# Automatización de informes en Python con Inteligencia artificial

**Víctor Hugo Muñoz Villa**

Epidemiólogo – Contratista – Equipo de Vigilancia en Salud Pública –
Secretaría Departamental de Salud del Valle del Cauca

THE GLOBAL HEALTH NETWORK
LATIN AMERICA AND THE CARIBBEAN

Universidad del Valle

GOBERNACIÓN
Departamento del
Valle del Cauca

paraíso de todos

# Automatización de informes en Python con Inteligencia artificial

www.valledelcauca.gov.co/

# Contenido de la Sesión

1. Algunos conceptos de Inteligencia Artificial: **Víctor Muñoz**
2. Un estudio de caso: automatización de informes analíticos de EISP: **Víctor Muñoz**
3. Datos no estructurados: Procesamiento de Lenguaje Natural*: **Víctor Muñoz**

*También existen los datos semiestructurados

# 1. Algunos conceptos de Inteligencia Artificial

*También existen los datos semiestructurados

# Que hay detrás de los grandes modelos de lenguaje (LLM)?

**1. Entrenamiento con Datos a Escala Masiva**

Son alimentados con casi la totalidad de internet, libros y bases de datos (cientos de miles de millones de palabras). Este es el **combustible** que les da su vasto conocimiento del mundo, del lenguaje y de los patrones de razonamiento.

**2. La Arquitectura Transformer** 🧠

Es el **motor** revolucionario que les permite procesar toda esa información. Su mecanismo de **"atención"** les da la capacidad de entender el contexto, la gramática y las relaciones sutiles entre palabras, superando las limitaciones de arquitecturas anteriores. "Attention Is All You Need"

**3. Un Modelo de Miles de Millones de Parámetros**

El resultado del entrenamiento es un modelo con una escala sin precedentes. Estos parámetros son como las **conexiones neuronales** del modelo donde se almacena todo el "aprendizaje". Su inmenso número es lo que permite la complejidad y la flexibilidad en sus respuestas.

# El "Sí" de la IA: ¿Por Qué los Modelos de Lenguaje Parecen Poder con Todo?

## El Origen de la Confianza Artificial

•**Son "Máquinas de Predicción":** Su objetivo fundamental no es "saber", sino predecir la siguiente palabra más lógica y probable en una frase. Actúan como un autocompletar superavanzado.

•**Entrenados con el "Conocimiento Experto":** Aprenden de millones de textos (tutoriales, libros, foros) donde los expertos humanos *sí* resuelven los problemas. El patrón dominante es la solución, no la duda.

•**Programados para Agradar (RLHF):** Durante su ajuste fino, son recompensados por evaluadores humanos por ser útiles, serviciales y sonar seguros. Decir "No puedo" a una tarea alcanzable a menudo se penaliza.

•**Resultado:** El modelo aprende que la respuesta más "correcta" y recompensada es actuar como un asistente optimista y capaz, incluso si no comprende realmente sus propias limitaciones.

# Realidad Práctica:
# Simulación vs. Ejecución y la Regla de Oro

## Cómo Usar la IA de Forma Inteligente y Segura

**1. El Peligro de la Confianza (Simulación vs. Ejecución):** Un LLM no *ejecuta* tu análisis de datos, sino que *simula* el texto y el código que un experto humano crearía al hacerlo. No procesa los números, genera la apariencia de haberlo hecho.

**2. Riesgo de "Alucinación":** Al no tener conciencia real de sus límites, si no conoce la respuesta, la inventará. Puede generar conclusiones falsas, datos incorrectos o código con errores sutiles que parecen lógicos.

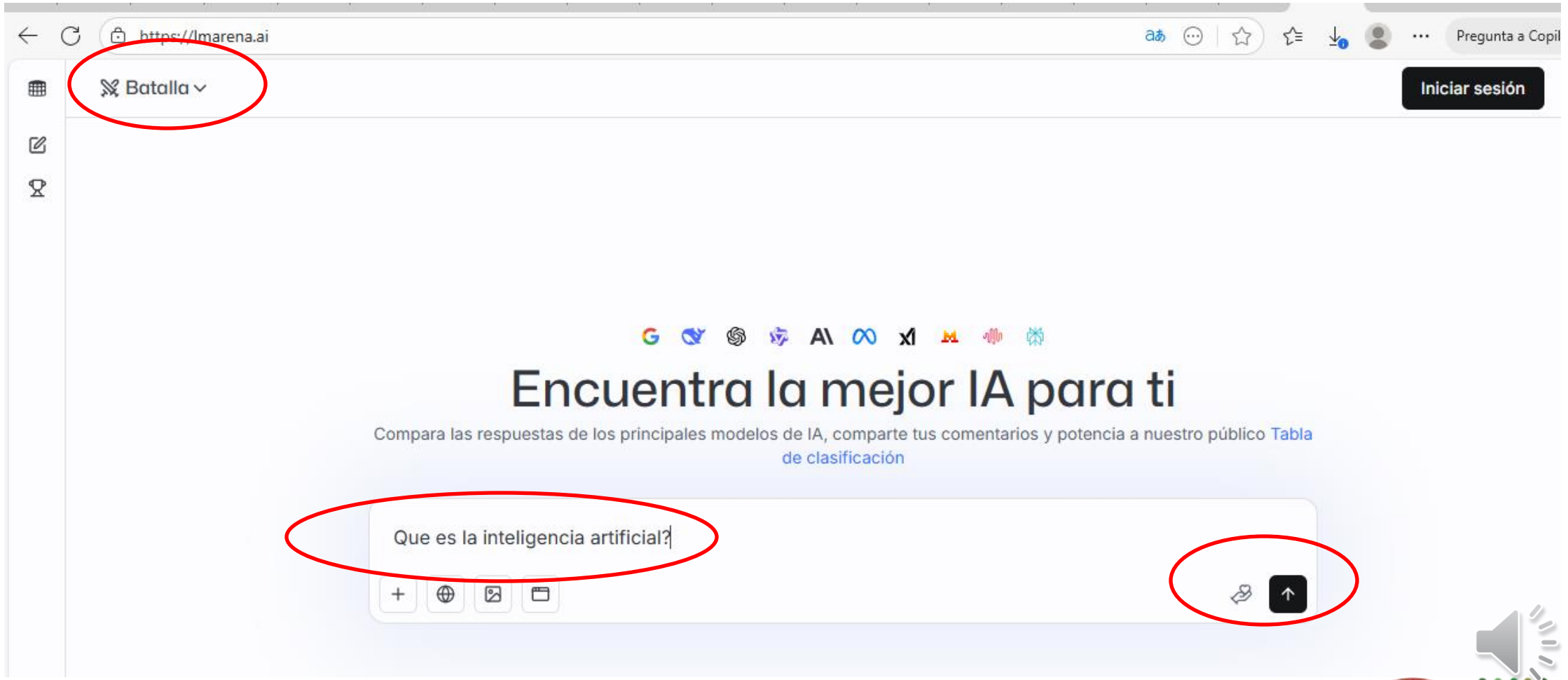
**3.LA REGLA DE ORO: Usa la IA como un Copiloto, no como el Piloto.**

•**Tu Rol es Crítico:** Pídele que genere el código, que estructure un informe o que te explique un método. Pero **SIEMPRE** debes ser tú quien **ejecute el código, verifique los datos y valide las conclusiones finales.**

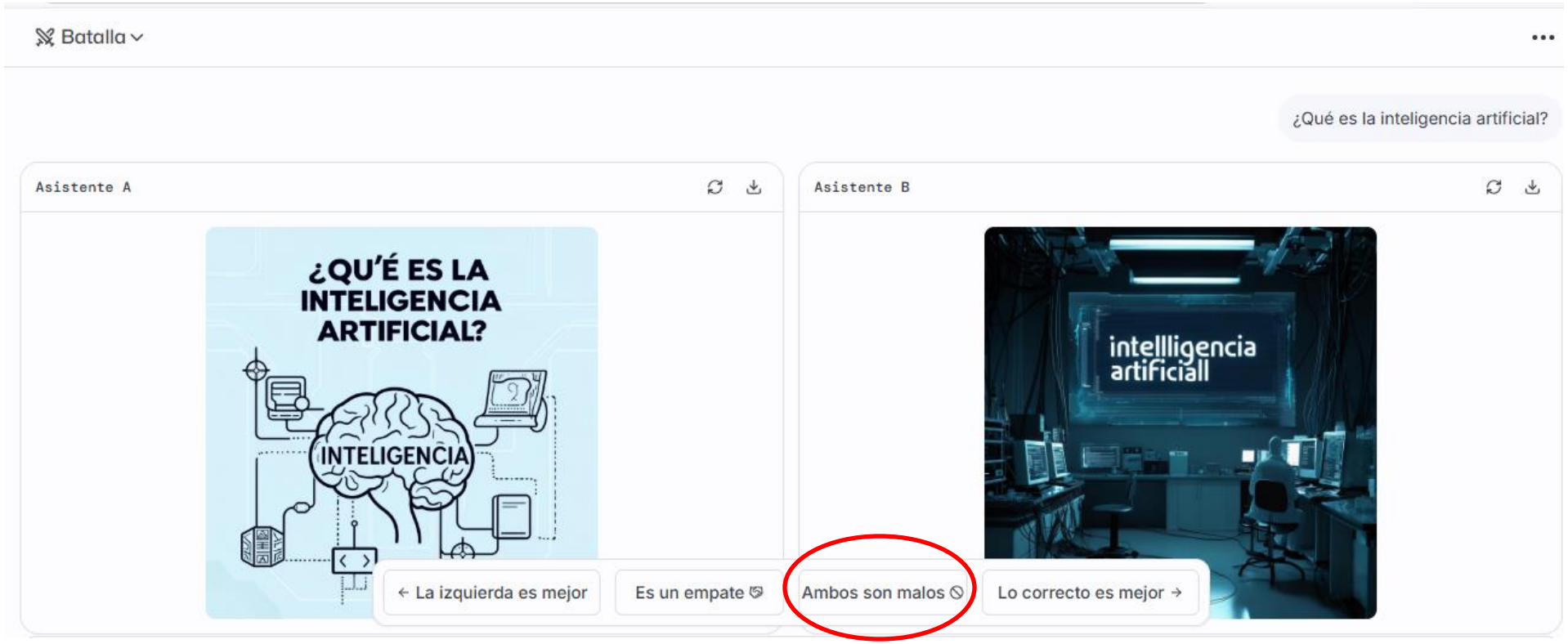
# Como escoger tu IA?

Link Arena: https://lmarena.ai/

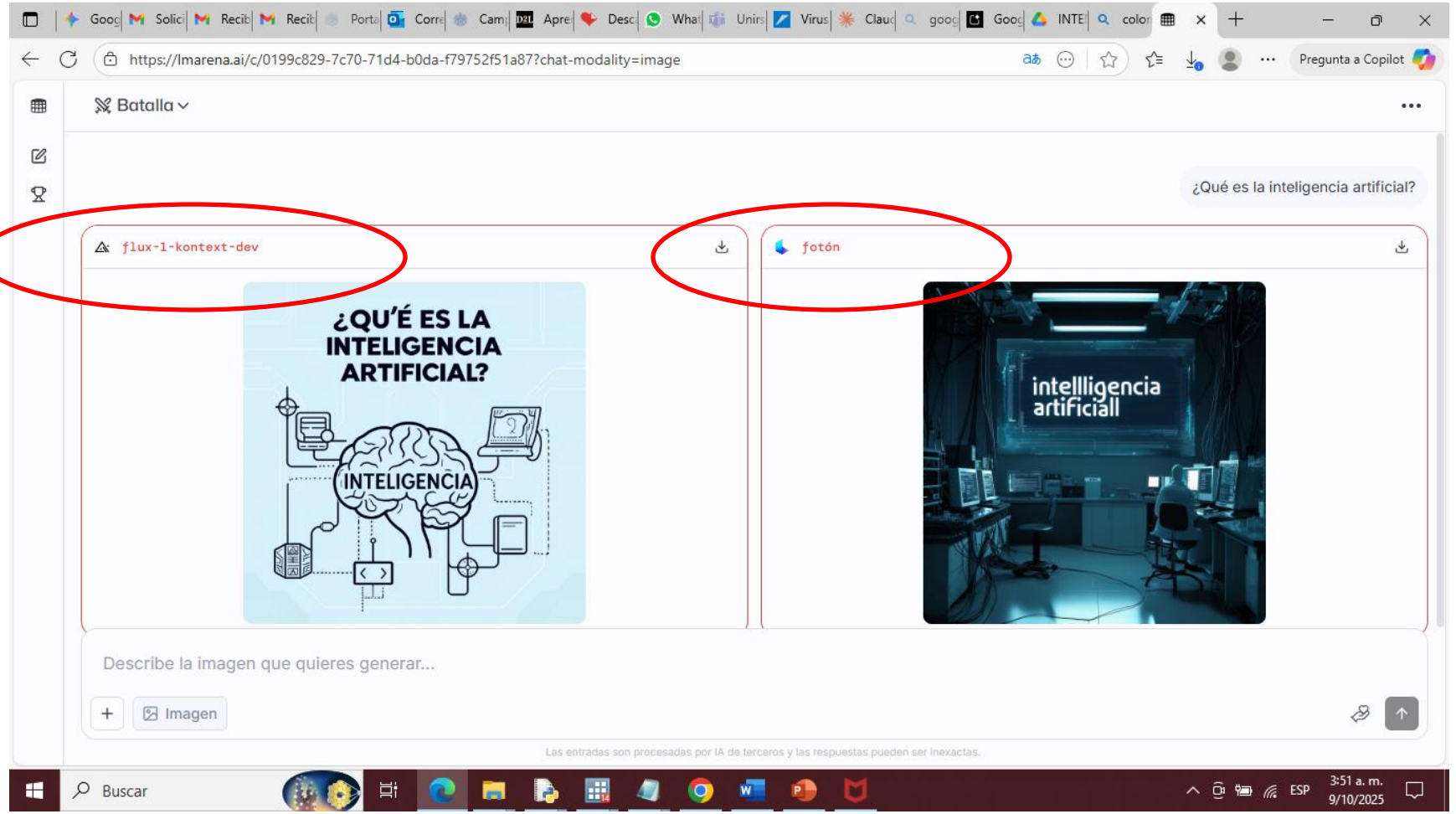
# Link Arena: https://lmarena.ai/



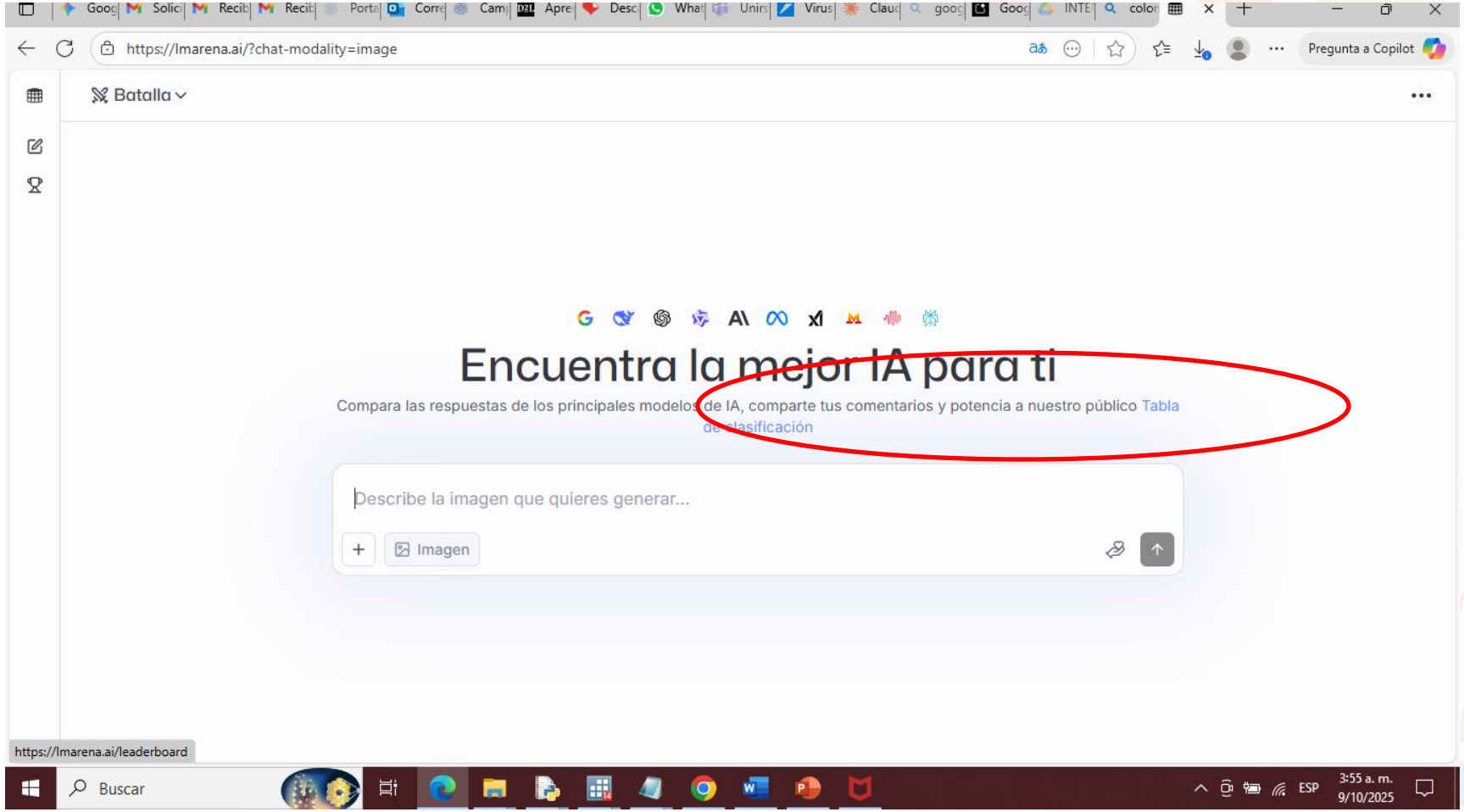
# Link Arena: https://lmarena.ai/



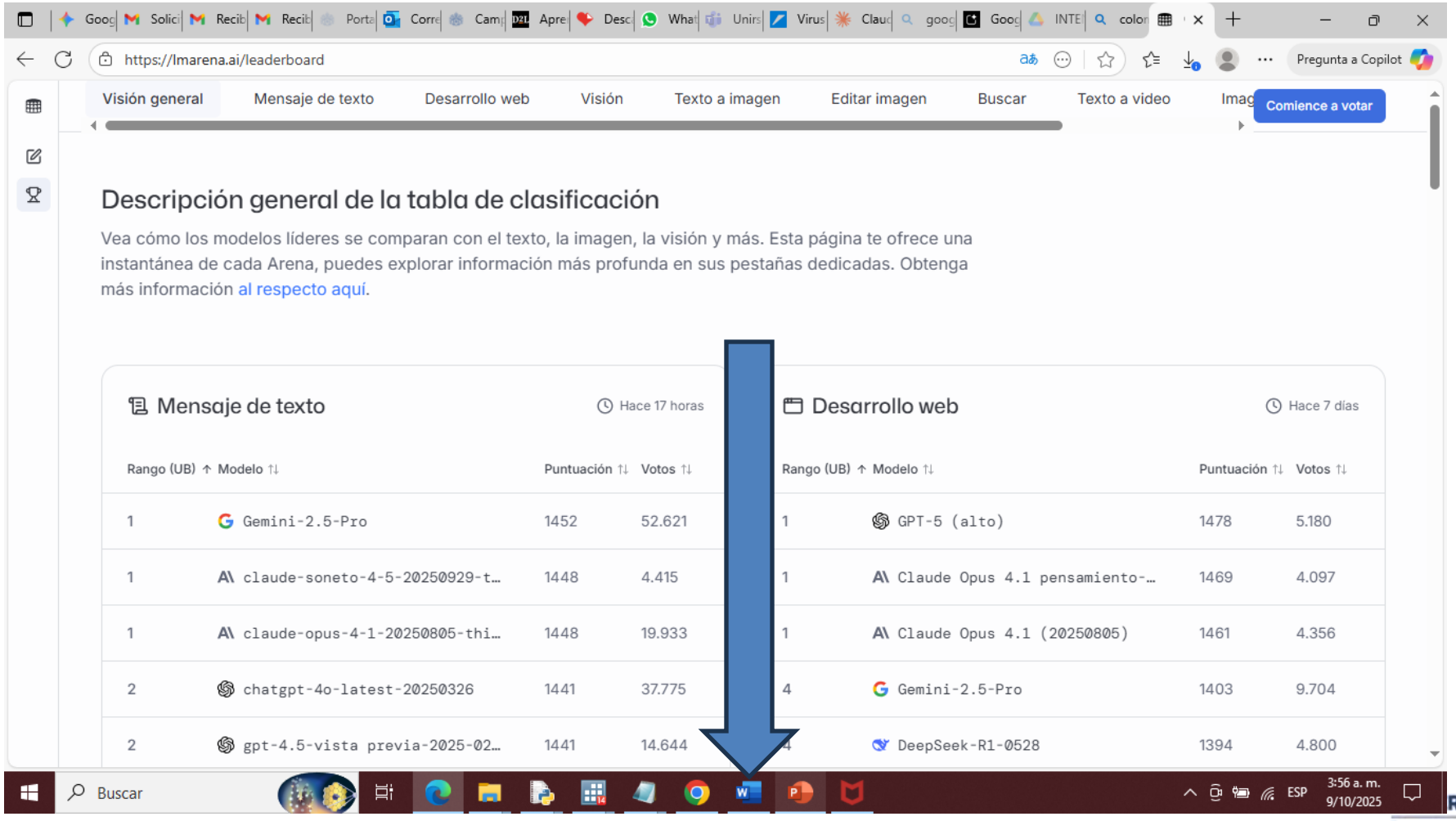
# Link Arena: https://lmarena.ai/



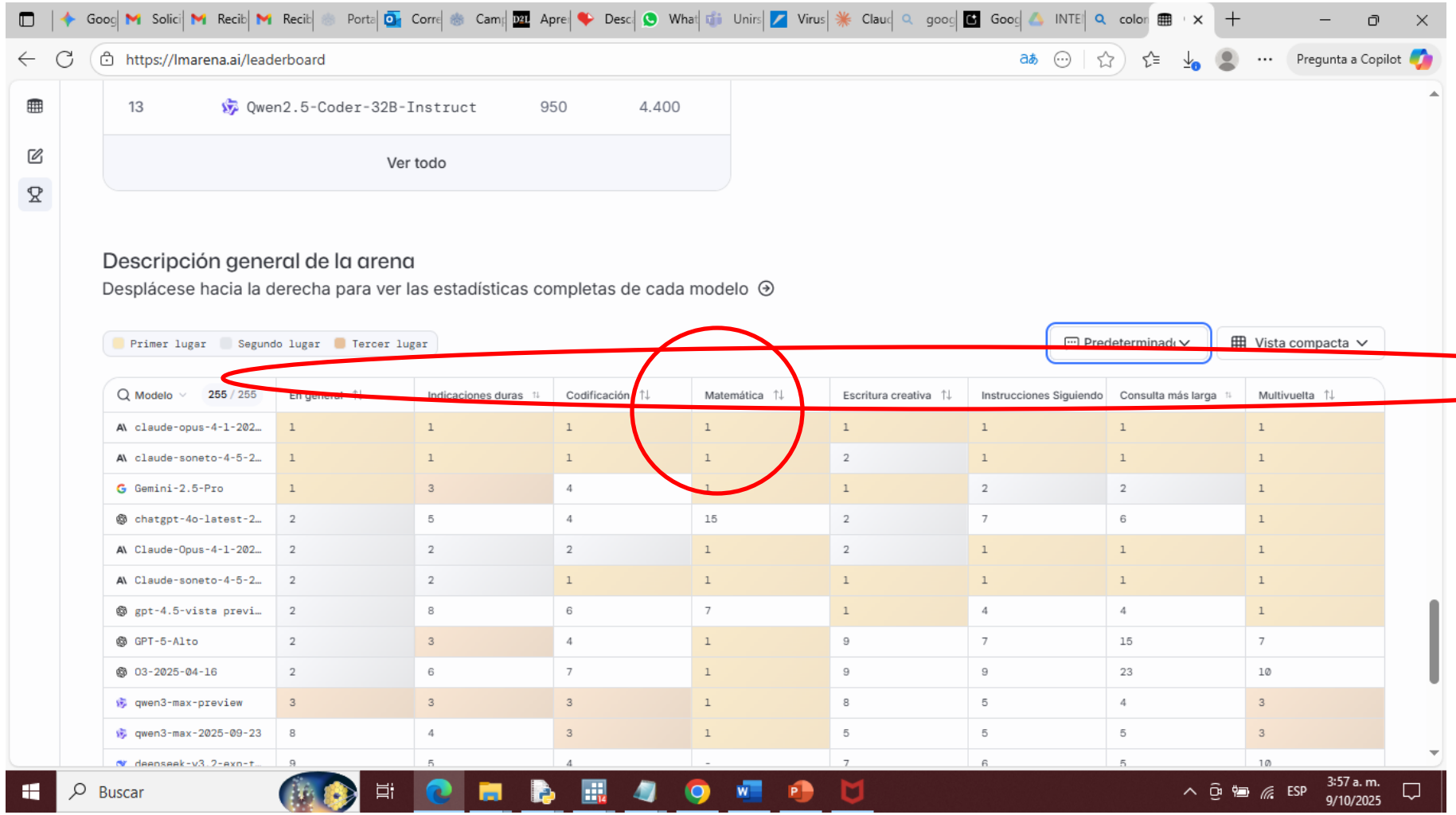
# Link Arena: https://lmarena.ai/



# Link Arena: https://lmarena.ai/



# Link Arena: https://lmarena.ai/



# 2. Un estudio de caso: automatización de informes analíticos de EISP

# Arquitectura Inteligente:
# El Panel de Control y la Sala de Máquinas

- **Estructura en diferentes archivos:** Un "Panel de Control" (Script_principal.py) fácil de usar para definir QUÉ analizar, y una "Sala de Máquinas" con diferentes Script que tienen la lógica compleja de CÓMO hacerlo.

- **Grandes beneficios:** Este diseño hace que el sistema sea fácil de usar (solo se edita un archivo), reutilizable en su gran mayoría para otros eventos de salud y simple de actualizar.

- **Principio clave:** Se separa la configuración del usuario (los filtros) de la lógica de programación, reduciendo errores y agilizando su elaboración.

# El Corazón del Sistema: Informes Dinámicos con Multi-Filtros

- **Informes a la medida en segundos:** El usuario solo necesita ajustar variables sencillas en un bloque de configuración para definir el análisis exacto que necesita.

- **Potentes capas de análisis:** Permite combinar filtros geográficos (Municipio), temporales (Semana), demográficos (Sexo, Curso de Vida) y específicos del evento (Desencadenante, Reincidencia).

- **Adaptación total y automática:** Al cambiar un filtro, todo el informe (textos, tablas, gráficas y hasta el nombre del archivo) se reconstruye instantáneamente para reflejar la nueva consulta.

# Automatización Total: Del Dato Crudo al Informe Profesional

- **Inteligencia más allá de los números:** El sistema realiza análisis estadísticos avanzados según el tamaño de la “muestra” (pruebas Z y Fisher), genera conclusiones escritas automáticamente (narrativas) y crea visualizaciones epidemiológicas clave.

- **Producto final de alta calidad:** Genera automáticamente un documento profesional en Word y PDF, con estilos, tablas con formato de color, encabezados y numeración de página.

- **El máximo valor**: Transforma días de trabajo manual en minutos, estandariza la calidad de los informes y asegura la precisión de los datos, liberando tiempo para….

# Ejemplo:
# Intento de suicidio

# INTENTO DE SUICIDIO

VALLE DEL CAUCA (sin distritos)

Semana epidemiológica 39 de 2025

*Introducción*

El intento de suicidio, definido como una conducta autoinfligida no fatal, constituye un problema crítico de salud pública en Colombia. Desde 2016, el Sistema Nacional de Vigilancia en Salud Pública (Sivigila) monitorea sistemáticamente este evento a través de la ficha de notificación 356, permitiendo caracterizar su comportamiento epidemiológico y orientar las intervenciones preventivas.

Este informe analiza los casos de intento de suicidio ocurridos en el Valle del Cauca (sin distritos) hasta la semana epidemiológica 39 de 2025, con énfasis en la identificación de tendencias temporales, características sociodemográficas y factores asociados que permitan fortalecer la respuesta intersectorial y la focalización de estrategias preventivas.

THE GLOBAL HEALTH NETWORK
LATIN AMERICA AND THE CARIBBE

Universidad del Valle

**Solicitar el cálculo de la variación porcentual y la diferencia absoluta entre los casos de dos años**

*Análisis de la situación actual*

En el Valle del Cauca sin distritos, hasta la semana epidemiológica 39 del año 2025, se han registrado 1.143 casos de intento de suicidio, lo que representa un incremento del 9,7% (101 casos más) en comparación con 2024. El promedio semanal durante 2025 es de 29,3 casos.

**Archivos externos con las poblaciones**

*Análisis Comparativo de Tasas por Municipio*

En el análisis departamental acumulado a la semana 39, los municipios que se destacan por tener las tasas más altas son: Alcalá (134,0 por 100.000 hab.), Vijes (112,6 por 100.000 hab.) y El Dovio (108,2 por 100.000 hab.). La tasa para el total del Valle del Cauca (sin distritos) es de 56,0, comparada con 51,1 en 2024.

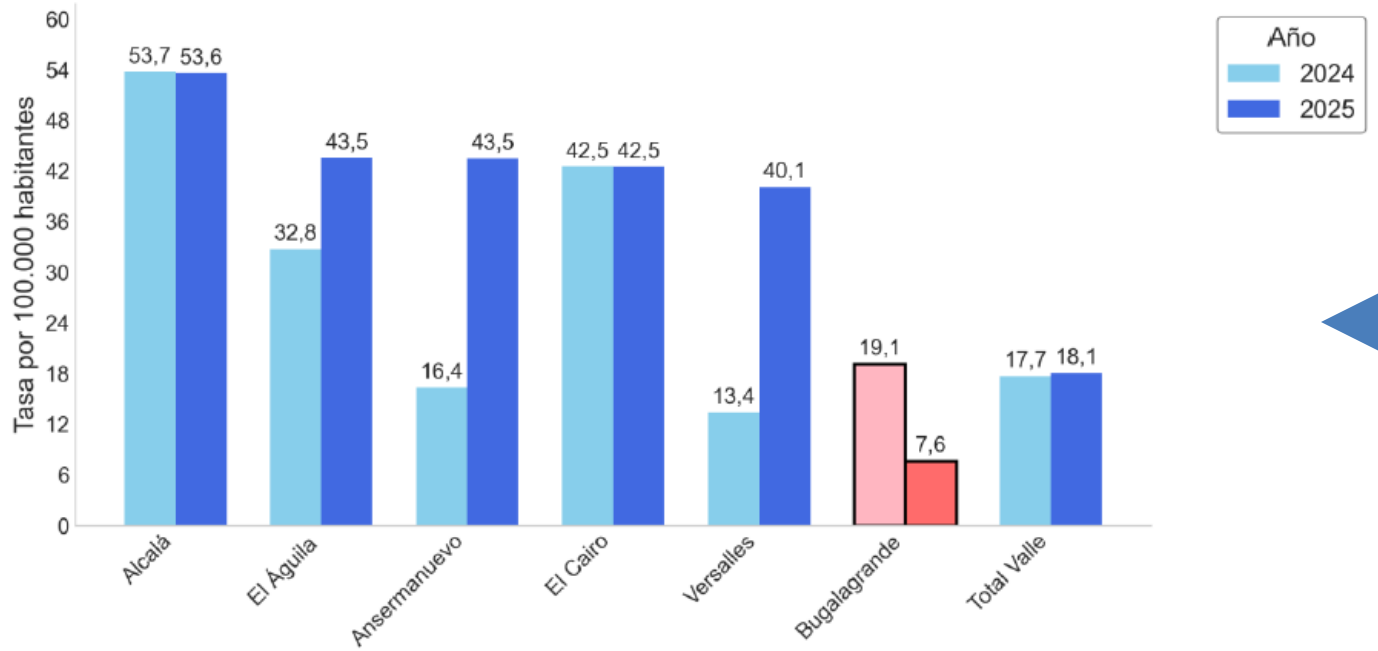
**Pedir la creación de un ranking de municipios por tasa para extraer el top 3**

GOBERNACIÓN Departamento del Valle del Cauca

paraíso de todos

# Ejemplo: Bugalagrande

*Gráfica 1: Municipios con mayores Tasas de Intento de Suicidio y Total Valle, Semanas 1-39, 2025 vs 2024. Filtros: Desencadenante: Problemas de pareja*




Fuente: Sivigila. Elaboración propia.

**Graficar solo los 5 municipios con las tasas más altas y añadir el total departamental. El municipio filtrado debe verse también reflejado en colores diferentes**

THE GLOBAL HEALTH NETWORK
LATIN AMERICA AND THE CARIBBEAN

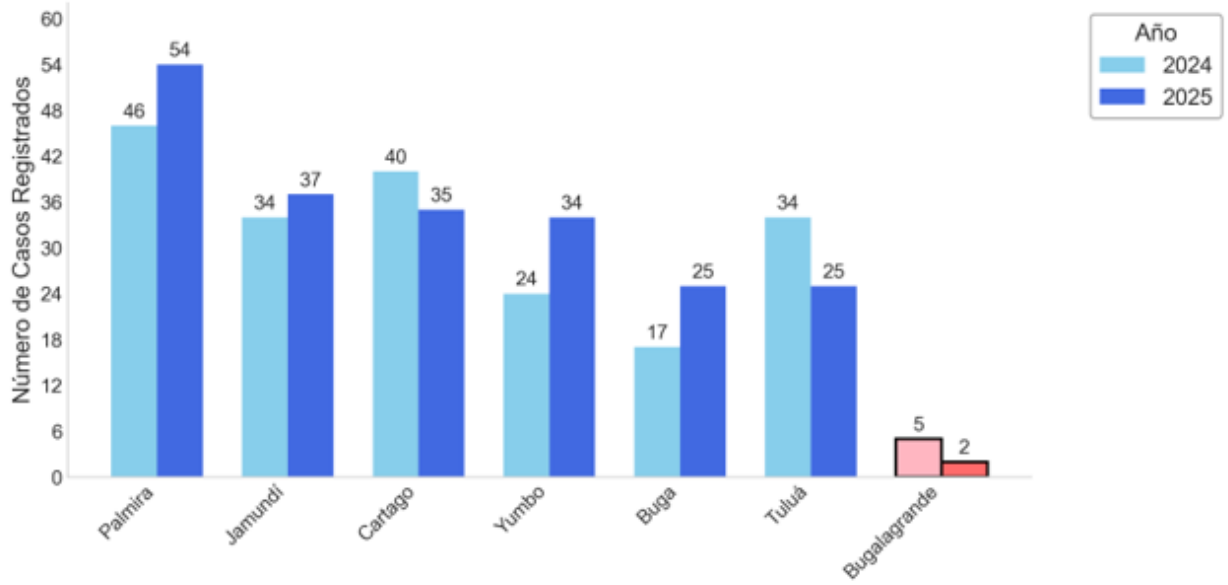
Universidad del Valle

*Análisis de Casos Absolutos de Intento de Suicidio con desencadenante de problemas de pareja por Municipio*

En este contexto, el municipio de Bugalagrande se encuentra entre los 3 municipios que comparten la posición número 32 (2 casos cada uno) de un total de 40 municipios del departamento.

**Generar la misma gráfica anterior, pero usando casos absolutos en lugar de tasas**

*Gráfica 2: Municipios con mayor Número de Casos de Intento de Suicidio, Semanas 1-39, 2025 vs 2024. Filtros: Desencadenante: Problemas de pareja*




GOBERNACIÓN Departamento del Valle del Cauca

paraíso de todos

**Pedir el cálculo de la mediana excluyendo los años de Pandemia y los años iniciales por bajo registro, Pedir el cálculo del promedio para la prueba de significancia estadística***

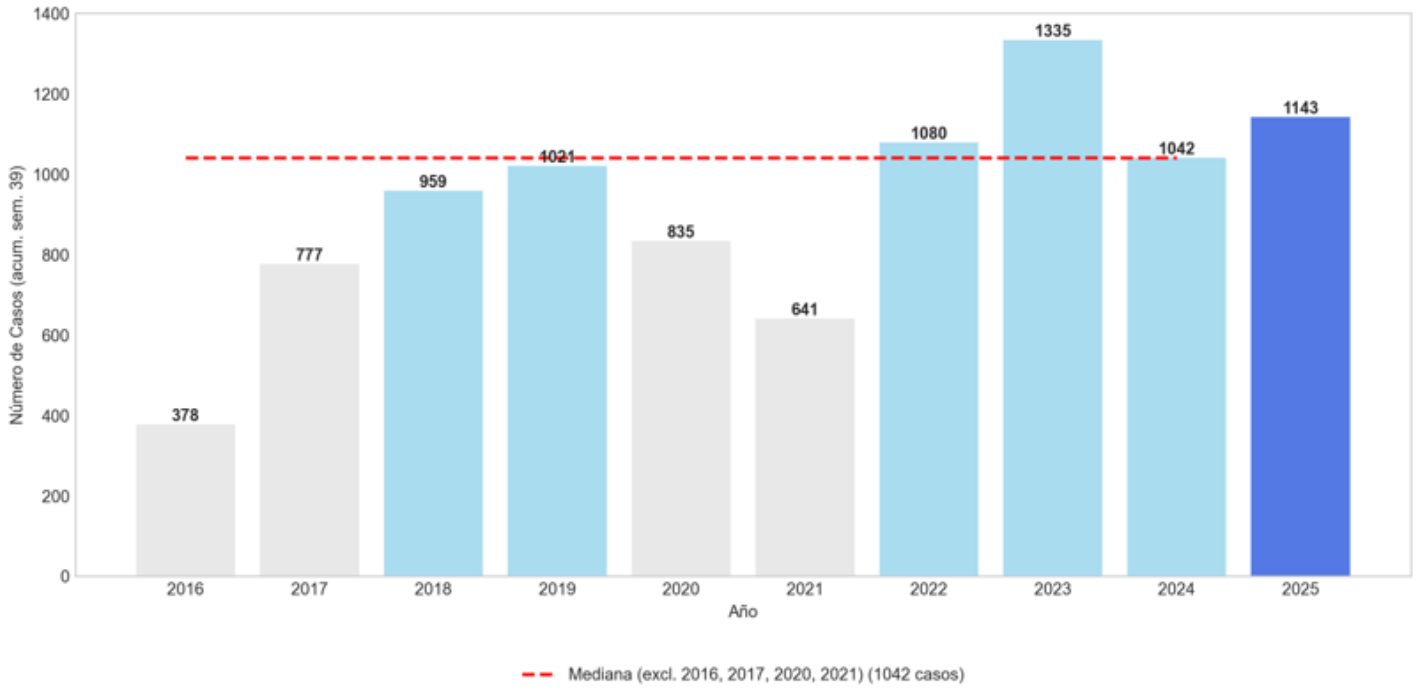
En el Valle Del Cauca Sin Distritos, hasta la semana 39, se han registrado 1.143 casos. Este valor supera la mediana histórica (1.042) para el mismo periodo (2016-2024, excluyendo 2016, 2017, 2020, 2021). Durante 2016-2024, el máximo se registró en 2023 (1.335), mientras que 2016 presenta el mínimo (378). Al comparar con el promedio histórico (1.087 casos), la diferencia no es estadísticamente significativa.

*Nota: Se utilizan umbrales ultra-conservadores. Para n≥1000: α=0.001 (prueba Z). Umbral muy estricto para prevenir significancia espuria en muestras muy grandes. IC95%: Rango donde, con 95% de certeza, se encuentra la verdadera diferencia. Si el intervalo incluye cero, sugiere que podría no haber diferencia real. RP (Razón de Proporciones) compara 2025 vs 2024: RP<1 indica mayor proporción en 2024; RP>1 indica mayor proporción en 2025; RP=1 indica proporciones iguales. Nota metodológica: Para el cálculo de la mediana histórica y las pruebas de significancia estadística, se excluyeron los años 2016, 2017 por subregistro inicial del sistema de vigilancia y los años 2020, 2021 debido a efectos de la pandemia COVID-19 en la notificación.*

**Pedir el cálculo de pruebas de significancia de acuerdo al tamaño poblacional del filtro aplicado**

# Pedir el gráfico de barras con todas especificaciones observadas

*Gráfica 3: Comportamiento Histórico Casos Intento de Suicidio (2016-2025) y Mediana (2016-2024), Valle Del Cauca Sin Distritos, Acumulado Semanas 1-39*




Fuente: Sivigila. Elaboración propia

**Compara la proporción de casos por sexo entre 2024 y 2025. Indica si la diferencia es estadísticamente significativa y crea una tabla con los resultados. Pedir el formato que se requiera**

*Análisis de Datos Básicos*

El sexo predominante es el femenino, representando el 64,4 % (736 casos). Esta proporción disminuyó 0,4 puntos porcentuales respecto a 2024 (675 casos). IC95%: -4,4 a 3,6, RP=0,99, p = 0,850. Esta diferencia no es estadísticamente significativa, .

*Tabla 1: Distribución por Sexo, Intento de Suicidio, Valle del Cauca sin distritos, Semanas 1-39, 2025 vs 2024*

| Sexo | Casos 2024 | % 2024 | Casos 2025 | % 2025 | Variación 2025 vs 2024 |
|---|---|---|---|---|---|
| Femenino | 675 | 64,8 % | 736 | 64,4 % | +61 (+9,0 %) |
| Masculino | 367 | 35,2 % | 407 | 35,6 % | +40 (+10,9 %) |
| **Total** | **1.042** | **100,0 %** | **1.143** | **100,0 %** | **+101 (+9,7 %)** |

Fuente: Sivigila. Elaboración propia.

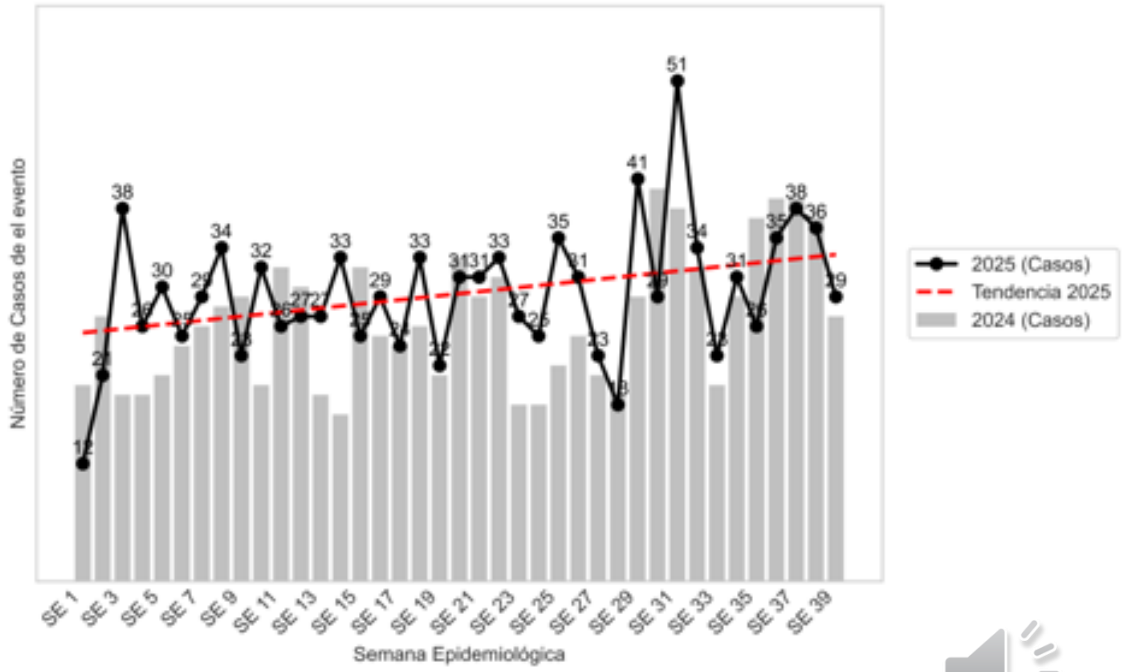
THE GLOBAL HEALTH NETWORK
LATIN AMERICA AND THE CARIBBEAN

Universidad del Valle

*Comportamiento Semanal de los Casos de Intento de Suicidio*

La gráfica presenta la distribución semanal de los casos del evento en el Valle del Cauca (sin Distritos) hasta la semana epidemiológica 39. Se compara el comportamiento del año 2025 (línea) con el del año 2024 (barras). La semana con el mayor número de reportes para el año 2025 es la semana 31, donde se registran 51 casos. Se observa una tendencia moderada al aumento, con un cambio promedio de 0,21 casos por semana (p = 0,029).

*Gráfica 5: Casos Semanales de Intento de Suicidio, Valle Del Cauca (Sin Distritos), Semanas 1-39, 2025 vs 2024*




GOBERNACIÓN Departamento del Valle del Cauca

paraíso de todos

**Para todas las variables clave. Genera una tabla resumen con el tipo de prueba, el valor p y la conclusión para cada una**

*Tabla 11: Resumen de Pruebas de Significancia Estadística - Intento de Suicidio, Valle del Cauca sin distritos, Comparación 2024 vs 2025 (Semanas 1-39)*

| Aspecto analizado | Tipo de análisis | Tamaño del efecto | Estadístico | Valor p | Conclusión |
|---|---|---|---|---|---|
| Comparación histórica (2016-2024) | Prueba Z (α=0,001) | Δ=+246 casos vs promedio | Z=8,23 | p<0,001 | *** Altamente significativo |
| Sexo predominante: Femenino | Prueba Z (α=0,05) | Δ=-0,4% | Z=-0,19 | p=0,85 | No significativo |
| Curso de vida predominante: La juventud (18-28 años) | Prueba Z (α=0,05) | Δ=-1,9% | Z=-0,92 | p=0,357 | No significativo |
| Área predominante: Cabecera mpal | Prueba Z (α=0,05) | Δ=-7,7% | Z=-4,8 | p<0,001 | *** Altamente significativo |
| Régimen predominante: Subsidiado | Prueba Z (α=0,05) | Δ=0,1% | Z=0,04 | p=0,967 | No significativo |
| Hospitalización: Sí | Fisher exacta (α=0,1) | Δ=-45,2% | --- | p<0,001 | *** Altamente significativo |
| EAPB predominante: Nueva eps | Prueba Z (α=0,05) | Δ=2,0% | Z=1,04 | p=0,3 | No significativo |
| Mecanismo: Intoxicación | Prueba Z (α=0,05) | Δ=0,5% | Z=0,22 | p=0,826 | No significativo |
| Reincidencia: Sí | Prueba Z (α=0,05) | Δ=-0,5% | Z=-0,21 | p=0,83 | No significativo |
| Factor de riesgo: Antecedente de trastorno mental | Prueba Z (α=0,05) | Δ=1,5% | Z=0,72 | p=0,472 | No significativo |
| Tendencia semanal (aumento) | Regresión lineal | +0,21 casos/sem | Pendiente=0,210 | p=0,029 | * Tendencia significativa |
| Total de casos | Prueba Z para conteos (α=0,01) | Δ=101 casos | Z=2,16 | p=0,031 | Sin cambio significativo |

Fuente: Sivigila. Elaboración propia.

**Definir las reglas para interpretar los resultados, como los umbrales de significancia (alfa) según el tamaño de la muestra.**

| | | | | | |
|---|---|---|---|---|---|
| trastorno mental | | | | | |
| Tendencia semanal (aumento) | Regresión lineal | +0,21 casos/sem | Pendiente=0,210 | p=0,029 | * Tendencia significativa |
| Total de casos | Prueba Z para conteos (α=0,01) | Δ=101 casos | Z=2,16 | p=0,031 | Sin cambio significativo |

Fuente: Sivigila. Elaboración propia.

*Interpretación de la tabla:*

- *Δ (Delta): Representa la diferencia absoluta en puntos porcentuales o casos entre los dos años*
- *α (Alfa): Nivel de significancia según estándares científicos internacionales:*
  - *α=0,10: Para muestras <30 casos (análisis exploratorio)*
  - *α=0,05: Para 30-1000 casos (estándar científico universal)*
  - *α=0,01: Para 1000-5000 casos (ajuste por múltiples comparaciones)*
  - *α=0,001: Para >5000 casos (ajuste para big data)*
- *Valor p: Probabilidad de obtener esta diferencia por azar*
- *Significancia: p < α indica diferencia estadísticamente significativa*
- *Metodología: Prueba Z para n≥5 en cada celda, Fisher exacta para n<5 (estándar OMS/CDC)*
- *Nota: Se utiliza el menor tamaño muestral entre los dos años comparados para determinar α (enfoque conservador)*

THE GLOBAL HEALTH NETWORK
LATIN AMERICA AND THE CARIBBEAN

Universidad

**Extraer los resultados clave de cada sección (total de casos, perfil demográfico, características del evento) y presentarlos de forma concisa.**

*Síntesis de Hallazgos: Intento de Suicidio Año 2025*

- **En el Valle del Cauca sin distritos, hasta la semana epidemiológica 39, se han registrado 1.143 casos, lo que representa un incremento del 9,7% (101 casos más) en comparación con 2024. La tasa de incidencia es de 56,0 por 100.000 habitantes, un aumento de 4,9 puntos respecto a 2024 (tasa:** 51,1).
- El perfil sociodemográfico indica que el sexo femenino es el más afectado (64,4%), con un aumento del 9,0% respecto a 2024; el curso de vida predominante es juventud (18-28 años) (33,2%).
- En cuanto a las características del intento, el mecanismo más utilizado es la intoxicación (64,7%), con un aumento del 10,5% respecto a 2024; el 41,3% presenta reincidencia, con un aumento del 8,5% respecto a 2024; el factor de riesgo principal es antecedente de trastorno mental (39,0%), con un aumento del 14,1% respecto a 2024.

GOBERNACIÓN Departamento del Valle del Cauca
paraíso de todos

THE GLOBAL HEALTH NETWORK
LATIN AMERICA AND THE CARIBBEAN

Universidad del Valle

**Genera una tabla con el histórico de casos por EAPB de los últimos 5 años. Ordénala por los casos de 2025 y aplica un mapa de calor de rojo a verde**

*Tabla Anexa 1: Top 30 EAPB con Mayor Número de Casos de Intento de Suicidio en Valle del Cauca (sin distritos) (Acumulado hasta SE 39)*

| EAPB | 2021 | 2022 | 2023 | 2024 | 2025 | % 2025 |
|---|---|---|---|---|---|---|
| Nueva EPS | 97 | 209 | 298 | 281 | 331 | 29,0 % |
| Emssanar | 140 | 227 | 223 | 156 | 168 | 14,7 % |
| S.O.S | 84 | 132 | 151 | 144 | 156 | 13,6 % |
| Coosalud | 70 | 99 | 144 | 101 | 117 | 10,2 % |
| Sura EPS | 22 | 73 | 106 | 62 | 80 | 7,0 % |
| EPS Sanitas | 29 | 96 | 120 | 73 | 70 | 6,1 % |
| Salud Total EPS | 13 | 41 | 59 | 33 | 49 | 4,3 % |
| Comfenalco | 19 | 36 | 45 | 45 | 42 | 3,7 % |
| Asmet Salud | 30 | 32 | 60 | 44 | 40 | 3,5 % |
| Fiduprevisora S.A | 7 | 9 | 14 | 9 | 24 | 2,1 % |
| Policía Nacional | 5 | 12 | 10 | 9 | 16 | 1,4 % |
| Fuerzas Militares | 2 | 10 | 14 | 14 | 9 | 0,8 % |
| Compensar EPS | 6 | 19 | 11 | 17 | 5 | 0,4 % |
| Asociación Indígena Del Cauca | 5 | 11 | 8 | 7 | 5 | 0,4 % |
| Coomeva EPS | 66 | 17 | 1 | 1 | 3 | 0,3 % |

GOBERNACIÓN Departamento del Valle del Cauca
paraíso de todos

# 3. Datos no estructurados (Procesamiento de Lenguaje Natural)

*También existen los datos semiestructurados

# Situación problema:

La Alcaldía de Cali quiere saber qué opinan los ciudadanos en redes sociales sobre los hurtos en la ciudad. Necesitan identificar cómo se siente la gente con respecto a la seguridad. Con esta información podrán tomar mejores decisiones para prevenir delitos y mejorar la seguridad ciudadana.

# Antes de Analizar: ¿De Dónde Sacamos los Datos? 📜

•**El Primer Paso**: Para recolectar datos de internet, existen dos métodos eficientes principales con filosofías muy diferentes.

1. **La Vía Oficial - Las APIs (La Puerta Principal 🚪):** Una API es la forma oficial en que una plataforma (como X/Twitter o Facebook) te permite acceder a sus datos. Es como **pedir directamente del menú** de un restaurante: recibes la información de forma limpia, estructurada y con permiso. Es la opción ideal para grandes plataformas.

## 2. Web Scraping

API: Interfaz de Programación de Aplicaciones

# 1. La Vía Oficial - Las APIs

## La puerta SÍ existe, PERO está semi-cerrada:

### Facebook/Meta API:

- ✅ **SÍ funciona** para acceder a **TU PROPIO** contenido
- ✅ **SÍ funciona** si eres administrador de una página
- ❌ **NO funciona** para leer contenido público de otros usuarios
- ❌ **NO funciona** para "escuchar" conversaciones generales

### Twitter/X API:

- ✅ **SÍ funciona** (1.500 tweets/mes gratis)
- ✅ **SÍ funciona** completamente si pagas ($100-$5.000/mes)
- ⚠️ **Limitado** en plan gratuito

# Ejemplo: Twitter/X API:

## 📋 LISTA COMPLETA DE REQUISITOS:

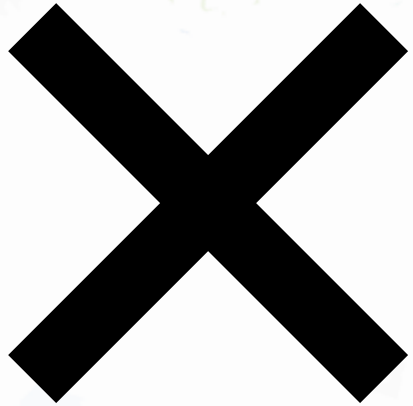
### I. Tener una cuenta de Twitter/X ✅

- La que ya tienes (tu cuenta personal normal)
- **Tiempo:** Depende si ya la tienes o debes crearla
- **Costo:** Gratis

# Ejemplo: Twitter/X API:

## II. Crear cuenta de DESARROLLADOR de Twitter

- **NO es lo mismo** que tu cuenta normal
- Es una cuenta especial para usar la API
- **Proceso:**
  1. Ir a: https://developer.twitter.com/
  2. Clic en "Sign up" o "Apply"
  3. Iniciar sesión con tu cuenta normal de Twitter
  4. **Llenar formulario detallado:**
     - ¿Para qué usarás la API? (Di: "Análisis académico/educativo")
     - ¿Qué datos analizarás? (Di: "Tweets públicos sobre salud pública")
     - ¿Compartirás datos con terceros? (Di: "No")
     - ¿Usarás datos de gobierno? (Di: "No")

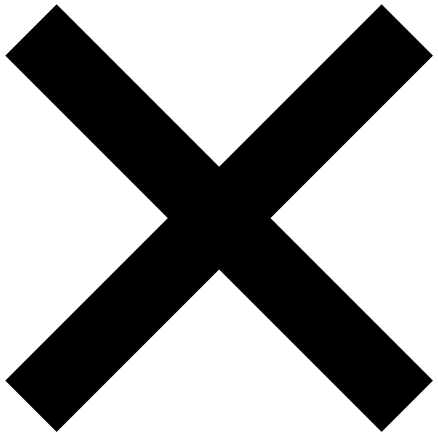


# Ejemplo: Twitter/X API:

5. Aceptar términos

6. **ESPERAR APROBACIÓN** ⏰

- ⏱️ **Tiempo:** 20-30 minutos llenar formulario + **2-7 DÍAS de espera** para aprobación
- 💰 **Costo:** Gratis (plan Free)
- ⚠️ **IMPORTANTE:** Twitter revisa manualmente tu solicitud



# Ejemplo: Twitter/X API:

## III. Crear un Proyecto y una App

(Solo después de que te aprueben)

- En el Developer Portal
- Crear "Project" → Crear "App"
- ⏱️ **Tiempo:** 5 minutos
- 💰 **Costo:** Gratis

## IV. Obtener el Bearer Token

- Dentro de tu App, ir a "Keys and Tokens"
- Generar "Bearer Token"
- **COPIARLO Y GUARDARLO** (no lo vuelves a ver)
- ⏱️ **Tiempo:** 2 minutos

💰 **Costo:** Gratis

# Ejemplo: Twitter/X API:

## V. Instalar librería Python (para uso en maquina local)

bash

pip install tweepy

- En tu terminal/consola
- ⏱️ **Tiempo:** 1 minuto
- 💰 **Costo:** Gratis

## VI. Reemplazar el Bearer Token en el código

python

*# Línea 37 del Notebook 2:*

BEARER_TOKEN = "AAAA...%2BuSeid..."

*# Tu token real*

- ⏱️ **Tiempo:** 30 segundos
- 💰 **Costo:** Gratis

# Antes de Analizar: ¿De Dónde Sacamos los Datos? 📜

**2. La Vía Alternativa - Web Scraping (La Ruta Creativa 🤖):** El scraping es automatizar un navegador para que "lea" y copie la información pública de una página web. Es tu **única opción para la gran mayoría de sitios que NO tienen una API**, como periódicos locales, foros o tiendas online.

# 2. La Vía Alternativa - Web Scraping

## ¿QUÉ ES WEB SCRAPING?

**Imaginen que quieren información de 100 páginas web, pero copiar y pegar manualmente tomaría mucho tiempo. El web scraping es como tener un robot que:**

1.Abre el navegador
2.Va a la página web
3.Lee el contenido
4.Extrae solo lo que necesitas
5.Lo guarda en una tabla ordenada

**Todo esto automáticamente en minutos."**

# 2. La Vía Alternativa Web Scraping

## PASO 1: ENTENDER LA ESTRUCTURA DE UNA PÁGINA WEB

**Concepto clave:** Toda página web está escrita en HTML, que es como el 'código fuente' que el navegador lee para mostrar la información visualmente.

Lo que VES:

```
┌──────────────────┐
│ @usuario123      │
│ Me robaron en    │
│ Cali ayer 😔     │
│ 15 likes         │
└──────────────────┘
```

→

Lo que HAY DETRÁS:

```
<div class="tweet">
   <span>@usuario123</span>
   <p>Me robaron en Cali</p>
   <span>15 likes</span>
</div>
```

# 2. La Vía Alternativa Web Scraping

## PASO 2: IDENTIFICAR QUÉ QUEREMOS EXTRAER

De un tweet sobre hurtos, ¿qué información queremos?

- ✅ Texto del tweet
- ✅ Fecha de publicación
- ✅ Usuario que lo publicó
- ✅ Número de likes/retweets
- ✅ Ubicación mencionada

El scraping necesita que le digamos EXACTAMENTE qué queremos extraer

# 2. La Vía Alternativa Web Scraping

## PASO 3: PROCESO DE SCRAPING

```
1. NAVEGADOR AUTOMATIZADO
   ↓
   [Robot abre Chrome/Firefox]
   ↓

2. IR A LA PÁGINA
   ↓
   [Robot va a Twitter.com/search?q=hurtos+Cali]
   ↓

3. ESPERAR QUE CARGUE
   ↓
   [Robot espera 3 segundos para que aparezcan los tweets]
   ↓
```

```
4. BUSCAR ELEMENTOS EN EL HTML
   ↓
   [Robot busca: <div class="tweet">]
   ↓

5. EXTRAER INFORMACIÓN
   ↓
   [Robot lee el texto de cada tweet]
   ↓

6. GUARDAR EN TABLA
   ↓
   [Robot guarda en Excel/CSV]
```

# 2. La Vía Alternativa Web Scraping

## PASO 4: HERRAMIENTAS NECESARIAS

**1. Selenium** (El navegador robot)

Es como un programa que controla Chrome automáticamente. Hace TODO lo que harías tú: hacer clic, escribir, scroll, etc.

**2. Beautiful Soup** (El lector de HTML)

Lee el código HTML de la página y encuentra la información específica que necesitas.

**3. Pandas** (El organizador de datos)

Toma todos los datos extraídos y los organiza en una tabla como Excel.

THE GLOBAL HEALTH NETWORK
LATIN AMERICA AND THE CARIBBEAN

Universidad del Valle

# Comparativo API - Web Scraping

| Aspecto | API | Web Scraping |
|---|---|---|
| Legalidad | ✅ Legal | ⚠️ Zona gris |
| Dificultad | 🟢 Media | 🔴 Alta |
| Datos disponibles | ⚠️ Limitados | ✅ Todos los públicos |
| Estabilidad | ✅ Estable | ⚠️ Se rompe si cambian HTML |
| Velocidad | ✅ Rápido | ⚠️ Más lento |
| Costo | 🟡 A veces $$ | 🟢 Gratis |
| Uso en producción | ✅ Recomendado | ⚠️ Con precaución |

GOBERNACIÓN Departamento del Valle del Cauca
paraíso de todos

# 2. La Vía Alternativa - Web Scraping

## PASO 5: ANÁLISIS POSTERIOR A LA OBTENCIÓN DEL ARCHIVO EXCEL O CVS

- Solicitar un script para análisis narrativo académico mediante modelos de Transformers (Hugging Face): generar síntesis coherentes, extraer palabras clave e identificar entidades por documento.

- Se recomienda usar Google Colab debido al tamaño de las librerías.

- Especificar formato de salida: reporte Word con párrafos fluidos por registro, Excel enriquecido con metadatos, y análisis agregado del documento completo.

- Requerir adaptabilidad automática: detección inteligente de columnas, manejo de datos incompletos y robustez para cualquier temática académica.

# 2. La Vía Alternativa Web Scraping

## PASO 5: ANÁLISIS POSTERIOR A LA OBTENCIÓN DEL ARCHIVO EXCEL O CVS

## Prompt ejemplo:

Necesito un script Python completo para Google Colab que analice documentos académicos de un Excel/CSV mediante IA.
Requisitos técnicos:

- Usar modelos de Hugging Face (Transformers) para generar síntesis narrativas de 200-300 palabras por cada documento
- Extraer palabras clave automáticamente con KeyBERT
- Identificar entidades (organizaciones, lugares, personas) con spaCy
- Detectar automáticamente columnas relevantes (título, contenido, URL)
- Manejar datos incompletos de forma inteligente

# 2. La Vía Alternativa Web Scraping

## PASO 5: ANÁLISIS POSTERIOR A LA OBTENCIÓN DEL ARCHIVO EXCEL O CVS

## Prompt ejemplo:

**Outputs esperados:**

- Reporte Word profesional con párrafos narrativos fluidos por documento, incluyendo referencias a fuentes
- Excel enriquecido con síntesis, palabras clave y entidades por fila
- Análisis agregado del corpus completo (conceptos más frecuentes)

**Flujo de trabajo**:

1. Solicitar carga de archivo (CSV/Excel) mediante files.upload()
2. Procesar automáticamente todos los registros
3. Descargar resultados a la carpeta Descargas automáticamente
4. El script debe ser robusto, adaptable a cualquier temática académica y mostrar progreso durante el procesamiento.

THE GLOBAL HEALTH NETWORK
LATIN AMERICA AND THE CARIBBEAN

Universidad del Valle

www.valledelcauca.gov.co/

# GRACIAS

GOBERNACIÓN
Departamento del
Valle del Cauca

paraíso de todos

Elaborado por:

**Víctor Hugo Muñoz Villa**

The Global Health Network LAC – Universidad del Valle



Mayor información en: proyecto.globalhealthnetworklac@correounivalle.edu.co