Christopher Jose Rodolfo Barrios Solis 18207
Diego Andre Castellanos Castro 20082

Proyecto 1

1. Descripción del set de datos

La base de datos que se proporciona está compuesta por 23 bases de datos diferentes, cada una correspondiente a un departamento de Guatemala, con una pequeña excepción. Guatemala consta de 22 departamentos. Sin embargo, en set tenemos que uno más ya que se tiene la variable de “Guatemala” y “Ciudad capital” probablemente debido a que hay más instituciones en esta región, esto nos va a proveer una buena información y es interesante como se ha organizado el almacenamiento de estos datos.

Esta base de datos contiene 17 variables, que son: CODIGO, DISTRITO, DEPARTAMENTO, MUNICIPIO, ESTABLECIMIENTO, DIRECCION, TELEFONO, SUPERVISOR, DIRECTOR, NIVEL, SECTOR, AREA, STATUS, MODALIDAD, JORNADA, PLAN y DEPARTAMENTAL. A continuación, se describen los datos que abarca cada una de estas variables.

- CODIGO: Es un número único asignado a cada institución, utilizado como identificador exclusivo para cada una de ellas.
- DISTRITO: Es un código que se refiere a la ubicación geográfica de la institución educativa.
- DEPARTAMENTO: Indica el departamento al que pertenece la institución.
- MUNICIPIO: Indica el municipio al que pertenece la institución dentro de su departamento.
- ESTABLECIMIENTO: Es el nombre completo de la institución educativa.
- DIRECCION: Representa la ubicación precisa de la institución.
- TELEFONO: Proporciona uno o más números de teléfono para contactar con la institución.
- SUPERVISOR: Hace referencia al nombre del supervisor de la institución.
- DIRECTOR: Indica el nombre del director de la institución.
- NIVEL: Se refiere al grado de estudio ofrecido (en este caso, se espera que siempre sea diversificado).

- SECTOR: Indica si la institución es pública u oficial, o privada.
- AREA: Se relaciona con si la institución se encuentra en un área urbana o rural.
- STATUS: Hace referencia si la institución está abierta o cerrada.
- MODALIDAD: Indica el tipo de clases impartidas en la institución, ya sea monolingüe o bilingüe.
- JORNADA: Es el período del día en el que se llevan a cabo las clases en la institución (mañana, tarde, noche o doble turno).
- PLAN: Se refiere a los días de la semana en los que se imparten las clases en la institución (plan diario o fines de semana).
- DEPARTAMENTAL: Indica una subdivisión del departamento (por ejemplo, norte, sur, este, oeste, entre otras posibilidades).

En conjunto, con todos los datos recopilados en bruto, el número total de registros en este conjunto de datos es de 8381. No obstante, es posible que existan duplicados, así como filas incompletas o filas que contienen texto no relevante ya que al final de toda la data guardada dice que tienen cierta cantidad de número de institutos encontrados, este tipo de datos no es relevante para nosotros.

Estrategia para limpiar el Conjunto de Datos

Debido a que se encuentra presente la posibilidad de que hayan datos duplicados, la mejor estrategia para prevenir que esto afecte el flujo del proyecto sería eliminando todos los datos duplicados mediante el uso de la limpieza de datos (Data Cleansing). La cual nos permitirá identificar y eliminar todos los datos duplicados, además de aquellos que sean incorrectos, incompletos, inexactos o irrelevantes para el análisis. Otra parte importante de la estrategia sería el uso del manejo de datos faltantes para las filas incompletas. Es bien sabido que los algoritmos no aceptan valores vacíos. Por lo cual, la mejor opción sería eliminar las observaciones que tengan valores perdidos.

Se debe unificar categorías que puedan variar solo por errores ortográficos, por ejemplo, palabras comunes en nombres de establecimientos que puedan tener o no tildes.

Es fundamental estandarizar el significado de "departamento" y "municipio" en todos los conjuntos de datos, especialmente en el de la ciudad capital, donde estas variables pueden tener interpretaciones diferentes.

Se deben evitar columnas con datos ambiguos o que contengan mensajes irrelevantes para el lector, asegurando que cada fila represente únicamente instituciones distintas.

Variables con mayor necesidad de limpieza

- CODIGO: Se eliminarán los guiones “-” para poder facilitar el código a números y para evitar malas interpretaciones de estos signos al momento de aplicar el algoritmo.
- DISTRITO: Se eliminarán los guiones para mayor facilidad de manipulación.
- DEPARTAMENTO: Estos se convertirán de tipo string a tipo int con la ayuda de un catálogo de códigos de departamentos para una interpretación más sencilla.
- MUNICIPIO: Se eliminarán caracteres especiales para evitar que en este campo haya valores difíciles de interpretar.
- ESTABLECIMIENTO: Se eliminarán caracteres especiales para mayor facilidad de manipulación.
- DIRECCION: Se eliminarán caracteres especiales para mejor manipulación y mayor facilidad de interpretación.
- TELEFONO: En algunos establecimientos, se encuentran campos vacíos, mientras que en otros, se registran dos números telefónicos en un solo campo.
- SUPERVISOR: Se eliminarán los posibles caracteres especiales para evitar posibles fallos al momento de aplicar el algoritmo.
- DIRECTOR: Se eliminarán los posibles caracteres especiales para evitar posibles fallos al momento de aplicar el algoritmo.
- NIVEL: Se cambiará de tipo string a tipo int utilizando un catálogo de nivel de escolaridad para facilitar la interpretación de los resultados.
- SECTOR: Se cambiarán de tipo string a tipo int utilizando un catálogo para mejorar la eficiencia del modelo.
- AREA: Se cambiará de tipo string a tipo int utilizando un catálogo de nivel de escolaridad para facilitar la interpretación de los resultados.
- STATUS: Se cambiarán de tipo string a tipo int utilizando un catálogo para mejorar la eficiencia del modelo.
- MODALIDAD: Se cambiará de tipo string a tipo int utilizando un catálogo de nivel de escolaridad para facilitar la interpretación de los resultados.
- JORNADA: Se cambiarán de tipo string a tipo int utilizando un catálogo para mejorar la eficiencia del modelo.
- PLAN: Se cambiará de tipo string a tipo int utilizando un catálogo de nivel de escolaridad para facilitar la interpretación de los resultados.
- DEPARTAMENTAL: Se cambiarán de tipo string a tipo int utilizando un catálogo para mejorar la eficiencia del modelo.

Limpieza de los Datos

Como se puede observar en el proceso realizado, se efectuó la conversión de los archivos .xls a formato .csv con el propósito de lograr un control más efectivo sobre los datos. Posteriormente, después de importar estos archivos, se procedió a unificar la información de todos los departamentos en un único conjunto de datos. A continuación, se llevó a cabo un análisis detallado de la calidad de los datos. Para facilitar esta evaluación, se generó un archivo .csv adicional con el objetivo de visualizar de manera más clara la estructura de los datos. Además, se generó un informe utilizando una librería específica.

Dentro de los datos, se identificaron diversas situaciones que requirieron atención. Entre ellas, destaca el caso relevante de la columna "DIRECTOR", en la cual se presentaban múltiples instancias de valores nulos. Estos valores fueron excluidos debido a la variedad de formas en las que los nulos se manifestaban en dicha columna. Asimismo, se detectaron duplicados en los datos, los cuales fueron eliminados. Se realizó un análisis de la cantidad de valores nulos presentes en distintas columnas. Específicamente, se observó que las columnas con la mayor cantidad de valores nulos eran "DIRECTOR" y "TELEFONO".

Adicionalmente, se tomaron medidas para eliminar filas que contenían información irrelevante, como detalles sobre la fuente de los datos. Estos registros fueron convertidos a valores nulos. Se continuó con la revisión del estado de otras columnas, como "ESTABLECIMIENTOS" y "DIRECTOR". En esta etapa, se contabilizó la cantidad de valores y se identificó que la presencia de tildes en el formato afectaba negativamente la legibilidad y escritura. Por lo tanto, se procedió a remover los acentos y a ajustar los índices.

Una vez completadas las correcciones mencionadas, se procedió a agrupar los valores duplicados en cada columna para identificar posibles datos anómalos. Durante este proceso, se identificaron algunas filas con información incorrecta que se pasó por alto previamente, las cuales fueron eliminadas. Otra anomalía detectada se encontró en la columna de "TELEFONOS", ya que algunos números tenían una cantidad atípica de dígitos (tanto más como menos de ocho dígitos). Estos casos se consideraron posiblemente erróneos y se excluyeron del análisis.

Una dificultad adicional se presentó en la columna de "DIRECCION", la cual tenía múltiples formatos de entrada. Dada esta complejidad, la corrección profunda de las direcciones resultó un desafío. Sin embargo, en términos generales, los demás aspectos parecen estar en orden. Se infiere que el programa original contaba con valores predefinidos y solo enfrentó dificultades en situaciones donde los usuarios podían ingresar datos personalizados.

Code Book

El conjunto de datos comprende 17 atributos, a saber: Código, distrito, departamento, municipio, establecimiento, dirección, teléfono, supervisor, director, nivel, sector, área, estado, modalidad, jornada, plan y departamento.

Variable	Valor
CODIGO	Datos insertados no predefinidos donde lo primeros dos numero significan el departamento, los siguientes dos el municipio y los 4 el numero de identificación del establecimiento: xx-xx-xxxx-xx
DISTRITO	Datos insertados no predefinidos que indican el distrito del pais en el que se ubica el establecimiento: xx-xxx
DEPARTAMENTO	Datos que poseen valores predefinidos que indican el departamento de el pais de Guatemala en el que se ubica el establecimiento, estos son los siguientes: Alta Verapaz, Baja Verapaz, Chimaltenango, Chiquimula, Ciudad Capital, El Progreso, Escuintla, Guatemala, Huehuetenango, Izabal, Jalapa, Jutiapa, Petén, Quetzaltenango, Quiché, Retalhuleu, Sacatepéquez, San Marcos, Santa Rosa, Sololá, Suchitepéquez, Totonicapán, Zacapa.
MUNICIPIO	Datos que poseen valores predefinidos que indican el municipio del departamento del pais de Guatemala en el que se ubica la institucion, estos son los siguientes:Acatenango, Agua Blanca, Aguacatán, Alotenango, Amatitlán, Antigua Guatemala, Asunción Mita, Atescatempa, Ayutla, Barberena, Cabañas, Cabricán, Cajola, Camotán, Canilla, Cantel, Casillas, Catarina, Chahal, Chajul, Champerico, Chiantla, Chicacao, Chicamán, Chiche, Chimaltenango, Chinautla, Chinique, Chiquimula, Chiquimulilla, Chisec, Chuarrancho, Ciudad Vieja, Coatepeque, Cobán, Colomba Costa Cuca, Colotenango, Comapa, Comitancillo, Concepción Chiquirichapa, Concepción Huista, Concepción Las Minas, Concepción Tutuapa, Conguaco, Cubulco, Cuilapa, Cuilco, Cunén, Cuyotenango, Dolores, El

	Adelanto, El Asintal, El Chal, El Estor, El Jícaro, El Palmar, El Progreso, El Quetzal, El Tejar, El Tumbador, Escuintla, Esquipulas, Esquipulas Palo Gordo, Estanzuela, Flores, Flores Costa Cuca, Fraijanes, Fray Bartolomé de las Casas, Génova Costa Cuca, Granados, Gualán, Guanagazapa, Guastatoya, Guatemala, Guazacapán, Huehuetenango, Huitán, Huite, Ipala, Ixcán, Ixchiguán, Iztapa, Jacaltenango, Jalapa, Jalpatagua, Jerez, Jocotán, Jocotenango, Joyabaj, Jutiapa, La Blanca, La Democracia, La Esperanza, La Gomera, La Libertad, La Reforma, La Tinta, La Unión, Lanquín, Las Cruces, Livingston, Los Amates, Magdalena Milpas Altas, Malacatán, Malacatancito, Masagua, Mataquescuintla, Mazatenango, Melchor de Mencos, Mixco, Momostenango, Monjas, Morales, Morazán, Moyuta, Nahualá, Nebaj, Nentón, Nueva Concepción, Nueva Santa Rosa, Nuevo Progreso, Nuevo San Carlos, Ocos, Olintepeque, Olopa, Oratorio, Pachalum, Pajapita, Palencia, Palestina de los Altos, Palin, Panajachel, Panzós, Parramos, Pasaco, Pastores, Patulul, Patzicía, Patzité, Patzún, Petatán, Poptún, Pueblo Nuevo, Pueblo Nuevo Viñas, Puerto Barrios, Purulhá, Quesada, Quetzaltenango, Quezaltepeque, Rabinal, Raxruhá, Retalhuleu, Río Blanco, Río Bravo, Río Hondo, Sacapulas, Salamá, Salcajá, Samayac, San Agustín Acasaguastlán, San Andrés, San Andrés Itzapa, San Andrés Sajcabajá, San Andrés Semetabaj, San Andrés Villa Seca, San Andrés Xecul, San Antonio Aguas Calientes, San Antonio Huista, San Antonio Ilotenango, San Antonio La Paz, San Antonio Palopó, San Antonio Sacatepéquez, San Antonio Suchitepéquez, San Bartolomé Aguas Calientes, San Bartolomé Jocotenango, San Bartolomé Milpas Altas, San Benito, San Bernardino, San Carlos Alzatate, San Carlos Sija, San Cristóbal Acasaguastlán, San Cristóbal Cucho, San Cristóbal Totonicapán, San Cristóbal Verapaz, San Diego, San Felipe, San Francisco, San Francisco El Alto, San Francisco La Unión, San Francisco Zapotitlán, San Gabriel, San Ildefonso Ixtahuacán, San Jacinto, San Jerónimo, San Jorge, San José, San José Acatempa, San José Chacaya, San José

	del Golfo, San José El Ídolo, San José El Rodeo, San José La Arada, San José La Máquina, San José Ojetenam, San José Pinula, San José Poaquil, San Juan Atitán, San Juan Chamelco, San Juan Comalapa, San Juan Cotzal, San Juan Ermita, San Juan Ixcoy, San Juan La Laguna, San Juan Ostuncalco, San Juan Sacatepéquez, San Juan Tecuaco, San Lorenzo, San Lucas Sacatepequez, San Lucas Tolimán, San Luis, San Luis Jilotepeque, San Manuel Chaparrón, San Marcos, San Marcos La Laguna, San Martín Jilotepeque, San Martín Sacatepequez, San Martín Zapotitlán, San Mateo, San Mateo Ixtatán, San Miguel Acatán, San Miguel Chicaj, San Miguel Dueñas, San Miguel Ixtahuacán, San Miguel Petapa, San Miguel Pochuta, San Miguel Siguilá, San Miguel Tucurú, San Miguel Uspantán, San Pablo, San Pablo Jocopilas, San Pablo La Laguna, San Pedro Ayampuc, San Pedro Carchá, San Pedro Jocopilas, San Pedro La Laguna, San Pedro Necta, San Pedro Pinula, San Pedro Sacatepéquez, San Pedro Soloma, San Pedro Yepocapa, San Rafael La Independencia, San Rafael Las Flores, San Rafael Petzal, San Rafael Pie de la Cuesta, San Raymundo, San Sebastián, San Sebastián Coatán, San Sebastián Huehuetenango, San Vicente Pacaya, Sanarate, Sansare, Senahú, Sibilia, Sibinal, Sipacapa, Sipacate, Siquinalá, Sololá, Sumpango, Tacaná, Tactic, Tajumulco, Taxisco, Tecpán Guatemala, Tectitán, Teculután, Tejutla, Tiquisate, Todos Santos Cuchumatán, Totonicapán, Unión Cantinil, Usumatlán, Villa Canales, Villa Nueva, Yupiltepeque, Zacapa, Zacualpa, Zapotitlán, Zaragoza, Zona 1, Zona 10, Zona 11, Zona 12, Zona 13, Zona 14, Zona 15, Zona 16, Zona 17, Zona 18, Zona 19, Zona 2, Zona 21, Zona 24, Zona 3, Zona 4, Zona 5, Zona 6, Zona 7, Zona 8, Zona 9, Zunil, Zunilito.
ESTABLECIMIENTO	Datos insertados no predefinidos que nos indican el nombre del establecimiento.
DIRECCION	Datos insertados no predefinidos que indican el domicilio del establecimiento.
TELEFONO	Datos insertados no predefinido del número para contactar con el establecimiento.

SUPERVISOR	Datos insertados no predefinido del nombre del encargado en que es el responsable del cumplimiento de las políticas educativas y de la ejecución de sus estrategias
DIRECTOR	Datos insertados no predefinido del nombre del encargado en dirigir y coordinar la actividad del centro, sin perjuicio de las competencias atribuidas al Claustro del profesorado y al Consejo Escolar.
NIVEL	Dato predefinido que indica el grado académico hasta el que posee el establecimiento: Diversificado
SECTOR	Datos predefinidos que indica tiene una institución o un encargado de financiar el establecimiento que se dividen en los siguientes: Cooperativa, Municipal, Oficial, Privado
AREA	Datos predefinidos que nos indica el desarrollo del lugar en donde se ubica el establecimiento de dividen el las siguientes: Rural, Sin especificar, Urbana.
STATUS	Datos predefinidos que indican la situación actual del establecimiento, estos son los siguientes: Abierta, Cerrada temporalmente, Temporal títulos.
MODALIDAD	Datos predefinidos que indican si se enseñan uno o más idiomas, se divide en los siguientes: Bilingüe, Monolingüe.
JORNADA	Datos predefinidos que indican el momento en el día en el cual el establecimiento atiende su jornada, son los siguientes: Doble, Intermedia, Matutina, Nocturna, Sin jornada, Vespertina.
PLAN	Datos predefinidos que nos indican la manera o días que llevan acabo las jornadas los establecimientos: A distancia, Diario (regular), Dominical, Fin de semana, Intercalado, Irregular, Mixto, Sabatino, Semipresencial, Semipresencial (dos días a la semana), Semipresencial (fin de semana), Semipresencial (un día a la semana), Virtual a distancia.
DEPARTAMENTAL	Datos predefinidos que nos indican la institución de cada departamento encargada del establecimiento: Alta Verapaz, Baja Verapaz,

	Chimaltenango, Chiquimula, El Progreso, Escuintla, Guatemala Norte, Guatemala Occidente, Guatemala Oriente, Guatemala Sur, Huehuetenango, Izabal, Jalapa, Jutiapa, Petén, Quetzaltenango, Quiché, Quiché Norte, Retalhuleu, Sacatepéquez, San Marcos, Santa Rosa, Sololá, Suchitepéquez, Totonicapán, Zacapa.

Github:https://github.com/CristopherBarrios/Proyecto-ObtencionYLimpiezaDeLosDatos-DataScience/tree/main