Universidad del Valle de Guatemala
Data Science
Sección 10

Avances del proyecto 1

María José Gil -20337
Fabián Juárez - 21440
Joshua Chicoj -20566
Sofi Lam Méndez - 21548

27 de julio 2024

Descripción del conjunto de datos:
El conjunto de datos fue recolectado de la página del Ministerio de Educación de Guatemala, se tomaron en cuenta aquellos datos de los establecimientos educativos a nivel diversificado de todo el país. Es importante mencionar que los datos crudos cuentan con 17 variables (columnas) y 9356 observaciones (filas).

Variables:

1. Código: Número único de identificación de cada centro educativo
2. Distrito: Código del distrito donde se encuentra el centro educativo.
3. Departamento: Nombre del departamento en el que está ubicado el centro educativo.
4. Municipio: Nombre del municipio donde se encuentra el centro educativo.
5. Establecimiento: Nombre del centro educativo.
6. Dirección: Domicilio física del centro educativo.
7. Teléfono: Número de teléfono del centro educativo.
8. Supervisor: Nombre del supervisor encargado del centro educativo.
9. Director: Nombre del director del centro educativo.
10. Nivel: Grado educativo que ofrece el centro (e.g., primaria, secundaria, diversificado).
11. Sector: Categoría al que pertenece el centro (e.g., público, privado, oficial).
12. Área: Espacio geográfico donde se encuentra el centro (e.g., urbana, rural).
13. Status: Situación operativa del centro educativo (e.g., abierta, cerrada temporalmente).
14. Modalidad: Tipo de enseñanza del centro educativo (e.g., monolingüe, bilingüe).
15. Jornada: Horario de clases que ofrece el centro educativo (e.g., matutina, vespertina, etc).
16. Plan: Tipo de programa de estudios que sigue el centro educativo (e.g., diario, regular).
17. Departamental: Oficina departamental a la que está adscrito el centro educativo.

Variables que más operaciones de limpieza necesitan:
Las variables que requieren más operaciones de limpieza en este conjunto de datos serán ordenadas de mayor a menor necesidad de limpieza, con el objetivo de identificar aquellas que necesitan más atención:

1. Director: tiene 920 datos faltantes
2. Teléfono: tiene 560 datos faltantes
3. Distrito: tiene 230 datos faltantes
4. Supervisor: tiene 232 datos faltantes
5. Establecimiento: posee varias faltas de ortografía ya que carece de diéresis en algunas ocasiones y en otra
6. Modalidad: los datos de esta variable no están distribuidas equitativamente, ya que el 79.9% de los datos están representados por la misma categoría.
7. Sector: los datos de esta variable no están distribuidas equitativamente, ya que el 61.6% de los datos están representados por la misma categoría.
8. Área: los datos de esta variable no están distribuidas equitativamente, ya que el 56.3% de los datos están representados por la misma categoría.

9. Plan: los datos de esta variable no están distribuidas equitativamente, ya que el 55.4% de los datos están representados por la misma categoría.
10. Status: los datos de esta variable no están distribuidas equitativamente, ya que el 51.9% de los datos están representados por la misma categoría.

Link al repositorio de github:

https://github.com/SofiLam13/Proyecto1-DataScience-.git