

Proyecto final

# STEMWOMEN

**Sistema de recomendación de  
iniciativas STEM según perfil de usuario**

*Andrea Galvão*

*Natalia García*

*Mayo 2025*



STEM, acrónimo de:

**SCIENCE  
TECHNOLOGY  
ENGINEERING  
MATHEMATICS**

## 1. CONTEXTUALIZACIÓN Y DEFINICIÓN DEL PROBLEMA

## Marco actual

A finales del 2015 se acuerda la **Agenda 2030 para el Desarrollo Sostenible**. Un plan de acción que aborda 17 objetivos, entre ellos el ODS 5 que **promueve la igualdad de género y el empoderamiento de mujeres y niñas**.



En 2017, el movimiento **#MeToo** marcó un antes y un después al visibilizar desigualdades y violencias de género, impulsando una ola de cambio social que también alcanzó al ámbito STEM.

**En el 2050, el 75% de los trabajos estará relacionado con el área STEM.**

Datos de la Organización de las Naciones Unidas (2021)

**Las nuevas tecnologías generarán 170 millones de nuevos empleos antes de 2030.**

“Informe sobre el futuro del empleo 2025” del Foro Económico Mundial

**Los empleos que crecerán con mayor rapidez requerirán capacidades tecnológicas en IA, big data, redes y ciberseguridad.**

“Informe sobre el futuro del empleo 2025” del Foro Económico Mundial



## Situación de la mujer en STEM

**Internacional:** *Entre 2011 y 2021, el porcentaje de mujeres licenciadas en STEM a nivel mundial se ha mantenido estancado en un 35%.*

*En 2022, las mujeres ocuparon menos del 25% de los puestos en ciencia, ingeniería y TIC.*

*En Estados Unidos, solo el 16% de los profesores titulares cuya principal área de investigación es la inteligencia artificial son mujeres.*

*Gender report 2024: Technology on her terms (UNESCO, 2024)*

**España:** *El porcentaje de mujeres en una ocupación STEM en España es del 5,5%. Aunque ha aumentado sustancialmente desde 2011, cuando estaba en 3,3% (ESADE).*

*Las tasas de mujeres matriculadas en grados universitarios STEM es muy inferior al de los hombres. En Matemáticas (36%), Física (27%), Telecomunicaciones (23%), o Informática (13%) son especialmente bajas.*

*En Bachillerato, la presencia de chicas en las ramas científico-técnicas es notablemente inferior, pese a que el porcentaje que completa los estudios exitosamente es superior al de los chicos.*

## ¿QUÉ PODEMOS HACER?

Este proyecto intenta desarrollar **un sistema de recomendación de iniciativas STEM personalizadas** según tipo de actividad, formato y edad del usuario.

El objetivo es acercar estas iniciativas a un público más amplio. Facilitando el acceso a recursos relevantes para distintos perfiles, contribuyendo a una mejor difusión de las oportunidades STEM.

## 2. ANÁLISIS EXPLORATORIO DE LOS DATOS (EDA)

## DEFINICIONES

Iniciativa STEM:	Cualquier acción, programa o proyecto orientado a promover el desarrollo de conocimientos y habilidades en Ciencia, Tecnología, Ingeniería y Matemáticas
Persona impactada:	Es aquella que ha participado, se ha beneficiado o ha sido influenciada de forma significativa por una actividad STEM.
Segmentos:	Los iniciativas se clasifican en dos segmentos: <b>Inspiración</b> y <b>Carrera</b> , según en qué etapa, educativa o profesional, se encuentren las personas que han sido impactadas por la iniciativa.
<b>Segmento de Inspiración</b> →	De 1 a 16 años. Comprende las etapas educativas comunes a todos los estudiantes, cuyo objetivo es despertar interés en disciplinas STEM.
<b>Segmento de Carrera</b> →	<ol style="list-style-type: none"><li>1. <u>Educación opcional</u>: Bachillerato, FP y Formación superior.</li><li>2. <u>La carrera profesional</u>: Desde la incorporación al mercado laboral, hasta los últimos años de la jubilación.</li></ol>

## COLUMNAS DE LA BASE DE DATOS

Base de datos propia que recoge los datos de actividad de las iniciativas que trabajan para reducir la brecha de género en los campos de la ciencia y la tecnología.

['año']:	Año al que corresponden los datos reportados por las iniciativas.
['año_inicio']:	Año en el que empieza su actividad la iniciativa.

# El dataset mide el impacto de las iniciativas STEM Women en España

## El DATASET se compone de:

- 6 bases de datos (2019-2024). Datos reportados por las iniciativas y recogidos mediante un cuestionario que consta de 37 preguntas.
- Datos descargados del Instituto Nacional de Estadística. Población de niñas y niños de 1 a 16 años de España (2019-2024).
- Datos del Ministerio de Ciencia Innovación y Universidades de mujeres matriculadas en estudios STEM (2019-2024)
- Datos del Ministerio de Educación Formación Profesional y Deportes de mujeres matriculadas en FP STEM (2019-2024)

**6a Edición del STEM Women Annual Report - España**

Este cuestionario se ha distribuido a todas las iniciativas que trabajan para impulsar y empoderar el talento femenino en el ámbito STEM en España. A partir de los datos recogidos y de la información publicada por los organismos oficiales, se elaborará el **STEM Women Annual Report**. El análisis nos permitirá conocer en qué territorios y en qué etapas de la trayectoria formativa o profesional se está impactando, para saber dónde se ha de actuar y así contribuir a minimizar la brecha de género.

**IMPORTANTE:**

- Todos los datos solicitados corresponden a la actividad realizada durante el año 2024 en España.
- El cuestionario se ha de responder por cada programa de la iniciativa. Si en la iniciativa se realiza más de un programa, se debe responder nuevamente.

*Nota: los datos proporcionados por las iniciativas se tratarán con absoluta confidencialidad. Únicamente se publicarán los datos agregados obtenidos en el análisis.*

Si deseas consultar los informes de años anteriores, puedes descargarlos desde el siguiente enlace: <https://www.globalstemwomen.org/swar>

Muchas gracias por vuestra colaboración.

DATOS INICIALES

1 Nombre de la iniciativa \*

Escribe su respuesta

2 Página web o redes sociales de la iniciativa \*

Escribe su respuesta

3 Breve descripción de la iniciativa \*

Escribe su respuesta

20. Número de personas impactadas por la iniciativa por franja de edad y en base a los datos de 2024.

- Si la iniciativa es también internacional, aportar únicamente los datos a nivel nacional.
- No incluir el número de seguidores en las redes sociales.
- Seleccionar una opción por franja de edad.

**IMPORTANTE: desliza la barra hacia la derecha para ver todas las categorías (de "0 impactos" a "> de 5.000 impactos").**

	0 impactos	< de 50 impactos	de 50 a 100 impactos	de 101 a 500 impactos	de 501 a 1.000 impactos	de 1.001 a 5.000 impactos	> de 5.000 impactos
De 1 a 5 años	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>				
De 6 a 9 años	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>				
De 10 a 12 años	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>				
De 13 a 14 años	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>				
De 15 a 16 años	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>				
Bachillerato	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>				
En los 2 primeros años de estudios de Grado y FP	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>				
Final de estudios: Grado, Máster, Doctorado, etc.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>				
Etapa júnior (primeros años de la carrera profesional)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>				
Etapa de consolidación de la carrera profesional	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>				
Etapa de madurez profesional (últimos años de carrera profesional)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>				

+ Agregar una instrucción

Obligatoria

# DataFrame

Año	¿A qué tipo de organización pertenece tu iniciativa?	Según el género, ¿a quién va dirigida la iniciativa?	¿En qué comunidades autónomas impacta la iniciativa?	Según las fuentes de financiación, la iniciativa es:	Indica el número de profesionales, asalariados y/o voluntarios, que trabajan activamente en la iniciativa	Formación (curso, taller, etc.)	Evento (Jornada, congreso, ponencia, talks, etc.)	Mentoring	Recursos: material de formación o similar	...	De 6 a 9 años	De 10 a 12 años	De 13 a 14 años	De 15 a 16 años	Bachillerato	En los 2 primeros años de estudios de Grado y FP	Final de estudios: Grado, Máster, Doctorado, etc.	Etapa júnior (primeros años de la carrera profesional)	Etapa de consolidación de la carrera profesional	Etapa de madurez profesional (últimos años de carrera profesional)
										...	...	...	...	...	...	...	...	...	...	...
0 2024	Empresa privada	Ambos	Andalucía;Aragón;Asturias;Baleares;Canarias;Ca...	Privada: su capital proviene exclusivamente de...	De 11 a 50	Importante	Poco importante	Muy importante	Muy importante	...	de 101 a 500 impactos	de 501 a 1.000 impactos	> de 5.000 impactos	> de 5.000 impactos	> de 5.000 impactos	de 501 a 1.000 impactos	de 101 a 500 impactos	de 50 a 100 impactos	de 50 a 100 impactos	de 50 a 100 impactos
1 2024	Organización sin ánimo de lucro (ONG, fundació...	Ambos	Aragón;Andalucía;Asturias;Baleares;Cantabria;C...	Mixta: combina las dos formas de financiación	Más de 100	Muy importante	Muy importante	Importante	Importante	...	0 impactos	de 501 a 1.000 impactos	de 501 a 1.000 impactos	de 501 a 1.000 impactos	0 impactos	0 impactos	0 impactos	0 impactos	0 impactos	0 impactos
2 2024	Organización sin ánimo de lucro (ONG, fundació...	Ambos	Cataluña;	Mixta: combina las dos formas de financiación	De 11 a 50	Muy importante	Muy importante	Muy importante	Importante	...	0 impactos	de 101 a 500 impactos	de 101 a 500 impactos	de 101 a 500 impactos	de 101 a 500 impactos	de 50 a 100 impactos	de 50 a 100 impactos	< de 50 impactos	< de 50 impactos	< de 50 impactos
3 2024	Empresa privada	Femenino	Madrid, Comunidad de;	Privada: su capital proviene exclusivamente de...	De 11 a 50	Muy importante	Importante	Importante	Importante	...	de 50 a 100 impactos	de 50 a 100 impactos	de 50 a 100 impactos	de 50 a 100 impactos	0 impactos	< de 50 impactos	0 impactos	0 impactos	0 impactos	0 impactos
4 2024	Institución pública y gubernamental	Femenino	Andalucía;Aragón;Asturias;Baleares;Canarias;Ca...	Privada: su capital proviene exclusivamente de...	De 11 a 50	No se realiza	Muy importante	No se realiza	No se realiza	...	0 impactos	0 impactos	0 impactos	0 impactos	0 impactos	0 impactos	de 1.001 a 5.000 impactos	de 1.001 a 5.000 impactos	de 1.001 a 5.000 impactos	de 1.001 a 5.000 impactos
5 2024	Organización sin ánimo de lucro (ONG, fundació...	Femenino	Andalucía;Aragón;Castilla y León;Comunidad Val...	Mixta: combina las dos formas de financiación	Más de 100	Muy importante	Muy importante	Muy importante	Poco importante	...	0 impactos	de 501 a 1.000 impactos	de 501 a 1.000 impactos	de 501 a 1.000 impactos	de 501 a 1.000 impactos	de 101 a 500 impactos	de 50 a 100 impactos	< de 50 impactos	0 impactos	0 impactos
6 2024	Institución educativa (colegio, universidad, e...	Ambos	Baleares;	Mixta: combina las dos formas de financiación	De 51 a 100	Muy importante	Muy importante	No se realiza	Importante	...	< de 50 impactos	de 101 a 500 impactos	de 101 a 500 impactos	de 101 a 500 impactos	de 1.001 a 5.000 impactos	de 1.001 a 5.000 impactos	de 1.001 a 5.000 impactos	de 50 a 100 impactos	de 50 a 100 impactos	de 50 a 100 impactos
7 2024	Institución educativa (colegio, universidad, e...	Ambos	Extremadura;	Pública: se financia en su totalidad con fondo...	De 6 a 10	Muy importante	Muy importante	Muy importante	Importante	...	0 impactos	de 101 a 500 impactos	de 50 a 100 impactos	de 50 a 100 impactos	< de 50 impactos	< de 50 impactos	< de 50 impactos	< de 50 impactos	< de 50 impactos	< de 50 impactos
8 2024	Organización sin ánimo de lucro (ONG, fundació...	Ambos	Cataluña;	Mixta: combina las dos formas de financiación	De 6 a 10	Muy importante	Poco importante	No se realiza	Poco importante	...	de 501 a 1.000 impactos	de 1.001 a 5.000 impactos	de 1.001 a 5.000 impactos	de 1.001 a 5.000 impactos	de 50 a 100 impactos	< de 50 impactos	0 impactos	0 impactos	0 impactos	0 impactos
9 2024	Institución educativa (colegio, universidad, e...	Femenino	Cataluña;	Pública: se financia en su totalidad con fondo...	De 2 a 5	Poco importante	Importante	Muy importante	Muy importante	...	0 impactos	0 impactos	0 impactos	0 impactos	0 impactos	0 impactos	< de 50 impactos	< de 50 impactos	< de 50 impactos	< de 50 impactos
10 2024	Institución educativa (colegio, universidad, e...	Ambos	Andalucía;Aragón;Asturias;Baleares;Canarias;Ca...	Pública: se financia en su totalidad con fondo...	De 2 a 5	Importante	Importante	Importante	Poco importante	...	0 impactos	de 101 a 500 impactos	de 101 a 500 impactos	de 101 a 500 impactos	de 50 a 100 impactos	de 101 a 500 impactos	0 impactos	0 impactos	0 impactos	0 impactos
11 2024	Institución pública y gubernamental	Ambos	Castilla y León;Comunidad Valenciana;Cataluña;...	Mixta: combina las dos formas de financiación	De 2 a 5	No se realiza	Muy importante	No se realiza	Poco importante	...	de 50 a 100 impactos	de 50 a 100 impactos	de 50 a 100 impactos	de 50 a 100 impactos	de 50 a 100 impactos	de 50 a 100 impactos	de 50 a 100 impactos	de 50 a 100 impactos	de 50 a 100 impactos	de 50 a 100 impactos
12 2024	Institución pública y gubernamental	Ambos	Andalucía;Aragón;Asturias;Baleares;Canarias;Ca...	Pública: se financia en su totalidad con fondo...	De 51 a 100	No se realiza	Poco importante	Importante	Muy importante	...	0 impactos	de 101 a 500 impactos	de 101 a 500 impactos	de 101 a 500 impactos	de 101 a 500 impactos	< de 50 impactos	0 impactos	de 50 a 100 impactos	de 50 a 100 impactos	de 50 a 100 impactos

## PRINCIPALES PROBLEMAS DEL DATASET

- Posible sesgo en las respuestas
- Posibles preguntas mal diseñadas
- Podrían ser datos no representativos
- Las bases de datos no tienen el mismo número de columnas.
- Dificultad del análisis de las preguntas no codificadas
- Preguntas con muchas categorías difíciles de manejar.
- Nombres de las columnas larguísimos
- Texto sucio procedente de web
- ...

Shape del df\_2024: (126, 60)  
 Shape del df\_2023: (114, 54)  
 Shape del df\_2022: (72, 55)  
 Shape del df\_2021: (61, 50)  
 Shape del df\_2020: (41, 46)  
 Shape del df\_2019: (43, 39)

#	Column	Non-Null Count	Dtype
0	ID	126 non-null	int64
1	Hora de inicio	126 non-null	datetime64[ns]
2	Hora de finalización	126 non-null	datetime64[ns]
3	Correo electrónico	126 non-null	object
4	Nombre	0 non-null	float64
5	Hora de la última modificación	0 non-null	float64
6	Año	126 non-null	int64
7	Nombre de la iniciativa	126 non-null	object
8	Página web o redes sociales de la iniciativa	126 non-null	object
9	Breve descripción de la iniciativa	126 non-null	object
10	¿A qué tipo de organización pertenece tu iniciativa?	126 non-null	object
11	Indica el nombre de la organización	126 non-null	object
12	¿La iniciativa es puramente STEM o incide en el ámbito pero no es su foco principal?	126 non-null	object
13	Según el género, ¿a quién va dirigida la iniciativa?	126 non-null	object
14	Ámbito geográfico	126 non-null	object
15	¿En qué comunidades autónomas impacta la iniciativa?	126 non-null	object
16	¿En qué año inicia su actividad la iniciativa?	126 non-null	object
17	Indica el número de profesionales, asalariados y/o voluntarios, que trabajan activamente en la iniciativa	126 non-null	object
18	Según las fuentes de financiación, la iniciativa es:	126 non-null	object
19	Formación (curso, taller, etc.)	126 non-null	object
20	Evento (Jornada, congreso, ponencia, talks, etc.)	126 non-null	object
21	Mentoring	126 non-null	object
22	Recursos: material de formación o similar	126 non-null	object
23	Blog	126 non-null	object
24	Redes sociales	126 non-null	object
25	Bolsa de trabajo	126 non-null	object
26	Networking	126 non-null	object
27	Fomento y ayuda al emprendimiento	126 non-null	object
28	Otras	126 non-null	object
29	Periodicidad de la actividad. En el 2024, ¿aproximadamente, con qué frecuencia se realizó la actividad?	126 non-null	object
30	Indica el número total de repeticiones de la actividad dentro del periodo seleccionado.	125 non-null	object
31	Duración de las sesiones. Aproximadamente, ¿cuánto tiempo dura cada sesión de la actividad?	126 non-null	object
32	¿En qué formato se realiza la iniciativa?	126 non-null	object
33	Si mides el éxito de tu iniciativa ¿qué indicadores utilizas?	126 non-null	object
34	Las personas impactadas son en su mayoría (más del 50%):	126 non-null	object
35	De 1 a 5 años	126 non-null	object
36	De 6 a 9 años	126 non-null	object
37	De 10 a 12 años	126 non-null	object
38	De 13 a 14 años	126 non-null	object
39	De 15 a 16 años	126 non-null	object
40	Bachillerato	126 non-null	object
41	En los 2 primeros años de estudios de Grado y FP	126 non-null	object
42	Final de estudios: Grado, Máster, Doctorado, etc.	126 non-null	object
43	Etapa júnior (primeros años de la carrera profesional)	126 non-null	object
44	Etapa de consolidación de la carrera profesional	126 non-null	object
45	Etapa de madurez profesional (últimos años de carrera profesional)	126 non-null	object
46	En la pregunta anterior, ¿has reportado impactos de 1 a 16 años?	126 non-null	object
47	Impacto en niñas. Indica qué aporta la iniciativa y cómo ayuda/contribuye a que estudien carreras STEM.	76 non-null	object
48	Impacto en familia. Si tu iniciativa impacta en la familia, ¿nos indicas de qué manera?	76 non-null	object
49	Impacto en profesorado. Si tu iniciativa impacta en el profesorado, ¿nos indicas de qué manera?	76 non-null	object
50	De la siguiente lista, indica que necesitas para llevar a cabo la iniciativa.	74 non-null	object
51	Otros impactos o comentarios	25 non-null	object
52	¿La iniciativa está impulsada por alguna empresa o empresas en concreto?	126 non-null	object

## LIMPIEZA INICIAL

- Títulos de las columnas → Eliminar espacios en blanco y caracteres especiales “\xa0”
- Eliminar columnas innecesarias (23 columnas)
- Renombrar columnas y unificar títulos en todos los df
- Valores nulos:

CCAA → *Impacto en todas las comunidades autónomas (3 nulos)*

*se decide no hacer nada con estos valores porque corresponden a años en los que no se pedía dicha información.*

*El cuestionario ha ido evolucionado, por lo que algunas preguntas y respuestas no están presentes en todos los años, generando valores nulos.*

→ Concatenar los 6 DataFrames y crear el df\_2019\_2024

Valores nulos:		
año	43	←
2019	41	←
2020	0	←
2021	2	←
2023	0	
2024	0	
Name: formato, dtype: int64		

Col con valores nulos:
df_2019 → 24
df_2020 → 31
df_2021 → 30
df_2022 → 37
df_2023 → 10
df_2024 → 12

<class 'pandas.core.frame.DataFrame'>		
RangeIndex: 457 entries, 0 to 456		
Data columns (total 37 columns):		
#	Column	Non-Null Count Dtype
---	---	---
0	id	457 non-null int64
1	año	457 non-null int64
2	Nombre_iniciativa	457 non-null object
3	genero	457 non-null object
4	ambito_geografico	457 non-null object
5	CCAA	454 non-null object
6	año_inicio	457 non-null object
7	num_profesionales	457 non-null object
8	financiacion	457 non-null object
9	1_5_años	393 non-null object
10	6_9_años	365 non-null object
11	10_12_años	365 non-null object
12	13_14_años	365 non-null object
13	15_16_años	365 non-null object

Col con valores nulos:
df_2019 → 14
df_2020 → 15
df_2021 → 23
df_2022 → 23
df_2023 → 3
df_2024 → 3



## TRANSFORMACIÓN DE LOS DATOS

**Formato inicial:** 1 columna por etapa

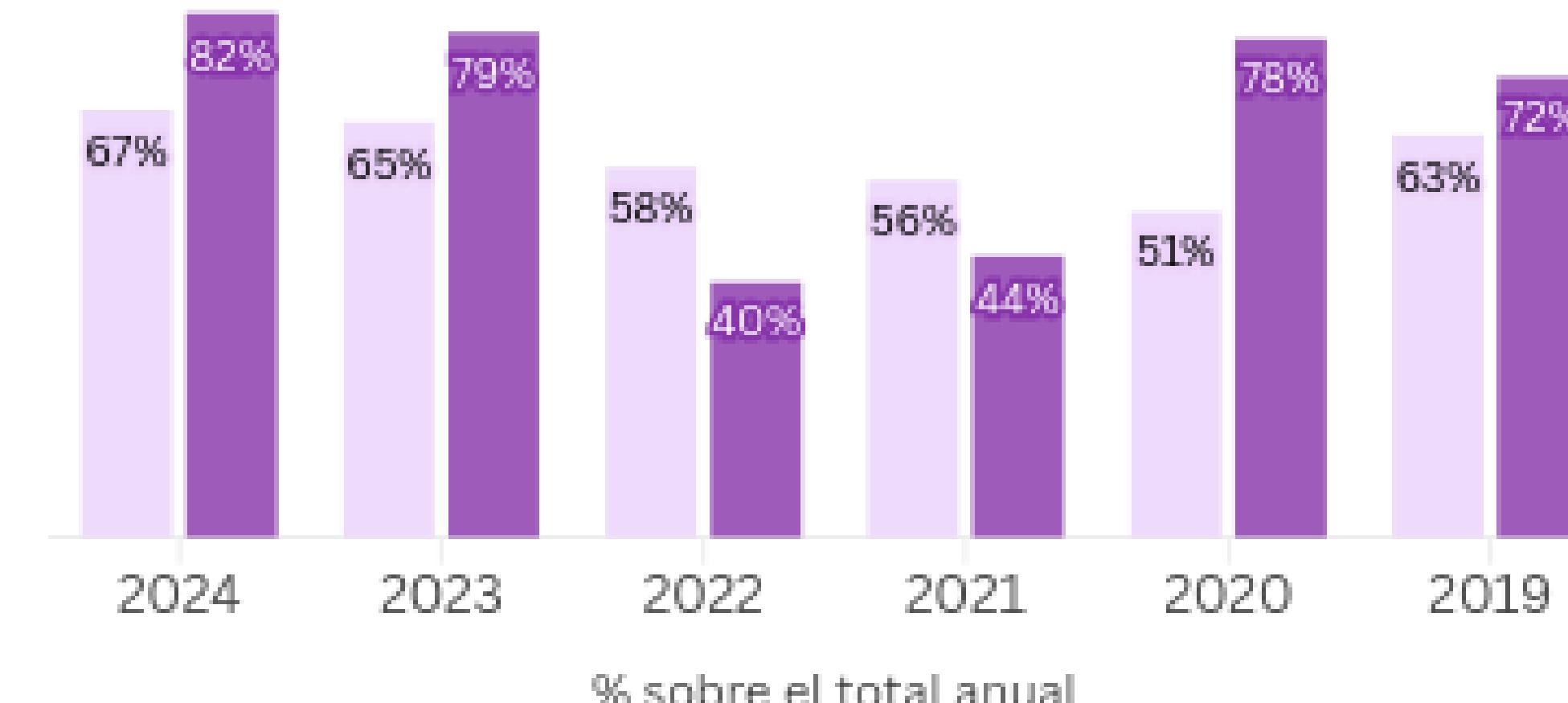
### Transformaciones:

- Crear col ['Total\_ipactos\_Ins']
- Crear col ['Total\_ipactos\_Carr']
- Crear col ['Total\_impactos']
- Crear col ["Segmento"]
- Crear df\_impactos

Categoría	Valor medio
0 impactos	0
< de 50 impactos	25
de 50 a 100 impactos	75
de 101 a 500 impactos	300
de 501 a 1.000 impactos	750
de 1.001 a 5.000 impactos	3000
> de 5.000 impactos	5001

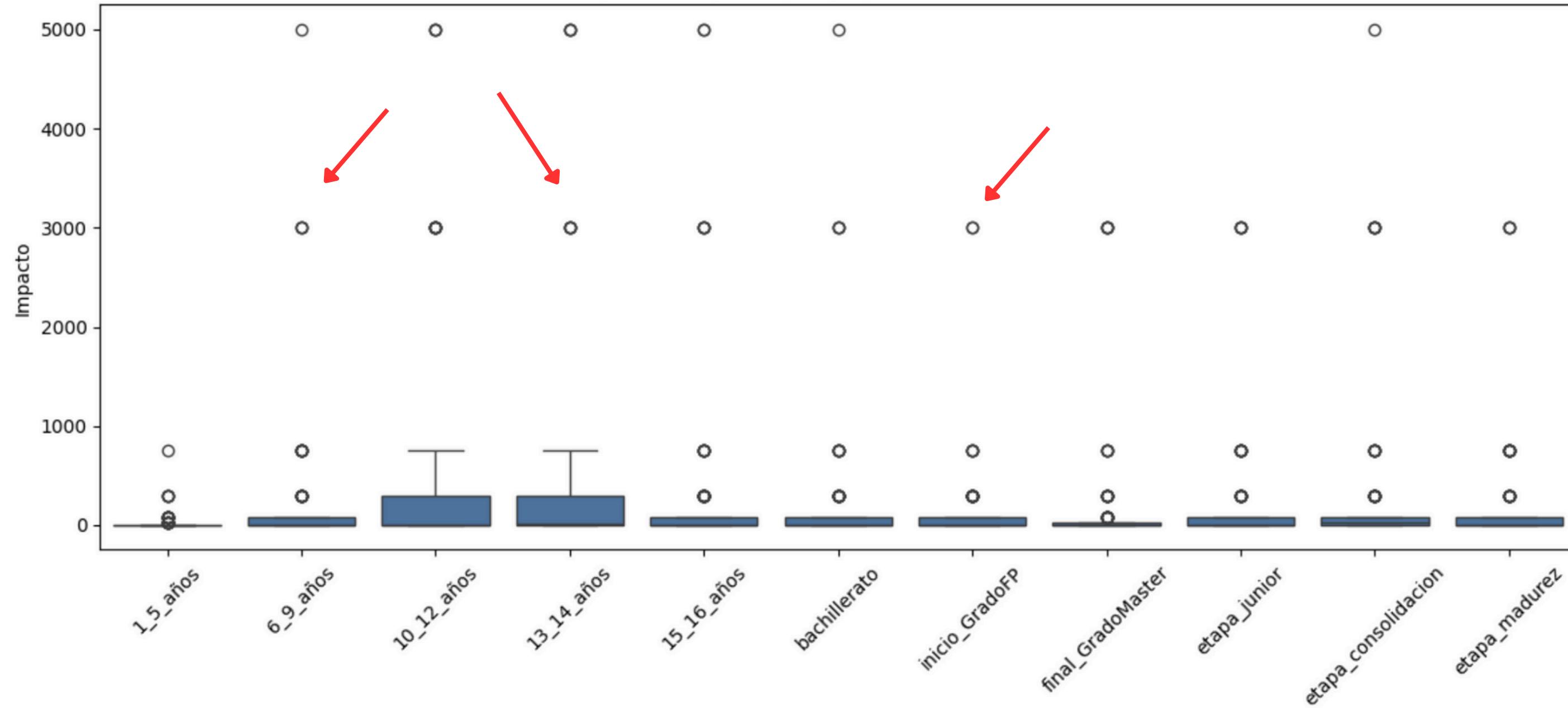
## Distribución de iniciativas por segmento

■ Inspiración ■ Carrera



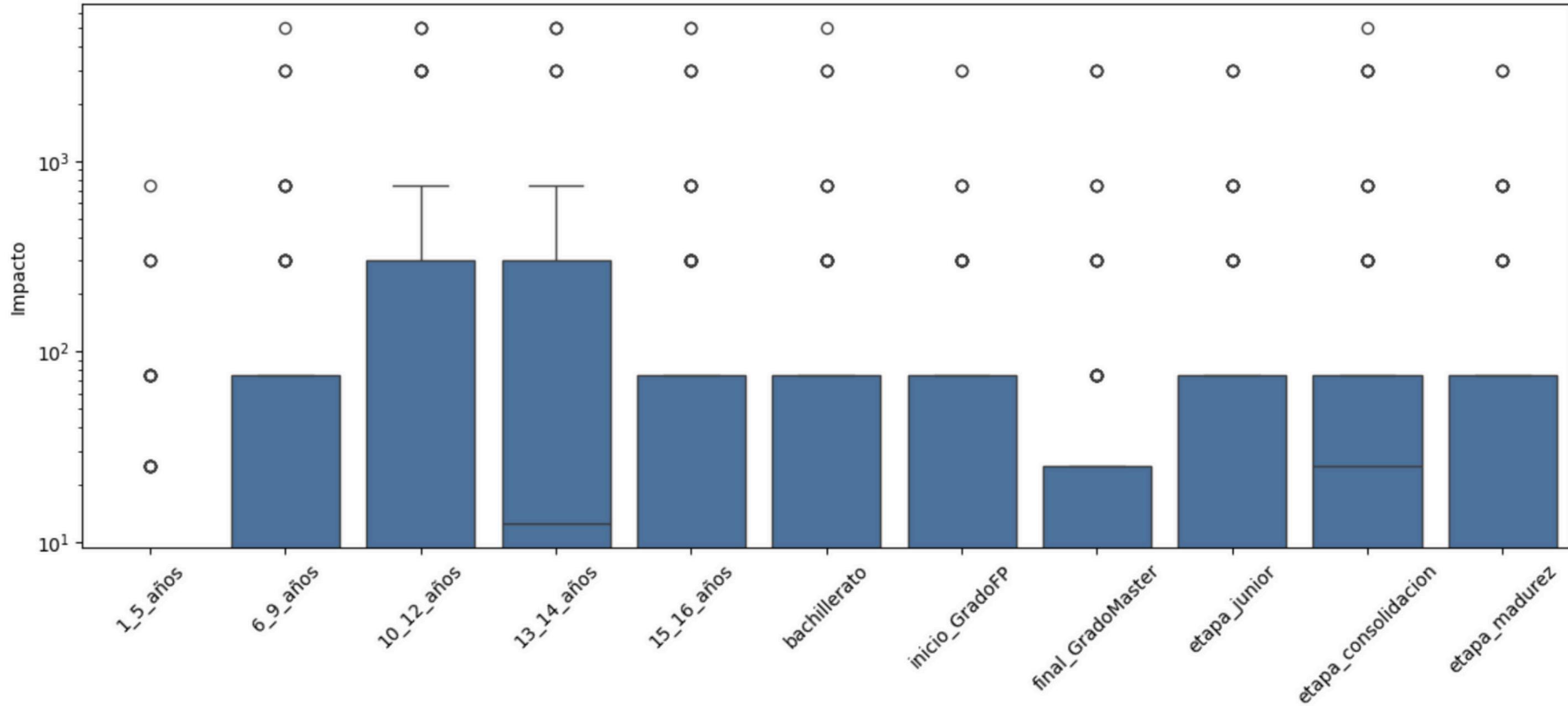


### Box plot de personas impactadas por grupo de edad (2024)





## Box plot de personas impactadas por grupo de edad - 2024 (escala logarítmica)

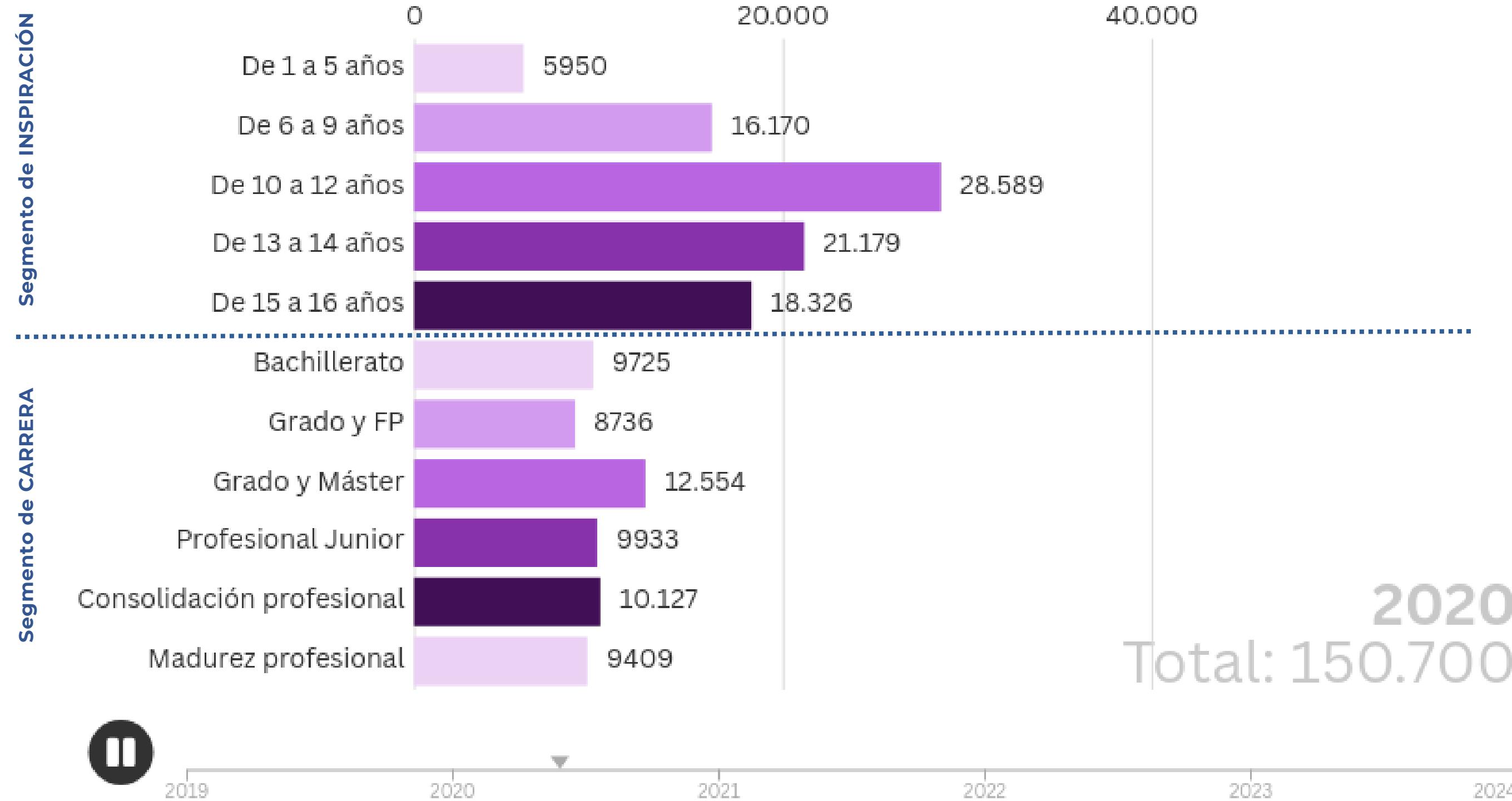


Mucha variabilidad entre las iniciativas



## Evolución del número de personas impactadas por las iniciativas (2019-2024)

### Según etapa educativa o profesional





## Segmento de INSPIRACIÓN

- EN ESPAÑA  
Las iniciativas han logrado impactar al menos en  
**151.309 niñxs**  
de 1 a 16 años durante el año 2024.

Representan un  
**2,11%**  
de una población de  
**7.178.832 niñ@s**

Fuente: Instituto Nacional de Estadística (INE)

## Segmento de CARRERA

Etapa de formación superior

- EN ESPAÑA  
Las iniciativas han logrado impactar aproximadamente a:  
**57.251 mujeres**  
A partir de 16 años durante el año 2024.

Representan un  
**9,70%**  
de una población de  
**590.323 mujeres estudiantes**

Fuente: Catálogo de datos. Estadística e Informes Universitarios. Ministerio de Ciencia e Innovación

## Segmento de CARRERA

Etapa profesional

- EN ESPAÑA  
Las iniciativas han logrado impactar aproximadamente a:  
**71.251 mujeres**  
de 20 a 65 años durante el año 2024.

Representan un  
**8,13%**  
de una población de  
**876.251 mujeres STEM**

Fuente: Fuente: Instituto Nacional de Estadística (INE). Población activa por grupo de edad, sexo y rama de actividad (CNAE: Información y comunicaciones y Actividades profesionales, científicas y técnicas)



## Distribución de iniciativas por Comunidad autónoma



2024

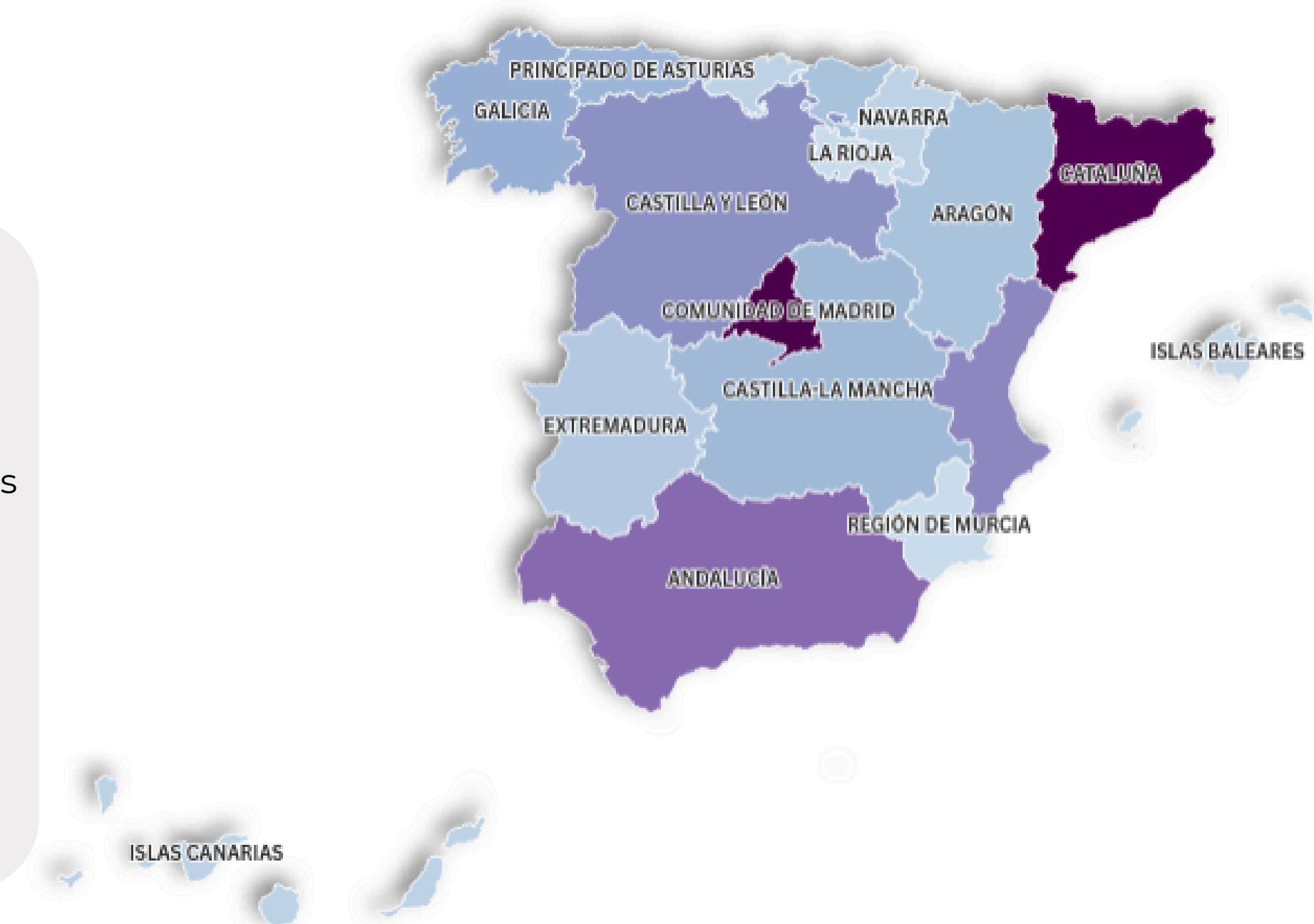
**Formato inicial:** 1 columna con múltiples categorías en forma de string.

### Transformaciones:

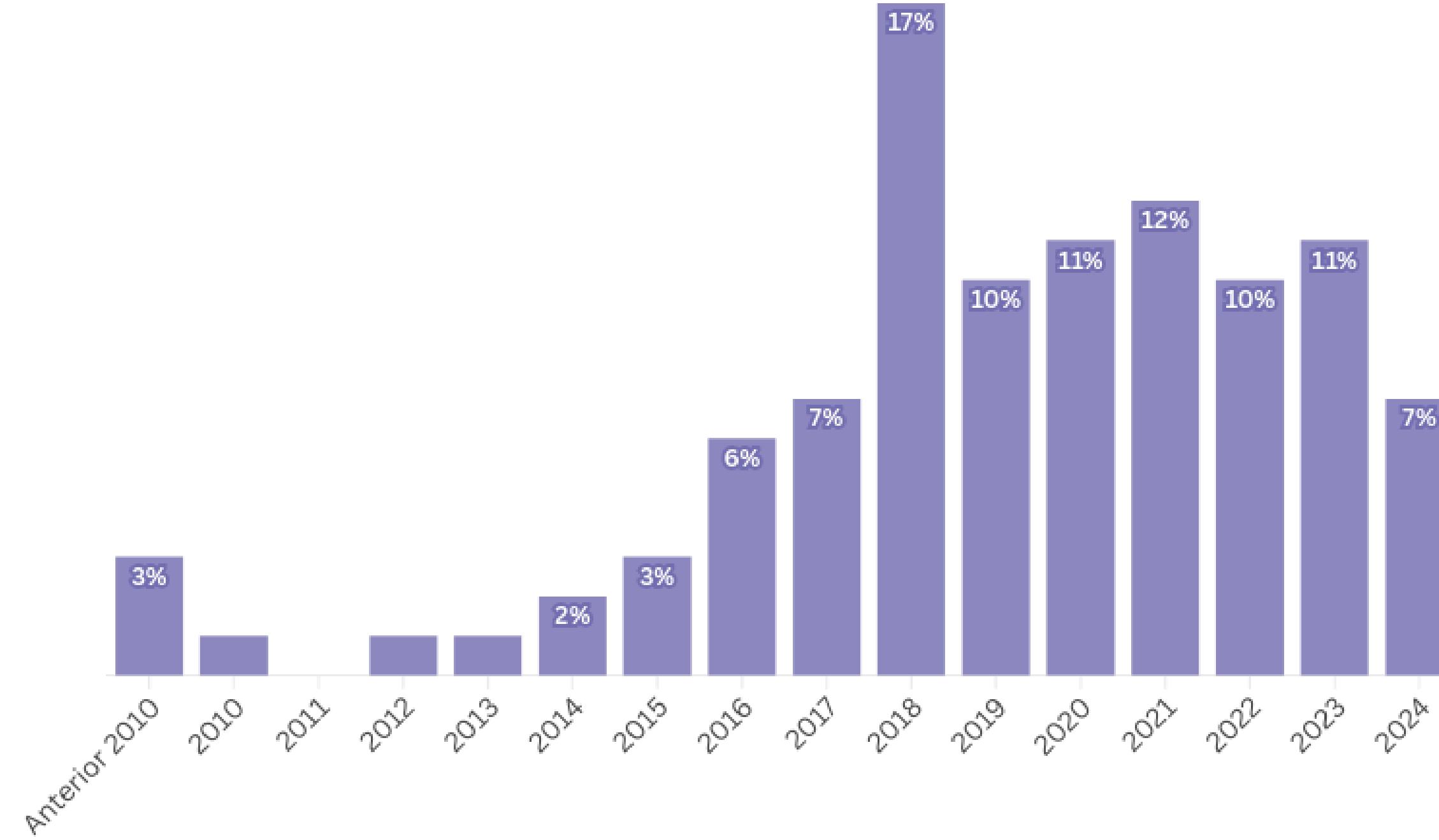
- sustituir 3 nulos por impacto en todas las CCAA.
- Convertir en 1 columna por CCAA y eliminar la original.

### `.get_dummies()`

- Renombrar los títulos de las columnas según GeoJSON
- Crear df\_CCAA

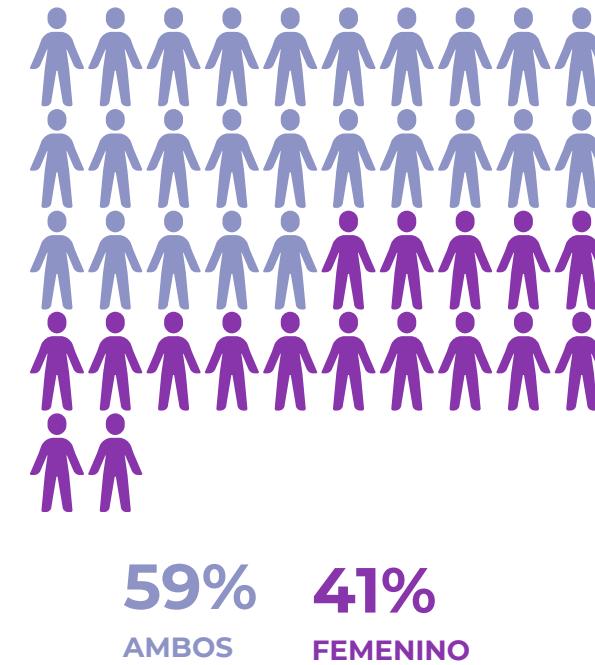


## Distribución de las iniciativas según el año de creación

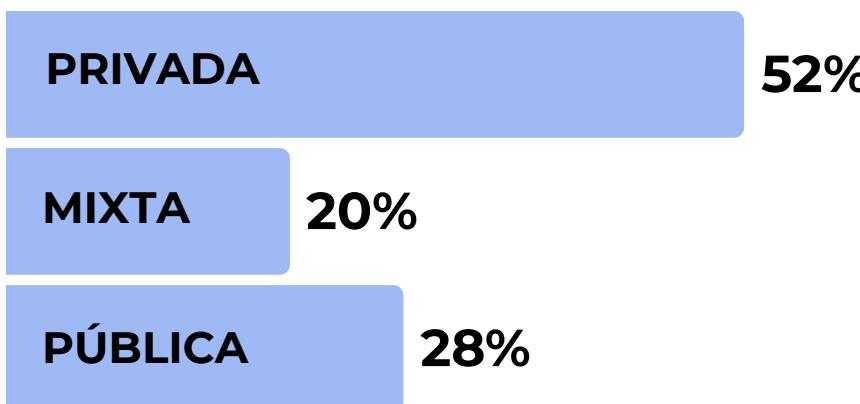




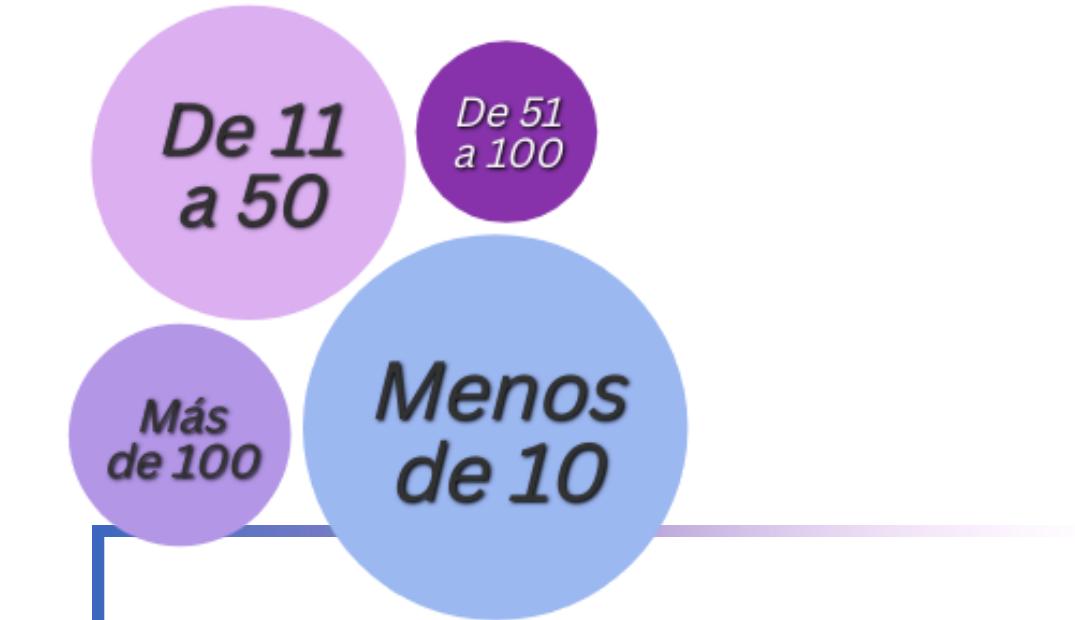
## Según el género a quién va dirigida la iniciativa



## Fuente de financiación de la iniciativa



## Número de profesionales que trabajan en la iniciativa (%)



### Transformaciones:

- Calcular el promedio de trabajadores por rango y multiplicarlo por el conteo.

Calculamos que en España en el 2024, al menos, **4.400 personas** trabajaron activamente para empoderar a las niñas y mujeres STEM

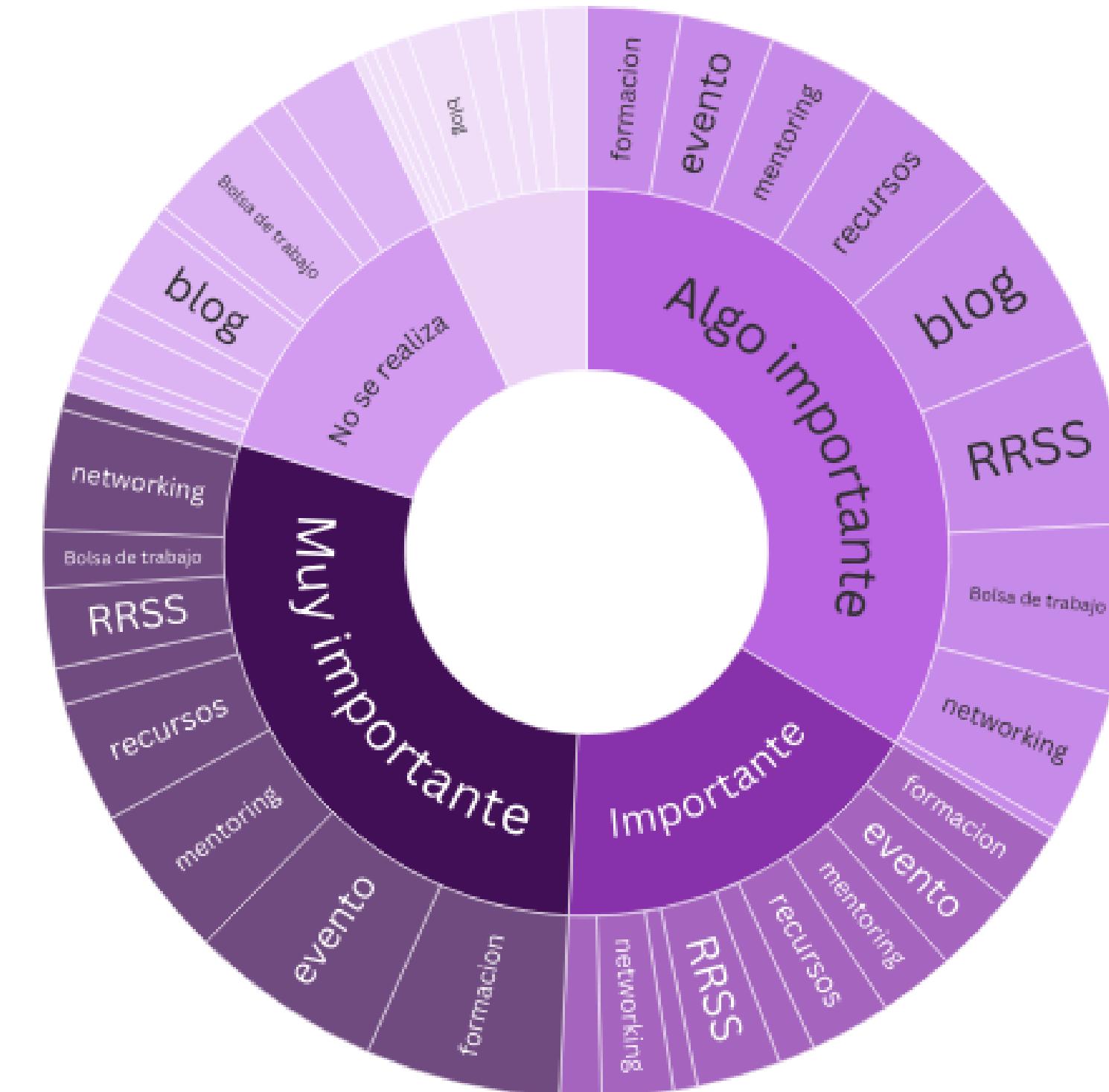


## Actividad realizada por la iniciativa según importancia

All

### Transformaciones:

- Categorías:
  - Muy importante
  - Importante
  - Algo importante
  - Poco importante
  - No se realiza
- Unificar categorías. No son las mismas para todos los años.



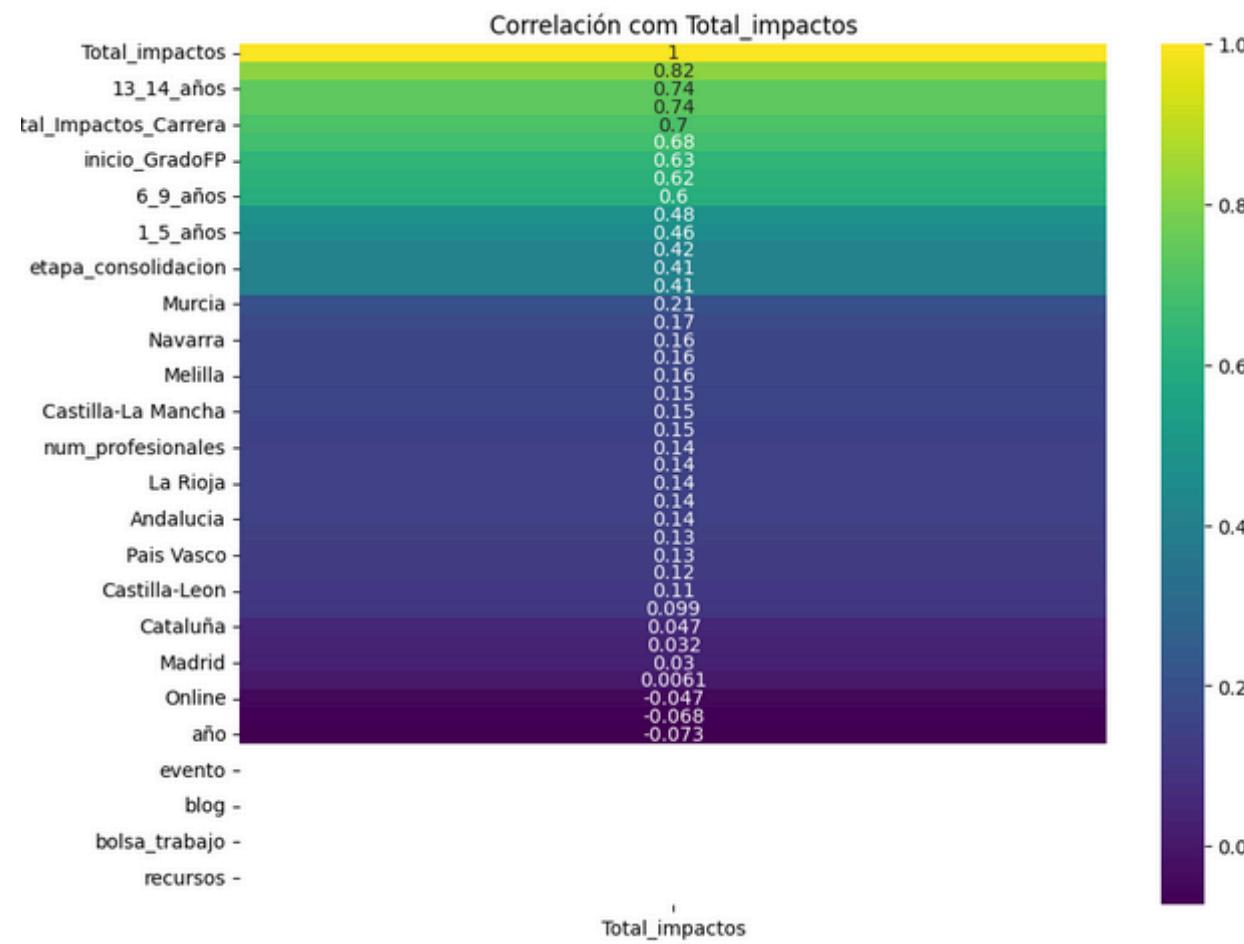
# DataFrame limpio

	genero	ambito_geografico	año_inicio	num_profesionales	financiacion	1_5_años	6_9_años	10_12_años	13_14_años	15_16_años	...	Ceuta	Valencia	Extremadura	Galicia	Madrid	Melilla	Murcia	Navarra	Pais Vasco	La Rioja
0	Ambos	Nacional	2017	30.5	Mixta	0	25	300	300	75	...	0	0	0	0	0	0	0	0	0	0
1	Femenino	Nacional	2018	3.5	Privado	0	25	25	25	25	...	0	0	0	0	0	0	0	0	0	0
2	Ambos	Nacional	2019	101.0	Mixta	0	25	25	75	75	...	0	1	0	0	1	0	0	0	0	0
3	Ambos	Nacional	2016	30.5	Publica	0	5001	5001	5001	5001	...	1	1	1	1	1	1	1	1	1	1
4	Ambos	Nacional	2012	3.5	Privado	0	0	3000	3000	0	...	0	0	0	0	1	0	0	0	0	0
5	Ambos	Nacional	2015	30.5	Privado	0	0	0	300	300	...	0	0	0	0	1	0	0	0	0	0
6	Femenino	Ambos	2015	3.5	Mixta	0	0	0	300	300	...	0	0	0	0	1	0	0	0	0	0
7	Femenino	Ambos	2016	8.0	Privado	0	0	750	750	750	...	0	0	0	0	1	1	0	0	0	0
8	Femenino	Ambos	2016	101.0	Mixta	0	0	3000	0	0	...	0	0	0	0	1	1	0	1	0	0
9	Ambos	Ambos	2016	30.5	Mixta	0	0	300	300	750	...	1	1	1	1	1	1	1	1	1	1
10	Femenino	Ambos	2016	101.0	Mixta	0	0	750	750	750	...	1	1	0	0	1	0	1	0	0	0
11	Ambos	Nacional	2018	30.5	Privado	0	0	300	25	0	...	0	1	0	0	1	0	0	0	1	0
12	Ambos	Nacional	2018	30.5	Publica	0	0	3000	3000	0	...	0	0	0	0	0	0	0	0	0	0
13	Ambos	Nacional	2018	75.5	Publica	0	0	3000	25	0	...	0	0	0	0	0	0	0	0	0	0
14	Ambos	Nacional	2019	75.5	Publica	0	0	75	75	25	...	0	0	0	0	0	0	0	0	0	0
15	Ambos	Nacional	2019	8.0	Privado	0	0	25	25	25	...	0	0	0	0	1	0	0	0	0	0
16	Ambos	Nacional	2019	3.5	Mixta	0	0	0	0	3000	...	0	0	0	0	0	0	0	0	0	0
17	Ambos	Nacional	2016	8.0	Privado	0	300	300	300	300	...	0	0	0	0	1	0	0	0	0	0
18	Femenino	Nacional	2017	75.5	Privado	0	300	750	75	75	...	0	1	0	1	1	0	0	0	1	0
19	Ambos	Internacional	2018	3.5	Privado	0	300	3000	3000	25	...	1	1	1	1	1	1	1	1	1	1

## 3. MODELOS DE MACHINE LEARNING



## Heatmap de Correlación con Total de Impactos



## Mayores correlaciones:

- 13\_14\_años (0.74)
  - Total\_Impactos\_Carrera (0.72)
  - inicio\_GradoFP (0.68)

Bajas correlaciones con:

- evento, blog, recursos, algunas regiones.

**Conclusión:** Las franjas etarias específicas y etapas tempranas de la carrera tienen mayor influencia en el impacto total, mientras que las variables de preferencia guían la personalización de las recomendaciones.

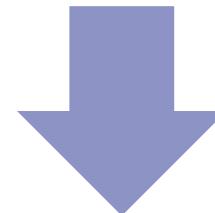
## Importación y entrenamiento del modelo

Se utiliza un RandomForestRegressor para entrenar el modelo con los datos. y calcular la importancia de cada variable con la columna Total\_impactos.

**Resultado: Se muestran las 10 variables más importantes.**

```
Total_Impactos_Inspiracion      0.531784  
Total_Impactos_Carrera         0.368725  
6_9_años                      0.022103  
inicio_GradoFP                 0.018657  
bachillerato                   0.014820  
13_14_años                    0.010623  
1_5_años                      0.008000  
final_GradoMaster              0.004165  
año                            0.003688  
10_12_años                    0.003238  
dtype: float64
```

- Para un modelo de recomendación, variables como formación, mentoría y evento deben seguir usándose, aunque no estén muy correlacionadas con el impacto total.
  - Esto se debe a que el objetivo es alinear las recomendaciones con los intereses del usuario.



## Mapeo de importancia: mapeando respuestas categóricas a valores numéricos



## Elección del Modelo

### 1. Modelo de Recomendación

- Objetivo: Sugerir iniciativas personalizadas para cada usuario o grupo, aumentando el impacto y la participación.
- Aplicaciones: Recomendación de actividades, contenidos educativos o conexiones de networking

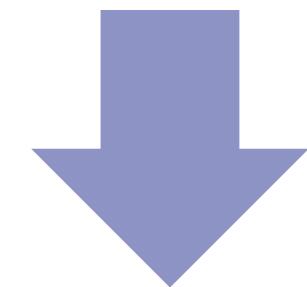
### 2. Modelo de Predicción

- Objetivo: Predicción de Matrículas a partir del Impacto de las Iniciativas
- Aplicaciones: Anticipar resultados, identificar factores de éxito y optimizar la toma de decisiones.

- Para el **modelo de recomendación**, variables como:

el tipo de actividad de las iniciativas  
franja de edad  
formato de presentación.

- Esto se debe a que el objetivo es alinear las recomendaciones con los intereses del usuario.



- Estandarizar

- **Mapeo de importancia: mapeando respuestas categóricas a valores numéricos**



## Modelo de Recomendación

Algoritmo elegido: cosine similarity y MinMaxScaler

### Pasos Principales:

#### Filtrado por Edad

Se filtran las iniciativas según la columna de la edad del perfil del usuario.

#### Filtrado por Formato

Filtrar según el formato preferido (Híbrida, Online, Presencial).

#### Filtrado por Actividad Preferida

Se filtran las actividades que el usuario considera "importantes" (valor  $\geq 2$ ).

#### Cálculo de Similaridad

Usamos cosine similarity para calcular la similitud entre el perfil del usuario y las iniciativas.

#### Resultados

Se muestran las iniciativas más similares según el perfil.

## Dificultades y Soluciones:

### Padronización de Nombres

Se normalizaron los nombres de las columnas con `.strip()` y `.lower()` para evitar inconsistencias.

### Clasificaciones Numéricas vs. Binarias

Las actividades se mapearon a valores numéricos para poder filtrarlas correctamente.

### Filtrado de Actividades

Aseguramos que solo se seleccionaran actividades con valor  $\geq 2$  ("importante" o "muy importante").

### Normalização dos vetores com MinMaxScaler

### Datos Faltantes o Erróneos

Se revisaron y ajustaron las columnas para evitar valores nulos o erróneos.

### Ajuste de Similaridad

Se corrigió la definición del vector y la comparación de similaridad, ajustando la matriz de actividades de manera que los resultados fueran más relevantes.

### Ejemplo de resultado:

- Total após filtro de edad: 190
- Total após filtro de formato: 67
- Total após filtro de actividad preferida: 55

nombre_iniciativa	
109	#ChicasImparablesTech
350	Charla STEM
307	Dia Internacional de la Mujer y la Niña en la ...
324	L'enginy (in)visible - El ingenio (in)visible ...
392	Escape Road: A la búsqueda de las científicas ...

formato_simplificado	Similitud
Presencial	0.666667
Hibrida;Presencial	0.609994
Presencial	0.577350
Online;Presencial	0.577350
Hibrida;Online;Presencial	0.557086



## HIPERPARÁMETROS DEL MODELO:

- **top\_n:** Define el número de recomendaciones a mostrar. Por defecto es 5, pero se puede ajustar según la necesidad.
- **vector:** Representa las preferencias del usuario en actividades. El valor de cada actividad (como 3 para preferida) se puede modificar según cómo se deseen ponderar las preferencias.
- **Método de Similaridad:** Se usa cosine\_similarity para comparar las preferencias.
- **Filtros y Preprocesamiento:** La forma en que se filtran los datos, como las actividades preferidas y las edades, también influye en los resultados.

## AFINACIÓN DE HIPERPARÁMETROS

La afinación principal se centra en la **preparación de los datos**, como el **mapeo de valores** y la **definición de características**. Esto incluye cómo se asignan los valores a las actividades y qué características se consideran más relevantes. Además, **ajustar el cálculo de la similaridad** entre el perfil del usuario y las iniciativas puede mejorar las recomendaciones.

## EVALUACIÓN DEL MODELO

### Métricas usadas:

Precisión y recall basados en coincidencias simuladas (actividad + formato).

### Resultados:

```
- Total após filtro de edad: 190  
- Total após filtro de formato: 67  
- Total após filtro de actividad preferida: 55  
Precision: 0.0  
Recall: 0.0
```

### Posibles causas:

Criterio de relevancia muy estricto.  
Ausencia de datos reales de usuario.  
Simulación poco representativa.  
Escalado con advertencias (input sin nombres).  
Evaluación circular (criterio = filtro).



## POSIBLES CAUSAS DEL BAJO RENDIMIENTO DEL MODELO

- No hay datos reales de usuarios (**ground truth**).
- **y\_true** y **relevancia\_preditiva** fueron definidos con las mismas reglas del modelo → genera circularidad

### Resultado:

- Métricas como Precision = 0.0, Recall = 0.0 no reflejan la calidad real del sistema.

### Evaluación inválida porque:

- El modelo no está siendo probado contra elecciones reales.
- Solo confirma lo que él mismo ya filtró.
- Las métricas supervisadas solo tienen sentido con retroalimentación del usuario.

## POSIBLES SOLUCIONES

### Evaluación cualitativa

- Verifica manualmente si las recomendaciones tienen sentido.
- Pide evaluación a especialistas.

### Pruebas de coherencia

- Cambia el perfil → observa si las recomendaciones cambian lógicamente.
- Verifica la diversidad en las recomendaciones.

### Futuro (con datos reales)

- Recoger interacciones (clics, elecciones, valoraciones).
- Aplicar métricas como Precision@N, Recall@N, NDCG



### SOLUCIÓN FINAL

#### Evaluación del Modelo de Recomendación

Dado que no tenemos datos reales de usuario, utilizamos una evaluación simulada:

- Precisión simulada: Compara las recomendaciones con las iniciativas relevantes según el perfil del usuario.
- RMSE simulado: Mide la diferencia entre las preferencias del usuario y las recomendaciones.
- Engagement simulado: Genera valores aleatorios para simular la tasa de interacción.
- Evaluación Cualitativa:
- Verifica manualmente si las recomendaciones tienen sentido.
- Pide la evaluación de expertos.

#### Limitaciones:

- Las métricas no reflejan interacciones reales, solo aproximaciones del modelo.
- Próximos pasos: Evaluar con datos reales para usar métricas como Precision@N, Recall@N y NDCG.



## Modelo de Predicción

### Algoritmo

- Regresión Lineal (LinearRegression de scikit-learn)
- Relación lineal entre impacto de iniciativas y matrículas

Objetivo: Predicción de Matrículas a partir del Impacto de las Iniciativas

#### 1. Unión de DataFrames:

- Combinamos df\_iniciativas (impacto de iniciativas) con df\_matriculadas (matrículas en bachillerato) usando la columna año.

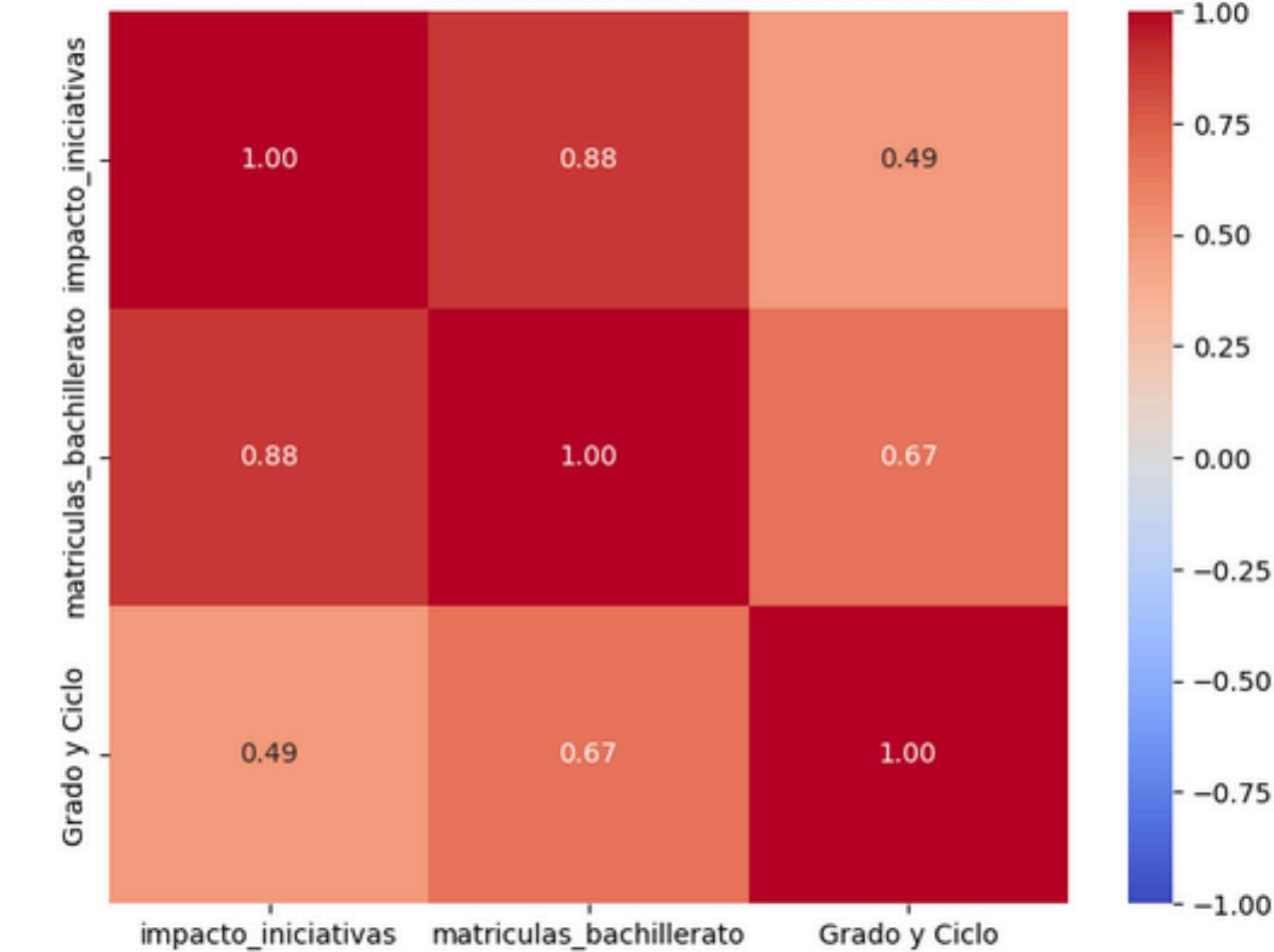
#### 2. Modelo de Predicción:

- Utilizamos regresión lineal para predecir las matrículas en los próximos 5 años basándonos en el impacto de las iniciativas pasadas.

#### 3. Resultados:

- Predicción de matrículas para los próximos años y visualización en un gráfico.
- La matriz de correlación mostró la relación entre el impacto de las iniciativas y las matrículas en bachillerato.

Matriz de Correlación de las Variables Seleccionadas



Qué muestra el resultado?

- impacto\_iniciativas y matriculas\_bachillerato tienen una correlación muy alta (0.88), lo que indica que cuando aumenta una, la otra también tiende a aumentar.
- matriculas\_bachillerato y Grado y Ciclo también tienen una correlación positiva considerable (0.67).
- impacto\_iniciativas y Grado y Ciclo tienen una correlación positiva moderada (0.49).

Esto sugiere que existe una relación importante entre el impacto de las iniciativas y el número de matrículas, y que también hay relación con el grado/ciclo educativo.



## HIPERPARÁMETROS DEL MODELO:

Se usaron los valores por defecto:

- **fit\_intercept=True**
- **normalize=False**

Sin normalización ni ajustes especiales

### R<sup>2</sup> (Coeficiente de Determinación)

Valor entre 0 y 1 (cuanto más cercano a 1, mejor).

Mide la proporción de la varianza explicada por el modelo.

Ideal para evaluar el ajuste general.

### MAE (Error Absoluto Medio)

Promedio de las diferencias absolutas entre las predicciones y los valores reales.

Fácil de interpretar (misma unidad que la variable predicha).

### MSE (Error Cuadrático Medio)

Los errores grandes son penalizados más fuertemente.

Sensible a valores atípicos (outliers).

### RMSE (Raíz del MSE)

Interpretación más intuitiva del MSE (misma unidad que las matrículas).

## AFINACIÓN DE HIPERPARÁMETROS

- Técnica utilizada: GridSearchCV
- Validación cruzada: LeaveOneOut (ideal para pocos datos)

R<sup>2</sup>: 0.7677642354192395

MAE: 1590.692537957168

RMSE: 2072.693916783087

Modelo afinado: Ridge (Regresión con regularización)

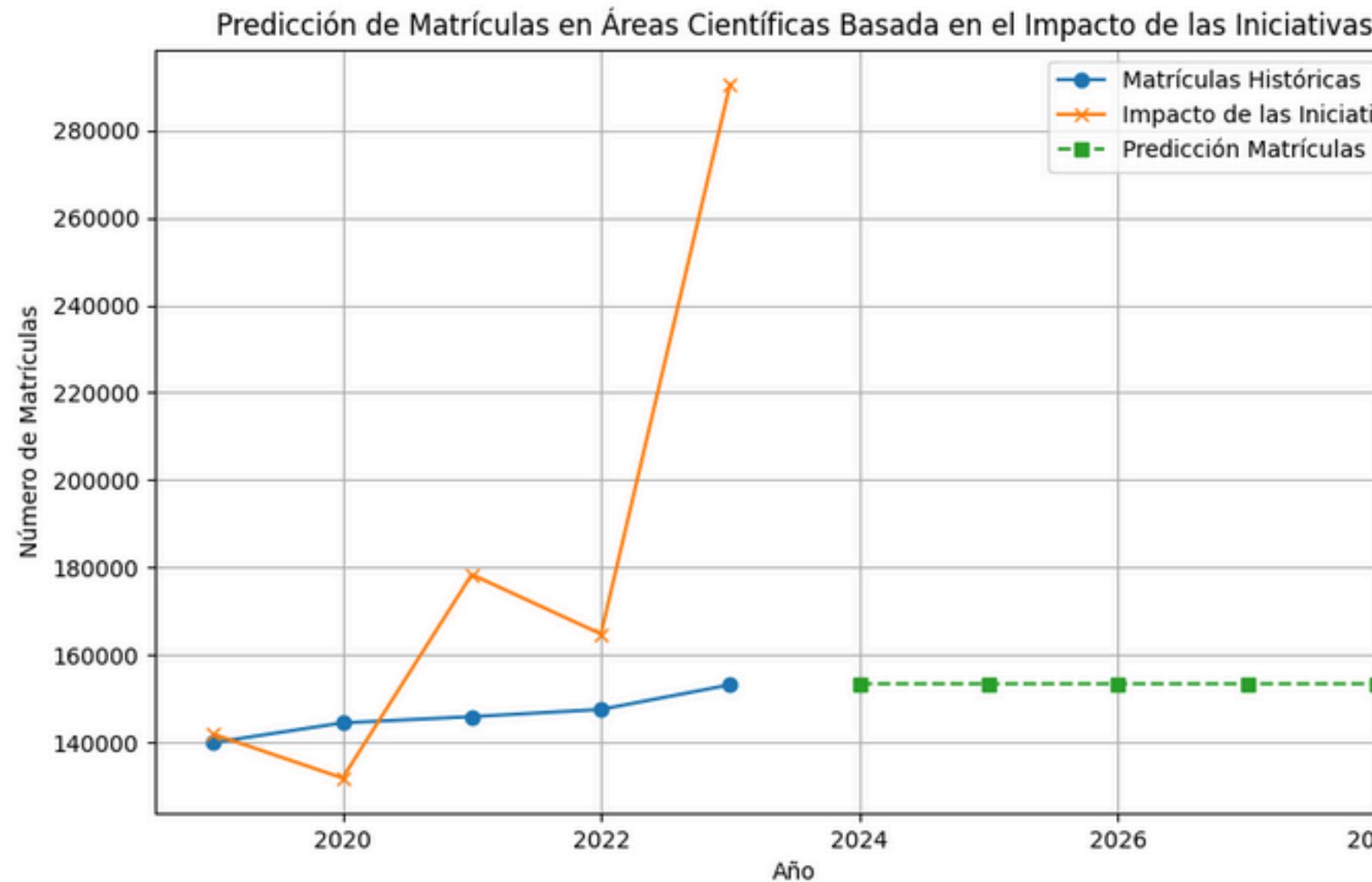
- Hiperparámetro ajustado: alpha
- Valores probados: 0.01, 0.1, 1, 10, 100
- Mejor resultado: alpha = 0.01

## Desafíos Encontrados

- Base de datos pequeña.
- Falta de datos antes de las iniciativas y poca diversidad en las características.
- Alta variabilidad en el impacto de las iniciativas, pero estabilidad en las matrículas.
- Riesgo de sobreajuste debido al tamaño de los datos y variabilidad del impacto.



## Análisis de Predicción de Matrículas Científicas (2024–2028)



### ¿Qué muestra el gráfico?

- Matrículas históricas: crecimiento leve y estable hasta 2022.
- Impacto de iniciativas: gran variabilidad, fuerte aumento en 2023.
- Predicción 2024–2028: matrículas se mantienen constantes (~153.398).

### Interpretación

- El modelo predice estabilidad, sin reflejar el aumento del impacto.
- Indica que el modelo es conservador o insensible a cambios recientes.
- Podría haber factores no incluidos que influyen más que el impacto.

### Limitaciones del análisis

- Falta de datos previos a las iniciativas.
- Escasez de datos desagregados por género (mujeres).
- Esto limita la evaluación real del impacto de las iniciativas.

## 5. CONCLUSIONES



## CONCLUSIONES

- El modelo de predicción muestra estabilidad en las matrículas para 2024–2028, sin reflejar el reciente aumento por impacto de iniciativas, lo que indica un enfoque conservador y posibles limitaciones por falta de datos históricos y desagregados.
- El modelo de recomendación, evaluado de forma simulada ante la ausencia de datos reales de usuario, ofrece resultados prometedores, aunque las métricas actuales solo aproximan el rendimiento real.
- Ambos modelos enfrentan desafíos por la escasez y calidad de los datos, lo que limita la evaluación precisa del impacto y la personalización.
- Como próximos pasos, es clave incorporar datos más completos y reales para validar y mejorar la capacidad predictiva y la relevancia de las recomendaciones, utilizando métricas especializadas como Precision@N, Recall@N y NDCG.
- Somos conscientes de las carencias que tiene la base de datos: principalmente, la alta disparidad en los valores de impacto entre iniciativas, muchas con bajo alcance y unas pocas con impactos muy elevados pero que distorsionan la media. Por ello, como mejora futura que ya se está trabajando, se plantea la **creación de un indicador de impacto ponderado**, que tenga en cuenta factores como la duración de la iniciativa, la frecuencia con la que se repite y el tipo de actividad. Esto permitiría obtener una visión más justa y representativa del impacto real. Sin embargo, resulta muy complejo, ya que cada iniciativa suele englobar múltiples actividades con características distintas, dificultando una respuesta clara sobre duración y repetición.

# GRACIAS