Minería de Datos

Proceso CRISP-DM

TAREA 1

Participantes:
Rincon Ramírez Victor Francisco

Práctica CRISP-DM

Instrucciones

- Considere los siguientes conjuntos de datos, para responder las preguntas propuestas ó usted se puede formular las preguntas que pueda responder a partir de los datos.
- Recuerden que las preguntas se pueden responder o no depende de la calidad de los datos proporcionados.
- Pueden complementar con otros conjuntos de datos.
- Debe realizar un proceso CRISP-DM, por lo tanto, los pasos deben estar explícitos en la tarea.

Desarrollo

Problemas que se deben abordar

LOCATEL fue la Institución que brindo apoyo y orientación a la Ciudadanía ante las grandes emergencias, así como pionera en México en la implementación de servicios de acuerdo a las necesidades de las y los usuarios de la Ciudad de México.

Las problemáticas a las que se debe de enfrentar LOCATEL abarcan desde la limitación de recursos financieros que puede restringir la implementación efectiva de los servicios prestados hasta carecer de la infraestructura necesaria para responder a la demanda.

Así como gestionar correctamente sus recursos para enfocarlos de manera optima por lo que requiere identificar los grupos poblacionales y tipos de servicio más demandado sin descuidar los demás aspectos funcionales de la Institución.

- Identificar patrones, el tipo y localización de usuarios para diseñar su infraestructura y gestionar equitativamente sus recursos

Comprensión del negocio

Fuente del dataset: Gobierno de la Ciudad de México. (2023). Servicios integrales 2022-2023 [Data set]. Archivo de Datos de la Ciudad de México. https://archivo.datos.cdmx.gob.mx/bases_integrales/servicios_integrales_2022-2023.csv

En esta base podrás encontrar toda la información de las llamadas realizadas al servicio LOCATEL en torno a sus servicios integrales desde noviembre de 2016. Esta información se actualizará semanalmente a partir del 6 de abril de 2020. Los servicios integrales de LOCATEL son:

- Atención y asesoría jurídico
- Atención y asesoría psicológica
- Atención y asesoría médica

La captura de información se realiza a través del Sistema de Registro de Información de Locatel (SIRILO). Los datos presentados en esta base de datos se obtienen durante la entrevista que realizan las operadoras y operadores, a través del consentimiento de las personas usuarias que se comunicaron al servicio de asesorías integrales de LOCATEL*. Esta información es otorgada de forma opcional por las usuarias y no es de carácter obligatorio para proporcionar la atención. La temática de cada atención es determinada por un protocolo de interacción que, a través

de la exploración del motivo de llamada, permite focalizar la demanda de la persona usuaria para elaborar un plan de acción.[1]

Descripción de datos

Index	Campo	Non-null	Tipo de dato	Descripción
0	folio	173913	float64	ID de la fila
1	fecha_alta	173932	objeto	Fecha del contacto
2	aÃ±o_alta	173932	int64	Año del contacto
3	mes_alta	173932	objeto	Mes del contacto
4	hora_alta	173932	objeto	Hora del contacto
5	edad	173932	int64	Edad del usuario
6	sexo	173932	objeto	Sexo del usuario
7	estado_civil	170991	objeto	Estado civil del usuario
8	ocupacion	170767	objeto	Ocupación de usuario
9	escolaridad	169864	objeto	Nivel de escolaridad del usuario
10	estado_usuaria	173932	objeto	Estado MX/USA de donde se encuentra el usuario
11	municipio_usuaria	173671	objeto	Municipio MX/USA de donde se encuentra el usuario
12	colonia_usuaria	173671	objeto	Colonia MX/USA de donde se encuentra el usuario
13	cp_usuaria	173930	float64	Código postal MX/USA de donde se encuentra el usuario
14	estado_hechos	11161	objeto	Estado en el cual se reporta ocurre un evento puede ser o no igual al estado_usuaria
15	municipio_hechos	11161	objeto	Municipio en el cual se reporta ocurre un evento puede ser o no igual al municipio_usuaria
16	colonia_hechos	11161	objeto	Colonia en la cual se reporta ocurre un evento puede ser o no igual a colonia_usuaria
17	cp_hechos	11161	float64	Código postal del lugar donde ocurre un evento puede ser o no igual al cp_usuaria

[1] Gobierno de la Ciudad de México. (2023). Servicios para la población en general. Datos CDMX. https://datos.cdmx.gob.mx/dataset/servicios-para-la-poblacion-en-general

18	origen	5731	objeto	Procedencia de la llamada: llamada directa o transferencia desde otro servicio de emergencia
19	servicio	173932	objeto	Tipo de servicio solicitado: Jurídico, Médico o Psicológico
20	tematica_1	173925	objeto	Especificaciones estandarizadas sobre el tipo de contacto
21	tematica_2	168395	objeto	Especificaciones estandarizadas sobre el tipo de contacto
22	tematica_3	147226	objeto	Especificaciones estandarizadas sobre el tipo de contacto
23	tematica_4	82559	objeto	Especificaciones estandarizadas sobre el tipo de contacto
24	tematica_5	49281	objeto	Especificaciones estandarizadas sobre el tipo de contacto
25	tematica_6	44244	objeto	Especificaciones estandarizadas sobre el tipo de contacto
26	tematica_7	40426	objeto	Especificaciones estandarizadas sobre el tipo de contacto

Seleccionamos los campos de 'fecha_alta', 'aÃ±o_alta', 'mes_alta', 'hora_alta', 'sexo', 'edad','estado_usuaria', 'servicio' por ser los campos más completos del dataset con 173932 objetos no nulos.

Tipo de datos: float64(3), int64(2), object(22)

Consumo de memoria: 35.8+ MB

IDE: Google Colaboratory

Lenguaje(s): Python 3, v3.10

Módulos: Pandas, Matplotlib, Numpy, ScikitLearn, Seaborn

Equipo: Thinkpad Lenovo T470

Espacio de almacenamiento requerido: 3GB

Memoria RAM requerida: 12GB

Procesador Local: i5-7300U, 2.60- 2.71 GHz

SO Local: Windows Pro 10 v22H2 x64

Navegadore(s): Google Chrome

Tipo de archivos: CSV, Py

Repositorio del proyecto: Github

Los campos de 'aÃ±o_alta' y 'mes_alta' serán omitidos por ser innecesarios al tener un campo que ya contiene ambos datos de nombre 'fecha_alta' para brindar una cronología a la sucesión de eventos y permita identificar el orden correcto.

Se omitirá el campo 'hora_alta' ya que la conversión correcta a enteras puede ser complicada e involucrará demasiado tiempo para una sola tarea

También se usará el campo: 'sexo', 'edad', 'estado_usuaria' y 'servicio' ya que sus datos no null corresponden al total de registros, por lo que cuenta con los suficientes datos para devolver información de valor y nos permita encontrar patrones entre los grupos poblacionales, su distribución geográfica y los servicios solicitados.

Se almacenarán en un *dataset* de nombre 'locatel_' que cuente solo con los campos mencionados y sobre el cual se trabajara.

Preparación de los datos

Extracción:

Se extrajeron los campos "fecha_alta","estado_usuaria","edad","servicio","sexo" que son campos sin datos nulos y se almacenaron en un dataset de nombre 'df_'

Se generó una copia de seguridad sobre la que se trabajara 'locatel_'

Transformación:

Se organizará por el orden cronológico del campo 'fecha_alta'

Se validan los datos de "fecha_alta","estado_usuaria", "edad", "servicio", "sexo" y se transformaran a datos cuantitativos (int64) para obtener los agrupamientos que nos indiquen los grupos de mayor incidencia.

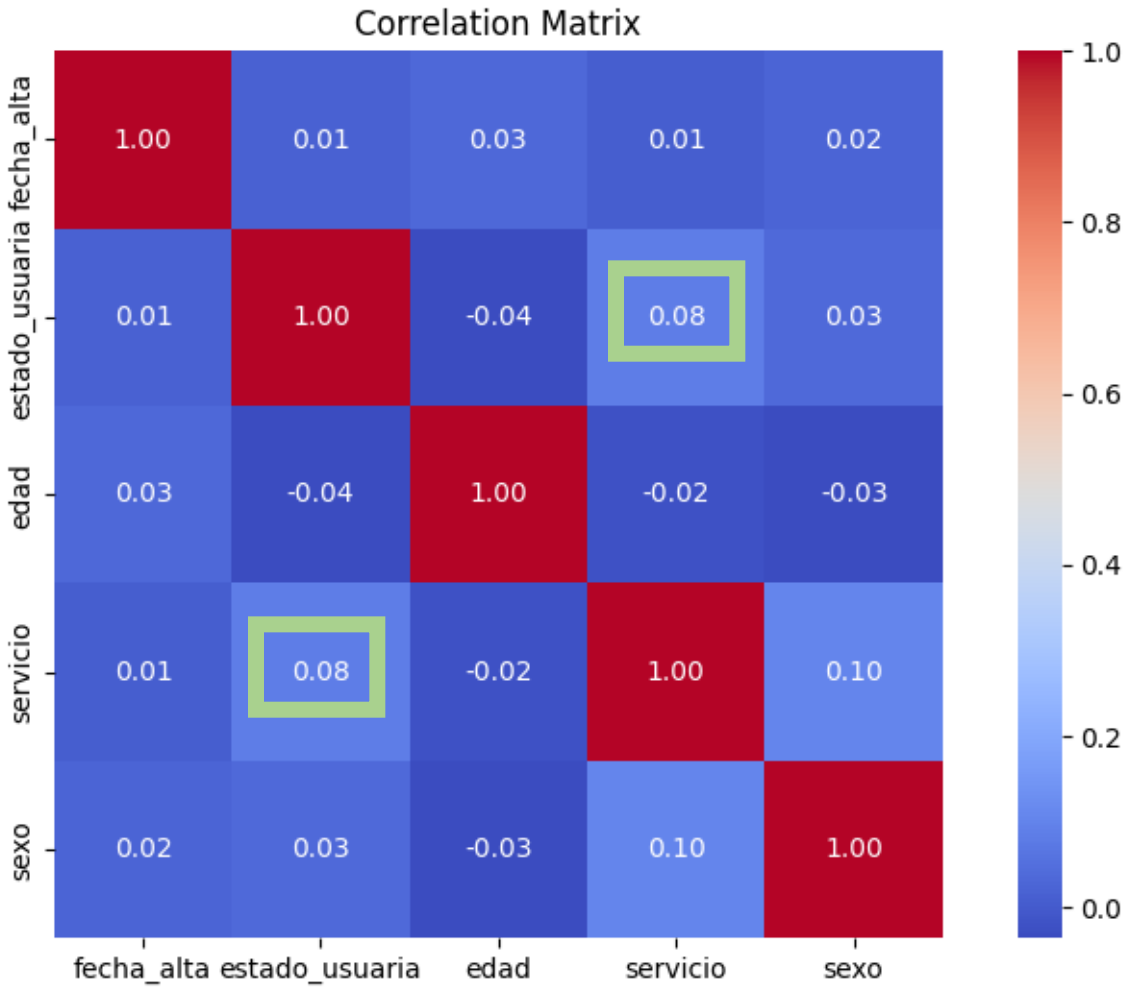
Almacenamiento:

Se genera un datatset con los campos seleccionados pero transformados a datos cuantitativos con el nombre de locatel_.csv y almacenado en el repositorio del proyecto

Modelado de datos

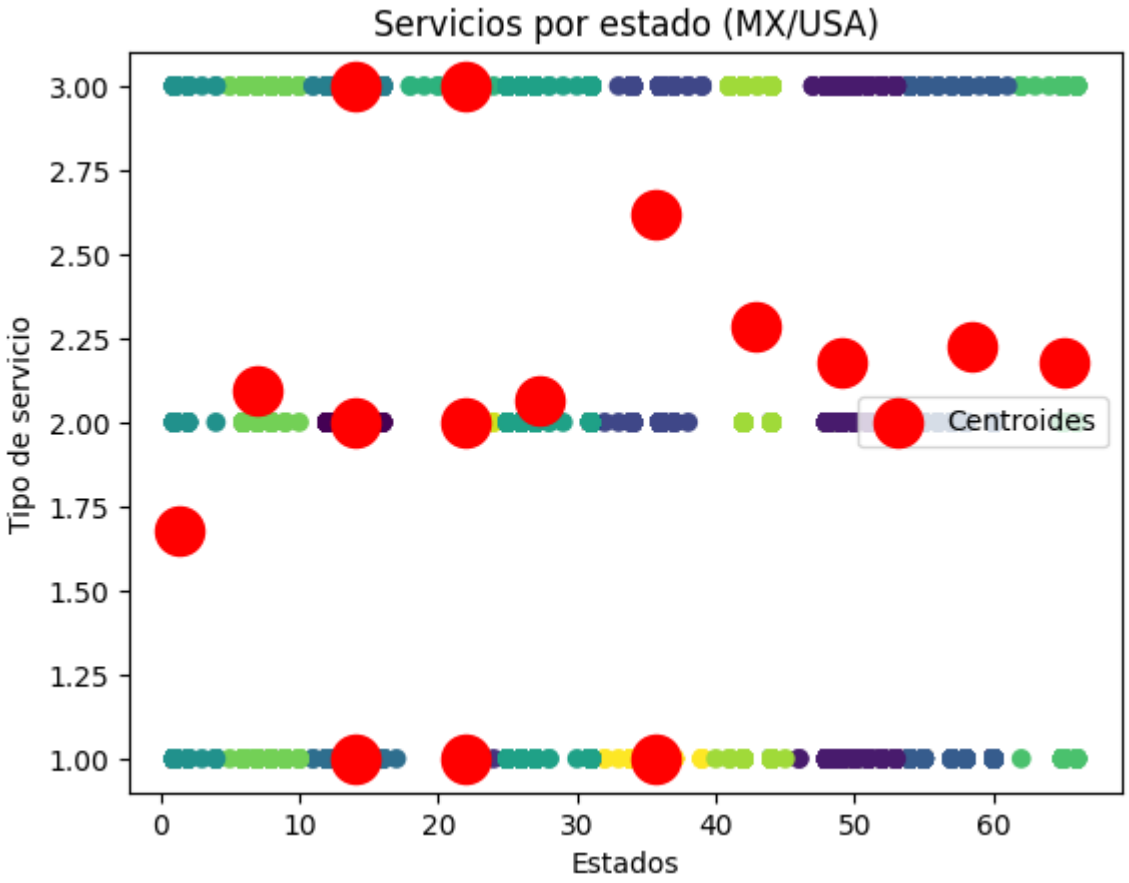
Ya con el *dataset* locatel_.csv se carga y se realiza una correlación entre los campos para conocer la relación entre ellos y determinar cuales serian los campos que se usaran para modelar.

Con una correlación entre 'servicios' y 'estado_usuaria' identificamos los campos con mayor relación (0.8)



Utilizaremos el algoritmo K-means con el módulo ScikitLearn el cual es un método de agrupamiento (clustering) para aprendizaje no supervisado y agrupar datos en k-grupos basados en características parecidas, es utilizado en Aprendizaje Automático (Machine Learning) para ver segmentaciones de los datos e inferir la estructura de los datos.

"El algoritmo busca dividir un conjunto de datos en k grupos o clusters, donde cada punto de datos se asigna al cluster más cercano según alguna medida de distancia".[2] (Evangelista, M., 2024)



La concentración de centroides en el en el rango de 45 a 60 indica que en estos estados es donde más se reportan más casos tanto médicos como de asistencia psicológica en los estados del 45 a 59. (ver Anexo)

Evaluación e Implementación

Expectativamente con estos datos podríamos determinar la procedencia y el tipo de servicios más solicitados

Las llamadas provenientes de Estados Unidos, pueden deberse a los migrantes que al llegar a ese país tiene pocas opciones tanto de atención médica como de pasar por el proceso de adaptación a la nueva cultura y modo de vida.

[2] Evangelista, M. (n.d.). K-means. En Minería de datos. Recuperado 2 de noviembre, 2024, de URL: https://miguelevangelista.gitbook.io/mineria-de-datos/tecnicas-de-mineria-de-datos/agrupamiento/k-means

Conclusión

La metodología CRISP-DM nos permite visualizar previo al inicio del proyecto los requerimientos y necesidades de: software, hardware, acceso a la información y técnicas que deberemos cubrir para ejecutar el proyecto correctamente, así como ejecutarlo de manera más procedural que nos permitirá identificar fallas y donde corregirlas.

Así mismo al contar con una documentación podemos anexar apéndices y anexos que faciliten su comprensión y lectura para futuras revisiones y mantenimiento.