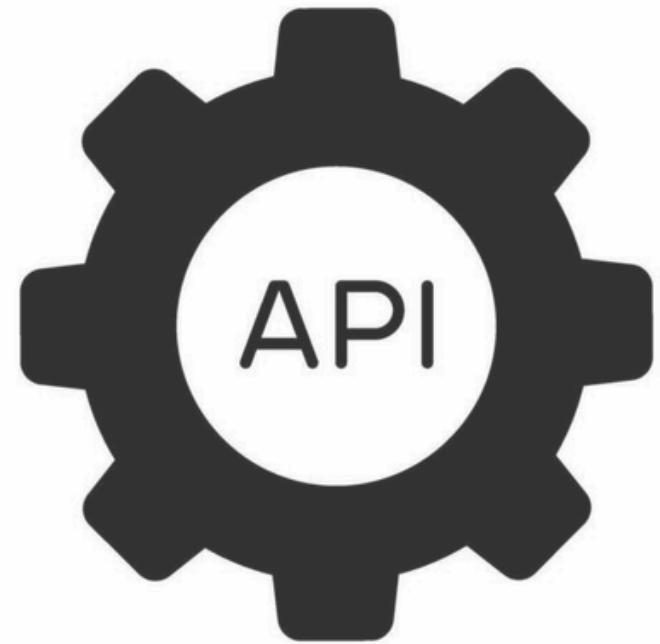


Project_Break_I_EDA

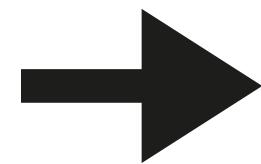
Data Science Repository Trends on GitHub:
An Insightful EDA



Lucas Perez Barone



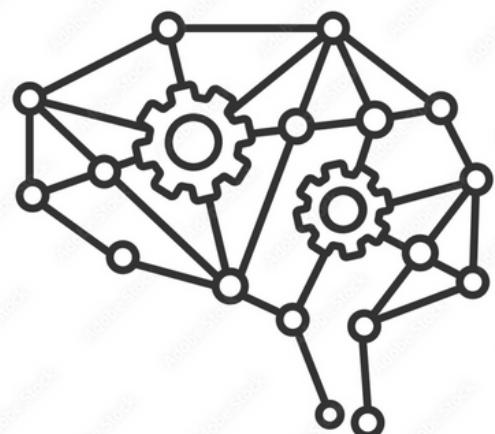
Panorama



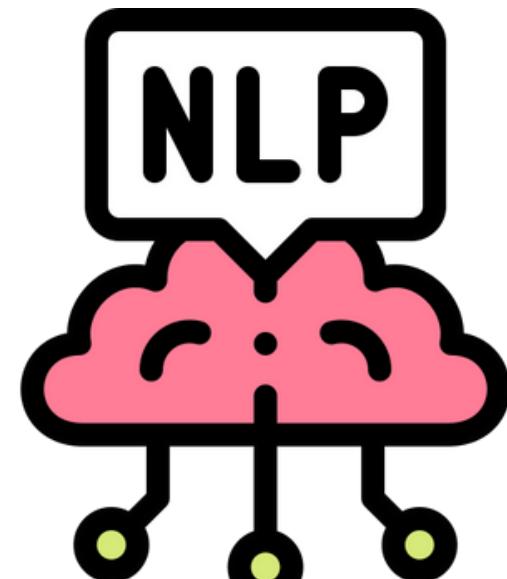
El mercado de la ciencia de datos sigue evolucionando de manera acelerada



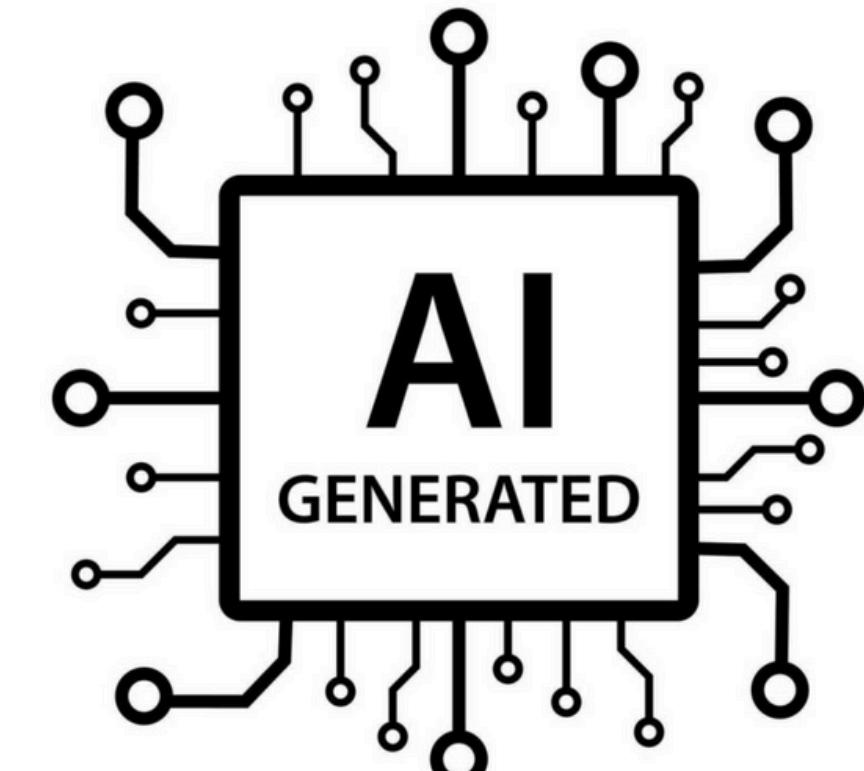
BIGDATA



DEEP LEARNING



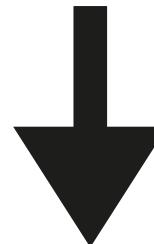
BUSINESS INTELLIGENCE



tiempo

Objetivos

- Identificar cuáles de estos temas y subáreas de la Ciencia de Datos han incrementado su relevancia.
- Identificar posibles nuevas tendencias para el mercado actual y futuro.



Cuáles están más asociadas con la **popularidad** y el **compromiso activo** de los repositorios de GitHub.



Principal plataforma de desarrollo
y control de versiones de código

¿Por qué es importante?

Profesionales/Estudiantes



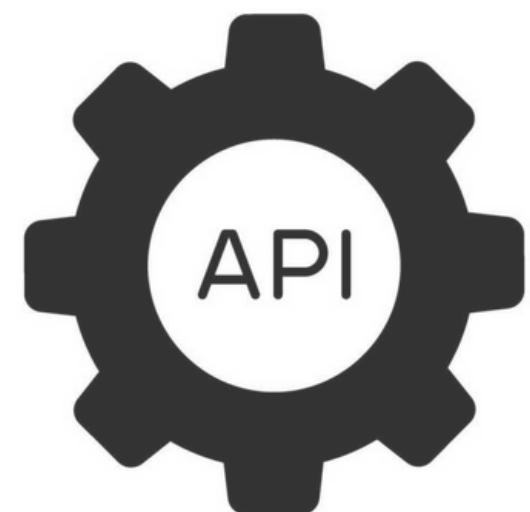
→ Planificar portafolio y desarrollo profesional/técnico

Empresas/Instituciones



→ Planificar desarrollo interno;

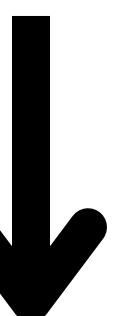
→ Actualizarse con el mercado, mejorando la competitividad.



Conjunto de Datos

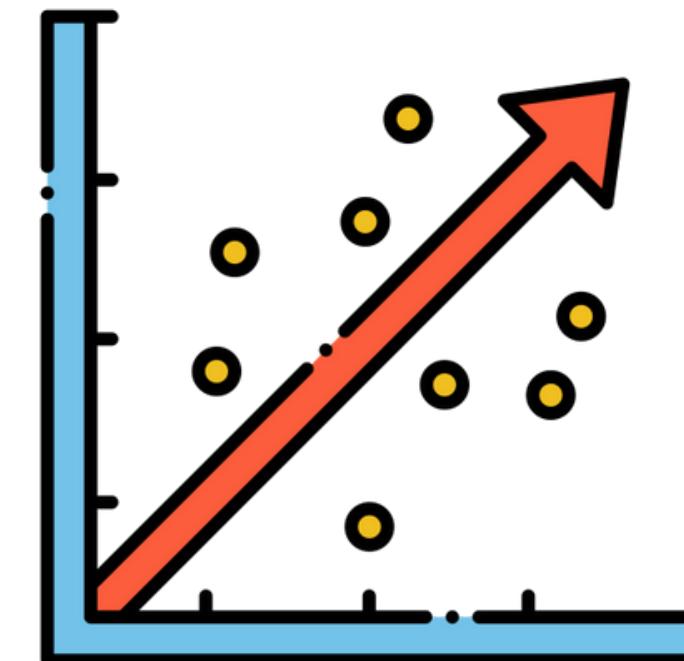


Procesamiento/ Limpieza



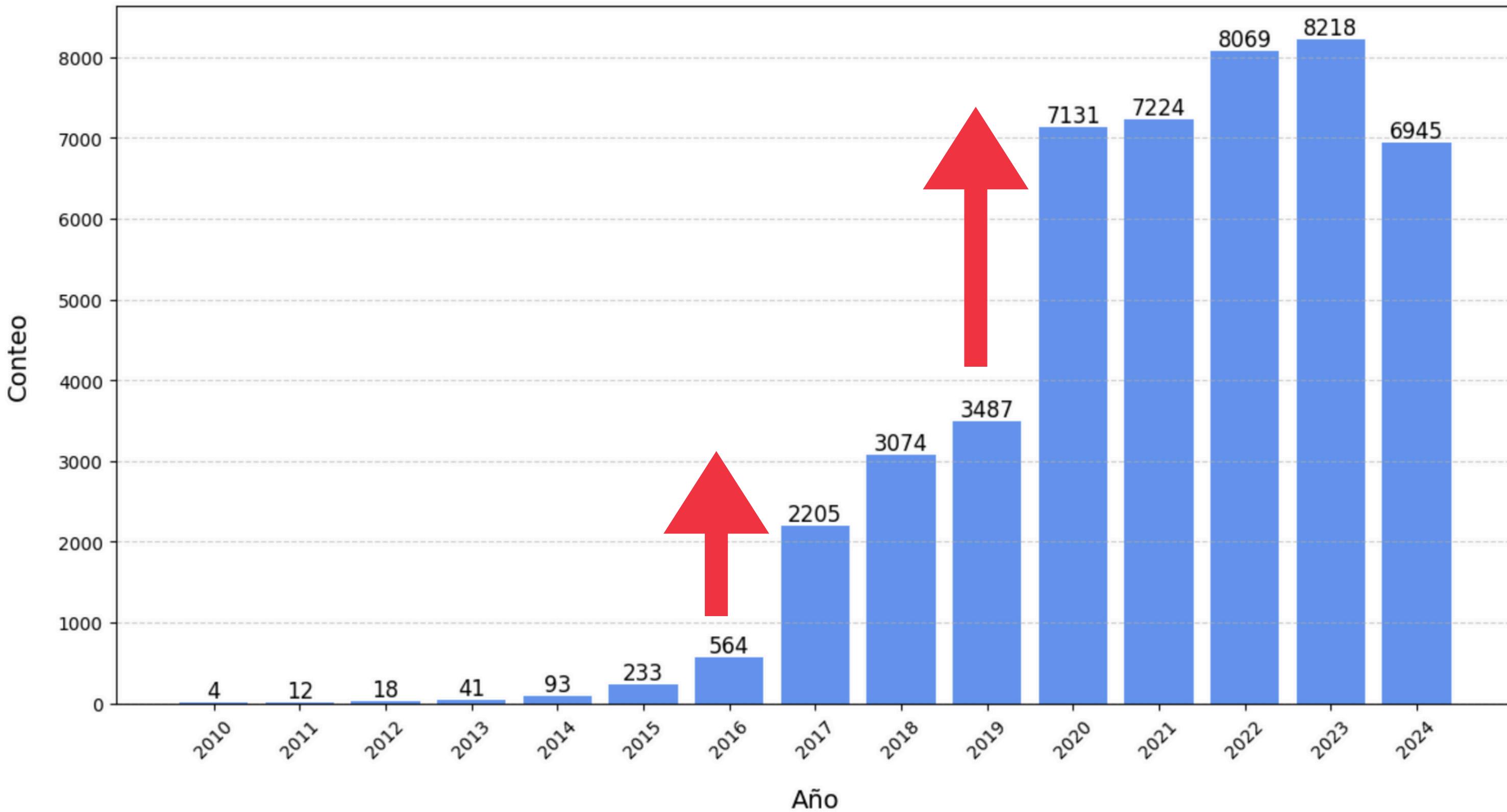
EDA

Análisis Exploratorio de Datos



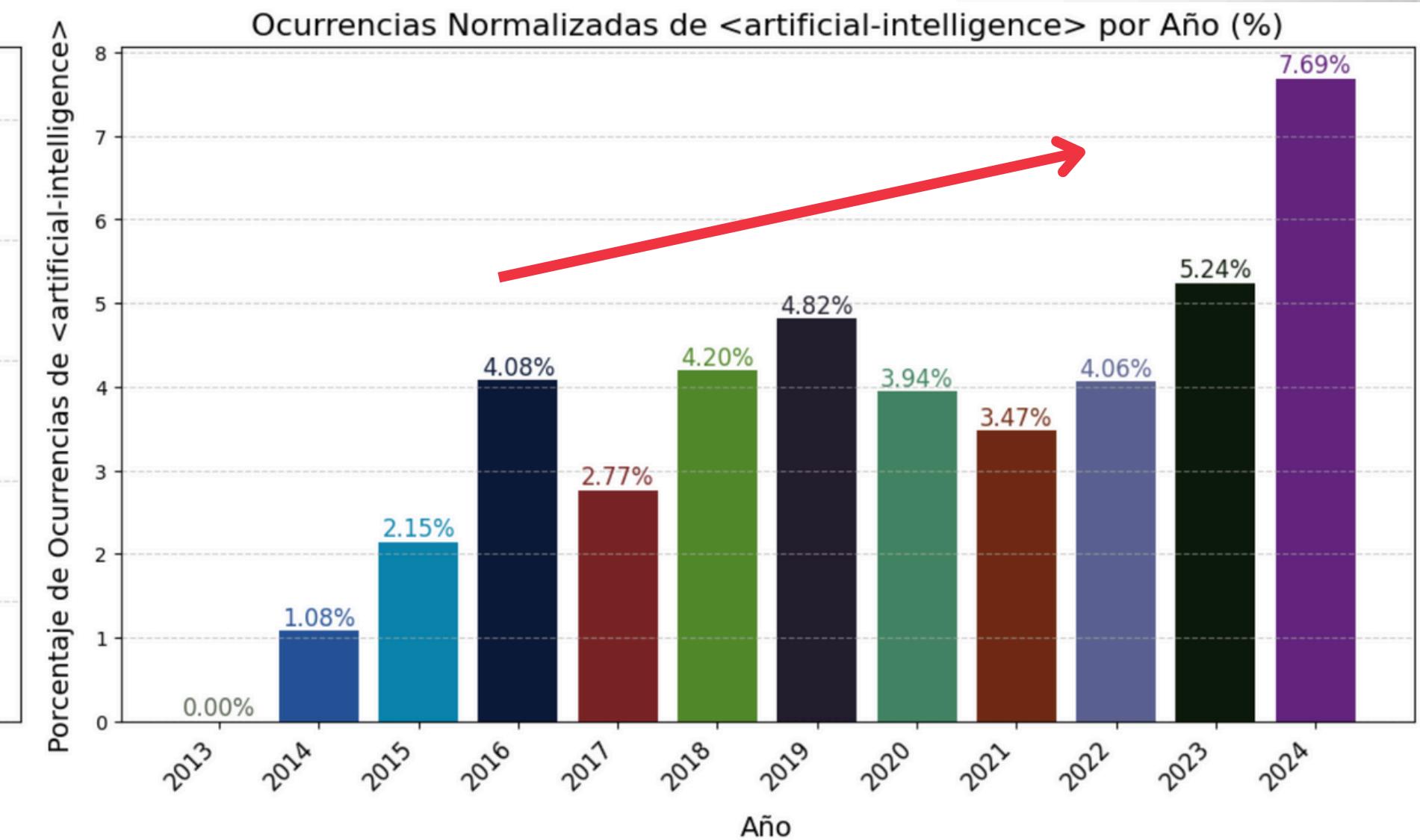
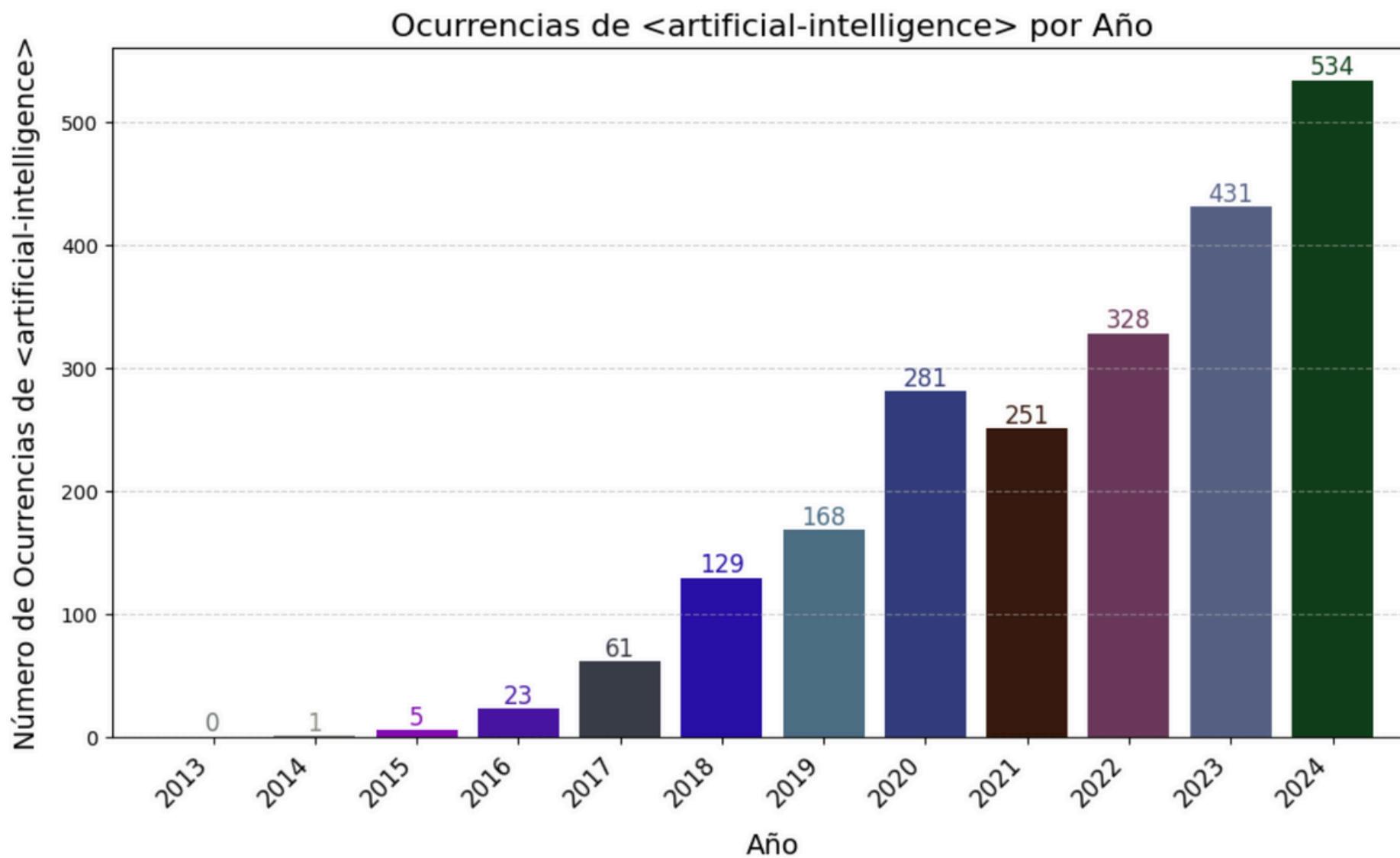
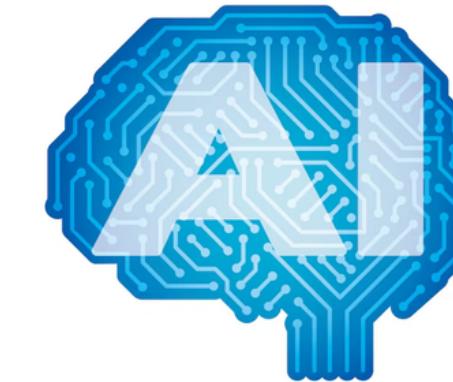
Visualización

Número de Repositorios por Año (2010-2024)



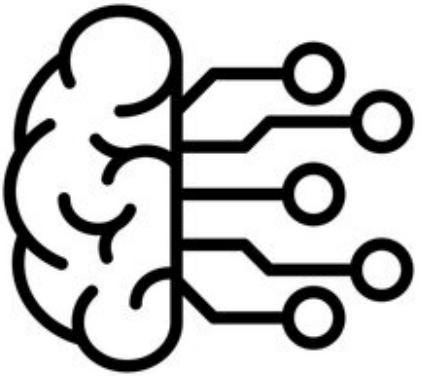
Visualización

→ Tema: Artificial Intelligence (AI)

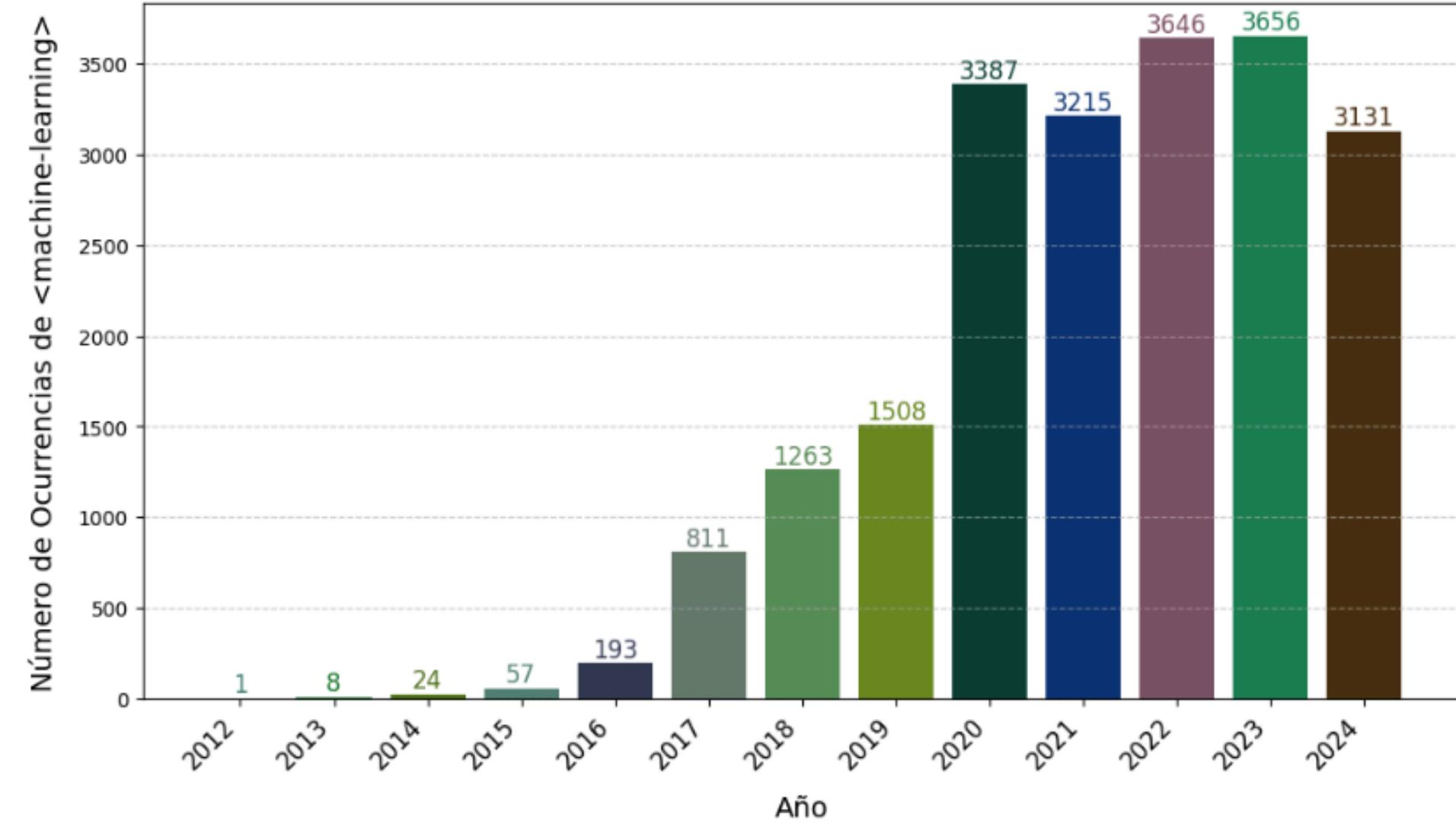


Visualización

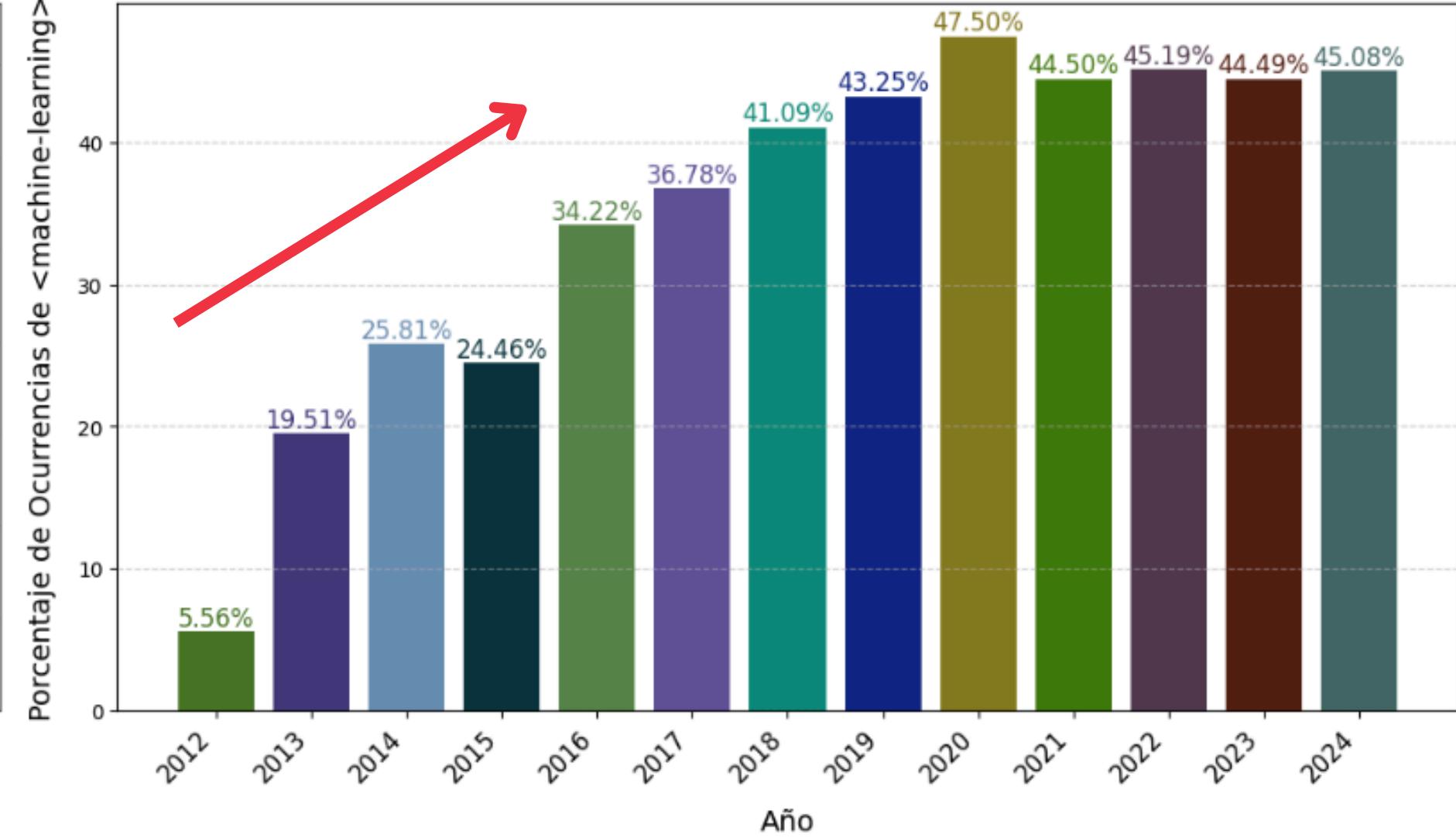
→ Tema: Machine-Learning (ML)



Ocurrencias de <machine-learning> por Año

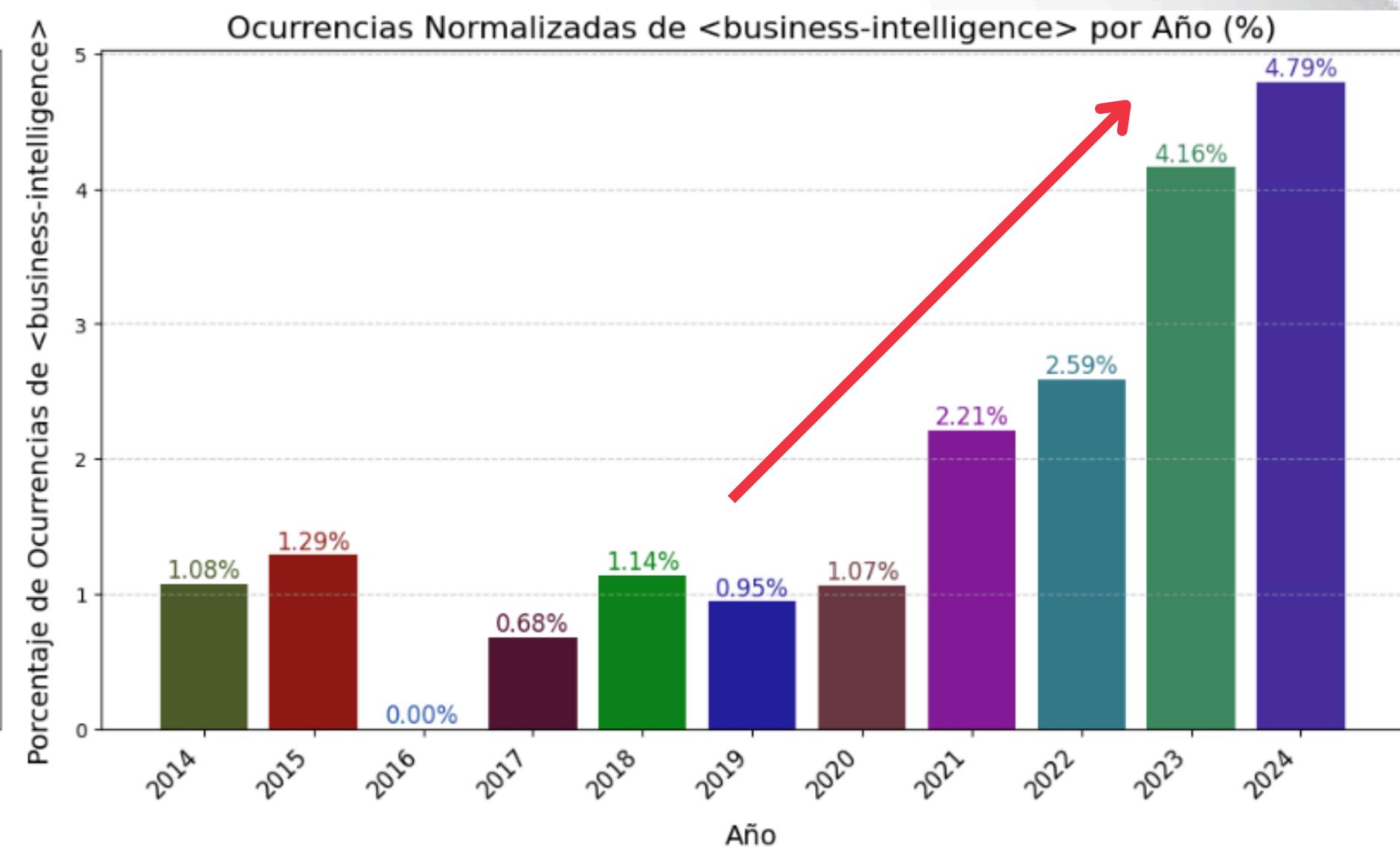
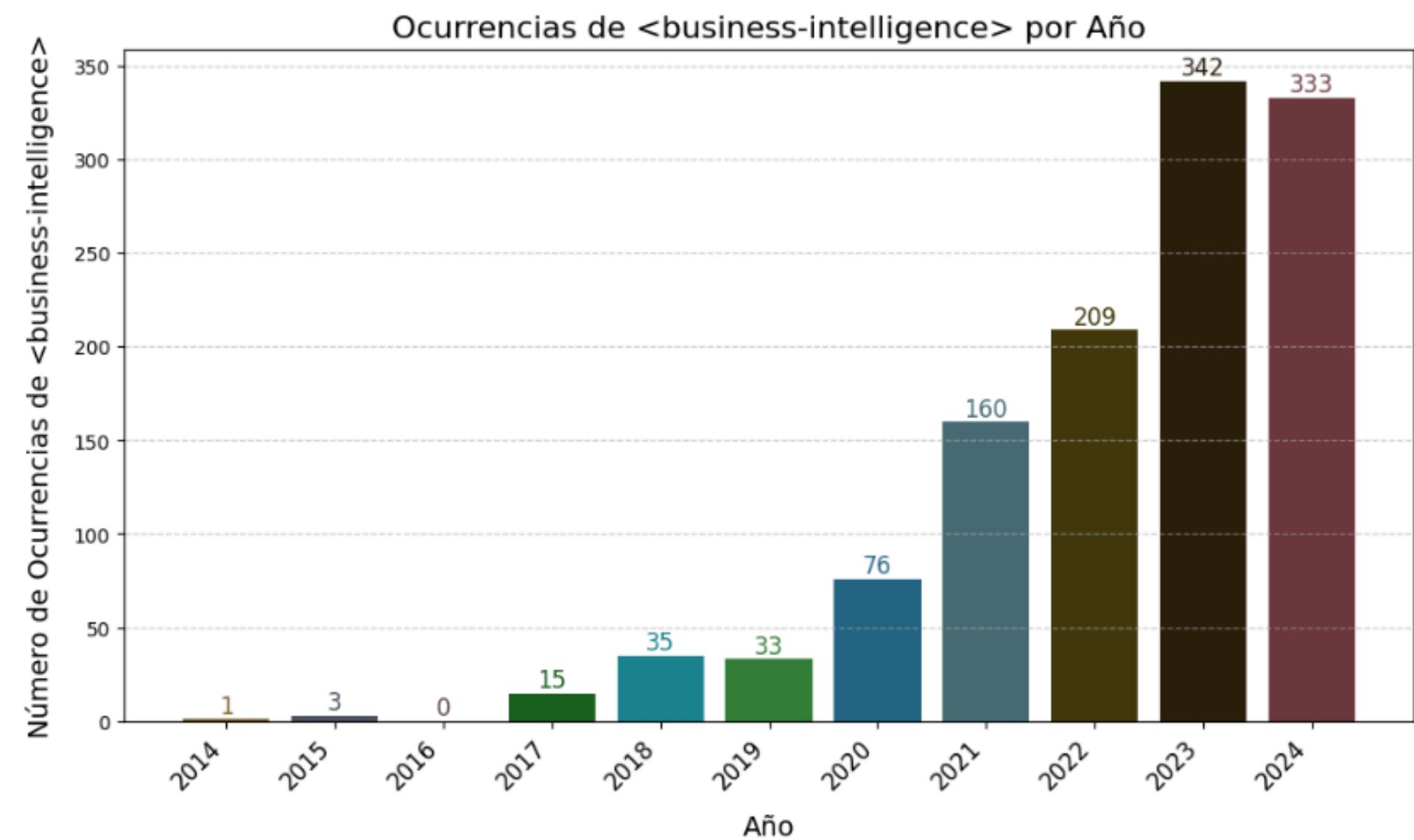


Ocurrencias Normalizadas de <machine-learning> por Año (%)



Visualización

→ Tema: Business Intelligence (BI)

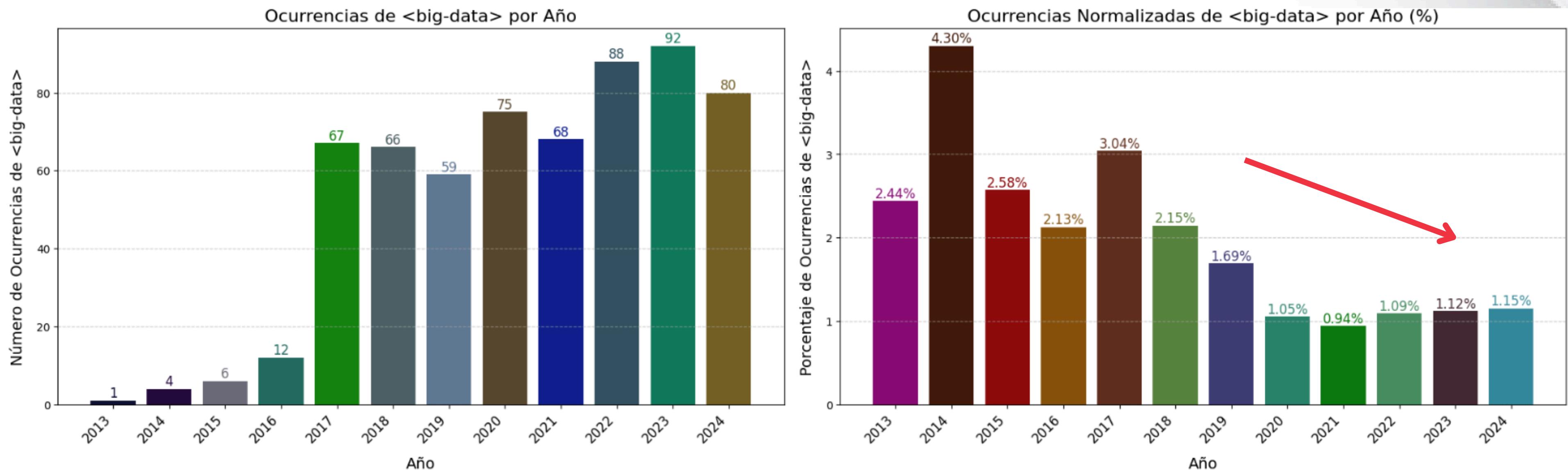


Visualización

→ Tema: Big-Data

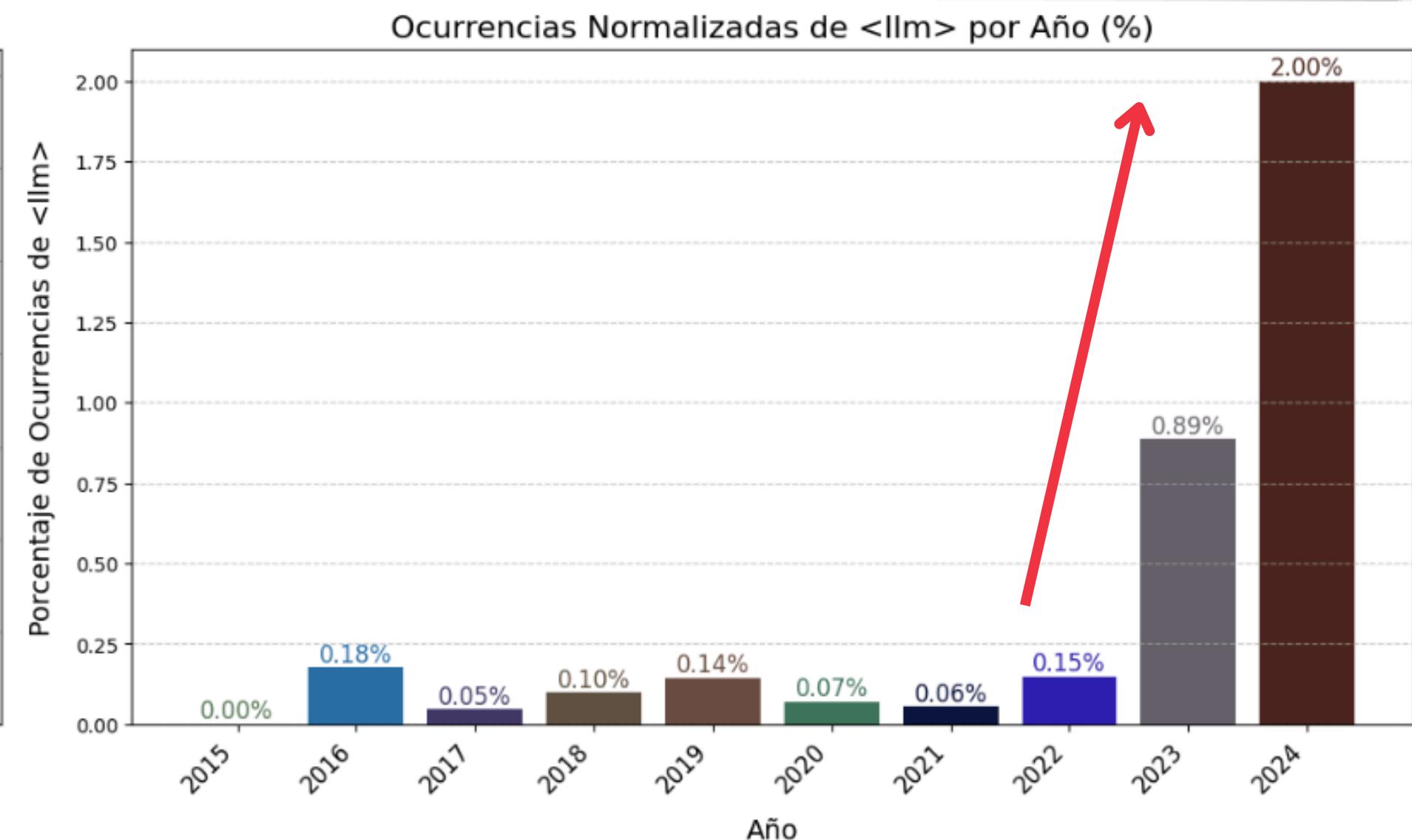
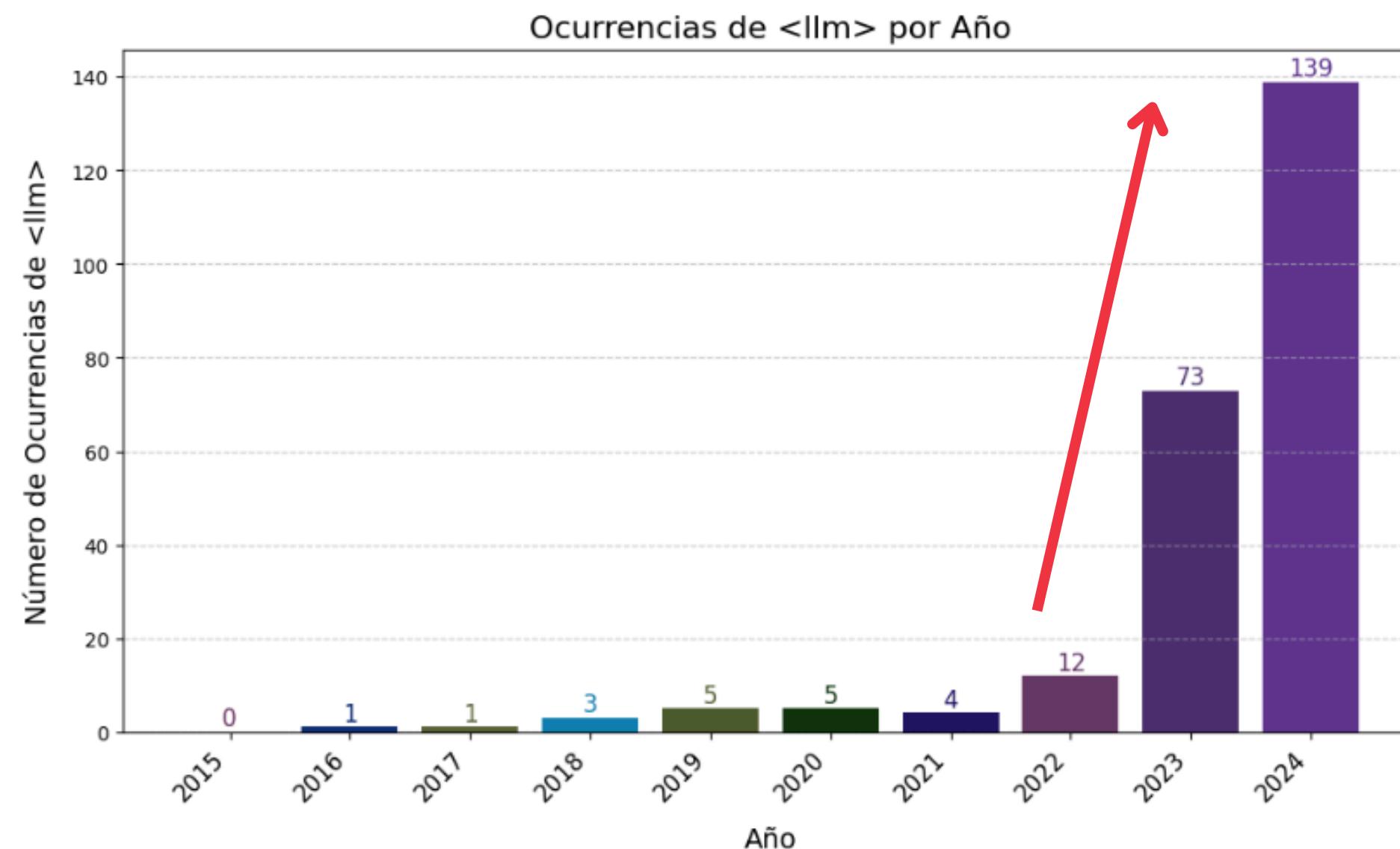


BIGDATA



Visualización

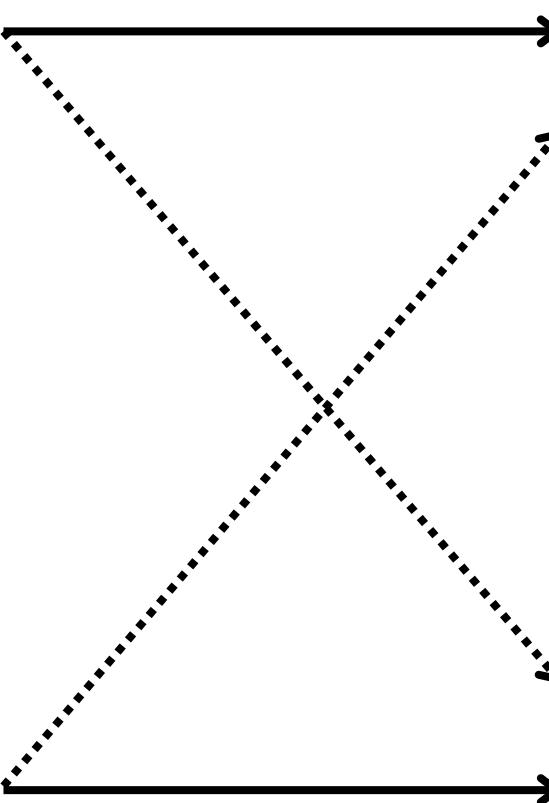
→ Tema: Large-Language Models (LLM)



Éxito de un repositorio



Número de Estrellas
(Stargazers)



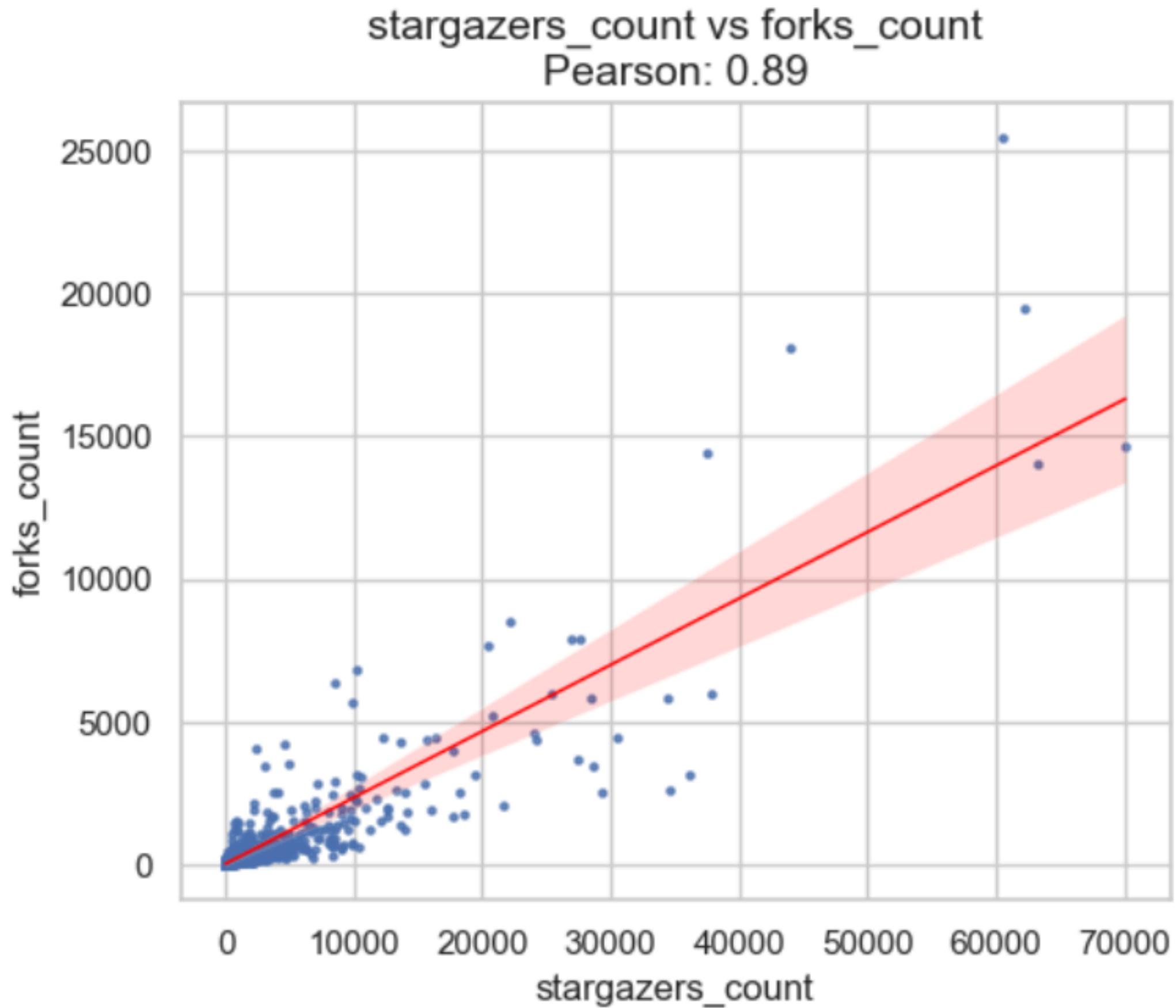
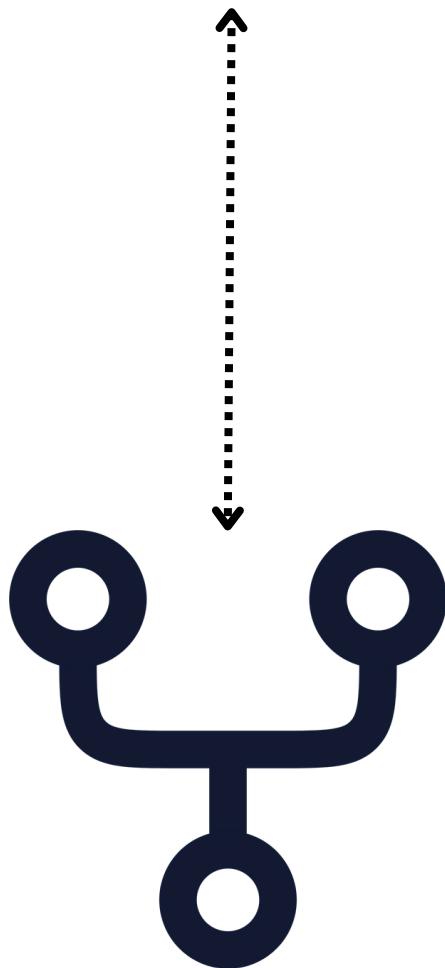
Número de Ramas
(Forks)

Popularidad

Compromiso
activo



Éxito de un repositorio



Éxito de un repositorio



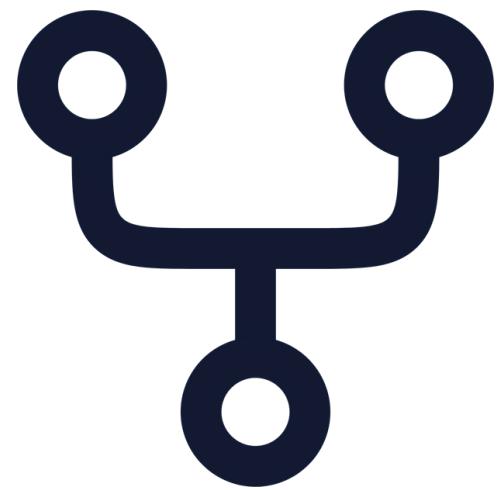
natural-language-processing
data-visualization
machine-learning
tensorflow
data-mining
data
pandas
python3
jupyter-notebookai

deep-learning
r
pytorch
mlops
statistics
artificial-intelligence
data-analysis
nlp

A word cloud centered around the word "machine-learning". The words are colored in various shades of purple, green, blue, and yellow. The size of each word represents its frequency or importance. A large white rectangular box with a thin black border is located in the upper right area, containing a red diagonal bar at the top right corner.



Éxito de un repositorio

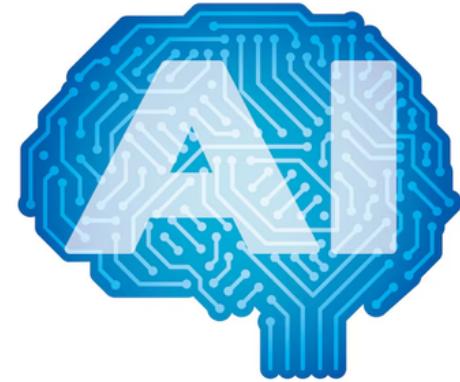


jupyter-notebook
machine-learning-algorithms
numpy
statistics
artificial-intelligence
r natural-language-processing
data
python data-analysis
ai deep-learning python3
machine-learning
computer-vision data-visualization

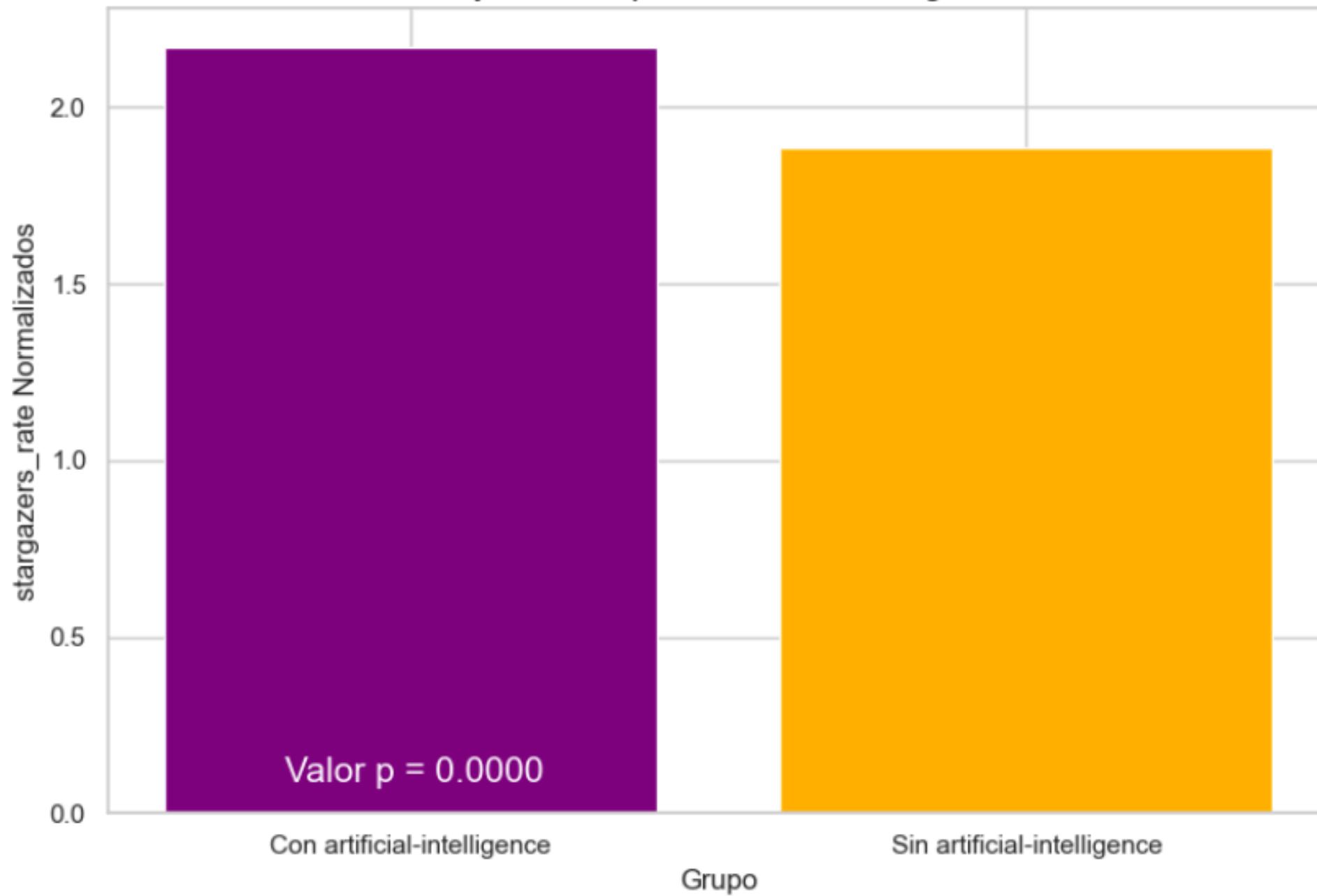


Visualización

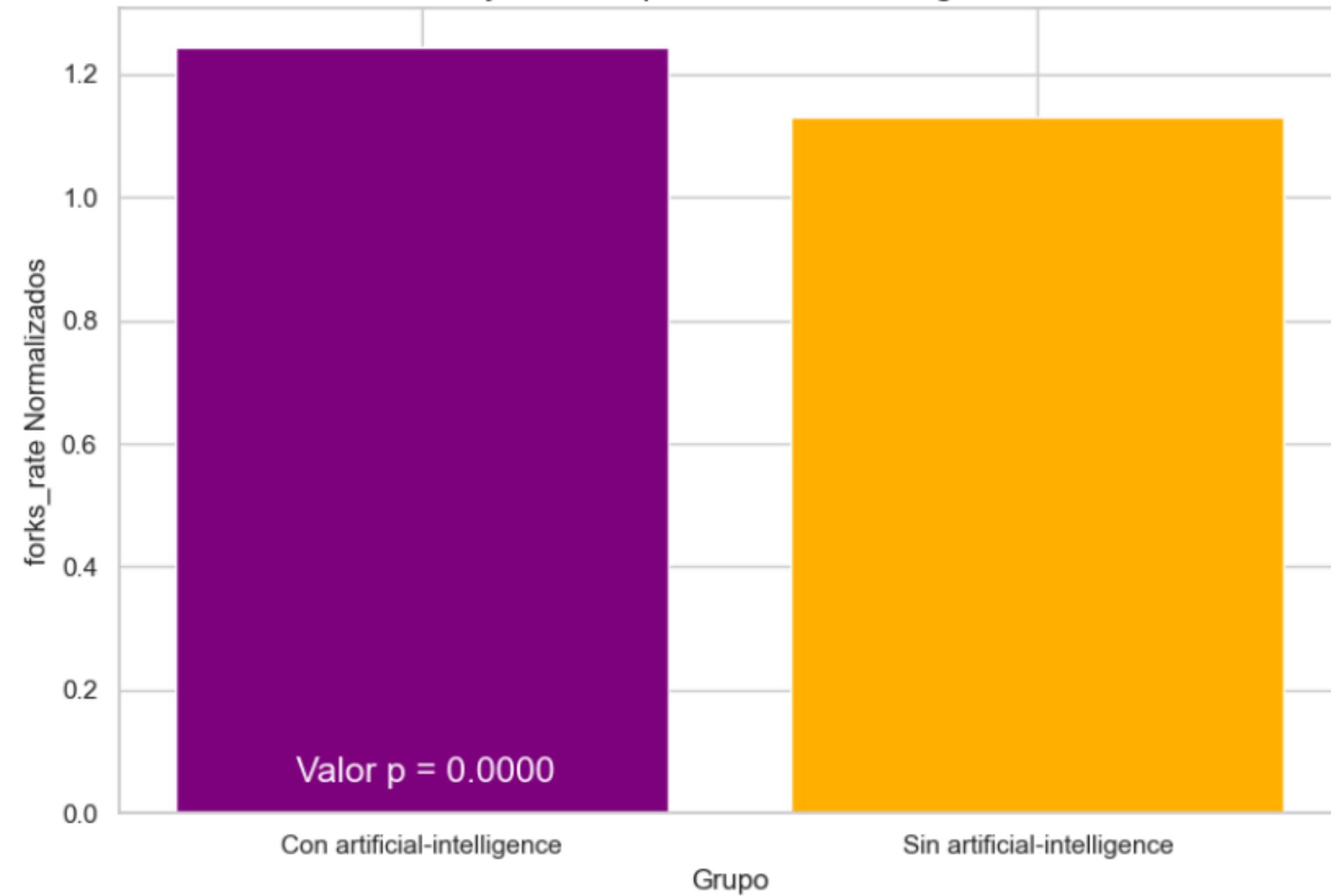
→ Tema: Artificial Intelligence (AI)



Comparación de stargazers_rate Normalizados
con y sin la etiqueta "artificial-intelligence"

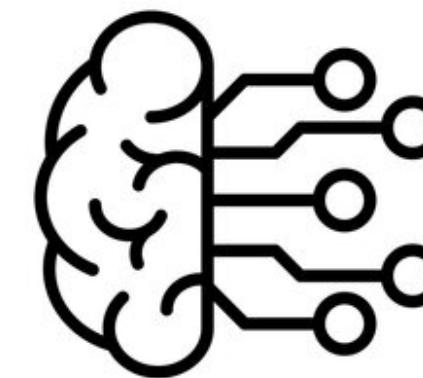


Comparación de forks_rate Normalizados
con y sin la etiqueta "artificial-intelligence"

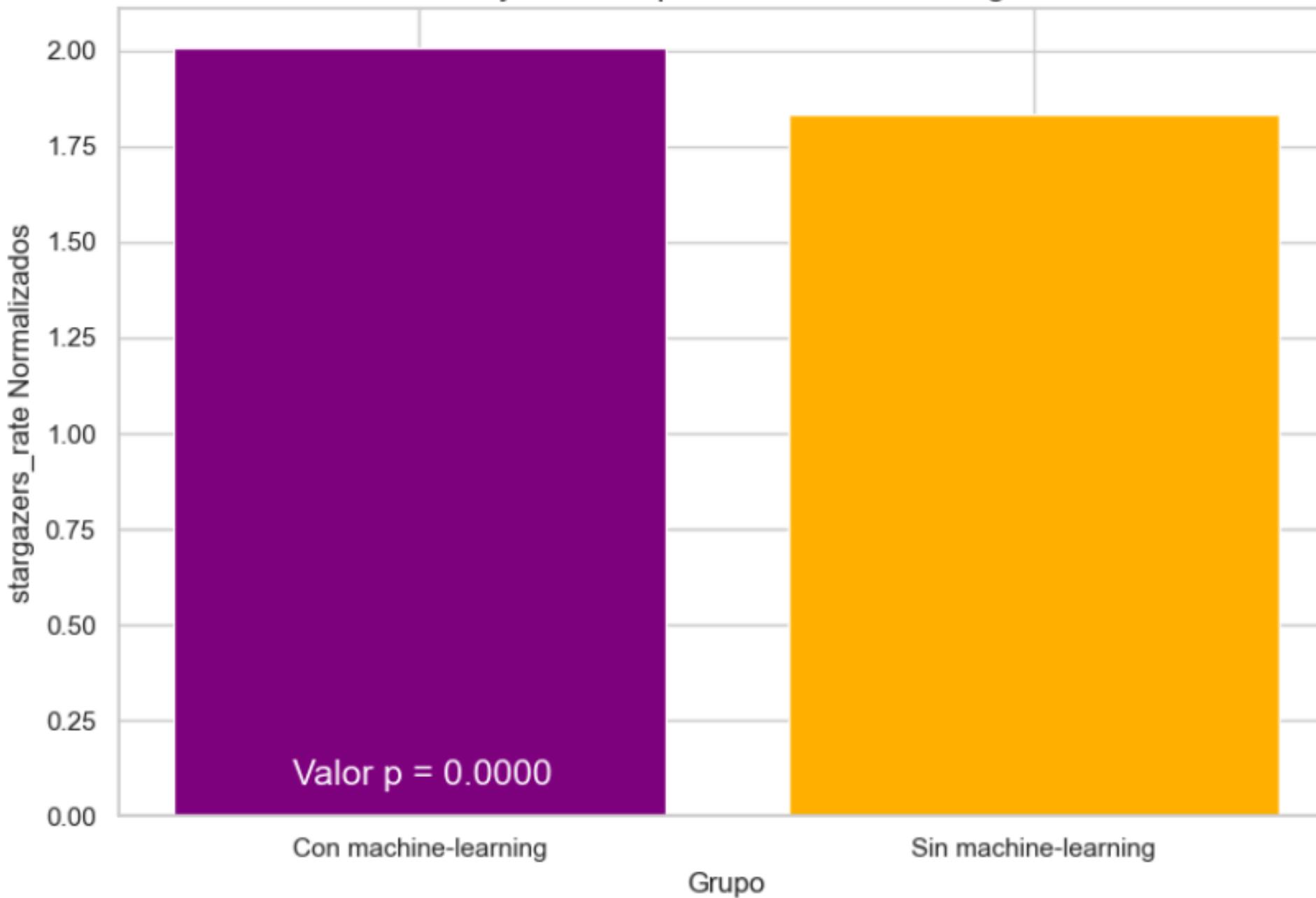


Visualización

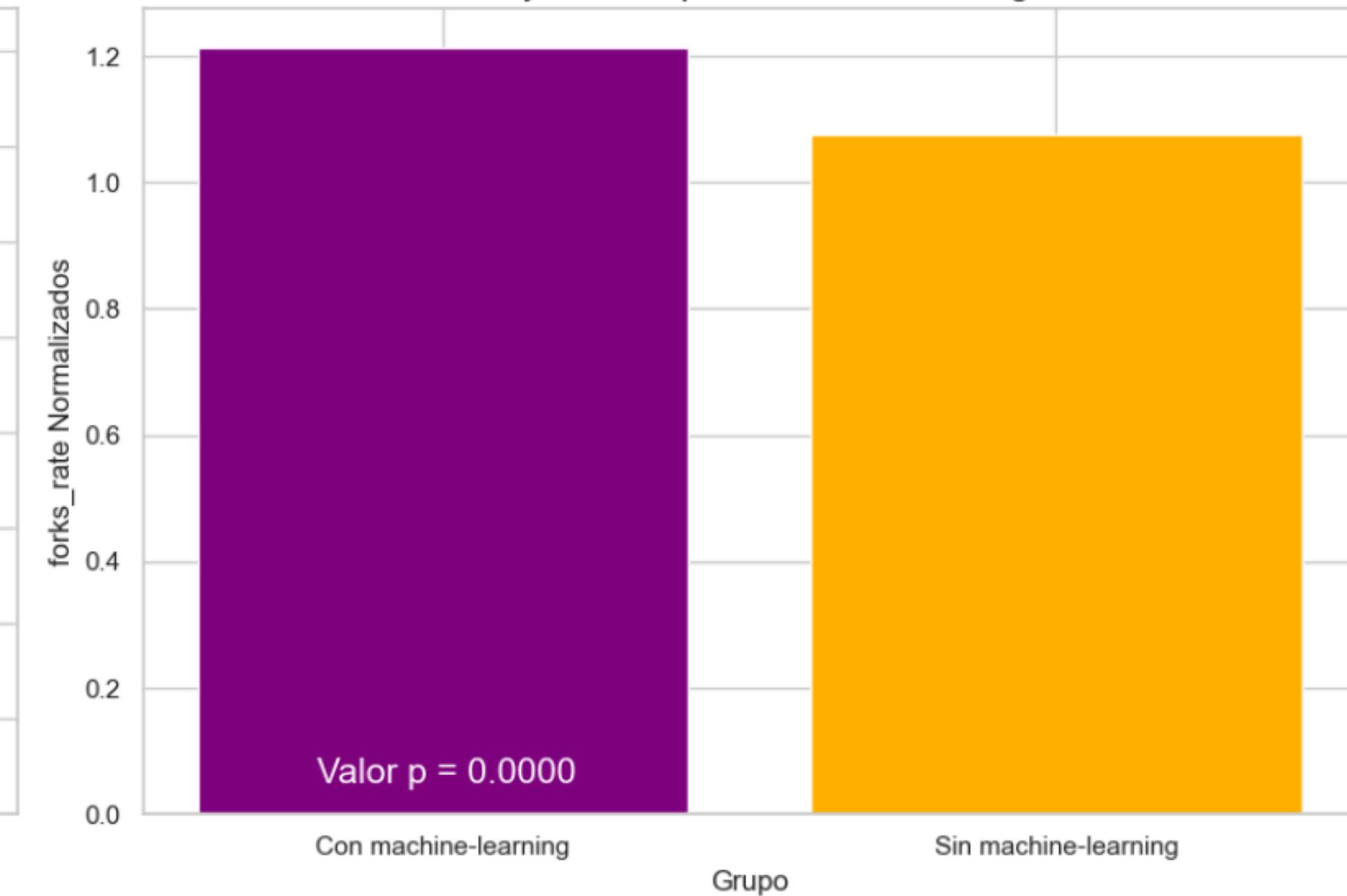
→ Tema: Machine-Learning (ML)



Comparación de stargazers_rate Normalizados
con y sin la etiqueta "machine-learning"



Comparación de forks_rate Normalizados
con y sin la etiqueta "machine-learning"

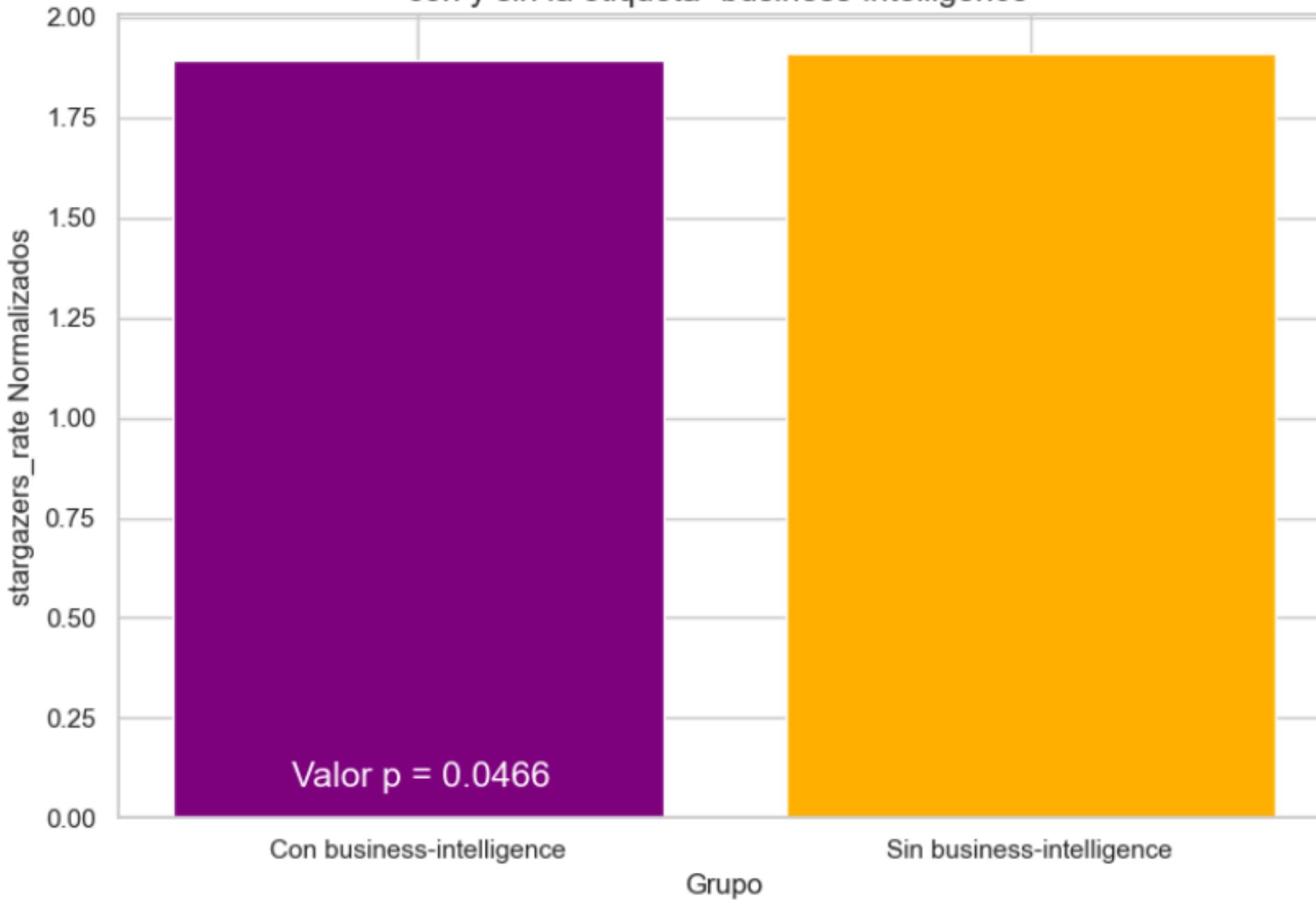


Visualización

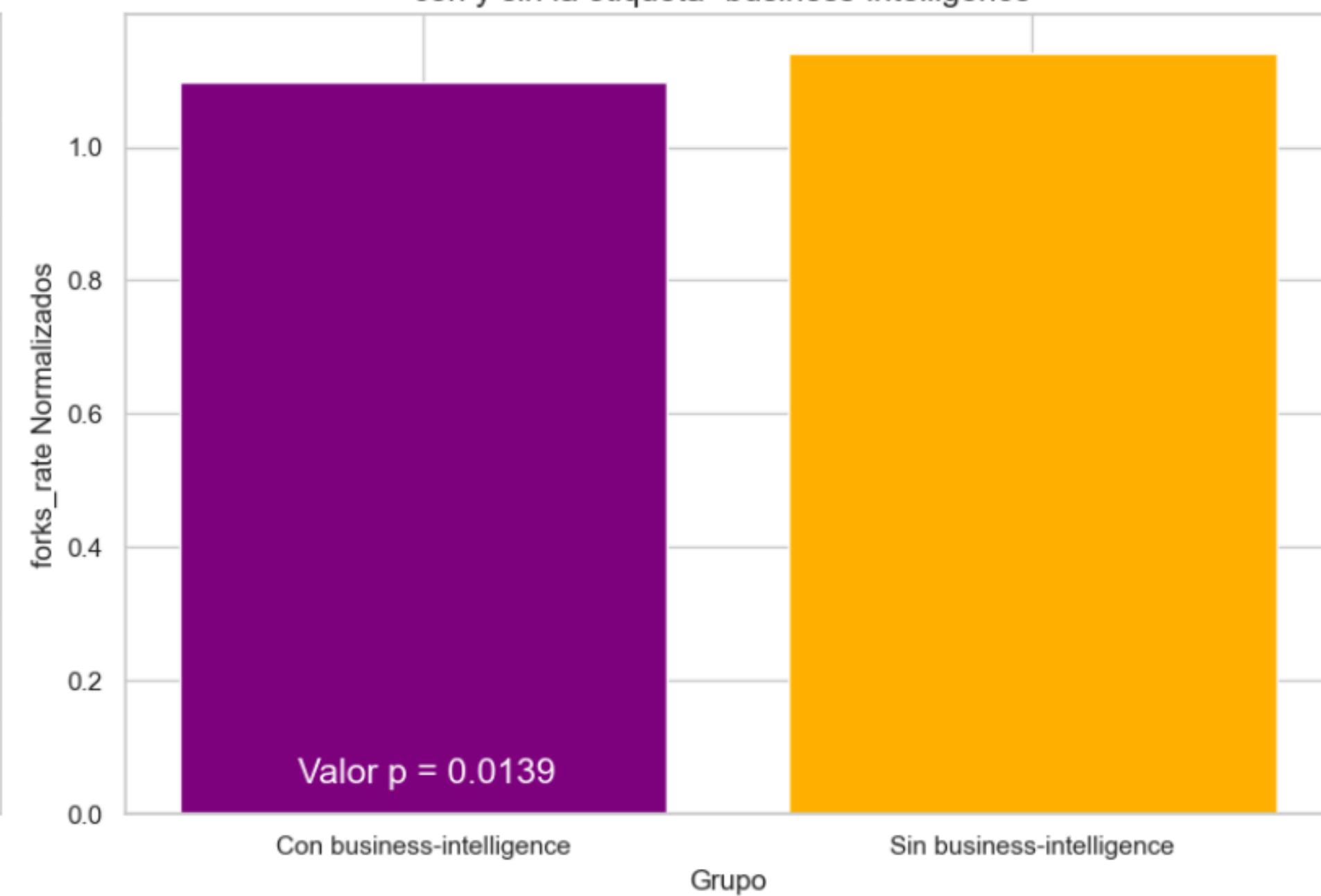
→ Tema: Business Intelligence (BI)



Comparación de stargazers_rate Normalizados
con y sin la etiqueta "business-intelligence"



Comparación de forks_rate Normalizados
con y sin la etiqueta "business-intelligence"

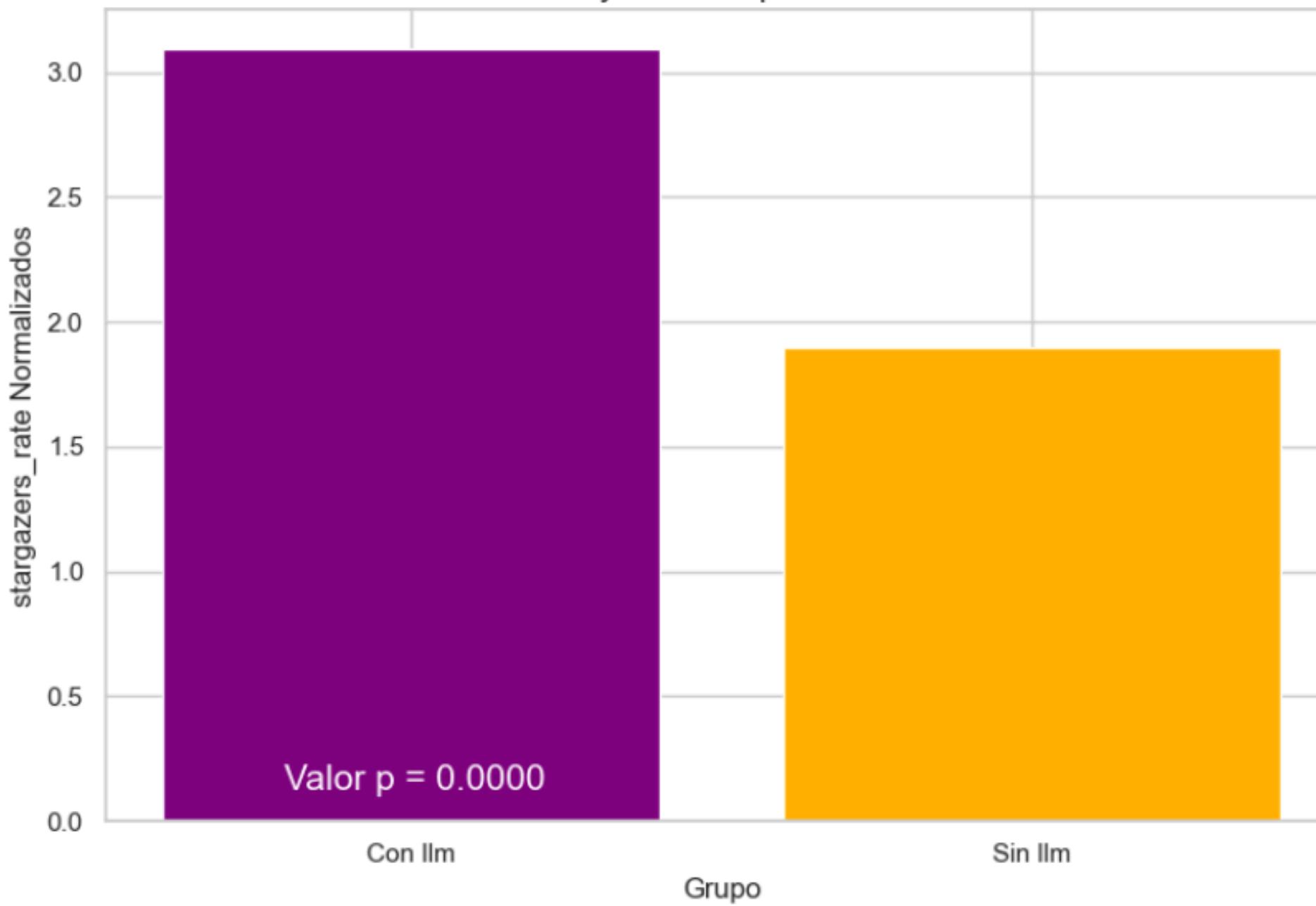


Visualización

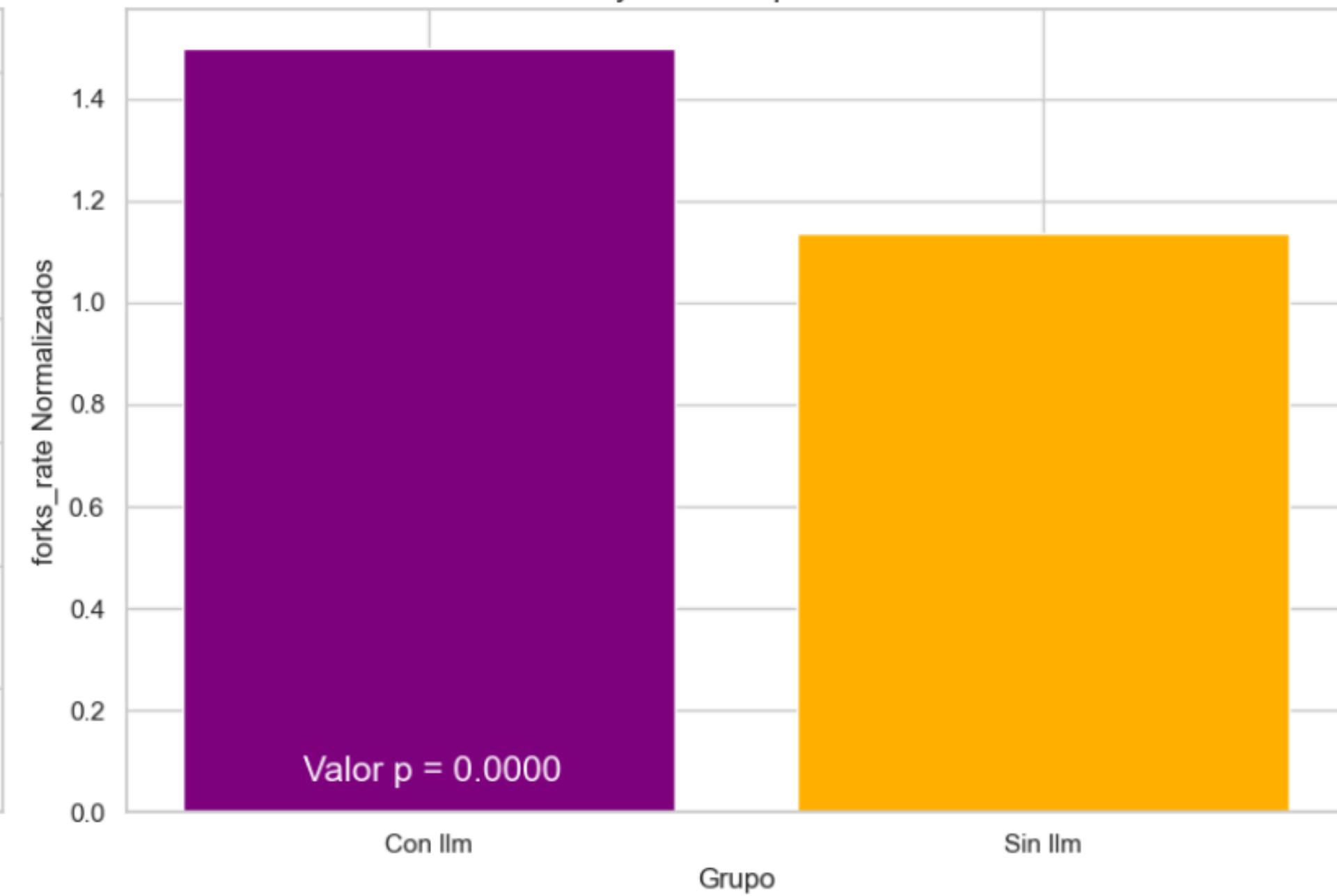
→ Tema: Large-Language Models (LLM)



Comparación de stargazers_rate Normalizados
con y sin la etiqueta "llm"



Comparación de forks_rate Normalizados
con y sin la etiqueta "llm"



Visualización

→ Tema: Deploy

Comparación de stargazers_rate Normalizados
con y sin la etiqueta "deploy"



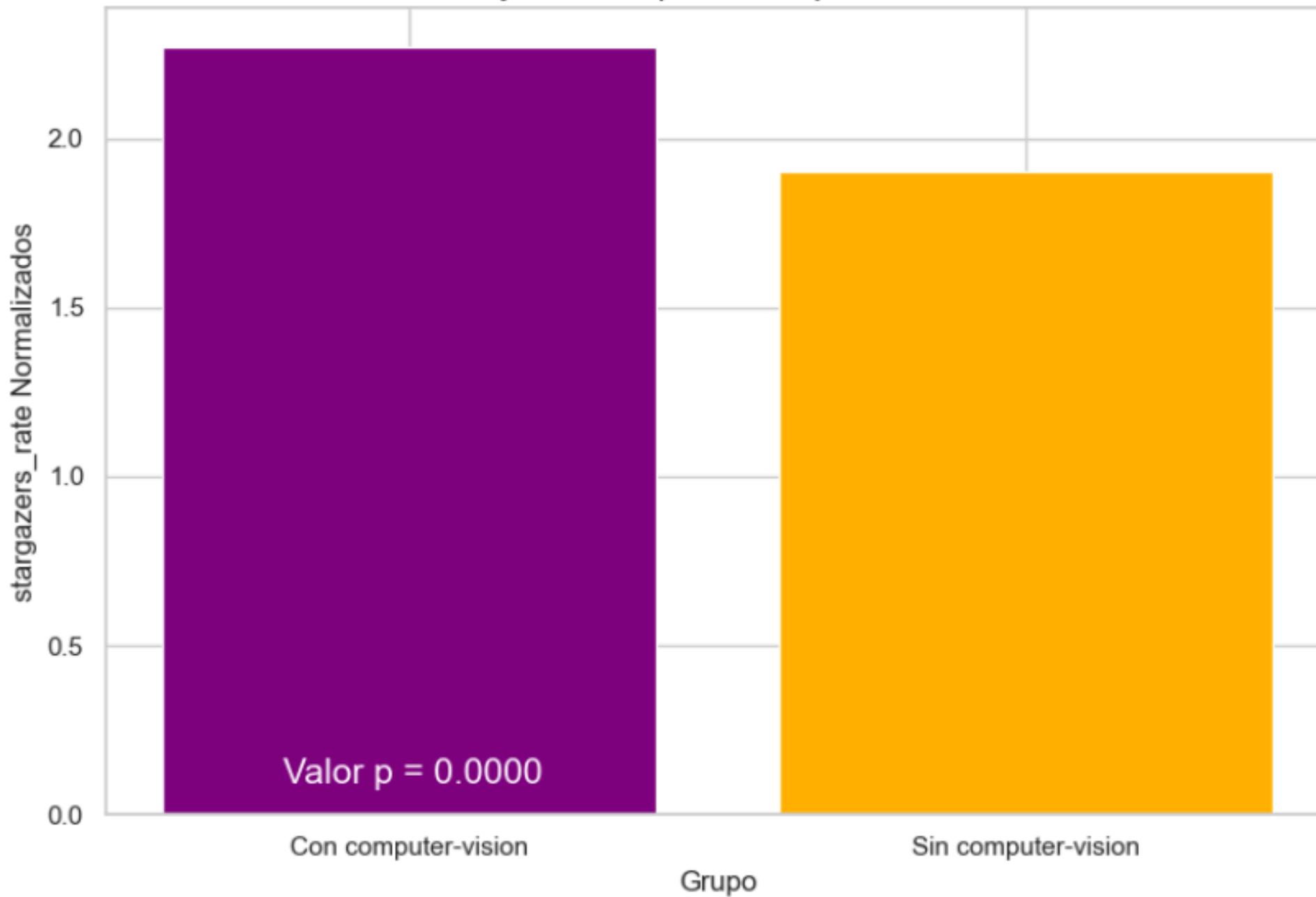
Comparación de forks_rate Normalizados
con y sin la etiqueta "deploy"



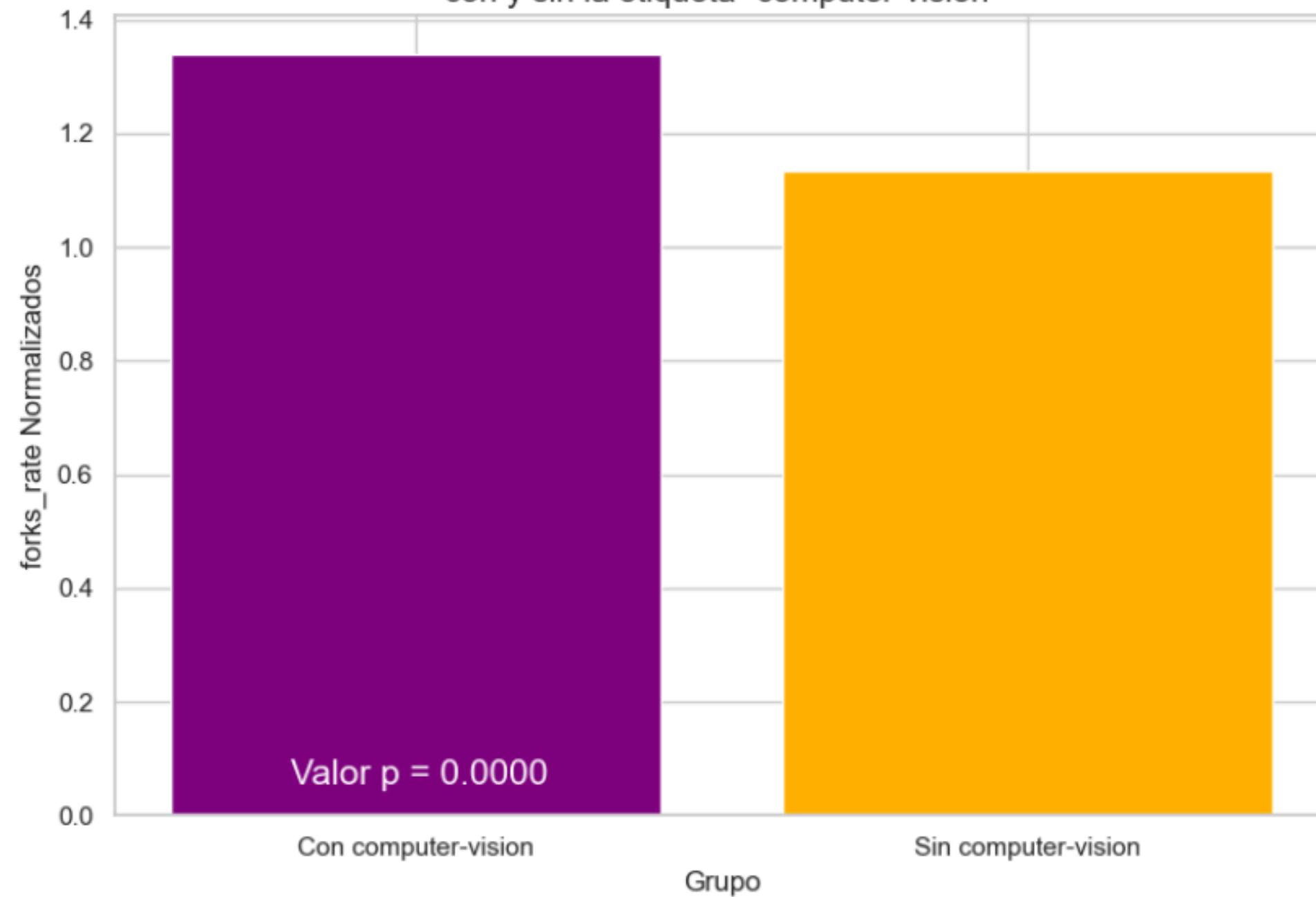
Visualización

→ Tema: Computer-Vision

Comparación de stargazers_rate Normalizados
con y sin la etiqueta "computer-vision"



Comparación de forks_rate Normalizados
con y sin la etiqueta "computer-vision"

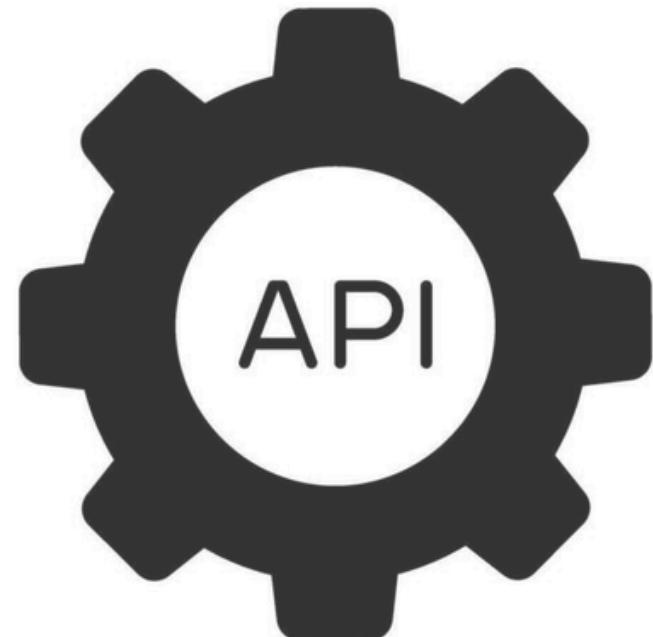


Conclusiones

- Determinados temas están ganando relevancia en el mercado de datos
 - cantidad
 - repercusión
- Temas destacados: LLMs, despliegue (deploy) y Visión por computadora (Computer Vision)
- Otros temas relevantes: Generative AI, Deep Learning y Procesamiento de Lenguaje Natural (NLP)

Project_Break_I_EDA

¡Muchísimas Gracias!



Lead Instructor: Rodrigo Oliver

Teacher Assistants: Ángela López Molina
Luismi Miguel Andújar
Iván Cordero

