UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

SEGMENTACIÓN DE CLIENTES BASADOS EN PATRONES DE MOVILIDAD PARA MEJORAR LA SATISFACCIÓN DE LOS CLIENTES EN UNA EMPRESA DE TELECOMUNICACIONES

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

CRISTIAN ANDRÉS BARRERA MUÑOZ

PROFESOR GUÍA:
NICOLÁS CISTERNAS GONZÁLEZ

MIEMBROS DE LA COMISIÓN:

RUBÉN DAZA BARRA

DANIEL VARELA LÓPEZ

SANTIAGO DE CHILE
2024

RESUMEN DE LA MEMORIA PARA OPTAR AL TÍTULO DE: Ingeniero Civil Industrial
ESTUDIANTE: Cristian Andrés Barrera Muñoz
FECHA: 2024
PROFESOR GUÍA: Nicolás Cisternas

SEGMENTACIÓN DE CLIENTES BASADOS EN PATRONES DE MOVILIDAD PARA MEJORAR LA SATISFACCIÓN DE LOS CLIENTES EN UNA EMPRESA DE TELECOMUNICACIONES

En respuesta a la creciente competencia en telecomunicaciones, Entel ha lanzado un proyecto estratégico para mejorar la satisfacción del cliente mediante la segmentación basada en patrones de movilidad. El objetivo es “desarrollar un modelo de clusterización para identificar el impacto en la satisfacción de la movilidad en la red móvil de los clientes pospago de Entel”, permitiendo a la empresa obtener información relevante para mejorar la satisfacción del cliente mediante el análisis de patrones de movilidad.

Para este proyecto, se desarrollaron distintos modelos de segmentación de clientes siguiendo la metodología CRISP-DM. Se emplearon dos algoritmos de clustering: k-means y DBSCAN, con los cuales se crearon tres modelos con k-means, configurados con 3, 4 y 5 clusters, y un modelo con DBSCAN. No obstante, los resultados del modelo DBSCAN fueron deficientes debido a la generación excesiva de clusters, lo que complicó el análisis de los grupos generados. Por esta razón, se optó por trabajar únicamente con los tres modelos de k-means, entrenándolos con los mismos datos para que cualquier diferencia en el rendimiento se pudiera atribuir a las características de los modelos y no a variaciones en los datos. Además, se llevaron a cabo transformaciones de variables y eliminación de valores atípicos para optimizar la capacidad de agrupamiento de los modelos.

Estos modelos fueron evaluados mediante 3 métricas de evaluación de clusters, las cuales miden de distintas formas la cohesión, o cuánto se parecen los datos de un mismo cluster, y la separación, o cuánto se diferencian los datos de distintos clusters. A partir de esto, se eligió el modelo de k-means con 4 clusters, dado que este modelo presenta un equilibrio entre estas tres métricas y un equilibrio entre complejidad y simplicidad para el análisis, logrando identificar diversos segmentos de clientes con patrones de movilidad distintos.

Finalmente, se realiza un análisis de satisfacción de cada grupo creado por el modelo. Aunque no se logra validar la hipótesis inicial la cual dice que “a mayor movilidad, menor satisfacción”, el proyecto permite identificar segmentos de clientes con satisfacción crítica, generando información como priorizar inversiones, diseñar campañas específicas y personalizar ofertas según cada segmento identificado.

DEDICATORIA

A todas las metas y sueños que han guiado mi camino y a los que seguirán inspirándome a seguir adelante.

AGRADECIMIENTOS

Esta sección va dirigida principalmente a mi núcleo familiar. Mi mamá, mi papá, y mis hermanos, quienes me han acompañado todo este largo camino. Desde que tengo memoria siempre me han apoyado y confiado en mí. Han sido una fuente de aprendizaje infinita para el “ramo” más difícil que ninguna entidad educativa imparte, el cual es la vida. Sin dudas son el pilar fundamental de todo este proceso, y también son la principal motivación para llegar al final de este camino. Muchas gracias por todo y espero algún día poder devolverles todo lo que me han dado.

También, dedicada a toda mi familia. Mis primos, primas, tíos, tías, abuelas, etc. Quienes, aunque el contacto no sea tan seguido, siempre están cuando uno lo necesita, brindando apoyo sin pedir nada a cambio, aún más en los momentos difíciles. Muchas gracias por siempre estar presentes.

A todos mis amigos, los del colegio, los de la infancia, los que hice en la universidad, en trabajos, en mis prácticas profesionales, a los que siguen conmigo, y a los que, por alguna otra razón u otra ya no están cerca. Muchas gracias por todos los buenos momentos y recuerdos.

A los profesores guía y co-guía, Nico y Rubén, quienes me acompañaron en este último trazo para poder terminar de la mejor manera posible. Poseen un conocimiento enorme y me hacían ver cosas que yo no veía en relación con este trabajo. Muchas gracias por el apoyo en esta etapa final.

Finalmente, y no menos importante, a todas las personas que aportan valor a la universidad, desde los profes, alumnos y funcionarios, hasta las personas que venden almuerzos afuera de la universidad que siempre te atienden de muy buena manera. Dentro de este último grupo, me gustaría destacar al “tío” Juan de los camarines, a quien conocí el primer año de universidad y siempre ha sido una excelente persona, preocupado por su trabajo y por todas personas que lo rodean. Espero que algún día lo puedan reconocer más allá de una mención en una memoria o tesis, y muchas gracias por todas las conversas y palabras de apoyo cuando notaba que algo no estaba bien con uno.

TABLA DE CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. ANTECEDENTES GENERALES

1.1. CONTEXTO DE LA EMPRESA Y MERCADO DE TELEFONÍA MÓVIL

Entel, una empresa con una trayectoria de 59 años en el campo de la tecnología y las telecomunicaciones, destaca como una de las principales sociedades anónimas en la Bolsa de Santiago, con operaciones en Chile y Perú, una base de más de 20,2 millones de abonados móviles e ingresos anuales por CLP 2.573.142 millones en 2023 (Empresa Nacional de Telecomunicaciones, 2024).

Reconociendo su papel más allá de la conectividad, Entel se compromete a acercar la tecnología a las personas y promover el desarrollo socioeconómico y la sostenibilidad. Su enfoque en la digitalización inclusiva y la generación de oportunidades se refleja en su amplia oferta de servicios de conectividad móvil y fija, además de una amplia gama de soluciones digitales e informáticas para personas y empresas de todos los tamaños. Esta variedad de servicios busca ofrecer experiencias sorprendentes respaldadas por una infraestructura de última generación y una sólida reputación de marca, lo que garantiza experiencias de calidad para sus clientes y consolida su posición como líder en el sector (Empresa Nacional de Telecomunicaciones, 2024).

En el mercado de telefonía móvil, Chile cuenta con un total de 133,4 líneas de telefonía móvil por cada 100 habitantes, lo que se traduce en 26,7 millones de líneas totales de telefonía móvil a diciembre de 2023, lo que refleja un mercado considerablemente grande y en constante evolución. En este mercado, Entel emerge como el líder indiscutible con una participación del 32,03%, seguido de cerca por Movistar con el 26,64%, WOM con el 21,55% y Claro con el 17,92%, entre los actores con mayor participación de mercado de telefonía móvil en Chile. Estos cuatro operadores principales abarcan el 98,14% del mercado de telefonía móvil en Chile (Subsecretaría de Telecomunicaciones, 2024).

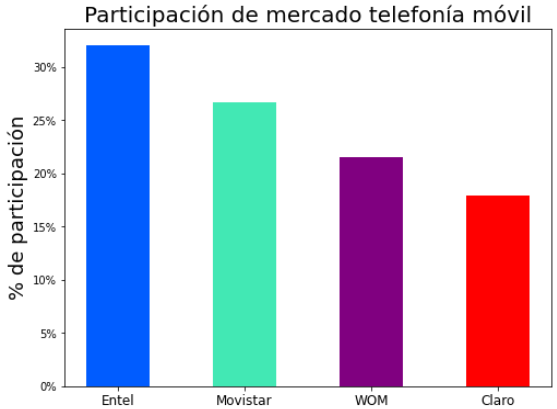

Figura 1: Participación de mercado de telefonía móvil

1.1.1. MERCADO DE TELEFONÍA MÓVIL POSPAGO

Como se mencionó anteriormente, Chile cuenta con 26,7 millones de líneas de telefonía móvil, de los cuales, el 70,19% corresponde a líneas de servicios de telefonía móvil pospago o con contrato, lo que se refiere al segmento de clientes que pagan por el servicio después de haberlos consumido, y el 29,81% restante corresponde a líneas de servicios de telefonía prepago, lo que corresponde al segmento de clientes que pagan por adelantado el servicio (Empresa Nacional de Telecomunicaciones, 2024).

Ahondando en el mercado de telefonía móvil pospago, Chile cuenta con un total de 93,63 de líneas de telefonía móvil pospago por cada 100 habitantes, lo que corresponde a 18,75 millones de líneas totales en este mercado a diciembre de 2023 (Subsecretaría de Telecomunicaciones, 2024), generando ingresos de $1.773.871 millones de pesos (Empresa Nacional de Telecomunicaciones, 2024).

En este mercado de telefonía móvil pospago, Entel lidera con una participación del 32,33%, seguido por Movistar con el 25,9%, WOM con el 21,39% y, finalmente, Claro con el 18,05%. Estos cuatro operadores principales abarcan el 97,66% del mercado de telefonía móvil pospago en Chile, demostrando la sólida posición de Entel en este mercado (Subsecretaría de Telecomunicaciones, 2024).

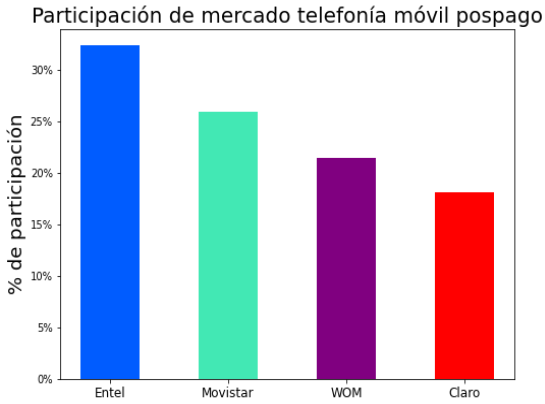

Figura 2: Participación de mercado de telefonía móvil pospago

1.2. RED MÓVIL Y PATRONES DE MOVILIDAD

La red móvil de Entel en Chile se fundamenta en una sólida infraestructura de antenas distribuidas estratégicamente en todo el país, abarcando tanto áreas urbanas como rurales. Estas antenas generan zonas de cobertura que garantizan la factibilidad del servicio, permitiendo a los usuarios una conexión estable y confiable. Cuando un dispositivo móvil se encuentra dentro del alcance de estas zonas de cobertura, se establece una comunicación con la red de Entel, donde los datos son transmitidos a

través de antenas y centrales telefónicas para su procesamiento y posterior envío al destino deseado. Este sistema de red proporciona una conexión eficiente y de alta calidad, facilitando la comunicación y el acceso a servicios en línea para los usuarios de Entel en todo el territorio chileno (Entel, n.d.).

Figura 3: Ejemplo de zonas de cobertura

En este contexto, los patrones de movilidad se refieren a los comportamientos y tendencias en el desplazamiento de los usuarios dentro de la red. Estos patrones incluyen cómo y cuándo los usuarios se mueven entre diferentes áreas de cobertura, y cómo este movimiento se relaciona con la utilización de los recursos de la red.

1.3. DESCRIPCIÓN DEL PROBLEMA U OPORTUNIDAD

Entel se enfrenta al desafío de comprender y mejorar la satisfacción de sus clientes pospago de servicios móviles. Este objetivo se alinea estrechamente con la misión de la empresa, que busca crear valor y contribuir al bienestar de la comunidad, incrementando la productividad y calidad de vida de las personas. En este sentido, la empresa reconoce la necesidad de comprender cómo la movilidad influye en el nivel de satisfacción de sus clientes. La movilidad, entendida como la frecuencia y ubicación de los desplazamientos de los clientes, puede afectar significativamente su experiencia con los servicios de telecomunicaciones. Por ejemplo, una persona con alta movilidad puede experimentar cambios en las conexiones a las antenas mientras viaja, lo que puede resultar en interrupciones en la conectividad cuando pasa de una antena a otra.

La hipótesis sugiere una relación inversa entre la movilidad de los clientes y su satisfacción, es decir, a menor movilidad, mayor satisfacción percibida, y viceversa. Esto se debe a que los clientes con una movilidad reducida pueden experimentar menos interrupciones en el servicio y una mayor consistencia en la calidad de la conexión. Esta estabilidad en la experiencia del usuario tiende a generar una percepción más positiva de la calidad del servicio y, por ende, una mayor satisfacción por parte del cliente.

Para abordar esta problemática, un modelo de clusters o grupos de clientes basados en patrones de movilidad y experiencia del cliente puede proporcionar una comprensión más precisa de cómo la movilidad afecta la satisfacción del cliente en los servicios de telecomunicaciones de Entel debido a que categoriza a los clientes en grupos con comportamientos de movilidad similares. Al identificar grupos de clientes con comportamientos de movilidad similares y evaluar la experiencia de servicio en cada grupo, la empresa puede dirigir de manera efectiva sus recursos de inversión hacia áreas específicas donde se necesite mejorar la calidad de la conexión y reducir interrupciones. Esto permite una asignación más eficiente de recursos, asegurando que las mejoras se implementen donde tengan el mayor impacto en la satisfacción del cliente.

Es importante destacar que este enfoque en la mejora de la satisfacción del cliente, en lugar de centrarse en la retención de clientes, refleja el compromiso de Entel con su misión empresarial y su deseo de generar valor para la comunidad a través de la mejora continua de sus servicios.

2. DESCRIPCIÓN Y JUSTIFICACIÓN DEL PROYECTO

El proyecto se centra en la implementación de un modelo de clusterización de clientes, con el fin de identificar grupos de clientes con distintos niveles de satisfacción y patrones de movilidad. Esta solución se presenta como una estrategia efectiva para abordar la falta de claridad sobre la relación entre movilidad y satisfacción del cliente, debido a que se pueden detectar segmentos de clientes que comparten características similares en términos de movilidad y analizar la satisfacción de estos grupos, permitiendo comprender de mejor manera los comportamientos de cada segmento.

Al comprender mejor estos patrones, la empresa estará en una mejor posición para tomar decisiones informadas que mejoren la experiencia del cliente y optimicen sus operaciones. Además, es relevante ya que proporciona información valiosa sobre cómo distribuir las inversiones de la empresa, permitiendo asignar recursos de manera más eficiente y dirigirlos hacia áreas que realmente impacten positivamente en la satisfacción del cliente y en el rendimiento general de la empresa.

La implementación de esta técnica de clusterización se justifica por la necesidad de Entel de identificar y comprender las necesidades y preferencias específicas de diferentes segmentos de clientes, lo cual es crucial debido a la complejidad del mercado de las telecomunicaciones y la importancia estratégica de mantener la satisfacción del cliente como un pilar fundamental de la competitividad empresarial. Esto no solo mejorará la calidad del servicio ofrecido, sino que también contribuirá a fortalecer la relación entre la empresa y sus clientes, promoviendo así un crecimiento sostenible y una ventaja competitiva en el mercado de las telecomunicaciones.

Al recopilar y analizar datos relevantes sobre la movilidad y la satisfacción del cliente, Entel podrá desarrollar recomendaciones específicas destinadas a mejorar la experiencia del cliente y optimizar las operaciones de la empresa. Estas recomendaciones se basarán en una comprensión profunda de los patrones de comportamiento del cliente, lo que permitirá a la empresa abordar las necesidades individuales de manera más precisa y eficiente, permitiendo a Entel destinar inversión y adaptar sus servicios y estrategias de atención al cliente de manera más efectiva, lo que a su vez mejorará la retención de clientes y aumentará la lealtad a la marca.

Actualmente, el churn rate, o tasa de deserción, mensual de clientes en el mercado de telefonía móvil pospago orbita alrededor del 2,5%, lo que significa que aproximadamente 117.500 clientes dejan de usar los servicios de Entel cada mes en este mercado. Considerando que el ingreso promedio por usuario es de $12.000 pesos chilenos, esta pérdida de clientes se traduce en pérdidas mensuales de $1.410 millones de pesos chilenos aproximadamente.

La implementación del proyecto de clusterización tiene el potencial de reducir esta tasa de deserción al identificar segmentos críticos de clientes con alta probabilidad de churn y aplicar estrategias específicas para mejorar su satisfacción. Si se logra una reducción del 10% en la tasa de churn, esto permitiría retener aproximadamente 11.750 clientes adicionales por mes, lo cual se traduciría en un incremento mensual de $141 millones de pesos chilenos en ingresos, lo que representaría un aumento anual de $1.692 millones. Este análisis muestra el impacto financiero que el proyecto puede tener, fortaleciendo tanto la rentabilidad como la sostenibilidad de la empresa en un mercado altamente competitivo.

Este análisis muestra cómo la hipótesis planteada anteriormente puede ser clave para mejorar la rentabilidad de Entel. Al segmentar a los clientes con alta movilidad y riesgo de churn, el proyecto de clusterización permite desarrollar estrategias focalizadas que podrían reducir la tasa de churn. Además, este proyecto podría ser un punto de partida para analizar la relación entre satisfacción y tasa de churn, ya que actualmente no existe una conexión clara entre estas métricas.

3. OBJETIVOS

El objetivo general del proyecto es desarrollar un modelo de clusterización para identificar el impacto en la satisfacción de la movilidad en la red móvil de los clientes pospago de Entel.

Para esto, los objetivos específicos son los siguientes:

1. Identificar variables relevantes que puedan explicar la movilidad y satisfacción de los clientes del servicio móvil de Entel.

2. Desarrollar un modelo de clusterización que integre los datos recopilados de movilidad y satisfacción de los clientes.

3. Evaluar la efectividad del modelo mediante métricas de validación.

4. Analizar la satisfacción dentro de cada grupo identificado por el modelo.

5. Realizar un análisis del modelo con datos actualizados para identificar patrones estadísticamente significativos entre movilidad y satisfacción de clientes.

4. ALCANCES

- Debido a restricciones de tiempo, el proyecto se limitará a la evaluación y elección del modelo de clusterización que ofrezca el mejor rendimiento. Los análisis posteriores y las recomendaciones para la implementación del modelo serán entregados a la empresa como una propuesta, dejando a su discreción, y quedando fuera de los alcances de este proyecto, la implementación.

- Se trabajará exclusivamente con datos de clientes pospago de servicios móviles de Entel en Chile, abarcando aproximadamente 4.7 millones de clientes. Dado este enfoque, los resultados obtenidos en el modelo no serán extrapolables a otros mercados debido a las diferencias de comportamiento. Estas diferencias en el comportamiento se deben principalmente a que los clientes de servicios móviles prepago pueden tener comportamientos distintos a los de servicios pospago debido a la naturaleza de sus planes y el control de gastos asociados. Por ejemplo, los clientes prepago pueden utilizar los servicios de manera más selectiva, lo que puede influir en sus patrones de uso y movilidad de manera diferente a los clientes pospago, que tienen un compromiso contractual más estable y pueden acceder a servicios adicionales de manera más continua. Además, es importante considerar que los servicios fijos, como la telefonía fija o la conexión a Internet residencial, no comprenden tanta movilidad como los servicios móviles, lo que implica que los factores que influyen en los patrones de comportamiento pueden ser considerablemente diferentes.

- Los datos disponibles comprenden información recopilada por la empresa desde julio de 2023 hasta marzo de 2024. Esta limitación temporal se debe a la disponibilidad de información completa de los clientes en este periodo. Se considera información

completa a datos que son integrales y de alta calidad, garantizando la disponibilidad de la información necesaria y confiable para realizar un análisis preciso y significativo.

- La meta principal del proyecto es proporcionar información relevante para las decisiones de inversión de la empresa. Los indicadores de éxito incluyen mejoras en la satisfacción del cliente, optimización de inversiones y calidad del modelo de clusterización. La calidad del modelo se medirá mediante métricas de validación, como la precisión, la cohesión y la separación de los clusters identificados.

5. MARCO CONCEPTUAL

En el ámbito de proyectos de clusterización de clientes en el sector de las telecomunicaciones, la ingeniería industrial desempeña un papel fundamental al proporcionar un marco metodológico y herramientas analíticas para entender y mejorar los procesos empresariales relacionados con la satisfacción del cliente. La aplicación de la ingeniería industrial en este contexto implica la utilización de técnicas de análisis de datos y modelado para segmentar a los clientes en grupos homogéneos con el fin de comprender mejor sus comportamientos y necesidades. Específicamente, áreas como Data Science y análisis de datos son esenciales para identificar patrones en los datos de los clientes y extraer insights accionables para la toma de decisiones estratégicas.

A continuación, se detallarán las herramientas utilizadas en cada proceso de este proyecto.

5.1. OBTENCIÓN Y PROCESAMIENTO DE DATOS

- SQL: Lenguaje de programación diseñado para gestionar datos en bases de datos relacionales. Estas bases de datos organizan la información en tablas, donde las filas y columnas representan diferentes características y relaciones entre los datos. Con SQL, se puede ejecutar operaciones como almacenar, actualizar, eliminar, buscar y recuperar datos, así como también optimizar el rendimiento de la base de datos (Amazon Web Services, n.d.).

- Python: Lenguaje de programación que se destaca en el procesamiento de grandes volúmenes de datos debido a su facilidad de uso, su comunidad activa, sus bibliotecas especializadas y su flexibilidad para adaptarse a diferentes necesidades y escenarios. Entre las bibliotecas destacadas para el manejo de grandes conjuntos de datos están Pandas, para análisis, NumPy, para cálculos numéricos, SciPy, para cálculos científicos, y Matplotlib, para visualización de datos (Juanweb, 2023).

- Análisis exploratorio de datos: El análisis exploratorio de datos (EDA) es una técnica utilizada para examinar conjuntos de datos, identificar sus características principales y emplear métodos visuales para resumirlas. Ayuda a determinar cómo manipular los datos para obtener respuestas deseadas, permitiendo descubrir patrones, anomalías, validar hipótesis o revisar supuestos. Principalmente se usa para explorar más allá de las formalidades de modelado o pruebas de hipótesis, brindando una comprensión más profunda de las variables y sus relaciones, así como para evaluar la eficacia de técnicas estadísticas para el análisis de datos (International Business Machines Corporation, n.d.).

Para el desarrollo del proyecto actual, se utilizará SQL para acceder a los datos de los clientes de Entel en Chile y Python para la manipulación de datos, incluyendo la exploración de datos (EDA), y para desarrollar modelos de clustering, calculando las métricas necesarias. Esto permitirá una comprensión profunda de los patrones en los datos y una identificación eficiente de segmentos de clientes.

5.2. MODELADO

5.2.1. PARÁMETROS E HIPERPARÁMETROS

- Parámetros: En los modelos de machine learning, los parámetros son las variables cuyos valores se ajustan durante el proceso de entrenamiento, utilizando conjuntos de datos. A diferencia de los hiperparámetros, estos valores no son configurados manualmente por el científico de datos, sino que son obtenidos a partir de los datos. Los parámetros son esenciales en los modelos de machine learning, ya que representan la parte del modelo que aprende de los datos y son fundamentales para realizar predicciones y determinar la capacidad del modelo para resolver un problema específico. Por lo tanto, estimar correctamente estos parámetros es una tarea crucial en el proceso de entrenamiento (Rodríguez, 2019).

- Hiperparámetros: Los hiperparámetros son ajustes externos que los científicos de datos establecen para controlar el proceso de entrenamiento de modelos de machine learning. A diferencia de los parámetros, que se determinan internamente durante el aprendizaje del modelo, los hiperparámetros deben configurarse manualmente antes del entrenamiento. Estos hiperparámetros, influyen en cómo se desarrolla el proceso de aprendizaje del modelo, pero no son derivados automáticamente durante ese proceso (Amazon Web Services, n.d.).

5.2.2. MÉTODOS PARA LA ELECCIÓN DE HIPERPARÁMETROS

- Método del codo: Este método ayuda a identificar un número apropiado de clusters para modelos que tengan como hiperparámetro el número de clusters, por ejemplo, el modelo K-means. Implica calcular la inercia al aplicar el mismo modelo con diferentes cantidades de clústeres, de 1 a N. La inercia es la suma de las distancias al cuadrado de cada objeto al centroide de su clúster. Luego, se grafica la inercia en función del número de clústeres y se busca un cambio brusco en la gráfica, como el codo de un brazo. Ese punto dice cuántos clusters son óptimos para los datos (Moya, 2016). Un ejemplo gráfico de este método se ilustra en la figura 11 del anexo A.

- Método Gap Statistics: Al igual que el método del codo, este método ayuda a identificar un número óptimo de clusters, comparando la dispersión dentro del conjunto de datos original con la dispersión esperada en conjuntos de datos generados aleatoriamente. Busca el punto donde la diferencia entre estas dispersiones es máxima para determinar la cantidad óptima de clústeres. Los pasos del método incluyen seleccionar un número máximo de clústeres para evaluar, entrenar modelos de k-means para diferentes números de clústeres, calcular la diferencia de logaritmos de la varianza intra-cluster para cada número de clústeres y determinar el valor óptimo de k que maximiza esta diferencia (Rodríguez, 2023). En la figura 12 del anexo A se incluye un ejemplo gráfico de este método.

- Método de la rodilla: El método de la rodilla es una técnica utilizada para identificar el hiperparámetro óptimo épsilon en el algoritmo de clusterización DBSCAN. Este parámetro determina la distancia máxima entre dos puntos para que sean considerados parte del mismo grupo. El método calcula las distancias de cada punto a sus k vecinos más cercanos, donde k es un valor especificado por el usuario. Estas distancias se ordenan y se grafican para formar una curva. La "rodilla" en esta curva representa un punto donde hay un cambio significativo en las distancias, indicando así el valor óptimo de épsilon (González, 2020), tal como se muestra en la figura 13 del anexo A.

5.2.3. MODELOS

- K-means: El algoritmo K-means es una técnica de aprendizaje no supervisado que busca identificar grupos en datos no etiquetados mediante la asignación de datos a K clusters basados en similitudes. Opera iterativamente, calculando centroides y etiquetas para cada grupo. Destaca por su simplicidad y eficiencia computacional, produciendo clusters compactos, pero enfrenta desafíos como la dificultad para determinar el número óptimo de clusters (K) y la sensibilidad a la escala de los datos. Su rendimiento puede verse afectado por una mala inicialización de los centroides (González, 2020; Andrés, 2022).

- DBSCAN (Density-Based Spatial Clustering of Applications with Noise): DBSCAN es un algoritmo de clustering basado en la densidad que identifica clústeres al buscar áreas densas de puntos, sin necesidad de predefinir el número de clústeres. Utiliza dos hiperparámetros: Épsilon, que define la proximidad necesaria para considerar puntos vecinos, y Puntos mínimos, que establece el número mínimo de puntos para formar una región densa. Entre sus ventajas están no requerir la especificación previa del número de clústeres, la capacidad de identificar clústeres de cualquier forma, y la habilidad para manejar ruido y valores atípicos. Sin embargo, puede tener dificultades con clústeres de densidades diferentes y requiere una correcta configuración de los parámetros, siendo sensible a pequeños cambios en los mismos (Rueda, 2024; González, 2020).

Se decidió utilizar K-means como primera opción debido a su eficiencia y capacidad para manejar grandes volúmenes de datos, ideal para la base de 4,7 millones de clientes. K-means identifica grupos compactos con patrones similares, es fácil de implementar, y converge rápidamente, lo que significa que llega a una solución estable en un número reducido de iteraciones, facilitando el análisis continuo de los datos de movilidad.

Por otro lado, DBSCAN es ideal dado que no requiere especificar el número de clusters y maneja bien el ruido y valores atípicos, comunes en los datos de movilidad. También es capaz de detectar patrones de movilidad con formas complejas y arbitrarias, ofreciendo una visión más detallada de los comportamientos de los clientes.

5.3. EVALUACIÓN

5.3.1. CRITERIOS DE EVALUACIÓN

- Separación: la distancia entre los centroides, o centro, de los clusters. El propósito de este criterio es determinar clusters donde los centroides estén lo más separados posible entre sí (Ramos Ponce, 2020).

- Cohesión: Mide qué tan próximos están los datos respecto al centroide de su cluster correspondiente. El propósito es generar clusters donde se maximice la cohesión interna, es decir, que los datos estén lo más cercanos posible dentro de un mismo cluster. Una forma común de medir esta cohesión interna es a través de la suma de errores cuadráticos (Ramos Ponce, 2020).

5.3.2. MÉTRICAS DE EVALUACIÓN

- Puntuación de Silueta (Silhouette Score): El puntaje de silueta, también conocido como "Silhouette Score" en inglés, ofrece una evaluación promedio de la cohesión dentro de cada grupo y la separación entre los grupos. Cuantifica la relación entre la separación de los distintos clústeres y la similitud entre los puntos dentro de un mismo clúster, expresándose en un valor que oscila entre -1 y 1. Valores cercanos a 1 indican una excelente separación entre clústeres, mientras que los cercanos a -1 sugieren una separación deficiente (Rodríguez, 2023).

  La fórmula para calcular esta métrica es la siguiente:

$$s(i) = \frac{b(i) - a(i)}{max(a(i)\,, b(i))}$$

  donde, $a(i)$ es la distancia promedio entre el punto i y todos los otros puntos en el mismo cluster; $b(i)$ representa la distancia promedio entre el punto i y todos los puntos del cluster más cercano al que no pertenece.

- Coeficiente de Davies-Bouldin: El coeficiente de Davies-Bouldin utiliza la relación entre la dispersión dentro de los clústeres y la separación entre clústeres para evaluar la calidad del clustering. La dispersión intra-cluster evalúa la proximidad entre los puntos dentro de cada grupo, siendo deseable una baja dispersión para un agrupamiento efectivo. El índice de Davies-Bouldin se calcula como el cociente entre ambos valores, minimizándose cuando los clústeres están bien separados y compactos (Rodríguez, 2023).

  Para un conjunto de k clusters, el coeficiente de Davies-Bouldin se calcula como:

$$DB = \frac{1}{k}\sum_{i=1}^{k} max_{j \neq i} \left(\frac{S_i + S_j}{d(c_i\,, c_j)}\right)$$

  donde, k representa el número de clusters; S_i la dispersión promedio dentro del cluster i, medida generalmente como la distancia promedio entre cada punto y el centroide del cluster i; $d(c_i\,, c_j)$ es la distancia entre los centroides de los clusters i y j; $max_{j \neq i}\left(\frac{S_i + S_j}{d(c_i\,, c_j)}\right)$ es la medida de la similitud entre el cluster i y el cluster j.

- Índice Calinski-Harabasz (Calinski-Harabasz Index): El índice de Calinski-Harabasz, también conocido como el Criterio de Relación de Varianza o Variance Ratio Criterion, es una métrica utilizada para evaluar la cohesión de un conjunto de datos agrupado. Se considera una agrupación mejor cuando el valor del índice es mayor. Este índice se calcula como el cociente entre la dispersión entre clústeres, medida como la suma de las distancias al cuadrado entre los centroides de cada clúster y el centroide global, y la dispersión promedio dentro de cada clúster, evaluada mediante la suma de las distancias al cuadrado de cada punto con respecto al centroide de su clúster. Se espera que esta ratio aumente con una mejora en la agrupación, donde un valor debe aumentar mientras que otro debe disminuir (Rodríguez, 2023).

  Para un conjunto de k clusters, el índice Calinski-Harabasz se calcula como:

$$CH = \frac{B(k)}{W(k)} \times \frac{N-k}{k-1}$$

  donde, k es el número total de clusters; $B(k)$ es la dispersión entre clusters (dispersión entre los centroides de los clusters); $W(k)$ representa la dispersión dentro de los clusters (dispersión promedio dentro de cada cluster); N es el número total de muestras en el conjunto de datos.

5.4. METODOLOGÍAS

Para el desarrollo del proyecto existen distintas metodologías factibles para abordarlo, entre las que destacan, debido a su enfoque sistemático y estructurado, teniendo un gran reconocimiento dentro de la industria de la ciencia de datos, los siguientes:

- KDD (Knowledge Discovery in Databases): Es un proceso que se centra en la extracción de información útil y conocimiento significativo a partir de grandes conjuntos de datos. Se enfoca en identificar patrones, tendencias y relaciones ocultas en los datos para generar nuevos conocimientos.

  Esta metodología ofrece una aproximación integral para extraer conocimiento a partir de grandes conjuntos de datos, permitiendo la identificación de patrones y tendencias ocultas que facilitan la toma de decisiones informadas. Sin embargo, su implementación eficaz puede requerir un alto nivel de expertise técnico, y el proceso puede resultar complejo y exigir recursos considerables, como tiempo y dinero. La interpretación de los resultados también puede ser subjetiva, dependiendo en gran medida de la experiencia del analista, lo que podría afectar la objetividad de las conclusiones alcanzadas (Moine, n.d.).

- CRISP-DM (Cross-Industry Standard Process for Data Mining): Es una metodología estándar utilizada en proyectos de minería de datos que consta de seis fases principales: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue. Proporciona una estructura sistemática para guiar a los equipos a lo largo del proceso de minería de datos.

  Las principales ventajas de esta metodología se basan en proporcionar una estructura clara y sistemática para guiar a los equipos a lo largo del proceso de minería de datos, y su amplio reconocimiento en la industria facilita la colaboración y comunicación entre equipos. Además, permite una evaluación continua de los resultados y la iteración en las fases del proyecto. Sin embargo, puede resultar rígida en contextos que requieren flexibilidad en el proceso de minería de datos. La implementación completa de todas las fases puede llevar tiempo y recursos adicionales, y no aborda específicamente los aspectos técnicos detallados de la minería de datos (Moine, n.d.).

- SEMMA (Sample, Explore, Modify, Model, Asesas): Es una metodología desarrollada por el SAS Institute que se enfoca en las etapas técnicas de la minería de datos, como el muestreo, la exploración de los datos, la modificación de variables, la creación de modelos predictivos y la evaluación de los resultados. Está diseñada para trabajar de manera eficiente con las herramientas de minería de datos de SAS.

  Esta metodología se enfoca en las etapas técnicas de la minería de datos, lo cual es beneficioso para usuarios con experiencia técnica, y está especialmente diseñada para trabajar de manera eficiente con las herramientas de minería de datos de SAS. Además, proporciona un marco claro y estructurado para llevar a cabo proyectos de minería de datos. Sin embargo, puede carecer de enfoque en aspectos no técnicos, como la comprensión del negocio y la gestión de proyectos. Su orientación hacia usuarios que utilizan herramientas específicas de SAS puede limitar su aplicabilidad en entornos con otras herramientas de minería de datos, y puede no ser tan ampliamente adoptada en la industria como CRISP-DM (Moine, n.d.).

Para este proyecto, se elige la metodología CRISP-DM por su estructura clara, que facilita la colaboración, la comunicación y la evaluación continua, permitiendo que los modelos se adapten al mercado de telecomunicaciones. KDD se descartó por su complejidad y alta demanda de recursos, mientras que SEMMA fue excluida por su falta de enfoque en la comprensión del negocio y su limitada adopción fuera de las herramientas de SAS.

6. METODOLOGÍA

Profundizando en las fases del Proceso Estándar Intersectorial para la Minería de Datos, CRISP-DM por sus siglas en inglés, esta metodología consta de seis pasos, que van

desde la comprensión del problema hasta la implementación y seguimiento. A continuación, se detallarán estos pasos para el desarrollo de este proyecto.

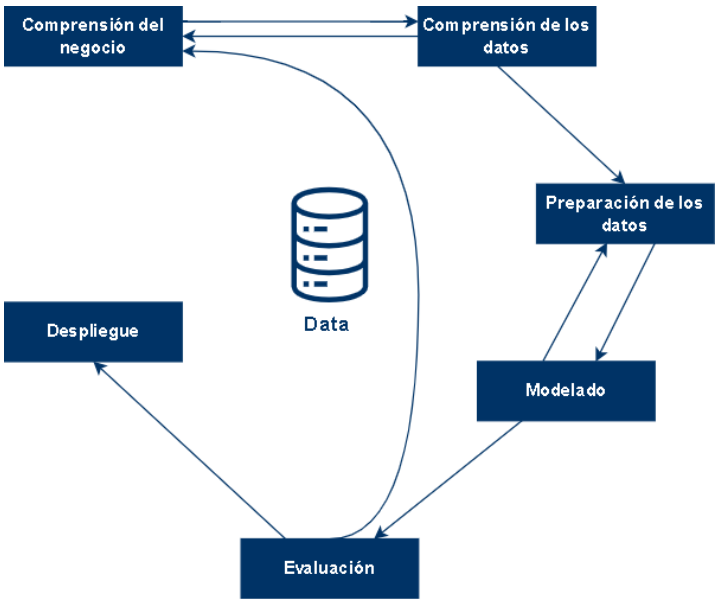

Figura 4: Metodología CRISP-DM

6.1. COMPRENSIÓN DEL NEGOCIO

En esta etapa inicial del proyecto, se comenzará identificando cómo definir la movilidad de los clientes utilizando los datos disponibles en las bases de datos de Entel. Además, se extraerán datos relacionados con la satisfacción del cliente. Se buscarán las variables o combinaciones de variables que puedan estar correlacionadas con la movilidad y satisfacción, además de algunas variables que puedan influir para caracterizar más a los clientes, y representarlas adecuadamente. El proceso de extracción de las tablas se llevará a cabo mediante consultas en SQL a la base de datos de clientes de Entel.

6.2. COMPRENSIÓN DE LOS DATOS

Se realizará un análisis exploratorio de los datos recopilados para comprender su estructura y características. Esta exploración inicial permite identificar patrones, tendencias y posibles anomalías en los datos, proporcionando información valiosa para el análisis posterior.

Se buscarán patrones y tendencias en los datos que puedan ofrecer información sobre el comportamiento de los clientes en relación con las variables de movilidad. Además, se enfocará en prestar atención a la identificación de posibles anomalías o valores atípicos que puedan afectar la calidad de los datos.

También, se analizará la distribución de las variables relevantes para comprender la variabilidad de los datos e identificar posibles sesgos en la muestra. También se examinarán las relaciones entre las variables para identificar correlaciones significativas.

Por último, se utilizarán técnicas de visualización de datos, como gráficos y diagramas, para representar visualmente las características clave de los datos. Esto facilitará la interpretación de los resultados y la comunicación de hallazgos importantes.

6.3. PREPARACIÓN DE LOS DATOS

Los datos recopilados se someterán a procesos de limpieza y transformación para garantizar su calidad y prepararlos adecuadamente para el análisis. Este paso es esencial para asegurar que los datos sean coherentes y estén en el formato adecuado para el modelado posterior.

Se eliminarán los datos incompletos o incorrectos, así como los valores atípicos que puedan afectar la precisión de los resultados del análisis. Además, se realizarán conversiones de formato si es necesario.

También se aplicarán técnicas de ingeniería de características para crear nuevas variables o modificar las existentes con el fin de mejorar la capacidad predictiva del modelo, como la normalización de variables numéricas.

6.4. MODELADO

Se aplicarán técnicas de clusterización para identificar grupos de clientes con patrones de movilidad similares y niveles de satisfacción diferentes. La clusterización permite segmentar la base de clientes, lo que facilita la personalización de servicios y estrategias de retención.

Los modelos de clusterización a utilizar serán K-means y DBSCAN. Estos modelos asignarán a cada cliente a un grupo en función de su comportamiento de movilidad y otros factores relevantes.

Para estos modelos se calcularán primero sus hiperparámetros óptimos para maximizar la calidad y la eficacia del clustering. En el caso de K-means y K-medoids, se calculará el K o número de clusters óptimo, a través del método del codo y el método de Gap

Statistics. Para el caso del modelo DBSCAN se calculará el épsilon a través del método de la rodilla.

6.5. EVALUACIÓN

Se evaluará la efectividad del modelo de clusterización utilizando métricas de validación. Esta evaluación garantiza que los grupos identificados sean significativos y relevantes para la empresa, proporcionando una base sólida para la toma de decisiones. Las métricas que se calcularán para la evaluación serán la Puntuación de Silueta (Silhouette Score), Coeficiente de Davies-Bouldin y Índice Calinski-Harabasz (Calinski-Harabasz Index).

Luego de obtener las distintas métricas para cada modelo, se elegirá el modelo con mejor desempeño a partir de estas métricas.

6.6. DESPLIEGUE

Se identificarán diferencias significativas en la satisfacción entre los grupos de clientes identificados, lo que permite formular recomendaciones específicas para mejorar la experiencia del cliente y optimizar las operaciones de la empresa. Posteriormente, se podrán analizar las características distintivas de cada grupo para comprender mejor las necesidades y preferencias de los clientes en cada segmento, lo que facilitará la personalización de estrategias y servicios para atender de manera más efectiva a cada segmento de clientes.

7. DESARROLLO Y RESULTADOS

A continuación, se presentan los resultados obtenidos como parte del proceso de análisis de datos y modelado llevado a cabo en el marco de este proyecto. Estos resultados ofrecen una visión detallada de los procesos realizados mediante la metodología descrita en la sección anterior.

7.1. COMPRENSIÓN DEL NEGOCIO

En esta sección, se presentan los hallazgos y análisis realizados sobre los factores relacionados con la movilidad de los clientes y la experiencia, basados en los datos extraídos de las bases de datos de Entel.

7.1.1. FACTORES RELACIONADOS CON MOVILIDAD DE LOS CLIENTES

Para poder seleccionar y/o crear las variables para el modelo, se analizaron los factores que podrían estar correlacionados con la movilidad de los clientes y que puedan ser obtenidos a partir de las bases de datos que se tienen a disposición. Estos factores son los siguientes:

- Cantidad de antenas a las que se conectan los clientes: Un mayor número de antenas conectadas puede indicar una mayor movilidad geográfica, ya que los clientes que se desplazan más tienden a conectarse a una mayor cantidad de antenas.

- Cantidad de antenas únicas a las que se conectan los clientes: Un mayor número de antenas únicas, o distintas, a las que se conecta cada cliente puede reflejar una mayor diversidad en las áreas visitadas por los clientes.

- Distancia máxima entre las antenas a las que se conecta cada cliente: Una mayor distancia entre antenas sugiere que el cliente se mueve a lo largo de distancias más grandes, lo que indica una mayor movilidad.

- Dispersión de la distancia de las antenas más utilizadas por cada cliente: La dispersión en las distancias de las antenas más utilizadas puede reflejar la regularidad y el alcance de los movimientos del cliente.

- Cantidad de comunas distintas en las que se encuentran las antenas más utilizadas por cada cliente: Esto indica la diversidad geográfica de las áreas frecuentadas por los clientes, sugiriendo una mayor movilidad si las antenas se encuentran en diferentes comunas.

7.1.2. FACTORES RELACIONADOS CON CARACTERÍSTICAS DE LOS CLIENTES Y LUGARES FRECUENTES

Además de los factores relacionados con la movilidad, también es crucial analizar las características de los clientes y los lugares que estos frecuentan, debido a que estos aspectos no solo proporcionan una visión más completa y detallada del comportamiento de los clientes, sino que también pueden robustecer los modelos en futuras iteraciones del proyecto. Además, estas características son fundamentales para caracterizar y diferenciar los grupos resultantes, permitiendo una segmentación más precisa y efectiva.

Para esto, se analizaron qué aspectos pueden estar relacionados con las características de los clientes y los lugares que frecuentan. Los factores considerados incluyen:

- Porcentaje del tiempo pasado en cada celda o antena: El tiempo que un cliente pasa conectado a una determinada celda o antena puede indicar la importancia de esa área en su vida diaria. Un alto porcentaje de tiempo en una celda específica sugiere que el cliente frecuenta esa ubicación, lo cual puede ser su hogar, lugar de trabajo o un lugar de interés recurrente. Esto proporciona una idea sobre los patrones de movilidad y las áreas clave en la vida del cliente, ayudando a caracterizar sus hábitos y rutinas.

- Calidad de cobertura de lugares frecuentes: La calidad de la cobertura (score 3G y 4G) en los lugares frecuentes de un cliente puede influir significativamente en su satisfacción y experiencia de uso. Clientes que pasan mucho tiempo en áreas con buena cobertura probablemente tengan una mejor experiencia de servicio, lo que puede correlacionarse con su lealtad y satisfacción. Además, las diferencias en las calidades de cobertura pueden ayudar a identificar áreas de mejora en la infraestructura de red.

- Tipo de área de lugares frecuentes (urbano o rural): Saber si los lugares frecuentes de un cliente están en áreas urbanas o rurales puede ofrecer información valiosa sobre su estilo de vida y necesidades. Los clientes en áreas urbanas pueden tener diferentes expectativas y patrones de uso de servicios móviles en comparación con los clientes en áreas rurales.

- Cantidad de antenas que un cliente se conecta por hora: La frecuencia con la que un cliente cambia de antena o se reconecta a la misma puede ser un indicador clave de su comportamiento de conexión. Un cliente que se conecta a muchas antenas por hora o que se conecta frecuentemente a la misma antena muestra una alta dinámica en su uso de la red. Esto puede reflejar patrones de uso intensivo y necesidades específicas de conectividad continua y de alta calidad.

  Por ejemplo, un cliente que permanece conectado a una sola antena durante largos períodos puede tener hábitos de uso estacionario, como trabajar desde casa o estar en un lugar fijo. En contraste, un cliente que cambia de antena o se conecta varias veces puede estar en constante movimiento, dependiendo de la red para diferentes actividades en múltiples ubicaciones.

- Tipo de plan que utiliza el cliente (precio del plan): El tipo y precio del plan que un cliente utiliza puede reflejar su perfil económico y sus expectativas de servicio. Los clientes con planes más caros suelen esperar una mejor calidad de servicio y una experiencia de usuario superior. Debido a estas expectativas elevadas, estos clientes

tienden a ser más exigentes y sensibles a la calidad del servicio que reciben. Por lo tanto, cualquier deficiencia en el servicio puede tener un impacto significativo en su nivel de satisfacción. Por otro lado, los clientes con planes más económicos pueden tener diferentes patrones de uso y prioridades, enfocándose más en la relación calidad-precio.

Estos factores fueron seleccionados debido a su capacidad para proporcionar una visión integral del comportamiento, expectativas y necesidades de los clientes, así como su relevancia directa para la calidad de la experiencia del servicio móvil.

7.1.3. VARIABLES

A partir del análisis anterior, se crea una tabla que recopila las variables a utilizar en la creación del modelo. Estas variables se obtienen de distintas bases de datos contenidas en tablas que se tienen a disposición, las cuales contienen información detallada sobre los clientes, sus patrones de uso y calidad de la red móvil, así como características técnicas y geográficas de las celdas, o antenas, que componen la infraestructura de la red. Se consideró la información de noviembre de 2023, comprendiendo un total de 4,7 millones de clientes en total.

A continuación, se presenta la descripción de cada variable y su método de construcción o extracción:

- q_celdas: Cantidad de antenas a las que se conecta cada cliente. Esta variable se obtiene directamente a partir de los registros de conexión en la base de datos.

- q_celdas_unicas: Cantidad de antenas únicas a las que se conecta cada cliente. Identificada y contada a partir de los registros de conexión en la base de datos.

- score_3g: Evaluación general de la calidad de la cobertura 3G. Obtenida de la base de datos que evalúa la intensidad y calidad de la señal.

- score_4g: Evaluación general de la calidad de la cobertura 4G. Obtenida de la base de datos que evalúa la intensidad y calidad de la señal.

- p_urbano: Porcentaje del tiempo que un cliente pasa en áreas urbanas. Calculado a partir de las 10 celdas que más trafica el cliente, determinando cuántas están en zonas urbanas.

- celdas_por_hora_activo: Cantidad de antenas a las que se conecta un cliente por hora de actividad. Se calcula dividiendo la cantidad total de antenas por el tiempo que el cliente está activo.

- precio_plan: Precio del plan del cliente. Extraído directamente de la base de datos de productos y planes de servicio.

- q_comunas_distintas: Cantidad de comunas distintas en las que se encuentran las 10 antenas que más trafica cada cliente. Calculada a partir de la ubicación geográfica de las antenas frecuentadas.

- fl_ownr_sbsc_cvm: Esta variable corresponde a una variable binaria que indica si el cliente es el propietario de la suscripción. Extraída de la base de datos de clientes y suscripciones.

- d_max_cm (Km): Distancia máxima entre las antenas que más trafica el cliente respecto al centro de masa. Calculada tomando las distancias entre las 10 antenas más utilizadas y el centro de masa de estas antenas. El centro de masa corresponde al punto medio de estas antenas.

- radio_de_giro (Km): Indica cuánto se mueve cada cliente respecto a su centro de masa. Se calcula de acuerdo con la siguiente fórmula:

$$r_g^a = \sqrt{\frac{1}{n_c^a}\sum_{i=1}^{n_c^a}(\overline{r_i^a} - \overline{r_{cm}^a})}$$

  donde, r_g^a representa el radio de giro del cliente a; n_c^a es la cantidad total de posiciones distintas del usuario a; $\overline{r_{cm}^a}$ es el centro de masa de las posiciones del usuario a.

Cabe mencionar que la información respecto a la ubicación geográfica de las celdas contiene su posición geolocalizada (latitud y longitud), y se transformó a una escala de distancia para calcular las distancias entre ellas. Esta conversión se realiza utilizando la proyección Mercator, que transforma las coordenadas geográficas en una representación en metros. Esto permite calcular de manera precisa las distancias entre diferentes puntos en la superficie terrestre.

7.1.4. SATISFACCIÓN

El indicador que se tomará como satisfacción de los clientes será el Índice de Satisfacción del Cliente (CSAT). Este indicador refleja el nivel general de satisfacción de los clientes con la empresa y se calcula mediante distintos modelos de predicción que tiene la empresa a partir de resultados de encuestas CSAT.

Estas encuestas CSAT se realizan a una muestra estadísticamente significativa de clientes y miden la satisfacción del cliente con la empresa, producto o servicio mediante preguntas específicas que los clientes evalúan en una escala de 1 a 7. Los aspectos evaluados en estas encuestas incluyen la calidad del servicio, la atención al cliente, la velocidad de la red, la resolución de problemas y la relación calidad-precio.

Dado que no se pueden obtener las respuestas de todos los clientes, se utilizan los resultados de estas encuestas, realizadas a una muestra representativa de clientes, y se ingresan en un modelo de predicción para estimar las puntuaciones individuales de todos los clientes. Este modelo de predicción se basa en diversas variables descriptivas de los clientes, como el uso de servicios, el historial de interacciones con la empresa y datos demográficos, entre otros.

A partir de los resultados predichos, se calcula la probabilidad de que cada cliente sea detractor, promotor o neutro. Un detractor es un cliente insatisfecho que puede dañar la reputación de la empresa, un promotor es un cliente muy satisfecho que puede recomendar la empresa a otros, y un cliente neutro es aquel que no tiene una opinión fuerte ni positiva ni negativa.

Finalmente, se calcula la probabilidad promedio de ser promotores y detractores de los clientes, y se restan estos valores para obtener la satisfacción general de la empresa.

Actualmente, Entel cuenta con una satisfacción promedio del 70% en cuanto a telefonía móvil pospago. En el rubro de las telecomunicaciones se espera estar en un rango mínimo de un 65% a 80% en esta métrica, indicando que las empresas están cumpliendo con las expectativas de la mayoría de sus clientes, pero aún tienen margen para mejorar. Este rango sugiere que, aunque la satisfacción es aceptable, los clientes podrían estar buscando ofertas mejores o servicios más satisfactorios si se les presenta una alternativa. Es por esto por lo que la meta es mejorar los valores actuales y no bajar de este rango.

7.2. COMPRENSIÓN DE LOS DATOS

Luego de la elección de las variables en el apartado anterior, se realizó un EDA para identificar las características principales de estas variables. Principalmente, se estudiaron las distribuciones de las variables a partir de diagramas de cajas y bigotes o gráfico de frecuencia para el caso de la variable binaria. Además, se estudió la existencia de valores nulos.

7.2.1. DISTRIBUCIÓN Y CORRELACIÓN DE VARIABLES DE MOVILIDAD

A continuación, se presenta la distribución que tiene cada variable relacionada con la movilidad.

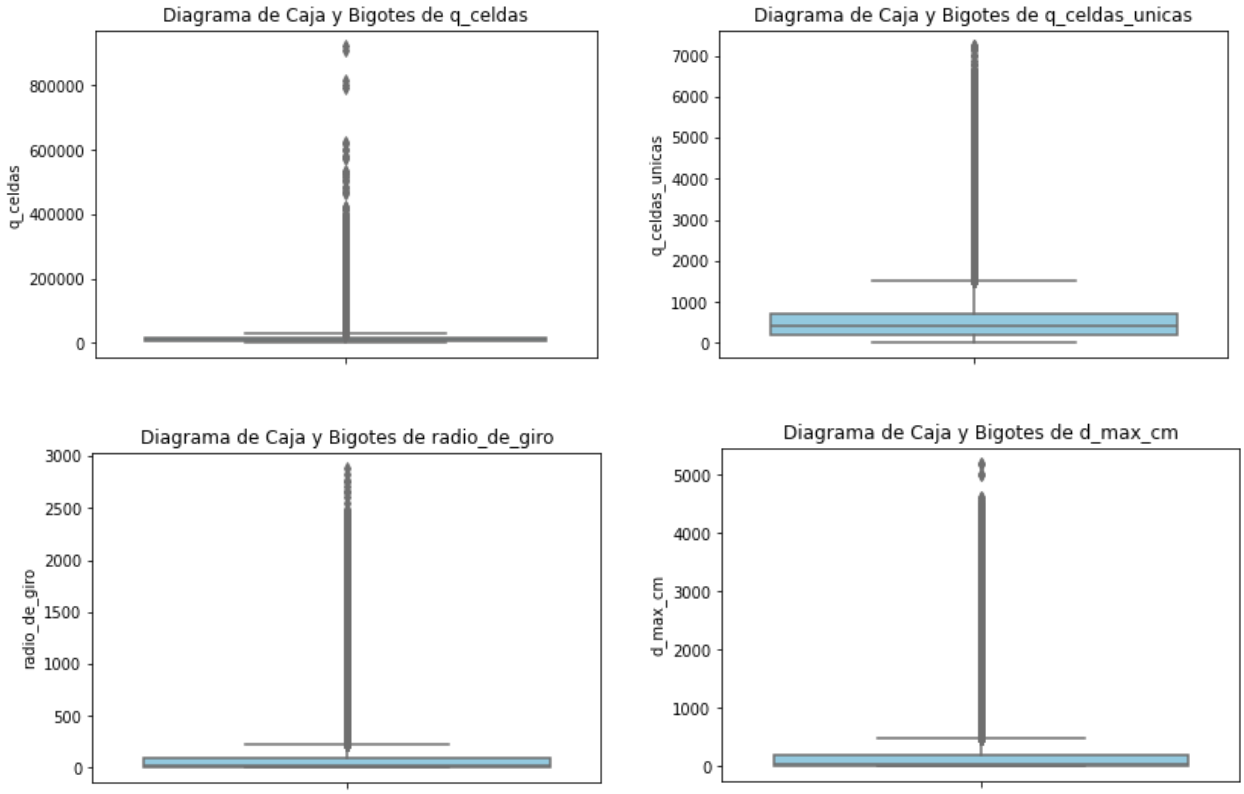

Figura 5: Distribución variables de movilidad

A partir de estos gráficos, se puede observar que la variable q_celdas presenta outliers que se desvían considerablemente de los datos centrales, los cuales se concentran en valores menores a 100,000 celdas. De manera similar, las variables radio_de_giro y d_max_cm también muestran valores extremos que se desvían considerablemente de sus datos centrales, los cuales se concentran en valores menores a 500 y 1,000

respectivamente. La variable q_celdas_unicas también presenta outliers, aunque estos no se desvían tanto de los datos centrales en comparación con las otras variables. Sin embargo, es importante identificar estos outliers para comprender su impacto en el análisis, determinar si representan errores, casos excepcionales o datos válidos, y decidir si es necesario aplicar técnicas de manejo de outliers para asegurar la precisión y la robustez de los resultados del modelo

Además del análisis de los outliers, se presenta la matriz de correlación de las variables para evaluar el grado de correlación entre ellas.

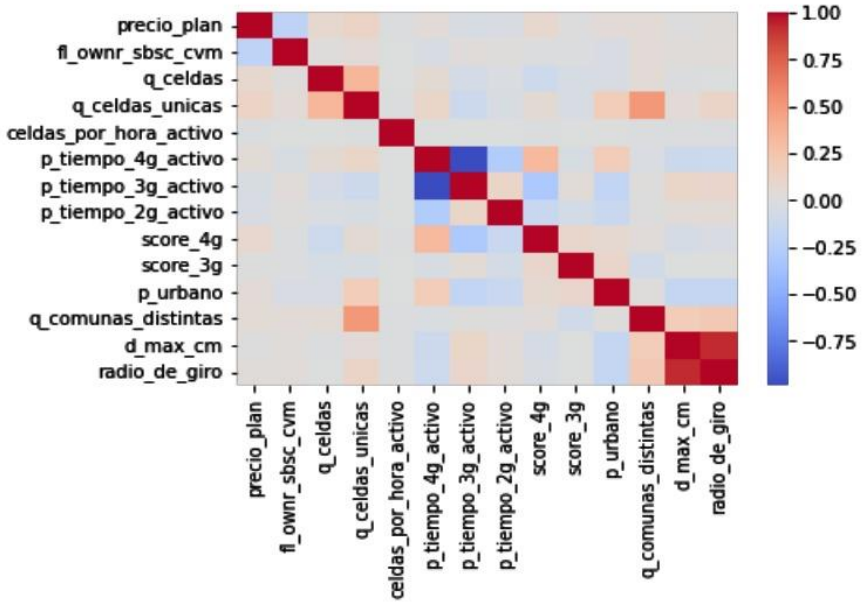

Figura 6: Matriz de correlación de variables

A partir de esto, se destacan las siguientes correlaciones:

- (Inversa) p_tiempo_4g_activo y p_tiempo_3g_activo debido a la naturaleza complementaria del uso de estas tecnologías.

- (Directa) q_comunas_distintas y q_celdas_unicas ya que, al transitar por más comunas distintas, la conexión a distintas celdas es mayor.

- (Directa) d_max_cm y radio_de_giro debido a la relación matemática entre ambas.

- (Directa) p_urbano y q_celdas_unicas, (directa) p_urbano y p_tiempo_4g_activo, (inversa) p_urbano y p_tiempo_3g_activo. Esto debido a que, las zonas urbanas tienden a tener mayor cantidad de celdas, y a tener mejor conexión.

- (Directa) score_4g y p_tiempo_4g_activo debido a que, al ser una mejor tecnología, tiende a tener una mejor evaluación de los clientes.

- (Inversa) score_4g y p_tiempo_3g_activo ya que los clientes pueden no estar relacionando una baja calidad de la tecnología con 4g, evaluando de mala manera a la red 4g, y no a 3g.

Identificar estas correlaciones es importante para el modelo, ya que se deben excluir aquellas variables que presentan alta correlación entre sí, dejando solo una de ellas. De este modo, se evita la multicolinealidad, que puede distorsionar los resultados del modelo y reducir su capacidad de generalización al introducir redundancia y posibles sesgos en la interpretación de los efectos de cada variable

7.3. PREPARACIÓN DE LOS DATOS

Luego del estudio de la distribución de las variables y el análisis de los valores atípicos, se procedió con la eliminación de aquellos que podrían influir de manera negativa en el modelo. Esta decisión se justifica debido a que tales valores pueden distorsionar los resultados del modelo al afectar los centros de los clusters, incrementar el error del modelo y crear grupos no representativos. Además, estos valores pueden sesgar las métricas de distancia y densidad utilizadas por los algoritmos de clustering, lo que puede llevar a una identificación incorrecta de patrones en los datos. Por lo tanto, eliminar estos valores extremos ayuda a mejorar la precisión y la fiabilidad del modelo, permitiendo una segmentación más precisa y efectiva de los datos.

Esta eliminación se hizo de acuerdo con los siguientes criterios:

- q_celdas > 600.000: Es improbable que un cliente se conecte a más de 600,000 antenas en un periodo de tiempo de un mes. Tales valores son inusuales y probablemente errores en los datos, lo que podría sesgar el modelo.

- q_celdas_unicas > 6.000: Conectarse a más de 6,000 antenas únicas es extremadamente raro en Chile, dada la infraestructura disponible (existen, en promedio, 2000 antenas por provincia en chile) y el comportamiento típico del usuario.

- radio_de_giro > 2000 (Km): Un radio de giro superior a 2,000 es raro para la mayoría de los clientes, ya que la mayoría de los usuarios no se desplazan tanto en su uso diario.

Esto resultó en la eliminación de un total de 2,127 datos, lo que representa solo el 0.045% del total de 4.7 millones de datos. Dado que esta proporción es mínima, el resto de los datos sigue siendo adecuado y relevante para el análisis. Por lo tanto, es poco probable que la pequeña cantidad de datos eliminados tenga un impacto significativo en los resultados o en la calidad de los modelos construidos.

Posteriormente a la eliminación de estos outliers, se normalizaron las variables numéricas para asegurar que todas tuvieran una media de 0 y una desviación estándar de 1. Esto se hizo con el propósito de garantizar la equidad entre las variables, ya que las diferentes escalas podrían influir directamente en la formación de los clusters por parte de los modelos.

7.4. MODELADO

La construcción de los modelos se realizó con una muestra del 25% del total de datos para reducir los tiempos y recursos necesarios para la construcción y análisis. Además, se tomó una muestra del 25% de los datos completamente distinta a la anterior para realizar las validaciones. Principalmente, para estudiar la capacidad de generalización del modelo.

Se usó la misma muestra para todos los modelos, o submuestras aleatorias de estos para los modelos más complejos, con el fin de atribuir cualquier diferencia en el rendimiento directamente a las características de los modelos, en lugar de a variaciones en los datos.

Los modelos se crearon sólo considerando variables de movilidad ('q_celdas', 'q_celdas_unicas', 'q_comunas_distintas' y 'radio_de_giro'). Considerando sólo una variable que tuviera una alta correlación con alguna otra variable, con el fin de evitar redundancia en el modelo, lo cual puede afectar negativamente la precisión, interpretabilidad y eficiencia del modelo. Por ejemplo, como radio_de_giro tiene una correlación muy alta con d_max_cm, entonces se consideró sólo una de estas variables para los modelos.

7.4.1. K-MEANS

Se decidió utilizar el modelo K-means en primera instancia, debido a que este modelo es eficiente y puede manejar grandes volúmenes de datos, ideal para la muestra de 1,12

millones de clientes. Además, este algoritmo es fácil de implementar y converge rápidamente, lo que significa que llega a una solución estable en un número reducido de iteraciones, facilitando el análisis continuo de los datos.

Este algoritmo requiere que se defina el número de clusters a formar. Esta elección puede ser determinada por el científico de datos según criterios específicos, como los requerimientos del negocio, o mediante técnicas que calculan el número óptimo de clusters. Un buen modelo de clustering logra encontrar un equilibrio al capturar las diferencias significativas entre grupos con un número moderado de clusters, evitando tanto la sobre división como la subdivisión excesiva de los datos. Esto asegura que los clusters sean coherentes y útiles para el análisis y la toma de decisiones.

Para estimar el número óptimo de clusters, se utilizó el método del codo y el método Gap Statistics. Obteniéndose los siguientes resultados:

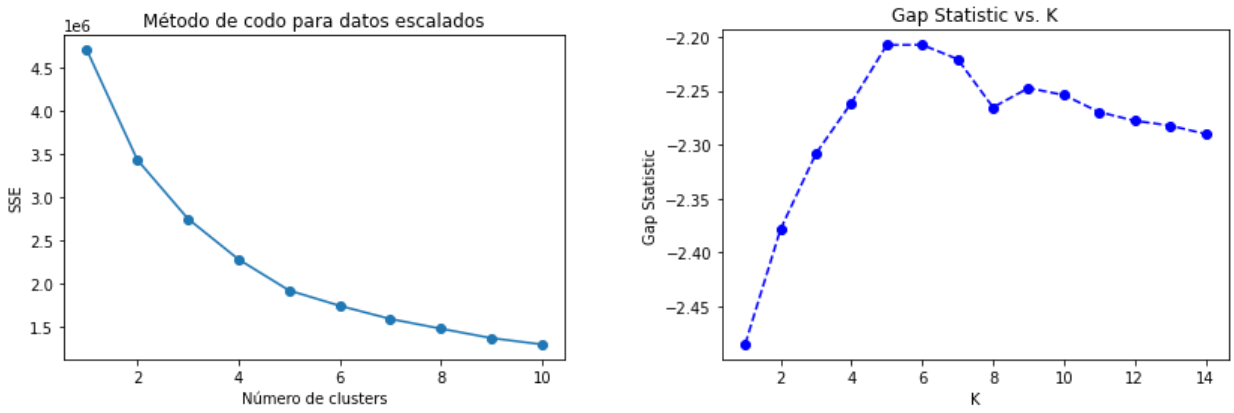

Figura 7: Número óptimo de clústers

Como se puede observar, en el método del codo, el punto donde la suma de los errores cuadráticos (SSE) comienza a disminuir con menor fuerza se observa en aproximadamente K = 5, indicando una disminución significativa en la variación dentro de los clusters hasta ese punto. Asimismo, en el método de Gap Statistics, el punto con la menor diferencia también se encuentra en aproximadamente K = 5, lo cual refuerza esta elección. Por lo tanto, ambos métodos coinciden en que la mejor partición de los datos se logra con 5 clusters.

A pesar de que el número óptimo de clusters se determinó como 5 mediante el método del codo y el método de Gap Statistics, se consideraron también modelos con 4 y 3 clusters. Esto se hizo con el objetivo de explorar cómo variaciones en el número de clusters afectan la estructura de los datos y el rendimiento del modelo. Evaluar múltiples configuraciones permite una comprensión más amplia de la agrupación de los datos,

proporcionando información adicional sobre la estabilidad y la robustez del modelo. Además, puede revelar subestructuras relevantes que no son evidentes cuando se utiliza únicamente el número óptimo de clusters o evaluando sólo una métrica para determinar el número óptimo de clusters.

7.4.2. DBSCAN

Como se mencionó en el marco teórico, el algoritmo DBSCAN no requiere la especificación previa del número óptimo de clusters. En su lugar, los hiperparámetros clave que deben ajustarse son el épsilon y el número mínimo de puntos. El épsilon determina la distancia máxima entre dos puntos para que se consideren vecinos directos, mientras que min samples establece el número mínimo de puntos que deben estar dentro de esta distancia para que un punto sea considerado núcleo.

El número mínimo de puntos se determinó como el doble del número de características en los datos, ya que esta práctica inicial busca asegurar que cada punto central (core point) tenga suficientes vecinos dentro de su épsilon para formar clusters significativos y evitar clasificar ruido como clusters.

Por otro lado, el épsilon se determinó mediante el método de la rodilla, obteniéndose el siguiente gráfico:

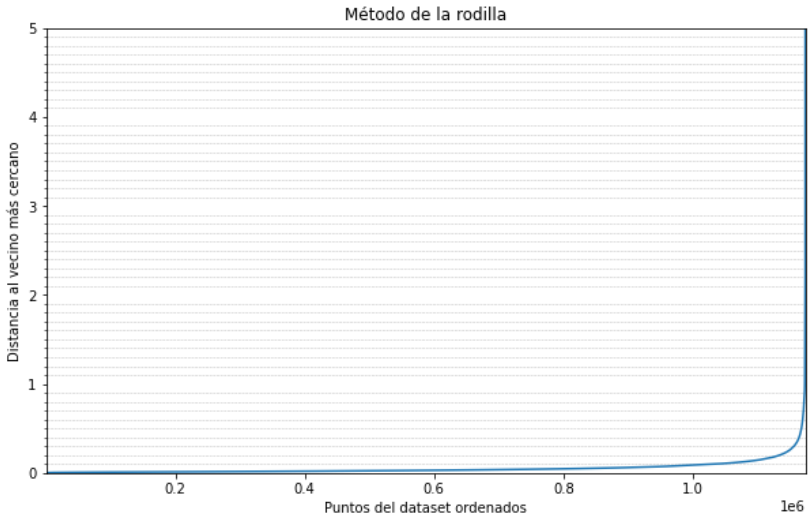

Figura 8: Método de la rodilla modelo DBSCAN

Este modelo generó un total de 42 clusters. Este número elevado de clusters puede dificultar la capacidad de extraer conclusiones claras y significativas sobre la estructura subyacente de los datos. Es por esto por lo que se decide descartar este modelo del análisis, continuando sólo con los modelos de K-means.

7.5. EVALUACIÓN

Luego de tener los modelos listos, se calcularon las métricas de evaluación de cada uno, obteniéndose los siguientes resultados:

Métrica	KMeans		
	5 clusters	4 clusters	3 clusters
Silhouette Score	0,30	0,32	0,35
Davies-Bouldin Index	1,06	1,10	1,34
Calinski-Harabasz Index	423968,22	414697,54	356488,73

Tabla 1: Métricas de evaluación de clusters

Es importante recordar que, se considera una mejor evaluación cuando el Silhouette score y el Índice Calinski-Harabasz son más altos, y con valores de Davies-Bouldin más bajos.

A partir de estos resultados, se decide elegir el modelo K-means con 4 clusters. Esta elección se justifica dado que este modelo representa un equilibrio en estas métricas. Además, elegir 4 clúster facilita el análisis al reducir la complejidad y mejorar la interpretación, ya que habría menos grupos para analizar y visualizar, lo que podría ayudar a identificar patrones distintivos y características de cada clúster de manera más clara.

7.6. DESPLIEGUE

7.6.1. ANÁLISIS CARACTERÍSTICAS POR GRUPO

A partir de los grupos creados en el modelo K-means con 4 clusters, se obtuvieron las siguientes estadísticas y distribuciones de las variables dentro de cada grupo. Cabe destacar que estas variables están en su escala normal (se revirtió la normalización).

Grupo	% Base	Media			
		q_celdas	q_celdas_unicas	q_comunas_distintas	radio_de_giro
0	24%	12322,8	1046,9	4,7	84
1	56%	7116,4	296,4	2,3	45,1
2	13%	25870,3	554,9	2,4	47,8
3	7%	10525,8	625,9	3,6	658,1

Tabla 2: Estadísticas de variables por cluster

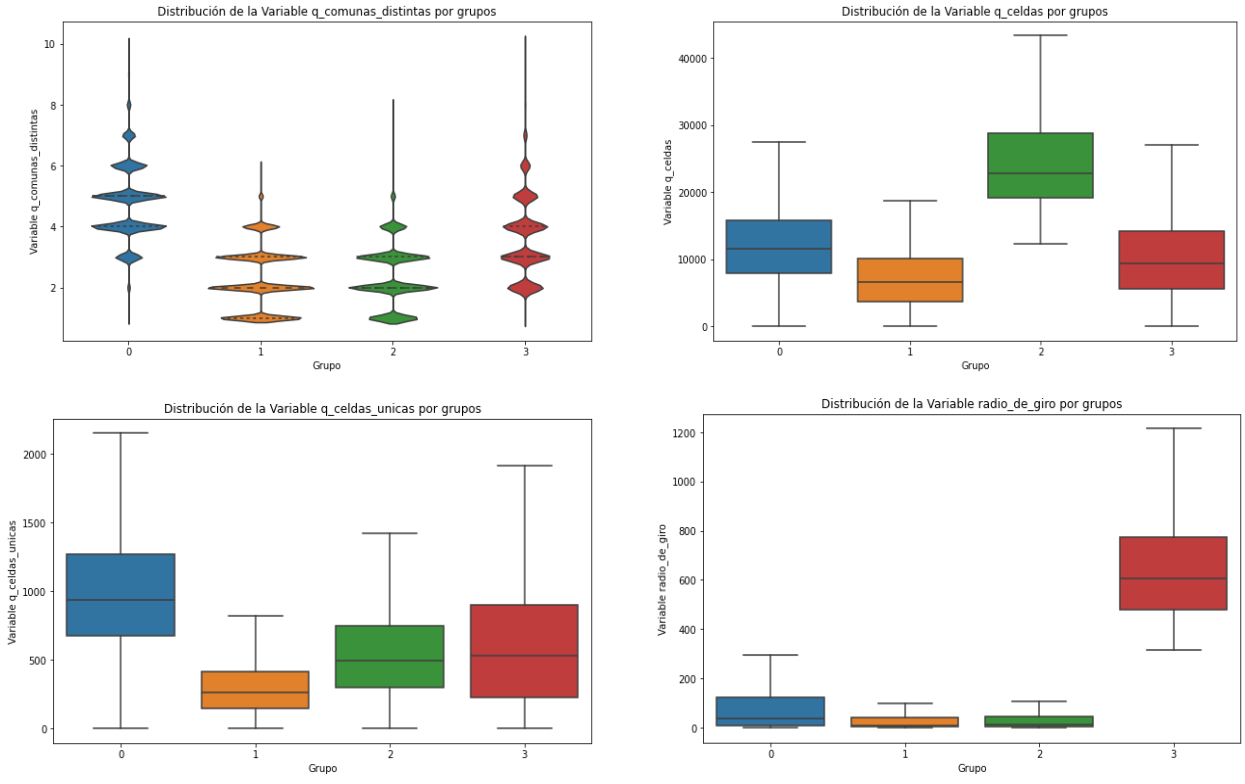

Figura 9: Gráficos de distribución de variables por grupo

Para proporcionar una mejor visualización de cómo varía esta distribución en cada grupo, se muestra el siguiente gráfico de coordenadas paralelas.

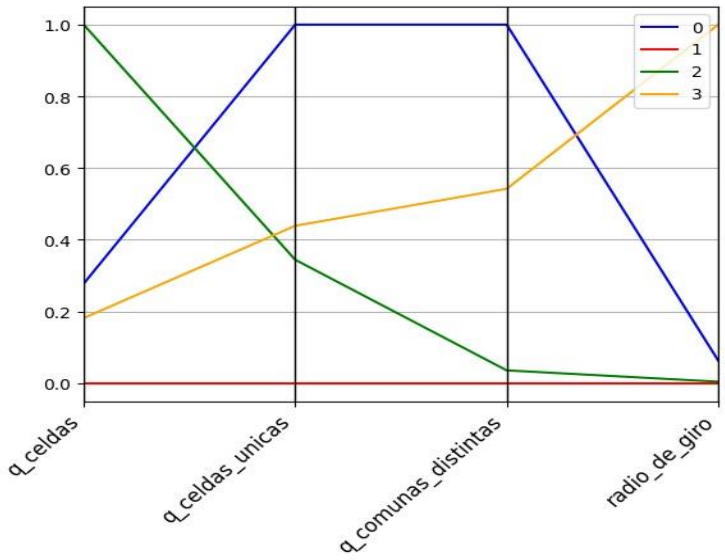

Figura 10: Gráfico de coordenadas paralelas

Este gráfico representa cada variable en un eje vertical separado y cada línea a través de los ejes muestra una observación individual, permitiendo identificar patrones y relaciones entre las variables en diferentes clusters.

Al analizar los gráficos se pueden sacar las siguientes conclusiones respecto a los grupos:

Grupo 0: gran número de celdas, celdas únicas y de comunas distintas. El radio de giro es relativamente alto, indicando que hay variabilidad en los desplazamientos. Este grupo se denominará “Viajeros urbanos”.

Grupo 1: Este grupo tiene las medidas de las variables más bajas de todas, indicando que tienen la movilidad más baja de todos. Este grupo se denominará “Homeoffice”.

Grupo 2: Tiene un número muy alto de celdas visitadas, aunque la cantidad de celdas únicas y la diversidad de comunas distintas es baja. El radio de giro es bajo a moderado, indicando que, aunque se mueven moderadamente, estos movimientos están concentrados en áreas específicas. Este grupo se denominará “Commuters” .

Grupo 3: Este grupo tiene una movilidad extrema con un número moderado de celdas, celdas únicas, pero con un radio de giro extremadamente alto. Esto sugiere que los clientes de este grupo realizan desplazamientos muy grandes y dispersos, cubriendo una amplia área geográfica. Este grupo se denominará “Viajeros extremos”.

7.6.2. ANÁLISIS DE SATISFACCIÓN POR GRUPO

Finalmente, luego de tener segmentados los clientes de acuerdo con su movilidad, se realiza el análisis de la satisfacción en cada grupo. Para esto, se cruzó la tabla de satisfacción con la tabla de datos de movilidad con la etiqueta que se le asignó al realizar la segmentación. Luego, se calculó la probabilidad de detractor y promotor de cada grupo. Cabe destacar que, además de ser promotor y detractor, también existe una probabilidad de ser neutro, quienes son aquellos clientes que están satisfechos, pero no lo suficiente como para recomendar activamente el servicio. Los clientes neutros se excluyen de esta probabilidad debido a que no influyen de manera significativa en la reputación de la marca ni en la adquisición de nuevos clientes. Se enfocan principalmente en los promotores y detractores para obtener una visión más clara del impacto en la satisfacción y lealtad de los clientes.

Los resultados obtenidos se muestran a continuación:

Grupo	%Base	Media Promotor	Media Detractor	SAT
0	24%	70,9%	10,1%	60,8%
1	56%	70,1%	10,3%	59,8%
2	13%	65,3%	12,3%	53,0%
3	7%	65,7%	11,7%	53,9%

Tabla 3: Satisfacción por cluster

A partir de esta tabla se obtienen las siguientes deducciones sobre los grupos:

Grupos 0: Con una base del 24% y una satisfacción general (SAT) de 60.8%, este grupo muestra una alta satisfacción. La buena satisfacción en este grupo podría estar relacionada con una cobertura robusta, buena estabilidad en la conexión y velocidad adecuada en áreas urbanas, donde la movilidad es alta. La percepción positiva del servicio en estas áreas sugiere que los clientes experimentan una conexión fiable y consistente, con menos problemas de intermitencia.

Grupo 1: Representa el 56% de la base y tiene una satisfacción general (SAT) de 59.8%. A pesar de tener una movilidad baja, la satisfacción es alta, lo que indica que los clientes en este grupo están bastante satisfechos con el servicio. Esto podría estar relacionado con una cobertura sólida y una buena calidad de conexión en áreas residenciales. La alta

satisfacción sugiere que, aunque la movilidad es baja, el servicio cumple con las expectativas de los clientes en términos de estabilidad y velocidad, sin problemas significativos de intermitencia o cobertura insuficiente.

Grupos 2: Con una base del 13% y una satisfacción general (SAT) de 53.0%, este grupo tiene una satisfacción relativamente baja. A pesar de tener una movilidad moderada, la baja satisfacción podría estar ligada a problemas de cobertura y estabilidad en áreas específicas que frecuentan. La movilidad moderada sugiere que, aunque no están restringidos a una sola área, podrían enfrentar problemas de conexión intermitente o velocidad variable en las áreas que visitan con frecuencia.

Grupo 3: Tiene una base del 7% y una satisfacción general (SAT) de 53.9%. Aunque la satisfacción es ligeramente superior a la del Grupo 2, la baja proporción de clientes indica que la experiencia del servicio podría estar afectada por problemas de cobertura y estabilidad a lo largo de largas rutas. Los clientes que realizan desplazamientos extensos pueden enfrentar problemas de intermitencia, variabilidad en la velocidad de conexión y cobertura inadecuada en zonas remotas o entre áreas geográficas dispersas.

Este procedimiento establece una metodología para la clasificación y análisis continuo de datos futuros basados en movilidad, utilizando el modelo de clustering K-means con 4 clusters. Los clusters permitirán categorizar nuevos datos según patrones de comportamiento similares. Al combinar estos segmentos con métricas de satisfacción del cliente, como la identificación de detractores o promotores, se facilita el análisis para identificar áreas de mejora y oportunidades estratégicas con mayor precisión.

8. DISCUSIONES

8.1. ALCANCES

El proyecto presenta varios alcances importantes que determinan el marco y los límites de su implementación y aplicabilidad. Debido a restricciones de tiempo, el proyecto se centra únicamente en la evaluación y selección del modelo de clusterización que ofrece el mejor rendimiento. Las recomendaciones para la implementación del modelo en la práctica serán presentadas como una propuesta separada a la empresa, y la ejecución de estas recomendaciones quedará fuera del alcance de este proyecto. Esto implica que el impacto real del modelo en la práctica dependerá de la capacidad de la empresa para traducir los resultados analíticos en acciones concretas y efectivas. Esto significa que, aunque se logren identificar segmentos clave de clientes y se genere información al respecto, la verdadera utilidad del proyecto dependerá completamente de la disposición y capacidad de Entel para adoptar y aplicar estas recomendaciones. La falta de un plan de implementación detallado podría llevar a que los resultados del modelo no se traduzcan en mejoras reales en la satisfacción del cliente o en la optimización de la red,

reduciendo así el potencial impacto del proyecto. Además, sin un seguimiento riguroso de la implementación, es difícil medir y validar el éxito del modelo en un entorno real, lo que puede limitar la capacidad de la empresa para justificar futuras inversiones basadas en estos análisis. Por lo tanto, aunque el proyecto ofrece un marco sólido para la toma de decisiones estratégicas, su éxito dependerá en gran medida de cómo Entel decida proceder con la información obtenida.

Además, el proyecto se basa exclusivamente en datos de clientes pospago de servicios móviles de Entel en Chile, que abarcan aproximadamente 4.7 millones de clientes. Este enfoque restringido significa que los resultados obtenidos no se pueden extrapolar directamente a otros mercados o segmentos de clientes, como los usuarios de servicios móviles prepago, debido a que los clientes prepago o clientes de servicios fijos presentan comportamientos y patrones de uso diferentes, lo que limita la generalización de los hallazgos. Por ejemplo, los clientes prepago pueden tener un comportamiento más selectivo en el uso de servicios, lo cual puede influir en sus patrones de movilidad de forma distinta en comparación con los clientes pospago.

La calidad de los datos es fundamental para el análisis, y el proyecto se basa en datos completos y de alta calidad para garantizar resultados precisos y significativos. Sin embargo, una limitante importante del proyecto es la restricción temporal de los datos, que abarcan desde julio de 2023 hasta marzo de 2024, y aún más si el entrenamiento de los modelos se realizó sólo con información de noviembre de 2023. Esta ventana temporal es un factor crítico, ya que la información utilizada es representativa de este periodo específico, pero puede no reflejar cambios en el comportamiento de los clientes posteriores a este intervalo. Por esto, se plantea realizar nuevos modelos con una base de datos más robusta, intentando representar lo que más se pueda los cambios de comportamientos de los clientes en una ventana más grande de tiempo.

8.2. ELECCIÓN DE VARIABLES

En el modelo de segmentación, se seleccionaron las variables de movilidad como la cantidad de antenas y antenas distintas, la cantidad de comunas distintas, la distancia máxima al centro de masa, y el radio de giro. Estas variables fueron elegidas porque ofrecen una visión de la movilidad de los clientes. La cantidad de antenas y antenas distintas refleja la diversidad de ubicaciones a las que se conecta un cliente, sugiriendo un mayor movimiento geográfico. La cantidad de comunas distintas agrega contexto geográfico, permitiendo identificar si el cliente se mueve en un área amplia o concentrada. La distancia máxima al centro de masa y el radio de giro proporcionan una medida de la dispersión y la extensión del movimiento, lo cual ayuda a entender cómo se distribuye la movilidad en el espacio.

Sin embargo, estas variables tienen limitaciones. Como, por ejemplo, la cantidad de antenas no diferencia entre conexiones breves y prolongadas, lo que podría dar una

imagen poco precisa de la movilidad real. Del mismo modo, la distancia máxima entre antenas podría no capturar con precisión el movimiento en zonas urbanas densas o rurales complejas, donde la distancia no siempre corresponde al desplazamiento real del cliente. En áreas urbanas densas, las antenas pueden estar muy cerca unas de otras, lo que significa que incluso un cliente que se mueve bastante podría estar siempre conectado a antenas próximas, reduciendo la distancia máxima calculada a pesar del considerable desplazamiento. Por otro lado, en zonas rurales, donde las antenas pueden estar más dispersas, un cliente podría recorrer largas distancias, pero seguir conectado a la misma antena o a unas pocas antenas distantes entre sí, lo que podría dar una impresión errónea de su movilidad real.

Variables adicionales que capturen la calidad de la cobertura o los horarios de conexión podrían haber mejorado la segmentación, ya que ofrecerían insights sobre la experiencia del cliente en ubicaciones clave. Por ejemplo, un cliente que experimenta mala cobertura en su hogar podría estar menos satisfecho, lo que afectaría su lealtad. Además, eventos como viajes ocasionales o emergencias pueden introducir ruido en los datos, distorsionando la percepción del comportamiento típico del cliente si no se filtran adecuadamente, lo que podría llevar a segmentaciones incorrectas. Para mitigar estas limitaciones, se podrían ajustar las variables para considerar el propósito del uso o la hora del día, lo que permitiría una captura más precisa de los patrones de movilidad. También sería beneficioso incluir datos sobre la percepción de calidad del servicio, lo que podría ayudar a identificar problemas no evidentes en la conectividad y, en última instancia, mejorar la relevancia y efectividad de la segmentación.

A partir de la correlación moderada (0.5) que presentan las variables "cantidad de antenas únicas" y la "cantidad de comunas distintas" es necesario analizar cada variable y cómo ayudan en el análisis de los resultados del modelo. Aunque existe una correlación moderada entre ellas, ambas capturan aspectos diferentes y complementarios de la movilidad.

La "cantidad de antenas únicas" mide la diversidad de conexiones que un cliente tiene a lo largo de su movimiento. Esta variable es particularmente relevante en áreas urbanas densas, donde un cliente puede moverse dentro de una zona geográficamente limitada, pero conectarse a muchas antenas diferentes debido a la alta densidad de la infraestructura de red. En estos casos, un cliente puede estar recorriendo distancias relativamente cortas, pero conectándose a múltiples antenas, lo que refleja un patrón de movilidad intenso y localizado.

Por otro lado, la "cantidad de comunas distintas" refleja la extensión geográfica del movimiento de un cliente, capturando la diversidad en las áreas que frecuenta. Esta variable es especialmente útil en zonas rurales o en áreas con menor densidad de antenas, donde recorrer la misma distancia podría resultar en conexiones a un número limitado de antenas, pero a través de diferentes comunas. Aquí, la "cantidad de comunas

distintas" ofrece una perspectiva más amplia del movimiento, indicando que el cliente está cubriendo una mayor variedad geográfica, a pesar de conectarse a menos antenas.

Considerando la variabilidad en la densidad de antenas, es crucial mantener ambas variables en el modelo. En áreas urbanas densas, la "cantidad de antenas únicas" es fundamental para capturar la granularidad del movimiento, mientras que la "cantidad de comunas distintas" permite identificar desplazamientos significativos en áreas con menor infraestructura de red. Este enfoque evita malinterpretaciones sobre la movilidad de los clientes en diferentes entornos.

Es importante destacar que, aunque estas variables están relacionadas de manera moderada, estas no son redundantes. Cada una aporta información sobre el comportamiento del cliente, tanto en términos de movilidad localizada como de desplazamiento geográfico amplio. Sin embargo, se recomienda monitorear cómo cada variable impacta en el modelo, asegurando que ambas sigan proporcionando valor y ayudando como diferenciador en el análisis de movilidad, teniendo cierto conocimiento sobre los sectores en donde se aplique el modelo.

8.3. MODELADO

El desarrollo y evaluación de modelos de clustering utilizando K-means y DBSCAN se basan en una metodología sólida y bien fundamentada, buscando un balance entre eficiencia y precisión en el análisis. La decisión de usar un 25% de los datos para la construcción de los modelos, junto con una muestra distinta para la validación, está justificada por la necesidad de reducir tiempos y recursos sin comprometer la representatividad, lo que facilita la evaluación de la capacidad de generalización de los modelos. No obstante, esta estrategia podría limitar la detección de patrones presentes en el 75% restante de los datos, por lo que, como alternativa, podría considerarse el uso de técnicas de validación cruzada en lugar de utilizar una sola muestra, lo que permitiría un análisis más exhaustivo. En cuanto a la selección de variables de movilidad, se eligieron aquellas no redundantes para evitar la multicolinealidad, que puede distorsionar la precisión y la interpretabilidad del modelo. Sin embargo, la eliminación de variables con alta correlación también puede omitir información relevante, una limitación que es mitigada mediante técnicas de reducción de dimensionalidad como PCA (Análisis de Componentes Principales) que conservan la varianza de los datos mientras se reducen las dimensiones como se muestra en el anexo B.

El K-means fue elegido por su eficiencia en la gestión de grandes volúmenes de datos, siendo ideal para la muestra de 1,12 millones de clientes. Además, su rápida convergencia permite un análisis continuo, lo que es esencial en contextos de datos masivos. No obstante, su principal limitación radica en la necesidad de especificar el número de clusters a priori, lo que puede llevar a una sobre o subsegmentación si no se elige adecuadamente. Aunque los métodos del codo y Gap Statistics fueron utilizados

para determinar el número óptimo de clusters, y ambos coincidieron en sugerir 5 clusters, estos métodos tienen limitaciones inherentes, como la distribución no esférica de los clusters. Como alternativas, se plantea explorar algoritmos como Gaussian Mixture Models (GMM), que permiten una mayor flexibilidad en la forma de los clusters, o el uso de técnicas avanzadas de clustering jerárquico que no requieren una predefinición del número de clusters.

Por otro lado, el modelo DBSCAN se seleccionó por su capacidad de identificar clusters de forma adaptativa, sin requerir un número de clusters predefinido, lo que es particularmente útil en datos con densidades variables. Sin embargo, la configuración de los hiperparámetros, como el épsilon, resultó en la creación grandes cantidades de clusters, lo que complicó la interpretación del modelo y su aplicabilidad. Esta sobre segmentación revela la limitación de DBSCAN cuando los parámetros no se ajustan con precisión para una cantidad de clusters deseada, especialmente en datasets con patrones de densidad complejos. Como alternativas a estos modelos, están algunos como HDBSCAN, una extensión de DBSCAN que selecciona el épsilon de manera adaptativa, o incluso técnicas basadas en clustering espectral, que son capaces de manejar la complejidad de estructuras de datos más intrincadas.

8.4. EVALUACIÓN Y RESULTADOS

La elección del modelo K-means con 4 clusters se basa en un equilibrio evaluativo entre las métricas de Silhouette Score, Davies-Bouldin Index y Calinski-Harabasz Index. La decisión se justificó por la necesidad de encontrar una segmentación que ofreciera una combinación adecuada de cohesión y separación sin complicar excesivamente el análisis. Aunque el modelo con 3 clusters mostró un Silhouette Score ligeramente superior, su Davies-Bouldin Index más alto indicaba una separación menos clara entre clusters, lo que podría llevar a interpretaciones menos robustas. Por otro lado, el modelo con 5 clusters, a pesar de tener un Calinski-Harabasz Index más alto y un Davies-Bouldin Index más bajo, mostró un Silhouette Score más bajo, lo que sugiere que la adición de clusters introdujo complejidad adicional sin una mejora significativa en la calidad del clustering.

El modelo seleccionado, con 4 clusters, muestra un Silhouette Score de 0,32, un Davies-Bouldin Index de 1,10, y un Calinski-Harabasz Index de 414,697.54. Estos valores indican que el modelo ofrece una separación razonable entre clusters con una cohesión interna adecuada, sin caer en la sobre segmentación. La elección de 4 clusters permite un análisis manejable, facilitando la identificación de patrones y características sin la complejidad excesiva de modelos con más clusters. A pesar de que los valores obtenidos son relativamente buenos, considerando la complejidad de los datos, la evaluación indica que hay espacio para mejorar. La calidad del clustering podría beneficiarse de la exploración de otros métodos o ajustes de parámetros, como el uso de técnicas de clustering jerárquico o modelos de mezcla gaussiana, los cuales proporcionan una

segmentación más matizada o validar la robustez del modelo mediante técnicas adicionales de evaluación.

A partir de los resultados del modelo K-means con 4 clusters, se puede apreciar que la movilidad por sí sola no es un predictor fiable de la satisfacción del cliente. La alta satisfacción en los Grupos 0 y 1, que presentan alta y baja movilidad respectivamente, indica que factores como la calidad del servicio y la estabilidad de la red tienen un impacto importante en la satisfacción. Los Grupos 2 y 3, con movilidad moderada y extremadamente alta, muestran niveles de satisfacción más bajos, lo que refuerza la idea de que la movilidad no está directamente relacionada con la satisfacción.

Este análisis revela que las etapas previas del modelo de clustering podrían haberse beneficiado de un enfoque más integral. Primero, se debió considerar la inclusión de variables adicionales que reflejen la calidad del servicio, como cobertura, estabilidad de la conexión y velocidad de datos. Estas variables son críticas para comprender cómo influyen en la satisfacción del cliente y deberían haberse incorporado desde el inicio para obtener una segmentación más precisa y relevante.

Además, el uso exclusivo de K-means, el cual asume que los clusters tienen forma esférica y tamaños similares, puede no haber capturado todas las complejidades de los datos. La integración de métodos de clustering alternativos que no requieran suposiciones sobre la forma de los clusters, como DBSCAN, podría haber revelado patrones más detallados. Aunque el modelo DBSCAN inicialmente generó una cantidad excesiva de clusters, ajustar sus hiperparámetros y aplicar técnicas de validación más exhaustivas podría haber proporcionado una segmentación más adecuada.

8.5. APLICACIONES E IMPACTO

El proyecto ofrece aplicaciones concretas que pueden transformar significativamente la estrategia de atención al cliente y la gestión de recursos de Entel. Al segmentar a los clientes en grupos específicos basados en patrones de movilidad, la empresa puede optimizar la red dirigiendo inversiones en infraestructura hacia áreas y segmentos críticos, como mejorar la cobertura en zonas con alta movilidad para los "Viajeros extremos" o en estabilidad de conexión para los "Homeoffice". Además, permite personalizar servicios, desarrollando ofertas que respondan a las necesidades de cada segmento, como paquetes de datos adicionales para los "Commuters". Las estrategias de marketing pueden enfocarse en reforzar la lealtad de los segmentos con alta satisfacción y abordar a aquellos con menor satisfacción mediante promociones o mejoras en el soporte. Finalmente, ajustar el soporte al cliente basado en la movilidad y satisfacción asegura que los problemas críticos se atiendan con urgencia en los segmentos necesarios. Aunque la hipótesis inicial sobre la relación directa entre alta movilidad y baja satisfacción no se validó, el proyecto demuestra que una segmentación precisa permite a Entel desarrollar estrategias adaptadas a las características específicas

de cada grupo de clientes, lo cual es crucial para reducir la tasa de churn. Esto proporciona una base para analizar más a fondo la relación entre satisfacción y churn, y ajustar las estrategias de retención en consecuencia. En definitiva, estas aplicaciones mejoran la satisfacción del cliente al abordar de manera más efectiva sus necesidades específicas, optimizan los recursos y costos operativos, y fortalecen la lealtad y retención, lo que mejora la posición competitiva de Entel en el mercado.

9. CONCLUSIÓN

El proyecto ha proporcionado resultados significativos en términos de comprensión y análisis de la satisfacción por cada segmento de clientes. Los objetivos planteados se centran en identificar segmentos de clientes con necesidades y comportamientos específicos, lo cual permite desarrollar estrategias personalizadas para mejorar la red móvil de Entel. Aunque se lograron avances importantes en la segmentación y se proporcionaron insights valiosos, la hipótesis principal del proyecto no fue validada, ya que no se encontró una relación directa entre la movilidad y la satisfacción del cliente.

Sin embargo, los resultados obtenidos permiten a Entel diseñar estrategias efectivas para cada segmento de clientes de acuerdo con las características de cada grupo, con el propósito de mejorar la satisfacción del usuario y optimizando la asignación de recursos para mejoras en la red. La identificación de segmentos específicos ofrece una base sólida para personalizar la oferta de servicios, lo que puede conducir a una mayor lealtad y retención de clientes.

Los objetivos del proyecto se alcanzaron parcialmente. Se logró identificar segmentos clave y proporcionar recomendaciones para mejorar la satisfacción del cliente. Sin embargo, falta de otras variables de comportamiento de los clientes, como la cantidad de tiempo que se conecta por antena o análisis geoespacial, impidió identificar problemas específicos de conectividad que podrían afectar la satisfacción del cliente.

Se propuso ampliar el estudio para incluir a usuarios prepago y la integración de más variables en el análisis, lo que proporcionaría una visión más completa y precisa del mercado total de telefonía móvil. También implementar otras variables para abordar problemas específicos de conectividad. Además, se propuso implementar otros algoritmos de aprendizaje no supervisado con el fin de encontrar modelos con mejor rendimiento, intentando capturar la mayor naturalidad en los datos. Estas propuestas permitirán a Entel desarrollar estrategias más inclusivas y efectivas para mejorar la satisfacción del cliente en todas las áreas.

Analizar la segmentación de clientes de acuerdo con su movilidad y satisfacción, en conjunto con la tasa de deserción de clientes es crucial para mejorar la rentabilidad de Entel, permitiendo desarrollar estrategias específicas que reduzcan la tasa de deserción. Este proyecto no solo facilita la identificación de segmentos críticos, sino que también abre la puerta a un análisis más profundo de la relación entre satisfacción, churn y movilidad, áreas donde actualmente no se observa una conexión clara, lo que podría tener un impacto significativo en la optimización de la retención de clientes y en el fortalecimiento de la estrategia comercial de la empresa.

10. BIBLIOGRAFÍA

Amazon Web Services. (n.d.). ¿En qué consiste el ajuste de hiperparámetros? AWS. 2024.
https://aws.amazon.com/es/what-is/hyperparameter-tuning/

Amazon Web Services. (n.d.). ¿Qué es SQL (lenguaje de consulta estructurada)? Amazon AWS. 2024.
https://aws.amazon.com/es/what-is/sql/

Andrés, D. (2022, Octubre 21). Clustering K-Means. Machine Learning Pills. 2024.
https://mlpills.dev/machine-learning-es/clustering-k-means/

Empresa Nacional de Telecomunicaciones. (2024). Memoria Integrada 2023. 2024.
https://entel.modyocdn.com/uploads/5e637cce-97a7-49a1-9ca2-dce0b54dadfb/original/Memoria_Entel_2023.pdf

Entel. (n.d.). Redes móviles: qué son, cómo funcionan y qué tipos existen. Comunidad Empresas de Entel. 2024.
https://ce.entel.cl/articulos/redes-moviles/

González, L. (2020, June 9). Algoritmo KMeans - Teoría. Aprende IA. 2024.
https://aprendeia.com/algoritmo-kmeans-clustering-machine-learning/

González, L. (2020, Diciembre 8). DBSCAN Teoría. Aprende IA. 2024.
https://aprendeia.com/dbscan-teoria/

González, L. (2020, Diciembre 8). DBSCAN Teoría. Aprende IA. 2024.
https://aprendeia.com/dbscan-teoria/

International Business Machines Corporation. (n.d.). ¿Qué es el análisis exploratorio de datos? IBM. 2024.
https://www.ibm.com/es-es/topics/exploratory-data-analysis

Juanweb. (2023, Diciembre 4). Procesamiento de Grandes Volúmenes de Datos con Python. CodigosPython. 2024.
https://codigospython.com/procesamiento-de-grandes-volumenes-de-datos-con-python/

Moine, J. M., Silva Haedo, A., & Silva Gordillo. (n.d.). Estudio comparativo de metodologías para minería de datos. 2024. https://sedici.unlp.edu.ar/bitstream/handle/10915/20034/Documento_completo.pdf?sequence=1&isAllowed=y

Moya, R. (2016, Septiembre 12). Selección del número óptimo de Clusters. Jarroba. 2024. https://jarroba.com/seleccion-del-numero-optimo-clusters/

Ramos Ponce, O. (2020). Tema 9: Agrupamiento (Clustering). 2024. http://oramosp.epizy.com/teaching/202/topicos/clases/9_Clustering.pdf?i=1

Rodríguez, D. (2019, December 16). ¿Cuál es la diferencia entre parámetro e hiperparámetro? Analytics Lane. 2024. https://www.analyticslane.com/2019/12/16/cual-es-la-diferencia-entre-parametro-e-hiperparametro/

Rodríguez, D. (2023, June 2). Optimizar el número de clústeres con gap statistics. Analytics Lane. 2024. https://www.analyticslane.com/2023/06/02/optimizar-el-numero-de-clusteres-con-gap-statistics/

Rodríguez, D. (2023, June 16). Identificar el número de clústeres con Calinski-Harabasz en k-means e implementación en Python. Analytics Lane. 2024. https://www.analyticslane.com/2023/06/16/identificar-el-numero-de-clusteres-con-calinski-harabasz-en-k-means-e-implementacion-en-python/

Rodríguez, D. (2023, June 23). Número óptimo de clústeres con Silhouette e implementación en Python. Analytics Lane. 2024. https://www.analyticslane.com/2023/06/23/numero-optimo-de-clusteres-con-silhouette-e-implementacion-en-python/

Rodríguez, D. (2023, June 30). El índice de Davies-Bouldinen para estimar los clústeres en k-means e implementación en Python. Analytics Lane. 2024. https://www.analyticslane.com/2023/06/30/el-indice-de-davies-bouldinen-para-estimar-los-clusteres-en-k-means-e-implementacion-en-python/

Rueda, J. F. (2024, Abril 6). Clusterización espacial: K-means vs DBSCAN. Descripción y ejemplo práctico con QGIS – COLUMNA DE INVESTIGACIÓN DATLAS. Blog Datlas. 2024.

https://blogdatlas.wordpress.com/2024/04/06/clusterizacion-espacial-k-means-vs-dbscan-descripcion-y-ejemplo-practico-con-qgis-columna-de-investigacion-datlas/

Subsecretaría de Telecomunicaciones. (2024). Estadísticas - Telefonía Móvil. 2024. https://www.subtel.gob.cl/estudios-y-estadisticas/telefonia/

Rodrigo, J. A. (2017). Análisis de Componentes Principales (Principal Component Analysis, PCA) y t-SNE. https://cienciadedatos.net/documentos/35_principal_component_analysis

ANEXOS

ANEXO A: MÉTODOS PARA ELECCIÓN DE HIPERPARÁMETROS

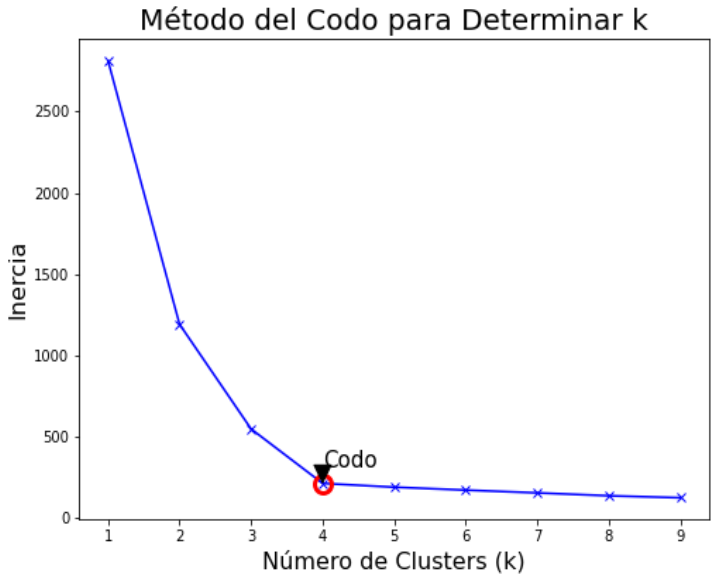

Figura 11: Ejemplo método del codo

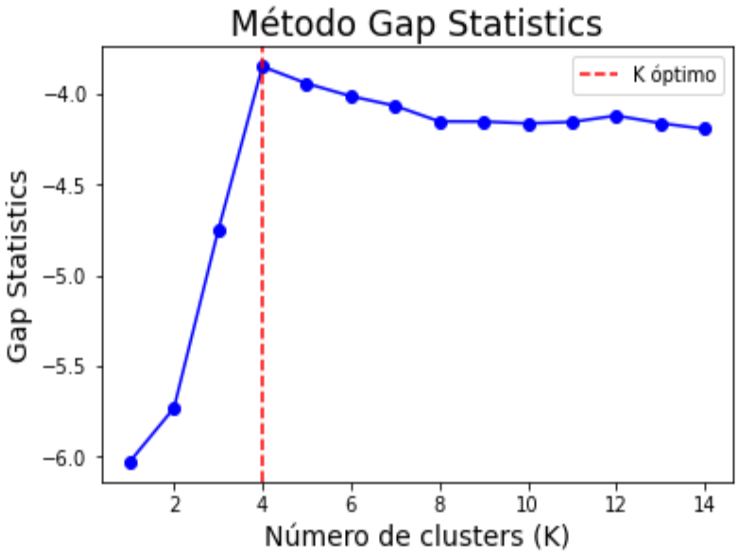

Figura 12: Ejemplo método Gap Statistics

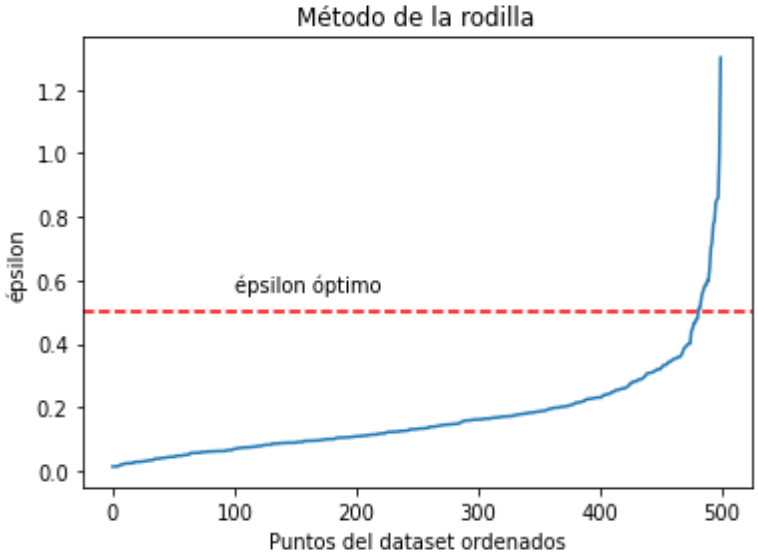

Figura 13: Ejemplo método de la rodilla

ANEXO B: ANÁLISIS DE COMPONENTES PRINCIPALES

El Análisis de Componentes Principales (PCA) es una técnica estadística utilizada para reducir la dimensionalidad de un conjunto de datos con muchas variables, manteniendo la mayor cantidad de información posible. Este método pertenece al aprendizaje no supervisado y es útil para simplificar la complejidad de los datos, identificando factores subyacentes que explican la mayor parte de la variabilidad.

PCA se basa en los conceptos de eigenvectors y eigenvalues de álgebra lineal. Un eigenvector es un vector que, al ser multiplicado por una matriz, produce un múltiplo del mismo vector. En PCA, cada componente principal es un eigenvector, y el orden de estas componentes se determina por sus eigenvalues, que indican cuánta varianza explica cada componente en los datos.

Geométricamente, las componentes principales representan direcciones en el espacio de las variables originales donde los datos muestran mayor variabilidad. La primera componente principal sigue la dirección de mayor varianza, mientras que las siguientes componentes son ortogonales entre sí, capturando la varianza restante. Esta interpretación permite visualizar cómo las componentes principales resumen la estructura de los datos en un espacio de menor dimensión.

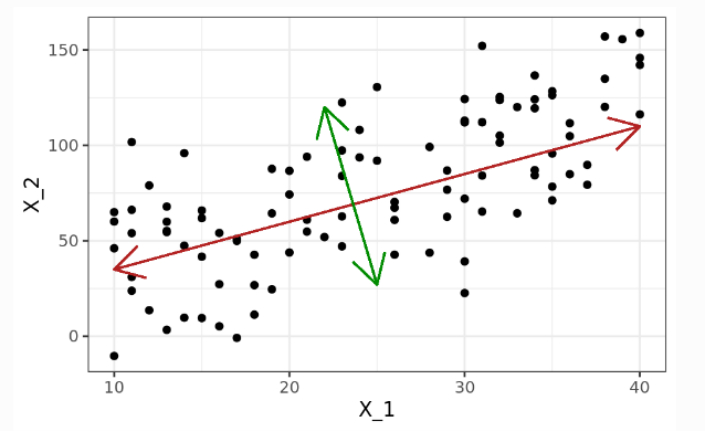

Figura 14: Ejemplo gráfico interpretación de PCA

El cálculo de las componentes principales se realiza combinando linealmente las variables originales. Primero se centralizan las variables restando la media, y luego se resuelve un problema de optimización para maximizar la varianza capturada por la primera componente. Este proceso se repite para calcular las siguientes componentes, asegurando que sean ortogonales a las anteriores. El cálculo de las componentes implica encontrar los eigenvectors de la matriz de covarianza de los datos, que representan las direcciones de mayor varianza (Rodrigo, 2017).