UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
ESCUELA DE POSTGRADO Y EDUCACIÓN CONTINUA
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

NUEVO MÉTODO PARA LA IDENTIFICACIÓN DE CLIENTES INSATISFECHOS EN EL ÁMBITO DE REDES MÓVILES

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN CIENCIA DE DATOS

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

SEBASTIÁN IGNACIO URBINA GAJARDO

PROFESOR GUÍA:
SEBASTIÁN RÍOS PÉREZ

MIEMBROS DE LA COMISIÓN:
CESAR AZURDIA MEZA
PABLO LEMUS HENRÍQUEZ
SANDRA CÉSPEDES UMAÑA

SANTIAGO DE CHILE
2024

RESUMEN DE LA TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN CIENCIA DE DATOS Y MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL
POR: SEBASTIÁN IGNACIO URBINA GAJARDO
FECHA: 2024
PROF. GUÍA: SEBASTIÁN RÍOS

NUEVO MÉTODO PARA LA IDENTIFICACIÓN DE CLIENTES INSATISFECHOS EN EL ÁMBITO DE REDES MÓVILES

En el ámbito de las redes móviles, la experiencia del cliente es crucial para la retención, siendo esencial anticipar e identificar previamente clientes insatisfechos. En un entorno competitivo, donde los ingresos por cliente disminuyen, las empresas buscan diferenciarse comprendiendo y anticipando las necesidades de los usuarios. La calidad y cobertura de la red son fundamentales, se estima que un 30 % de los clientes se fuga por problemas de red, por lo que la convergencia de aspectos técnicos de la red con la experiencia de los clientes resulta fundamental para la toma de decisiones estratégicas.

Tradicionalmente, la industria de las telecomunicaciones ha adoptado un enfoque reactivo, priorizando sitios según la insistencia de reclamos de clientes, sin considerar directamente la experiencia percibida por los usuarios. Este enfoque limita la comprensión de la satisfacción del cliente, que generalmente se monitorea a través de un número reducido de encuestas, representando menos del 0.05 % del total de clientes.

Este trabajo se centra en el desarrollo de herramientas analíticas avanzadas para abordar estos desafíos en dos frentes principales. El primero introduce una metodología para explotar los datos disponibles y así identificar zonas con deficiencias en calidad y cobertura de red, lo que impacta directamente en la satisfacción del cliente. Este enfoque se basa en el análisis de cientos de millones de registros para priorizar las inversiones en infraestructura de red, utilizando un indicador denominado Score 4G, que proporciona una perspectiva clara sobre la cobertura en todo el país.

En paralelo, se desarrolla un modelo de satisfacción de dos pasos, primero, para distinguir entre clientes detractores y no detractores, y luego para diferenciar clientes neutros y promotores. Ambos modelos logran una sensibilidad del 57 %. Con sólo siete variables es posible identificar clientes potencialmente insatisfechos, lo que facilita su aplicación a la amplia base de clientes de la compañía. Este enfoque es innovador, ya que sólo se utilizan datos de la red para evaluar la experiencia de los clientes.

Los resultados obtenidos en este trabajo ya están influyendo en la toma de decisiones proactivas, el modelo de zonas de interés ya se encuentra implementado en la compañía, priorizando proyectos de mejora de red en zonas críticas y utilizando el Score 4G para evaluar el impacto del crecimiento urbano en la cobertura y la satisfacción del cliente. Este estudio representa un avance significativo en la integración de análisis de datos y estrategias orientadas al cliente en el sector de las telecomunicaciones.

A mi madre Irma que me protege, guía e inspira desde el cielo.

Te Amo

Agradecimientos

Me gustaría dejar en escrito una mención especial a los pilares de mi vida, esas personas que me han visto crecer, me han apoyado, han confiado en mi y han sido el combustible de mi llama interior.

Primero, me gustaría agradecer a mi madre, Irma, quien luchó durante toda su vida para que saliera adelante, estudiara y fuera otro en la vida. Nada de lo que he logrado habría sido posible sin ella. Estoy seguro que desde el cielo, sigue cuidándome, guiándome e iluminando cada uno de mis días.

También a mi tío Seba, quien ha sido un pilar fundamental en mi crecimiento como persona. Agradezco enormemente su orientación y apoyo constante en mis decisiones, así como estar siempre disponible para lo que fuera necesario. Su presencia ha dejado una marca invaluable en mi vida.

Asimismo, quiero agradecer a Don Pato, una persona de bien, esforzada y trabajadora, quien me brindo su apoyo en mis estudios y estuvo presente gran parte de mi vida.

Igualmente, deseo expresar mi gratitud a mi madre biológica, Luz María, y a mi hermano José Vicente, quienes han sido una fuente constante de inspiración, motivación y apoyo emocional en mi vida. Su presencia ha enriquecido estos últimos años.

Quiero extender mi agradecimiento a mis amigos, Daniel, Eduardo y Benjamín, quienes han estado a mi lado en los momentos más felices y duros de mi vida. Su amistad ha sido una parte fundamental de este logro, y estoy agradecido por tenerlos en mi vida.

Además, quiero agradecer a María Paz, por ser mi contención emocional, una persona increíble que me ha enseñado una manera diferente de ver la vida.

Finalmente, quiero agradecer también a todas esas personas que han confiado en mi y han sido un aporte en mi desarrollo como persona y profesional. En particular, deseo dar las gracias a Martina, por dedicar parte de su tiempo a colaborar en la mejora de este documento.

Tabla de Contenido

Índice de Tablas

Índice de Ilustraciones

Capítulo 1

Introducción

1.1. Contexto

En un mundo globalizado e interconectado, para nadie es indiferente el rol fundamental que juegan las redes móviles de telecomunicaciones. En este contexto, la satisfacción del cliente resulta imprescindible para su retención[1]. Un cliente insatisfecho tiene mayor tendencia a fugarse de la compañía, por lo que es sumamente relevante poder anticiparse e identificar aquellos clientes con una mala experiencia, con el fin de tomar acciones que los mantengan fidelizados.

La entrada de nuevos competidores a un entorno de por sí altamente competitivo como es el de las telecomunicaciones, contrasta con la tendencia a la baja que ha experimentado el ingreso promedio por cliente [2]. Por tal razón, las empresas de telecomunicaciones se ven obligadas a esforzarse por diferenciarse y proporcionar un valor agregado a sus clientes. Por lo tanto, comprender y anticipar las necesidades de los usuarios se ha convertido en una prioridad crucial para las empresas dentro de este contexto dinámico.

La calidad de la red es un componente esencial en la experiencia del cliente. Los usuarios esperan una conectividad constante, rápida y confiable, considerando una cobertura adecuada como un requisito básico.

A pesar de los avances tecnológicos en el sector de las telecomunicaciones, persiste una desconexión entre las dimensiones operativas y de negocio. Mientras que el ámbito operativo se centra en los indicadores claves de rendimiento (KPIs) relacionados con la infraestructura de red, el área de negocio se enfoca en la retención y satisfacción del cliente y en el desarrollo de nuevos servicios [3]. Esta divergencia resulta en una brecha significativa entre las experiencias percibidas por los clientes y las expectativas manejadas por los operadores.

Esta división de roles dentro de las organizaciones revela oportunidades para mejorar la integración de los datos de la red con la experiencia del cliente. El análisis de grandes volúmenes de datos se presenta como una herramienta valiosa para vincular la satisfacción del cliente, medida a través de encuestas, con su experiencia en la red, integrando así el aspecto operativo y de negocio. Este enfoque analítico busca no solo comprender la experiencia del cliente desde un punto de vista técnico, sino también transformar insights en estrategias y acciones concretas para mejorar la satisfacción de los clientes.

1.1.1. Comprensión del Negocio

El crecimiento sostenido del mercado de redes móviles de telecomunicaciones en Chile se puede evidenciar principalmente a través de dos grandes aristas: La primera, en el aumento anual de un 29 % en el tráfico, evidenciando el masivo consumo de internet móvil y la creciente relevancia de los operadores en la vida cotidiana. La segunda, en la penetración de mercado, alcanzando 133 abonados por cada 100 habitantes, lo que sugiere que la mayoría de las personas disponen de más de un servicio de telefonía móvil, ya sea por poseer múltiples dispositivos o por el uso de líneas secundarias. En total, el número de abonados a servicios de telefonía móvil asciende a 26,4 millones [4].

Dentro del mercado de internet móvil, cuatro empresas predominan en Chile: Entel, Movistar, Claro y WOM. La evolución de las conexiones en las redes 3G, 4G y 5G, ilustrada en la Figura 1.1, muestra el notable crecimiento de WOM desde su ingreso al mercado en 2015. WOM ha desafiado a los operadores establecidos, fomentando una competencia que ha llevado a una reducción en los precios de los planes móviles, y como resultado, a una disminución en el ingreso promedio por cliente [2].

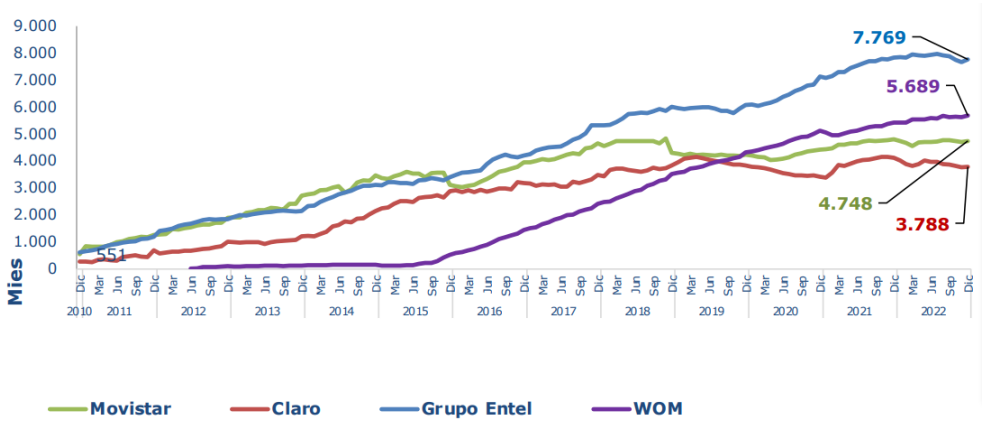


Figura 1.1: Conexiones 3G+4G+5G por empresa Internet Móvil [4]

Respecto al total de clientes abonados, el liderazgo lo mantiene Entel con 10 millones. Así lo indica la distribución de los abonados entre estas empresas desde el año 2010 al año 2022 representada en la Figura 1.2. Asimismo, es interesante como se estrecha la competencia a lo largo del tiempo.

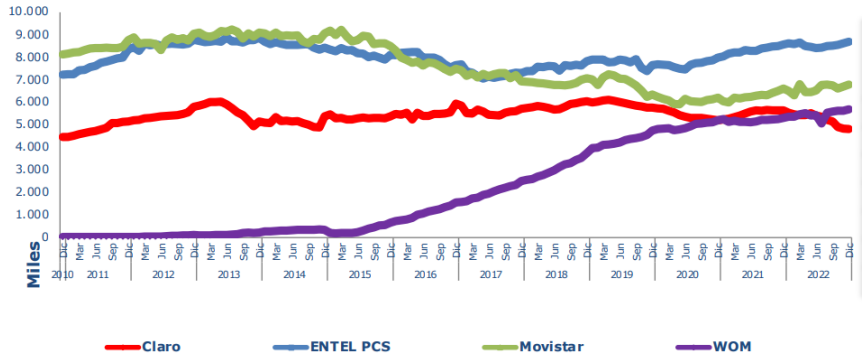


Figura 1.2: Total abonados Telefonía Móvil por empresa [4]

Ambos gráficos proporcionan una panorámica del dinamismo y la intensa competitividad en la industria de los operadores de redes móviles en Chile. En este entorno, es crucial profundizar en la experiencia del cliente como un diferenciador clave en este entorno altamente disputado. De acuerdo con [5], aspectos como la calidad y cobertura de la red son primordiales y se encuentran entre los factores más valorados al evaluar un servicio de red móvil. Esta interrelación entre la competitividad del mercado y la calidad de la experiencia del cliente refuerza la necesidad de una constante innovación y mejora en el sector.

El presente trabajo abordará dos áreas principales. La primera área de estudio se centrará en la correlación entre la calidad y cobertura de la red y la satisfacción del cliente, con la finalidad de desarrollar una metodología para identificar zonas geográficas con deficiencias en estos aspectos que afectan la satisfacción. Este enfoque se basa en el análisis de millones de registros diarios de clientes relacionados con indicadores de red.

En la segunda área de este estudio, se diseñará y desarrollará un modelo de Machine Learning orientado a clasificar a los clientes en tres categorías: detractores, neutros y promotores. Este modelo utilizará datos de la red que reflejan la calidad y cobertura de sus conexiones en un período determinado. Con la finalidad de comprender cómo estas variables de red influyen directamente en la percepción y satisfacción del cliente.

La integración de las dos áreas principales abordadas en este trabajo busca ofrecer herramientas para mejorar la toma de decisiones de cara a la experiencia de los clientes. La primera área permite la priorización efectiva de mejoras en la red y servicios, especialmente en aquellas zonas críticas identificadas, con el fin de elevar la satisfacción del cliente. Por otro lado, el modelo de satisfacción del cliente ofrece insights fundamentales para la identificación y atención proactiva de clientes insatisfechos, potenciando estrategias de retención y mejora de la experiencia del cliente.

1.1.2. Net Promoter Score (NPS)

Comprender la satisfacción del cliente es fundamental en el ámbito empresarial, debido a que las decisiones estratégicas de la compañía deben tener siempre como foco a los clientes. Una herramienta clave para realizar seguimiento a la experiencia es el *Net Promoter Score* (NPS), el cual permite medir satisfacción de los clientes con respecto a una marca, producto o servicio. Fue desarrollado por Fred Reichheld, Bain & Company y Satmetrix a principios de la década de 2000 [6, 7].

Según este modelo, las calificaciones pueden dividirse en cuatro rangos distintos. Un rango que va de -100 a -50 se define como deficiente, reflejando una percepción extremadamente negativa de los clientes hacia el servicio o producto. Una calificación entre -49 y 0 se considera insuficiente, indicando aún una percepción negativa, pero menos severa. Por otro lado, un rango de 0 a 49 se etiqueta como suficiente, mostrando una aceptación o satisfacción moderada, y finalmente, una puntuación de 50 a 100 se clasifica como excelente, lo que denota una percepción altamente positiva y una probable lealtad hacia la marca. En este sistema, cualquier puntuación por debajo de 0 se interpreta como insatisfactoria, mientras que una puntuación por encima de 0 se considera como indicativa de satisfacción del cliente.

La medición de satisfacción se basa en las respuestas de los clientes a través de encuestas que consideran tres aspectos; general, internet y llamadas. Los clientes son consultados a través de llamadas telefónicas y responden las siguientes preguntas.

- ¿Qué nota le pondría a la compañía? (Nota CIA)
- ¿Qué nota le pondría al internet móvil? (Nota IM)
- ¿Qué nota le pondría al servicio de llamadas? (Nota VOZ)

En una escala de evaluación es de 1 a 7, donde 1 es la peor nota y 7 la mejor se categorizan los clientes bajo los siguientes criterios, inspirados en el NPS.

- Detractores: Nota menor a 5
- Neutros: Nota igual a 5
- Promotores: Nota mayor a 5

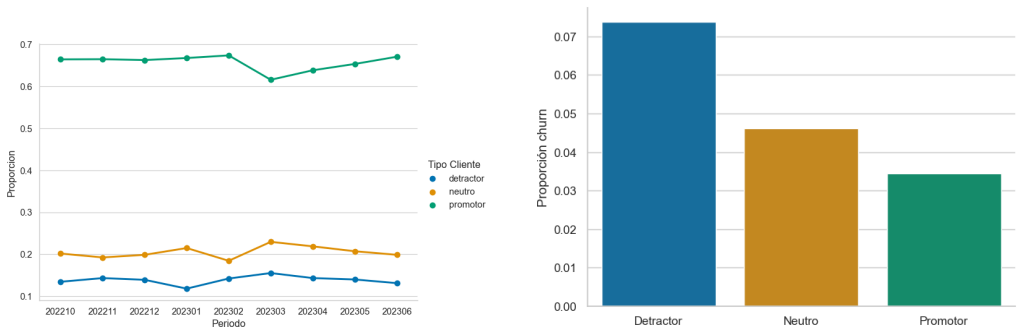
De esta manera se puede calcular el *Net Promoter Score* o NPS como sigue:

$$NPS = \frac{\#\text{Promotores} - \#\text{Detractores}}{\text{Total encuestados}} * 100 \tag{1.1}$$

Satisfacción y Fuga

El análisis histórico de la proporción de encuestados por tipo de cliente en la compañía revela una tendencia interesante: la proporción de los clientes según las categorías de satisfacción ha mostrado una consistencia notable a lo largo de los últimos meses, tal y como se ilustra en la Figura 1.3.a. Esta estabilidad en las proporciones ofrece una perspectiva valiosa sobre la dinámica de la satisfacción del cliente en la empresa.

Además, los datos indican que ha habido una fuga proporcionalmente mayor de clientes detractores, aquellos que manifiestan insatisfacción con el servicio. Este patrón subraya una correlación significativa entre la insatisfacción del cliente y su propensión a abandonar la compañía, como se evidencia en la Figura 1.3.b.



(a) Proporción de tipo de cliente por periodo

(b) Proporción de fuga por tipo de cliente último semestre

Figura 1.3: Satisfacción y Fuga

1.2. Motivación

El desaprovechamiento evidente de gran volumetría de información que maneja el operador de telecomunicaciones para la toma de sus decisiones de negocio, enfocadas en la satisfacción del cliente es la motivación principal de este trabajo. Se utilizarán datos relacionados a la conexión que experimenta el cliente con la red móvil celular, por ejemplo, indicadores de cobertura, calidad, tráfico y movilidad.

Según se indica en [3], se prevee que durante la próxima década haya un cambio paradigmático en el sector de las telecomunicaciones, desde un enfoque centrado en las infraestructuras de antenas a uno que priorice la experiencia y satisfacción de los clientes.

Este cambio de paradigma implica que las empresas de telecomunicaciones deben adaptarse rápidamente para satisfacer las demandas de los clientes en un entorno altamente competitivo. En este contexto, la utilización eficiente de la gran cantidad de información disponible se vuelve crucial para tomar decisiones estratégicas que mejoren la calidad de los servicios ofrecidos y fortalezcan la lealtad de los usuarios. u

1.2.1. Hipótesis

La idea principal que se aborda en el presente trabajo es la necesidad de integrar los aspectos técnicos de la red y la satisfacción del cliente. De esta necesidad surge la hipótesis de que el análisis de datos de red y la aplicación de técnicas de analítica avanzada pueden mejorar la identificación de clientes insatisfechos a través de relacionar la cobertura con la satisfacción de los clientes.

Este estudio propone comprender cómo los clientes experimentan y perciben los servicios de la red móvil de un operador de telecomunicaciones y presentar herramientas para mejorar la toma de decisiones estratégicas, proponiendo considerar dos enfoques:

1. **Zonas de Interés de Calidad y Cobertura:** La cobertura de red, siendo un fenómeno físico que se extiende a través del espacio mediante ondas electromagnéticas, es esencial para proporcionar acceso y conectividad a los usuarios. Este enfoque busca la identificación de áreas densamente transitadas por clientes que experimentan baja calidad y cobertura de red mediante la explotación de cientos de millones de datos generados por los teléfonos de los clientes. El desarrollo de este enfoque se detallará en el Capítulo 3.
2. **Modelo de Satisfacción:** Desde el cliente se caracterizará a los usuarios utilizando variables relacionadas con la red, con el fin de identificar aquellos que potencialmente están insatisfechos. Para ello se propondrá un modelo de clasificación de dos etapas que sólo utiliza variables de la red para explicar la experiencia de los clientes. El desarrollo de este enfoque se presentará en el Capítulo 4.

1.3. Objetivos

1.3.1. Objetivo General

Desarrollar un nuevo método de identificación de clientes insatisfechos del servicio móvil de una empresa de telecomunicaciones mediante el análisis de datos de calidad de red y técnicas de analítica avanzada.

1.3.2. Objetivos Específicos

Para lograr el objetivo general, se presentan los siguientes objetivos específicos.

1. Diseño y evaluación de metodología para identificar zonas de interés de cobertura móvil utilizando técnicas de análisis geoespacial.
2. Diseño de variables que caractericen a los clientes en base a datos de calidad y cobertura de red.
3. Diseño y evaluación de un modelo de *Machine Learning* que permita cuantificar la percepción de la red en función de datos de potencia y calidad móvil.

1.4. Resultados Esperados y Alcances

De acuerdo con la investigación y análisis realizados en el presente trabajo, se plantean como resultado esperados los siguientes:

1. Modelo de *Machine Learning* para identificar clientes detractores, neutros y promotores en función de datos de calidad y cobertura de red. El modelo quedará a disposición de la compañía quedando en ella la decisión de utilizarlo para la toma de decisiones.
2. Desarrollo de una metodología que permita identificar áreas de interés de cobertura móvil de manera eficiente utilizando técnicas de análisis geoespacial.
3. Creación de mapas y visualizaciones que muestren la distribución espacial de la cobertura de red y que resalten las áreas de interés. El alcance de este punto estará acotado a la región metropolitana debido a que concentra la mayor cantidad de clientes de la compañía.
4. Generar un KPI de cobertura que tenga estrecha relación con la satisfacción y permita cuantificar la cobertura a nivel de comuna o región. El alcance del KPI corresponderá a un procesamiento a nivel nacional.

1.5. Estructura del Trabajo

Para alcanzar el objetivos generales y lograr los objetivos específicos señalados, el presente trabajo se estructura en 5 capítulos que organizan de una manera clara el contenido del mismo, haciéndolo a su vez más ordenado e inteligible.

1. Capitulo 1: Introducción, objetivos y resultados esperados.
2. Capitulo 2: Metodología, revisión bibliográfica y marco teórico.
3. Capitulo 3: Metodología de Zonas de Interés de Calidad y Cobertura de red.
4. Capitulo 4: Modelo de Satisfacción
5. Capitulo 5: Conclusiones, trabajo futuro y propuestas para la organización.

Capítulo 2

Metodologías y Técnicas

2.1. Calidad de Servicio y Calidad de la Experiencia

La Experiencia de Usuario, también conocida como Calidad de la Experiencia (QoE, por sus siglas en inglés), es un concepto integral que abarca todos los elementos percibidos por el cliente respecto a un servicio y su capacidad para satisfacer sus expectativas. Este concepto, según Nokia [8], adquiere una relevancia particular en el sector de las telecomunicaciones debido a su impacto directo en el negocio.

Según se indica en [9] la calidad de la experiencia es un área de investigación emergente que evalúa distintos aspectos de la percepción de un usuario frente a un servicio. Se mide principalmente a través de dos métodos: subjetivo y objetivo. Por un lado, la medición subjetiva involucra la participación del usuario y busca cuantificar la QoE en términos de la puntuación MOS [10], utilizando una escala discreta de 5 puntos en un entorno controlado y sin considerar las expectativas de la persona. Por otro lado, la medición objetiva busca estimar la QoE a través de un modelo paramétrico sin la necesidad de contar con la participación del usuario y se fundamenta en mediciones de Calidad de Servicio (QoS, por sus siglas en inglés) de indicadores de red.

En el desarrollo del presente trabajo, se adoptará un enfoque integrador que combina la medición objetiva y subjetiva, con un interés particular en el análisis y utilización de datos relacionados con la cobertura y calidad de red, además de encuestas de satisfacción. Este enfoque busca capturar la experiencia móvil tal como la perciben los usuarios, reconociendo que esta percepción puede diferir del rendimiento técnico esperado de la red, tal como indica [11].

Este trabajo se distingue en el ámbito de las telecomunicaciones por su enfoque dual. En el primer enfoque, se explorará la identificación de "Zonas de Interés de Calidad y Cobertura". Aquí, se analizará la cobertura de red geoespacialmente como un fenómeno físico crucial para proporcionar acceso y conectividad, utilizando datos georeferenciados de las sesiones de conexión de los clientes para identificar áreas densamente transitadas con baja calidad y cobertura de red. Este análisis, detallado en el Capítulo 3, va más allá de los enfoques tradicionales centrados únicamente en aspectos técnicos.

Por ejemplo, en el estudio [12], se propone una metodología para detectar agujeros de cobertura en las celdas, definiendo un agujero como un debilitamiento de señal recibida por la celda de servicio, estando bajo los niveles mínimos requeridos para ofrecer un servicio de calidad y rendimiento de radio adecuados. Mientras tanto, [13] propone identificar áreas con problemas de cobertura, utilizando DBSCAN para agrupar zonas con bajo RSRP y generar recomendaciones sobre posibles causas. Estos estudios, aunque valiosos, se centran principalmente en los aspectos técnicos de la cobertura de red a nivel de celda y no de cliente.

Adicionalmente, en [14] se propone un modelo para predecir la cobertura en áreas rurales, integrando datos de múltiples fuentes con un historial de seis meses. Por su parte, [15] aborda la detección autónoma de celdas defectuosas en redes LTE, empleando datos de móviles de usuarios y un software de simulación, con el uso de K-Means y N-Gramas para identificar anomalías en la red.

A diferencia de estos estudios, los cuales utilizan datos desde las celdas y se enfocan solamente en QoS, el presente trabajo considera la interacción entre la calidad de la red y la experiencia del usuario, combinando el análisis técnico con la percepción del cliente. Esto permite no solo identificar áreas de baja cobertura y calidad, sino también comprender cómo estas afectan la experiencia del usuario, ya que los datos provienen desde los dispositivos de los clientes, un aspecto crucial que ha sido menos explorado en investigaciones previas, posiblemente al costo económico de obtener encuestas representativas y de infraestructura de datos.

El segundo enfoque, que se presentará en el Capítulo 4, se centra en un "Modelo de Satisfacción" desde la perspectiva del cliente. Aquí, se propondrá un modelo de clasificación de dos etapas utilizando exclusivamente variables de red para discernir usuarios potencialmente insatisfechos. Este enfoque representa una innovación significativa en comparación con estudios anteriores. Por ejemplo, [16] explora el uso de redes neuronales para estimar la QoE, correlacionando indicadores de red con la experiencia de los clientes y definiendo umbrales para ciertos tipos de experiencia. Mientras tanto, [17] se enfoca en modelar la QoE percibida a través del análisis de big data, considerando factores como el tiempo, la ubicación y diferentes servicios móviles. Además, [18] utiliza la teoría de decisión y redes Bayesianas para abordar las interdependencias complejas entre los parámetros de QoE y el contexto del usuario. Un enfoque similar se presenta en [19], donde se predice la QoE para servicios de voz y navegación web utilizando mediciones de RF y métricas de QoS.

A diferencia de los estudios mencionados en el párrafo anterior, que aunque relacionan la Calidad de Servicio (QoS) con la Calidad de la Experiencia (QoE), no incorporan una evaluación directa de la satisfacción del cliente a través de encuestas, sino mas bien utilizan modelos paramétricos para estimar el MOS de QoE y en función de ello relacionar variables de red. De los pocos trabajos similares se tiene el presentado en [20], que utiliza encuestas para diferenciar clientes según su percepción de la red, sin embargo, no evalúa distintos algoritmos de clasificación y sólo usa un modelo directamente sobre tres clases de clientes. El presente trabajo se beneficia del análisis de más de 20.000 encuestas de clientes. Esto proporciona una perspectiva única y valiosa sobre la percepción del cliente, ofreciendo una comprensión más profunda de la satisfacción del usuario en relación con la calidad y cobertura de la red.

Los principales aspectos que diferencian al presente trabajo tienen que ver con considerar encuestas de satisfacción, datos de cobertura y datos de calidad. Una tabla resumen de algunos artículos analizados se presenta en al siguiente Tabla 2.1.

Tabla 2.1: Comparativa Artículos

Autor, año	Breve Descripción	Encuestas de Satisfacción	Datos de Cobertura	Datos de Calidad	Comentarios
Gómez-Andrades et al., 2016[12]	Se propone una metodología para detectar agujeros de cobertura en las celdas. Un agujero es cuando la señal recibida de la celda de servicio y los vecinos está bajo niveles que permitan ofrecer el servicio a un mínimo nivel de calidad y rendimiento de radio.	No	Si	No	Definen umbrales para calificar las conexiones de clientas buenas/malas. No consideran el impacto de la cobertura en la experiencia ni utilizan algún modelo, solo umbrales.
Qiao et al., 2023[13]	Proponen una metodología para identificar áreas con problemas de cobertura y generan una recomendación sobre posibles causas. Utilizan DBSCAN para agrupar zonas con bajo RSRP.	No	Si	No	Consideran una mala cobertura bajo un RSRP de -110. Filtro que aplicar para agrupar zonas con problemas. Además, usan una grilla cuadrada para promediar el RSRP por zonas.
Lyu et al., 2022[14]	Proponen un modelo de cobertura, para predecir la cobertura en ciertas áreas rurales. Utilizan datos de múltiples fuentes de información con 6 meses de historia.	No	Si	No	El modelo funciona como una alternativa para estimar la cobertura donde no existen datos.
Chernov et al., 2015[15]	Aborda el tema de detección autónoma de celdas defectuosas en red LTE utilizando datos de los móviles de usuarios desde un software de simulación. Utilizan K-Means para identificar anomalías en la red y lo combinan con N-Gramas.	No	Si	No	No consideran el impacto de las celdas defectuosas en la satisfacción. Sólo se aborda la detección de celdas anómalas.
Pierucci and Micheli, 2016[16]	Explora el uso de redes neuronales para estimar la QoE. Se centra en correlacionar indicadores de red con la experiencia de los clientes y define umbrales para ciertos tipos de experiencia.	No	Si	No	No utilizan encuestas como variable objetivo de experiencia, sólo definen umbrales bajo la hipótesis que deberían impactar en satisfacción.
Yusuf-Asaju et al., 2017[17]	Se busca modelar la QoE percibida a través de análisis de big data. El estudio propone un marco para modelar la QoE percibida, considerando factores como el tiempo, la ubicación y diferentes servicios móviles.	No	Si	No	Este estudio sirve como guía para entender los principales factores que pueden influir la satisfacción de los clientes.
Mitra et al., 2011[18]	Este enfoque utiliza la teoría de decisión y redes Bayesianas para manejar interdependencias complejas entre parámetros de QoE y contexto del usuario, incluyendo aspectos ambientales y de dispositivo. Se enfoca en medir y predecir QoE en una escala única	No	No	No	Es un trabajo teórico y no utiliza datos de satisfacción reales.
Pedras et al., 2018[19]	Se presenta un modelo de Calidad de Experiencia (QoE) para servicios de voz y navegación web en redes 3G y 4G. Este modelo predice la QoE percibida por los usuarios en una escala de Puntuación de Opinión Media (MOS), utilizando mediciones de Radiofrecuencia (RF) y métricas de Calidad de Servicio (QoS).	No	Si	Si	Las variables más relevantes que identifican tienen que ver con cobertura.
Qiao et al., 2022[20]	Se propone un modelo de clasificación de random forest para clasificar clientes según su percepción de la red utilizando sólo datos de redes.	Si	Si	No	Sólo se evalúa un modelo de clasificación y las variables se calculan en un periodo estadístico de 7 días. La variable más relevante para clasificar la percepción resulta ser la cobertura.
Wu et al., 2009[21]	Proponen una definición clara de QoE y un modelo conceptual.	No	No	No	Este trabajo sirve como guía para entender las diferencias entre QoE y QoS. Es un trabajo teórico.

2.2. Redes Móviles

Para efectos del trabajo no se profundizará los aspectos técnicos de las redes móviles y únicamente se mencionarán los conceptos y/o fundamentos claves para el desarrollo de la tesis.

Una red de telecomunicaciones es un conjunto de dispositivos conectados entre sí que permite la transmisión de datos, voz y otros tipos de señales. Las redes de telecomunicaciones se utilizan en una amplia gama de aplicaciones, desde la comunicación personal hasta el transporte de datos a gran escala. Este trabajo está acotado a las redes móviles celular, las cuales son una infraestructura de telecomunicaciones que permite a los usuarios comunicarse sin la necesidad de conectarse a una linea fija [22].

A lo largo del continuo desarrollo de las telecomunicaciones, se han experimentado transformaciones significativas en las tecnologías que respaldan la comunicación móvil. La primera generación de telefonía móvil, conocida como 1G, surgió en la década de los 80, marcando un hito inicial en la capacidad de realizar llamadas móviles. Esta tecnología se caracterizaba por la transmisión analógica de voz y una cobertura limitada. [3]

Con el avance del tiempo, las sucesivas generaciones de telefonía móvil, como la 2G, 3G y 4G o LTE (*Long Termn Evolution)*, introdujeron mejoras sustanciales en las capacidades de transmisión de datos y la calidad de las comunicaciones. La segunda generación (2G) permitió la transición a la comunicación digital, brindando mejor calidad de voz y la capacidad de enviar mensajes de texto. La tercera generación (3G) llevó consigo velocidades de transmisión de datos más rápidas y habilitó servicios multimedia, como la transmisión de vídeo. La cuarta generación (4G) supuso un salto significativo en la velocidad de conexión, posibilitando la proliferación de servicios avanzados, como la navegación web móvil y la transmisión de datos a alta velocidad. [3]

El despliegue de la quinta generación (5G) en los últimos años representa el último hito en esta evolución. La tecnología 5G se caracteriza por velocidades de datos ultra rápidas, baja latencia y una capacidad de conexión masiva de dispositivos.

Los teléfonos móviles han evolucionado paralelamente a las generaciones de telecomunicaciones, mejorando notablemente en términos de capacidad de procesamiento, funcionalidades y usabilidad. Estos dispositivos son mucho más que simples herramientas de comunicación; se han convertido en centros de información, entretenimiento y gestión personal, todo accesible desde la palma de la mano.

La experiencia del usuario con su equipo móvil depende en gran medida de la tecnología de la red a la que se conecta. Por ejemplo, las redes 4G permiten a los usuarios disfrutar de una navegación web más rápida y eficiente respecto a la generación predecesora.

En el entorno de las telecomunicaciones móviles, el usuario interactúa con la red a través de su equipo móvil conectándose a una celda, la unidad mínima de cobertura en la cual el cliente obtiene recursos para navegar por internet. Estas celdas pertenecen a un sector específico y están presentes en un Punto de Presencia (POP) o estación base (BS), que es el lugar donde conviven múltiples celdas de diferentes tecnologías. De esta manera el usuario recibe

la cobertura y puede acceder a internet. (Figura 2.1)

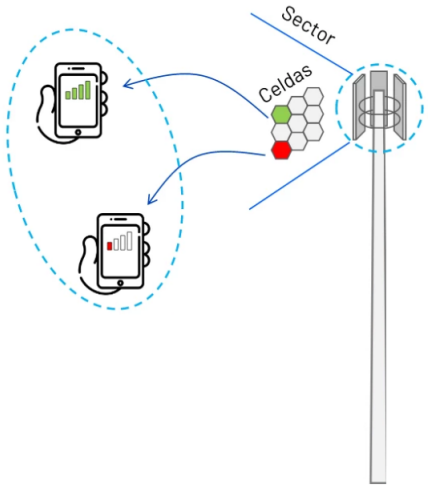


Figura 2.1: Esquema POP o BS

Cobertura y Calidad Móvil

En el ámbito de la cobertura móvil, especialmente en redes 4G, el indicador primordial es la señal de referencia de potencia recibida (RSRP, por sus siglas en inglés). Este parámetro se mide en decibelios-milivatio (dBm) y constituye una métrica clave para evaluar la intensidad o potencia de la señal que capta el teléfono móvil del usuario desde la celda de servicio. Su importancia radica en que el RSRP proporciona una medida directa de la cantidad de potencia de la señal que llega al teléfono móvil desde la estación base (BS), siendo esencial para determinar la cobertura de la red. Esta métrica no solo es crucial para la planificación y optimización de la red, sino que también influye significativamente en la calidad de la comunicación de voz y datos en la red móvil, ya que una señal fuerte y clara es fundamental para una transmisión eficiente [23].

En paralelo a la intensidad de la señal, otro elemento esencial es su calidad, la cual se mide a través del indicador de la relación de calidad de señal recibida (RSRQ, por sus siglas en inglés). El RSRQ, medido en decibelios (dB), es una métrica específica de las redes 4G diseñada para evaluar la calidad de la señal recibida en comparación con el nivel de interferencia que experimenta el teléfono móvil. La relación entre RSRP y RSRQ es fundamental, ya que el RSRP indica la intensidad de la señal, mientras que el RSRQ evalúa la calidad de esta señal teniendo en cuenta las interferencias presentes [23].

$$RSRQ = \frac{N \cdot RSRP}{RSSI} \tag{2.1}$$

La ecuación anterior relaciona RSRQ y RSRP. N representa el número de bloques de recursos físicos (PRB, por sus siglas en inglés) y el Indicador de la Intensidad de la Señal Recibida (RSSI, por sus siglas en inglés) provee información sobre la potencia total recibida, incluyendo interferencias [23].

Asimismo, es pertinente mencionar que en las redes 3G, los equivalentes a estos indicadores son el RSCP (Received Signal Code Power) para la intensidad de la señal y el EC/NO (Energy per Chip/Noise Ratio) para la calidad.

En última instancia, para garantizar una experiencia de usuario óptima en redes 4G (y 3G, con RSCP y EC/NO respectivamente), es imprescindible que tanto el RSRP como el RSRQ se mantengan en niveles adecuados. Un alto valor en estas métricas asegura una señal robusta y de calidad. Sin embargo, una excelente cobertura no siempre se traduce en una mayor velocidad de descarga, dado que factores como la congestión de la celda pueden afectar el rendimiento de la red.

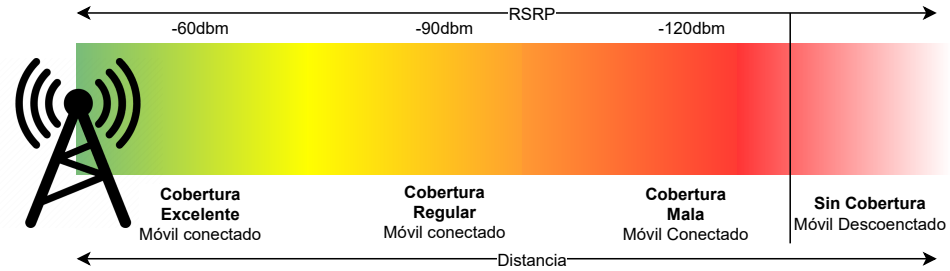


Figura 2.2: Esquema Conceptual de RSRP [24]

En la Figura 2.2 se muestra de manera esquemática como se degrada el RSRP a medida que un móvil se aleja de la celda de servicio, llegando a un punto donde deja de tener cobertura. Cabe mencionar que no es el único factor que afecta el nivel de RSRP, por ejemplo, también influye el entorno de la celda (urbano, rural), la velocidad del teléfono móvil y características propias del canal de transmisión.

Además, es relevante mencionar que la cobertura de la señal móvil también depende de la morfología del lugar. Por ejemplo, en la Figura 2.3 se ilustra el caso de un teléfono móvil en 4 posiciones distintas, las barras del móvil representan el RSRP. Muy cerca de la antena (a), con cobertura excelente, al lado de una casa con árboles (b) donde la cobertura no es tan buena porque hay árboles entre la antena y el dispositivo. En (c) presenta mejor cobertura que en (b) porque está en un lugar más libre y en (d) no tiene cobertura porque está lo suficientemente lejos de la antena.

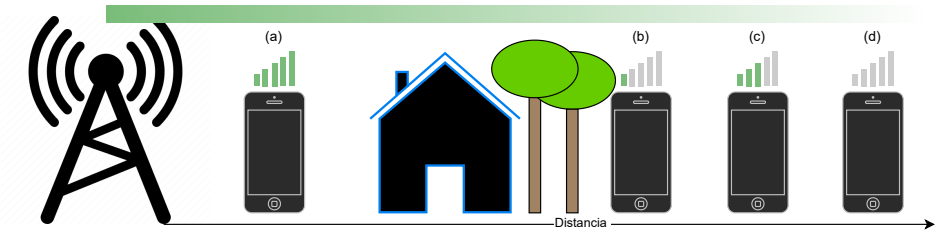


Figura 2.3: Esquema atenuación de la señal de cobertura | Elaboración propia

La cobertura tiene distintos matices que dependen del tipo de tecnología. En lo específico, la unidad mínima de cobertura es una celda asociada sólo a un tipo de tecnología, que puede

ser 2G, 3G, 4G o 5G con una respectiva banda de frecuencia, En el caso de 4G, en Chile se operan las bandas de 700 MHz, 1900 MHz y 1700-2100 MHz masivamente por los operadores de telecomunicaciones. [25]

Un aspecto relevante sobre las bandas es que tienen estrecha relación con la cobertura que se provee físicamente en el lugar, es decir, una banda con menor frecuencia, por ejemplo, de 700 MHz, tendrá un mayor alcance en términos de cobertura como se ilustra en la Figura 2.4.

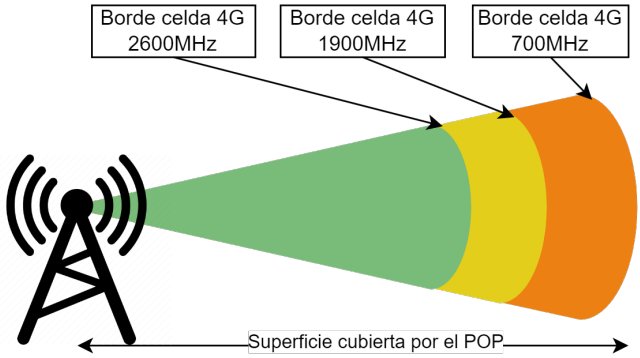


Figura 2.4: Esquema cobertura POP | Elaboración Propia

Es importante mencionar que en este trabajo no existirá distinción de las frecuencias a las cuales navega el cliente debido a que no se cuenta con esa información.

2.3. Modelos de Machine Learning

2.3.1. Regresión Lineal

Un aspecto relevante del negocio, mas allá de encontrar correlaciones entre red y satisfacción, consiste es entender la sensibilidad de una variable especifica con respecto a la satisfacción de los clientes. Con este propósito se utilizará una regresión lineal, la cual permitirá cuantificar el impacto esperado sobre el NPS ante una variación en la calidad y cobertura percibida por los clientes.

Una regresión lineal permite modelar la relación entre una variable dependiente y una o más variables independientes. Su objetivo es encontrar la mejor línea recta que se ajuste a los datos y así explicar la variable y en términos de x. [26]

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i \tag{2.2}$$

Donde:

- y_i es el valor observado de la variable dependiente en la i-ésima observación.
- β_0 es la ordenada al origen, que representa el valor estimado de y cuando todas las variables independientes son iguales a cero.
- $\beta_1, \beta_2, \ldots, \beta_p$ son los coeficientes de regresión, que indican el cambio esperado en y por un aumento unitario en cada variable independiente, manteniendo las otras constantes.

- $x_{i1}, x_{i2}, \ldots, x_{ip}$ son los valores de las variables independientes en la i-ésima observación.
- ε_i es el término de error, que representa la diferencia entre el valor observado y_i y el valor predicho $\hat{y}_i$ por el modelo.

Los coeficientes de regresión se estiman a partir de los datos de entrenamiento utilizando el método de mínimos cuadrados. El objetivo es minimizar la suma de los cuadrados de los residuos (diferencias entre los valores observados y predichos) para obtener los coeficientes β_i.

2.3.2. Clasificación

Uno de los objetivos relevantes del presente trabajo consiste en desarrollar un modelo de *Machine Learning* para cuantificar la percepción de la red pudiendo diferenciar clientes detractores, neutros y promotores. Esta tarea de clasificación se abordará mediante modelos de *Machine Learning* basados en árboles, seleccionados debido a su facilidad de interpretación y destacado desempeño en tareas de clasificación. Es por ello que a continuación se describe a grandes rasgos el funcionamiento de los modelos: Random Forest, XGBoost y LightGBM.

Random Forest

Random Forest es un algoritmo de aprendizaje supervisado utilizado para problemas de clasificación y regresión. Esu un modelo de ensamblaje (*ensemble*) que combina múltiples árboles de decisión (modelos simples) para mejorar la precisión de la predicción. Cada árbol en el bosque es construido de manera independiente utilizando una muestra aleatoria de los datos y un subconjunto aleatorio de las características. Luego, los árboles individuales votan por la clase o valor de la variable objetivo y la predicción final se calcula como la media o la moda de los votos de los árboles individuales. [27]

En la Figura 2.5 se visualiza el diagrama que muestra de manera general el funcionamiento de *Random Forest*.

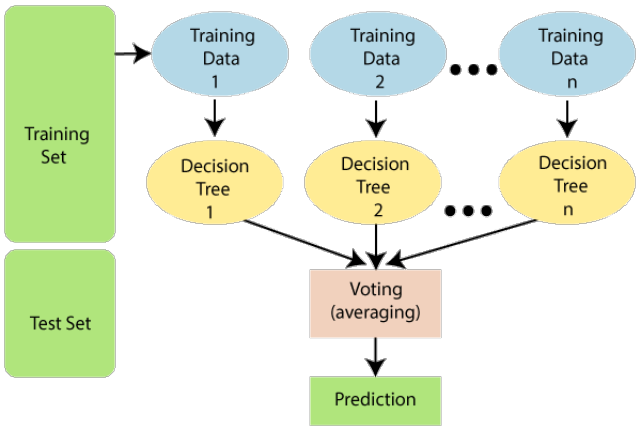


Figura 2.5: Funcionamiento *Random Forest*[28]

El algoritmo *Random Forest* tiene varias ventajas sobre los árboles de decisión individuales, como la reducción del sobreajuste (*overfitting*) y la mejora de la estabilidad y precisión de la predicción. Además, es capaz de manejar grandes conjuntos de datos con múltiples

características y de detectar la importancia de cada característica para la predicción.

Desde una perspectiva técnica, Random Forest implementa la técnica de *bagging* (Bootstrap Aggregating) para generar múltiples árboles de decisión. Estos árboles se construyen utilizando distintos subconjuntos de entrenamiento, derivados del conjunto de datos original. La clave de esta técnica radica en la reducción de la correlación entre los árboles individuales, lograda a través de la selección aleatoria tanto de muestras como de características en cada iteración. Este proceso comprende etapas sucesivas: la selección aleatoria de muestras y características, la construcción de cada árbol de decisión, la asignación de pesos específicos a las muestras, y finalmente, la ponderación de las predicciones provenientes de cada árbol para formar una predicción consolidada y robusta.

Boosting

La técnica de *boosting*, propuesta por Friedman en 2001 [29] a diferencia de *bagging* crea árboles de manera secuencial de modo que cada árbol subsiguiente tiene por objetivo reducir los errores del árbol anterior.

Este proceso funciona mediante la asignación de pesos a cada observación en el conjunto de datos de entrenamiento. En cada iteración, el modelo se enfoca en las observaciones que fueron mal clasificadas en la iteración anterior, aumentando su peso para que sean más importantes en el próximo modelo. De esta manera, el modelo posterior se enfoca en corregir los errores cometidos por el modelo anterior, lo que permite que el modelo final tenga un rendimiento significativamente mejor.

Existen varios algoritmos basados en *boosting*, sin embargo, en este estudio se utilizará *XGBoost* y *LightGBM* dado que son considerados algoritmos estado del arte y han demostrado ser los algoritmos por excelencia en tareas de clasificación. [30]

XGBoost

eXtreme Gradient Boosting (XGboost) es un algoritmo de aprendizaje automático ampliamente utilizado en problemas de clasificación que ha demostrado ser altamente efectivo en diversas tareas de predicción. XGBoost se destaca por su capacidad para producir modelos de alta calidad y su eficiencia computacional. [31]

Este algoritmo utiliza un enfoque de crecimiento de árboles nivel por nivel (*level-wise)*, construyendo cada nivel por completo antes de pasar al árbol secuencial. Además, utiliza la técnica de *boosting*, donde los modelos se construyen de manera secuencial, centrándose en corregir los errores cometidos por los modelos anteriores.

LightGBM

Light Gradient Boosting Machine (LightGBM) pertenece a la familia de los algoritmos de *boosting*. LightGBM se destaca por su eficiencia y velocidad, especialmente en conjuntos de datos grandes. Su eficacia se debe en parte a la forma en que trata con la construcción de árboles durante el proceso de *boosting*.

LightGBM utiliza un enfoque de tipo *leaf-wise* para construir árboles, en lugar del enfoque

level-wise utilizado por XGBoost y por la mayoría de los algoritmos de *boosting*. Este enfoque permite construir árboles de manera más eficiente al expandir los nodos que reducen la pérdida de manera más significativa. Además, LightGBM utiliza histogramas para encontrar las mejores características y realizar divisiones en los nodos de los árboles de manera más rápida. [32]

2.4. Evaluación de Modelos

Para evaluar el rendimiento de un modelo, es indispensable definir una métrica de decisión que permita comparar y discriminar los modelos para concluir cual de ellos ofrece los mejores resultados de acuerdo con el objetivo establecido.

Matriz de Confusión

La matriz de confusión, como se observa en la Tabla 2.2, es una organización que compara los resultados predichos del modelo con los resultados reales cuando se está trabajando en un problema de clasificación de dos clases, donde una es negativa y la otra positiva.

Tabla 2.2: Matriz de confusión Modelo Detractores

		Real	
		Negativo	Positivo
Predicción	Negativo	VN	FN
	Positivo	FP	VP

De esta tabla se desprenden los siguientes campos:

1. **Verdadero Positivo (VP):** El modelo predice la observación como positiva y realmente era positiva.
2. **Falso Positivo (FP):** El modelo predice la observación como positiva y realmente era negativa.
3. **Falso Negativo (FN):** El modelo predice la observación como negativa y realmente era positiva.
4. **Verdadero Negativo (VN):** El modelo predice la observación como negativa y realmente era negativa.

Métricas de desempeño

A partir de la matriz de confusión se pueden calcular las siguientes métricas de desempeño.

Precision (Precisión): Esta métrica evalúa la proporción de predicciones positivas realizadas por el modelo que son verdaderamente positivas. Se calcula utilizando la siguiente fórmula:

$$\text{Precisión} = \frac{VP}{VP + FP} \tag{2.3}$$

Una alta precisión indica que el modelo tiene una baja tasa de falsos positivos, lo que significa que, cuando predice una observación como positiva, tiene una alta probabilidad de ser correcta.

Recall (Sensibilidad): El Recall mide la proporción de observaciones positivas reales que el modelo es capaz de identificar correctamente. Se calcula mediante la fórmula:

$$\text{Recall} = \frac{VP}{VP + FN} \tag{2.4}$$

Un Recall elevado indica que el modelo tiene una baja tasa de falsos negativos, lo que implica que es efectivo para detectar observaciones positivas en el conjunto de datos.

Accuracy (Exactitud): La exactitud evalúa la proporción de predicciones correctas en comparación con el total de observaciones. Se calcula como:

$$\text{Accuracy} = \frac{VP + VN}{VP + VN + FP + FN} \tag{2.5}$$

La exactitud proporciona una medida general del rendimiento del modelo y es especialmente útil cuando las clases están balanceadas en el conjunto de datos.

Curva ROC - AUC

La curva ROC (*Receiver Operating Characteristics*) representa la relación entre la Tasa de Falsos Positivos (t_{fp}) y la Tasa de Verdaderos Positivos (t_{vp}). De esta manera se muestra la compensación entre el beneficio de clasificar bien a la clase positiva (VP) versus el costo de equivocarse (FP). Para cada umbral de decisión en que decida clasificar a una observación como positiva se obtiene un punto en la curva con coordenadas (t_{fp}, t_{vp}). [33]

La utilidad de la curva ROC es que un clasificador aleatorio tendrá una curva igual a la función identidad, que conectará los puntos $(0, 0)$ y $(1, 1)$. De esta manera se define el Área bajo la curva ROC (AUC), lo que representa la mejora del modelo respecto al azar. Por lo tanto, un AUC de 1 es un modelo perfecto que clasifica todas las observaciones correctamente y un modelo cuyo AUC es 0.5 indica que está prediciendo un comportamiento aleatorio.

Curva Precision - Recall - AUCPR

Una de las principales desventajas del AUC es su sensibilidad ante conjuntos de datos con desequilibrio de clases. El AUC tiende a dar resultados optimistas cuando se enfrenta a clases desequilibradas. Esto significa que incluso un modelo que predice la clase minoritaria con precisión muy baja puede tener un AUC alto si la clase mayoritaria es mucho más grande.

En contraste, la curva Precision-Recall se enfoca en la capacidad del modelo para identificar con precisión ejemplos de la clase positiva, lo que la hace más adecuada para evaluar modelos en situaciones de desequilibrio de clases. La Precision-Recall puede proporcionar un resultado más realista del rendimiento del modelo cuando existe un desequilibrio entre las clases.

De manera similar al AUC, la curva Precision - Recall muestra la compensación de Precision (Ecuación 2.3) y Recall (Ecuación 2.4) a diferentes umbrales de decisión y se puede obtener el área bajo la curva (AUCPR) lo que cuantifica la mejora respecto a un modelo aleatorio.[34]

Curva Lift

La curva Lift, o análisis de ganancia, permite cuantificar cuán mejor es un modelo en comparación con la elección aleatoria de clientes. Se representa el porcentaje acumulado de la muestra en relación con la ganancia obtenida mediante el Lift.

La ganancia Lift se calcula como la proporción de observaciones de la clase positiva sobre el total de observaciones en un determinado percentil dividido por la proporción de la clase positiva en toda la muestra. Se puede expresar como: [35].

$$Lift_i = \frac{\#Positivos_i}{\#Positivos_i + \#Negativos_i} \cdot \frac{\#Positivos + \#Negativos}{\#Positivos} \tag{2.6}$$

Donde, # representa el conteo de observaciones según la categoría.

Una ganancia Lift de 1 representa ninguna ganancia respecto a una clasificación de clientes de manera aleatoria, ya que se captura la proporción de clientes de clase positiva en igual magnitud que si se hiciera de manera aleatoria. Cada Lift mayor a 1 representa una ganancia respecto a no tener un modelo de discriminación.

2.5. Selección de Variables

Un aspecto fundamental en el entrenamiento de modelos es la selección de variables ya que, una selección eficaz de variables contribuye a la mejora de la eficiencia computacional y a la interpretación del modelo, al tiempo que evita el sobreajuste (*overfitting*) al ruido presente en variables menos relevantes. Esto se traduce en modelos más precisos y robustos, capaces de generalizar de manera efectiva a datos no vistos durante el entrenamiento.

En el presente trabajo se utilizaran las técnicas de Weight of Evidence e Information Value con el objetivo de seleccionar las variables más relevantes en el modelo de satisfacción.

Weight of Evidence

El peso de la evidencia (WoE por sus siglas en inglés) es una transformación que mide la fortaleza entre una variable predictora y una variable objetivo binaria. En la Ecuación 2.7 se tiene su expresión matemática.

$$WoE = \ln \frac{\%\text{Clase Positiva}}{\%\text{Clase Negativa}} \tag{2.7}$$

Para calcular Woe, primeramente se estratifica una variable en *bins* y por cada *bin* se calcula la proporción de muestras de cada clase respecto a su propia clase, para luego aplicar

logaritmo al cociente entre la proporción de clase positiva y la proporción de la clase negativa. Asi, por ejemplo, si se obtiene un WoE alto, significa que en ese *bin* hay una mayor cantidad de observaciones de la clase positiva respecto a la negativa.

Como es de esperar, cuando se estratifica una variable en *bins* es bastante probable que en cierto *bin* queden pocas observaciones, por lo que de manera razonable sólo se considerarán los casos, para ambas clases, en que haya más de 10 observaciones en un *bin*.

Information Value

El valor de la información (IV, por sus siglas en inglés) aprovecha la utilidad de WoE para resumir en un sólo número el poder predictivo de la variable en estudio. Su fórmula se presenta en la Ecuación 2.8

$$IV = \sum_{i}^{h} (WoE_i \cdot (\,\%\text{Clase Positiva}_i - \,\%\text{Clase Negativa}_i) \tag{2.8}$$

Para interpretar IV se utilizará la regla de decisión definida en [36], por lo que se utilizarán todas las variables con un IV mayor o igual a 0.02, umbral que indica que la variable tiene algún grado de poder predictivo.

Tabla 2.3: Interpretación de Information Value [36]

Information Value	Poder de predicción
<0.02	Irrelevante para la predicción
0.02 - 0.1	Predictor débil
0.1 - 0.3	Predictor medio
0.3 - 0.5	Predictor fuerte
>0.5	Predictor Sospechoso

2.6. Explicabilidad de Modelos

Según Miller [37], la interpretabilidad se define como el grado en que un ser humano puede comprender la causa de una decisión. En este contexto, es fundamental comprender las decisiones que toma un modelo, particularmente cuando se trabaja con personas, para identificar posibles sesgos o simplemente entender qué variables permiten generar acciones de cara a los usuarios.

SHAP

En este trabajo, se adoptará el enfoque de SHAP (SHapley Additive exPlanations), uno de los métodos más ampliamente utilizados en la industria en la actualidad.

El método SHAP se basa en la teoría de juegos y proporciona explicaciones aditivas para los resultados de modelos de aprendizaje automático. Este método considera todas las posibles combinaciones de características, asignando a cada característica un valor de Shapley que cuantifica su contribución al resultado del modelo. [38]

Una característica innovadora de SHAP es la representación de la explicación del valor de Shapley como un método aditivo de atribución de características, similar a un modelo lineal. Esto aporta claridad a la interpretación de la contribución de cada característica al resultado del modelo.

Las propiedades mas relevantes de SHAP son tres:

1. Precisión local: los valores SHAP son localmente precisos, lo que significa que son precisos para una predicción concreta. Esto contrasta con las medidas de importancia de características globales, como la importancia de permutación, que son precisas para todo el conjunto de datos.

2. Missigness: Los valores SHAP son capaces de explicar las predicciones de los modelos incluso cuando faltan características. Esto se debe a que los valores SHAP se calculan utilizando una representación lineal de las características, lo que les permite tener en cuenta los valores que faltan.

3. Consistencia: La propiedad de consistencia dice que si un modelo cambia de modo que la contribución marginal del valor de una característica aumenta o se mantiene igual (independientemente de otras características), el valor de Shapley también aumenta o se mantiene igual.

2.7. Análisis Geoespacial

El análisis geoespacial tiene como objetivo describir y visualizar información que puede representarse en un contexto geoespacial. En este trabajo, se utilizarán cientos de millones de registros de clientes que reportan indicadores de red con los que se buscará identificar áreas de baja cobertura y calidad de red. A continuación, se introducirán algunos conceptos esenciales necesarios para comprender el desarrollo de la metodología.

2.7.1. Índice espacial jerárquico hexagonal de Uber: H3

Procesar millones de registros resulta complejo, sobretodo cuando se quiere visualizar geoespacialmente. Es por ello que en el campo del análisis espacial, existen diversos tipos de grillas, las cuales son fundamentales para analizar grandes conjuntos de datos y permiten rasterizar el espacio en polígonos regulares y aplicar agregaciones sobre los datos.

Una de las librerías más populares que permite hacer una rasterización es H3, la cual permite dividir el globo terrestre en hexágonos de manera eficiente y precisa. Además de H3, otras técnicas de grillas ampliamente utilizadas incluyen las grillas rectangulares. Cada tipo de grilla tiene sus propias ventajas y desventajas en términos de resolución espacial, complejidad computacional y capacidad de representar datos geoespaciales. [39]

La elección de una grilla hexagonal se fundamenta en que todos sus vecinos están a una distancia uniforme. Esto los hace más apropiados que los cuadrados, por ejemplo, que tienen vecinos a diferentes distancias.

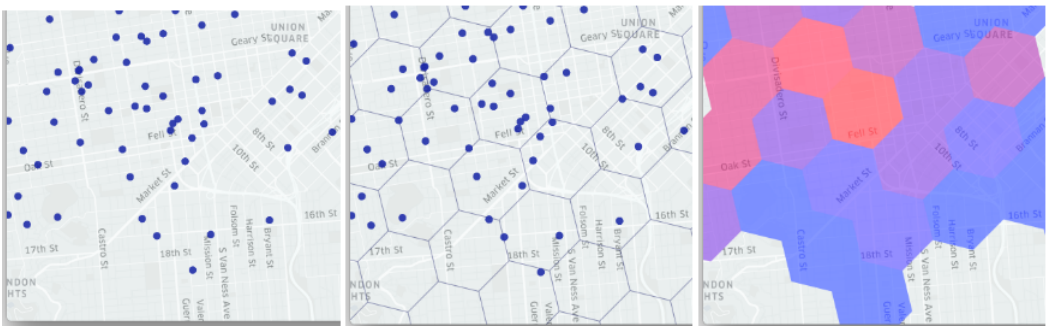


Figura 2.6: H3 Uber [39]

Parte fundamental de este trabajo consiste en el uso de H3, ya que permitirá agregar cientos de millones de datos en hexágonos, los cuales cubren una pequeña área geográfica.

2.7.2. Autocorrelación espacial

La autocorrelación espacial es una herramienta que proporciona una comprensión de cómo varía un fenómeno en un marco geográfico de análisis. Permite examinar patrones y relaciones entre los datos en función de su ubicación espacial, lo que resulta útil para entender mejor la distribución geográfica del fenómeno estudiado. Al analizar la autocorrelación espacial, se pueden identificar agrupaciones, zonas o *clusters* que permite obtener información rápida y valiosa sobre la variable de estudio. [40]

Cuando existe un patrón espacial, puede ser de agrupación (autocorrelación espacial positiva, los valores similares están próximos entre sí) o de competencia (autocorrelación espacial negativa, disimilitud entre vecinos, los valores altos repelen a otros valores altos). Dicha representación visual se puede observar en la Figura 2.7.

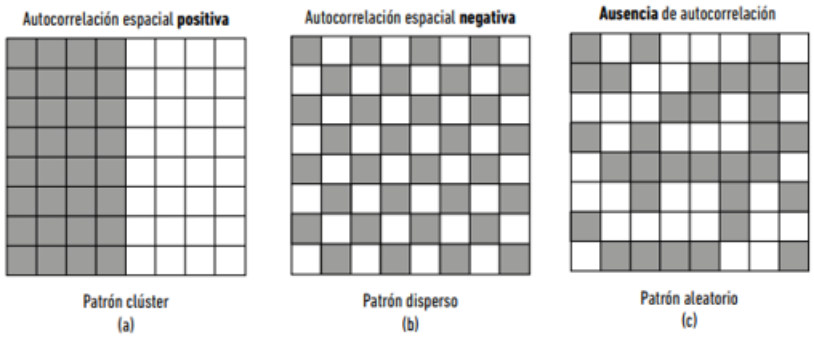


Figura 2.7: Autocorrelación espacial [40]

2.7.3. Índice de Moran

Luego de entender el concepto de autocorrelación espacial es necesario diferenciar entre dos tipos de índices para cuantificar un fenómenos espacial: índices globales e índices locales.

Los índices globales evalúan la asociación general para el conjunto de unidades analizadas, lo cual puede ser a través de la media global de la variable analizada espacialmente.

Por otro lado, Luc Anselin en 1997 presentó los Indicadores locales de asociación espacial (LISA, por sus siglas en inglés) [41] que permiten identificar zonas donde se presenta agrupamiento o dispersión de cierto fenómeno.

Sea N el número de las unidades espaciales de análisis (hexágonos, cuadrados o cualquier geometría espacial) $w_{i,j}$ la relación entre la unidad i y j, es decir, $w_{i,j}$ toma el valor de 1 si i y j son vecinos.

Considerando que se va a trabajar con hexágonos, para definir el estadístico de Moran Local se necesitan 2 componentes. Una atributo de similitud. ¿Qué tanto se parecen los hexágonos? y un atributo espacial. ¿Quienes son vecinos de cada hexágono?

Bajo la lógica de Moran, la similitud se puede obtener comparando un hexágono central versus el ponderado de sus vecinos a través de los pesos espaciales $w_{i,j}$ y la variable de interés X_i del hexágono i.

Sea I el Indice de Moran, su fórmula matemática suele escribirse como sigue:

$$I = \frac{N}{\sum_i \sum_j w_{ij}} \cdot \frac{\sum_i \sum_j w_{ij} \cdot (X_i - \bar{X}) \cdot (X_j - \bar{X})}{\sum_i (X_i - \bar{X})^2} \quad \forall i \neq j \tag{2.9}$$

cuyos dominio es $I \in [-1, 1]$.

Se puede identificar la varianza y reemplazar en la fórmula anterior,

$$I = \frac{1}{\sum_i \sum_j w_{ij}} \cdot \frac{\sum_i \sum_j w_{ij} \cdot (X_i - \bar{X}) \cdot (X_j - \bar{X})}{\sigma_X^2} \tag{2.10}$$

Así, se puede distribuir la desviación típica entre los factores del producto cruzado:

$$I = \frac{1}{\sum_i \sum_j w_{ij}} \cdot \sum_i \sum_j w_{ij} \cdot \frac{X_i - \bar{X}}{\sigma_X} \cdot \frac{X_j - \bar{X}}{\sigma_X} \tag{2.11}$$

Y, por último, reescribir la fórmula utilizando puntuaciones *z-scores*:

$$I = \frac{1}{\sum_i \sum_j w_{ij}} \cdot \sum_i \sum_j w_{ij} \cdot z_i \cdot z_j \tag{2.12}$$

Como se van a considerar unidades espaciales con geometrías de hexágonos las ponderaciones de contigüidad son de $\frac{1}{6}$ para cada vecino. Lo que significa que $w_{i,j} = \frac{1}{6}$ y cuya suma da 1.

Se puede aplicar la normalización por filas, es decir, cada vecino pesa lo mismo por lo que

se obtiene: $\sum_i \sum_j w_{ij} = N$. Así, la fórmula del Indice de Moran se convierte en:

$$I = \frac{\sum_i z_i \cdot \sum_j w_{ij} \cdot z_j}{N} \tag{2.13}$$

Para evaluar la presencia de un patrón espacial, se puede comparar el valor observado de I_i con el valor esperado bajo la hipótesis nula de aleatoriedad. El procedimiento implica fijar un hexágono i y permutar los $N-1$ hexágonos restantes K veces. Esto genera una distribución de referencia utilizada para comparar el valor observado en i con el caso aleatorio en la misma ubicación.

Posteriormente, se utiliza la distribución de referencia para calcular:

$$p = \frac{R+1}{K+1} \tag{2.14}$$

Aquí, R representa el número de veces que el índice de Moran local (I) calculado utilizando datos permutados es más extremo que el valor observado en el hexágono i, y K denota el número total de permutaciones.

El valor de p se denomina pseudo p-valor y se utiliza en contraste con un nivel de significancia α para filtrar aquellas zonas con mayor autocorrelación espacial. Esto genera un efecto de filtro sobre las áreas más interesantes.

2.7.4. Herramienta de visualización: Kepler [42]

Kepler es una herramienta de código abierta desarrollada por Uber para el análisis de datos geoespaciales que permite realizar visualizaciones de manera interactiva, personalizable y fácil teniendo los datos correctamente preprocesados.

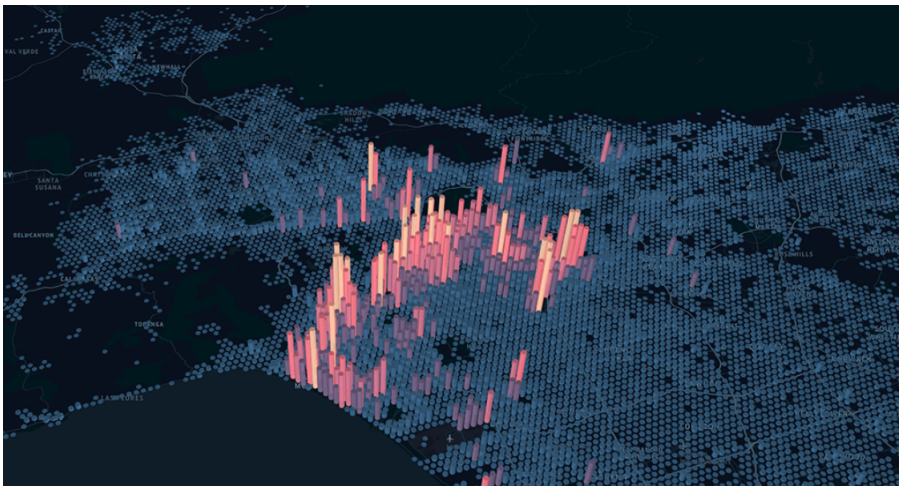


Figura 2.8: Ejemplo de visualización usando Kepler

Esta herramienta será utilizada para generar las visualizaciones que serán entregadas a la compañía de telecomunicaciones.

2.8. Metodologías: Proyecto Ciencia de Datos

Dentro del ámbito de ciencia de datos existen varias metodologías comúnmente utilizadas en desarrollos analíticos. Entre las principales metodologías se puede encontrar KDD (*Knowledge Discovery in Databases*) [43], CRISP-DM (*Cross-Industry Standard Process for Data Mining*)[44] y adquiriendo mayor relevancia en el último tiempo CRISP-ML (*Cross-Industry Standard Process for the development of Machine Learning applications*) [45].

En primer lugar, KDD proporciona un marco general para el descubrimiento de conocimiento a partir de bases de datos, sin embargo, carece de especificidades en cuanto a la ejecución de cada una de las etapas. Es más un concepto general que un conjunto específico de pasos.

En segundo lugar, CRISP-DM es una metodología ampliamente utilizada en la industria y se centra en la minería de datos, proporcionando una guía detallada para cada fase del proceso. Es conocida por su flexibilidad al permitir iteraciones entre las fases.

Por último, CRISP-ML es una extensión de CRISP-DM e incorpora dos aspectos relevantes que no son considerados por esta. Considera una metodología entre fases de aseguramiento de calidad de los datos e incluye una fase de monitorio y mantención de modelos de *Machine Learning*.

Tabla 2.4: Tabla comparativa de metodologías

KDD	CRISP-DM	CRISP-ML(Q)
	1. Comprensión del Negocio	1. Objetivo de Negocio
1. Selección de los Datos	2. Comprensión de los Datos	2. Comprensión de los Datos
2. Preparación de los Datos	3. Preparación de los Datos	3. Preparación de los Datos
3. Transformación de los Datos	4. Modelamiento	4. Modelamiento
4. Minería de Datos	5. Evaluación	5. Evaluación
5. Interpretación y Evaluación	6. Despliegue	6. Despliegue
		7. Monitoreo y Mantención

Como el enfoque principal del presente trabajo no es alcanzar una etapa de monitoreo de los modelos a proponer, las metodologías KDD y CRISP-ML no resultan precisamente adecuadas. Por el contrario, se busca lograr una comprensión profunda y fundamental del negocio, proponiendo herramientas para la toma de decisiones a través de la explotación de datos. Es por ello que se utilizará como marco de desarrollo analítico la metodología CRISP-DM.

Las principales fases que considera CRISP-DM son las siguientes:

- Comprensión del Negocio: Esta etapa busca obtener una mirada general de la empresa en la cual se realizará el trabajo. Es fundamental para el avance del proyecto entender las necesidades y dolores de la compañía en el ámbito del trabajo a realizar.
- Comprensión de los Datos: Los datos son el fundamento de la analítica, por lo que en

esta etapa se busca comprender los datos que se tendrán a disposición para poder llevar adelante el trabajo. La compañía cuenta con muchas fuentes de información, por lo que se buscará procesar muestras y entender que representan ciertas bases de datos.

- Preparación de los Datos: Esta etapa se centra en la limpieza, transformación y selección de la información con la cual se trabajará, lo cual es fundamental para las conclusiones posteriores. Dado que se manejan grandes volúmenes de información, será esencial buscar herramientas que permitan procesar los datos de manera eficiente.
- Modelamiento: Esta etapa busca la experimentación de modelos y técnicas para dar solución al problema planteado. Como se consideran dos áreas, una a nivel geoespacial y otra a nivel de cliente se definirá un modelamiento apropiado para cada caso.
- Evaluación: En esta fase se busca evaluar la solución propuesta. Para ello, se contrastará el análisis de variables de red con satisfacción.
- Despliegue: En esta última fase, se busca implementar o automatizar la solución propuesta. Para efectos del proyecto, se espera que al menos una propuesta sea implementada.

El desarrollo metodológico se abordará a través de dos enfoques. Por un lado, el desarrollo de una metodología para identificar zonas de interés de calidad y cobertura de red a nivel geoespacial procesando cientos de millones de registros. Y por otro lado, un enfoque a nivel de cliente, con el desarrollo de un modelo de dos etapas para diferenciar clientes según las categorías de detractores, neutros y promotores aprovechando distintas bases de datos con información histórica de los dispositivos móviles de los usuarios. tal como se muestra en la Figura 2.9.

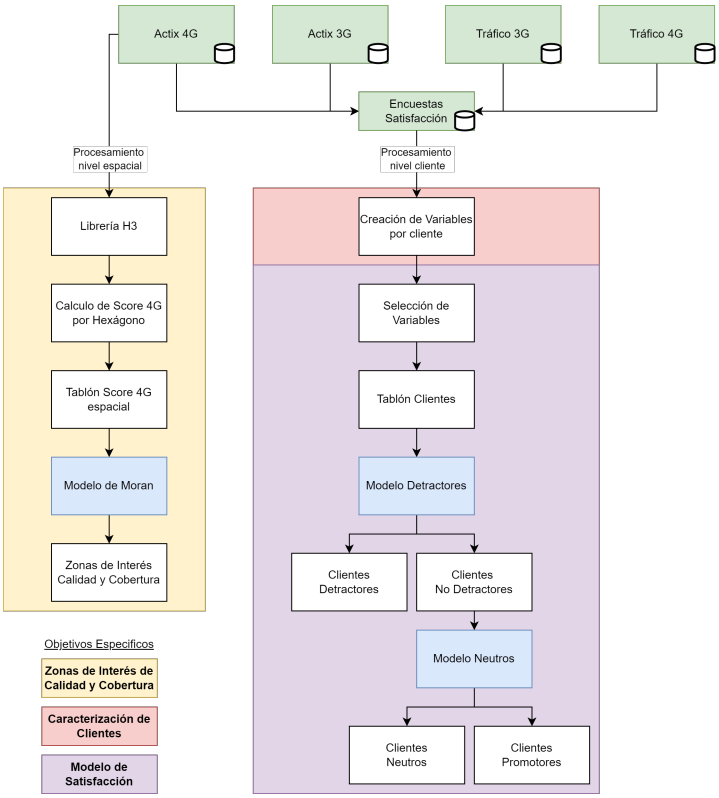


Figura 2.9: Desarrollo metodológico

Capítulo 3

Zonas de Interés de Calidad y Cobertura

3.1. Comprensión de los Datos

Durante una conexión entre un teléfono móvil y una antena de red móvil se registra información que lleva el nombre Call Data Records y se almacena en dos bases de datos. Cada una contiene información según la tecnología en que trafica el cliente.

Para el modelamiento geoespacial de zonas de interés de calidad y cobertura de red se trabajará con Actix 4G, ya que parte de la estrategia de la empresa es ir reemplazando 3G por 4G. Parte de las cuatro fuentes de información con las que se cuenta para el proyecto. La totalidad de fuentes serán usadas en el siguiente capítulo, referente al modelo de satisfacción. (Figura 3.1)

Las bases de datos Actix 3G/4G almacenan información sobre las sesiones entre los dispositivos móviles y la red dependiendo de la tecnología en que navegue. Estos registros pueden ser tanto de tráfico o llamadas. Los datos se almacenan entre las 07:00 horas y las 23:00 horas y se generan alrededor 400 millones de registros diarios a nivel de país para más de 10 millones de usuarios.

Lo valioso de esta información es que permite cubrir el país con datos de calidad y cobertura de red, siempre y cuando exista una antena a la que el cliente se haya conectado.

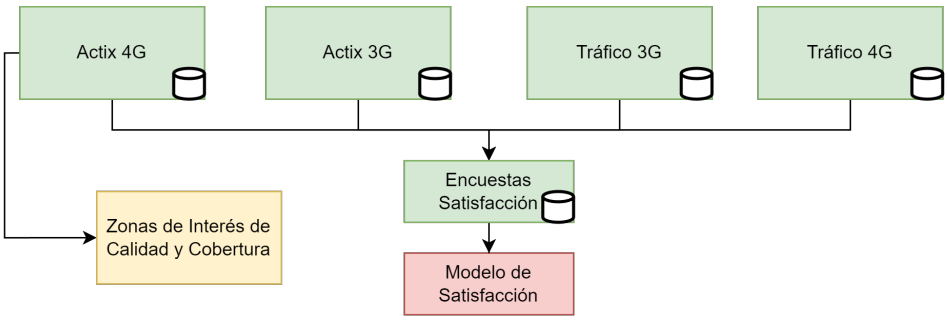


Figura 3.1: Relación fuentes de información y objetivos

Estos datos representan la huella digital de los clientes con la riqueza que presentan no sólo

una ubicación de donde se genera el dato, sino que también ciertos indicadores de calidad y cobertura de la red. Con esta información se puede mapear geoespacialmente la red móvil del operador de telecomunicaciones e identificar sólo a través de los datos zonas potencialmente deficientes en términos de calidad y cobertura de red. Las variables presentes en las bases que serán utilizadas son las siguientes.

- id cliente: Identificador único del cliente
- absolutetime: Tiempo exacto en que se genera el datos
- latitud, longitud: Coordenadas aproximadas en el momento del dato
- RSRQ: Indicador referente a la calidad de la conexión en el momento del dato
- RSRP: Indicador referente a la cobertura de la conexión en el momento del dato

La razón de su elección es porque permiten asociar geoespacialmente los indicadores de RSRQ y RSRP de los clientes. Estas variables permitirán ajustar una ventana de tiempo, contar clientes y obtener un novedoso indicador mostrado mas adelante.

Asimismo, se cuenta con datos de la calidad de llamadas, tiempos de llamadas, tráfico de volumen de datos, etc, pero no serán utilizados para el objetivo específico que se trabajará en este capitulo, sino para la creación de un modelo de satisfacción que se detallará en el Capitulo 4. La lista detallada de variables se puede consultar en la Sección A de Anexos.

De las variables a considerar en este capitulo, el id del cliente se encuentra anonimizado y se utilizará sólo para cuantificar clientes dentro de cada área como se verá más adelante. El absolutetime permitirá agrupar las ventanas temporales de información. El dato de latitud y longitud que tiene la base se obtiene mediante triangulación de antenas a través de un algoritmo específico de la plataforma que suministra dicha información. De la totalidad de datos registrados en un día, aproximadamente el 70 % están correctamente geolocalizados, lo que equivale a 280 millones de registros. (Figura 3.2)

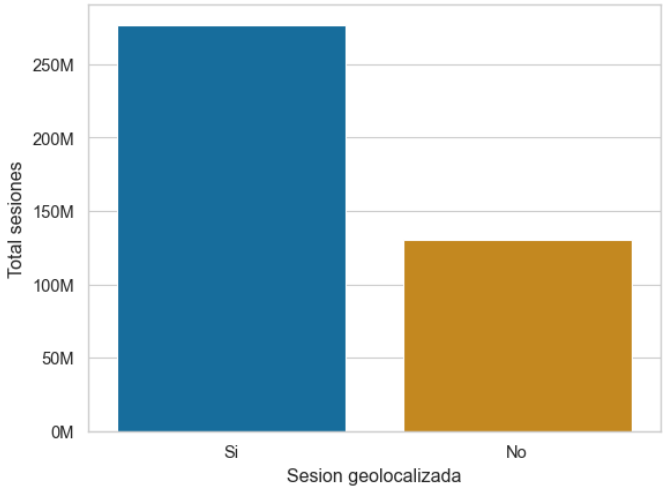


Figura 3.2: Conteo sesiones Actix 4G para el día 01-07-2023

Respecto al RSRP y RSRQ, estos indicadores serán los más relevantes para el trabajo. Ya que con ellos se creará un novedoso indicador que los relaciona y con lo que se permitirá cuantificar la calidad y cobertura de áreas geográficas. Un ejemplo del comportamiento del RSRP y RSRQ a través del tiempo se ilustra en la Figura 3.3.

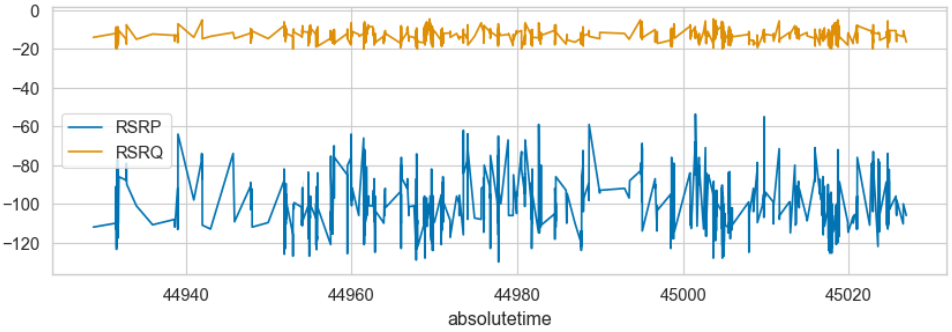


Figura 3.3: RSRP y RSRQ para un cliente en particular

En la Figura 3.3 *absolutetime* se refiere al tiempo absoluto medido en días desde el año 1900 en que se almacena el registro, es sólo referencial para visualizar la secuencia temporal del RSRP y RSRQ. Asimismo, es importante recordar que RSRP se mide en la unidad dBm y RSRQ en dB.

3.2. Preparación de los Datos

La preparación de los datos es fundamental en cualquier desarrollo analítico, sobretodo cuando se trabaja con cientos de millones de registros. A continuación se detallará el metodología definida para procesar la gran volumetría de información que almacena Actix 4G con la finalidad de identificar zonas de interés de calidad y cobertura.

El primer paso en el procesamiento de Actix 4G implica la selección exclusiva de sesiones que cuenten con geolocalización precisa, garantizando así la certeza en la ubicación donde se generan los datos.

Asimismo, los ingredientes fundamentes del trabajo se desprenden del uso del RSRP y RSRQ. A continuación, se presenta el cálculo detallado para la obtención de estos indicadores, tomando como referencia las columnas disponibles en Actix 4G.

$$RSRP = \frac{session\_av\_rsrp\_sev\_sum}{session\_av\_rsrp\_sev\_count} \tag{3.1}$$

$$RSRQ = \frac{session\_av\_rsrq\_sev\_sum}{session\_av\_rsrq\_sev\_count} \tag{3.2}$$

En este cálculo se asegura que el denominador sea distinto de 0. Los datos de RSRP y RSRQ representan la cobertura y calidad respectivamente empírica en las zonas en las que se genera el dato reportado por los móviles.

Una vez calculado el RSRP y RSRQ para cada sesión se obtiene una estructura de datos tabulares como se puede observar en la Tabla 3.1.

Tabla 3.1: Datos extraidos de Actix 4G

Dia	Latitud	Longitud	RSRP	RSRQ
2023-04-05	-33.41	-70.55	-112.5	-33.5
...	...	...	...	...

Con esta información ya es posible asignar a un punto geoespacial (x,y) los indicadores de RSRP y RSRQ. Sin embargo, se está trabajando con cientos millones de datos y se requiere la generación de áreas o zonas de interés. Así es como nace la importancia de usar una rasterización espacial hexagonal utilizando la librería H3 desarrollada por Uber.

Para ello, se trabajó con hexágonos de resolución 9, ya que no es un área ni tan grande ni tan pequeña como se observa en la Figura 3.4, su lado es en promedio de 200 metros y el área de 105.000 m^2. En Santiago Centro un hexágono de esas dimensiones representa un área aproximada de 4 manzanas.

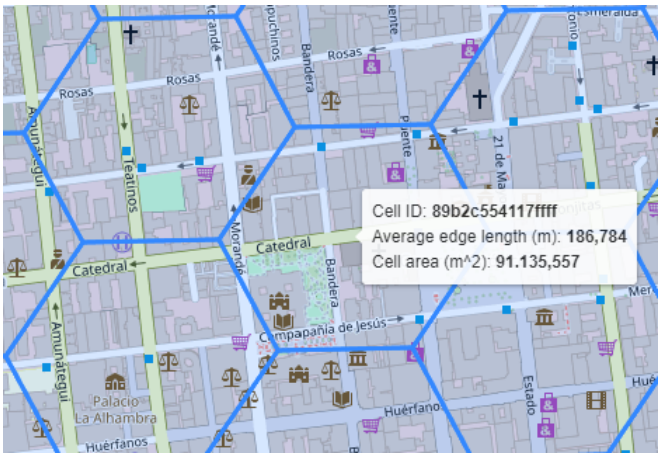


Figura 3.4: Hexágono resolución 9 H3 | Santiago Centro

Con esta división hexagonal se pueden contar las coordenadas (x,y) que caen dentro de cada hexágono y aplicar una agregación inteligente que permita transformar a un sólo indicador la calidad y cobertura de cada hexágono.

Esta agregación espacial de los datos no se centra en las antenas sino en los registros reportados por los clientes. Es por ello que se decide usar una rasterización hexagonal uniforme, que permite comparar zonas de igual área y forma.

3.2.1. Score 4G

Una etapa crucial en la preparación de los datos para el análisis geoespacial es la formulación de una variable que refleje de manera efectiva la experiencia de los clientes en términos de calidad y cobertura de la red. En este contexto, se introduce el "Score 4G" como un indicador innovador, que sintetiza de manera integral los indicadores de calidad (RSRQ) y cobertura (RSRP) en una única métrica.

El desarrollo de esta variable se basa en una serie de umbrales específicos para cada indicador, los cuales han sido cuidadosamente definidos en colaboración con expertos técnicos

en el ámbito de las redes móviles. Para el calculo del Score 4G se debe utilizar la Tabla 3.2.

Tabla 3.2: Matriz de RSRP y RSRQ

	RSRQ [dB]		
RSRP [dBm]	(, -16]	(-16, -12]	(-12,)
(-105,)	Regular	Regular Alta	Alta
(-115, -105]	Regular Baja	Regular	Regular Alta
(, -115]	Baja	Regular Baja	Regular

De esta manera, utilizando la matriz (Tabla 3.2) se cuenta para cada hexágono las sesiones que caen en cada cuadrante y luego se calcula la siguiente nota ponderada que se denomina "Score 4G". Una nota cercana a 5 indica que la mayoría de las sesiones ocurridas en el hexágono estuvieron en un rango de calidad y cobertura alto, y si la nota es cercana a 1 indica que la mayoría de sesiones en el hexágono estuvieron en rangos de calidad y cobertura bajos.

$$\text{Score 4G}_{Hex_i} = \frac{5 \cdot \#Alta + 4 \cdot \#RegularAlta + 3 \cdot \#Regular + 2 \cdot \#RegularMala + 1 \cdot \#Mala}{\#\text{Sesiones hexágono } i} \tag{3.3}$$

El símbolo # representa el conteo de sesiones en ese cuadrante/hexágono.

Un diagrama de la metodología de las zonas de interés de calidad y cobertura se muestra en la Figura 3.5. Los millones de registros de calidad (RSRQ) y cobertura (RSRP) son procesados utilizando la librería H3 en *python* con lo que se les asigna un identificador único que representa una área hexagonal. Luego, dentro de cada área se calcula el Score 4G utilizando la matriz de la Tabla 3.2. Finalmente, sólo para mostrar un ejemplo visual, se utiliza la herramienta de Kepler para plasmar el Score 4G espacialmente bajo una escala de colores.

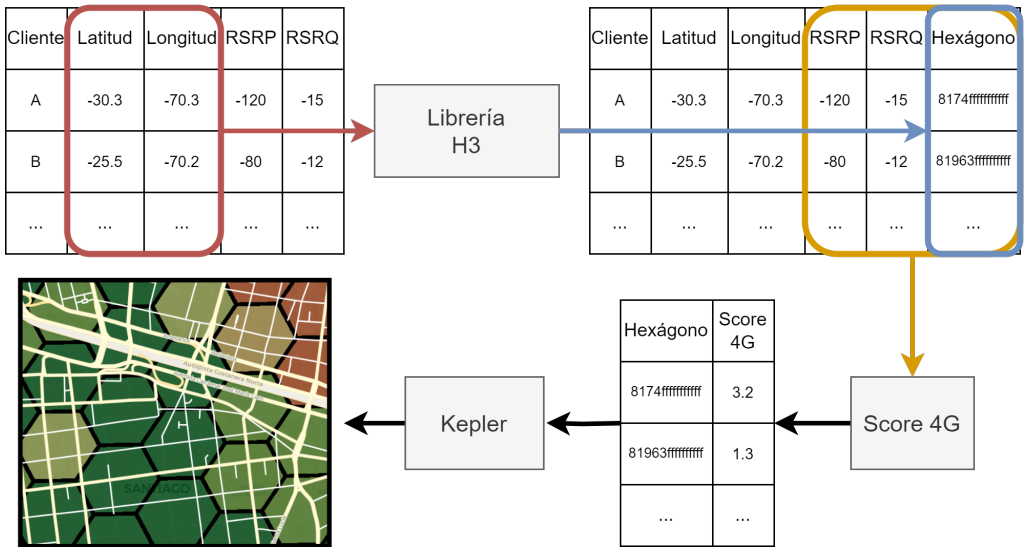


Figura 3.5: Diagrama Metodología

3.3. Modelamiento

Como se mencionó en el Capitulo 2, el Índice de Moran permite generar agrupaciones geoespaciales similares en función de una variable de interés. En ese contexto, y dado la preparación de datos realizada en la parte anterior es que se aplicará en el contexto de cobertura y calidad de red con el objetivo de identificar zonas de interés.

Para aplicar Moran, es necesario contar con la geometría o unidad de análisis, una métrica de similitud y una forma de establecer la vecindad. En este contexto, la unidad de análisis serán hexágonos que poseen un Score 4G como atributo, representando en una escala del 1 al 5 la calidad y cobertura en cada zona. La métrica de similitud definida por Moran implica comparar cada hexágono central con el promedio ponderado de sus vecinos. Utilizar hexágonos presenta la ventaja de que los 6 vecinos se identifican directamente, y se les asocia un peso igual, cada uno representando $\frac{1}{6}$.

Para aplicar Moran Local se utilizará la librería *pysal* de *python* que tiene implementado el modelo. Para ello se le debe entregar un *DataFrame* con las geometrías de los hexágonos(asociados a un único identificador) y la variable de análisis, que en este caso corresponde al Score 4G calculado en la sección anterior.

Luego de aplicar el modelo se pueden obtener 4 tipo de clusters:

- LL: Zonas con Alto Score 4G rodeadas de zonas de Alto Score 4G
- HH: Zonas con Bajo Score 4G rodeadas de zonas de Bajo Score 4G
- LH: Zonas con Alto Score 4G rodeadas de zonas de Bajo Score 4G
- HL: Zonas con Bajo Score 4G rodeadas de zonas de Alto Score 4G

La identificación de clusters Moran es valiosa, sin embargo, no todas las zonas poseen la misma relevancia para la compañía. Una forma de asignar importancia a una zona desde la perspectiva de la empresa es cuantificar el número de clientes que frecuentan dicha área. Por ende, las zonas con una mayor cantidad promedio de clientes deberían considerarse más relevantes para enfocar mejoras en la red.

Se define la densidad promedio de clientes como el producto ponderado entre los clientes únicos observados en el hexágono i y el promedio de clientes diarios durante el periodo de estudio.

$$DPC_i = \frac{\sum_d q_{i,d}}{\sum_d d} \tag{3.4}$$

Donde DPC_i representa el número promedio diario de clientes únicos presentes en el área dentro del hexágono i y $q_{i,d}$ representa el número de clientes únicos observados en el hexágono i en el día d.

De esta manera, se puede utilizar este indicador por hexágonos como medida de importancia, ya que es relevante centrarse en las áreas con la mayor densidad de clientes.

Para efectos de este estudio el análisis se realizará en un mes particular, por lo que se usará $d \in \{1, \cdots, 30\}$.

3.3.1. Área de estudio

Para aplicar Moran es necesario definir un área de análisis, la cual puede ser tan grande como se quiera. Sin embargo, dado que el foco de negocio usualmente está donde se concentra la mayoría de cliente, el área de análisis serán las zonas urbanas de la Región Metropolitana cuya área corresponde a 916 km^2, como se ilustra en la Figura 3.6. El mes de estudio abarca el mes de mayo de 2023.

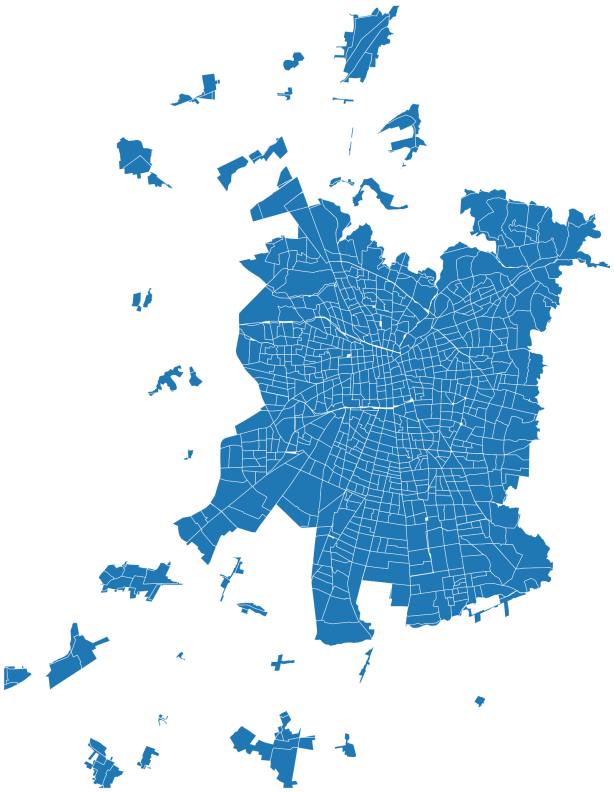


Figura 3.6: Zonas Urbanas Región Metropolitana

Cabe destacar que se trabajará con todos los datos de Actix 4G en un mes en particular, tomando una muestra del 30 % de clientes comerciales cuyas sesiones hayan sido geolocalizadas dentro del área de estudio. En términos de magnitud, esto representa alrededor de 300 millones de registros.

3.3.2. Visualización Score 4G

La variable de estudio corresponderá al Score 4G, el cual se puede visualizar geoespacialmente en la Figura 3.7. A primera vista se pueden identificar algunas zonas con una cobertura y calidad de red potencialmente inferior. Asimismo, se pueden generar visualizaciones interactivas utilizando la herramienta de Kepler como se muestra en la Sección B de Anexos.

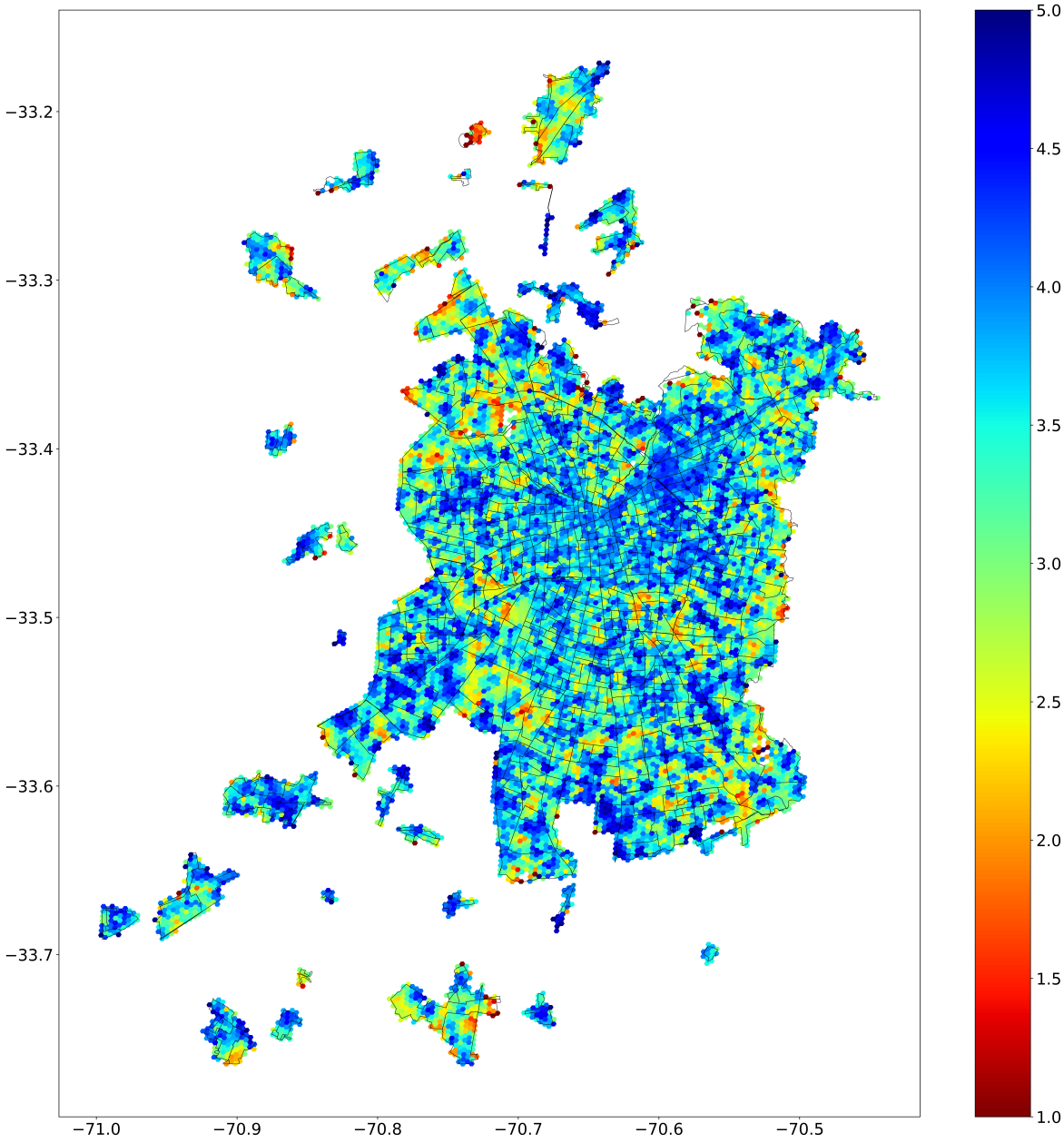


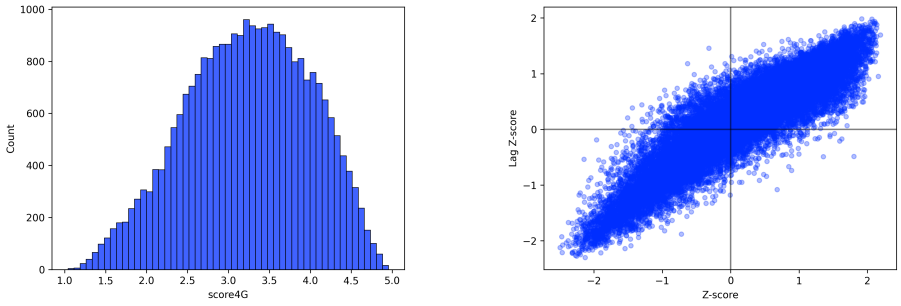
Figura 3.7: Score 4G Zonas Urbanas Región Metropolitana

3.3.3. Identificación de Zonas de Interés de Calidad y Cobertura

Para identificar las zonas de interés de calidad y cobertura se va a utilizar el Modelo del Indice De Moran. Esto permitirá generar clusters de interés en términos de la calidad y cobertura de red a nivel hexagonal en el área de estudio.

Se puede notar en la Figura 3.8.a que el Score 4G sigue una distribución prácticamente normal, lo que indica que la mayoría de las notas de calidad y cobertura se concentran entre un Score 4G de 3.0 y 3.5.

Intuitivamente Moran lo que hace es comparar el Score 4G de cada zona con el promedio de las zonas vecinas, lo que se conoce como lag espacial. Si se contrasta el Score 4G versus el Score 4G lagueado para cada hexágono se obtiene como resultado el diagrama de dispersión de la Figura 3.8.b. Desde ahí se pueden definir los 4 cuadrantes mencionados anteriormente. Zonas LL (abajo a la izquierda), Zonas HH (arriba a la derecha), Zonas HL (abajo a la derecha) y Zonas LH (arriba a la izquierda).



(a) Distribución Score 4G

(b) Score 4G Normalizado versus lag espacial Score 4G

Figura 3.8: Score 4G

Al aplicar Moran sobre el área de estudio (Figura 3.7), se obtienen las siguientes zonas coloreadas según el cuadrante correspondiente, como se muestra en la Figura 3.9.

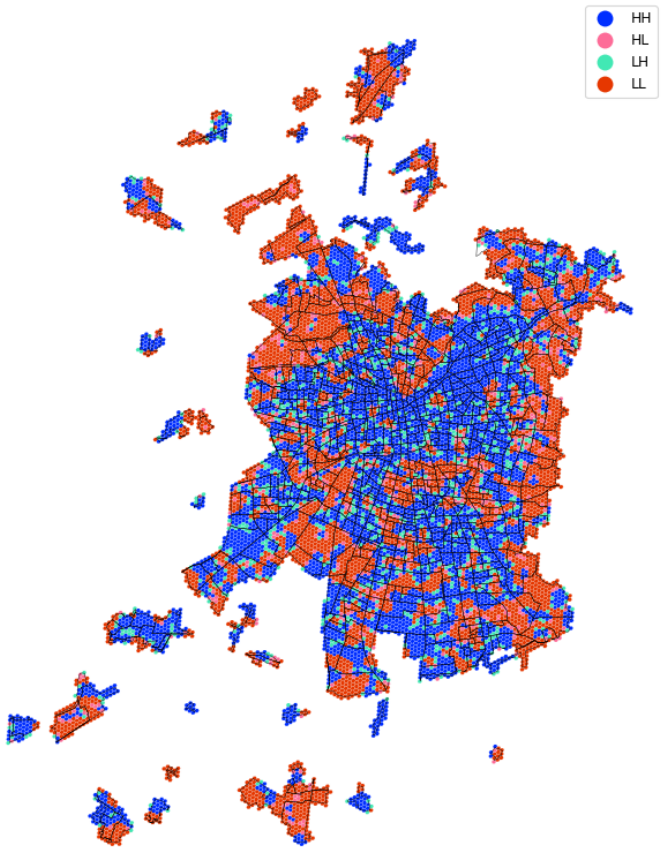


Figura 3.9: Clusters de interés Moran Local Zonas Urbanas Región Metropolitana

Si bien ya se pueden visualizar zonas con baja calidad y cobertura relativa (LL), sería útil de alguna manera filtrar las zonas más relevantes o grandes. Para ello Moran permite aplicar un nivel de significancia que genera un efecto de filtrado sobre las áreas y se concentran los cluters más relevantes.

La Figura 3.10 muestra el efecto de aplicar el parámetro de significancia a diferentes niveles. Lo interesante es que a mayor significancia se concentran las áreas más críticas donde se debería esperar peor satisfacción por parte de los clientes.

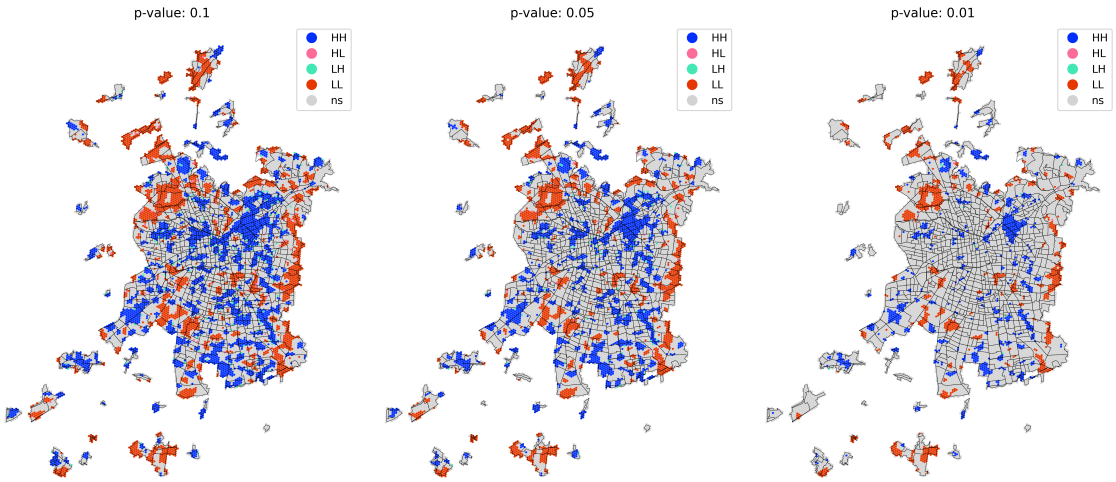


Figura 3.10: Clusters de interés Moran Local según nivel de significancia

Si bien de esta manera ya es posible identificar zonas de interés de calidad y cobertura, no todas las zonas son igual de relevantes. Por lo que se calculará la densidad promedio de clientes diarios definido en la Ecuación 3.4. El resultado se puede visualizar en escala logarítmica en la Figura 3.11, un tono de azul más intenso señala una densidad promedio de clientes más alta, mientras que un tono de rojo más intenso indica una densidad promedio de clientes más baja.

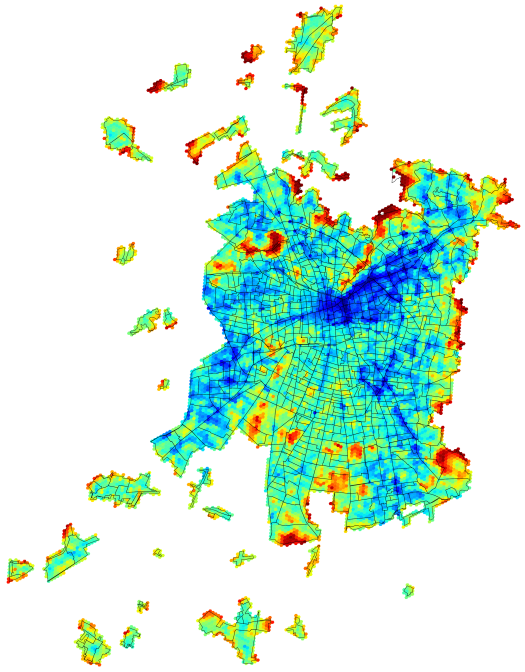


Figura 3.11: Densidad promedio de clientes en escala logarítmica

Es interesante destacar que se marcan aquellas zonas donde los clientes pasan más tiempo en promedio, lo cual corresponde en este caso al eje Alameda-Providencia. Además, se señalan otros lugares que también poseen una densidad promedio significativa de clientes.

Siguiendo con este razonamiento, se debe establecer un umbral para filtrar las zonas más densas. Para ello, se decide tomar el percentil 90 de la distribución del conteo de clientes únicos promedio por día, como se observa en la Figura 3.12.

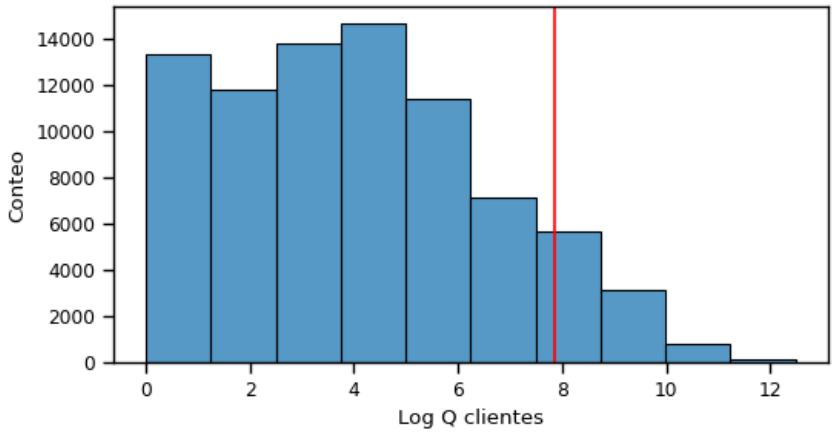


Figura 3.12: Distribución logaritmica densidad promedio diaria de cliente

El siguiente paso consiste en aplicar el filtro de densidad a las Zonas LL, las que fueron categorizadas con Moran, utilizando un nivel de significancia de 0.05. El resultado se muestra en la Figura 3.13.

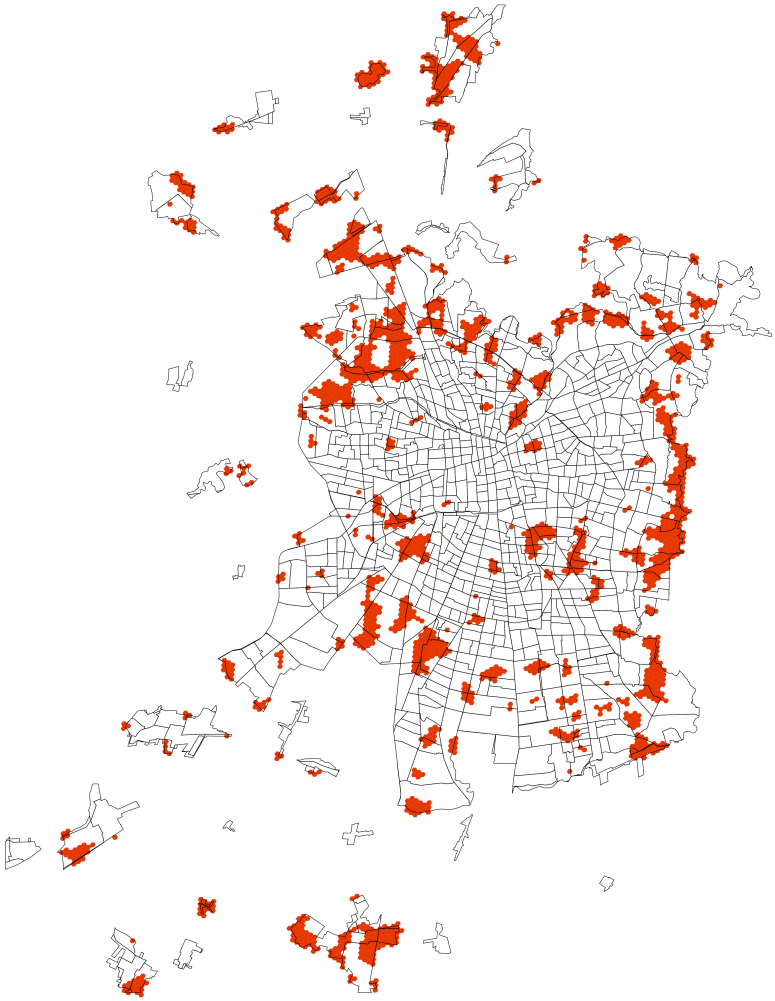


Figura 3.13: Zonas LL

Finalmente, al aplicar el filtro de densidad promedio de clientes a las Zonas LL, que según el Score 4G poseen baja calidad y cobertura móvil, se obtiene el resultado de la Figura 3.14 utilizando la herramienta de visualización de Kepler, lo que permite interactivamente revisar las zonas. Se identifican algunos clusters más grandes que otros cuya intensidad de color representa que la mayoría de las sesiones en cada área son pésimas en términos de calidad y cobertura. Destaca en el norte de la zona de estudio un lugar con bastante densidad de clientes y un bajo Score 4G, lo que podría indicar problemas de calidad y cobertura en ese sector.

Figura 3.14: Clusters de interés Moran Local Visualizados con Kepler

3.4. Análisis Score 4G

En las siguientes secciones, se explorarán dos análisis fundamentales que vinculan el Score 4G con la satisfacción del cliente, diferenciando entre enfoques a nivel individual y geoespacial. Se debe recordar que, como se estableció en el Capítulo 1, la satisfacción se mide mediante el Net Promoter Score (NPS).

Los análisis se estructuran de la siguiente manera:

- Nivel de Cliente: Este análisis involucra a 23.000 clientes, a quienes se les ha calculado el Score 4G basándose en registros históricos de RSRP y RSRQ. El objetivo es entender cómo este score individual se relaciona con su satisfacción personal.
- Nivel Geoespacial: Utilizando las mismas encuestas de los 23.000 clientes, se contrasta el Score 4G a nivel regional, que se definirá en profundidad más adelante, con la satisfacción general de la región. Es relevante mencionar que el Score 4G a nivel regional se calcula utilizando cientos de millones de registros de RSRP y RSRQ y no sólo los de los clientes encuestados como lo es el caso anterior.

3.4.1. Score 4G y Satisfacción nivel cliente

Comprender la experiencia del cliente con la red es un aspecto crucial de este trabajo. El Score 4G, definido en una escala del 1 al 5, se desarrolló para cuantificar la calidad de la cobertura tanto a nivel espacial como individual. Por lo tanto, es esencial investigar cómo este indivador se correlaciona con la satisfacción del cliente.

Se ha investigado exhaustivamente el uso del Score 4G para identificar zonas críticas en términos de calidad y cobertura de la red. Surge la interrogante de si este indicador se relaciona de alguna manera con la satisfacción del cliente. Los resultados son reveladores: el Score 4G muestra una correlación significativa con la satisfacción.

Al calcular el Score 4G a nivel individual, agrupándolo en percentiles y evaluando el NPS para cada grupo, se observan un resultado notable, como se muestra en la Figura 3.15. Estos hallazgos confirman que la cobertura y calidad de la red tienen un impacto directo y cuantificable en la experiencia del cliente.

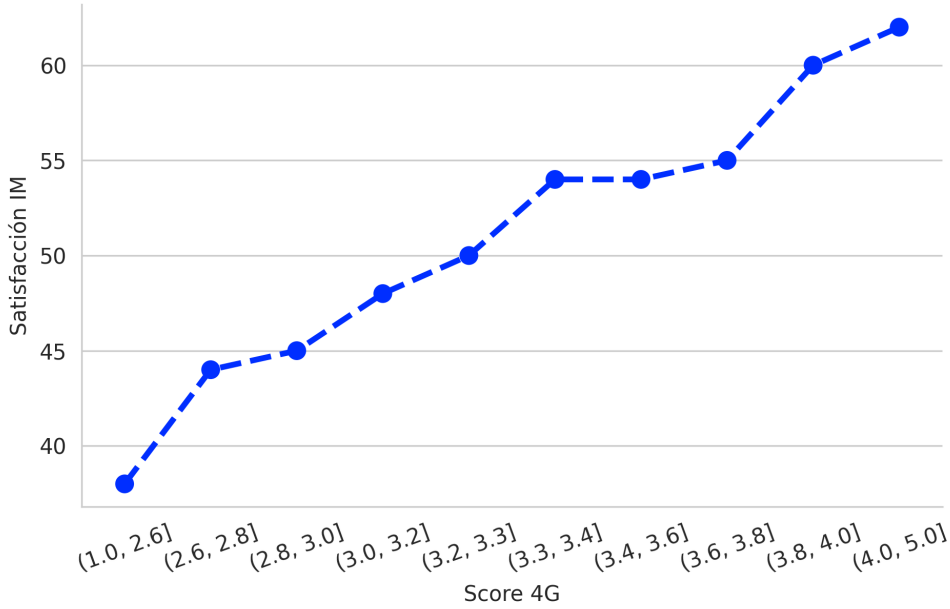


Figura 3.15: NPS versus Score 4G

El resultado es notablemente interesante, ya que evidencia que la satisfacción correlaciona fuertemente con el indicador creado. Lo que cuantifica como la cobertura y calidad influye en la experiencia de los clientes.

Ahora bien, La relación entre el Score 4G y la satisfacción se puede modelar mediante una regresión lineal:

$$NPS_{IM} = \beta_0 + \beta_1 \cdot Score4G \tag{3.5}$$

$$\Delta Score4G = \frac{1}{\beta_1} \Delta NPS_{IM} = 10.07 \cdot \Delta NPS_{IM} \tag{3.6}$$

Este modelo arroja un $\beta_0 = 0.185$ y $\beta_1 = 0.099$, con un $R^2 = 0.9469$, sugiriendo que un incremento de un punto en el Score 4G a nivel de grupo de clientes podría resultar en un aumento esperado de 10 puntos en la satisfacción. Esto implica que mejoras en la cobertura pueden tener un efecto significativo y positivo en la satisfacción del cliente, como se demuestra tanto lógica como empíricamente.

3.4.2. Score 4G y Satisfacción nivel regional

Profundizando aún más, se aprovecha el vasto conjunto de datos disponibles (cubriendo el país completo) para calcular un Score 4G a nivel regional. Siguiendo la metodología utilizada para identificar zonas de interés y abarcando todo el país como área de estudio, se procesan los datos de Actix 4G correspondientes a los últimos tres meses, lo que representa más de 1,000 millones de registros.

En este enfoque, se emplea una metodología que asigna mayor importancia a ciertos hexágonos sobre otros, similar al enfoque empleado para la densidad promedio de clientes. Debido a la magnitud del volumen de datos, se opta por calcular un valor ponderado que combina el Score 4G con el conteo de clientes únicos en cada hexágono, en lugar de un promedio diario de clientes, para optimizar la capacidad computacional. Así, el Score 4G a nivel regional se define como:

$$Score4G_{Region} = \frac{\sum_i Score4G_{Hex_i} \cdot Q_{Hex_i}}{\sum_i Q_{Hex_i}} \tag{3.7}$$

Aquí, $Score4G_{Hex_i}$ representa el Score 4G del i-ésimo hexágono en la región de estudio, y Q_{Hex_i} denota el número de clientes únicos observados en dicho hexágono durante el mes de análisis.

Los resultados se presentan en la Figura 3.16, evidenciando una correlación notable entre el Score 4G y la satisfacción a nivel regional. Es importante señalar que en esta análisis se consideran solo las regiones con más de 50 encuestados por mes, excluyendo así casos como la Región de Magallanes y la Región de Aysén por su menor cantidad de datos.

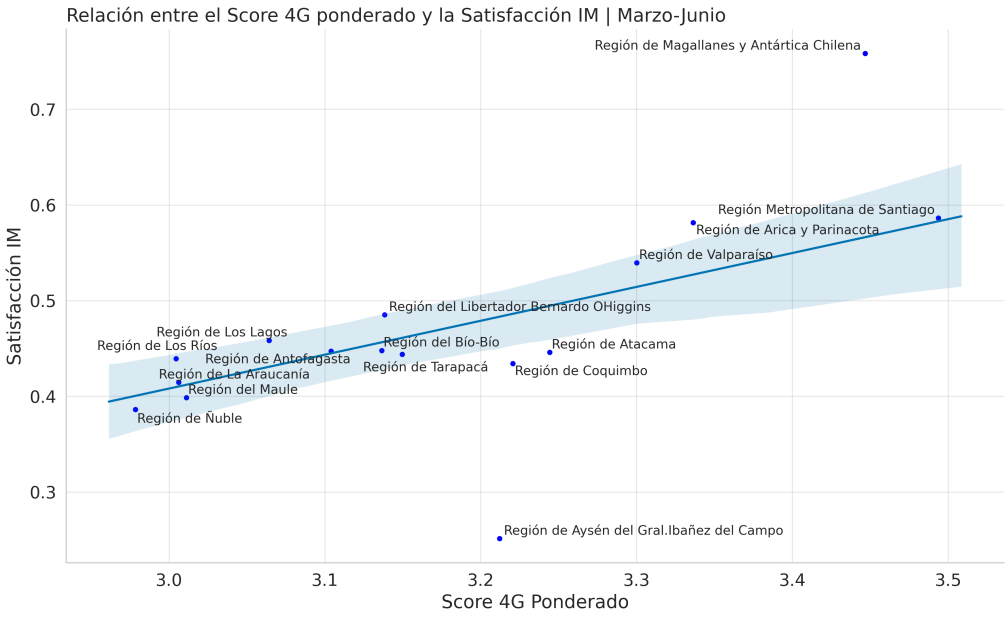


Figura 3.16: Relación Score 4G Ponderado y Satisfacción Promedio Internet Móvil Marzo a Junio

La correlación de Pearson entre el Score 4G y el NPS a nivel regional alcanza un valor de 0.9, manteniendo así la correlación a nivel global y reafirmando la importancia de la calidad y cobertura de la red en la satisfacción del cliente a través de distintas escalas geográficas.

3.5. Comentarios Finales

La utilización de cientos de millones de registros con indicadores de calidad y cobertura de red almacenados en la base de datos Actix 4G ha permitido identificar áreas geográficas con potenciales deficiencias de servicio sin la necesidad de enviar gente del área operativa a terreno.

La creación del Score 4G, que combina los indicadores RSRP y RSRQ, ha sido un avance significativo. Este indicador ha permitido cuantificar la calidad y cobertura en términos que son relevantes y comprensibles tanto para técnicos como para responsables de la toma de decisiones.

Los resultados demuestran una correlación notable entre el Score 4G y el NPS tanto a nivel individual como regional. Este hallazgo es fundamental, ya que subraya la importancia directa de la calidad y cobertura de la red en la percepción y satisfacción del cliente. Un aumento en el Score 4G se asocia consistentemente con una mejora en el NPS, lo que indica que los esfuerzos enfocados en mejorar la calidad de la red pueden tener un impacto significativo en la experiencia general del cliente.

La utilización de herramientas de visualización como Kepler (Véase Sección B de Anexo) ha permitido no solo una representación gráfica intuitiva de los datos, sino también la posibilidad de realizar análisis interactivos, facilitando así la interpretación y comprensión de los resultados.

En general, este capitulo ha mostrado como desde el dato bruto se puede llegar a obtener herramientas poderosas con las que se puede acerca la comprensión de la experiencia del cliente al asociar distintos indicadores que usualmente sólo se observan en las áreas técnicas de redes. Estos resultados son una guía hacia un enfoque más centrado en el usuario.

Capítulo 4

Modelo de Satisfacción

En el capítulo dedicado al Modelo de Satisfacción, se realizará un tratamiento exhaustivo de los datos enfocado específicamente en comprender la relación entre la satisfacción del cliente y diversas variables de telecomunicaciones.

4.1. Comprensión de los datos

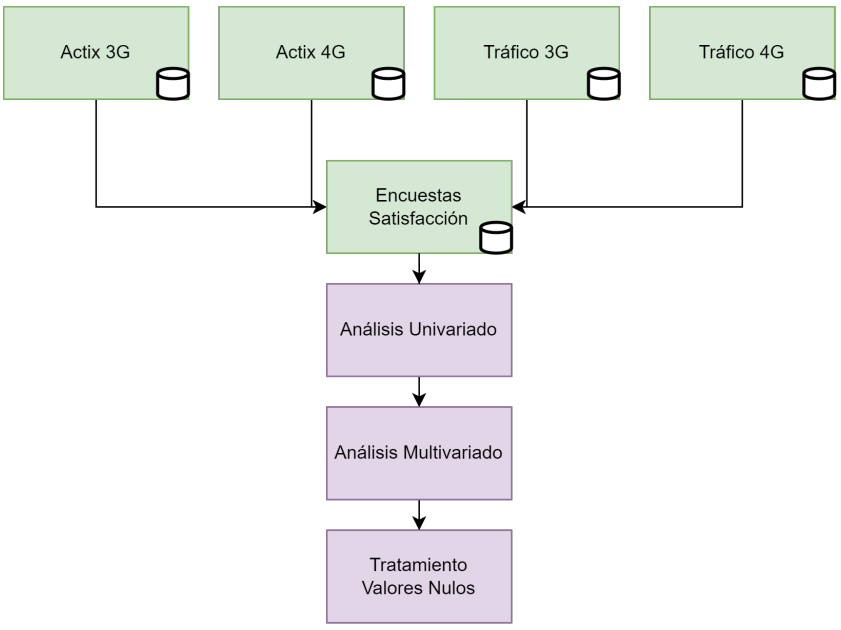


Figura 4.1: Comprensión de los Datos

El tratamiento de datos para el modelo de satisfacción considerará como base el universo de clientes presentes en la encuesta de satisfacción. La lógica de extracción de los datos consiste en recopilar información desde Actix 3G, Actix 4G, Tráfico 3G y Tráfico 4G para cada cliente 30 días previos al día en que es encuestado, como se observa en la Figura 4.2. Posteriormente, se realizará un análisis univariado, multivariado y de valores nulos para asegurar calidad en la información. (Figura 4.1)

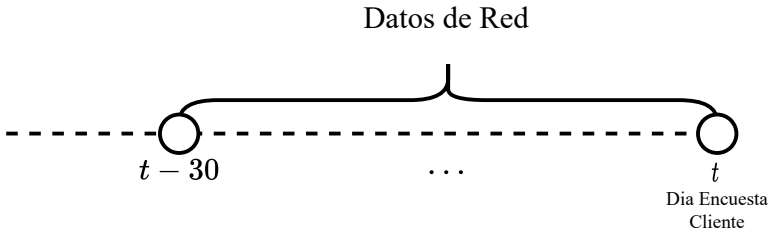


Figura 4.2: Esquema obtención datos para cada cliente

La muestra comprende un total de 23.246 clientes con planes móviles(pago mensual), quienes fueron encuestados entre noviembre de 2022 y junio de 2023. Para cada uno se cuenta con la nota de internet móvil y la fecha exacta en que fue encuestado.

Los clientes son categorizados en detractores, neutros y promotores dado su nota de satisfacción en donde responden a la pregunta. "¿Qué tan satisfecho se encuentra con el servicio de internet móvil en una escala de 1 a 7".

La proporción de clientes por categoría se muestran en la Tabla 4.1.

Tabla 4.1: Distribución porcentual tipo cliente 2022-10 al 2023-06

Tipo Cliente	Porcentaje
Promotor	65.6
Neutro	20.5
Detractor	13.8

4.1.1. Análisis Univariado

Es fundamental en la ciencia de datos comprender y explorar en profundidad los datos disponibles. Por lo tanto, en esta sección, se detallará el análisis de las variables de clientes Actix 4G. La lista completa de variables disponibles por cliente se puede consultar en la Sección A de Anexos. Asimismo, el detalle del análisis univariado para Actix 3G se puede encontrar en la Sección C de Anexos.

Actix 4G

Actix 4G, al igual que su contraparte 3G, registra datos de sesiones que abarcan tanto tráfico de datos como llamadas. En esta sección, se llevará a cabo un análisis de las variables disponibles a nivel de cliente en Actix 4G. Es importante destacar que las variables en Actix 4G pueden diferir de las de Actix 3G debido a las diferencias en las tecnologías.

Conteo Sesiones

Se realiza un análisis del conteo de sesiones en 4G, similar al efectuado para 3G. Para una mejor interpretación y manejo de la variabilidad en los datos, se presenta el conteo en una escala logarítmica, como se muestra en la Figura 4.3. Esta representación permite apreciar más claramente la distribución y frecuencia de las sesiones entre los clientes.

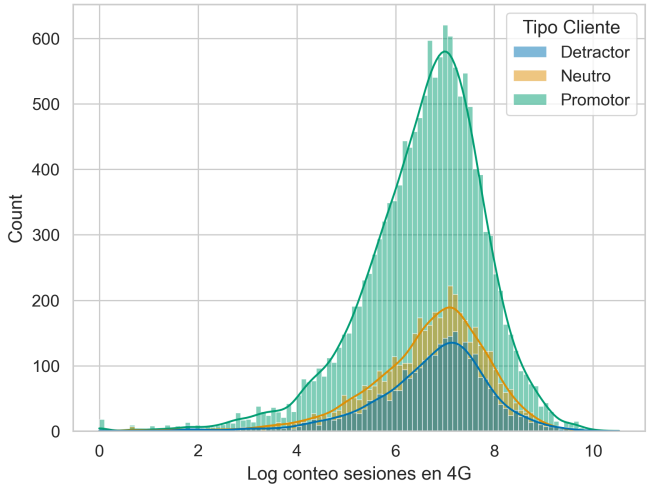


Figura 4.3: Log conteo sesiones en tecnología 4G

A primera vista, no se perciben diferencias significativas en las distribuciones del conteo de sesiones entre los distintos tipos de clientes.

Sesiones Geolocalizadas

Al calcular el porcentaje de sesiones geolocalizadas por cliente en la red 4G, se obtiene la distribución mostrada en la Figura 4.4. Esta distribución proporciona una visión detallada de la frecuencia con la que las sesiones de los clientes son geolocalizadas con éxito en la tecnología 4G.

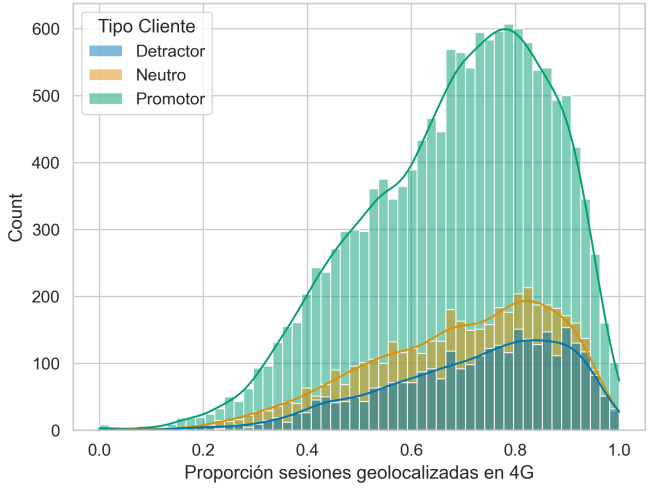


Figura 4.4: Histograma porcentaje sesiones geolocalizadas en tecnología 4G por tipo de cliente

Los resultados indican que una mayoría sustancial de los clientes tiene más del 60 % de sus sesiones correctamente geolocalizadas en 4G.

Llamadas en 4G

En el análisis de las llamadas en la red 4G, la columna *csduration time* es fundamental, ya que indica la duración de cada llamada en milisegundos. A continuación, se presenta un análisis descriptivo de esta variable:

Tabla 4.2: Análisis descriptivo variable csduration time [ms]

Conteo	251921
Promedio	25554
Desv. Est.	55917
Mínimo	0
25 %	4188
50 %	9270
75 %	25761
Máximo	3598329

De un total de sesiones, 251.921 se identifican como llamadas, con la llamada más larga registrada durando 3.598.329 milisegundos, equivalente aproximadamente a una hora.

El siguiente paso es analizar qué porcentaje de las sesiones en 4G corresponden a llamadas de voz a través de la red móvil, es decir, aquellas sesiones con un valor distinto de nulo en *csduration time_cs*.

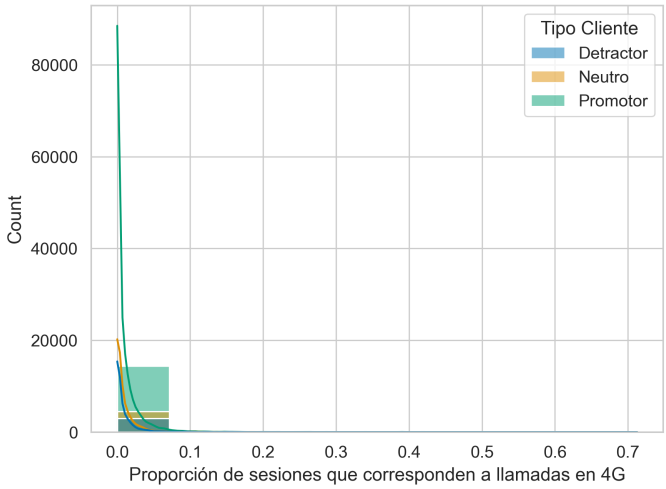


Figura 4.5: Histograma proporción de sesiones que corresponden a llamadas en tecnología 4G por tipo de cliente

La Figura 4.5 muestra que, para la mayoría de los clientes en 4G, las sesiones que corresponden a llamadas representan menos del 10 % del total de sus registros.

Un aspecto adicional de interés es el promedio de llamadas realizadas durante diferentes horas del día, segmentado por tipo de cliente. La Figura 4.6 detalla esta información, mos-

trando el promedio de llamadas en la red 4G según la hora del día y el tipo de cliente.

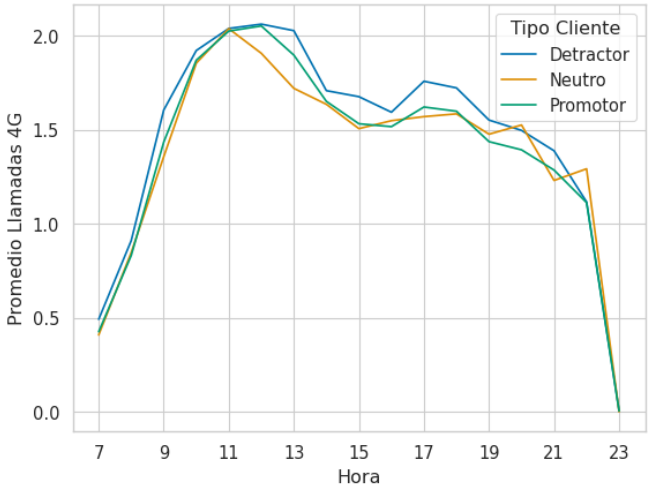


Figura 4.6: Promedio llamadas en 4G hora del día por tipo de cliente

Se observa que el máximo promedio de llamadas en 4G ocurre alrededor del mediodía. Sin embargo, al igual que en 3G (Sección C de Anexos), no se identifican patrones distintivos que diferencien a los clientes según el volumen de sus llamadas.

Estado conexión

En la red 4G, se disponen de variables similares a las de 3G, como *connectiondropped_cs* y *connectionfailed_cs*, que indican el estado de las llamadas. No obstante, en este caso, la cantidad de registros no nulos en estas columnas es mínima, con solo 18 sesiones para *connectiondropped_cs* y 2 para *connectionfailed_cs*. Debido a esta limitada cantidad de datos, se decide no proceder con un análisis detallado de estas variables en el contexto de 4G

Tráfico 4G

En lo que respecta al tráfico en la red 4G se aplica una transformación logarítmica para manejar adecuadamente la variabilidad en los volúmenes de tráfico entre diferentes clientes. Esta transformación ayuda a normalizar los datos y facilitar su interpretación.

La Figura 4.7 muestra la distribución del tráfico 4G por tipo de cliente, a través de un histograma que permite visualizar de manera efectiva las diferencias en el uso de datos.

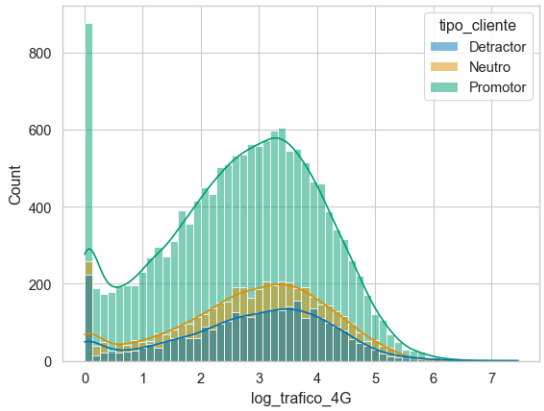


Figura 4.7: Histograma del logaritmo del tráfico en bytes de 4G por tipo de cliente

Al analizar la distribución del tráfico 4G, no se observan diferencias significativas en el volumen de datos utilizados entre los distintos tipos de clientes. Esto sugiere que el patrón de uso de datos en 4G es relativamente homogéneo entre los clientes, independientemente de su nivel de satisfacción.

4.1.2. Análisis Multivariado

El análisis multivariado se enfoca en explorar las interacciones y correlaciones entre varias variables dentro de un conjunto de datos. Este tipo de análisis es crucial para comprender cómo las diferentes variables se influyen mutuamente y para identificar patrones subyacentes. En este contexto, el objetivo es investigar las relaciones entre varias variables clave relacionadas con la cobertura y la calidad de la red, utilizando los datos de Actix 3G y Actix 4G. Además, se realizará un contraste entre el tráfico en 3G y 4G para proporcionar una comprensión más profunda de estos aspectos en diferentes tecnologías. El detalle del análisis multivariado para Actix 3G se puede encontrar en la Sección D de Anexos.

Actix 4G

Se analizan los indicadores RSRP y RSRQ para evaluar la cobertura y calidad. Al comparar estos dos indicadores, como se muestra en la Figura 4.8, se observa una tendencia similar a la encontrada en 3G. Los clientes clasificados como detractores tienden a agruparse en niveles más bajos de RSRP, lo que podría indicar que estos clientes están más frecuentemente en áreas con cobertura más débil o problemática.

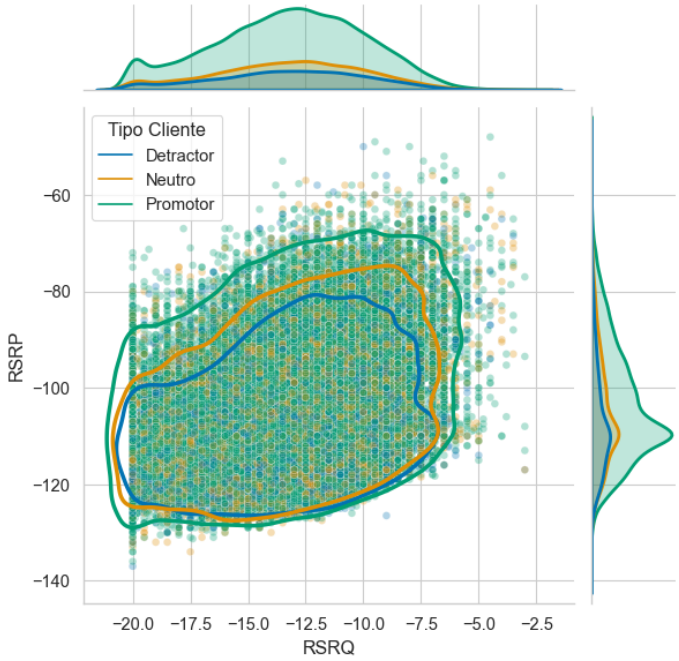


Figura 4.8: Scatterplot RSRP vs RSRQ con KDE

Tráfico 3G y 4G

Un aspecto relevante de la experiencia es el tráfico en cada tecnología dado que cada una tiene limitantes de velocidad y capacidad. A continuación se compara para cada cliente el tráfico en 3G versus el tráfico en 4G por tipo de cliente. (Figura 4.9)

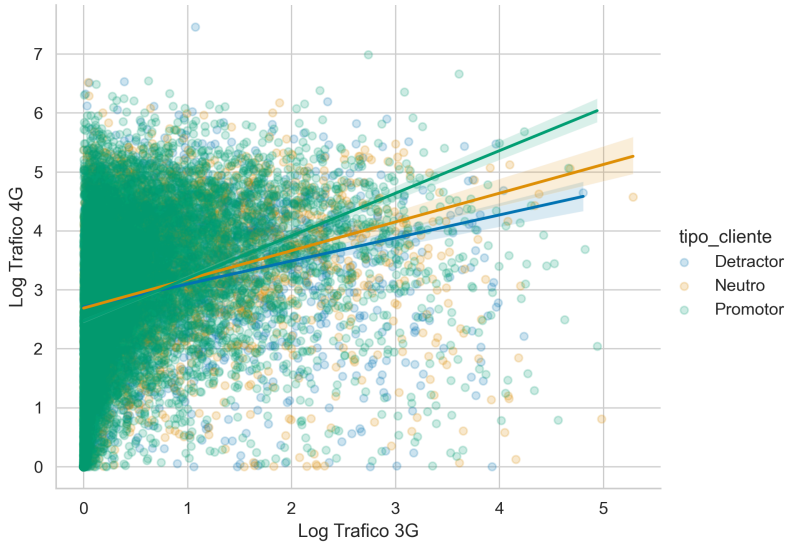


Figura 4.9: Logaritmo del Tráfico 3G en bytes vs Logaritmo del Tráfico 4G en bytes por tipo de cliente

Se puede notar que los clientes detractores presentan una menor sensibilidad de su tráfico 3G versus el tráfico 4G respecto a los clientes promotores. Esto sugiere que los clientes detractores tienden a navegar más en la red 3G que la 4G.

4.1.3. Tratamiento de Valores Nulos

El tratamiento de valores nulos es un proceso fundamental para entender la presencia de valores faltantes. Como se está trabajando con datos masivos a nivel de cliente de red, es importante entender la falta de ciertos campos en la base de datos de estudio, ya que un valor nulo podría ser indicio de algún problema referente al cliente que afecte su experiencia.

Para un análisis en profundidad de los datos nulos se utilizará la librería de *python* llamada *missingno*. Se usará la función del mapa de calor sobre los datos.

El mapa de correlación *missingno* mide la correlación de nulidad: la intensidad con que la presencia o ausencia de una variable afecta a la presencia de otra:

- Los valores cercanos a 1 indican que la presencia de valores nulos en una columna está correlacionada con la presencia de valores nulos en otra columna.
- Los valores cercanos a -1 indican que la presencia de valores nulos en una columna está inversamente correlacionada con la presencia de valores nulos en otra columna. En otras palabras, cuando los valores nulos están presentes en una columna, hay valores de datos presentes en la otra columna, y viceversa.
- Los valores cercanos a 0, indican que hay poca o ninguna relación entre la presencia de valores nulos en una columna en comparación con otra.

El análisis de valores nulos en Actix 3G se puede encontrar en la sección E del Anexo.

Actix 4G

Se procede a calcular el porcentaje de valores nulos por columna para la base Actix 4G.

Tabla 4.3: Variables con valores nulos en Actix 4G

Variable	Porcentaje Valores Nulos
connectionfailed_cs	99.999
connectiondropped_cs	99.997
connectionok_cs	99.155
csdurationtime	99.152
rsrp	40.009
rsrq	40.009
endlat	4.874
endlon	4.874

Se puede observar en la Tabla 4.3 la presencia de muchos valores de gran cantidad de valores nulos en las variables *connectionfailed_cs*, *connectiondropped_cs*, *connectionok_cs* y *csdurationtime_cs*. Cuya razón, según el entendimiento de los datos se debe a que existen registros que son de llamadas y estas variables son no nulas en esos casos. Como la mayoría de las sesiones son de tráficos gran parte de estas sesiones quedan como nulas.

Asimismo, es interesante notar que el alrededor 40 % de las sesiones tienen nulos sobre las variables de RSRP y RSRQ, que hacen referencia a la cobertura y calidad de la sesión.

Por último, sólo alrededor del 5 % de las sesiones no tienen una ubicación determinada.

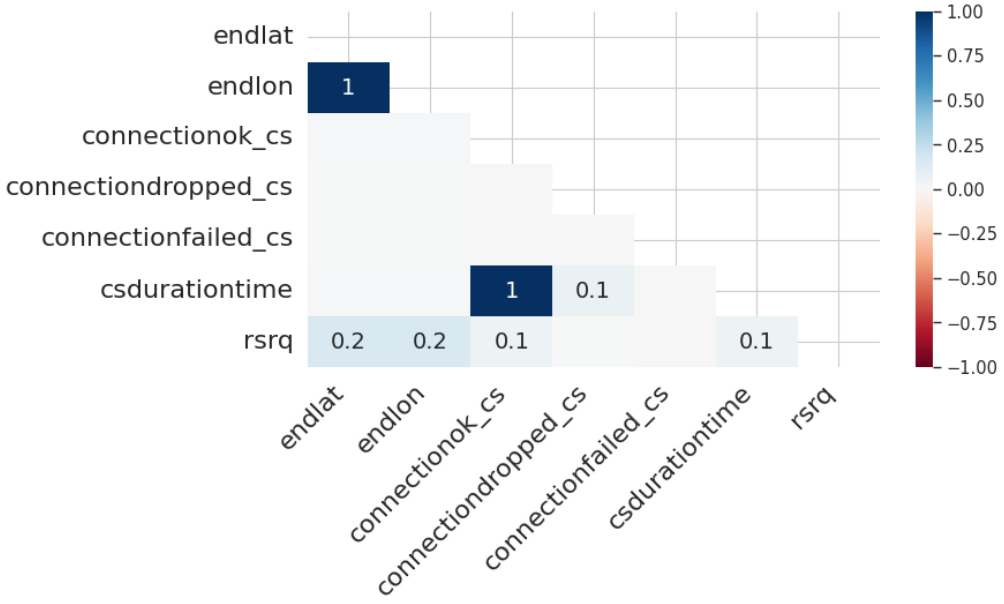


Figura 4.10: Correlación Valores Nulos Actix 4G

Se puede observar en la Figura 4.10 que la presencia de un valor nulo en la variable *endlat* implica un valor nulo sobre *endlon*, ya que la correlación es 1. Esto indica que hay sesiones en las cuales no se logra establecer la ubicación del cliente, lo que tiene relación por definición con la columna *geolocationflag* que toma el valor de 0 cuando no se logran establecer las coordenadas geográficas de la sesión.

Asimismo, ocurre lo mismo con la variable *csdurationtime* sobre la variable *connectionok_cs*. Esto se debe a que Actix almacena sesiones que pueden ser tanto de tráfico como de llamadas. Como se está anlizando Actix 4G, esas columnas son nulas en los casos en que la sesión corresponde a tráfico.

4.2. Preparación de los Datos

La preparación de los datos implica el proceso de seleccionar, transformar y crear características relevantes a partir de datos de entrada, con el objetivo de construir modelos de aprendizaje automático precisos y eficientes.

Muchos de los algoritmos de aprendizaje automático poseen un mejor desempeño cuando los valores de las variables se transforman a un valor fácil de interpretar por los modelos. Por otro lado, datos sucios pueden entorpecer las predicciones generadas por nuestros modelos, al igual que las escalas en que se presentan los datos. Por esto, es relevante que las variables que se utilizan se encuentren en escalas similares y con distribuciones relativamente similares a la distribución normal. [46]

4.2.1. Caracterización de Clientes

Un objetivo clave dentro del modelo de satisfacción es la caracterización de clientes en función de datos de red que permitan diferenciar a los tipos de clientes. Según se exploró en la Sección Comprensión de los Datos hay ciertos indicadores que dan a luz diferencias entre tipos de clientes, como por ejemplo, el número de sesiones y tráfico en 3G.

Para lograr una caracterización efectiva de los clientes, se procederá a crear agregaciones de la información previamente mencionada, enfocándose en los siguientes aspectos:

1. Llamadas: Agregaciones diversas relacionadas con llamadas para 3G y 4G.
2. Tráfico: Cálculo del tráfico total del cliente para 3G y 4G.
3. Calidad y Cobertura: Estimación de un indicador de calidad y cobertura para 3G y 4G.
4. Movilidad: Inclusión de una métrica que cuantifique la movilidad del cliente.

Para llamadas y tráfico, se realizarán agregaciones simples basadas en los datos de los clientes.

Por el lado de calidad y cobertura, se calculará un Score 4G a nivel de cliente basado en sus sesiones, utilizando la información de la Tabla 3.2. Además, se incluirá el cálculo de un Score 3G, siguiendo la misma lógica del Score 4G pero ajustando los umbrales definidos en la Tabla 3.2. En este caso, se emplearán umbrales específicos para RSCP (equivalente a RSRP) establecidos en -101 y -95, mientras que para ECNO (equivalente a RSRQ) se utilizarán los umbrales de -15 y -11. De esta manera, se asignará una puntuación tanto para la cobertura en red 4G como en red 3G.

$$\text{Score } (3\text{G})4\text{G}_{C_i} = \frac{5 \cdot \#Alta + 4 \cdot \#RegularAlta + 3 \cdot \#Regular + 2 \cdot \#RegularMala + 1 \cdot \#Mala}{\#\text{Sesiones cliente } i} \tag{4.1}$$

El símbolo # representa el conteo de sesiones en cada cuadrante para el cliente i.

El último aspecto es movilidad. Por lo que, se creará una variable que cuantifique la movilidad del cliente en función de sus posiciones geolocalizadas. Para ello, se calculará el radio de giro r_g^a para el cliente a, definido en [47] y utilizado en contextos similares con datos de telefonía [48]. La fórmula para el radio de giro es la siguiente:

$$r_g^a = \sqrt{\frac{1}{n_c^a} \sum_{i=1}^{n_c^a} (\vec{r_i^a} - \vec{r_{cm}^a})} \tag{4.2}$$

Donde, n_c^a representa el total de posiciones distintas del usuario a y r_{cm}^a representa el centro de masa de las posiciones del cliente a. De esta manera se puede cuantificar qué tanto se mueve un cliente respecto a su centro de masa de sus posiciones geolocalizadas.

La Tabla 4.4 resume las variables que se crearán para cada cliente, abarcando aspectos de llamadas, tráfico, calidad, cobertura y movilidad. Esta compilación proporcionará una vista

integral del comportamiento y experiencia de los usuarios en la red. La lista detallada de variables se puede encontrar en la Sección F de Anexos.

Tabla 4.4: Variables modelo de satisfacción

Variable	Descripción
Tráfico en 3G/4G	Total tráfico en gigabytes por tecnología
Score Cobertura 3G/4G	Nota de Cobertura entre 1 a 5
Minutos llamadas 3G/4G	Total de minutos en llamadas
Nº Sesiones 3G/4G	Total de registros en Actix
Nº llamadas caidas/fallidas 3G/4G	Total de llamadas caídas o fallidas
Nºconexiones indoor/outdoor 3G	Total de sesiones indoor o outdoor
Tiempo en establecer llamada 3G	Tiempo promedio de RRC
Nº días sin datos 3G/4G	Total de días sin datos
% registros nulos 3G/4G	Porcentaje de sesiones cuyo valor de ECNO/RSCP ó RSRP/RSRQ es nulo
Nº cambios de 4G->3G	Total de veces que se experimentó un cambio de 4G a 3G
Radio de Giro	Radio de Giro definido en la ecuación 4.2

Esta tabla proporciona una base sólida la construcción del modelo de satisfacción. La inclusión de un amplio espectro de variables garantiza una comprensión detallada de la experiencia y el comportamiento de los clientes en relación con los servicios de red.

4.3. Modelamiento

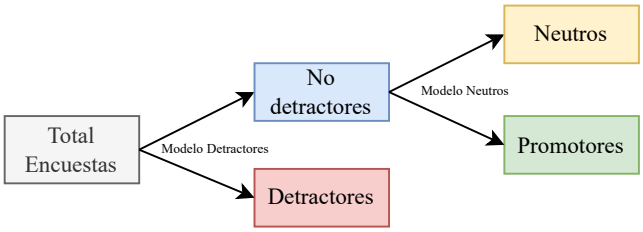


Figura 4.11: Estructura Modelo Satisfacción

En la etapa de modelamiento se procederá a la construcción de un modelo de satisfacción, como se visualiza en la Figura 4.11, que seguirá un enfoque de dos etapas. En primer lugar, se calibrará un modelo para distinguir entre clientes detractores y no detractores. Posteriormente, se aplicará un segundo modelo para diferenciar a los clientes neutros de promotores. Este enfoque permitirá clasificar a los clientes en las tres categorías fundamentales para el negocio: detractores, neutros y promotores.

Selección de Variables

El proceso de selección de variables es un paso crucial en el modelamiento y se enfoca en identificar aquellas características más relevantes y con mayor poder predictivo para los modelos. Este proceso se llevó a cabo en dos fases: eliminación de variables altamente correlacionadas y evaluación de su *Information Value* sobre los conjuntos de entrenamiento de los modelos.

Para descartar variables redundantes, se eliminaron aquellas con una correlación de Pearson superior a 0.8 en valor absoluto. Luego, se empleó el *Information Value* (IV) para evaluar el poder predictivo de las variables restantes. Siguiendo la guía proporcionada en la Tabla 2.3 (Página 20), se seleccionaron solo aquellas variables con un $IV \geq 0.02$, es decir, variables con un grado significativo de relevancia predictiva.

Modelo Detractores

Para la selección de variables en el modelo de detractores se evaluó el*Information Value* y se obtuvieron las siguientes variables relevantes que serán utilizadas para entrenar el modelo.

Tabla 4.5: Information Value conjunto entrenamiento modelo detractores

Variable	Information Value
Porcentaje sesiones con nulo RSRP/RSRQ	0.023
Total tráfico en 4G	0.021
Score 3G	0.033
Total cambios de 4G a 3G	0.041
Score 4G	0.065
Total de sesiones en 3G	0.075
Total de días sin sesiones en 3G	0.095
Total tráfico en 3G	0.134

Modelo Neutros

Para el modelo enfocado en diferenciar clientes neutros de promotores, el *Information Value* reveló que las siguientes 5 variables eran las más relevantes (Tabla 4.6).

Tabla 4.6: Information Value conjunto entrenamiento modelo neutros

Variable	Information Value
Total de sesiones en 3G	0.022
Score 3G	0.024
Total tráfico en 4G	0.024
Total días sin datos en 3G	0.029
Total tráfico en 3G	0.032

La cuidadosa selección de estas variables garantiza que los modelos para identificar clientes detractores y neutros estén fundamentados en datos significativos y relevantes. Este enfoque ayuda a maximizar la precisión y eficacia de los modelos.

4.3.1. Entrenamiento de Modelos

El proceso de entrenamiento de los modelos se centró en evaluar tres algoritmos: *Balanced Random Forest, LightGBM y XGBoost*, aplicados a los conjuntos de variables seleccionadas para ambos modelos de detractores y neutros. Para optimizar los hiperparámetros de cada modelo, se utilizó *HalvingGridSearchCV* de *scikit-learn* [34], un método de búsqueda en grillas eficiente para el ajuste de hiperparámetros. Asimismo, se utilizó el 80 % del total de datos

en cada modelo para entrenar.

La selección del modelo óptimo se basó en lograr un equilibrio entre un alto valor de Recall y la minimización del *overfitting*, utilizando como referencia las curvas ROC y Precision-Recall. Los resultados obtenidos se presentan en la Tabla 4.7, resumiendo el desempeño de los modelos en términos de Recall para ambas clasificaciones.

Tabla 4.7: Resultados Modelos

Modelo	Recall Modelo Detractores	Recall Modelo Neutros
Balanced Random Forest	0.569	0.568
LightGBM	0.563	0.525
XGBoost	0.719	0.888

Aunque *XGBoost* mostró el mejor rendimiento en términos de Recall, su alta diferencia entre las curvas de entrenamiento y prueba (como se detalla en la Sección de Anexos 5.2.3) indicó un significativo *overfitting*. Por lo tanto, se seleccionó el *Balanced Random Forest* como el modelo definitivo para identificar tanto a los detractores como a los neutros. Este modelo no solo proporcionó un buen Recall, sino que también demostró un ajuste más consistente entre los conjuntos de entrenamiento y prueba, indicando una generalización más robusta.

La elección del*Balanced Random Forest* subraya la importancia de considerar no solo la métrica de decisión de un modelo, sino también su capacidad para generalizar a nuevos datos, un aspecto crucial para garantizar la confiabilidad y aplicabilidad práctica del modelo en la industria.

4.3.2. Resultados Modelo Detractores

El modelo *Balanced Random Forest* seleccionado para identificar a los clientes detractores ha demostrado ser eficaz en la discriminación de este grupo específico. La Tabla 4.8 presenta la matriz de confusión que resume el desempeño del modelo.

Tabla 4.8: Matriz de confusión Modelo Detractores

		Predicción	
		No Detractor	Detractor
Real	No Detractor	3.674	2.208
	Detractor	405	535

El Recall para la clase de detractores es del 57 %, indicando que el modelo identifica correctamente a más de la mitad de los clientes detractores. Sin embargo, el Precision de solo el 20 % sugiere que hay un margen considerable de falsos positivos.

Análisis de Ganancia Lift

Si se realiza un análisis de ganancia de Lift se obtienen los resultados presentes en la Tabla 4.9.

Tabla 4.9: Ganancia Lift Modelo Detractores

Decil	No Detractores	Detractores	Tasa Respuesta	Lift
1	516	167	0.24	1.74
2	540	142	0.21	1.52
3	569	113	0.17	1.23
4	571	111	0.16	1.16
5	579	103	0.15	1.09
6	602	80	0.12	0.87
7	615	67	0.10	0.73
8	623	59	0.09	0.65
9	635	47	0.07	0.51
10	632	51	0.07	0.51

Al examinar la Ganancia Lift para el modelo detractores, se destaca que el modelo presenta un rendimiento significativamente mejor que una clasificación aleatoria en los primeros cinco deciles, donde la propensión a ser detractores es mayor. En el primer decil, el modelo logra un impresionante Lift de 1.78, indicando que supera en 1.8 veces la eficacia de una clasificación aleatoria. Esta tendencia positiva se mantiene en los siguientes deciles, con el quinto decil aún mostrando un Lift respetable de 1.07. Además, se puede observar en la Figura I.1 de la Sección de Anexos que el modelo efectivamente prioriza la clasificación de los clientes detractores. Si se ordenan por decil en función de su probabilidad de ser detractores, se capturan más clientes detractores en los primeros deciles.

La curva Lift se puede observar en la Figura 4.12, la cual es decreciente y para el primer decil logra una mejora de alrededor 1.5 veces sobre la clasificación aleatoria.

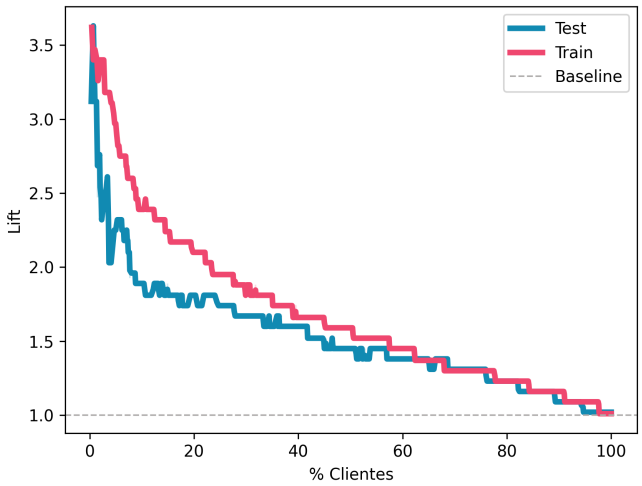


Figura 4.12: Curva Lift Modelo Detractores

Curva ROC y PR

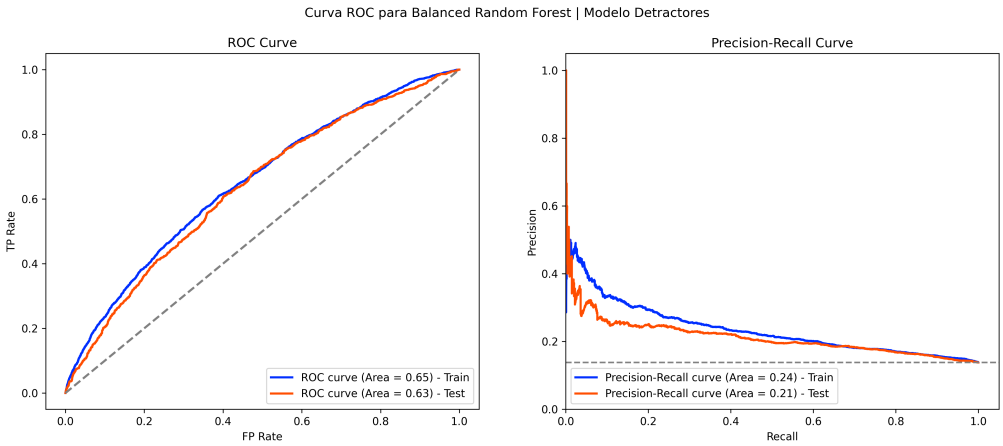


Figura 4.13: Curva ROC y Curva PR Modelo Detractores

Las curvas ROC y Precision-Recall proporcionan un resumen de los resultados del modelo. Se observa que tanto para el conjunto de entrenamiento como para el de prueba, las curvas son notablemente similares, indicando la capacidad del modelo para evitar el sobreajuste (*overfitting*). Se obtiene un área bajo la curva (AUC) de 0.63 en el conjunto de prueba, lo cual se considera un desempeño satisfactorio y sugiere que el modelo posee capacidad predictiva superior a un modelo aleatorio.

Explicabilidad

Parte fundamental del utilizar modelos en la industria es entender la razón de sus resultados. Para comprender la sensibilidad de las variables en la clasificación se utilizó la biblioteca SHAP de *python*. Se generó el siguiente gráfico que representa la importancia de las variables. Este gráfico condensa el aporte marginal de cada variable en relación con la tendencia a predecir la clase positiva.

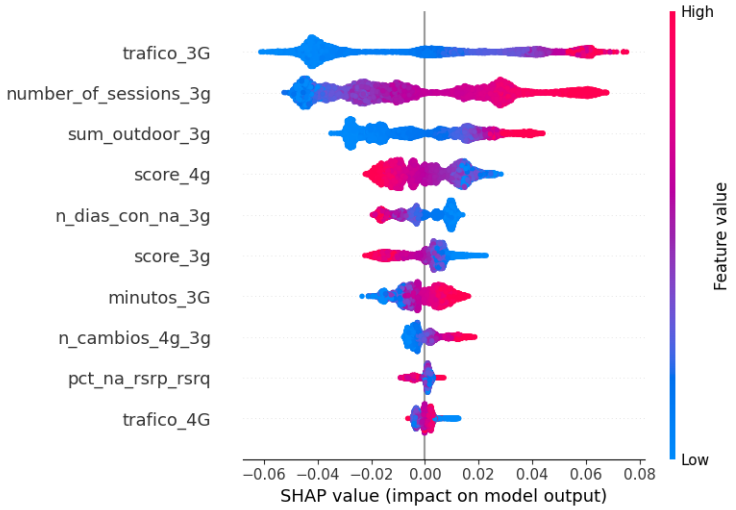


Figura 4.14: Importancia de SHAP Modelo Detractores

De esta observación, se destaca que las personas con un elevado tráfico, minutos y número de sesiones en la red 3G tienden a clasificarse como detractores. Por otro lado, aquellos con un destacado Score 4G, es decir, aquellos que frecuentemente se encuentran en áreas con una sólida cobertura y calidad de la red 4G, tienden a mostrar menor tendencia a ser detractores.

También es interesante notar que a mayor cantidad de días sin datos en 3G hay mayor probabilidad de no ser detractor, es decir, un cliente que rara vez aparece traficando en 3G es poco probable que sea detractor. Lo que se relaciona directamente con el tráfico en esta tecnología.

En términos generales, la experiencia predominante en un cliente clasificado como detractor se caracteriza por una mayor utilización de la red 3G, acompañada de un significativo número de transiciones de 4G a 3G.

4.3.3. Resultados Modelo Neutros

Tabla 4.10: Matriz de confusión Modelo Neutros

		Predicción	
		Promotor	Neutro
Real	Promotor	2.434	2.000
	Neutro	624	822

De la matriz de confusión presente en la Tabla 4.10 se puede calcular el Recall de la clase neutro, con lo que se obtiene un 57 % de Recall. Lo que indica que el modelo logra identificar al 57 % del universo de clientes neutros. Asimismo, si se analiza el Precision, se obtiene un 29 %, es decir, del conjunto de clientes no detractores, el modelo identifica al 29 % de los que predice como neutros.

Análisis de Ganancia

Tabla 4.11: Ganancia Lift Modelo Neutros

Decil	Promotores	Neutros	Tasa Respuesta	Lift
1	371	217	0.37	1.50
2	402	186	0.32	1.30
3	429	159	0.27	1.10
4	447	141	0.24	0.98
5	450	138	0.23	0.94
6	448	140	0.24	0.98
7	454	134	0.23	0.94
8	462	126	0.21	0.85
9	484	104	0.18	0.73
10	487	101	0.17	0.69

Al examinar la Tabla 4.11 de Ganancia Lift, el modelo presenta una mejora en comparación con la selección aleatoria en los primeros tres deciles, donde se obtienen Lifts de 1.5, 1.3 y 1.1, indicando un rendimiento superior a una clasificación aleatoria. Además, se puede observar en la Figura I.2 de la Sección de Anexos 5.2.3 que el modelo efectivamente prioriza la clasificación de los clientes neutros. Si se ordenan por decil en función de su probabilidad de ser neutros, se capturan más clientes neutros en los primeros deciles.

La curva Lift se puede observar en la Figura 4.15, la cual es decreciente y para el primer decil logra una mejora de alrededor 1.7 veces sobre la clasificación aleatoria.

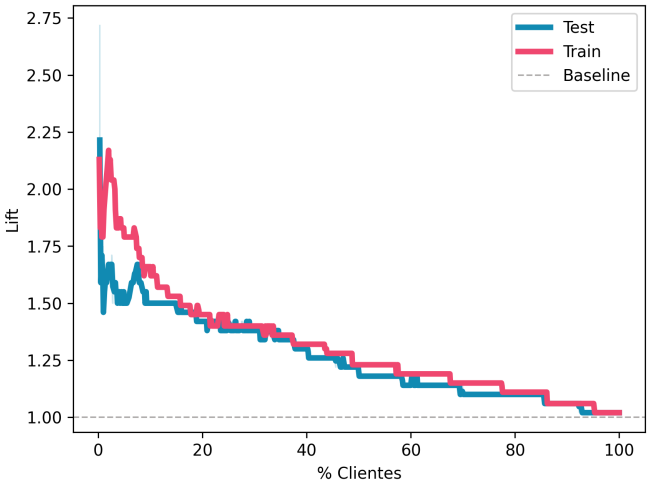


Figura 4.15: Curva Lift Modelo Neutros

Curva ROC y PR

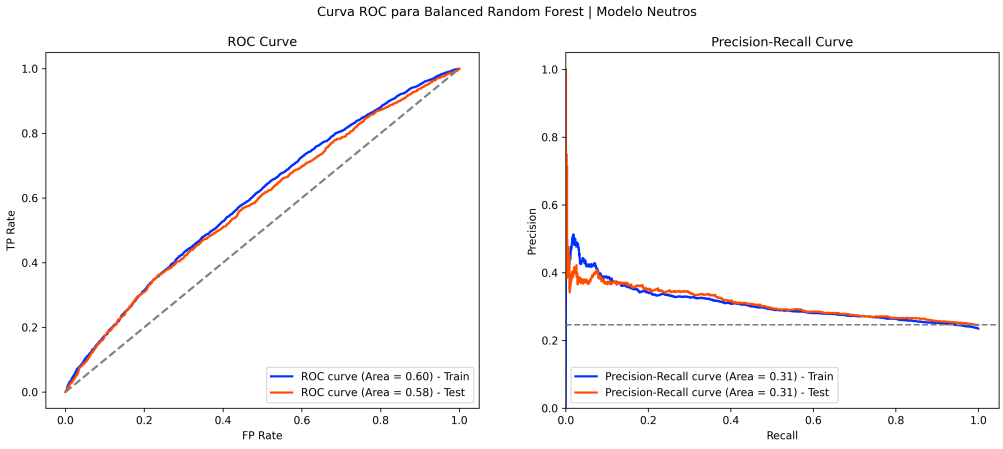


Figura 4.16: Curva ROC y Curva PR Modelo Neutros

En este escenario, las curvas ROC y Precision-Recall revelan un rendimiento inferior del modelo en la categoría de neutro en comparación con la de detractores. Esta discrepancia es comprensible, ya que la clasificación de neutro se refiere únicamente a una nota en la escala

de 1 a 7, lo que implica una tarea más compleja. A pesar de esto, se ha logrado un AUC de 0.58 en el conjunto de prueba, indicando que el modelo posee una capacidad de clasificación superior al azar.

Explicabilidad

Los resultados de SHAP para este modelo se tienen en la Figura 4.17.

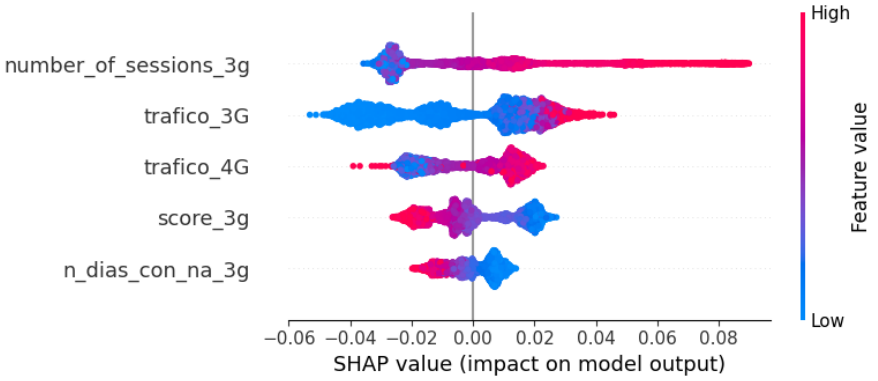


Figura 4.17: Importancia de SHAP Modelo Neutros

Resulta interesante observar un patrón similar al modelo de detractores. Los clientes con un elevado número de sesiones y tráfico en la red 3G tienden a clasificarse como neutros. Curiosamente, la variable de tráfico en la red 4G también se destaca, ya que un valor elevado aumenta la propensión del cliente a ser clasificado como neutro. Además, un buen Score 3G, interpretado como el cliente que pasa la mayor parte del tiempo en áreas con buena cobertura y calidad en la red 3G, indica una disminución en la tendencia a ser clasificado como neutro.

4.4. Comentarios Finales

Este capítulo ha abordado con éxito el desarrollo de un modelo de satisfacción para diferenciar clientes detractores, neutros y promotores a través de un minucioso proceso de comprensión, preparación y modelado de datos.

El análisis univariado y multivariado ha permitido comprender las distribuciones de los datos así como también identificar a priori pequeñas diferencias de las variables por tipo de cliente. Lo que ha dado ideas para crear variables agregadas de clientes basadas en llamadas, tráfico, calidad, cobertura y movilidad.

El proceso meticuloso de selección de variables mediante el *Information Value* ha resultado en la creación de modelos eficientes y significativos, cuyas variables no solo son estadísticamente relevantes sino también intuitivamente comprensibles en términos de la experiencia del cliente.

Los modelos han mostrado una capacidad notable para clasificar correctamente a los clientes en las categorías de detractores y neutros, logrando un Recall en cada caso de 57 % y 57 % respectivamente. Este nivel de recall en la identificación de clientes insatisfechos y neutrales es un activo valioso para cualquier estrategia de retención de clientes y optimización de la

red. Esta capacidad de identificación proactiva de clientes potencialmente insatisfechos abre la puerta a intervenciones dirigidas y oportunas para mejorar la experiencia del cliente y, por ende, su satisfacción general.

En resumen, los modelos desarrollados ofrecen una herramienta poderosa para mejorar la estrategia de retención de clientes y la optimización de la red. La identificación proactiva de clientes potencialmente insatisfechos permite acciones dirigidas para mejorar su experiencia.

Capítulo 5

Conclusiones y Propuestas

5.1. Conclusiones

El vertiginoso aumento del tráfico de datos en dispositivos móviles y el auge del Big Data han resaltado no solo la importancia, sino también la urgencia de desarrollar soluciones innovadoras y basadas en datos en el ámbito de las redes móviles de telecomunicaciones. En este escenario dinámico y altamente competitivo, este trabajo no solo ha respondido a esta demanda, sino que también ha establecido nuevos estándares para la toma de decisiones impulsada por datos en la industria.

Se presentaron dos herramientas innovadoras para mejorar la toma de decisiones de un operador de redes móviles, centradas en la experiencia del cliente. La primera, una metodología para la identificación geoespacial de zonas de interés de calidad y cobertura de red. Este enfoque ofrece no solo una imagen detallada de la situación actual en la Región Metropolitana, sino que también sienta las bases para realizar el análisis para las diferentes comunas del país. El descubrimiento más notable de este enfoque fue la alta correlación de 0.9 entre el Score 4G propuesto y la satisfacción de los clientes, subrayando la potencialidad del Score 4G como un indicador crítico y transformador para el negocio desde la perspectiva de los clientes.

El segundo enfoque, un Modelo de Satisfacción del Cliente basado en solo 7 variables, ha demostrado ser no solo eficiente sino también poderoso. Con un Recall del 57 % para el modelo de detractores y 57 % para neutros, este modelo abre una ventana hacia la comprensión y extrapolación de la satisfacción del cliente a toda la base de clientes de la compañía. Este trabajo es particularmente crucial, ya que permite la implementación de campañas de retención más dirigidas y eficaces, enfocadas en clientes que experimentan problemas de red.

Estos resultados no solo demuestran la viabilidad de tomar decisiones fundamentadas en datos en la industria de las telecomunicaciones, sino que también ilustran el poder de los análisis avanzados para transformar grandes volúmenes de información en ventajas competitivas tangibles. Al combinar el análisis del Score 4G con el Modelo de Satisfacción, se abre un camino hacia una gestión más inteligente y centrada en el cliente, permitiendo a las empresas no solo retener a sus clientes sino también mejorar su experiencia.

5.2. Propuestas

5.2.1. Propuesta 1: Oportunidades de Mejoras de Red

Se propone expandir el cálculo del Score 4G a un nivel más granular, específicamente a nivel comunal. Esta expansión permitirá identificar con precisión áreas críticas, ofreciendo al operador de redes móviles una herramienta avanzada para una asignación de recursos más estratégica y focalizada en su planificación anual. Al concentrar los esfuerzos en comunas con Scores 4G bajos, la empresa podrá abordar eficazmente las deficiencias en cobertura y calidad, mejorando directamente la experiencia del cliente.

5.2.2. Propuesta 2: Oportunidad de mejoras comerciales

La aplicación del Score 4G en regiones con puntuaciones bajas en cobertura y calidad puede revelar oportunidades clave para mejorar la retención y la experiencia del cliente. Sugerimos un enfoque combinado que utilice el Score 4G y el Modelo de Satisfacción para identificar clientes potencialmente insatisfechos en estas áreas. Este método permitiría implementar estrategias proactivas y personalizadas, como comunicaciones directas con estos clientes, para abordar sus inquietudes específicas y mejorar su percepción y lealtad hacia el servicio.

5.2.3. Propuesta 3: Análisis en profundidad para monitoreo de KPI

Se recomienda el uso continuo del Score 4G para monitorear cómo el crecimiento urbano y otros cambios demográficos afectan la cobertura y calidad de la red a lo largo del tiempo. Este seguimiento permitirá a la empresa comprender y anticipar las tendencias en la evolución de la red, proporcionando datos valiosos para la planificación estratégica a largo plazo y la toma de decisiones informadas.

Adicionalmente, el Score 4G debe ser incorporado en el sistema de monitoreo de KPIs de la empresa, permitiendo una evaluación continua y ajustes dinámicos en respuesta a los cambios en la demanda del mercado y las condiciones de la red.

Estas propuestas mejoradas no solo subrayan la aplicación práctica de tus hallazgos, sino que también muestran un enfoque estratégico y proactivo hacia la mejora continua y la adaptación a las necesidades cambiantes del mercado y los clientes.

Bibliografía

[1] Gustafsson, A., Johnson, M. D., y Roos, I., "The effects of customer satisfaction, relationship commitment dimensions, and triggers on customer retention," Journal of Marketing, vol. 69, no. 4, pp. 210–218, 2005, doi:10.1509/jmkg.2005.69.4.210.

[2] Ostornol, J. T., "Celulares en chile superan los 33 millones y gasto promedio de usuarios cae con fuerza en 10 años," 2022, https://www.df.cl/celulares-en-chile-superan-los-33-4-millones-y-gasto-promedio-de (visitado el 2023-09-19).

[3] Ouyang Y, Wang L, Y. A. *et al.*, "Next decade of telecommunications artificial intelligence," SciOpen, pp. 20–22, 2022, doi:10.26599/AIR.2022.9150003.

[4] Subsecretaria de Telecomunicaciones, "Sector telecomunicaciones cierre 2022," 2023, https://www.subtel.gob.cl/estudios-y-estadisticas/informes-sectoriales-anuales/ (visitado el 2023-11-06).

[5] Entel, "Memoria integrada 2022," 2023, https://informacioncorporativa.entel.cl/ (visitado el 2023-11-06).

[6] Reichheld, F. F., "The one number you need to grow," Harvard business review, vol. 81, no. 12, pp. 46–55, 2003.

[7] Reichheld, F., The ultimate question 2.0 (revised and expanded edition): How net promoter companies thrive in a customer-driven world. Harvard Business Review Press, 2011.

[8] Nokia, "Quality of experience (qoe) of mobile services: Can it be measured and improved?," 2004, https://docplayer.net/25986899-White-paper-quality-of-experience-qoe-of-mobile-services-can-it-be-measured-and-improved.html.

[9] Moso, J. y Muange, W., "Quality of experience (qoe) measurement and its challenges in mobile networks for multimedia," vol. 7, 2018.

[10] "ITU-T P.800: Mean opinion score (mos) - definitions of terms related to the quality of a telecommunication service." ITU-T Recommendation, 1996, https://www.itu.int/rec/T-REC-P.800/es.

[11] Vasilios A. Siris, Konstantinos Balampekos, M. K. M., "Mobile quality of experience: Recent advances and challenges," IEEE Explore, 2014.

[12] Gómez-Andrades, A., Barco, R., y Serrano, I., "A method of assessment of lte coverage holes," EURASIP Journal on Wireless Communications and Networking, vol. 2016, no. 1, pp. 1–12, 2016, doi:10.1186/s13638-016-0733-y.

[13] Qiao, J., Zhu, J., Cheng, X., Xu, L., Meng, N., Cheng, L., Zhai, J., Di, Z., y Dong, F., "Weak coverage analysis method for mobile networks based on machine learning," en Signal and Information Processing, Networking and Computers (Wang, Y., Liu, Y.,

Zou, J., y Huo, M., eds.), (Singapore), pp. 1171–1178, Springer Nature Singapore, 2023.

[14] Feibi Lyu, Chen Cheng, J. Z. X. C. L. X. Z. W. J. Q. L. L. Z. D., "Coverage estimation of mobile network using supervised learning model on artificial estimation dataset," en 2021 International Conference on Information and Communication Technologies for Disaster Management (ICT-DM), pp. 1–6, IEEE, 2022, doi:10.1109/ICT-DM52643.2021.9664185.

[15] Chernov, S., Petrov, D., y Ristaniemi, T., "Location accuracy impact on cell outage detection in lte-a networks," IEEE, 2015, doi:10.1109/IWCMC.2015.7289247.

[16] Pierucci, L. y Micheli, D., "A neural network for quality of experience estimation in mobile communications," 2016, doi:10.1109/mmul.2016.21.

[17] Yusuf-Asaju, A. W., Dahalin, Z. B., y Ta'a, A., "Mobile network quality of experience using big data analytics approach," 2017, doi:10.1109/icitech.2017.8079923.

[18] Mitra, K., Ahlund, C., y Zaslavsky, A., "A decision-theoretic approach for quality-of-experience measurement and prediction," 2011, doi:10.1109/icme.2011.6012098.

[19] Pedras, V., Sousa, M., Vieira, P., Queluz, M., y Rodrigues, A., "A no-reference user centric qoe model for voice and web browsing based on 3g/4g radio measurements," pp. 1–6, 2018, doi:10.1109/WCNC.2018.8377198.

[20] Qiao, J. *et al.*, "Mobile network user perception prediction based on random forest algorithm," IEEE, 2022, doi:10.1109/ICT-DM52643.2021.9664054.

[21] Wu, W., Arefin, A., Rivas, R., Nahrstedt, K., Sheppard, R., y Yang, Z., "Quality of experience in distributed interactive multimedia environments," 2009, doi:10.1145/1631272.1631338.

[22] Entel, "Entel," 2023, https://ce.entel.cl/articulos/redes-moviles/ (visitado el 2023-08-19).

[23] S. Sesia, I. T. y M.Baker, LTE: The UMTS Long Term Evolution: From Theory to Practice. United Kingdom: John Wiley Sons Ltd., 2011.

[24] Telecom, V., "How to interpret rscp(3g) / rsrp(4g) and rsrq (lte) parameters.," 2021, https://help.venntelecom.com/support/solutions/articles/44001931273-how-to-interpret-rscp-3g-rsrp-4g-and-rsrq-lte-parameters- (visitado el 2023-08-19).

[25] de Telecomunicaciones, S., "Preguntas frecuentes," 2018, https://multibanda.cl/preguntas-frecuentes/#faq4.

[26] Wooldridge, J. M., Introducción a la Econometría: Un Enfoque Moderno. Cengage Learning Editores, 5th ed., 2012.

[27] Breiman, L., 2001, doi:10.1023/a:1010933404324.

[28] Salunkhe, V., "Random forest classification," 2021-07-21, https://medium.com/@viveksalunkhe80/random-forest-classification-c0afb1fb0430 (visitado el 2023-11-04).

[29] Friedman, J. H., "Greedy function approximation: A gradient boosting machine.," The Annals of Statistics, vol. 29, no. 5, pp. 1189 – 1232, 2001, doi:10.1214/aos/1013203451.

[30] Florek, P. y Zagdański, A., "Benchmarking state-of-the-art gradient boosting algorithms for classification," 2023.

[31] Chen, T. y Guestrin, C., "Xgboost: A scalable tree boosting system," CoRR, vol. abs/1603.02754, 2016, http://arxiv.org/abs/1603.02754.

[32] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., y Liu, T.-Y., "Lightgbm: A highly efficient gradient boosting decision tree," en Advances in Neural Information Processing Systems (Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., y Garnett, R., eds.), vol. 30, Curran Associates, Inc., 2017, https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

[33] Fawcett, T., "An introduction to roc analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, 2006, doi:https://doi.org/10.1016/j.patrec.2005.10.010. ROC Analysis in Pattern Recognition.

[34] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., y Duchesnay, E., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[35] Thorn, J., "The lift curve: Unveiled," 2022-01-02, https://towardsdatascience.com/the-lift-curve-unveiled-998851147871 (visitado el 2023-11-04).

[36] Kulkarni, K., "Understand weight of evidence and information value!," 2021-07-20, https://www.analyticsvidhya.com/blog/2021/06/understand-weight-of-evidence-and-information-value/ (visitado el 2023-11-06).

[37] Miller, T., "Explanation in artificial intelligence: Insights from the social sciences," Artificial Intelligence, vol. 267, pp. 1–38, 2019, doi:https://doi.org/10.1016/j.artint.2018.07.007.

[38] Štrumbelj, E. y Kononenko, I., "Explaining prediction models and individual predictions with feature contributions," Knowledge and Information Systems, vol. 41, pp. 647–665, 2014, doi:10.1007/s10115-013-0679-x.

[39] Uber Blog, "H3: Uber's hexagonal hierarchical spatial index," 2018, https://www.uber.com/en-CL/blog/h3/ (visitado el 17/07/23).

[40] Siabato, W. y Guzmán-Manrique, J., "La autocorrelación espacial y el desarrollo de la geografía cuantitativa," Cuadernos de Geografía: Revista Colombiana de Geografía, vol. 28, pp. 1–22, 2019, doi:10.15446/rcdg.v28n1.76919.

[41] Anselin, L., "Local indicators of spatial association—lisa," Geographical Analysis, vol. 27, no. 2, pp. 93–115, 1995, doi:https://doi.org/10.1111/j.1538-4632.1995.tb00338.x.

[42] GL, K., "Kepler gl," 2023, https://kepler.gl/ (visitado el 2023-08-19).

[43] Fayyad, U. M., Piatetsky-Shapiro, G., y Smyth, P., "From data mining to knowledge discovery in databases," AI Mag., vol. 17, pp. 37–54, 1996, https://api.semanticscholar.org/CorpusID:61287995.

[44] Rüdiger Wirth, J. H., "Crisp-dm: towards a standard process modell for data mining," Journal of Data Warehousig, 2000.

[45] Studer, S., Bui, T. B., Drescher, C., Hanuschkin, A., Winkler, L., Peters, S., y Müller, K., "Towards CRISP-ML(Q): A machine learning process model with quality assurance methodology," CoRR, vol. abs/2003.05155, 2020, https://arxiv.org/abs/2003.05155.

[46] Patel, H., "Data-centric approach vs model-centric approach in machine learning," 2023-08-01, https://neptune.ai/blog/data-centric-vs-model-centric-machine-learning

(visitado el 2023-11-06).

[47] González, M. C., Hidalgo, C. A., y Barabási, A.-L., "Understanding individual human mobility patterns," Nature, vol. 453, pp. 779–782, 2008, doi:10.1038/nature06958.

[48] Richardson Corvalán, C., "Construcción y caracterización de perfiles de clientes en base a su movilidad," 2014, https://repositorio.uchile.cl/handle/2250/132029.

Anexos

Anexo A. Columnas presentes en las bases de 3G/4G

3G/4G - imsi: Identificador anonimizado asociado al cliente

3G/4G - absolutetime: Tiempo exacto en días desde 1900 en que el registro fue almacenado

3G/4G - dia: Día en que se registra la sesión

3G/4G - hora: Hora en que se registra la sesión

3G/4G - imsi: Identificador único del móvil, asociado al chip del teléfono.

3G/4G - starlat, starlon: Latitud y longitud inicial donde se inicia la sesión

3G/4G - endlat, endlon: Latitud y longitud final donde termina la sesión

3G/4G - geolocationflag: Toma el valor de 3 si se logra geolocalizar la sesión en base a triangulación de antenas, 0 sino.

3G/4G - connectionok_cs: Indicador de que la llamada se ha establecido correctamente en la conexión.

3G/4G - connectiondropped_cs: Indicador de que se ha caído la llamada durante la conexión.

3G/4G - connectionfailed_cs: Indicador de que ha fallado la llamada durante la conexión.

4G - session_av_rsrp_sev_count: Se utiliza junto con session_av_rsrp_sev_sum para calcular el valor RSRP medio del sector de servicio durante la duración de la conexión.

4G - session_av_rsrq_sev_count: Se utiliza junto con session_av_rsrq_sev_sum para calcular el valor RSRP medio del sector de servicio durante la duración de la conexión.

4G - session_av_rsrp_sev_sum: Se utiliza junto con session_av_rsrp_sev_count para calcular el valor RSRP medio del sector de servicio durante la duración de la conexión.

4G - session_av_rsrq_sev_sum: Se utiliza junto con session_av_rsrq_sev_count para calcular el valor RSRP medio del sector de servicio durante la duración de la conexión.

3G - session_av_ecno_count: Se utiliza junto con session_av_ecno_sum para calcular el valor ECNO medio del sector de servicio durante la duración de la conexión.

3G - session_av_ecno_count: Se utiliza junto con session_av_ecno_sum para calcular el valor ECNO medio del sector de servicio durante la duración de la conexión.

3G - session_av_rscp_sum: Se utiliza junto con session_av_rscp_count para calcular el valor RSCP medio del sector de servicio durante la duración de la conexión.

3G - session_av_rscp_sum: Se utiliza junto con session_av_rscp_count para calcular el valor RSCP medio del sector de servicio durante la duración de la conexión.

3G - indoorconnection: Este indicador toma el valor de 1 cuando el algoritmo de localización indica que la conexión era interior.

3G - outdoorconnection: Este indicador toma el valor de 1 cuando el algoritmo de localización indica que la conexión era exterior.

Anexo B. Visualizaciones con Kepler

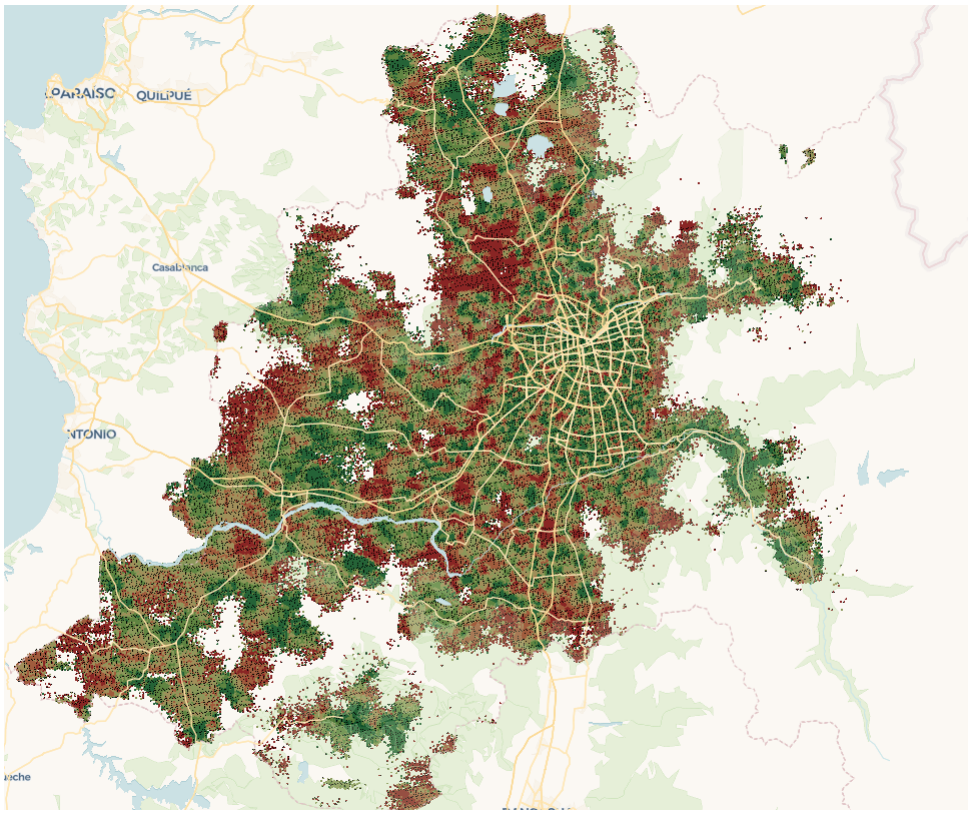


Figura B.1: Score 4G Región Metropolitana usando Kepler

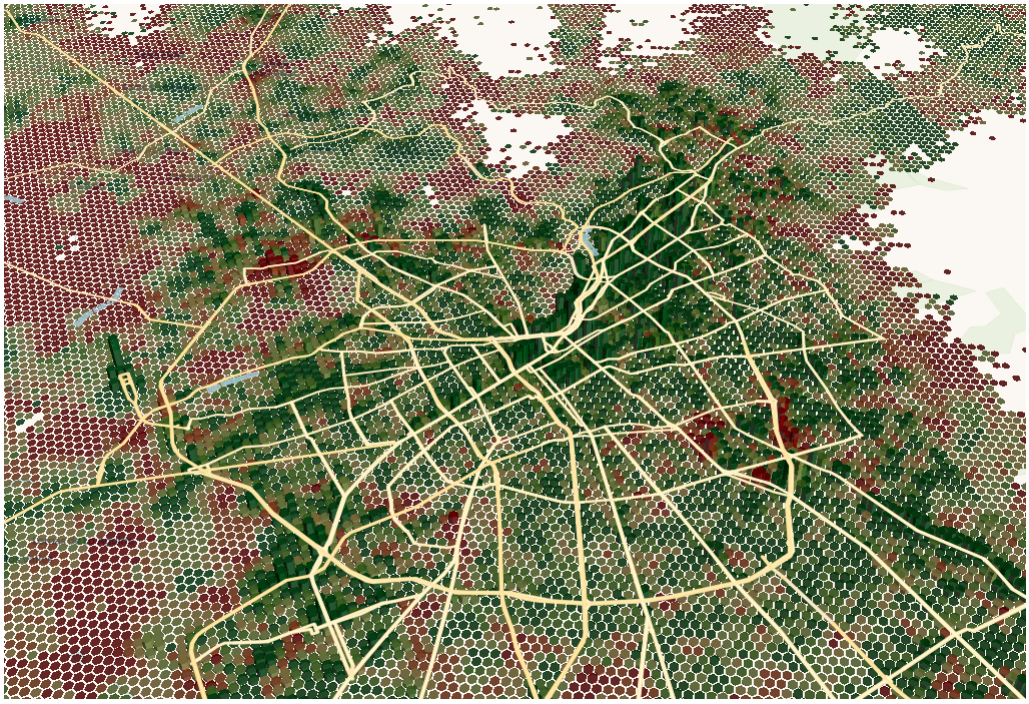


Figura B.2: Zonas Densas Región Metropolitana usando Kepler

Anexo C. Análisis Univariado 3G

Actix 3G

Conteo Sesiones

El conteo de sesiones representa el conteo de registros que contiene Actix 3G para cada cliente en el periodo estadístico definido de 30 días. Este cálculo permite comprender mejor las diferencias en el uso de la red entre promotores, neutros y detractores, proporcionando así una visión más clara de cómo la frecuencia de interacción con la red puede estar relacionada con los niveles de satisfacción del cliente.

A continuación se presenta la distribución en escala logarítmica diferenciada por tipo de cliente. (Figura C.1)

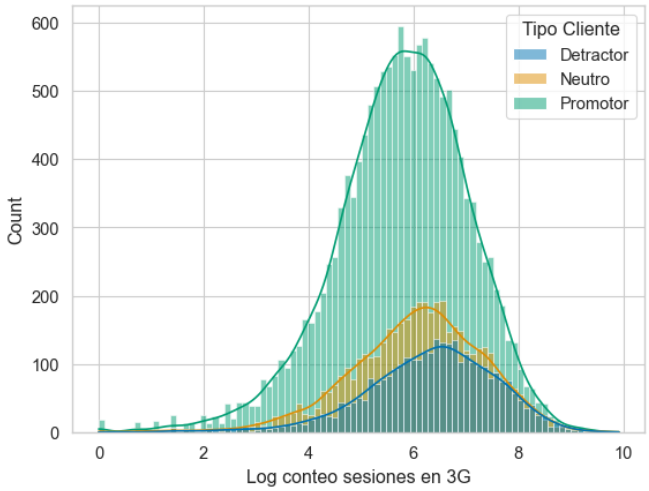


Figura C.1: Logaritmo del conteo sesiones en tecnología 3G

Se puede notar que la distribución entre los tipos de clientes es sutilmente distinta. Los clientes detractores tienden a tener más sesiones en 3G que los clientes promotores. Lo cual es interesante y puede significar que un cliente detractor tiende a experimentar más la red 3G.

Llamadas en 3G

Otro dato que contiene Actix 3G es relacionado a llamadas presente en la columna llamada *csdurationtime* que representa el tiempo de la llamada. Este dato permite evaluar la frecuencia de uso del servicio de llamadas por parte de los clientes. Para ello, se ha calculado la proporción de registros correspondientes a llamadas en relación con el total de registros por cliente. Esta distribución se visualiza en la Figura C.2, ofreciendo una perspectiva de la utilización del móvil en llamadas en la red 3G por tipo de cliente.

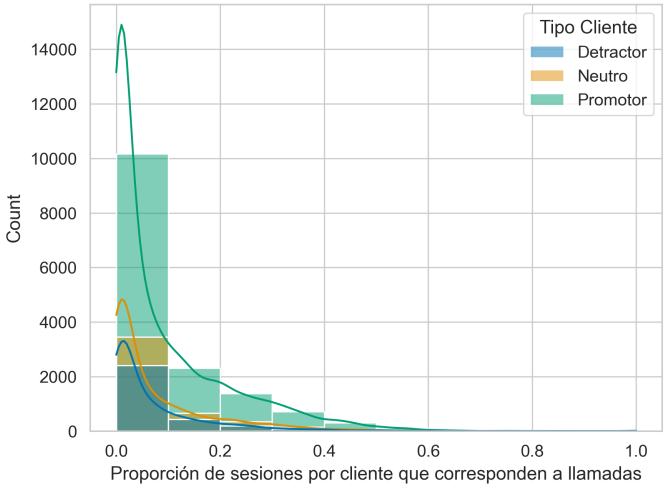


Figura C.2: Histograma proporción de sesiones que corresponden a llamadas en tecnología 3G por tipo de cliente

El análisis indica que no existen diferencias notables en la proporción de registros de lla-

madas entre los distintos tipos de clientes en Actix 3G. Además, se observa que la mayoría de las sesiones están relacionadas con tráfico de datos y no con llamadas.

Un análisis adicional interesante implica el estudio del promedio de llamadas realizadas a lo largo del día, distinguiendo entre diferentes tipos de clientes. La Figura C.3 muestra este patrón, destacando que el intervalo de tiempo con mayor promedio de llamadas se sitúa entre las 11 y las 13 horas. Sin embargo, este análisis no revela patrones distintivos que permitan diferenciar entre los tipos de clientes basados en el uso de llamadas.

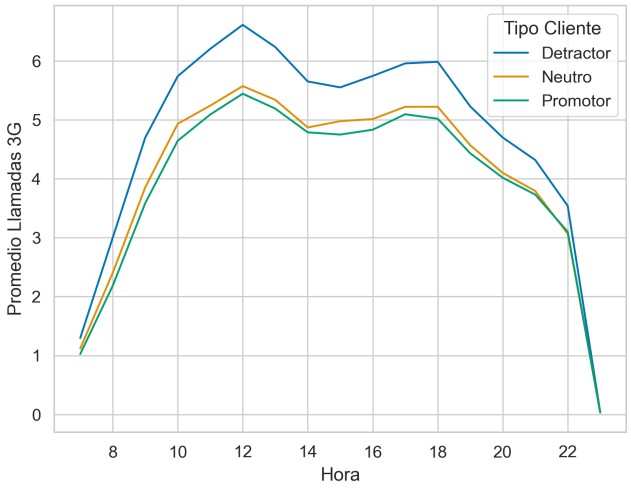


Figura C.3: Promedio llamadas en 3G hora del día por tipo de cliente

Sesiones Geolocalizadas

Una dimensión importante en el análisis de los datos de Actix 3G es la geolocalización precisa de las sesiones. Se examina el porcentaje de sesiones que han sido correctamente geolocalizadas, ya que una hipótesis inicial podría sugerir que los clientes insatisfechos experimentan un mayor número de sesiones sin geolocalización, posiblemente debido a problemas de cobertura, lo que podría indicar una experiencia de servicio deficiente. Sin embargo, al analizar los datos representados en la Figura C.4, se encuentra que no existen diferencias significativas en el porcentaje de sesiones geolocalizadas entre diferentes tipos de clientes.

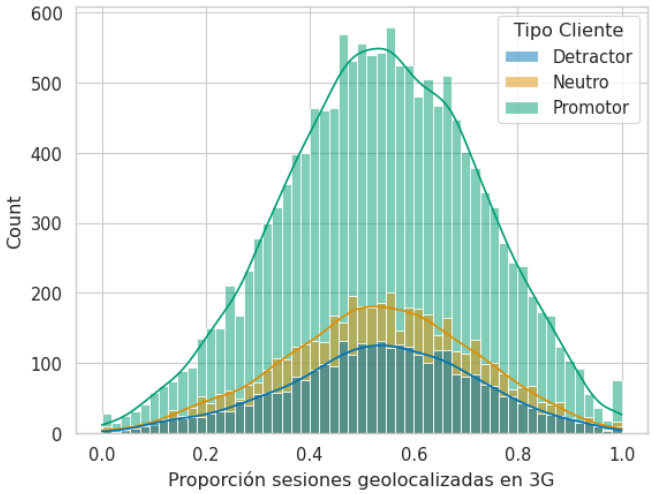


Figura C.4: Histograma porcentaje sesiones geolocalizadas en tecnología 3G por tipo de cliente

RRC Setup Time

Un aspecto técnico crucial en la experiencia de llamadas en 3G es el tiempo de configuración de la conexión RRC (Radio Resource Control), medido en milisegundos. Este tiempo, conocido como *rrcsetuptime*, es el período que tarda el teléfono del cliente en establecer una conexión durante una llamada. Intuitivamente, se puede inferir que un mayor tiempo de configuración de RRC podría impactar negativamente la satisfacción del cliente, especialmente si las llamadas no se conectan rápidamente, lo que podría conducir a repetidos intentos de llamada y, por tanto, a una experiencia desfavorable.

Tabla C.1: Análisis descriptivo variable rrcsetuptime

Conteo	12355730
Promedio	476
Desv. Est.	849
Minimo	0
25 %	324
50 %	381
75 %	435
Máximo	634576

La Tabla C.1 muestra un análisis descriptivo de la variable *rrcsetuptime*. Su valor máximo es 634576 milisegundos que corresponde a aproximadamente 11 minutos. Lo cual se entiende como una llamada que demoró en establecer la llamada alrededor de 11 minutos, lo cual es un caso bastante atípico.

El promedio es de 476 milisegundos que corresponde aproximadamente a 0.5 segundos. Tiene sentido, pues es un tiempo pequeño y se puede entender intuitivamente como lo que demora un teléfono en establecer la conexión de llamada con otro móvil.

Este análisis se profundiza agrupando los datos por cliente y calculando el promedio del *rrcsetuptime* durante el período estadístico considerado. Este enfoque permite examinar si hay variaciones significativas en el tiempo de configuración de RRC entre diferentes clientes y cómo esto podría correlacionarse con su nivel de satisfacción general con el servicio de llamadas en 3G.

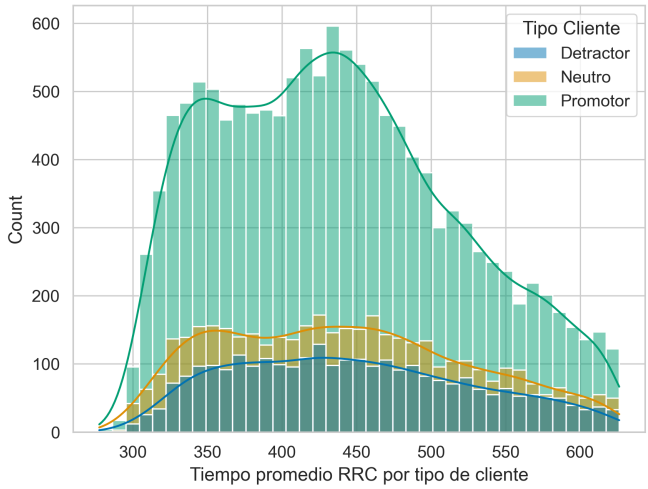


Figura C.5: Histograma promedio *rrc setup time* por tipo de cliente

A priori se puede analizar en la Figura C.5 que la distribución del promedio del *rrcsetuptime* es distinta entre clientes promotores y detractores, siendo la de estos últimos más plana. Al calcular el promedio de *rrcsetuptime* por cada categoría de cliente, se revelan diferencias sutiles pero significativas: los promotores tienen un promedio de 470 milisegundos, los neutros de 476 milisegundos, y los detractores de 487 milisegundos. Es particularmente revelador que los clientes menos satisfechos, o detractores, experimentan en promedio tiempos de establecimiento de llamadas ligeramente más largos.

Connection time

La base de datos Actix 3G incluye un parámetro clave denominado *connectiontime*, que mide la duración de cada sesión en milisegundos. Este tiempo representa la duración durante la cual un cliente permanece conectado a una misma celda de servicio. Se espera que los clientes con alta movilidad tengan un *connectiontime* promedio más corto, dado que cambian de celda con frecuencia.

Tabla C.2: Análisis descriptivo variable connectiontime

Conteo	16.776.815
Promedio	84.539
Desv. Est.	193.231
Minimo	0
25 %	1.592
50 %	13.455
75 %	50.240
Máximo	904.966

La Tabla C.2 proporciona un análisis descriptivo de esta variable. El valor máximo observado de 904.966 milisegundos, aproximadamente 15 minutos, es coherente con los intervalos de almacenamiento de sesiones de Actix, que son de 15 minutos. Este dato indica que es común que una sesión alcance la duración máxima registrada.

El promedio de *connectiontime* es de 84.539 milisegundos, lo que equivale aproximadamente a 1.4 minutos. Este valor sugiere que la mayoría de las sesiones son relativamente breves.

Al calcular el promedio de *connectiontime* por cliente y analizar su distribución (Figura C.6), se observa que no hay diferencias significativas en las distribuciones para los distintos tipos de clientes (promotores, neutros y detractores). Esto indica que la duración de la conexión no varía notablemente entre diferentes grupos de clientes en términos de su satisfacción con el servicio.

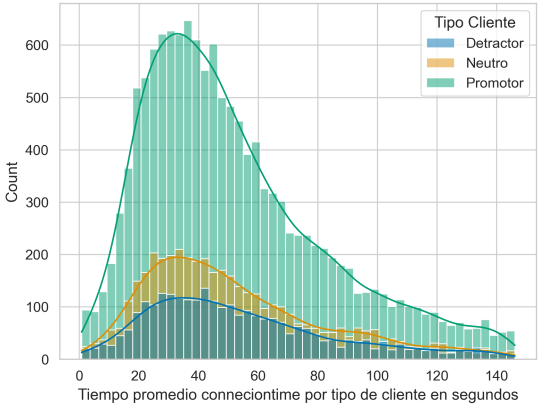


Figura C.6: Histograma y Boxplot variable connectiontime

Este hallazgo sugiere que la duración de las sesiones en la red 3G no es un factor diferenciador en la experiencia del cliente en términos de su percepción general del servicio.

Indoor/Outdoor connection

La base también contiene las siguientes variables: *indoorconnection* y *outdoorconnection*, representadas respectivamente con valores binarios (0 o 1). Estas variables indican si la sesión del cliente ocurrió en un entorno interior o exterior. La Tabla C.3 presenta un análisis

cruzado de estas variables con *geolocationflag*, identificando que las sesiones que no tienen valores asignados en *indoorconnection* y *outdoorconnection* corresponden a aquellas que no fueron geolocalizadas.

Tabla C.3: Tabla cruzada de indoorconnection, outdoorconnection y geolocationflag

		geolocationflag	geolocationflag
indoor connection	outdoor connection	0	3
0	0	8.009.017	0
	1	0	4.716.812
1	0	0	4.050.985

Al analizar la distribución de la proporción de sesiones clasificadas como *indoor* o *outdoor* para cada cliente, se observa en la Figura C.7 que no hay diferencias significativas en estas distribuciones entre los diferentes tipos de clientes.

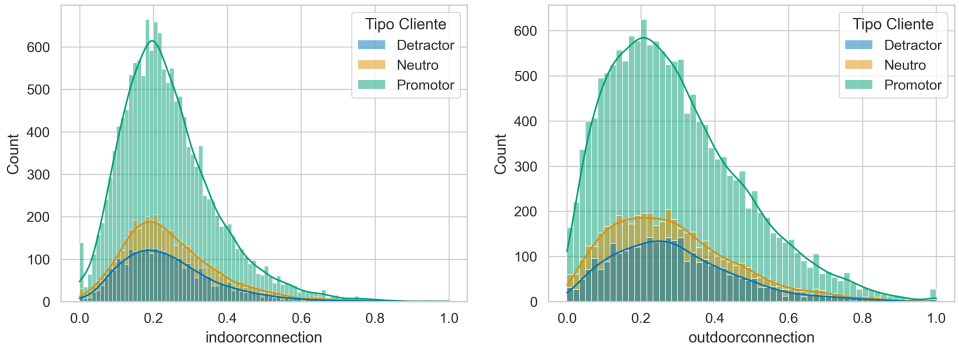


Figura C.7: Histograma del porcentaje de sesiones interiores (indoor) y exteriores (outdoor) por tipo de cliente

Estado conexión

Actix 3G también proporciona información sobre el estado de las conexiones durante las llamadas, mediante variables binarias como *connectionok_cs*, *connectiondropped_cs* y *connectionfailed_cs*. Estas representan si la llamada se realizó exitosamente, se cayó o falló, respectivamente. Aunque estos indicadores muestran una alta proporción de valores nulos, lo cual se explica por el predominio del tráfico de datos sobre las llamadas en los registros de Actix 3G, el análisis de estos datos podría revelar patrones significativos.

Tabla C.4: Porcentaje valores nulos variables sobre estado de conexiónt

Variable	% Valores Nulos
connectionok_cs	91.1
connectiondropped_cs	99.9
connectionfailed_cs	99.9

A pesar de la alta incidencia de valores nulos, un análisis más detallado se enfoca en el ratio de llamadas exitosas comparado con llamadas fallidas o caídas. Este ratio se calcula como $\frac{connectionok}{(connectionok+connectiondropped+connectionfailed)}$. La distribución de este ratio por tipo de cliente se observa en la Figura C.8

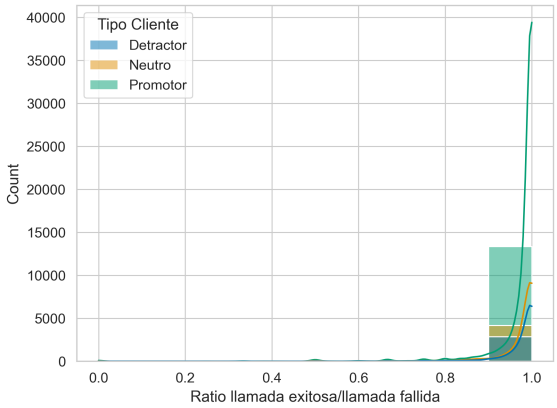


Figura C.8: Histograma ratio de conexiones en llamadas exitosas versus fallidas

Este resultado sugiere que cuando se realiza una llamada, la probabilidad de que sea exitosa es alta para todos los clientes, sin distinción significativa entre los diferentes tipos de clientes en términos de satisfacción. Esto podría indicar que la calidad de la conexión durante las llamadas no es un factor determinante en la percepción general de satisfacción, al menos en la red 3G.

Tráfico 3G

La fuente de datos denominada Tráfico 3G se enfoca exclusivamente en el volumen de datos utilizados por los clientes al navegar por Internet mediante la tecnología 3G. Esta variable registra el total de gigabytes transferidos por cada cliente durante el periodo de análisis. Dado que esta variable exhibe una gran variabilidad debido a la disparidad en los volúmenes de tráfico entre diferentes clientes, se aplicará una transformación logarítmica.

La Figura C.9 presenta la distribución del tráfico 3G, segmentada por tipo de cliente. En ella, se muestra el histograma correspondiente a los volúmenes de tráfico de datos en la red 3G.

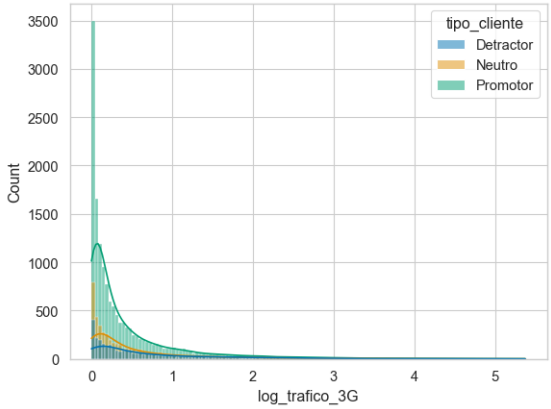


Figura C.9: Histograma del tráfico de 3G por tipo de cliente

A partir de esta visualización, no se logran observar diferencias significativas en el uso de datos en 3G entre los distintos tipos de clientes, sean estos promotores, neutros o detractores.

Anexo D. Analisis Multivariado 3G

Actix 3G

En 3G, la cobertura y calidad se miden a través de RSCP y ECNO. Indicadores que están presentes en Actix 3G a nivel de cliente. Se analizará la relación entre ambos indicadores por tipo de clientes.

La Figura D.1 muestra un gráfico de dispersión combinado con una estimación de la distribución de probabilidad utilizando la técnica de KDE (*Kernel Density Estimation*). Esta visualización revela que los clientes detractores tienden a tener valores más negativos de RSCP en comparación con los promotores, sugiriendo una posible correlación entre una menor cobertura (RSCP más negativo) y una menor satisfacción del cliente.

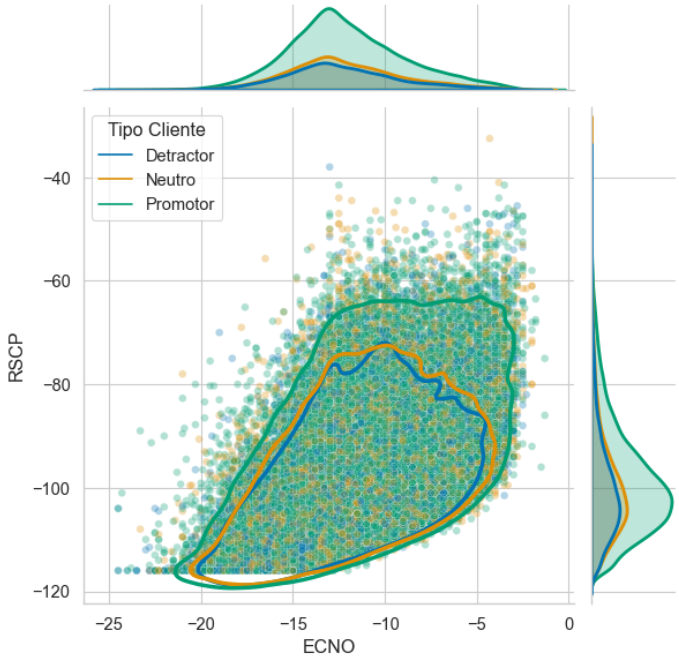


Figura D.1: Scatterplot RSCP vs ECNO con KDE

Anexo E. Tratamiento Valores Nulos 3G

Actix 3G

Se procede a calcular el porcentaje de valores nulos por columna para la base Actix 3G.

Tabla E.1: Variables con valores nulos en Actix 3G

Variable	Porcentaje Valores Nulos
connectiondropped_cs	99.9
connectionfailed_cs	99.9
csdurationtime	94.8
connectionok_cs	91.1
indoorconnection	75.9
outdoorconnection	71.9
rrcsetuptime	26.4
ecno	16.2
rscp	16.2
endlat	2.6
endlon	2.6

Una manera de analizar los valores nulos es buscando la correlación entre valores faltantes. La Figura E.1 muestra la correlación tanto positiva o negativa de cada par de variables

con valores nulos. Se puede notar una alta correlación de nulidad entre *csdurationtime* y *connectionok_cs* que tiene relación con las llamadas, pues si no existe un *connectionok_cs* entonces *csdurationtime*, que se refiere a la duración de llamada será nulo. Lo mismo ocurre entre *rrcsetuptime*, *ecno* y *rscp*, lo que parece indicar que cuando se genera un valor en la variable *rrcsetuptime* que tiene relación al tiempo en que se establece la llamada no se registran los indicadores de calidad y cobertura.

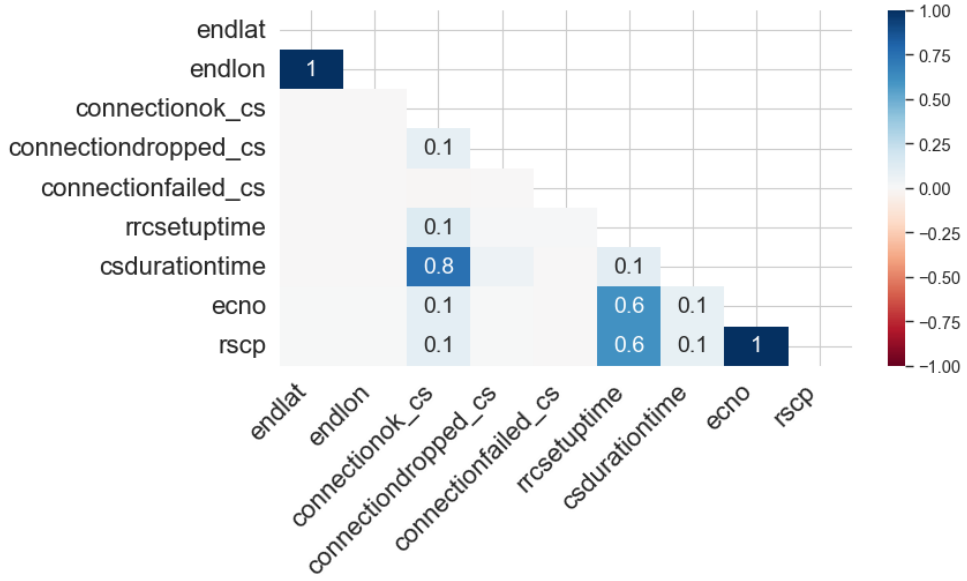


Figura E.1: Correlación Valores Nulos Actix 3G

Anexo F. Variables para caracterizar clientes

3G/4G - Número de sesiones 3G/4G: Conteo de sesiones en la ventana estadística de 30 días.

3G/4G - Promedio tiempo llamadas 3G/4G: Promedio del tiempo de llamadas en cada tecnología en la ventana estadística de 30 días.

3G/4G - Desviación estándar tiempo llamadas 3G/4G: Desviación estándar del tiempo de llamadas en cada tecnología en la ventana estadística de 30 días .

3G/4G - Tiempo total llamadas 3G/4G: Suma total de la duración de llamadas en cada tecnología en la ventana estadística de 30 días .

3G/4G - Cantidad llamadas 3G/4G: Conteo de llamadas exitosas en cada tecnología en la ventana estadística de 30 días .

3G/4G - Cantidad llamadas fallidas 3G/4G: Conteo de llamadas fallidas en cada tecnología en la ventana estadística de 30 días .

3G/4G - Cantidad llamadas caídas 3G: Conteo de llamadas caidas en cada tecnología en la ventana estadística de 30 días .

3G/4G - Score 3G/4G: Score de Calidad y Cobertura para cada tecnología en la ventana estadística de 30 días.

3G/4G - Tráfico 3G/4G: Tráfico en gigabytes en cada tecnología en la ventana estadística de 30 días.

3G - Promedio RRC Setup Time 3G: Promedio RCC Setup Time en tecnología 3G en la ventana estadística de 30 días.

3G - Desviación estándar RRC Setup Time 3G: Desviación estándar RCC Setup Time en tecnología 3G en la ventana estadística de 30 días.

3G - Promedio duración conexiones 3G: Promedio duración de sesiones en tecnología 3G en la ventana estadística de 30 días.

3G - Desviación estándar duración conexiones 3G: Desviación estándar duración de sesiones en tecnología 3G en la ventana estadística de 30 días.

3G - Cantidad conexiones interior 3G: Conteo de sesiones categorizadas como interior en tecnología 3G en la ventana estadística de 30 días.

3G - Cantidad conexiones exterior 3G: Conteo de sesiones categorizadas como exterior en tecnología 3G en la ventana estadística de 30 días.

Anexo G. Modelos testeados para clasificar clientes detractores

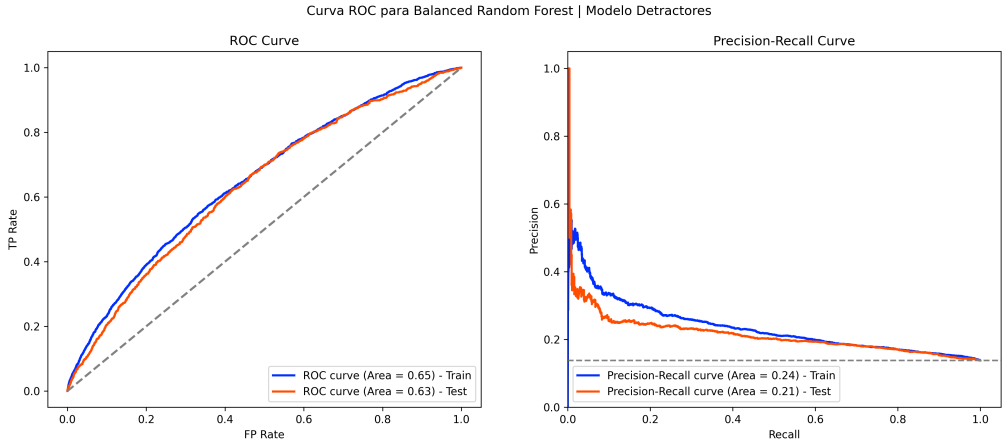


Figura G.1: Balanced Random Forest | Modelo detractores

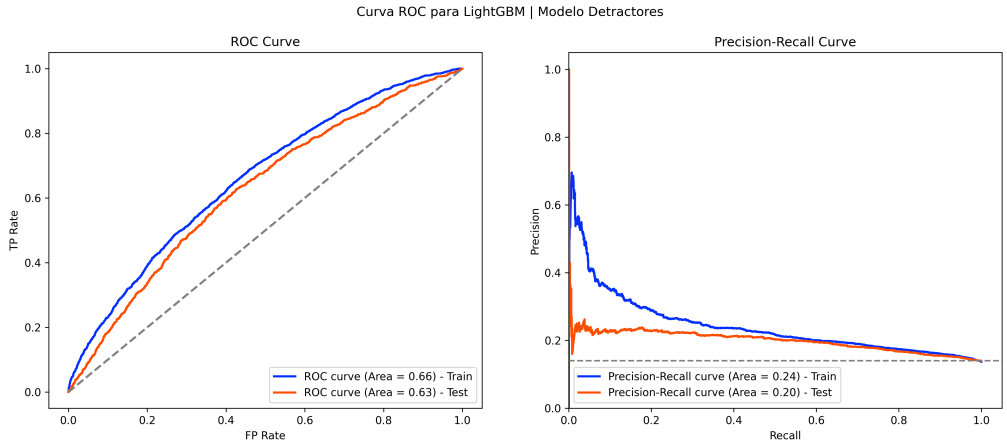


Figura G.2: LightGBM | Modelo detractores

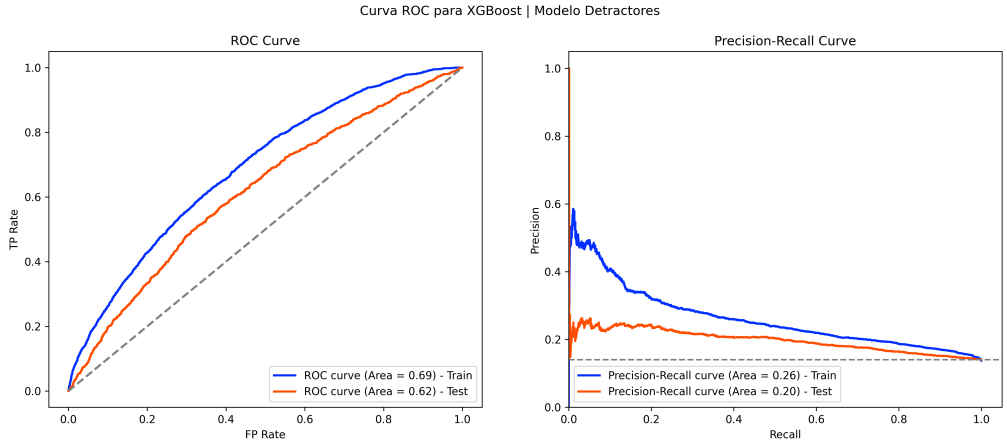


Figura G.3: XGBoost | Modelo detractores

Anexo H. Modelos testeados para clasificar clientes neutros

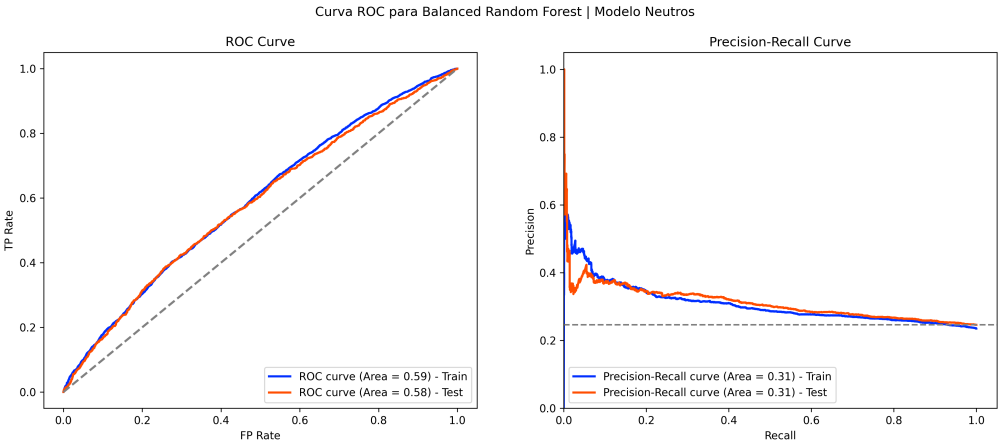


Figura H.1: Balanced Random Forest | Modelo Neutros

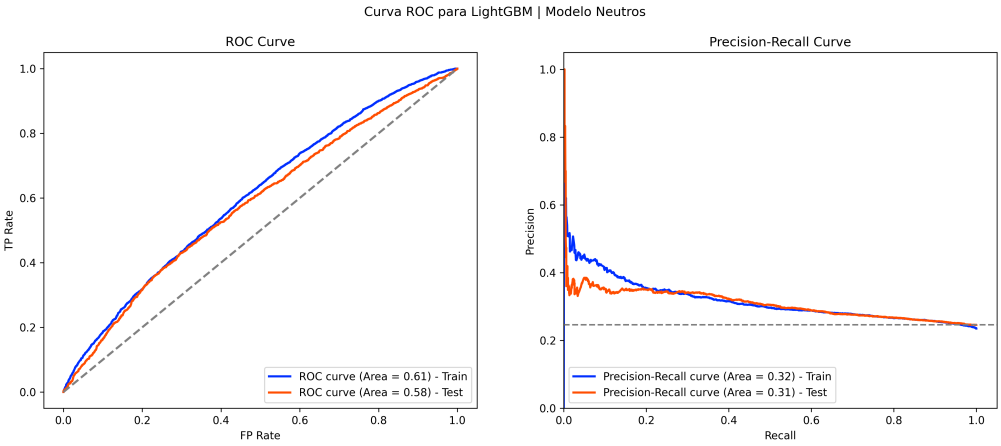


Figura H.2: LightGBM | Modelo Neutros

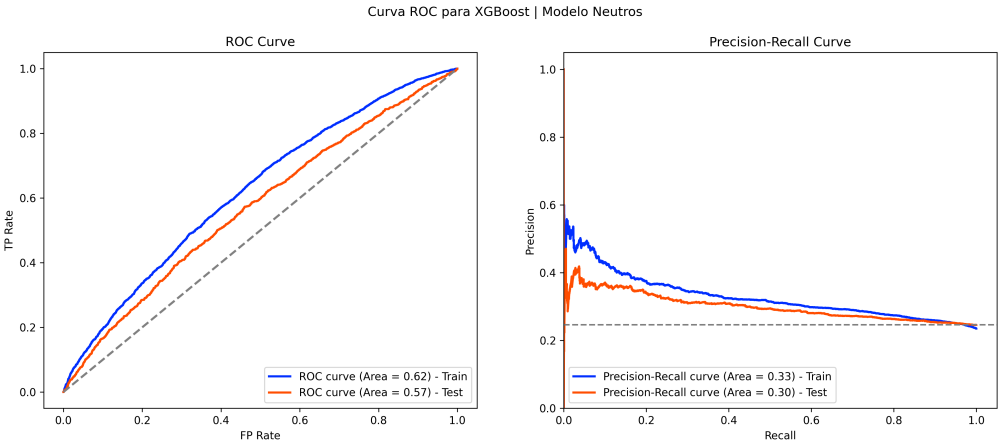


Figura H.3: XGBoost | Modelo Neutros

Anexo I. Análisis Ganancia Lift

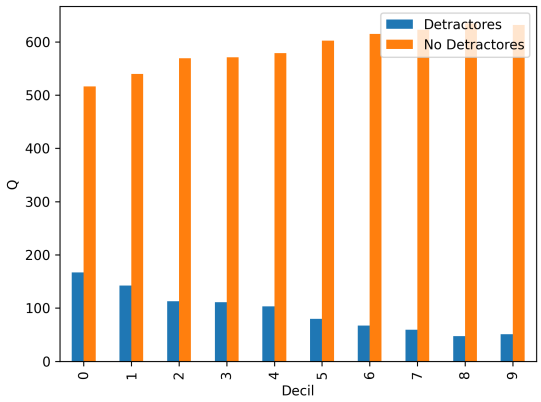


Figura I.1: Detractores y No detractores | Modelo Detractores

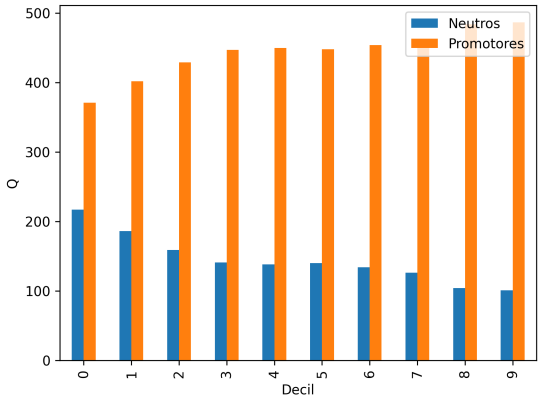


Figura I.2: Neutros y Promotores | Modelo Neutros