**Análisis de Base de Datos**

**Héctor Fabio Londoño Arbeláez**

**Objetivo:**  
Cada grupo deberá seleccionar una base de datos, realizar un análisis exploratorio básico, comprender su estructura y contenido, y presentar un informe con los hallazgos más relevantes.

Cada grupo debe realizar las siguientes tareas con la base de datos seleccionada:

1. **Selección de la Base de Datos**

* Nombre de la base de datos: CVD\_cleaned.csv
* Fuente (URL, repositorio, etc.):

https://www.kaggle.com/datasets/alphiree/cardiovascular-diseases-risk-prediction-dataset?resource=download

* Formato del archivo (.csv, .xlsx.): .csv

1. **Estructura de la Base de Datos**

* Número total de filas (registros): 308.854
* Número total de columnas (atributos o variables): 19

1. **Descripción de las Columnas**

Por cada columna se debe especificar:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Nombre de la columna** | **Tipo de dato** | **Descripción** | **Cantidad de datos** | **Valores únicos (si aplica)** |
| General\_Health | Texto | Salud General? | 308.854 |  |
| Checkup | Texto | Chequeo | 308.854 |  |
| Exercise | Booleano | Ejercicio | 308.854 | Si / No |
| Heart\_Disease | Booleano | Cardiopatía | 308.854 | Si / No |
| Skin\_Cancer | Booleano | Cáncer de Piel | 308.854 | Si / No |
| Other\_Cancer | Booleano | Otro Cáncer | 308.854 | Si / No |
| Depression | Booleano | Depresión | 308.854 | Si / No |
| Diabetes | Booleano | Diabetes | 308.854 | Si / No / Other |
| Arthritis | Booleano | Artritis | 308.854 | Si / No |
| Sex | Booleano | Género | 308.854 | Female / Male |
| Age\_Category | Texto | Edad | 308.854 |  |
| Height\_(cm) | Float | Altura (Cm) | 308.854 |  |
| Weight\_(kg) | Float | Peso (Kg) | 308.854 |  |
| BMI | Float | IMC | 308.854 |  |
| Smoking\_History | Booleano | Historial de Tabaco | 308.854 | Si / No |
| Alcohol\_Consumption | Float | Consumo Alcohol | 308.854 |  |
| Fruit\_Consumption | Float | Consumo de Frutas | 308.854 |  |
| Green\_Vegetables\_Consumption | Float | Consumo de Verduras | 308.854 |  |
| FriedPotato\_Consumption | Float | Consumo de Papas Fritas | 308.854 |  |

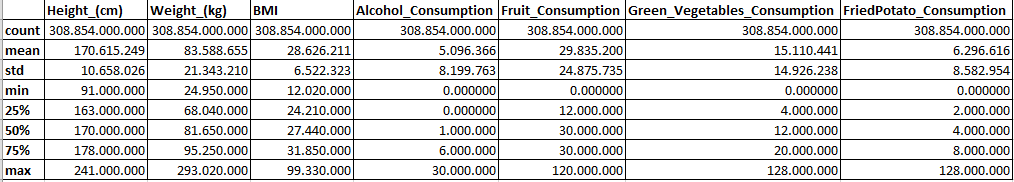
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |

1. **Calidad de los Datos**

* ¿Hay valores faltantes en alguna columna? ¿Cuántos? No hay registros para este criterio
* ¿Hay columnas con valores repetidos o inconsistentes? No hay registros para este criterio
* ¿Existen datos fuera de rango o errores evidentes? No hay registros para este criterio
* ¿Conoce la última fecha de actualización de los datos? Alphiree · Actualizado hace 2 años

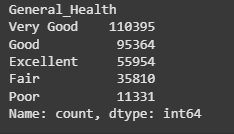
1. **Análisis Cuantitativo Básico**

Para columnas numéricas:

* Mínimo, máximo, media, mediana, desviación estándar.

Para columnas categóricas:

* Frecuencia de cada categoría.



* Número de categorías únicas. 5 Categorías para el campo General\_Health

1. **Conclusiones Iniciales del Grupo**

* ¿Qué información relevante puede extraerse de esta base de datos?

Se puede extraer información para predecir riesgos cardiovasculares por edades, por tendencias en hábitos saludables y por consumo de alimentos.

* ¿Qué problemas de calidad o estructura podrían afectar un análisis más profundo?

Se podrían presentar problemas en la calidad por valores faltantes o inconsistentes, valores fuera de rango o adición de nuevos conceptos no alineados al caso de estudio que puedan generar ruido a los datos resultados esperados.

* ¿Qué preguntas podría ayudar a responder esta base de datos?

Lo más relevante son todas aquellas que conlleve a la predicción para prevenir en la población en general que pueda en un futuro tener problemas relacionados con riesgos cardiovasculares.

* ¿Qué mejoras o limpiezas serían necesarias?

Segmentar la información en un rango de fechas estipulado para un estudio específico, filtrar esta data para clasificarla a una población específica y realizarle un tratamiento de limpieza de los datos analizando sus variables en cuanto eliminando o generando datos dummis para nulos o información inconsistente, estandarizando información y normalizando la estructura de datos en general.

1. **Entrega**

Cada grupo debe entregar este documento diligenciado y, si es posible, el archivo de base de datos con sus observaciones o anotaciones.

DataSet

https://drive.google.com/file/d/1uzwoJGWoCt4QXfI\_0kgErrzl6KHjM1Dy/view?usp=drive\_link