuevos centroides; si se encuentra que son más cercanos a otro *cluster* se reasignan.

4. → Esto se repite hasta que ninguno de los puntos cambia de *cluster*.

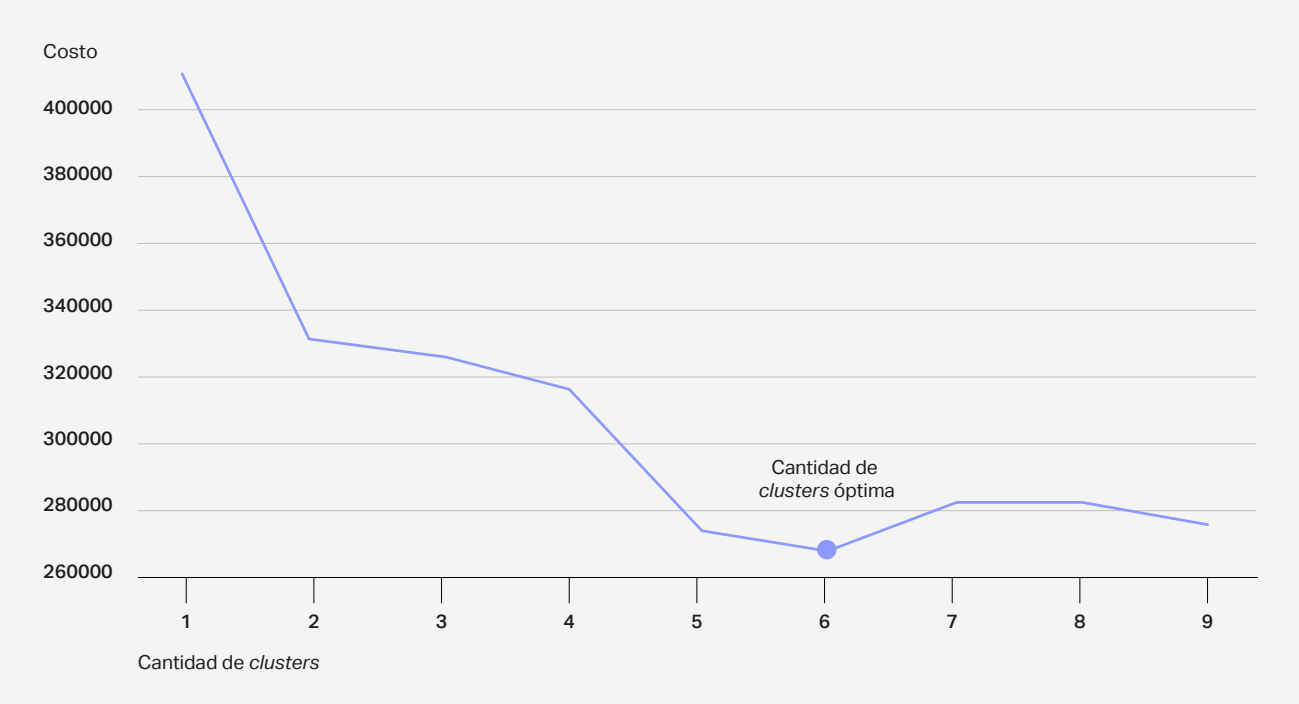
A la hora de evaluar políticas públicas o analizar datos de individuos, vamos a contar en la mayoría de los casos con variables categóricas, por lo que tener una herramienta específica para este tipo de datos es importante. Además, K Modas cuenta la ventaja de que es eficiente y puede procesar rápidamente un gran volumen de datos, a diferencia de otros métodos que también procesan datos categóricos, como los jerárquicos.

Paso 3 →

# ¿Cómo encontrar la cantidad de *clusters*?

Los algoritmos de K Medias y de K Modas necesitan que les indiquemos la cantidad de *clusters* que deseamos encontrar. **Aunque el número indicado no sea el correcto, los algoritmos se forzarán a encontrar esos k *clusters*.** Por esto es importante hallar el número correcto. El método del codo (o *elbow*) va a ayudarnos a elegirlo. Lo que hacemos es correr el algoritmo para un rango de valores k y evaluar en cada uno el costo asociado.

Idealmente, queremos un costo bajo para que los *clusters* sean más bien homogéneos, pero el costo va disminuyendo a medida que aumenta k, hasta ser 0 cuando tenemos 1 *cluster* por cada observación. Para "saber dónde parar" vamos a elegir el punto de inflexión del gráfico, aquel donde comenzamos a tener rendimientos decrecientes por aumentar k. O, en otras palabras, "el codo" de la curva graficada.



Box 1

# El Registro Federal de Cultura

**El Registro Federal de Cultura se creó en marzo de 2021 (Resolución 130/2021) y se puso en funcionamiento a través de la plataforma www.somos.cultura.gob.ar. Esta herramienta tiene como objetivo principal registrar a personas humanas y jurídicas que desarrollen sus actividades en el sector cultural en el territorio de la República Argentina; además, se utiliza para realizar la inscripción de las convocatorias del Ministerio de Cultura de la Nación.**

En noviembre de 2021, el Registro Federal de Cultura se encontraba integrado en un 75% por personas que se habían inscripto en el mes de marzo de 2021. En este sentido, puede inferirse que la mayoría lo hizo para postularse al Apoyo Extraordinario Cultura Solidaria (Resolución 210/2021), dirigida a trabajadores/as culturales que vieron afectada su actividad laboral por las medidas de aislamiento asociadas al COVID-19. Esta convocatoria estuvo abierta entre el 8 y el 14 de marzo de 2021[2]. En esa situación de emergencia, era fundamental conocer a quiénes se habían acercado al Ministerio, sus perfiles ocupacionales y en qué medida el contexto pandémico los/as había dejado en una situación de vulnerabilidad.

2 Otras convocatorias realizadas a través del registro fueron el de Cultura Solidaria y el Apoyo Fortalecer Cultura, otorgados durante 2020.

Paso 4 ⟶

# Análisis de *clustering*

Así como el número de *clusters* óptimo se puede definir a partir del método del codo, también se deben seleccionar las variables que se utilicen para la *clusterización*. La decisión es relevante porque va a determinar los resultados de los grupos, y su utilidad va a ser el resultado de considerar variables relevantes en el análisis. En nuestro caso, la selección de variables se realizó con base en los hallazgos de un análisis exploratorio; en este tipo de análisis, nos interesa revisar distribuciones, métricas de resumen y otras herramientas que permitan identificar patrones o anomalías en los datos. Para ello se utilizan herramientas gráficas y descriptivas.

A partir de este análisis preliminar se seleccionaron seis variables para realizar el *clustering*:

1. Provincia de residencia
2. Identidad de género
3. Porcentaje de ingresos proveniente de la actividad cultural
4. Situación ocupacional en el sector cultural
5. Área cultural principal
6. Inscripción a otros registros

## ¿Cómo se interpretan los *clusters*?

Una vez que tenemos definidos los grupos debemos analizar los resultados que obtuvimos. Para ello, vamos a revisar qué cosas quedaron dentro de cada uno analizando los centroides y algunos gráficos de utilidad.

Paso 4

# Observar los centroides

En primer lugar vamos a revisar cómo quedaron los centroides del gráfico, es decir aquellos puntos con las características promedio (en el caso de K Medias) o las más frecuentes (en el caso de K Modas). Estos puntos son relevantes porque nos indican el centro del *cluster* y nos sirven como referencia para evaluar los puntos y considerarlos parte de un grupo o del otro.

Por ejemplo, si revisamos los centroides del *cluster* 4 encontraremos lo siguiente:

| ***Cluster*** | **%** | **N** | **Provincias frecuentes** | **Identidad de género** | **Porcentaje de ingresos** | **Situación ocupacional** | **Áreas culturales frecuentes** | **Inscripción a registros** |
|---|---|---|---|---|---|---|---|---|
| *Cluster* 4 | 18 | 20.593 | Buenos Aires | Mujeres | 0% a 25% del ingreso | Trabajador informal | Escénicas | No |

Ahora bien, ¿esto significa que todas las personas pertenecientes a este grupo son mujeres de la provincia de Buenos Aires que se dedican a las artes escénicas de manera informal, que recibían un bajo porcentaje de su ingreso por actividades culturales y que no tenían una relación previa con el Ministerio? La respuesta es no. Si bien los centroides nos indican tendencias, también hay otras personas dentro de los subgrupos que no se corresponden con este "trabajador cultural promedio del *cluster* 4".

Esto sucede especialmente en aquellas variables donde tengamos muchas opciones distintas —como puede ser la de Provincias, que tiene 24 opciones diferentes en Argentina—, y por eso es valioso mirar en mayor detalle la composición de estos *clusters*.

Si agregamos cuánto representa cada variable en el total de personas que componen el *cluster* 4 (20.593), tenemos un panorama más completo, tal como se ve a continuación:

| ***Cluster*** | **%** | **N** | **Provincias frecuentes** | **Identidad de género** | **Porcentaje de ingresos** | **Situación ocupacional** | **Áreas culturales frecuentes** | **Inscripción a registros** |
|---|---|---|---|---|---|---|---|---|
| *Cluster* 4 | 18 | 20.593 | Mayoría **Buenos Aires (45%)** / luego CABA (7%), Santa Fe (5%) y Santiago del Estero (4%) | **Mujeres (82%)** / varones (14%) | **0% a 25% del ingreso** (72%) | **Trabajador informal (51%)** y voluntarix (23%) | **Escénicas (26%),** Visuales (12%) y Gastronomía (12%) | **No (94%)** |

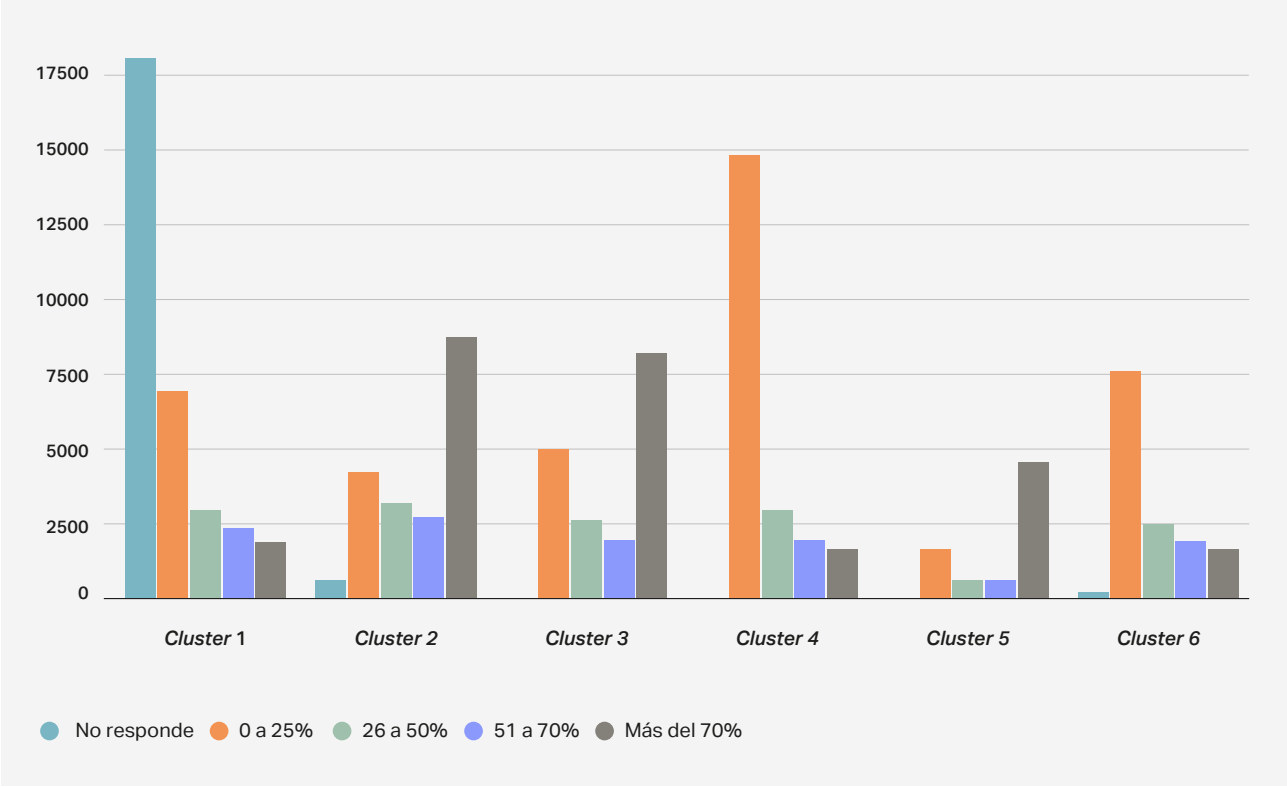
De esta manera, con variables relativizadas al tamaño total del cluster, observamos que este grupo está mayormente compuesto por trabajadoras informales y voluntarias con ingresos provenientes de su actividad cultural; que estos ingresos representan menos del 25% de sus ingresos totales; y que no están inscriptas en otros registros del Ministerio de Cultura de Nación. Es decir que no tenían una relación previa con la política ministerial. En cuanto a las provincias y áreas culturales a las que pertenecen, las variables están más atomizadas: solo un 45% del grupo reside en la provincia de Buenos Aires; el 55% restante se divide en las demás provincias.

Paso 4

# Graficar los *clusters*

De la misma manera, podemos mirar cómo se reparten las distintas características y variables elegidas dentro de los *clusters*, lo que nos va a ayudar a identificar posibles patrones que concentren algunas variables o estén más repartidas en otras.

En nuestro caso de estudio podemos observar, por ejemplo, tendencias en el porcentaje de ingresos que las personas perciben de su actividad cultural. Así, es bastante claro que quienes se encuentran desocupados en el sector se concentran particularmente en el primer *cluster*; asimismo, que el *cluster* 4 está "tomado" por personas para quienes los ingresos que provienen de su actividad cultural representan menos del 25% de sus ingresos totales. Estos hallazgos nos indican que hay una diferenciación bastante marcada en estos grupos, la cual hizo que se agrupen de esta manera: son muy distintos al resto en lo que refiere a los ingresos que reciben por su trabajo en el sector. En el caso de los *clusters* 2, 3 y 5, son mayoría quienes perciben más del 70% de sus ingresos de su actividad en el sector. Al encontrarse repartido en varios *clusters*, podemos pensar que esa característica es menos determinante a la hora de agrupar.



# La utilidad de este análisis para la política pública

Llevar a cabo una *clusterización* sobre una base de datos con información de personas nos permite encontrar grupos y perfiles distintos basados en las variables con las que se cuenta. Esta herramienta ha sido utilizada en *marketing* o comercialización para la segmentación de audiencias y clientes. En esta guía proponemos difundir su uso para analizar y conocer en mayor profundidad a la población objetivo de una política pública; también, su puede usar para otros casos como el de la evaluación de impacto, si se desea distinguir el efecto de una política entre distintos subgrupos.

**Las herramientas de aprendizaje supervisado son de gran ayuda para conocer a la poblaciones a quien el Estado les habla o con quienes se relaciona. Aportan conclusiones que muchas veces la estadística descriptiva no puede observar, porque analizar muchas variables juntas de esta manera puede tornarse engorroso y poco eficiente. Mediante el uso de *clusters*, podemos valernos de modelos estadísticos que relacionan muchas variables a la vez y encuentran patrones que son invisibles para el ojo humano. Son herramientas que nos ayudan a ordenar los datos y agruparlos fácilmente.**

En el caso de la experiencia colaborativa entre Fundar y la Dirección de Planificación y Seguimiento de Gestión del Ministerio de Cultura de Nación, el análisis de segmentación fue muy útil para identificar a un grupo de nuevos/as destinatarios/as de política que no tenían una relación previa con el Ministerio, y que se acercaron en el contexto pandémico (tal como observamos en el *cluster* 4). Asimismo, también fue útil para identificar grupos muy distintos entre sí por varias características, como los/las desocupados/as en el sector y quienes perciben más del 70% de ingresos de su actividad cultural. Tanto por sus ingresos provenientes de su trabajo en el sector como por las demás variables incluidas en el análisis, estos grupos están compuestos por personas de muy distinto perfil. A la hora de pensar en políticas públicas dirigidas al sector cultural, contar con esta información nos permite segmentar grupos y dirigir los esfuerzos de manera más eficiente.

# Acerca del equipo autoral

**Alejandro Avenburg**

**Investigador del Área de Datos**

Licenciado en Ciencia Política por la Universidad de Buenos Aires y doctor en Ciencia Política por Boston University. Fue becario post-doctoral de CONICET y por la Universidad Nacional de San Martín.

**Julia Houllé**

**Directora de Planificación y Seguimiento de Gestión de la Unidad de Gabinete de Asesores del Ministerio de Cultura de la Nación**

Licenciada en Sociología (UBA). Trabajó en el sector público como analista de datos y en gestión.

**Paula Luvini**

**Investigadora del Área de Datos**

Licenciada en Economía por la UBA y maestranda en Ciencia de Datos en la UdeSA. Trabajó en el sector público y en el privado y como docente.

**Magalí Rodrigues Pires**

**Analista de datos en la Dirección de Planificación y Seguimiento de Gestión (UGA)**

Licenciada en Sociología (UBA). Maestranda en Explotación de Datos y Descubrimiento del Conocimiento (UBA).

---

**Dirección ejecutiva:** Martín Reydó

**Coordinación editorial:** Gonzalo Fernández Rozas

**Diseño:** Micaela Nanni

**Revisión institucional:** Juliana Arellano

---

La Dirección de Planificación y Seguimiento de Gestión pertenece a la Unidad Gabinete de Asesores del Ministerio de Cultura de la Nación y tiene entre sus responsabilidades primarias asistir a la planificación estratégica de las políticas, y realizar el monitoreo y la evaluación de los programas y acciones implementadas. Para ello, produce y sistematiza información a través del Sistema de Información Cultural de la Argentina (SInCA) y del área de seguimiento y proyectos especiales, y centraliza la implementación de las convocatorias a través del Registro Federal de Cultura.

---

Fundar es un centro de estudios y diseño de políticas públicas que promueve una agenda de desarrollo sustentable e inclusivo para la Argentina. Para enriquecer el debate público es necesario tener un debate interno: por ello lo promovemos en el proceso de elaboración de cualquiera de nuestros documentos. Confiamos en que cada trabajo que publicamos expresa algo de lo que deseamos proyectar y construir para nuestro país. Fundar no es un logo: es una firma.



En Fundar creemos que el lenguaje es un territorio de disputa política y cultural. Por ello, sugerimos que se tengan en cuenta algunos recursos para evitar sesgos excluyentes en el discurso. No imponemos ningún uso en particular ni establecemos ninguna actitud normativa. Entendemos que el lenguaje inclusivo es una forma de ampliar el repertorio lingüístico, es decir una herramienta para que cada persona encuentre la forma más adecuada de expresar sus ideas.

## Modo de citar

Avenburg, A., Houllé, J., Luvini, P. y Rodrigues Pires, M. (2022). Guía práctica para caracterizar a la población objetivo de una política pública a partir de registros administrativos.

Disponible en https://www.fund.ar

Guía práctica para caracterizar a la población objetivo de una política pública a partir
de registros administrativos / Alejandro Avenburg ... [et al.]. - 1a ed. - Ciudad Autónoma de Buenos Aires : Fundar , 2023.
Libro digital, PDF

Archivo Digital: descarga y online
ISBN 978-987-48985-2-4

1. Análisis de Datos. 2. Bases de Datos. 3. Protección de Datos. I. Avenburg, Alejandro.
CDD 351.02854678

ISBN 978-987-48985-2-4

fundar