

# Manual de usuario — Dashboard Saber Pro (Streamlit)

Este documento describe cómo usar la aplicación web del proyecto Saber Pro: acceso, carga de datos, menú y contenido de cada sección.

## Tabla de contenidos

- 1. Introducción
- 2. Requisitos
- 3. Acceso a la aplicación
- 4. Carga del archivo
- 5. Menú y páginas
- 6. Ampliaciones del manual (interpretación de KPIs, gráficos EDA, Groq, glosario, etc.)
- 7. Posibles problemas y soluciones

## 1. Introducción

El dashboard Saber Pro es una aplicación interactiva que permite explorar los resultados de las Pruebas Saber Pro (2018-2021) y su relación con variables socioeconómicas y territoriales. Incluye:

- Resumen del dataset, auditoría de nulos y evidencia de missingness (MNAR).
- Procesamiento: trazabilidad de la imputación y validación.
- KPI: indicadores de brechas (estrato, computador, educación materna, región, impacto territorial).
- EDA: análisis exploratorio (cuantitativo, cualitativo y gráficos).
- Groq IA: consultas asistidas por modelo de lenguaje (requiere API key).

## 2. Requisitos

- Archivo del dataset: `saber_pro.csv` (mismo que se describe en el README del repositorio, sección Set de datos. El nombre del archivo debe ser exactamente ``saber_pro.csv`` (el dashboard lo valida al subir).

## 3. Acceso a la aplicación

- En línea: Saber Pro — Streamlit
- En local: tras clonar el repositorio e instalar dependencias (ver Clonación e instalación, ejecutar desde la carpeta del proyecto:

```
cd streamlit
```

```
streamlit run main_app.py
```

Se abre en el navegador la URL indicada en la terminal (por defecto `http://localhost:8501`).

## 4. Carga del archivo

1. En el menú lateral (sidebar) aparece la opción "Subir saber\_pro.csv".
2. Se hace clic en "Browse files" (o se arrastra el archivo) y se selecciona el archivo saber\_pro.csv.
3. Si el archivo tiene otro nombre, la aplicación muestra un mensaje de advertencia y no procesa los datos hasta que se suba un archivo llamado saber\_pro.csv.
4. Tras una carga correcta, se mostrará una barra de progreso y luego el menú de navegación y el contenido según la página seleccionada.

### 4.1 Flujo de uso recomendado

Para sacar el máximo partido al dashboard en la primera vez:

1. Se sube el archivo saber\_pro.csv (sección 4).
2. Se revisa Resumen para ver el tamaño del dataset, nivel de nulos y evidencia de missingness (MNAR).
3. Se consulta Procesamiento para entender cómo se imputaron los datos y que la validación Cramér antes/después sea estable.
4. Se accede a KPI para ver las brechas (KRAS, FD, KCC, KDS) y las tablas de equidad regional e impacto territorial.
5. Se explora EDA (pestañas Cuantitativo, Cualitativo y Gráficos) para análisis más detallado y visual.
6. Opcional: en Groq IA, con API key configurada, se pueden pedir insights ejecutivos o recomendaciones de política.

No es obligatorio seguir este orden; el menú permite saltar a cualquier página en cualquier momento.

## 5. Menú y páginas

En el sidebar, después del selector de archivo, aparece el menú "Ir a" con las opciones Resumen, Procesamiento, KPI, EDA y Groq IA. A continuación se describe cada página.

### 5.1 Sidebar

Selector "Subir saber\_pro.csv" con archivo cargado y menú "Ir a" con las cinco opciones (Resumen, Procesamiento, KPI, EDA, Groq IA).

## 5.2 Resumen

La página Resumen / Observaciones ofrece una vista general del dataset cargado y de la calidad de los datos, con énfasis en los valores faltantes (nulos) y en si su patrón está asociado al rendimiento (evidencia de missingness no aleatorio, MNAR). Consta de cuatro tarjetas superiores y tres bloques debajo.

Tarjetas superiores (métricas base)

Tarjeta	Qué muestra	Para qué sirve
Filas	Número total de filas del dataset original (antes de cualquier eliminación).	Conocer el tamaño de la muestras.
Columnas	Número total de columnas del dataset original.	Ver cuántas variables (identificación, socioeconómicas, target, indicadores) tiene el archivo.
Nulos promedio	Porcentaje medio de celdas vacías en todo el dataset (promedio del % de nulos por columna).	Valor global de "cuánto falta"; un valor alto indica que la imputación o el tratamiento de nulos será importante.
Kruskal (N_NULOS vs target)	Resultado del test de Kruskal-Wallis que compara la cantidad de nulos por fila (N_NULOS) entre los distintos niveles de RENDIMIENTO_GLOBAL. Si el p-value es muy bajo ( $p \approx 0$ ), se rechaza la hipótesis de que el número de nulos sea igual en todos los grupos.	Evidencia de MNAR: indica que los datos faltantes no son aleatorios; su cantidad está asociada al nivel de rendimiento (por ejemplo, los estudiantes de bajo rendimiento suelen tener más celdas vacías). Eso obliga a tratar los nulos con cuidado (imputación condicional, etc.) y a no ignorarlos.

Muestra de datos

- Slider "Filas a mostrar": permite elegir entre 5 y 200 filas (paso 5; por defecto 20) para previsualizar el dataset.
- Tabla: muestra esa cantidad de filas del dataset original (sin procesar), con todas las columnas. Sirve para revisar nombres de variables, tipos de valores y una primera idea del contenido.

Auditoría de nulos

- Slider "Filtrar: nulos\_% mínimo": de 0 % a 100 % (paso 1). Solo se muestran columnas cuyo porcentaje de nulos sea mayor o igual a ese valor. En 0 % se ven todas las columnas; subiendo el filtro se enfocan las columnas más afectadas por faltantes.
- Tabla: para cada columna aparecen el porcentaje de nulos (nulos\_%), el tipo de dato (tipo) y el número absoluto de nulos (nulos\_n). Las columnas se ordenan por nulos\_% descendente. Sirve para priorizar qué variables requieren imputación o revisión.

Missingness (observaciones)

- Descripción de N\_NULOS por clase de rendimiento: estadísticas (media, desviación, etc.) del número de nulos por fila (N\_NULOS) desglosadas por nivel de RENDIMIENTO\_GLOBAL. Permite ver si, por ejemplo, el grupo “bajo” tiene en promedio más nulos que el grupo “alto”.

- Asociación con MISSING\_HEAVY (Cramér's V): para cada variable candidata se muestra la asociación (Cramér's V) con MISSING\_HEAVY (indicador de si la fila tiene 2 o más nulos). Valores altos indican que esa variable está más relacionada con “tener muchos nulos”; ayuda a entender qué factores se asocian al patrón de datos faltantes y refuerza la lectura de MNAR.

En conjunto, la pantalla Resumen sirve para dimensionar el dataset, revisar la calidad de los datos y justificar que el missingness se trata como no aleatorio (MNAR) antes de pasar al detalle del Procesamiento y a los KPI.

## 5.3 Procesamiento

La página Procesamiento (trazabilidad) muestra el pipeline aplicado a los datos: desde la decisión de eliminar filas con muchos nulos hasta la imputación condicional y la validación. Todo está numerado en cinco bloques para seguir el orden lógico del tratamiento.

### 1) Evaluación de eliminación por nulos ( $k \geq 2$ )

- Qué hace: Se identifican las filas con 2 o más celdas vacías ( $k \geq 2$ ) y se compara la distribución del target (RENDIMIENTO\_GLOBAL) entre las filas que se eliminarían y el dataset completo.

- Tabla: Columnas como “Eliminados %”, “Total %” y “Diferencia\_abs\_pp” por nivel de rendimiento (bajo, medio-bajo, medio-alto, alto). Si las diferencias son grandes (por ejemplo, entre “Eliminados” y “Total” en la categoría “bajo”), eliminar esas filas podría sesgar el análisis.

- Para qué sirve: Decidir de forma transparente si es aceptable eliminar esas filas o si conviene ser más conservador; en el dashboard la eliminación ya está aplicada en el pipeline, pero aquí se ve la evidencia.

### 2) Evidencia de MNAR

- Qué muestra: El p-value del test de Kruskal-Wallis (N\_NULOS vs nivel de rendimiento) y una tabla con la asociación (Cramér's V) entre varias variables (estrato, computador, departamento, etc.) y el indicador MISSING\_HEAVY (filas con  $\geq 2$  nulos).

- Para qué sirve: Refuerza que los datos faltantes no son aleatorios: están asociados al rendimiento y a variables socioeconómicas. Justifica usar imputación condicional por grupo (Estrato, Departamento) en lugar de imputar sin contexto.

### 3) Imputación condicional por (Estrato, Departamento)

- Qué muestra: Lista de variables tratadas (las que tenían nulos y se imputan por grupo) y una tabla con el porcentaje de nulos restantes por variable después de la imputación.

- Cómo funciona: Para cada variable con faltantes, se rellena usando la moda o la mediana dentro del grupo definido por Estrato y Departamento del estudiante, de modo que la imputación sea coherente con el contexto socioeconómico y territorial.

- Para qué sirve: Ver qué columnas se imputaron y comprobar que, tras el proceso, el % de nulos restantes sea bajo o cero.

#### 4) Validación de estabilidad (Cramér's V Antes vs Después)

- Qué muestra: Una tabla que compara, para cada variable imputada, el Cramér's V con el target (RENDIMIENTO\_GLOBAL) antes y después de la imputación.
- Para qué sirve: Comprobar que la fuerza de asociación con el rendimiento no cambió de forma artificial: si las diferencias son pequeñas (p. ej. inferiores a 0,01 en valor absoluto), la imputación no distorsiona las relaciones que interesan para los KPI y el EDA.

#### 5) Dataset procesado (muestra)

- Qué muestra: Las primeras 30 filas del dataset ya procesado (después de eliminación, imputación, codificación de estrato/binarias/target y posible reducción de cardinalidad).
- Para qué sirve: Revisar que las transformaciones se aplicaron bien (columnas numéricas o categóricas codificadas, sin nulos donde corresponde) antes de pasar a KPI y EDA.

En conjunto, la pantalla Procesamiento permite auditar cada paso del tratamiento de datos y validar que la imputación es adecuada antes de interpretar los indicadores.

## 5.4 KPI

La página KPI concentra los indicadores de brechas socioeconómicas y territoriales en relación con el rendimiento académico (bajo vs resto). Arriba hay cinco tarjetas con los valores principales; debajo, el detalle de cada KPI con texto explicativo y tablas. Los indicadores son asociativos y diagnósticos, no causales: sirven para priorizar intervenciones y monitorear equidad.

#### Tarjetas superiores

Tarjeta	Qué muestra	Para qué sirve
P(Nacional) bajo	Porcentaje de estudiantes en bajo rendimiento en todo el dataset procesado.	Base de referencia: proporción nacional de riesgo.
KRAS (pp)	Diferencia en puntos porcentuales entre P(bajo) en estratos 1–2 y P(bajo) en estratos 5–6.	Brecha por estrato socioeconómico; valores altos = mayor desigualdad.
FD (pp)	Diferencia en pp entre P(bajo) sin computador y P(bajo) con computador.	Brecha digital; cuánto más riesgo asociado a no tener PC.
KCC (pp)	Diferencia en pp entre P(bajo) con madre de baja educación y P(bajo) con madre de alta educación.	Brecha por capital cultural (educación materna).
KDS	Promedio de Cramér's V entre variables socioeconómicas (estrato, tecnología, educación, región, etc.) y el target.	Qué tan fuerte es la asociación global entre condiciones estructurales y rendimiento (0–1).

#### Bloques detallados (debajo de la línea)

Cada KPI tiene un subtítulo, un texto explicativo (fórmula y propósito) y una tabla con los números:

- KRAS: Tabla con segmentos (Estrato 1–2, Estrato 5–6), P(bajo) y KRAS en pp.
- FD: Segmentos Sin computador / Con computador, P(bajo) y FD en pp.
- KCC: Segmentos Madre baja educación / Madre alta educación, P(bajo) y KCC en pp.
- KDS: Tabla de variables socioeconómicas con su Cramér's V frente al target y el KDS promedio.
- Equidad regional: Tabla con P(bajo) por región (agrupación de departamentos) y la brecha regional en pp; permite ver qué regiones tienen mayor y menor riesgo.
- Impacto territorial:
  - Top 15 departamentos por casos: número de estudiantes (N), P(bajo), casos en bajo rendimiento, exceso en pp y en casos respecto al promedio nacional.
  - Región (casos): N, P(bajo) y casos por región.

Sirve para priorizar: casos = volumen absoluto; exceso = dónde el riesgo es mayor al nacional (carga estructural).

## 5.5 EDA

La página EDA (Análisis Exploratorio de Datos) permite explorar el dataset procesado desde tres enfoques: numérico, categórico y visual. Tiene tres pestañas: Cuantitativo, Cualitativo y Gráficos.

### Pestaña Cuantitativo

- Distribución del target (%): Porcentaje de estudiantes en cada nivel de RENDIMIENTO\_GLOBAL (bajo, medio-bajo, medio-alto, alto). Sirve para ver el equilibrio de clases.
- Resumen estadístico: Estadísticas por variable numérica (media, desviación, percentiles, asimetría, curtosis). Solo columnas numéricas del dataset procesado.
- Pares correlacionados ( $|\text{corr}| > 0,7$ ): Pares de variables numéricas con correlación en valor absoluto mayor a 0,7. Ayuda a detectar multicolinealidad o redundancia.
- Kruskal numéricas vs target: Para variables como RENDIMIENTO\_GLOBAL e INDICADOR\_1 a 4, el p-value del test de Kruskal-Wallis indica si la distribución de la variable numérica difiere entre los niveles de rendimiento. p bajo = asociación significativa.
- Outliers por clase (IQR) — indicadores: Conteo de outliers (método IQR) por nivel de rendimiento para INDICADOR\_1 e INDICADOR\_2. Sirve para ver si hay valores extremos concentrados en algún grupo.

### Pestaña Cualitativo

- Cramér's V (categóricas vs target): Tabla con la asociación (Cramér's V) entre cada variable categórica y RENDIMIENTO\_GLOBAL. Valores altos = mayor asociación con el rendimiento.
- Perfiles categóricos (Top 15 por riesgo): Para las variables F\_ESTRATOVIVIENDA, F\_TIENECOMPUTADOR, F\_TIENEINTERNET y REGION, se muestra una tabla con las categorías ordenadas por riesgo (proporción de bajo rendimiento). Permite ver qué estratos, acceso a tecnología o regiones concentran más riesgo.

## Pestaña Gráficos

- Selector "Seleccione la gráfica": Desplegable con preguntas de negocio (p. ej. "¿Cómo está distribuido el rendimiento académico?", "¿Qué tan grande es la brecha por estrato?", "¿Hay brecha por acceso a computador?", "¿Qué regiones están mejor/peor en riesgo?", "¿Qué departamentos concentran más casos?", etc.). Al elegir una, se dibuja la gráfica correspondiente (barras, riesgo por estrato/región/departamento, burbujas impacto vs riesgo, etc.).
- Interpretación: Cada gráfica responde una pregunta concreta; en la sección 6.2 del manual hay una guía por tipo de gráfica. No deben interpretarse como causalidad.

En conjunto, EDA sirve para profundizar en distribuciones, asociaciones y visualizaciones antes o después de revisar los KPI.

## 5.6 Groq IA

La página Groq IA es un asistente de chat que ayuda a explicar hallazgos, redactar insights para política pública y sugerir análisis adicionales usando un modelo de lenguaje (Groq). Requiere una API key de Groq; si no está configurada en Secrets o variables de entorno, debe ingresarse en el sidebar de esta página.

### Sidebar — Configuración Groq

- Groq API Key: Campo de tipo contraseña donde se ingresa la API key si no está ya configurada. Se usa solo en la sesión actual si no existe en `secrets` o en variables de entorno.
- Modelo: Selector entre `llama-3.1-8b-instant` (más rápido) y `llama-3.3-70b-versatile` (más capaz). Afecta la calidad y la velocidad de las respuestas.
- Incluir contexto del dataset (recomendado): Si está activado, el modelo recibe un resumen del dataset (filas, columnas, % nulos, KPIs calculados: P(nacional), KRAS, FD, KCC, KDS) para que las respuestas se apoyen en los datos cargados. Si se desactiva, el asistente no ve esos números (útil cuando no se desea enviar ningún dato a Groq).

### Área principal — Chat

- Historial: Los mensajes del usuario y las respuestas del asistente se muestran en orden. El sistema usa un rol de "analista senior", pide respuestas en español y sin causalidad (solo asociaciones), con evidencia cuantitativa.
- Cuadro de escritura: Permite escribir una pregunta (ej.: "Redacte 5 insights ejecutivos para MinEducación"). Al enviar, la pregunta se añade al historial y se llama a la API de Groq; la respuesta aparece debajo. Si no hay API key válida, se muestra un error.
- Errores: Si la API falla (HTTP o otro), se muestra el mensaje de error en pantalla.

### Acciones rápidas (prompts)

Tres botones que insertan automáticamente una pregunta en el chat y disparan la respuesta:

Botón	Pregunta que envía
Insights ejecutivos	Redactar 7 insights ejecutivos accionables (no causales) para MinEducación basados en KRAS, FD, KCC, KDS y territorio.
Recomendaciones de política	Proponer 5 líneas de intervención (conectividad, tutorías, focalización territorial) priorizadas por impacto y equidad, justificando con números.
Próximos análisis	Sugerir próximos análisis estadísticos o de ML (sin implementarlos) que complementen el diagnóstico, indicando variables y métricas.

Sirve para obtener texto listo para informes o presentaciones sin escribir el prompt a mano. La privacidad de los datos enviados a Groq se explica en la sección 7.1.

## 6. Ampliaciones del manual

Esta sección desarrolla en detalle: interpretación de KPIs y tablas, guía de los gráficos del EDA, configuración de Groq, glosario, requisitos técnicos y diferencias entre uso en línea y local.

### 6.1 Explicación detallada de cada KPI e interpretación de tablas

Los KPIs del dashboard son asociativos y diagnósticos: miden brechas y asociaciones entre variables socioeconómicas/territoriales y el rendimiento académico, pero no establecen causalidad. Sirven para priorizar intervenciones y monitorear equidad.

KPI 1 — KRAS (Riesgo Académico Social)

- Fórmula:  $KRAS = P(\text{Bajo} \mid \text{Estrato } 1-2) - P(\text{Bajo} \mid \text{Estrato } 5-6)$ , en puntos porcentuales (pp).
- Qué mide: Cuánto mayor es la probabilidad de bajo rendimiento en estudiantes de estratos bajos (1–2) frente a estratos altos (5–6).
- Tabla: Columnas Segmento (Estrato 1–2, Estrato 5–6), P(bajo) (probabilidad de bajo rendimiento en ese segmento), KRAS\_pp (diferencia en pp).
- Interpretación: Si  $KRAS = 20$  pp, en promedio hay 20 puntos porcentuales más de probabilidad de bajo rendimiento en estratos 1–2 que en 5–6. Valores altos indican mayor brecha socioeconómica.

KPI 2 — FD (Fricción Digital)

- Fórmula:  $FD = P(\text{Bajo} \mid \text{Sin computador}) - P(\text{Bajo} \mid \text{Con computador})$ , en pp.
- Qué mide: Brecha de riesgo de bajo rendimiento entre quienes no tienen computador y quienes sí.
- Tabla: Segmento (Sin computador, Con computador), P(bajo), FD\_pp.
- Interpretación:  $FD = 18$  pp significa que no tener computador se asocia con 18 pp más de probabilidad de bajo rendimiento. Útil para priorizar programas de acceso a tecnología.



### KPI 3 — KCC (Capital Cultural — educación materna)

- Qué mide: Diferencia de riesgo entre estudiantes cuya madre tiene baja educación (ninguno, primaria incompleta/completa) y aquellos cuya madre tiene educación profesional o posgrado.
- Tabla: Segmento (Madre baja educación, Madre alta educación), P(bajo), KCC\_pp.
- Interpretación: KCC en pp indica cuánto mayor es el riesgo de bajo rendimiento cuando la educación materna es baja. Refleja asociación con capital cultural familiar, no causa directa.

### KPI 4 — KDS (Dependencia socioeconómica del rendimiento)

- Qué mide: Promedio de Cramér's V entre varias variables socioeconómicas (estrato, computador, internet, automóvil, educación padre/madre, región, departamento) y el rendimiento.
- Escala: Cramér's V entre 0 (sin asociación) y 1 (asociación fuerte). KDS es el promedio de esas asociaciones.
- Tabla: Variable, Cramér's V, KDS\_promedio. Las variables están ordenadas por Cramér's V descendente.
- Interpretación: KDS alto indica que el rendimiento está fuertemente asociado a condiciones socioeconómicas; KDS bajo sugiere que otras variables podrían pesar más. No implica que el estrato "cause" el rendimiento, solo asociación.

### KPI 5 — Equidad regional

- Qué mide: Riesgo de bajo rendimiento por región (agrupación de departamentos). La tabla muestra P(bajo) por región y la brecha regional en pp (diferencia entre la región con mayor y menor riesgo).
- Uso: Identificar regiones con mayor y menor riesgo; la brecha en pp cuantifica la desigualdad territorial.

### KPI 6 — Impacto territorial

- Qué muestra:
- Por departamento: número de estudiantes (N), probabilidad de bajo (p\_bajo), casos en bajo (casos), exceso en pp respecto al promedio nacional (exceso\_pp), exceso en número de casos (exceso\_casos).
- Por región: N, p\_bajo, casos.
- Interpretación: Casos = impacto absoluto (dónde hay más estudiantes en bajo rendimiento). Exceso\_pp y exceso\_casos = carga estructural (dónde el riesgo es mayor al nacional). Priorizar: alto impacto absoluto para volumen de intervención; alto exceso para focalizar donde la brecha es peor.

## 6.2 Guía de interpretación de los gráficos del EDA

En la pestaña Gráficos del EDA se elige una pregunta en el desplegable "Seleccione la gráfica". A continuación, qué muestra cada una y cómo interpretarla.

Gráfica	Pregunta	Qué ver	Cómo interpretar
Distribución del rendimiento	¿Cómo está distribuido el rendimiento académico?	Barras con porcentaje por categoría del target (bajo, medio-bajo, medio-alto, alto).	Proporción de estudiantes en cada nivel; si "bajo" es muy alto, hay una base grande de riesgo.

Gráfica	Pregunta	Qué ver	Cómo interpretar
Brecha por estrato	¿Qué tan grande es la brecha por estrato socioeconómico?	Dos barras: P(bajo) para Estrato 1–2 vs 5–6, y KRAS en pp.	Cuanto más altas la barra de estratos bajos y el valor KRAS, mayor brecha socioeconómica.
Riesgo por estrato	¿Cómo cambia el riesgo a través de los estratos?	Riesgo de bajo rendimiento (eje Y) por estrato (eje X).	Tendencia creciente al subir estrato indica que a mayor estrato, menor riesgo (asociación).
Brecha por computador	¿Hay brecha por acceso a computador?	Dos barras: Sin vs Con computador y FD en pp.	FD positivo: no tener computador se asocia a mayor riesgo.
Brecha computador por estrato	¿La brecha por computador cambia según el estrato?	Interacción: brecha Sin/Con PC en cada estrato.	Si la brecha es mayor en estratos bajos, la falta de computador impacta más ahí (asociación).
Brecha por educación madre	¿Hay brecha por educación de la madre?	Dos barras: madre baja vs alta educación y KCC en pp.	KCC en pp cuantifica la diferencia de riesgo asociada al nivel educativo materno.
Educación materna y riesgo	¿Qué niveles de educación materna tienen mayor riesgo?	Riesgo (o proporción bajo) por categoría de educación de la madre.	Identificar qué niveles tienen mayor riesgo para focalizar mensajes o programas.
Regiones en riesgo	¿Qué regiones están mejor/peor en riesgo?	Barras de P(bajo) por región, ordenadas. Métricas: mejor región, peor región, brecha en pp.	Regiones con barra más alta = mayor riesgo; la brecha en pp mide desigualdad regional.
Top departamentos (casos)	¿Qué departamentos concentran más casos (top)?	Barras con número de casos (estudiantes en bajo rendimiento) por departamento (top 10).	Dónde hay más volumen para intervenir en términos absolutos.
Impacto vs riesgo (burbujas)	¿Cómo priorizar departamentos por impacto vs riesgo?	Gráfico de dispersión/burbujas: eje X e Y (p. ej. riesgo vs casos o exceso), tamaño por N.	Combinar impacto absoluto (casos) y riesgo relativo (exceso) para priorizar departamentos.
Regiones (casos)	¿Qué regiones concentran más casos (impacto absoluto)?	Casos de bajo rendimiento por región.	Similar al top departamentos pero por región.
Departamentos en riesgo	¿Qué departamentos están mejor/peor en riesgo?	P(bajo) por departamento. Métricas: mejor y peor departamento, brecha en pp.	Identificar departamentos con mayor y menor riesgo para políticas territoriales.

## 6.3 Configuración de la API de Groq

La página Groq IA permite hacer consultas en lenguaje natural sobre el proyecto y los datos. Para usarla hace falta una API key de Groq.

### Obtener la API key

1. Se entra en Groq Console (registro o inicio de sesión).
2. Se crea o selecciona un proyecto y se accede a la sección de API keys.
3. Se genera una nueva API key y se copia (solo se muestra una vez).

### Dónde ingresar la key en el dashboard

- En la página Groq IA, en el sidebar aparece "Configuración Groq".
- Si no hay key configurada, se muestra el campo "Groq API Key" (tipo contraseña). Ahí se ingresa la key; se usa solo en esa sesión.

### Opciones en la página

- Modelo: se puede elegir entre `llama-3.1-8b-instant` (rápido) y `llama-3.3-70b-versatile` (más capaz).
- Incluir contexto del dataset: si está activado, el asistente recibe un resumen del dataset (filas, columnas, nulos, KPIs calculados) para responder con base en los datos cargados.

### Seguridad

- No se debe compartir la API key ni subirla a repositorios públicos.

### Uso

- Se escriben preguntas en el cuadro de chat (ej.: "Redacte 5 insights ejecutivos para MinEducación").
- Los botones Insights ejecutivos, Recomendaciones de política y Próximos análisis envían prompts predefinidos para obtener respuestas orientadas a política educativa y análisis complementarios.

## 7. Posibles problemas y soluciones

Situación	Qué hacer
La app no carga o tarda mucho	La primera carga procesa todo el CSV y puede tardar varios segundos; es normal. Si se usa la app en línea, se recomienda comprobar la conexión.
Mensaje "El archivo cargado se llama X"	El archivo debe llamarse saber_pro.csv. Se debe renombrar y volver a subir.
Error al subir el archivo	Se debe verificar que sea un CSV válido (separador coma, codificación UTF-8). Debe corresponder al dataset Saber Pro (2018-2021) con las columnas esperadas.
Groq IA no responde	Se debe revisar que se haya ingresado una API key de Groq válida en la página "Groq IA".