Luis Alejandro Azúa Urrutia
20/02/2026

Reporte de Análisis exploratorio para Prevención Cardiovascular.

Presentado por:

Luis Alejandro Azua Urrutia

Informe de hallazgos

Objetivos:

Identificar factores de salud, así como demográficos que pudieran influir en la presencia de enfermedades del corazón, concentrándose en la población de México. Con esta información se busca mejorar y optimizar las campañas de salud pública.

Conjunto de datos y variables:

El conjunto de datos “heart_2020_cleaned.csv” es una base de datos del Censo Nacional de Población y Vivienda 2020 del INEGI. Consta de 3119795 (filas) y 18 variables (columnas) relacionadas con el estilo de vida y salud de la población. En esta se muestra “HeartDisease” (variable objetivo) y otros factores que podrían tener correlación con su presencia.

A continuación, se presenta una muestra de los datos en formato de tabla, así como una breve descripción de todas las variables que se encuentran en ellos.

Luis Alejandro Azúa Urrutia
20/02/2026

Muestra de datos "heart_2020_cleaned.csv"

	Sujeto 1	Sujeto 2	Sujeto 3	Sujeto 4	Sujeto 5
HeartDisease	No	No	No	No	No
BMI	16.6	20.34	26.58	24.21	23.71
Smoking	Yes	No	Yes	No	No
AlcoholDrinking	No	No	No	No	No
Stroke	No	Yes	No	No	No
PhysicalHealth	3.0	0.0	20.0	0.0	28.0
MentalHealth	30.0	0.0	30.0	0.0	0.0
DiffWalking	No	No	No	No	Yes
Sex	Female	Female	Male	Female	Female
AgeCategory	55-59	80 or older	65-69	75-79	40-44
Race	White	White	White	White	White
Diabetic	Yes	No	Yes	No	No
PhysicalActivity	Yes	Yes	Yes	No	Yes
GenHealth	Very good	Very good	Fair	Good	Very good
SleepTime	5.0	7.0	8.0	6.0	8.0
Asthma	Yes	No	Yes	No	No
KidneyDisease	No	No	No	No	No
SkinCancer	Yes	No	No	Yes	No

HeartDisease: Indica si el paciente fue diagnosticado con enfermedad de corazón.

BMI: Índice de masa corporal.

Smoking: Indica si ha fumado al menos 100 cigarrillos o más en su vida.

AlcoholDrinking: Consumidor frecuente de alcohol.

Stroke: Si el paciente fue informado por medico de infarto cerebral.

PhyshicalHealth: Lesiones o enfermedades físicas en los últimos 30 días.

MentalHealth: Estrés, depresión o problemas emocionales en los últimos 30 días.

DiffWalking: Si el paciente tiene problemas serios al caminar.

Sex: Genero biológico

AgeCategory: Rango de edad del encuestado

Race: Identificación de grupo étnico

Diabetic: Si el paciente ha sido diagnosticado con diabetes o pre-diabetes.

PhysicalActivity: Si en el último mes realizo actividad física.

Genhealth: Evaluación personal de estado de salud.

SleepTime: Promedio de horas dormidas en 24 horas.

Asthma: Si alguna vez fue diagnosticado con asma.

KidneyDisease: Si el paciente ha padecido alguna enfermedad de los riñones.

SkinCancer: Si el paciente fue diagnosticado con cáncer de piel**.**

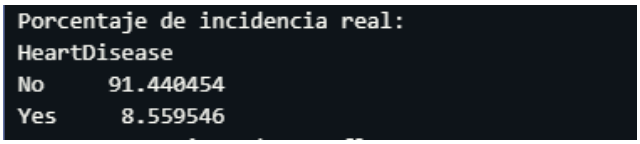
Luis Alejandro Azúa Urrutia
20/02/2026

Principales hallazgos desglosados para:

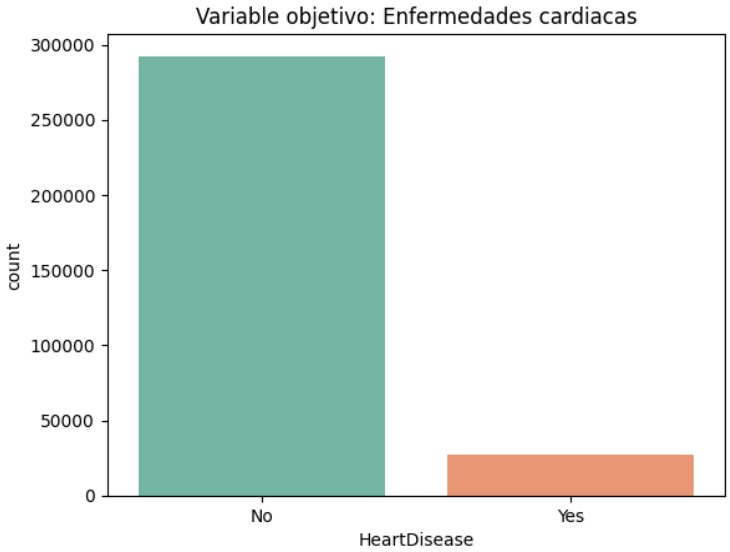
Análisis univariado:

Graficando los datos proporcionados observamos que muestra una población mayoritariamente sana, pero con un segmento que ya tiene un historial de enfermedades y afecciones cardiacas.

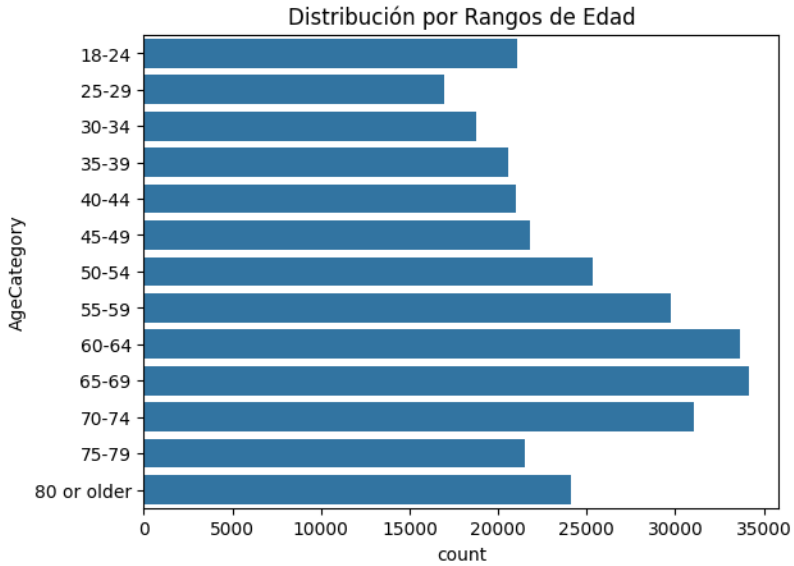
El **91.4%** de los encuestados parecen estar **sanos** y no padecer este tipo de enfermedad. Como se identificó un desbalance significativo dentro de la variable objetivo, pues solo **un 8.5%** de los participantes **muestran afecciones cardiacas**, el modelo de Machine Learning que buscamos implementar deberá ser ajustado para evitar sesgos y garantizar que se detecte de manera oportuna en este segundo grupo de alto riesgo.



Análisis de variable objetivo en porcentaje.



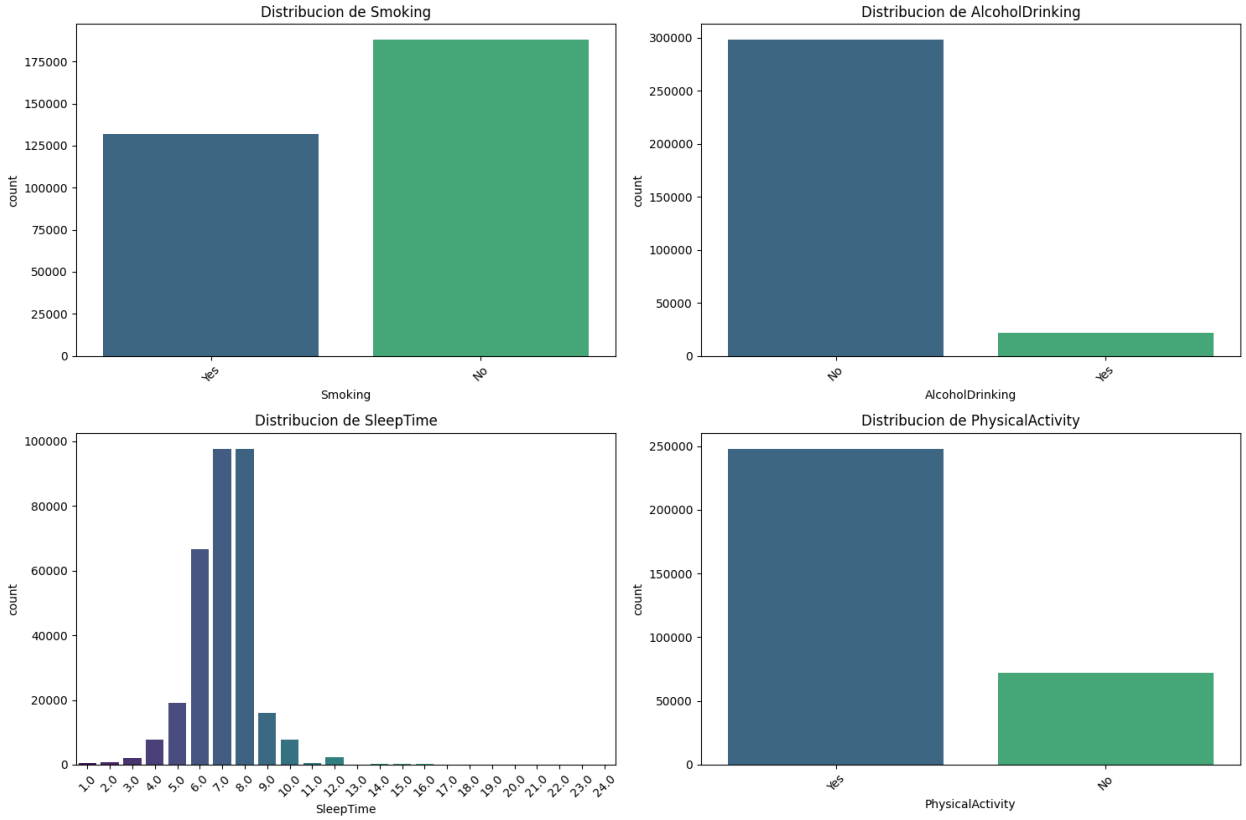
Otro factor importante a mencionar es la edad de los participantes. La distribución de edad nos deja observar que una parte muy grande se concentra en rangos de adultos mayores, de 55 años en adelante. Este factor será determinante en los resultados que veremos al relacionarlo con nuestra variable objetivo.



Distribución por rangos de edad (AgeCategory)

Al hacer un análisis en variables especificas relacionadas con el estilo de vida y salud general, podemos apreciar que la población tiene una

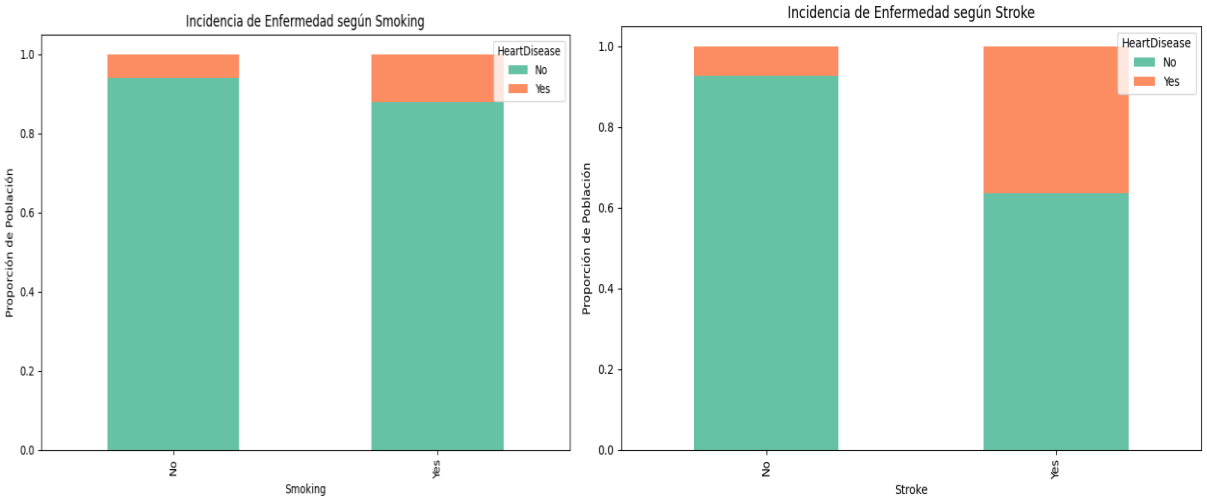
Luis Alejandro Azúa Urrutia
20/02/2026



Las gráficas de arriba muestran la distribución de variables relacionadas con estilo de vida y salud. Por si solas podrían mostrar que la mayoría evita hábitos que puedan ocasionar daños a la salud como beber o fumar, o dormir pocas horas, sin embargo, al relacionarlas con otras variables descubrimos algo diferente.

Análisis bivariado:

En esta sección realizaremos un análisis comparando nuestra variable objetivo (HeartDisease) con otras variables que podrían influir o no en esta.

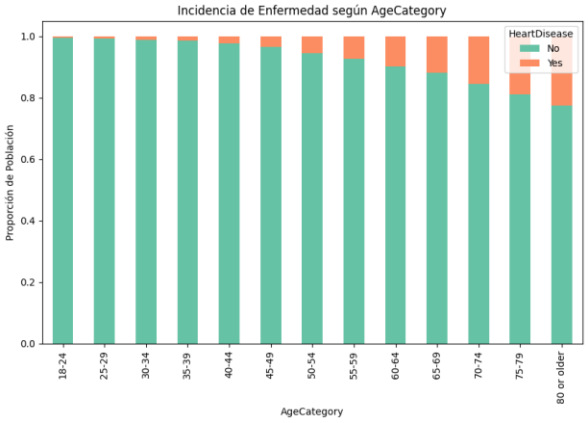


Luis Alejandro Azúa Urrutia
20/02/2026

Haciendo las primeras comparaciones, podemos observar como el hábito de fumar podría ser un factor a considerar en relacion con las enfermedades cardiacas, pues los casos aumentan en la población que ha fumado más de 100 cigarrillos en toda su vida.

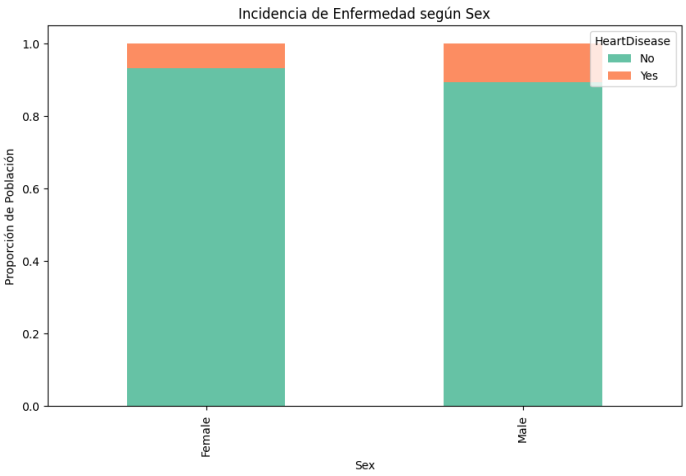
Es muy importante resaltar los resultados de la segunda gráfica, en la que comparamos las incidencias de enfermedad una vez que la persona ha presentado infartos cerebrales previamente. Este es un factor de extrema importancia. La posibilidad de padecer esta enfermedad se triplica en comparación con gente si este antecedente, sugiriendo que el sistema vascular ya fallo previamente y aumenta la vulnerabilidad del corazón.

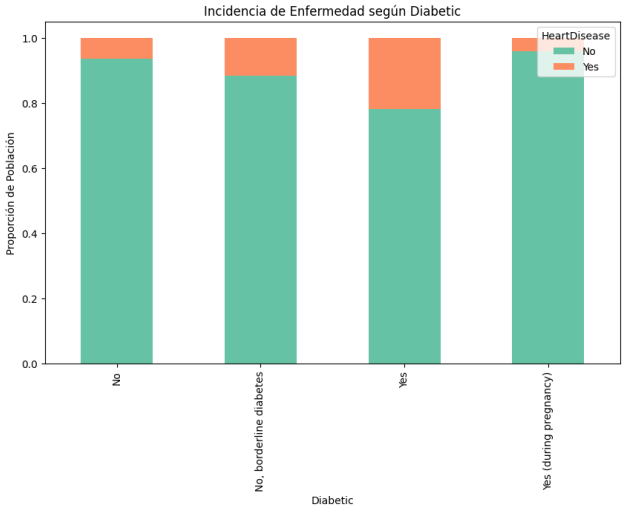
Otro factor determinante es la edad(AgeCategory), pues los datos muestran un aumento de riesgo exponencial de padecer afecciones cardiacas a partir de los 60 años aproximadamente.



Si observamos la tabla, se ve a simple vista que las incidencias van en considerable aumento conforme la persona envejece.

Al realizar un análisis con la variable de genero biológico, podemos apreciar también un factor interesante. Este factor biológico determina si la campaña deberá dirigirse de manera más agresiva al público masculino considerando que a pesar de los datos equilibrados entre hombre y mujer, podemos notar que la incidencia de enfermedad es más visible en el sector masculino.



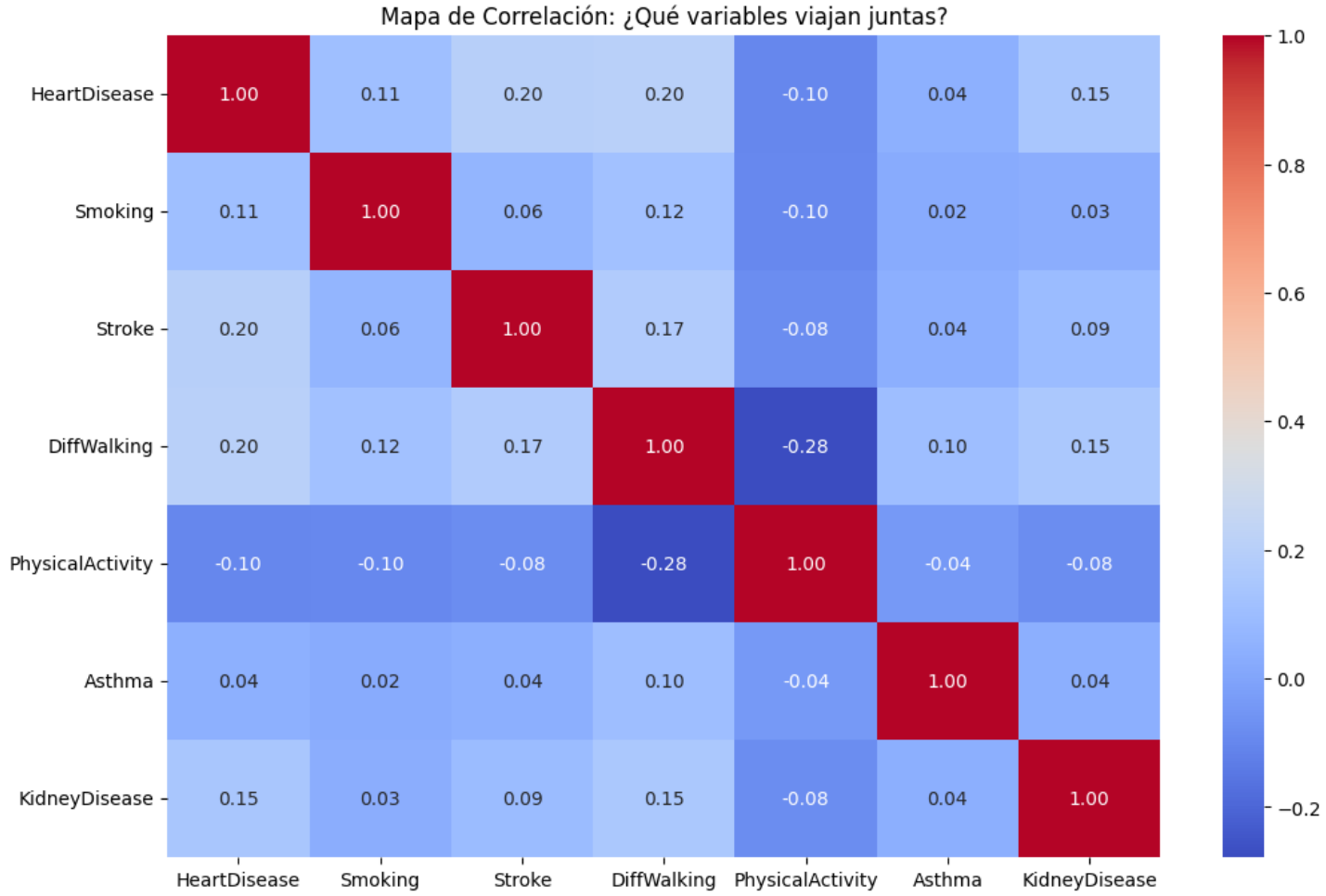


Por último, tras analizar la variable de diabetes confirmamos una estrecha correlación entre diabetes y afecciones al corazón. El diagnostico de diabetes actúa como un multiplicador, elevando los casos positivos de una enfermedad cardiaca drásticamente.

Luis Alejandro Azúa Urrutia
20/02/2026

Análisis multivariado:

Para finalizar el estudio implementamos un análisis multivariado de correlación, utilizando un Heatmap, con el fin de identificar como cada variable es capaz de afectar al corazón, así como la interacción que se presenta entre ellas.



Las 7 variables que se presentan en la gráfica fueron seleccionadas porque representan tres factores de riesgo a considerar. Los antecedentes clínicos, capacidad de movilidad y el estilo de vida. Para medir matemáticamente la relacion tuvimos que convertirlos en valores numéricos (0 y 1).

En la gráfica podemos observar como la variable HeartDisease y Stroke tienen una muy alta correlación, siendo .20 uno de los números más altos y confirmando el hecho de que dificultades vasculares previas tienden a ser una clara bandera roja y convirtiéndolo en un perfil de muy alto riesgo. Otro de los factores que podemos apreciar con más facilidad gracias a la gráfica es la correlación que tienen las enfermedades de corazón con la perdida de movilidad. En la variable DiffWalking vemos como hay números altos tanto en HeartDisease como en Stroke, pues la perdida de movilidad está fuertemente ligada con estas enfermedades. De lo contrario, vemos como la actividad física tiene una correlación negativa con las enfermedades podemos reforzar el hecho de que la movilidad tiene mucho que ver y se convierte en una importante defensa.

Luis Alejandro Azúa Urrutia
20/02/2026

En conclusión:

El análisis de los valores matemáticos encontrados en los datos nos permite concluir que el perfil de riesgo más alto que la Secretaria de Salud debe considerar, es un hombre desde 50 años de edad, que presente antecedentes de infarto cerebral, diabetes y baja movilidad.

La siguiente fase del proyecto será considerar los datos para entrenamiento de un modelo de regresión logística. Se deberán considerar los niveles de desbalance de datos, entre personas con enfermedades cardiacas y personas sanas, pues el porcentaje de 8.5% de incidencia es muy bajo. Se tendrán que realizar ajustes en métricas de sensibilidad para no omitir casos positivos.