

# Taller - Explorando la Salud de EE. UU. con Datos de NHANES

Por: Ing. Ricardo Naranjo Faccini, M.Sc.

[gerencia@skinait.com](mailto:gerencia@skinait.com)

<https://skinait.com/ricardo-naranjo-faccini/cms/>

## Sumario

Taller - Explorando la Salud de EE. UU. con Datos de NHANES.....	1
1. Objetivo general:.....	1
2. Metodología.....	2
2.a. Grupos.....	2
2.b. Desarrollo.....	2
2.c. Entrega.....	2
2.d. Herramientas permitidas:.....	2
2.e. Rol del equipo:.....	2
3. Fases del Taller.....	3
3.a. Exploración y Familiarización con los Datos.....	3
3.b. Preparación y Limpieza de los Datos.....	3
3.c. Construcción de la Vista Minable.....	3
3.d. Análisis y Visualización de los Datos.....	4
3.e. Presentación de Resultados.....	4
4. Evaluación.....	5
5. Licencia.....	5

---

## 1. Objetivo general:

Guiar a los estudiantes en el proceso completo de familiarización, limpieza, preparación, análisis y visualización de datos usando el conjunto de datos de NHANES 2009-2012, con el fin de extraer conclusiones útiles para la salud pública y generar hipótesis predictivas.

## 2. Metodología

### 2.a. Grupos

Los estudiantes se organizarán en los mismos grupos de 3 personas que han trabajado en la clase presencial para desarrollar el taller.

### 2.b. Desarrollo

El taller se realizará fuera del horario de clase y tendrán 15 minutos en clase para exponer frente a sus compañeros los resultados obtenidos y el procedimiento para llegar a ellos.

### 2.c. Entrega

La exposición y entrega se realizará en la clase del 15 de agosto.

### 2.d. Herramientas permitidas:

- Hoja de cálculo (Excel, Google Sheets, etc.)
- Lenguaje Python (recomendado: pandas, matplotlib/seaborn, sklearn)
- Software Orange (visual, sin programación)

### 2.e. Rol del equipo:

Cada grupo será un **equipo de analistas de salud pública** contratado por el CDC para investigar patrones de salud en la población estadounidense. Su tarea es preparar, analizar y presentar un informe que explore los datos, destaque hallazgos clave y proponga posibles factores predictivos de salud.

## 3. Fases del Taller

### 3.a. Exploración y Familiarización con los Datos

#### 1. Explora el conjunto de datos.

- ¿Cuántas filas y columnas hay?
- ¿Qué representan las variables?
- Identifica al menos **5 variables numéricas** y **2 categóricas**.
- Apoyandote en <https://www.cdc.gov/nchs/nhanes/index.html> construye el diccionario de datos (inventario de información).
- Realizar el análisis exploratorio de los datos, Mínimo incluir:
  - Tabla resumen
  - Métricas de tendencia central
  - Métricas de dispersión
  - Visualización de métricas de dispersión
  - Análisis de las visualizaciones
  - Análisis de Correlación
  - Recuerde incluir análisis de todo lo visualizado y no solo el gráfico

#### 2. Entregable:

- Descripción del conjunto de datos.
- Diccionario de datos.

### 3.b. Preparación y Limpieza de los Datos

1. Reconocimiento y tratamiento de atributos con valores únicos o distintos
2. Reconocimiento y tratamiento de atributos con valores faltantes
  - Identifica columnas con valores faltantes.
  - Describe qué estrategia aplicarán (eliminación, imputación, etc.) y por qué.
3. Reconocimiento y tratamiento de atributos con valores atípicos/incorrectos
4. Reconocimiento y tratamiento de registros atípicos
  - Asegúrate de que variables numéricas y categóricas estén correctamente tratadas.

5. Reconocimiento y tratamiento de atributos redundantes
6. Filtrado y normalización (si aplica).
  - Detecta y trata valores extremos (outliers).
  - Escala o transforma columnas si van a ser comparadas directamente.
- Entregable:
  - Archivo de datos limpio (.csv).
  - Informe de las decisiones tomadas.

### 3.c. Construcción de la Vista Minable

1. Define tres objetivos de análisis.
  - Por ejemplo:
    - “¿Qué variables se correlacionan más con la obesidad?”
    - “¿Podemos predecir la presión arterial a partir de hábitos alimenticios?”
2. Selecciona las variables más relevantes para cada objetivo.
  - Puedes usar estadísticas descriptivas, correlaciones, o simplemente hipótesis razonables.
3. Genera una “vista minable” para cada objetivo:
  - Mínimo incluir:
    - Generación de variables derivadas tipo 1 y 2
    - Normalización de al menos un atributo
    - Discretización de al menos un atributo
    - Numerización 1 a n de al menos un atributo
  - Un subconjunto del dataset con las variables necesarias para el análisis (puede incluir etiquetas si hay tareas de clasificación o regresión).
4. Entregable:
  - 3 Vistas minable guardada como archivo (.csv).
  - Justificación breve de las variables incluidas.

### 3.d. Análisis y Visualización de los Datos

1. Análisis exploratorio:
  - Usa gráficos: histogramas, cajas (boxplots), dispersión, mapas de calor, etc.
  - Compara grupos (ej.: por género, edad, condición médica).

**2. Obtén correlaciones relevantes.**

- ¿Qué relaciones encuentran entre las variables?
- ¿Se pueden identificar patrones preocupantes?

**3. Predicción:**

- Prueba un modelo predictivo simple (regresión lineal, árbol de decisión, etc.) para predecir alguna variable clave (como IMC, colesterol, presión arterial).

**4. Entregable:**

- Imágenes de las visualizaciones con pie de figura.
- Tabla de correlaciones principales.
- Resultados del modelo predictivo.

### **3.e. Presentación de Resultados**

**1. Prepara una mini-presentación o infografía.**

- Expón los hallazgos más interesantes.
- Incluye visualizaciones claras.
- Concluye con una recomendación basada en datos (por ejemplo, una política de salud, o un grupo de riesgo identificado).

**2. Entregable final:**

- 1 documento PDF con:
  - Resumen ejecutivo.
  - Vista minable final.
  - Principales gráficos y análisis.
  - Conclusión.
  - Propuesta de modelo predictivo.
- Adjuntar todos los archivos generados: .csv, notebooks .ipynb, hojas de cálculo, capturas de Orange, etc.

---

## **4. Evaluación**

Criterio	Puntaje
Limpieza y preparación de datos	20 pts

Criterio	Puntaje
Construcción de vista minable	15 pts
Análisis y visualizaciones	30 pts
Creatividad y profundidad	15 pts
Presentación y claridad	20 pts
Total	100 pts

---

## 5. Licencia



Diseño de bases de datos relacionales: Normalización está bajo una [licencia de Creative Commons Reconocimiento-CompartirIgual 4.0 Internacional](https://creativecommons.org/licenses/by-sa/4.0/).