



Análisis de Nacimientos en Argentina (2023)

Autor: Valentino Martin Lorenzati

Dataset: [Link](#)

Fecha: Diciembre - 2025

Contexto: (Argentina, nacimientos 2023)

1. Introducción	1
2. Descripción de los datos	2
3. Análisis exploratorio de datos (EDA)	3
Distribución general de los nacimientos	4
Maternidad adolescente	4
Nivel educativo y maternidad adolescente	7
4. Limitaciones e inconsistencias detectadas	8
Inconsistencias en los datos de nivel educativo	8
Datos faltantes o mal especificados	8
5. Ciencia de Datos: Clustering de perfiles maternos	9
Objetivo del clustering	9
Selección de variables	9
Metodología aplicada	9
6. Resultados del clustering y caracterización de los perfiles	10
Cluster 0	10
Cluster 1	11
Cluster 2	11
Cluster 3	11
Síntesis de los perfiles identificados	12
7. Discusión y conclusiones	13

1. Introducción

El presente trabajo tiene como objetivo analizar los nacimientos registrados en Argentina, poniendo especial interés en la edad materna, el nivel educativo de las gestantes y la distribución territorial. A partir de un enfoque basado en datos, se busca identificar patrones que no resultan evidentes a partir de estadísticas descriptivas simples, como el fenómeno de la maternidad adolescente y su relación con factores educativos.

La investigación se desarrolla en dos etapas. En primer lugar, se realiza un análisis exploratorio de datos, orientado a describir la estructura general del conjunto de datos, cuantificar distribuciones y detectar posibles inconsistencias administrativas.

En una segunda etapa, se incorpora una aproximación de ciencia de datos mediante técnicas de clustering no supervisado, con el objetivo de identificar perfiles maternos estadísticamente homogéneos a partir de variables sociodemográficas.

Los resultados obtenidos permiten entender la comprensión del fenómeno estudiado y sentar las bases para análisis más específicos o profundos en futuras investigaciones.

2. Descripción de los datos

El análisis se basa en datos oficiales de nacimientos provistos por el Sistema de Estadísticas Vitales del Ministerio de Salud de la Nación (DEIS), correspondientes a registros administrativos de nacimientos ocurridos en el año 2023, Argentina. El conjunto de datos contiene información sociodemográfica de las madres, así como variables asociadas al evento del nacimiento y a la jurisdicción de residencia.

Cada registro representa una agrupación de nacimientos con características comunes, incluyendo variables como el grupo etario de la madre, el nivel de instrucción alcanzado, la provincia de residencia, y la cantidad de nacimientos asociados a dicha combinación de atributos.

Las variables centrales consideradas en el análisis fueron:

- Edad materna, registrada en forma de grupos etarios predefinidos.
- Nivel de instrucción de la madre, clasificado según categorías educativas oficiales.
- Jurisdicción de residencia, correspondiente a la provincia o distrito de la madre.

Es importante señalar que, al tratarse de datos administrativos, el dataset puede presentar inconsistencias o limitaciones de registro. Estas situaciones fueron identificadas y consideradas en la interpretación de los resultados, sin comprometer la validez del análisis agregado.

En conjunto, el dataset proporciona una base sólida para el análisis exploratorio y la aplicación de técnicas de ciencia de datos, permitiendo abordar el fenómeno de los nacimientos desde una perspectiva cuantitativa, territorial y educativa.

3. Análisis exploratorio de datos (EDA)

El análisis exploratorio de datos tiene como objetivo describir la estructura general del conjunto de nacimientos y detectar patrones relevantes en relación con la edad materna, el nivel educativo y la distribución territorial. Esta etapa permite contextualizar el fenómeno estudiado y orientar la posterior aplicación de técnicas de ciencia de datos.

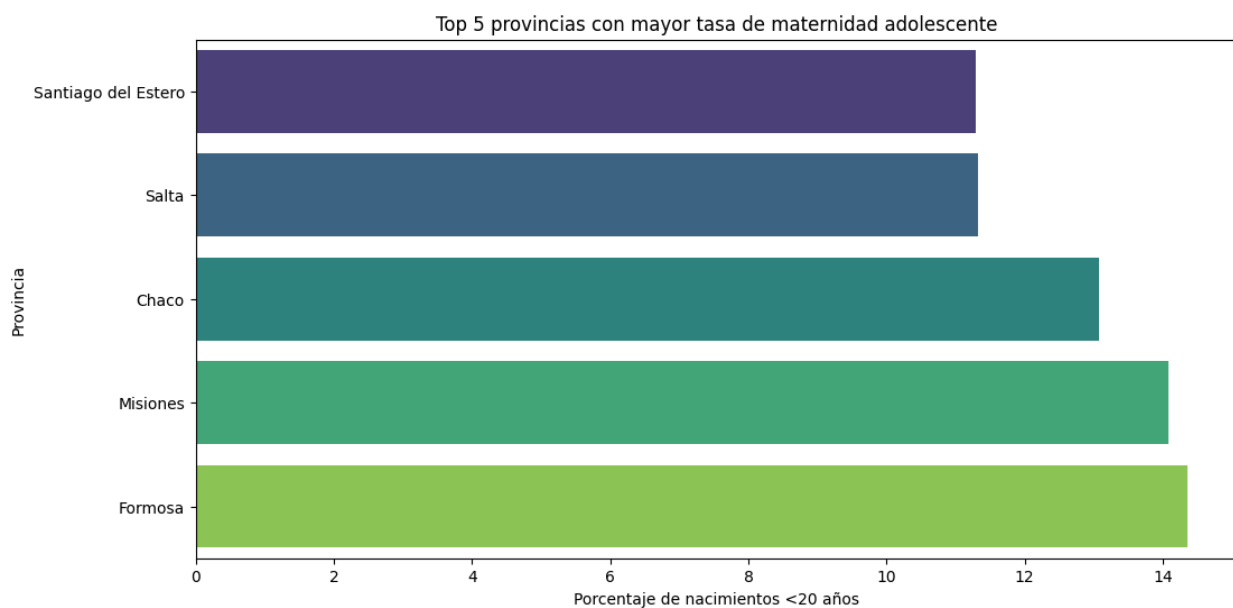
Distribución general de los nacimientos

En términos absolutos, la mayor cantidad de nacimientos se concentra en las provincias con mayor población, Buenos Aires, seguida por Santa Fe y Córdoba. Este comportamiento es esperable y refleja la estructura demográfica del país. Sin embargo, el análisis de valores absolutos resulta insuficiente para comprender dinámicas relativas, por lo que se incorporan posteriormente tasas y proporciones.

Con respecto de la edad materna, los nacimientos se concentran mayormente en rangos etarios de adultez temprana y media, mientras que en los extremos de edad presentan una participación considerablemente menor. Esta distribución coincide con patrones reproductivos, aunque exhibe variaciones relevantes al analizar subgrupos específicos.

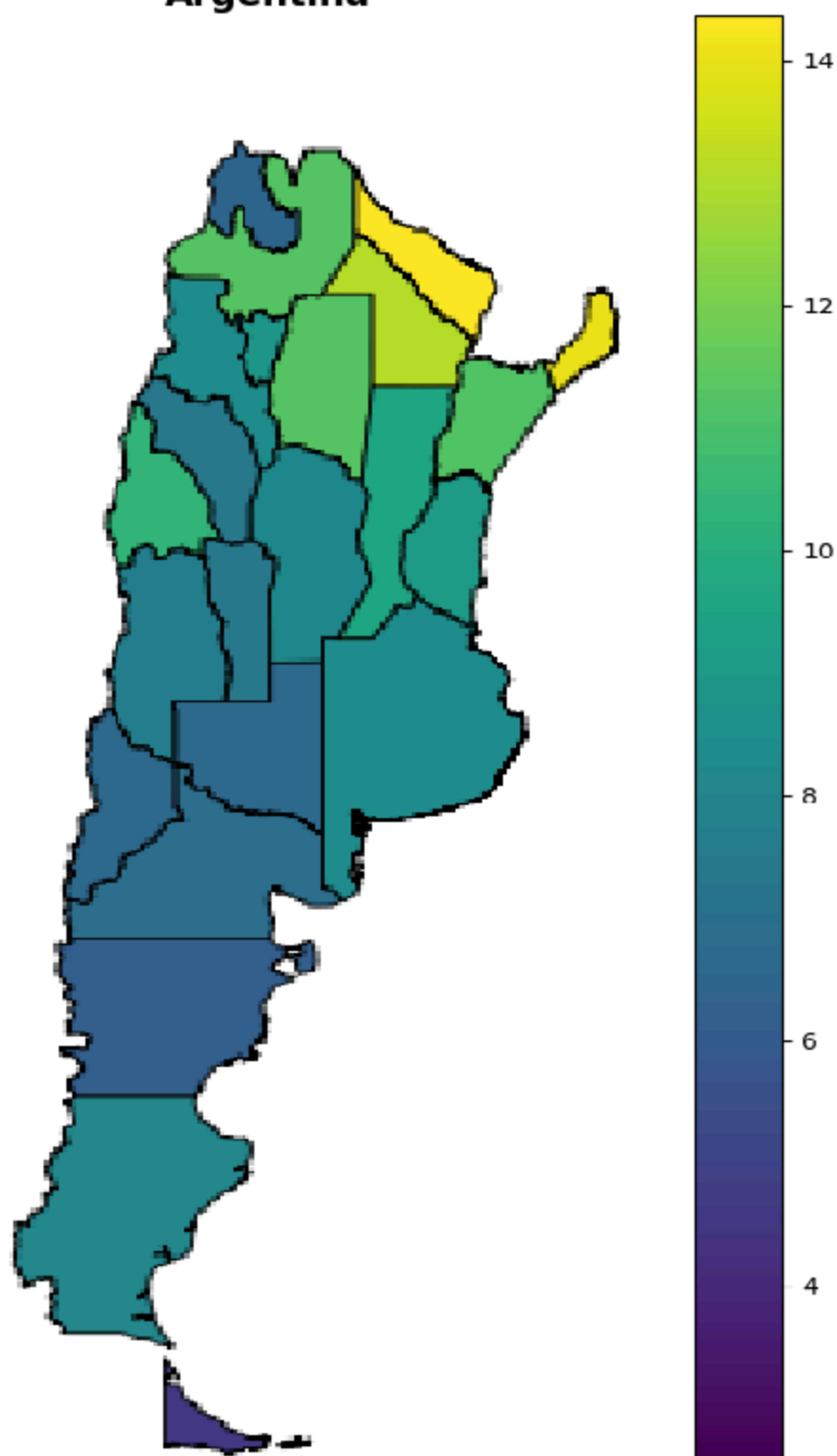
Maternidad adolescente

Para el presente análisis, se considera maternidad adolescente a los nacimientos correspondientes a madres menores de 20 años. Al observar la distribución territorial de estos casos, se advierte que, si bien Buenos Aires concentra el mayor número absoluto de nacimientos adolescentes, las tasas relativas más elevadas se registran en provincias del NEA y NOA, como Formosa, Misiones y Chaco.



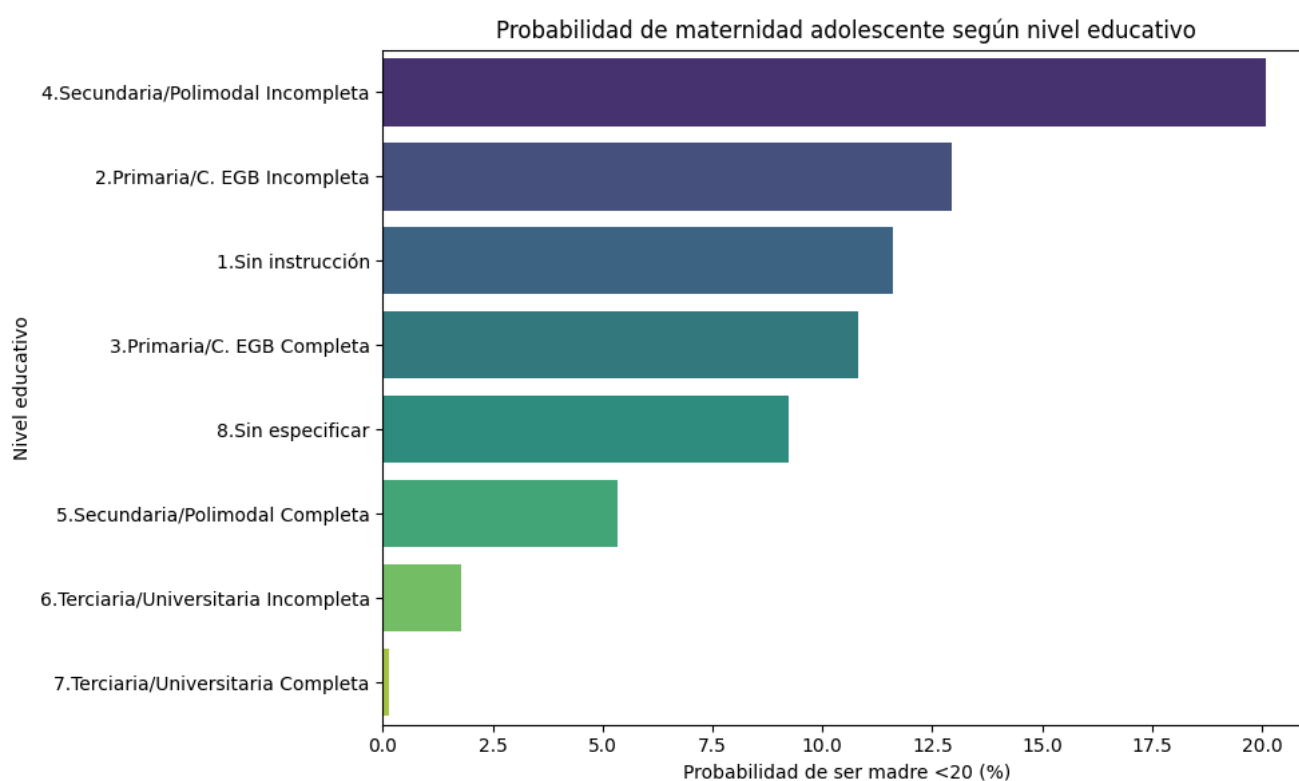
Este hallazgo demuestra la importancia de analizar proporciones en lugar de conteos absolutos, ya que permite identificar territorios donde la maternidad adolescente tiene un mayor peso relativo dentro del total de nacimientos, independiente del tamaño poblacional.

Tasa de maternidad adolescente por provincia Argentina



Nivel educativo y maternidad adolescente

El análisis del nivel educativo de las gestantes muestra una relación inversa entre educación y maternidad adolescente. Los mayores niveles de probabilidad de maternidad antes de los 20 años se observan en mujeres con secundario incompleto o niveles educativos inferiores, mientras que esta misma probabilidad disminuye notablemente a medida que aumenta el nivel de educación.



Las categorías terciaria y universitaria completa presentan valores marginales, los cuales se interpretan como posibles inconsistencias administrativas, sin impacto en el patrón general observado.

4. Limitaciones e inconsistencias detectadas

Durante el análisis se detectaron algunos registros con varias limitaciones e inconsistencias que deben ser consideradas al interpretar los resultados obtenidos. Las principales dificultades fueron las siguientes:

Inconsistencias en los datos de nivel educativo

Uno de los hallazgos más relevantes fue la identificación de registros de madres adolescentes (<20 años) que figuran con títulos terciarios o universitarios completos. Este comportamiento resulta ser una anomalía, ya que las trayectorias educativas típicas indican que es improbable que las mujeres menores de 20 años hayan completado estudios de nivel terciario o universitario.

Si bien algunos de estos casos podrían ser correctos, la baja frecuencia (aproximadamente 0.3% de los registros) y la discordancia con las trayectorias educativas esperables sugieren que estas entradas son probablemente el resultado de inconsistencias administrativas o errores de carga de datos.

Este tipo de inconsistencias puede surgir por diversas razones, como errores en la clasificación de los niveles educativos o problemas al momento de ingresar los datos en el sistema, lo que genera un sesgo en los resultados relacionados con la educación y maternidad adolescente.

Datos faltantes o mal especificados

El dataset también contiene una cantidad grande de registros con la variable "nivel educativo" sin especificar, alrededor del 8% en algunas categorías. La falta de información en esta columna puede afectar la precisión de las conclusiones y los patrones observados, ya que impide clasificar correctamente a una porción considerable de las madres según su nivel educativo.

5. Ciencia de Datos: Clustering de perfiles maternos

Con el objetivo de complementar el EDA y profundizar en la identificación de patrones, se incorporó una etapa de ciencia de datos basada en técnicas de aprendizaje no supervisado. En concreto, se utilizó clustering, orientado a segmentar los nacimientos en perfiles maternos homogéneos, considerando múltiples variables sociodemográficas.

A diferencia del análisis descriptivo, que examina variables de forma aislada, el clustering permite analizar la interacción conjunta entre múltiples atributos, revelando estructuras que no resultan evidentes a simple vista.

Objetivo del clustering

El objetivo principal de esta etapa fue identificar grupos de nacimientos con características similares en términos, de, no establecer relaciones causales ni realizar predicciones, sino explorar la heterogeneidad del fenómeno y concentrar la información en un número reducido de perfiles interpretables.

Selección de variables

Para la construcción de los clusters se seleccionaron variables representativas del perfil sociodemográfico de las madres. Estas variables fueron transformadas a un formato numérico (dummies) adecuado para su procesamiento por el algoritmo de clustering.

La variable jurisdicción de residencia fue evaluada inicialmente, pero posteriormente excluida del modelo principal con el fin de evitar que la segmentación estuviera dominada por factores territoriales, priorizando así la identificación de perfiles definidos por edad y educación.

Metodología aplicada

Se utilizó el algoritmo K-means, uno de los métodos de clustering más conocidos, se utilizó debido a su simplicidad, interpretabilidad y eficiencia computacional.

Dado que, K-means no funciona con eficiencia con muchas dimensiones, se aplicó previamente una técnica de reducción de dimensionalidad (PCA), que permitió proyectar la información en un espacio de menor dimensión conservando la mayor parte de la variabilidad. Esta transformación facilitó tanto la estabilidad del modelo como la visualización posterior de los resultados.

Interpretación general del enfoque

Es importante destacar que el clustering no busca reproducir clasificaciones predefinidas ni imponer una estructura jerárquica sobre los datos. Por el contrario, los grupos identificados emergen de patrones de similitud estadística, lo que explica que algunos clusters presenten combinaciones de edad y educación que no siguen un orden secuencial estricto.

Este comportamiento no constituye una limitación del método, sino una característica inherente a los algoritmos de aprendizaje no supervisado, que priorizan la similitud multivariada por sobre la continuidad ordinal de las categorías.

6. Resultados del clustering y caracterización de los perfiles

El análisis de clustering permitió identificar cuatro perfiles maternos diferenciados, definidos a partir de la combinación entre grupo etario y nivel educativo. Estos clusters no representan categorías normativas, sino agrupaciones estadísticas que sintetizan patrones recurrentes en los datos.

A continuación, se describen las características principales de cada perfil identificado.

Cluster 0 – Madres jóvenes adultas con educación secundaria (22,2 %)

Este cluster está compuesto exclusivamente por madres del grupo etario 20 a 24 años, lo que lo convierte en un perfil etariamente homogéneo. En términos educativos, se caracteriza por:

- Predominio de educación secundaria completa.
- Alta presencia de secundaria incompleta.
- Participación marginal de educación terciaria o universitaria.

Este perfil representa madres jóvenes adultas cuya maternidad ocurre mayormente tras finalizar o interrumpir el nivel secundario, reflejando trayectorias educativas medias.

Cluster 1 – Perfil mixto etario con educación media en transición (33,8 %)

El cluster 1 presenta una combinación de grupos etarios, principalmente en madres de 15 a 19 años y 25 a 29 años. Desde el punto de vista educativo, se observa:

- Predominio de educación secundaria, tanto completa como incompleta.
- Presencia moderada de educación terciaria o universitaria incompleta.
- Baja participación de niveles educativos superiores completos.

Este perfil refleja situaciones de transición, donde la maternidad ocurre tanto en etapas tempranas como en edades adultas, coexistiendo con trayectorias educativas en curso o recientemente finalizadas.

Cluster 2 – Madres adultas con alta educación superior (29,1 %)

El cluster 2 se concentra mayormente en madres de 30 a 34 años, con una participación relevante del grupo de 40 a 44 años. En términos educativos, se distingue por:

- Alta proporción de educación terciaria o universitaria, tanto completa como incompleta.
- Menor peso relativo de educación secundaria.
- Trayectorias educativas avanzadas al momento del nacimiento.

Este perfil representa maternidades asociadas a una postergación del evento reproductivo, coherente con la finalización o continuidad de estudios superiores.

Cluster 3 – Madres adultas tardías con educación superior completa(14,9 %)

El cluster 3 está compuesto exclusivamente por madres del grupo etario 35 a 39 años, presentando una homogeneidad etaria marcada. Desde el punto de vista educativo, se caracteriza por:

- Predominio claro de educación terciaria o universitaria completa.
- Escasa presencia de niveles educativos bajos o medios.
- Trayectorias formativas consolidadas previas a la maternidad.

Este perfil representa el segmento con mayor capital educativo y una maternidad claramente postergada.

Síntesis de los perfiles identificados

En conjunto, los clusters reflejan una relación consistente entre edad materna y nivel educativo. A medida que aumenta la edad al momento del nacimiento, se observa una mayor presencia de niveles educativos superiores, mientras que las maternidades más tempranas se asocian principalmente a trayectorias educativas medias o incompletas.

Estos perfiles permiten sintetizar la heterogeneidad del fenómeno y aportan una estructura interpretativa que complementa y refuerza los hallazgos del análisis exploratorio.

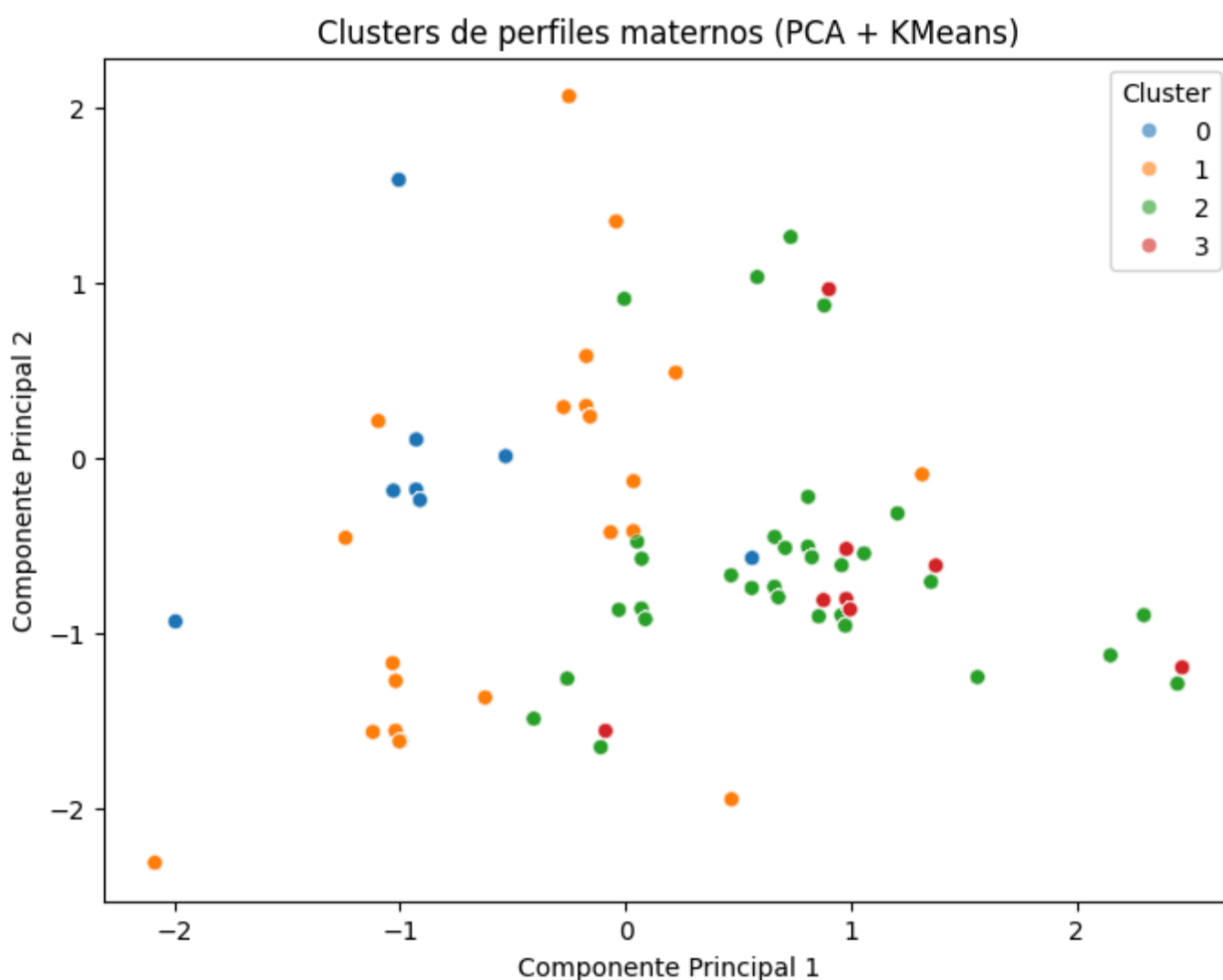


Diagrama de dispersión (scatter plot)

7. Discusión y conclusiones

Los resultados obtenidos a lo largo del análisis permiten abordar el fenómeno de los nacimientos desde una perspectiva de estadística descriptiva y técnicas de ciencia de datos. El análisis exploratorio evidenció patrones claros en la relación entre edad materna, nivel educativo y distribución territorial, mientras que el clustering aportó una síntesis estructurada de estos patrones mediante la identificación de perfiles maternos diferenciados.

En primer lugar, el análisis por tasas puso en evidencia que, si bien las provincias más pobladas concentran la mayor cantidad absoluta de nacimientos, las tasas más elevadas de maternidad adolescente se observan en provincias del NEA y NOA. Este hallazgo refuerza la necesidad de analizar proporciones en lugar de conteos absolutos para comprender las desigualdades territoriales.

En segundo lugar, se identificó una asociación robusta entre nivel educativo y maternidad adolescente, donde la probabilidad de maternidad antes de los 20 años disminuye de manera marcada a medida que aumenta el nivel de instrucción. Este patrón se mantiene consistente aun considerando las limitaciones detectadas en los datos, lo que sugiere que la educación constituye un factor estructural clave en la comprensión del fenómeno.

La etapa de clustering permitió integrar estas dimensiones y revelar perfiles maternos que combinan edad y educación. Los clusters identificados muestran una conexión clara desde maternidades tempranas asociadas a niveles educativos medios o incompletos, hacia maternidades más tardías vinculadas a trayectorias educativas avanzadas y consolidadas. Esta segmentación no debe interpretarse como una clasificación rígida, sino como una herramienta exploratoria que explica la heterogeneidad de los datos.

Es importante destacar que este trabajo adopta un enfoque exploratorio y descriptivo, sin intención de establecer relaciones causales. También, la presencia de inconsistencias administrativas, particularmente en variables educativas para edades tempranas, necesita cuidado en la interpretación de ciertos resultados puntuales, aunque no modifique importantemente los patrones generales observados.

En conjunto, el análisis demuestra el valor de combinar análisis exploratorio de datos y técnicas de ciencia de datos para estudiar fenómenos sociales complejos. Los resultados obtenidos permiten enriquecer la comprensión de la maternidad en Argentina y constituyen una base sólida para futuras investigaciones que incorporen variables adicionales, enfoques temporales o modelos explicativos más avanzados.