Rubén Fernández Lorenzo

Data Science

Eda - Inversión en ia

The Bridge | Digital Talent Accelerator

Contenido

[Objetivo 2](#_Toc173358186)

[Origen de los datos 2](#_Toc173358187)

[Análisis 2](#_Toc173358188)

[Bibliografía 4](#_Toc173358189)

Memoria EDA

# Objetivo

En este documento se recogerá toda la información con la que se ha trabajado a lo largo del EDA y las dificultades obtenidas y las principales conclusiones a las que se han llegado.

Al final del documento, se muestra un apartado de bibliografía donde se hace referencia a toda la labor de investigación que se ha realizado durante el desarrollo del trabajo.

# Origen de los datos

Los datasets utilizados en el proyecto han sido extraídos de Kaggle, y son los siguientes:

**AI Computation & Hardware Trends**: <https://www.kaggle.com/datasets/programmerrdai/ai-computation-and-hardware-trends>

**AI Global Index:** <https://www.kaggle.com/datasets/katerynameleshenko/ai-index>

# Análisis

Viendo los datos ofrecidos por los DataSets que he escogido, he desarrollado la hipótesis nula: No sale rentable invertir en IA en la actualidad.

A partir de ahí, como se ha pensado en ser una presentación para negocio, he decidido entrar en el menor número de detalles técnicos, es por ello que en la presentación lo único que se ve del código que he desarrollado son algunas de las gráficas resultado.

La presentación se ha planteado para ser llamativa y tener enganche desde un punto de vista cercano que genere comodidad y confianza.

Principalmente se introducen dos términos principales para poder desarrollar la presentación: MMLU y capacidad de computación (petaFLOPS).

En primer lugar, se ha analizado el *benchmark* **MMLU**, el cual es estándar que se utiliza para evaluar la comprensión de lenguaje de modelos de IA. Este nos permite comparar la capacidad de diferentes modelos de lenguaje en tareas que requieren una comprensión profunda y multifacética del lenguaje humano.

MMLU evalúa modelos en una variedad de tareas, desde preguntas de opción múltiple hasta análisis de texto complejo, lo que lo convierte en un indicador integral del rendimiento de la IA en la comprensión del lenguaje. Esta evaluación multifacética es esencial para desarrollar modelos de IA que puedan entender y procesar información de manera similar a los seres humanos.

Posteriormente se ha evaluado la cantidad de potencia computacional utilizada para entrenar el modelo de IA, medida en petaFLOPs, a lo largo de los años. En esta fase, se ha analizado un gráfico donde se puede observar a simple vista como han ido creciendo la capacidad computacional a lo largo de los años, notando un incremento muy considerable en el año 2023.

Las siguientes partes de la presentación, he planteado comparar la capacidad de computación con la máquina con mayor potencia de computación en la actualidad (Frontier) e investigando la inversión que se ha realizado he mostrado la cantidad de máquinas similares y el costo que han tenido en los años la inversión de la IA.

Finalmente, la última parte de la presentación ha sido la parte de mayor labor de investigación, informándome y obteniendo artículos y gráficas de diferentes artículos fiables para poder adjuntarlas a la presentación, y concluir con el rechazo de la hipótesis nula.

En el código también aparece la importación de dos DataSets los cuales no se han empleado para la presentación ya que o no era un tema que a negocio interese (producción de chipsets) o se alejaba del punto de vista principal planteado originalmente (empresas centradas en IA). A pesar de ello, son datasets, que se podrían utilizar para una mayor labor de investigación si se quiere analizar a una posible competencia directa (con el DataSet de las empresas que se dedican a la IA) o empezar a adquirir productos para computar (producción de chipsets).

# Principales dificultades

Las principales dificultades con las que se ha trabajado han sido la inconsistencia o datos nulos del dataset que se ha seleccionado, ya que había columnas que aparecían con una gran cantidad de nulos, y por tanto no se han podido considerar.

Había columnas, que el tipo del dato no era un valor cuantificable, y al hacer el ‘top 5’ de algo, no podía ordenarse de ningún modo. Esto se ha solucionado cambiando el tipo de la columna a entero, y entonces poder comparar cuantitativamente los datos.

Finalmente, había columnas que en sus datos aparecían caracteres especiales como ‘$’ o ‘<’ un hecho que no ayudaba a la hora de clasificar los datos. Entonces tras iterar por todas las ocurrencias de la columna, se ha realizado un *“replace”* para borrar todos los caracteres especiales que impedían trabajar con estos datos.

# Bibliografía

Información de MMLU: <https://www.victormolla.com/mmlu>

Noticia ingresos de Google: <https://www.businessinsider.es/resultados-google-q4-2023-record-ingresos-beneficios-trimestrales-1361079>

2021 año de los modelos de IA: <https://www.technologyreview.es/s/13901/2021-el-ano-de-los-modelos-de-inteligencia-artificial-gigantes>

Frontier: <https://www.xataka.com/ordenadores/frontier-nuevo-supercomputador-potente-mundo-ha-alcanzado-hito-romper-barrera-exaescala#:~:text=Fast%20food-,Frontier%2C%20el%20nuevo%20supercomputador%20m%C3%A1s%20potente%20del%20mundo%2C%20ha%20alcanzado,la%20barrera%20de%20la%20exaescala&text=M%C3%A1s%20de%201.000.000.000.000.000,total%20de%201%2C1%20exaFLOPS>

Gráficos económicos: <https://es.investing.com/>