Universidad Nacional Autónoma de México

Facultad de Estudios Superiores Acatlán

Licenciatura en Matemáticas Aplicadas y Computación

Equipo LLM (Línea Mujeres)

Materia:

Temas Selectos de Computación I - Métodos y Técnicas para la Gestión de Datos

Grupo: 1753

Profesor:

Reyes Hernández Sergio

Integrantes del Equipo:

Nombre
Soria Cabrera Andrés
Olvera Juárez Josué Israel
Godinez Rojas Alexis Omar
Xaltenco Aguilar Salvador
Mejia Trejo Hazel Jesús

Índice

1. Origen y Contexto de los Datos **2**
1.1. ¿Quién genera los datos? . . . 2
1.2. ¿Con qué propósito? . . . 2
1.3. ¿Cuánta información hay disponible? . . . 2
1.4. Fuente Oficial y Actualizaciones . . . 2

2. Desafíos en la Carga de Datos y Soluciones **2**
2.1. Solución 1: Muestreo Aleatorio (Exploratoria) . . . 2
2.2. Solución 2: Carga de Archivo Comprimido (Implementada) . . . 2

3. Procesamiento y Análisis en SAS **3**
3.1. Importación y Creación de la Tabla Permanente . . . 3
3.2. Verificación de la Estructura . . . 4
3.3. Análisis Exploratorio y Calidad de Datos . . . 4

4. Análisis Programático del Diccionario de Datos **4**
4.1. Hoja: CATÁLOGO . . . 4
4.2. Hoja: LISTADO DE TEMÁTICAS . . . 6
4.3. POSIBLES COMBINACIONES DE TEMÁTICAS . . . 6
4.4. Conclusiones de la Estructura . . . 7

1 Origen y Contexto de los Datos

1.1 ¿Quién genera los datos?

Los datos son generados por el **Gobierno de la Ciudad de México** a través de las interacciones registradas en el servicio "Línea Mujeres".

1.2 ¿Con qué propósito?

El propósito es doble:

1. **Transparencia:** Ofrecer un registro público y abierto de las llamadas y servicios proporcionados para atender la violencia contra las mujeres.
2. **Análisis y Políticas Públicas:** Permitir que investigadores, analistas y el propio gobierno estudien los patrones de violencia, identifiquen necesidades y diseñen políticas públicas más efectivas.

1.3 ¿Cuánta información hay disponible?

El conjunto de datos históricos completo contiene **501,833 registros** (observaciones), cada uno representando una llamada o interacción.

1.4 Fuente Oficial y Actualizaciones

La base de datos original fue completada en noviembre de 2020. La información y sus actualizaciones mensuales se pueden encontrar en el Portal de Datos Abiertos de la CDMX:

Enlace: `https://datos.cdmx.gob.mx/dataset/linea-mujeres`

2 Desafíos en la Carga de Datos y Soluciones

Durante el análisis, se presentó un desafío técnico: el archivo `linea-mujeres-cdmx.csv` (aproximadamente 137 MB) superaba el límite de carga de la plataforma SAS Studio (aproximadamente 95 MB). Se exploraron dos soluciones:

2.1 Solución 1: Muestreo Aleatorio (Exploratoria)

La primera estrategia fue reducir el tamaño del archivo mediante un **muestreo aleatorio**. Se creó una muestra representativa de 300,000 filas, resultando en el archivo `linea-mujeres-cdmx-reducido.csv`. Aunque este enfoque es estadísticamente válido para análisis exploratorios, no permite trabajar con la totalidad de las observaciones.

2.2 Solución 2: Carga de Archivo Comprimido (Implementada)

La solución definitiva y más efectiva fue **integrar Python en SAS Studio para manejar los datos completos**. El proceso consistió en subir el archivo `linea-mujeres-cdmx.zip`

(que es más ligero que el CSV) y luego ejecutar un script de Python dentro de SAS para descomprimirlo directamente en el servidor. Esto permitió el acceso a las **501,833 observaciones** sin pérdida de información.

El código utilizado para la descompresión fue el siguiente:

Listing 1: Código SAS para descompresión con Python

```
/* Archivo: descomprimir.sas */
proc python;
submit;
# c digo Python para descomprimir el archivo ZIP
import zipfile
import os

# Ruta donde se subi  el archivo ZIP en el servidor de SAS Viya
zip_path = '/export/viya/homes/421043394@pcpuma.acatlan.unam.mx/linea-mujeres-cdmx.zip'
# Ruta donde se extraer  el contenido
extract_path = '/export/viya/homes/421043394@pcpuma.acatlan.unam.mx'

with zipfile.ZipFile(zip_path, 'r') as zip_ref:
    zip_ref.extractall(extract_path)
    print(" Archivo  descomprimido exitosamente!")

endsubmit;
quit;
```

3 Procesamiento y Análisis en SAS

Una vez disponible el archivo CSV completo en el entorno de SAS, se procedió a su importación y análisis.

3.1 Importación y Creación de la Tabla Permanente

El siguiente código define una librería (`LM`) y luego importa los datos del archivo CSV a una tabla de trabajo, para finalmente guardarla como una tabla permanente en la librería del proyecto.

Listing 2: Importación de datos y creación de tabla permanente

```
/* Creaci n de la librer a para el proyecto */
libname LM "~/PG2V2/proyecto";

/*
  Importaci n del archivo CSV a una tabla temporal.
  Nota: El nombre del archivo puede ser el reducido o el completo,
  dependiendo de la soluci n utilizada.
*/
proc import datafile="~/PG2V2/proyecto/linea-mujeres-cdmx-reducido.csv"
dbms=csv out=work.tabla1
replace;
run;

/* Creaci n de la tabla permanente a partir de la temporal */
```

```
15 title "Tabla general de Linea Mujeres";
16 data LM.tabla1;
17 set work.tabla1;
18 run;
19 TITLE;
```

3.2 Verificación de la Estructura

Se utilizó `proc contents` para verificar la estructura de la tabla cargada y asegurar que los tipos de datos y formatos fueran correctos.

Listing 3: Verificación de estructura de datos

```
1 proc contents data=LM.tabla1;
2 format fecha_alta date10.;
3 run;
```

3.3 Análisis Exploratorio y Calidad de Datos

El análisis del conjunto de datos completo (501,833 registros) confirma los hallazgos preliminares:

- **Calidad de Datos:** Existe una alta proporción de valores nulos en columnas clave, especialmente las relacionadas con los detalles del hecho (`estado_hechos`, `municipio_hechos`) y las temáticas secundarias (`tematica_4` a `tematica_7`).
- **Variables Principales:** Las columnas `tematica_1` y `tematica_2` son las más completas y fiables para el análisis.
- **Datos Atípicos:** Campos como `edad` presentan un rango muy amplio (1 a 96 años), lo que sugiere la necesidad de una limpieza de datos antes de realizar modelos predictivos.

4 Análisis Programático del Diccionario de Datos

Este análisis se basa en el archivo:
`diccionario-de-datos-llamadas-realizadas-a-linea-mujeres-1.xlsx` y sus hojas internas.

4.1 Hoja: CATÁLOGO

Esta hoja define la estructura principal del conjunto de datos, detallando cada una de las 28 variables (columnas).

Cuadro 1: Estructura del Catálogo de Variables

Variable	**Descripción**	**Tipo SAS**
FOLIO	Identificador único de la llamada	Numérico
FECHA_ALTA	Fecha en la que se generó la llamada de la usuaria (día, mes y año)	Carácter
AÑO_ALTA	Año en que se generó la llamada de la usuaria	Carácter
MES_ALTA	Mes en que se generó la llamada de la usuaria	Numérico
DIAS_ALTA	Día en que se generó la llamada de la usuaria	Numérico
HORA_ALTA	Año en que se generó la llamada de la usuaria	Numérico
SEXO	Sexo de la usuaria que realizó la llamada	Carácter
EDAD	Edad de la usuaria que realizó la llamada	Numérico
ESTADO CIVIL	Estado civil de la usuaria que realizó la llamada	Carácter
OCUPACION	Ocupación de la usuaria que realizó la llamada	Carácter
ESCOLARIDAD	Escolaridad de la usuaria que realizó la llamada	Numérico
ESTADO_USUARIA	Entidad Federativa de la usuaria que realizó la llamada	Carácter
MUNICIPIO_USUARIA	Municipio de la usuaria que realizó la llamada	Carácter
COLONIA_USUARIA	Colonia de la usuaria que realizó la llamada	Carácter
CP_USUARIA	Código Postal de la usuaria que realizó la llamada	Carácter
ESTADO_HECHOS	Estado en el que se registraron los hechos	Carácter
MUNICIPIO_HECHOS	Municipio en el que se registraron los hechos	Carácter
COLONIA_HECHOS	Colonia en la que se resgistraron los hechos	Carácter
CP_HECHOS	Código Postal en el que se registraron los hechos	Carácter
ORIGEN	Plataforma desde la que se inicio la llamada	Carácter
SERVICIO	Servicio que se le proporcionó a la usuaria	Carácter
TEMATICA_1	Primera temática en la que se categoriza la llamada	Carácter
TEMATICA_2	Segunda temática en la que se categoriza la llamada	Carácter
TEMATICA_3	Tercera temática en la que se categoriza la llamada	Carácter
TEMATICA_4	Cuarta temática en la que se categoriza la llamada	Carácter
TEMATICA_5	Quinta temática en la que se categoriza la llamada	Carácter
TEMATICA_6	Sexta temática en la que se categoriza la llamada	Carácter
TEMATICA_7	Séptima temática en la que se categoriza la llamada	Carácter

4.2 Hoja: LISTADO DE TEMÁTICAS

Esta hoja funciona como una tabla de consulta ('lookup table'). Define los posibles valores que pueden tomar las columnas `TEMATICA_1` a `TEMATICA_7` en el dataset principal.

Cuadro 2: Listado de Temáticas Disponibles

Temática	**Descripción**
ADICCIONES	-
AMBIENTAL	-
CIVIL	-
FAMILIAR	-
FISCAL	-
LABORAL	-
MERCANTIL	-
MIGRATORIO	-
PENAL	-
SERVICIOS ASISTENCIALES	-
SERVICIOS DE SALUD	-
SEXUALIDAD	-
VIOLENCIA	-
PSICOLÓGICO	-
VIOLENCIA	PSICOLÓGICO

4.3 POSIBLES COMBINACIONES DE TEMÁTICAS

Esta hoja detalla las combinaciones más frecuentes o válidas entre `TEMATICA_1` y `TEMATICA_2`, mostrando cómo se co-clasifican las llamadas. Esto es clave para entender la interrelación de los problemas reportados.

Cuadro 3: Combinaciones de Temáticas Frecuentes

Temática 1	**Temática 2**
CIVIL	FAMILIAR
CIVIL	PSICOLÓGICO
FAMILIAR	PSICOLÓGICO
FISCAL	PSICOLÓGICO
LABORAL	PSICOLÓGICO
MERCANTIL	PSICOLÓGICO
MIGRATORIO	PSICOLÓGICO
PENAL	FAMILIAR
PENAL	PSICOLÓGICO
SERVICIOS ASISTENCIALES	PSICOLÓGICO
SERVICIOS DE SALUD	PSICOLÓGICO
SEXUALIDAD	PSICOLÓGICO
VIOLENCIA	PSICOLÓGICO

4.4 Conclusiones de la Estructura

El diccionario de datos está muy bien estructurado y va más allá de una simple lista de variables. Se organiza como una base de datos relacional en miniatura:

- **Hoja `CATÁLOGO` (Datos):** Actúa como la tabla de hechos principal, definiendo las columnas del dataset.
- **Hoja `LISTADO DE TEMÁTICAS` (Dimensiones):** Es una tabla de dimensiones que da significado a los valores categóricos de las temáticas. Permite enriquecer los datos, reemplazando un código o nombre corto con una descripción completa.
- **Hoja `POSIBLES COMBINACIONES DE TEMÁT` (Reglas de Negocio):** Define las relaciones o reglas de co-ocurrencia entre las temáticas. Esto es fundamental para análisis más profundos, como análisis de canasta de mercado (qué problemáticas suelen ir juntas) o para validar la calidad de la clasificación de datos.

En resumen, la estructura del diccionario proporciona un contexto rico que es crucial para un análisis de datos preciso y profundo. Facilita la comprensión no solo de *qué* datos se recogen, sino de *cómo se relacionan* entre sí.

Conclusión General

La carga exitosa del dataset completo de Línea Mujeres CDMX, con sus 501,833 registros, representa un hito importante en el desarrollo de este proyecto. La implementación de la solución técnica mediante Python integrado en SAS demostró ser efectiva para superar las limitaciones de tamaño de archivo.

El análisis del diccionario de datos revela una estructura bien organizada que facilitará los análisis posteriores. La documentación completa del proceso garantiza la reproducibilidad y transparencia del trabajo realizado por el equipo LLM.