

Análisis Exploratorio de la Encuesta Mundial de Salud Escolar (EMSE 2018)

Primera Exposición Grupal – Taller de Programación (UBA)

Julián Delgadillo Marín

Alejandro Alcocer

Christian Campos

Maestría en Economía Aplicada

Universidad de Buenos Aires

Miércoles 5 de noviembre de 2025

Estructura de la presentación

- 1 1. Fuente e institución
- 2 2. Nivel y representatividad
- 3 3. Limpieza y estructura de la base
- 4 4. Variables analizadas
- 5 5. Resultados descriptivos
- 6 6. Ideas de investigación
- 7 7. Proyección hacia la Aplicación Final
- 8 8. Cierre

Fuente e institución

- **Base de datos:** 3º Encuesta Mundial de Salud Escolar (EMSE 2018), Argentina.
- **Institución responsable:** Ministerio de Salud de la Nación – Secretaría de Acceso a la Salud, Subsecretaría de Estrategias Sanitarias, Dirección Nacional de Abordaje Integral de Enfermedades No Transmisibles.
- **Coordinación técnica:** Área de Vigilancia de la Salud y Monitoreo de Programas. Responsables: *Dra. Julieta Rodríguez Cámara* y *Dra. María Graciela Abriata*.
- **Objetivo principal:** Generar información precisa sobre comportamientos de salud, factores de riesgo y de protección en adolescentes escolarizados, para orientar políticas públicas y estrategias sanitarias.
- **Cobertura:** Representatividad nacional y provincial, con estudiantes de 13 a 17 años pertenecientes a escuelas de nivel medio.
- **Edición:** Tercera aplicación en el país (ediciones previas: 2007 y 2012).

Nivel y representatividad

- **Nivel de análisis:** Individual (estudiantes escolarizados de 13 a 17 años).
- **Diseño muestral:** Probabilístico, bietápico y estratificado, con selección de escuelas proporcional a la matrícula y selección sistemática de clases dentro de cada escuela.
- **Cobertura:** Representatividad nacional y provincial, abarcando instituciones públicas y privadas de nivel medio.
- **Tamaño muestral:** 600 escuelas seleccionadas; participaron 513 (86 %), con un total de 57.095 estudiantes encuestados y 56.981 cuestionarios válidos.
- **Tasa de respuesta general:** 63 % (86 % en escuelas \times 74 % en estudiantes).
- **Período de campo:** Mayo a diciembre de 2018.
- **Ponderación:** Cada cuestionario tiene un peso asociado que corrige diferencias de probabilidad y no respuesta, garantizando la representatividad de los resultados.
- **Instrumento:** Cuestionario estructurado con 72 preguntas y 81 variables derivadas, agrupadas en 12 módulos temáticos (alimentación, actividad física, salud mental, consumo, etc.).

Limpieza y estructura de la base

- **Carga y exploración inicial:** se importó el archivo `EMSE_DatosAbiertos.csv` y se verificaron dimensiones, tipos de variables y primeras filas.
- **Corrección de codificación:** se detectó el *encoding* real con `chardet` y se aplicó la función `fix_encoding()` para corregir caracteres mal decodificados (acentos y eñes).
- **Tratamiento de valores faltantes:** se identificaron mediante un *heatmap* y se filtraron las variables con menos del 20 % de datos ausentes, generando el subconjunto `df_reducido`.
- **Integración con el libro de códigos:** se automatizó la detección de columnas que contienen los nombres de variables y las preguntas, generando un **diccionario limpio** (*Variable*, *Descripción*) con las variables efectivamente presentes en la base.
- **Selección de variables clave:** se construyó un subconjunto temático de 10 pares de variables (`qX / texto_qX`), renombradas para análisis legible: *Edad*, *Sexo*, *Grado*, *Actividad física*, *Consumo de frutas*, *Lesiones*, *Tabaco*, *Amigos cercanos*, *Alcohol*, *Educación de los padres*.
- **Resultado:** un **DataFrame estructurado y limpio**, listo para análisis descriptivo y visualizaciones.

Variables analizadas

- **Edad:** distribución etaria de estudiantes (13–17 años).
- **Sexo:** identificación de género autorreportada.
- **Actividad física:** frecuencia semanal de ejercicio o movimiento activo.
- **Consumo de frutas:** número de días con ingesta de frutas en la última semana.
- **Tabaco:** consumo de cigarrillos en los últimos 30 días.
- **Alcohol:** frecuencia de consumo de bebidas alcohólicas.
- **Lesiones:** ocurrencia de lesiones graves en los últimos 12 meses.
- **Amigos cercanos:** red de apoyo social entre pares.
- **Educación de los padres:** nivel educativo máximo alcanzado por los progenitores.

Criterio de selección

Se seleccionaron variables representativas de tres dimensiones: *caracterización sociodemográfica, hábitos saludables y conductas de riesgo*, priorizando aquellas con baja proporción de valores faltantes y relevancia para el análisis exploratorio.

Resultados descriptivos

- **Cobertura de datos:** todas las variables presentan una completitud superior al 99.8 %, sin pérdida significativa de información.
- **Edad:** 8 categorías entre 13 y 17 años; moda en 14 años.
- **Sexo:** 3 categorías, predominio de femenino.
- **Grado escolar:** 6 categorías, moda en 2º año / 11º grado nivel Polimodal.
- **Hábitos saludables:**
 - **Actividad física:** 6 categorías; la más frecuente es "Nunca".
 - **Consumo de frutas:** 9 categorías; moda "0 días" durante la última semana.
- **Conductas de riesgo:**
 - **Tabaco:** 6 categorías; moda "Nunca".
 - **Alcohol:** 8 categorías; moda "0 días".
 - **Lesiones:** 9 categorías; moda "Ninguna".
- **Factores de protección:**
 - **Amigos cercanos:** 5 categorías; moda "3 o más".
 - **Educación de los padres:** moda en "Secundario completo".

Síntesis

Las variables presentan alta calidad y coherencia interna. Se observa una tendencia general a la baja práctica de hábitos saludables y escasa frecuencia de consumo de alcohol o tabaco declarada.

Distribución por sexo

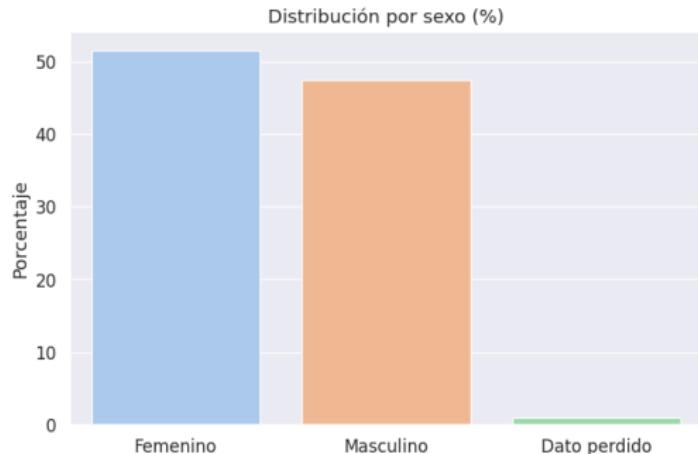


Figura 1: Distribución porcentual del sexo declarado por los estudiantes.

- La muestra presenta una proporción equilibrada entre sexos:
 - **Femenino:** 51.3 %
 - **Masculino:** 47.5 %
 - **Dato perdido:** ~1 %
- No se observan sesgos significativos de representación entre grupos.
- Refleja la composición típica de población escolar media en Argentina.

Distribución de edades

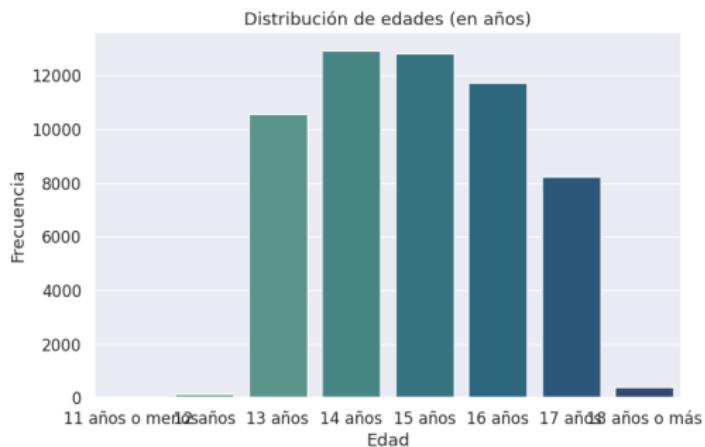


Figura 2: Distribución de edades (en años) – EMSE 2018

- La muestra incluye adolescentes de entre **13 y 17 años**, con presencia marginal de casos fuera del rango.
- Las edades **14 y 15 años** concentran la mayor frecuencia, representando el núcleo principal de la población encuestada.
- La distribución muestra una forma unimodal y coherente con la estructura del sistema educativo medio argentino.
- Estos resultados confirman la adecuación etaria de la encuesta al grupo objetivo definido por la OMS y el Ministerio de Salud.

Nivel educativo de los padres

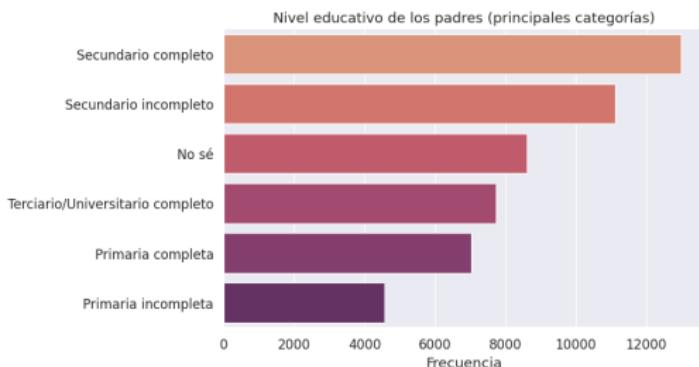


Figura 3: Nivel educativo de los padres – Principales categorías

- La mayoría de los estudiantes reporta que sus padres alcanzaron como máximo **Secundario completo** (categoría modal).
- Le siguen las categorías **Secundario incompleto** y “**No sé**”, lo que sugiere cierta brecha informativa o desconocimiento sobre el nivel educativo familiar.
- Los niveles **terciario/universitario completo** y **primaria completa** presentan frecuencias menores.
- En conjunto, los datos reflejan un contexto educativo familiar medio–bajo, coherente con la composición general del sistema educativo público.

Hábitos saludables

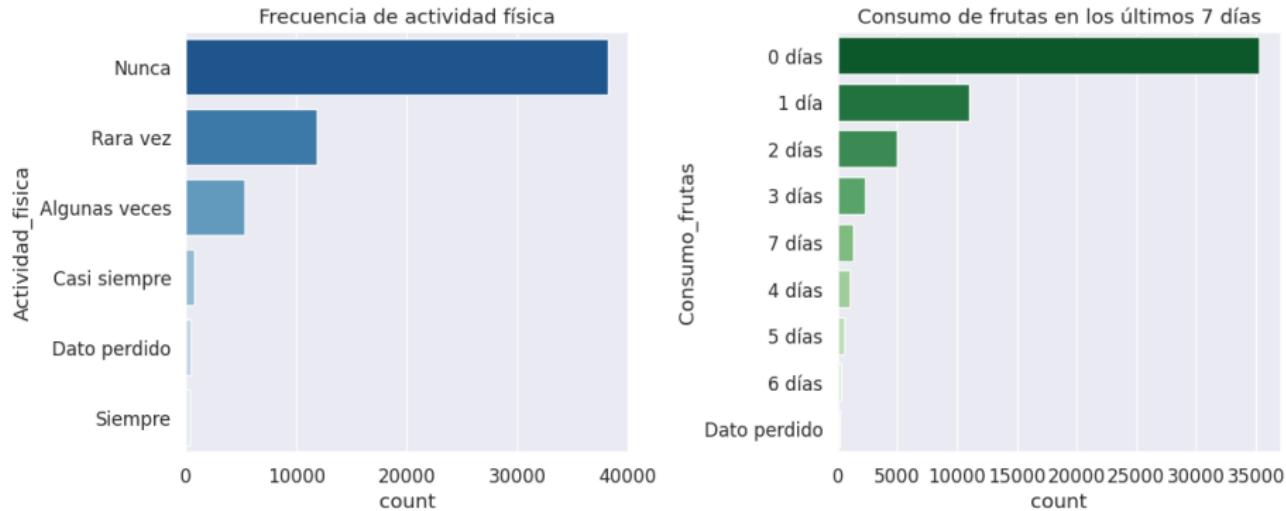


Figura 4: Frecuencia de actividad física y consumo de frutas – Últimos 7 días.

- La **actividad física** muestra una tendencia marcadamente baja: más del 60 % de los estudiantes declara “Nunca” realizar actividad física.
- El **consumo de frutas** también evidencia hábitos poco saludables: la categoría modal es “0 días”, seguida de “1 día”.
- Ambas variables revelan una baja incorporación de prácticas saludables, con escasa frecuencia de ejercicio y alimentación equilibrada.

Conductas de riesgo: tabaco y alcohol

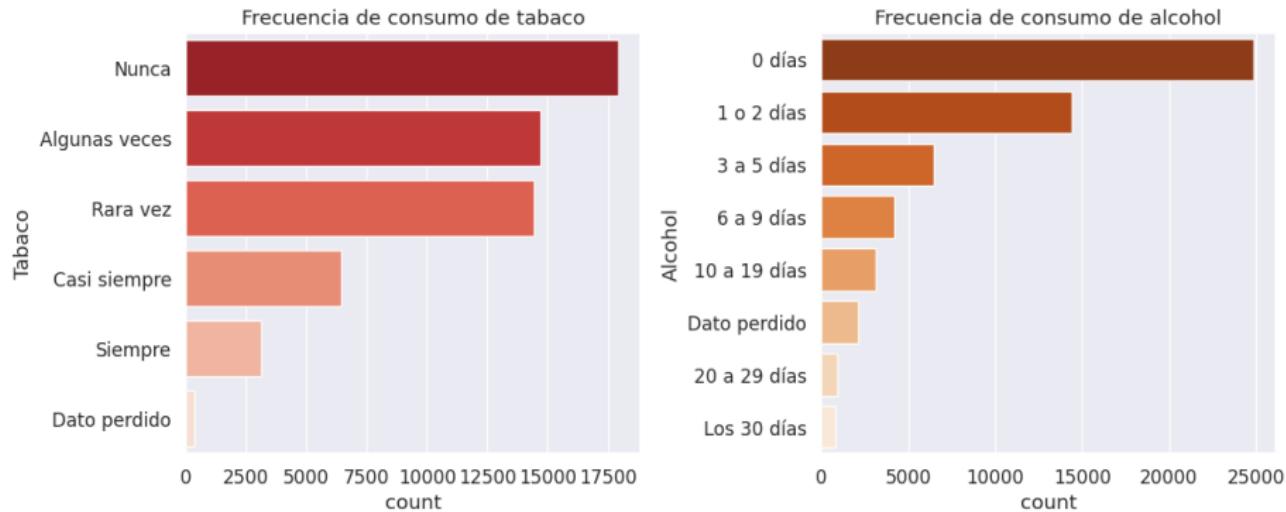


Figura 5: Frecuencia de consumo de tabaco y alcohol – Últimos 30 días.

- El **consumo de tabaco** presenta una distribución concentrada en la categoría “Nunca”, seguida de “Rara vez” y “Algunas veces”.
- En cuanto al **consumo de alcohol**, la mayoría reporta “0 días” en el último mes, aunque existe una fracción relevante (25 %) que declara consumo entre 1 y 9 días.
- Ambos indicadores confirman una baja prevalencia general de consumo

Ideas de investigación

① Relación entre el nivel educativo de los padres y las conductas de riesgo:

Analizar si un mayor nivel educativo parental se asocia con menor prevalencia de consumo de tabaco y alcohol en adolescentes.

② Asociación entre actividad física y consumo de sustancias:

Explorar la posible relación inversa entre la frecuencia de actividad física y la probabilidad de consumir tabaco o alcohol.

③ Diferencias regionales en hábitos saludables:

Evaluando la existencia de brechas territoriales en alimentación y ejercicio físico, considerando el tipo de escuela (urbana/rural, pública/privada).

④ Apoyo social y factores protectores:

Examinar si contar con más amigos cercanos o mayor supervisión familiar reduce la probabilidad de conductas de riesgo.

Estas son posibles hipótesis que pueden ser desarrolladas en el borrador analítico de tres páginas y en el póster final del taller.

Uso previsto de la base EMSE 2018 en la Aplicación Final

- En la segunda instancia del curso, el grupo trabajará nuevamente con la base de microdatos de la Encuesta Mundial de Salud Escolar (EMSE 2018), utilizando una herramienta de análisis aprendida en clase.
- **Herramienta de análisis:** modelo de clasificación supervisado (Logit o KNN) para explorar la relación entre variables sociodemográficas, hábitos saludables y conductas de riesgo.
- **Objetivo:** evaluar cómo el nivel educativo de los padres, la actividad física y el consumo de frutas se asocian con la probabilidad de que un estudiante haya consumido alcohol o tabaco en los últimos 30 días.
- **La aplicación incluirá:**
 - Visualizaciones interactivas y comparativas de predicción (curva ROC y matriz de confusión).
 - Validación cruzada para estimar el desempeño del modelo.
 - Discusión sobre los resultados, sus limitaciones y posibles implicancias para políticas de salud escolar.
- En conjunto, la base EMSE 2018 funcionará como caso aplicado, integrando limpieza, visualización y modelado predictivo en línea con los objetivos del Taller de Programación.

Conclusiones y próximos pasos

- Se logró construir una base de datos depurada, documentada y lista para análisis, integrando la información original del Ministerio de Salud con el libro de códigos oficial.
- El análisis exploratorio permitió identificar patrones relevantes en los hábitos de salud adolescente y establecer dimensiones clave para futuras investigaciones.
- Los resultados iniciales muestran una baja frecuencia de prácticas saludables (actividad física, consumo de frutas) y una prevalencia moderada de conductas de riesgo (tabaco y alcohol).
- Las variables sociodemográficas y familiares (edad, sexo, educación parental) ofrecen un marco sólido para el desarrollo de modelos explicativos y comparativos.
- **Próximos pasos:** elaboración de visualizaciones interactivas, análisis bivariado y modelado descriptivo para las hipótesis planteadas.

¡Gracias por su atención!

Taller de Programación – Maestría en Economía Aplicada, UBA (2025)