Universidad del valle de Guatemala
Ciencia de la computación
Data Science

Avances del proyecto #1

Jorge De León 19817
Pablo A. Coutinho 18817
Edgar Andree Toledo 18439
Andrés Quinto 18288

De la página del Ministerio de Educación se extrajeron todas las instituciones de nivel diversificado del país. Alrededor de 23 datasets fueron descargados, la pagina solo brinda información por departamento, cada dataset cuenta con 17 variables, sin embargo, la cantidad de filas si cambia dependiendo del departamento y dataset:

Alta Verapaz	298
Baja Verapaz	129
Chimaltenango	306
Chiquimula	173
Ciudad Capital	1383
El Progreso	118
Escuintla	518
Guatemala	1297
Huehuetenango	448
Izabal	319
Jalapa	138
Jutiapa	277
Petén	362
Quetzaltenango	438
Quiche	233
Retalhuleu	268
Sacatepéquez	297
San marcos	502
Santa rosa	150
Sololá	142
Suchitepéquez	382

Totonicapán	85
Zacapa	97

Todos los datasets utilizan mayúsculas para sus columnas, además de estar todo en español, para ello seguiremos este formato y las unificamos. Las variables son de tipo categórico. La suma total de las filas corresponden a 8360 luego de unificar el dataset.

Variables para limpieza:

1. Teléfono

En esta variable hay datos tipo strings, enteros y floats por lo tanto es necesario convertir todos los datos a string, debido a que no se pueden convertir a integer ya que hay caracteres en los datos.

2. Departamental

La función de esta variable es dividir ciertos departamentos por regiones. Al final de varios nombres hay carácter de separación y esto complica la limpieza ya que cuenta como otro dato, por lo tanto se decidirá no utilizar esta variable ya que es repetitiva con la variable Departamento.

3. Director

Faltan muchos nombres de los directores y esta información que hace falta está llenada con guiones. Esa fragilidad de los datos hace que no se puedan analizar correctamente. Por lo tanto, se reemplazarán estos datos con nan y así tener una variable completa.

4. Establecimiento

Los nombres de los establecimientos tienen carácter especial, al leer el archivo de la base de datos se pierden datos ya que cuenta con un carácter que causa conflicto al leerlo. Se reemplazará ese carácter por un signo diferente y así evitar ese problema.

5. Departamento

Esta variable se utilizará completa ya que no causa ningún problema y los datos se encuentran limpios al extraerlos del data set.

TRANSFORMACIONES

La información viene separada en 23 archivos que se necesitan limpiar, estandarizar y unificar para poder trabajar con la data. No todos los archivos fuente tienen completa la información del director/directora entonces esa columna no se tomará en cuenta. En algunos departamentos la variable de dirección especifica únicamente el nombre de la aldea a la que pertenece el instituto, por lo que se ignorara esta columna. Del mismo modo, el número de teléfono del establecimiento es poco estandarizado, frecuentemente no está en la base de datos y es muy poco relevante para el análisis; por lo que se lo ignorará.

Para normalizar el formato de los nombres en las variables categóricas, se convertirá todos los datos a mayúsculas y se resolverán las colisiones de nombres repetidos añadiendo un número arbitrario al final, dado que coincide con la convención de nombramiento que se maneja para muchos centros de estudios. En el dataset de Guatemala , la variable departamental indica “Guatemala norte”, mientras que en el dataset de Ciudad capital, la variable departamental indica ”Guatemala” , esto se estandarizará a “Guatemala” ya que no nos es de interés esta distinción. Y en su defecto, se asumirá que la modalidad de un establecimiento es monolingüe.

Nota: En el interior del país muchos establecimientos no tienen nombre propio, solo tienen como valor “EORM” o “EOUM” en la variable establecimiento, por las siglas Escuela oficial rural mixta/ Escuela oficial urbana mixta.