**Aprendizaje no Supervisado**

**Primera Entrega del Proyecto: Propuesta Inicial**

Grupo 11:

|  |  |
| --- | --- |
| Wendy Galvis Larios | José David Contreras Arenas |
| Miller Patiño Garzón | Cesar Martínez Parra |
| Luis Felipe González Vallejo |  |

**Aplicación del aprendizaje no supervisado en el estudio de quejas a la Superintendencia de Industria y Comercio en Colombia**

**Resumen**

El presente proyecto está planteado en torno al programa de acuerdos directos de la superintendencia de industria y comercio en Colombia, la cual es la entidad encargada de la protección de la competencia económica y de velar por los derechos de los consumidores, entre otras funciones. Por su parte, los acuerdos directos básicamente son instancias a las cuales los ciudadanos pueden acudir con quejas y/o reclamos respecto a algún tipo de bien o servicio que hayan adquirido. En dichas instancias, se establece una mediación entre el consumidor y la empresa para llegar a una solución respecto al problema. Se cuenta con un registro histórico de casos interpuestos ante esta entidad entre los años 2020 y 2024, con información sobre el departamento y ciudad/ municipio del hecho, el motivo de la queja, información socioeconómica sobre la persona que interpone la queja, el estado de la petición, y el tiempo total para la resolución de la queja, entre otros. Dicho esto, este proyecto está enmarcado en el interés por comprender la dinámica detrás de este mecanismo de resolución de problemas, así como por encontrar los factores que influencian el resultado final de una queja. Para este fin, se hace uso del aprendizaje no supervisado y otras herramientas matemáticas como reducción de la dimensionalidad, con el propósito de descubrir relaciones previamente no conocidas en los datos, y posiblemente descubrir ineficiencias y/o limitaciones en los procesos de la entidad en torno al manejo de diferentes circunstancias en esta instancia de acuerdos directos, que le permitan prestar un mejor servicio y asegurar el cumplimiento de sus funciones como entidad reguladora del comercio y la buena competencia en Colombia.

**Introducción**

El presente proyecto se enmarca en los servicios de atención al ciudadano prestados por la Superintendencia de Industria y Comercio de Colombia, entidad encargada de proteger los derechos de los consumidores conforme al Estatuto del Consumidor [1]. Uno de los servicios fundamentales ofrecidos por la Superintendencia es la facilitación de la resolución de conflictos entre consumidores y proveedores a través del mecanismo de "Arreglos Directos". Este proceso permite que las quejas y disputas, relacionadas con problemas como incumplimiento de garantías, publicidad engañosa, daños por servicios prestados o la existencia de cláusulas abusivas, sean resueltas de manera amistosa entre las partes involucradas con la mediación de la Superintendencia [2].

A pesar de los esfuerzos para mejorar la calidad del servicio y la eficiencia en la resolución de estos conflictos, se ha identificado una oportunidad de optimización mediante el análisis de los datos históricos de casos de "Arreglos Directos". Para este proyecto, como se menciona en el resumen, se cuenta con un conjunto de datos confidenciales proporcionado por la Superintendencia de Industria y Comercio, que incluye los casos interpuestos ante la entidad por la ciudadanía desde el año 2020 hasta el 2024. El objetivo de este proyecto es aplicar técnicas de aprendizaje no supervisado para descubrir patrones ocultos en estos datos, con el fin de mejorar los procesos de mediación y optimizar la asignación de recursos de la entidad. La pregunta clave que este proyecto busca responder es: **"¿Cómo podemos utilizar técnicas de aprendizaje no supervisado para identificar patrones en los datos de casos de 'Arreglos Directos' que puedan ayudar a mejorar la eficacia de la mediación?"**

Nuestro análisis se centra principalmente en el uso de algoritmos de *clustering*, que pertenecen al ámbito del aprendizaje no supervisado y nos permiten identificar agrupaciones en los datos sin necesidad de etiquetas predefinidas. Con esto, se espera no solo mejorar la tasa de resolución de conflictos, sino también proporcionar un mejor servicio a los consumidores y proveedores, mejorando la eficiencia operativa de la Superintendencia de Industria y Comercio. Este análisis permitirá a la entidad anticipar los resultados de las mediaciones, gestionar mejor los recursos y ofrecer una experiencia más transparente y predecible a los ciudadanos.

Respecto a la literatura relacionada, a nivel internacional se encuentra que el uso de técnicas avanzadas de análisis de datos, como el clustering, se ha convertido en una herramienta poderosa para comprender y mejorar la gestión de quejas. Estudios recientes, como el de Xu et al. (2018), han empleado técnicas de machine learning, como K-means, para clasificar órdenes de quejas en el servicio al cliente móvil. Investigaciones similares en diferentes sectores, como el análisis de Chugani et al. (2018) sobre quejas en el sector bancario y el trabajo de Hansen, Wilke y Zaichkowsky (2010) sobre la gestión de quejas en minoristas, evidencian cómo estas técnicas pueden proporcionar información valiosa para mejorar el desempeño empresarial. Estos enfoques apoyan la elección de K-means para nuestro análisis, aunque también se considerará la combinación de diversos algoritmos, como el clustering jerárquico, para enriquecer la interpretación de los datos.

A nivel nacional, se encontraron referencias en trabajos de grado, como el de López Stan (2022), quien buscó identificar la criticidad de los clientes con afectación de servicios en la corporación TIGO mediante el uso de K-means. Su motivación fue lograr que la compañía pudiera dirigirse adecuadamente a sus clientes según sus necesidades y preferencias, aumentando así la satisfacción del servicio. Por otra parte, De La Rosa Ramos (2023) utilizó datos de quejas interpuestas por los usuarios de las Entidades Promotoras de Salud (EPS) para realizar un análisis de *clusters* que identificara y caracterizara grupos representativos de EPS según su desempeño, con el fin de identificar fortalezas y debilidades de cada grupo.

**Descripción detallada de los datos**

El conjunto de datos que se analiza está contenido en el archivo ARREGLO\_DIRECTO.csv, que incluye 3991 registros de casos interpuestos ante la Superintendencia de Industria y Comercio de Colombia entre 2020 y 2024. Estos datos proporcionan información sobre diversos aspectos de los casos de quejas y reclamos, y están organizados en 34 variables diferentes. En el siguiente enlace, se detalla el diccionario de datos para comprender mejor el contenido y la estructura del archivo [Diccionario de Datos.](https://docs.google.com/document/d/16TRoUxR0aMxm12S2qj4xc_9TtB_cNOzi/edit?usp=sharing&ouid=112448246161884979270&rtpof=true&sd=true)

**Información Descriptiva de las Variables**

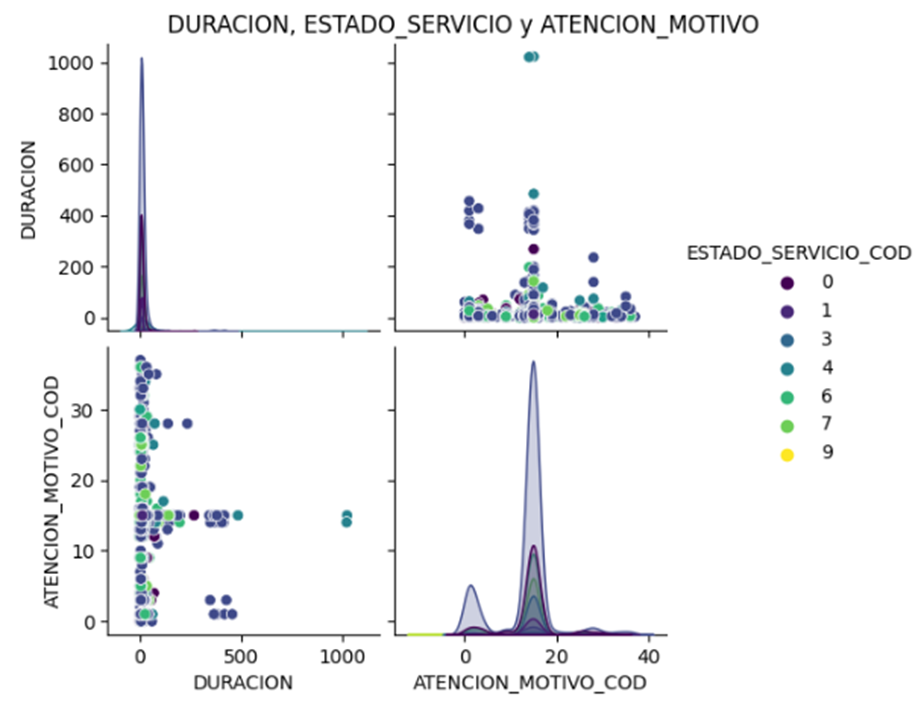
A continuación, se describen las variables más relevantes del conjunto de datos:

1. La variable **ATENCION\_MOTIVO** tiene 3991 observaciones en el conjunto de datos. Esta variable es categórica y presenta 38 categorías únicas, cada una representando un motivo específico para la atención o consulta. La categoría más frecuente es "Incumplimiento de los términos de garantía", que aparece en el 61% de los casos registrados. Este motivo es el más común entre los usuarios que han interpuesto quejas, lo que puede reflejar una tendencia predominante en las razones de atención a la entidad.
2. La variable **DURACIÓN** indica la cantidad de días que se demora un registro en ser finalizado y tiene 3991 observaciones. Su media es de 15.6 días, con una desviación estándar de 38.9 días, lo que sugiere una gran variabilidad en el tiempo de resolución de los casos. Los valores de duración varían desde un mínimo de 0 días hasta un máximo de 1022 días. El percentil 25 se sitúa en 6 días, la mediana es de 9 días, y el percentil 75 es de 15 días. Estos estadísticos proporcionan una idea clara de la distribución del tiempo requerido para resolver los casos, destacando tanto los tiempos cortos como los prolongados.

|  |  |
| --- | --- |
| **DURACIÓN (días)** | |
| Mínimo | 0 |
| Máximo | 1022 |
| Media | 15.6 |
| Desviación estándar | 38.9 |
| Percentil 25 | 6 |
| Percentil 50 | 9 |
| Percentil 75 | 15 |

1. La variable **ESTADO\_SERVICIO** se refiere al estado actual del servicio o caso y cuenta con 3991 observaciones. Es una variable categórica con 10 categorías únicas, representando distintos estados del caso. La categoría más frecuente es "CERRADO", que aparece en 49% casos. Esta categoría dominante indica que una parte significativa de los casos ha sido cerrada. Otras categorías son “ACORDADO”, “CERRADO-ICOM”, entre otras.

En la siguiente grafica se observa la relación entre la duración de los casos, el motivo de la atención y el estado del servicio, destacando una concentración significativa de casos con tiempos de resolución cortos asociados a motivos específicos. La codificación por colores sugiere que ciertos estados del servicio podrían estar correlacionados con tiempos de resolución. Además, la distribución de las variables indica posibles clústeres naturales y outliers, lo que sugiere la aplicabilidad de técnicas de clustering o reducción de dimensionalidad para identificar patrones y tendencias en la resolución de casos.



**Metodología**

Para la resolución del problema planteado, se proponen tres diferentes alternativas de algoritmos de aprendizaje no supervisado, cada una con sus ventajas y desventajas en la aplicación actual.

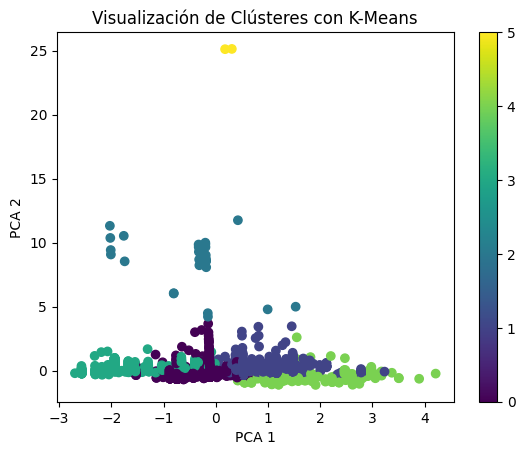
* **K-means Clustering:** Es un algoritmo eficiente, relativamente simple de implementar, y de fácil interpretabilidad, ya que proporciona clusters claramente definidos que pueden ser analizados para entender las características comunes de las observaciones dentro de cada grupo. Esto ayudará a identificar patrones problemáticos dentro del contexto del proceso.

Una desventaja es que requiere definir el número de clusters previamente, por lo cual este parámetro se ajustará según las necesidades y los resultados progresivamente obtenidos del análisis. Esto permitirá explorar diferentes niveles de granularidad en los datos e identificar la idoneidad del algoritmo para este tipo de datos.

* **K-medoides**: Este algoritmo funciona de forma similar a k-means, en cuanto está basado en centroides. Su ventaja principal respecto al anterior es la robustez a datos atípicos, puesto que se ubica el centroide en la observación central del cluster, evitando influencia de datos atípicos. Su desventaja principal es un costo computacional mayor que k-means.
* **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**: Este método es útil para encontrar clusters de formas variadas y manejar outliers de manera efectiva. En este caso, podría identificar casos excepcionales que no se ajustan bien a los patrones comunes, lo cual es relevante para identificar problemas atípicos. Una desventaja es que no reconoce bien clusters de densidad variable.

**Resultados y discusión**

* **K-means:** Se identificó 6 clusters con un Silhouette Score de 0.63, mostrando buena diferenciación entre quejas. Algunos clusters presentaron más casos cerrados, indicando eficiencia, mientras otros mostraron más reprogramaciones, señalando áreas de mejora en la gestión. Esto facilita optimizar recursos y tiempos de respuesta.



* **K-medoides:** Con relación a la implementación del algoritmo, se puede concluir que la repetida formación de un solo clúster sugiere que K-medoids podría no ser el algoritmo más adecuado para este conjunto de datos particular, posiblemente debido a su sensibilidad a la inicialización y a las características de los datos.
* **DBSCAN:** Se implementó el algoritmo DBSCAN utilizando la distancia de Gower para manejar variables mixtas. Para determinar el valor de eps, se utilizó el método del codo, identificando un valor óptimo de eps = 0.0214 y estableciendo min\_samples = 3 para capturar clusters pequeños y relevantes. El análisis reveló que en los casos de CHAPINERO, 5 de 6 casos presentaron duraciones significativamente más largas que la media, mientras que el caso con el VALOR\_PRODUCTO más alto tuvo una duración mucho más corta. Esto sugiere una posible disparidad en la gestión de los casos, donde aquellos de mayor valor parecen recibir prioridad, lo que indica una falta de equidad en los procesos de resolución, que debería ser analizada por la entidad.

Interfaz de usuario gráfica

Descripción generada automáticamente

**Conclusiones**

Para concluir, en el ejercicio actual se emplearon diversas técnicas de aprendizaje no supervisado aplicado a datos reales de quejas interpuestas por ciudadanos ante la Superintendencia de Comercio, con el fin de detectar patrones previamente desconocidos en los datos y posiblemente ofrecer recomendaciones a la entidad sobre cómo mejorar sus prácticas y garantizar de una mejor manera el cumplimiento de la libre competencia económica y todas las garantías que esta conlleva. Si bien no todos los métodos fueron exitosos, los resultados encontrados sugieren que los casos resueltos exitosamente tienen características similares entre sí en lo que respecta el valor del bien o servicio en cuestión, e incluso el área en donde se interpone la queja, y que los diferencian de los casos de mayor duración y los no resueltos. Esta información es de utilidad para la empresa, que puede usarla para plantearse presentar un servicio más equitativo a los ciudadanos, pues lo ideal es que todos los casos recibieran una atención y prioridad similar, sin discriminación hacia los casos que involucran, por ejemplo, mayores sumas de dinero en las transacciones. Por último, se aclara que este es un ejercicio netamente académico y los datos empleados en él son estrictamente confidenciales y sin permiso de publicación de parte de la SIC.

**Repositorio de Github**

Link: <https://github.com/Miller2915/ApreNoSupe>

**Bibliografía (formato APA)**

1. **Superintendencia de Industria y Comercio.** *Nuestra entidad*. Superintendencia de Industria y Comercio. Consultado el 31 de agosto de 2024, disponible en <https://sedeelectronica.sic.gov.co/nuestra-entidad#:~:text=La%20Superintendencia%20de%20Industria%20y,sus%20funciones%20administrativas%20y%20jurisdiccionales>.
2. **Superintendencia de Industria y Comercio**. *Arreglo directo*. Superintendencia de Industria y Comercio. Consultado Agosto 31, 2024, disponible en <https://www.sic.gov.co/red-nacional-de-proteccion-al-consumidor/servicios/arreglo-directo>
3. Xu, J., Zhao, J., Zhao, N., Xue, C., Fan, L., Qi, Z., & Wei, Q. (2018). The research and construction of complaint orders classification corpus in mobile customer service. En M. Zhang et al. (Eds.), *Natural Language Processing and Chinese Computing* (pp. 351–361). Springer. <https://doi.org/10.1007/978-3-319-99501-4_31>
4. Chugani, S., Govinda, K., & Ramasubbareddy, S. (2018). Data analysis of consumer complaints in banking industry using hybrid clustering. *In 2018 Second International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 848–763). IEEE. <https://doi.org/10.1109/ICCMC.2018.8487638>
5. Hansen, T., Wilke, R., & Zaichkowsky, J. (2010). Managing consumer complaints: differences and similarities among heterogeneous retailers. *International Journal of Retail & Distribution Management*, *38*(1), 6-23.
6. López Stan, S. (2022). *Segmentación de clientes con indisponibilidad o afectación en sus servicios, Área Data Analytics TIGO* (Trabajo de grado profesional, Ingeniería de Sistemas, pregrado, Universidad de Antioquia, Ciudad Universitaria).
7. De La Rosa Ramos, Y. R. (2023). *Análisis multivariado del desempeño de las EPS en Colombia: Un estudio basado en peticiones, quejas, reclamos y denuncias de usuarios (2019-2022)* (Tesis de maestría, Universidad Tecnológica de Bolívar). Universidad Tecnológica de Bolívar.