**Analítica de datos y herramientas de inteligencia artificial**

**TI 3002C.501**

**Actividad 4:**

**Extracción de características**

**Luis Pablo Padilla Barbosa A00572040**

28 de abril del 2023

**Limpieza de datos**

Para el tratamiento de los valores nulos, se aplicó un get\_numeric\_data para obtener todos los datos que sean numéricos y se agregaron a una variable para poderlos limpiar de forma segregada a los demás. Posteriormente para poder acceder a los datos categóricos se aplicó un dropna tomando en cuanta las columnas de la variable que contenía todos los valores numéricos y nuevamente se agregaron estos datos a una nueva variable.

Adicional a ello, para poder determinar las columnas que contaran con la mayor contidad de registros y poderlas usar al momento de hacer el análisis de los datos, utilizando la variable con los datos categóricos se aplicó un ciclo For en donde se iban eliminando todas aquellas columnas que contaran con menos de 90 registros no nulos, para de esta forma únicamente considerar dentro de nuestro análisis aquellas columnas que contaran con la mayor cantidad de datos posibles, ya que son las que realmente se podrían analizar.

En cuanto a la limpieza de datos nulos, se aplicó para los datos numéricos la función fillna en donde se rellenaban todos los valores nulos con la media de la columna en cuestión de forma general y posteriormente, se eliminaron aquellas columnas que no tuvieran ningún valor no nulo, ya que son columnas que se encuentran vacías y que no sirven para nuestro análisis.

Por otro lado, para los categóricos lo que se hizo fue de nueva cuanta usar la función fillna, pero ahora se completaban los valores nulos con la leyenda “Sin categoría” con de que al momento de que los vayamos a analizar, podamos conocer cuantas personas no contestaron esa pregunta o parte de la encuenta.

Una vez que se tenían todos los datos limpios, se empezó a hacer el tratamiento de los valores atípicos o outliers. Esto se hizo mediante el método de los cuantiles, puesto que de esta forma se abarcan muchos valores atípicos que por medio del método de varianza. Una vez que se cambiaron todos los outliers por valores nulos, estos se reemplazaron por el promedio de la columna original, dejando de esta forma los datos límpios.

Ya para terminar con la limpieza de los datos y poder usarlos para hacer el análisis de las columnas, se utilizó un concat con la finalidad de juntar los valores categóricos con los valores numéricos y de esta forma ahora si quedarnos únicamente con un DataFrame que contuviera todos los datos, pero ahora sin valores nulos, ni outliers.

**Análisis de los datos**

Para el análisis de las columnas, se consideraron al azar 20 columnas de la lista con el nombre de las mismas que se había generado al momento de hacer la limpieza para conocer cuales eran aquellas que contaban con más de 90 registros. Posteriormente, se generó para cada una de ellas una tabla de frecuencia y también un gráfico de pastel, ya que es el que se consideró como la mejor opción para visualizar los datos categóricos con la cual podría ser mucho más fácil interpretar los mismos.

Tomando en cuenta lo anterior, se llegaron a las siguientes conclusiones:

1. La mayoría de los datos recolectados provienen de tiendas de abarrotes como primer lugar y de fonditas, alcanzando casi un 50% de la muestra.
2. El 42% de la muestra no cambió el espacio de su tienda durante el año anterior a la aplicación de la encuesta y de los que si lo hicieron la moyoría lo incrementaron, lo que implica un ligero crecimiento en la industria.
3. Casi una tercera parte de los encuestados cuentan con una puerta enrejada, lo que se traduce en un incremento en la inseguridad percibida por los comerciantes.
4. Como complemento a la anterior, el 70% de los encuestados no exhibe sus productos en el exterior, por lo que se puede decir que no sienten la confianza en las personas para que nos les roben la mercancía.
5. El 60% de la muestra no vende productos frescos, por lo que se podría decir que para algunos comercios es complicado vender sus productos de forma rápida.
6. El 30% de los encuestados utiliza su teléfono para llevar un control administrativo de su negocio.
7. Solo el 70% de los encuestados llevan un control de sus finanzas mes con mes.
8. Son muy pocos los que utilizan un medio digital para llevar un control de sus gastos, la mayoría lo hace en papel o simplemente no lo hace.
9. Lo mismo del punto anterior aplica para el registro de las ventas, la mayoría lo hace con un medio análogo como el papel.
10. El 70% de los que contestaron la pregunta son un negocio formal, el resto no se encuentra registrado ante el SAT.
11. El 40% de los encuestados sufren de estrés laboral.
12. Casi el 80% de los negocios cuentan con una conexión a internet, dada ahí la importancia de la conectividad en la actualidad.
13. Cada vez más negocios pequeños están empezando a aceptar otros formas de pago adicionales al efectivo, lo que nos habla mucho sobre la inclusión financiera.
14. La mayoría de los negocios utilizan las redes sociales como un medio para gestionar la relación con sus consumidores.
15. Más del 70% de los pequeños comercios de acuerdo a la muestra, no otorgan crédito a los consumidores.
16. Solamente aproximadamente el 10% se encuentran vinculados con una plataforma de ecommerce.
17. La mayor parte de los comercios organizan sus productos al menos de forma semanal.
18. Los principales canales de comunicación con los clientes para estos pequeños comercios son las redes sociales y el Whatsapp.
19. De los negocios, la mayoría no utiliza ninguna herramienta de planeación de ventas o utiliza únicamente un cuaderno.
20. Casi el 50% de la muestra está totalmente en descuerdo con que se compartan datos sobre su inventario con los proveedores.