Libro de Códigos

Proyecto 1 Ciencia de Datos

Miembros del Equipo

- DIEGO ALEJANDRO PERDOMO SAGASTUME
- DIEGO ANDRÉS ALONZO MEDINILLA
- GUILLERMO ALFONSO FURLAN ESTRADA
- ROBERTO FRANCISCO RIOS MORALES

Introducción

Este documento describe el conjunto de datos utilizado en el proyecto de obtención y limpieza de datos sobre los establecimientos educativos en Guatemala. El objetivo del proyecto es preparar un conjunto de datos limpio y consistente que pueda ser utilizado para análisis posteriores. Los datos incluyen información detallada sobre los establecimientos que ofrecen educación a nivel diversificado, y se han obtenido de la fuente oficial del Ministerio de Educación de Guatemala.

Descripción General del Conjunto de Datos

El conjunto de datos contiene información relevante sobre los establecimientos educativos, abarcando aspectos administrativos, geográficos y operativos. Se han realizado varias operaciones de limpieza para asegurar la calidad y consistencia de los datos.

Variables y Descripción

1. **CODIGO**
    - **Descripción**: Identificador único del establecimiento.
    - **Valores Posibles**: Enteros únicos para cada establecimiento.
2. **DISTRITO**
    - **Descripción**: Distrito administrativo al que pertenece el establecimiento.
    - **Valores Posibles**: Nombres de distritos (cadena de texto).
3. **DEPARTAMENTO**
    - **Descripción**: Departamento donde se encuentra el establecimiento.
    - **Valores Posibles**: Cadenas de texto que representan los nombres de los departamentos de Guatemala, codificados según el archivo `encoding_map.json`.
4. **MUNICIPIO**
    - **Descripción**: Municipio donde se encuentra el establecimiento.
    - **Valores Posibles**: Cadenas de texto que representan los nombres de los municipios, codificados según el archivo `encoding_map.json`.
5. **ESTABLECIMIENTO**

    - **Descripción**: Nombre del establecimiento educativo.
    - **Valores Posibles**: Nombres de establecimientos (cadena de texto).
6. **DIRECCION**
    - **Descripción**: Dirección física del establecimiento.
    - **Valores Posibles**: Cadenas de texto que representan la dirección completa.
7. **TELEFONO**
    - **Descripción**: Número de teléfono del establecimiento.
    - **Valores Posibles**: Números de teléfono (cadena de texto).
8. **SUPERVISOR**
    - **Descripción**: Nombre del supervisor a cargo del establecimiento.
    - **Valores Posibles**: Nombres (cadena de texto).
9. **DIRECTOR**
    - **Descripción**: Nombre del director del establecimiento.
    - **Valores Posibles**: Nombres (cadena de texto).
10. **NIVEL**
    - **Descripción**: Nivel educativo que ofrece el establecimiento.
    - **Valores Posibles**: Codificado, por ejemplo, `0` para Diversificado según el archivo `encoding_map.json`.
11. **SECTOR**
    - **Descripción**: Tipo de administración del establecimiento.
    - **Valores Posibles**:
        - `0`: Cooperativa
        - `1`: Municipal
        - `2`: Oficial
        - `3`: Privado
12. **AREA**
    - **Descripción**: Ubicación geográfica del establecimiento.
    - **Valores Posibles**:
        - `0`: Rural
        - `1`: Sin especificar
        - `2`: Urbana
13. **STATUS**
    - **Descripción**: Estado operativo del establecimiento (activo, inactivo, etc.).
    - **Valores Posibles**: Cadenas de texto describiendo el estado.
14. **MODALIDAD**
    - **Descripción**: Tipo de enseñanza impartida en el establecimiento.
    - **Valores Posibles**:
        - `0`: Bilingüe
        - `1`: Monolingüe
15. **JORNADA**
    - **Descripción**: Horario de clases del establecimiento.
    - **Valores Posibles**:

        - `0`: Doble
        - `1`: Intermedia
        - `2`: Matutina
        - `3`: Nocturna
        - `4`: Sin jornada
        - `5`: Vespertina
16. **PLAN**
    - **Descripción**: Plan de estudios del establecimiento.
    - **Valores Posibles**:
        - `0`: A distancia
        - `1`: Diario (Regular)
        - `2`: Dominical
        - `3`: Fin de semana
        - `4`: Intercalado
        - `5`: Irregular
        - `6`: Mixto
        - `7`: Sabatino
        - `8`: Semipresencial
        - `9`: Semipresencial (dos días a la semana)
        - `10`: Semipresencial (fin de semana)
        - `11`: Semipresencial (un día a la semana)
        - `12`: Virtual a distancia
17. **DEPARTAMENTAL**
    - **Descripción**: Información sobre la organización departamental.
    - **Valores Posibles**: Cadenas de texto que describen la organización.

Estrategias de Limpieza

1. **Carga y Concatenación de Datos**
    - Se cargaron los datos de un archivo CSV que contiene información completa de los establecimientos educativos. Para unificar la información de múltiples fuentes, se concatenaron varios archivos en un solo DataFrame de pandas, permitiendo un análisis integrado y consistente.
2. **Eliminación de Filas Duplicadas**
    - Se eliminaron filas duplicadas para asegurar que cada registro sea único. Esto se logró utilizando la función `drop_duplicates` de pandas, especialmente en la columna `ESTABLECIMIENTO` y otras columnas críticas que podrían contener información redundante.
3. **Identificación de Valores Nulos**
    - Se realizó un análisis exhaustivo para identificar valores nulos en todas las columnas. Se utilizó la función `isnull` para contar los valores faltantes y

evaluar su impacto en el conjunto de datos. Esto ayudó a determinar las columnas críticas que requerían atención especial para mantener la integridad de los datos.

4. **Eliminación de Filas con Valores Nulos**
    - Las filas que contenían valores nulos en las columnas `CODIGO` y `ESTABLECIMIENTO` fueron eliminadas, ya que estos campos son fundamentales para identificar de manera única a cada registro. Se usó `dropna` en pandas para realizar esta operación, garantizando que los datos críticos estén completos.
5. **Normalización de Texto**
    - Se llevó a cabo un proceso de normalización de texto para eliminar acentos y caracteres especiales de las columnas de tipo cadena. Esto se realizó utilizando la función `str.normalize` de pandas, que permite transformar las cadenas de texto en un formato estándar, mejorando la consistencia de la representación de nombres y direcciones.
6. **Limpieza de Números de Teléfono**
    - Los números de teléfono se estandarizaron para asegurar un formato uniforme. Se utilizaron expresiones regulares para eliminar caracteres no numéricos y verificar que cada número cumpliera con la longitud estándar esperada. Este paso garantiza que los datos de contacto sean precisos y consistentes en todo el conjunto de datos.
7. **Transformación de Datos Categóricos con Encoders**
    - Las variables categóricas fueron transformadas en valores numéricos utilizando un encoder. Se implementó un mapeo desde el archivo `encoding_map.json`, que define un esquema de codificación para variables como `NIVEL`, `SECTOR`, `AREA`, `MODALIDAD`, `JORNADA`, `PLAN`, `MUNICIPIO` y `DEPARTAMENTO`. Este paso es crucial para preparar los datos para análisis estadísticos y modelado, facilitando el procesamiento de las categorías por algoritmos de aprendizaje automático.

Fecha de Extracción de Datos

- **Fecha**: Los datos utilizados en este proyecto se obtuvieron en Julio de 2024

Fuente

- **URL**: http://www.mineduc.gob.gt/BUSCAESTABLECIMIENTO_GE/