Informe: Procesos de ETL en el Proyecto de Estadística y Probabilidad

- David Alejandro Cajiao
- Juan Andrés López
- Joan Mateo Bermúdez

Introducción

En este informe se detallan los procesos de Extracción, Transformación y Carga (ETL) llevados a cabo en el marco del proyecto de Inferencia Estadística y Modelación Estadística. El objetivo del proceso ETL fue preparar un conjunto de datos limpio y confiable para el análisis estadístico y modelación.

Extracción de Datos

Los datos fueron extraídos de archivos CSV que contenían información sobre estudiantes de high school de tres estados diferentes. Cada archivo contenía varias variables relevantes para el análisis.

Transformación de Datos

El proceso de transformación incluyó varias etapas clave para asegurar la calidad y consistencia de los datos:

1. Eliminación de Valores Nulos

Primero, se eliminaron los registros con valores nulos en las columnas clave: 'Ageyears', 'Height_cm' y 'Armspan_cm'. Esta eliminación se realizó para evitar sesgos en el análisis debido a datos incompletos.

2. Limpieza de Valores Anómalos

Se desarrolló una función para limpiar rangos y textos no numéricos en las columnas 'Height_cm' y 'Armspan_cm'. Esta función eliminó cualquier texto no numérico y convirtió los rangos a valores numéricos únicos.

3. Conversión de Unidades

Los valores de altura y distancia entre brazos extendidos que estaban en pies y pulgadas fueron convertidos a centímetros. Este paso fue crucial para unificar las unidades de medida y facilitar el análisis posterior.

4. Imputación de Valores Anómalos

Durante el análisis inicial, se identificaron valores atípicos en las columnas de altura y distancia entre brazos extendidos. Para tratar estos valores, se calcularon las diferencias entre la altura y la distancia entre brazos. Aquellos registros que presentaban diferencias significativas fueron imputados con la media de la respectiva variable. Este proceso ayudó a estandarizar los datos y eliminar inconsistencias.

5. Filtrado de Datos

Se filtraron los datos para excluir registros con edades mayores a 20 años y alturas menores a 127 cm, basándonos en referencias de crecimiento normal para adolescentes. Esto aseguró que el conjunto de datos reflejara con precisión la población objetivo del estudio.

6. Conversión Final y Exportación

Finalmente, se realizaron conversiones adicionales para asegurar que todas las columnas estuvieran en el formato adecuado (númerico). Los datos transformados fueron exportados a un nuevo archivo CSV limpio, listo para el análisis.

Carga de Datos

El conjunto de datos limpio se cargó en un entorno de análisis (Python o R) para su exploración y análisis estadístico. Este paso incluyó la verificación final de la integridad de los datos y la confirmación de que todas las transformaciones se realizaron correctamente.

Conclusión

El proceso ETL realizado fue fundamental para asegurar la calidad y confiabilidad del conjunto de datos utilizado en el análisis estadístico y la modelación. La cuidadosa eliminación de valores nulos, limpieza de valores anómalos, conversión de unidades y la imputación de datos fueron pasos esenciales que permitieron obtener un conjunto de datos robusto y adecuado para el análisis.