Avances del proyecto 1.

Autores:

Josue Samuel Argueta Hernandez | Carné 211024
Alejandro José Martinez de León | Carné 21430
Kristopher Javier Alvarado López | Carné 21188
Astrid Marie Glauser Oliva | Carné 21299

Curso:

Data Science

Catedrático:

Lynette Garcia Perez

Sección:

10

Universidad del Valle de Guatemala
11 calle 15-79 Zona 15 Vista Hermosa lll
Guatemala, C. A.
Facultad de Ingeniería

REPOSITORIO

https://github.com/Gustixa/P1_DS.git

- **Describa el set de datos, cuantas filas tiene inicialmente con los datos crudos, y cuantas variables.**

  El set de datos consiste en la información recopilada de los centros educativos registrados por el Mineduc (ministerio de educación) distribuidos por departamento en Guatemala.

  Inicialmente posee 9400 filas en total siendo un archivo crudo. Esto tomando en cuenta que se hizo una unión de todos los archivos para tener uno solo con toda la información y simplificar el análisis.

Variable	Descripción
Código	ID para representar el establecimiento
Distrito	ID único para representar el sector o distrito en donde se ubica el establecimiento
Departamento	Nombre del departamento en donde se ubica el establecimiento
Municipio	Nombre del municipio en donde se ubica el establecimiento
Establecimiento	Nombre del establecimiento
Dirección	Dirección del establecimiento
Teléfono	Número de teléfono
Supervisor	Nombre del supervisor
Director	Nombre del director
Nivel	Nivel de estudios que imparte el establecimiento (primaria, básico, diversificado, etc)
Sector	Si pertenece al sector privado u oficial
Área	Si pertenece al área rural o urbana
Status	Si el establecimiento se encuentra abierto o cerrado
Modalidad	Monolingüe o bilíngue

Jornada	Matutina, vespertina o nocturna
Plan	Diario (regular) o fin de semana
Departamental	Indica si es un establecimiento departamental o único

- **Liste las variables que más operaciones de limpieza necesitarán.**

  DIRECTOR (tiene más del 10% de valores nulos)

  TELEFONO

  SUPERVISOR

  DISTRITO

  DIRECCION

  CODIGO

  ESTABLECIMIENTO

- **Especifique una estrategia para limpiar el conjunto de datos. Por ejemplo:**
- **Para la variable "Establecimiento" se planea:**
  1. **Convertir todo a mayúsculas o minúsculas**
  2. **Eliminar duplicados**
  3. **Revisar errores ortográficos o de cambios de letras en nombres.**
  4. **...**

Para la variable "DIRECTOR":

- **Convertir todo a mayúsculas o minúsculas:** Para mantener la consistencia.
- **Eliminar duplicados:** Asegurarse de que no haya registros duplicados.
- **Corregir errores ortográficos:** Corregir nombres mal escritos, incluyendo caracteres especiales mal representados.
- **Manejo de valores nulos:** Rellenar valores nulos con "DESCONOCIDO".

Para la variable "TELEFONO":

- **Eliminar caracteres no numéricos:** Dejar solo los dígitos.

- **Verificar longitud del número:** Asegurarse de que los números tengan la longitud correcta.
- **Manejo de valores nulos:** Rellenar valores nulos con un valor predeterminado como "00000000".

Para la variable "SUPERVISOR":

- **Convertir todo a mayúsculas o minúsculas:** Para mantener la consistencia.
- **Eliminar duplicados:** Asegurarse de que no haya registros duplicados.
- **Corregir errores ortográficos:** Corregir nombres mal escritos.

Para la variable "DISTRITO":

- **Convertir todo a mayúsculas o minúsculas:** Para mantener la consistencia.
- **Eliminar duplicados:** Asegurarse de que no haya registros duplicados.
- **Corregir errores ortográficos:** Corregir nombres mal escritos.

Para la variable "DIRECCION":

- **Normalizar las abreviaturas comunes:**Convertir "AV." a "Avenida".
- **Revisar y corregir errores ortográficos:** Corregir direcciones mal escritas.
- **Eliminar espacios en blanco adicionales y caracteres no deseados:** Limpiar la columna de caracteres innecesarios.

Para la variable "CODIGO":

- **Eliminar duplicados:** Asegurarse de que no haya registros duplicados.
- **Verificar formato:** Asegurarse de que todos los códigos sigan el mismo formato.

Para la variable "ESTABLECIMIENTO":

- **Convertir todo a mayúsculas o minúsculas:** Para mantener la consistencia.
- **Eliminar duplicados:** Asegurarse de que no haya registros duplicados.
- **Revisar errores ortográficos o de cambios de letras en nombres:** Corregir nombres mal escritos.