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

APLICACIONES DE DATA SCIENCE PARA LA MEJORA DE LA MEDICIÓN Y COBRO DE LA DISTRIBUCIÓN DE LA ENERGÍA ELÉCTRICA EN CONTEXTOS DE PANDEMIA MUNDIAL

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

BRAULIO MELQUÍADES MELLADO LEAL

PROFESOR GUÍA:
SEBASTIÁN RÍOS PÉREZ

MIEMBROS DE LA COMISIÓN:
RODRIGO VERSCHAE TANNENBAUM
MARCEL GOIC FIGUEROA

SANTIAGO DE CHILE
2021

RESUMEN DE LA MEMORIA PARA OPTAR AL TÍTULO DE: Ingeniero Civil Industrial
POR: Braulio Melquíades Mellado Leal
FECHA: 17/08/2021
PROFESOR GUIA: Sebastián Ríos Pérez

APLICACIONES DE DATA SCIENCE PARA LA MEJORA DE LA MEDICIÓN Y COBRO DE LA DISTRIBUCIÓN DE LA ENERGÍA ELÉCTRICA EN CONTEXTOS DE PANDEMIA MUNDIAL

El presente trabajo de investigación muestra cómo se abordaron cuatro problemas asociados a la dificultad para medir el consumo y cobrar a sus clientes la distribución eléctrica realizada por Enel Distribución S.A. en la región metropolitana.

El primer problema abordado corresponde al de la falta de precisión al momento de cobrar a los clientes cuyos consumos registrados en medidores convencionales no se han podido registrar en un mes por parte de los técnicos lectores, considerando lectura mensual. Actualmente, cuando esto sucede se utiliza una estimación basada en el promedio móvil de seis meses anteriores al mes de imposibilidad de lectura. Se demuestra que esta estimación es deficiente y se determina que, utilizando un modelo XGBoost Regressor, se rebajaría el MAPE asociado de un 39,04% a un 28,92%. Este resultado puede potencialmente influenciar políticas públicas que involucren cambios a nivel nacional.

El segundo problema corresponde a la alta cantidad de clientes cuyos consumos registrados de medidores convencionales no han podido ser registrados por parte de la empresa durante 3 meses o más, debido a la pandemia COVID-19 desde abril de 2020. El número de estos clientes es de alrededor de 26.035. Se contribuye a bajar este número a alrededor de cinco mil, para lo cual se utilizan tableros de visualización y cinco modelos de aprendizaje automático supervisado que utilizan la salida del algoritmo OPTICS sobre la georreferenciación de las viviendas como entrada. Con esto se logra una métrica de F1 de 61%. Como tercer problema, se abordan las viviendas sin lectura realizada por parte de la empresa, durante dos a cuatro meses, debido a la pandemia COVID-19 y el cambio de empresas subcontratadas para la lectura de medidores, desde diciembre de 2020. Para su resolución se aplica lo realizado en el segundo problema adaptándolo a la disponibilidad de datos y contexto de la situación, logando una métrica F1 promedio de 60%.

Por último, se aborda, de forma general, el problema del alto número de reclamos asociados a facturación en la compañía, cuya magnitud ascendía a 1.492.495 desde abril a octubre de 2020. Para lo anterior, se utilizó un enfoque predictivo basado en modelos de aprendizaje automático supervisado, considerando clasificación binaria. Se logró una precisión del 93%. Aun así, no se recomienda su aplicación directa, ya que, debido a la ley 21.340, que flexibiliza el pago por parte de los clientes, pudiendo sesgar de sobremanera los resultados a la fecha de utilización.

Con lo anterior señalado, se concluye que el resultado de este trabajo apoyaría, teóricamente, a que la empresa realice una mejor lectura en tiempos críticos provocados por la pandemia COVID-19. Si esta lectura, por un motivo exógeno, no puede ser buena, que pueda cobrar mejor el consumo correspondiente, aunque no se pueda registrar. Y si, por otro lado, hay errores de facturación por cualquier otro motivo, pueda anticiparse a los reclamos de los clientes, generando una mejor imagen para la empresa y tranquilidad para la población.

TABLA DE CONTENIDO

CAPÍTULO 1: CONTEXTO Y PRESENTACIÓN DEL TEMA

1.1 ANTECEDENTES GENERALES

1.1.1 Caracterización del cliente

El tema de investigación es realizado en conjunto con el CEINE (Centro de Investigación en Inteligencia de Negocios) de la Universidad de Chile. El cliente asociado es Enel Distribución Chile S.A., que esta descrito en este documento.

El rubro de la empresa corresponde a la distribución de la energía eléctrica, que corresponde a la etapa final en el suministro de electricidad a los usuarios finales. La misión declarada de la organización en su memoria del año 2019 corresponde a la (Enel Distribución, 2019), expresada en 5 postulados.

- Abrimos el acceso a la energía a más personas
- Abrimos el mundo de la energía a la nueva tecnología
- Nos abrimos al nuevo uso de la energía
- Nos abrimos a las nuevas formas de gestionar la energía para la gente
- Nos abrimos a nuevas alianzas

Por otro lado, la visión de la empresa está orientada a destacar el paradigma "Open Power" de Enel, que hace referencia a definir valores que aumenten los comportamientos y valores destinados a aumentar la implicación y participación de las personas que trabajan en Enel, para aumentar la comunicación dentro de la empresa e incrementar la generación de soluciones a los problemas (Grupo Enel, 2017). La visión de la empresa se declara como sigue.

"Open Power para resolver algunos de los más grandes retos de nuestro mundo"

Los principales valores de la empresa se definen como: confianza, proactividad, responsabilidad e innovación.

El servicio que ofrece la empresa es el de suministro de energía eléctrica tanto a particulares como empresas. Los clientes asociados pueden ser particulares, correspondientes a los distintos hogares donde la empresa posee distribución de energía, empresas y organismos públicos (para la iluminación de las luminarias públicas, por ejemplo).

Enel Distribución corresponde a la empresa de distribución de energía eléctrica más grande de Chile. La empresa opera en un área de concesión superior a los 2.105 Km2, bajo una concesión indefinida otorgada por el Gobierno de Chile, transmitiendo y distribuyendo electricidad en 33 comunas de la Región Metropolitana que incluyen las zonas de las subsidiarias Empresa Eléctrica de Colina Ltda. y Empresa de Transmisión Chena S.A. Durante 2019, la energía distribuida a través de las redes Enel Distribución Chile totalizó 17.107 GWh, lo que representa un 44% de las ventas de las distribuidoras a nivel nacional (Grupo Enel, 2019). El número de clientes se estima en casi 2 millones, según el sitio web de la descripción de Enel y la contraparte en la empresa. En el Anexo A se detallan las comunas en donde Enel Distribución Chile presta servicios.

La una ventaja competitiva relevante de la empresa es la pertenencia al grupo Enel, que cuenta con presencia en 32 países (Grupo Enel, 2020) con más de 70 millones de clientes, lo que permite

disponer de un mayor grupo de apoyo al generar proyectos, además de la capacidad de aprovechar el conocimiento adquirido por sus operaciones en distintos países, que se muestran en la ilustración 1. Adicionalmente, se asegura la generación y transmisión de la energía al ser producidas por empresas del mismo grupo.

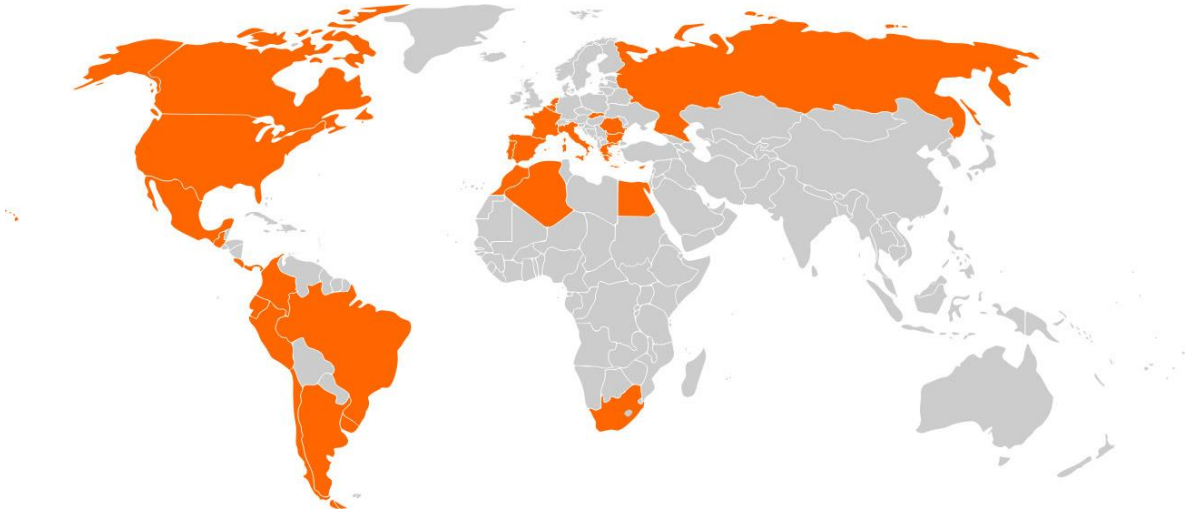

Ilustración 1: Presencia del Grupo Enel en el mundo

1.1.2 Mercado y marco institucional

Los principales actores asociados son los clientes, las empresas distribuidoras, transmisoras y los organismos reguladores.

La actual legislación eléctrica de Chile organiza el mercado eléctrico nacional en un sistema en el que las empresas privadas se encargan de prestar los servicios de electricidad. Estas empresas funcionan en mercados competitivos y no competitivos, pero todas cumplen con una regulación de precios y de calidad (CGE, 2020).

Las empresas generadoras de energía corresponden a las que transforman las fuentes de energía primaria (proveniente de los flujos de agua, en materiales, desde la energía solar, entre otras) en energía eléctrica transportable. En este segmento, la competencia y la existencia de diferentes actores es legal, aunque la ley faculta a la autoridad para obligar la interconexión de las instalaciones eléctricas, y así asegurar un sistema eficiente y seguro para todos.
La coordinación del sistema de generación está a cargo del Centro de Despacho Económico de Carga (CDEC), que determina planes de operación, líneas de transmisión y subestaciones de poder del sistema, para garantizar que el suministro sea seguro, al menor costo posible y que llegue a todas las personas. Los generadores enfrentan demandas que provienen de tres mercados básicos: Empresas Concesionarias de Distribución, Clientes no Sujetos a Fijación de Precios y Otros Generadores.

Por su parte, las empresas transmisoras de energía eléctrica se encargan del transporte de la energía desde los puntos en donde se genera hasta los centros de consumo masivo a través instalaciones de transmisión, que son las líneas y subestaciones de transformación que operan en tensión nominal (tensión eléctrica de trabajo para la cual fue diseñado un artefacto eléctrico) superior a 23 kilovoltios (kV). Las economías de este servicio operan como monopolio, ya que la inversión debe

ser única. Es por eso por lo que la legislación eléctrica lo define como un segmento regulado en el sistema. Los propietarios de sistemas de transmisión establecidos, como concesionarios de líneas de transporte de energía –cableado eléctrico- o cuyas instalaciones usen bienes nacionales de uso público (como, por ejemplo, cualquier espacio de la vía pública) deben permitir el paso de la energía a los interesados en transportarla. A cambio, estos deben usar estas instalaciones pagan al propietario a través de peajes, con precios unitarios de energía y potencia transportada.

Las empresas distribuidoras de energía eléctrica, como lo es, Enel Distribución Chile, deben suministrar energía eléctrica a sus clientes, para lo cual ellos cancelan su consumo de forma mensual, que es medido a través de medidores de energía eléctrica. La transparencia es clave en esta relación ya que todas las empresas del rubro deben cobrar exactamente lo que el usuario consumió, para no caer en ambigüedades. En la situación actual se identifica que se pueden producir reclamos asociados al cobro por parte de los clientes, o bien, fraudes por parte de estos. Como se mencionó en la sección anterior, Enel Distribución representa un 44% de las ventas, lo que implica que es el distribuidor de energía eléctrica con mayores ingresos de Chile, por lo que el presente estudio se realiza con una muestra considerable a nivel nacional.

Con respecto a las más importantes regulaciones de las empresas distribuidoras, estas se encuentran en el reglamento general de servicios eléctricos del Ministerio de Minería (publicado en el año 1998). Las regulaciones más importantes para este trabajo de título son las siguientes:

- Los concesionarios de servicio público de distribución deberán facturar en base a las cantidades que consten en el equipo que registra los consumos del usuario. Aun así, se podrá estimar el consumo cuando no se pueda medir correctamente
- La facturación de los consumos debe hacerse de manera mensual o bimestral
- El usuario tiene la responsabilidad de que el técnico de medición pueda constatar el consumo de energía eléctrica mensual. En caso de no ser posible, la empresa puede realizar un cobro por hasta dos meses utilizando el promedio móvil de los seis meses anteriores.

En el anexo B se pueden observar los artículos asociados a las regulaciones expuestas.

Las tendencias del mercado apuntan a varios focos (Enel Distribución, 2020)**,** los principales son:

- La medición inteligente, mediante nuevos medidores que permitan generar tarifas diferencias por horario y la medición mensual de forma automática
- La autogeneración de energía por parte de los usuarios, que involucra nuevos desafíos con respecto a la medición y cobro de energía eléctrica
- La contribución a la descontaminación de las ciudades, mediante soluciones de eficiencia en el consumo, además de implementación de transporte y luminarias de bajo consumo y eficientes.

Dado el tamaño de las ventas de la empresa y su presencia internacional, se define como líder, poseyendo gran influencia en las nuevas tendencias en la distribución de energía eléctrica a través de sus distintas iniciativas, como las primeras instalaciones de medidores inteligentes (Revistaei, 2016) utilizando un proveedor de medidores que es parte del Grupo Enel, liderando la tendencia a utilizar nuevas formas de medición de energía eléctrica en dirección al aumento de la eficiencia en el consumo y cobro.

1.1.2.1 Regulaciones adicionales a causa de la pandemia COVID-19

El 5 de agosto de 2020 se promulga la ley 21.249 por parte del poder legislativo. Esta ley implica que para las empresas de distribución eléctrica no pueden realizar cortes de suministros por mora en el pago a un número considerable de segmentos de personas e instituciones, hasta 260 días de publicación de la ley. Adicionalmente, deben dar facilidades para pagar los montos adeudados, como, por ejemplo, cancelar las deudas en 48 cuotas sin interés. El texto completo de esta ley se encuentra en el anexo C.

Una extensión a dicha ley corresponda a la n° 21.340 promulgada el 13 de mayo de 2021, que extiende el plazo de no corte de suministro hasta el 31 de diciembre de 2021. Es de vital importancia comprender esta ley y sus implicancias en este trabajo, ya que puede producir cambios en el comportamiento de pago y de reclamo de los clientes.

1.1.3 Desempeño de la organización

El crecimiento de la organización se ve representado en la siguiente tabla, con información extraída desde los informes de sostenibilidad de 2018 y 2019 de la empresa.

Tabla 1 Indicadores de crecimiento de la empresa

Ámbito	Valor 2018	Valor 2019	% de Crecimiento
Número de clientes	1.924.984	1.972.218	2,39%
Ventas de energía (GWh)	16.782	17.107	1,89%
EBIDTA (Millones de pesos)	200.614	221.607	9,47%
Número de colaboradores	681	743	8,34%

Fuente: Elaboración Propia

Se vislumbra con estas cifras que la empresa está en pleno crecimiento, observando el aumento del EBIDTA y el número de colaboradores. Aun así, cabe destacar que las pérdidas de energía han disminuido de manera despreciable (0,03%) entre los años 2018 y 2019. A pesar de lo anterior, dichas pérdidas han aumentado hasta un 5,25% en el primer trimestre de 2020 (Enel Chile, 2020), desde el 5,02% calculado a fines del 2019, lo que justifica que sea uno de los focos de la empresa en la actualidad el disminuir está métrica.

1.2 DESCRIPCIÓN DEL PROYECTO Y JUSTIFICACIÓN

1.2.1 Información del área

El área de trabajo corresponde a la de control de energía, medición y balance eléctrico de la empresa. Esta área tiene como fin el asegurar una correcta facturación a todos los clientes de la compañía, además de accionar el control de pérdidas de energía y la estimación del balance eléctrico para maximizar el margen de compra y venta de la compañía.

El alcance que tiene el área corresponde a:

- Administrar, de acuerdo con las instrucciones provenientes de las líneas de negocios, las actividades del balance energético, medición y recuperación de energía.
- Realizar la adquisición, certificación y almacenamiento de las medidas
- Definir los planes de verificación de los medidores, para realizar análisis asociados a fraudes y diversas actividades de minería de datos. Emitir las inspecciones y monitorear su

progreso. Realizar reconstrucciones del consumo de energía y asegurar el cobro de los procesos de recuperación de energía.
- Asegurar la ejecución de iniciativas con respecto a las pérdidas de energía no técnicas.

Según relata la contraparte, el área se compone de 25 trabajadores y se divide en 3 unidades: lectura, medición y balance. Aproximadamente un 50% de los trabajadores corresponden a ingenieros senior y especialistas (en su mayoría ingenieros civiles eléctricos e industriales). El otro 50% corresponde a personal de back office y técnicos eléctricos que trabajan en terreno. Cabe destacar que la documentación interna de la empresa con las estructuras de las áreas detalladas está en trámite de ser conseguida.

Con el tema de memoria se verán impactados los clientes de la empresa, al tener cobros más acertados y atención a sus reclamos de manera más veloz debido a la liberación de los recursos. También se verán impactados los colaboradores del área, al hacer más eficiente su trabajo disminuyendo tareas evitables. Adicionalmente, según declara la contraparte, estas ambigüedades en los cobros generan una mala imagen de la empresa debido a comentarios de clientes por distintas vías, que, con el trabajo a realizar, podrían verse potencialmente aminorados.

1.2.2 Problema a abordar y análisis causal

La contraparte manifiesta que existen alrededor de 1.300.000 clientes que utilizan medidores convencionales, vale decir, cuya medición se realiza de forma pedestre, para lo cual un funcionario de la empresa debe realizar la lectura a los clientes de manera mensual. Por otro lado, existen alrededor de 600.000 medidores telemedidos, de los cuales alrededor de 300.000 son Smart Meters, vale decir, incorporan la última tecnología introducida en el año 2016 por la empresa, el resto de los medidores corresponden a una tecnología que está fuera de uso y no permiten las ventajas de los Smart Meters, como el cobro según distintos horarios.

Para efectos de este trabajo de memoria se considerarán los clientes asociados a medidores convencionales (que deben ser leídos de forma pedestre para el cobro del consumo).

Cabe destacar que los clientes pueden cambiar sus medidores por Smart Meters de manera voluntaria, pero estos generan un gran rechazo en la población, debido a que existen sospechas por parte de los medios de que la intención detrás del uso de estos medidores es asegurar ganancias para las empresas distribuidoras (CIPER, 2019). Adicionalmente, el proyecto de ley presentado por el gobierno para cambiar estos medidores a todos los clientes pedestres en un plazo determinado se rechazó de manera categórica por parte del parlamento y la ciudadanía.

Las problemáticas por abordar serán cuatro y estarán descritas de forma específica en cada uno de los siguientes cuatro capítulos. A modo general, serán las siguientes:

1. La imprecisión en el cobro que se produce cuando los consumos de una vivienda no pueden ser registrados durante un mes, ya que actualmente se utiliza un promedio móvil simple de 6 meses para predecir dicho consumo, lo que no ha mostrado eficacia históricamente. Este problema quedó al descubierto debido a la pandemia COVID-19 y es de vital importancia una pronta solución, ya que se actúa en contextos donde se dificulta de sobremanera el trabajo de los técnicos lectores.
2. La cantidad considerable de viviendas cuyos consumos no han podido ser registrados en cuatro meses o más debido a la pandemia COVID-19, a fecha de octubre de 2020

3. La cantidad considerable de viviendas cuyos consumos no han podido ser registrados en un plazo de tiempo de dos a cuatro meses debido al cambio de licitación de empresas contratistas encargadas de la lectura, a fecha de principios de abril de 2021. Asimismo, se debe mejorar el problema de productividad asociado.
4. El potencial alto número de reclamos que se pueden producir en invierno de 2021 por motivos de facturación.

1.3 OBJETIVOS

1.3.1 Objetivo general

El objetivo general se enuncia como sigue.

"Aumentar la eficacia y eficiencia de los macroprocesos de medición y cobro de energía eléctrica de la organización, introduciendo herramientas analíticas y modelos predictivos en sus lógicas de negocio que permitan mitigar los efectos de la pandemia COVID-19"

Cabe destacar que se entiende por eficiencia como el rendimiento de los recursos utilizados por la empresa y eficacia a la consecución de objetivos de los macroprocesos.

1.3.2 Objetivos específicos

- Proponer un nuevo modelo de cobro a todos los clientes a nivel nacional y demostrar que es sustancialmente más preciso que el utilizado actualmente
- Construir herramientas analíticas y modelos predictivos que permitan aumentar la tasa de lectura de las viviendas que no han podido ser leídas a causa de la pandemia en curso
- Integrar modelos de aprendizaje automático supervisado y no supervisado en una herramienta que permita aumentar la productividad asociada al proceso de lectura en terreno
- Construir y evaluar modelos predictivos que permitan determinar la probabilidad de reclamo debido a facturación por parte de los clientes

1.4 MARCO CONCEPTUAL

A continuación, se describirá el estado del arte con respecto a las principales herramientas potenciales a utilizar para la construcción de los modelos matemáticos mencionados anteriormente durante el trabajo de memoria. Estas se pueden agrupar en dos grandes tipos: clasificación y clustering.

(Para abordar el trabajo de memoria en general, se utilizará el libro "Introduction to Machine Learning With Pyhon", de los autores Andreas Müller y Sarah Guido, debido a que este libro cumple con las exigencias y mezcla tanto la base teórica del tema como la aplicación en programación.)

1.4.1 Algoritmos de aprendizaje supervisado

Existen dos tipos de estos algoritmos: de clasificación y regresión. En clasificación el objetivo es predecir una clase a la que pertenece un registro desde una lista de posibilidades predefinidas. Esta puede ser binaria (solo dos clases) o multiclase (más de dos clases). En regresión se pretende predecir un número continuo a partir de ciertos atributos predefinidos.

Los algoritmos de clasificación que se utilizan durante el trabajo de memoria son Random Forest, Support Vector Machines y XGBoost. Por otro lado, para tareas de regresión de utilizaran versiones diferenciadas de los mismos anteriores, además de Desicion Tree, Regresión Lineal (OLS) y K-Nearest Neigbors.

A continuación, se detallarán los algoritmos que generan mejores resultados en este trabajo, Random Forest y XGBoost.

1.4.1.3 Random Forest

El algoritmo Random Forest[1] (Breiman, 2001) es una técnica de aprendizaje supervisado que genera múltiples árboles de decisión sobre un conjunto de datos de entrenamiento: los resultados obtenidos se combinan a fin de obtener un modelo único más robusto en comparación con los resultados de cada árbol por separado (Lizares, 2017).

Cada árbol se obtiene mediante un proceso de dos etapas:

1. Se genera un número considerable de árboles de decisión con el conjunto de datos. Cada árbol contiene un subconjunto aleatorio de variables m (predictores) de forma que m < M (donde M = total de predictores).-
2. Cada árbol crece hasta su máxima extensión.

Cada árbol generado por el algoritmo Random Forest contiene un grupo de observaciones aleatorias (elegidas mediante bootstrap, que es una técnica estadística para obtener muestras de una población donde una observación se puede considerar en más de una muestra).

Las observaciones no estimadas en los árboles (también conocidas como “out of the bag”) se utilizan para validar el modelo. Las salidas de todos los árboles se combinan en una salida final Y (conocida como ensamblado) que se obtiene mediante alguna regla (generalmente el promedio, cuando las salidas de los árboles del ensamblado son numéricas y, conteo de votos, cuando las salidas de los árboles del ensamblado son categóricas). Lo anterior se muestra gráficamente en la siguiente figura (Espinosa-Zúñiga, 2020).

[1] Extraído de: https://doi.org/10.22201/fi.25940732e.2020.21.3.022

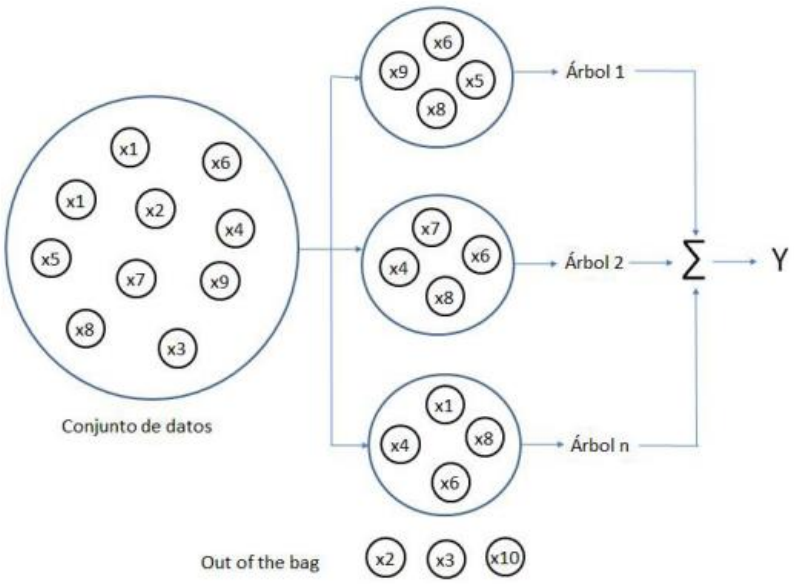

Ilustración 2 Funcionamiento Random Forest. Fuente: Espinosa-Zuñiga, 2020

Las principales ventajas del algoritmo Random Forest (Cánovas et al., 2017) son:

a) Pueden usarse para clasificación o predicción: En el primer caso, cada árbol "vota" por una clase y el resultado del modelo es la clase con mayor número de "votos" en todos los árboles, de forma que cada nueva observación se presenta a cada uno de los árboles y se asigna a la clase más "votada". En el segundo caso, el resultado del modelo es el promedio de las salidas de todos los árboles.
b) El modelo es más simple de entrenar en comparación con técnicas más complejas, pero con un rendimiento similar.
c) Tiene un desempeño muy eficiente y es una de las técnicas más certeras en bases de datos grandes.
d) Puede manejar cientos de predictores sin excluir ninguno y logra estimar cuáles son los predictores más importantes, es por ello por lo que esta técnica también se utiliza para reducción de dimensionalidad.
e) Mantiene su precisión con proporciones grandes de datos perdidos.

Por otra parte, sus principales desventajas son las siguientes:

a) La visualización gráfica de los resultados puede ser difícil de interpretar.
b) Puede sobre ajustar ciertos grupos de datos en presencia de ruido.

c) Las predicciones no son de naturaleza continua y no puede predecir más allá del rango de valores del conjunto de datos usado para entrenar el modelo. En el caso de predictores categóricos con diferente número de niveles, los resultados pueden sesgarse hacia los predictores con más niveles.
d) Se tiene poco control sobre lo que hace el modelo (en cierto sentido es como una caja negra).

Las ventajas de Random Forest hacen que se convierta en una técnica ampliamente utilizada en muchos campos, por ejemplo, teledetección (para clasificación de imágenes), bancos (para detección de fraudes y clasificación de clientes para otorgamiento de crédito), medicina (para analizar historiales clínicos a fin de identificar enfermedades potenciales en los pacientes), finanzas (para pronosticar comportamientos futuros de los mercados financieros) y comercio electrónico (para pronosticar si un cliente comprará, o no, cierto producto), entre otros.

1.4.1.4 XGBoost

El algoritmo XG Boost (Extreme Gradient Boosting)[2] es una técnica de aprendizaje supervisado (Chen y Guestrin, 2016) también basada en árboles de decisión y que es considerada el estado del arte en la evolución de estos algoritmos, como se muestra en la siguiente figura (Espinosa-Zúñiga, 2020).

Ilustración 3: Evolución de algortimos basados en decision trees. Fuente: Espinosa-Zuñiga, 2020

El algoritmo XG Boost tiene las siguientes características (Chen y Guestrin, 2016):

a) Consiste en un ensamblado secuencial de árboles de decisión (este ensamblado se conoce como CART, acrónimo de "Classification and Regression Trees"). Los árboles se agregan secuencialmente a fin de aprender del resultado de los árboles previos y corregir el error producido por los mismos, hasta que ya no se pueda corregir más dicho error (esto se conoce como "gradiente descendente" (Figura 3).
b) La principal diferencia entre los algoritmos XGBoost y Random Forest es que en el primero el usuario define la extensión de los árboles mientras que en el segundo los árboles crecen hasta su máxima extensión.
c) Utiliza procesamiento en paralelo, poda de árboles, manejo de valores perdidos y regularización (optimización que penaliza la complejidad de los modelos) para evitar en lo posible sobreajuste o sesgo del modelo.

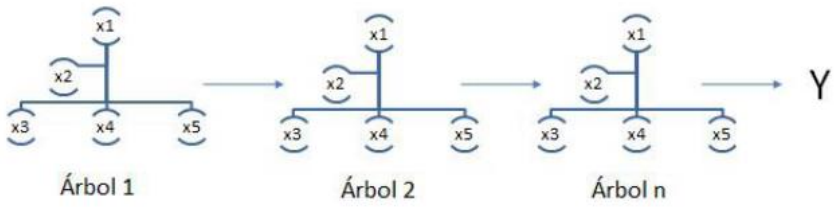

Ilustración 4: Funcionamiento algoritmo XGBoost. Fuente: Espinosa-Zuñiga, 2020

[2] Extraído de: https://doi.org/10.22201/fi.25940732e.2020.21.3.022

El algoritmo XGBoost funciona así:

a) Se obtiene un árbol inicial F0 para predecir la variable objetivo "y", el resultado se asocia con un residual $(y - F_0)$.
b) Se obtiene un nuevo árbol h_1 que ajusta al error del paso previo.
c) Los resultados de F_0 y h_1 se combinan para obtener el árbol h_1, donde el error cuadrático medio de F_1 será menor que el de F_0:

$$F_1(x) < -F_0(x) + h_1(x)$$

d) Este proceso se sigue iterativamente hasta que el error es minimizado lo más posible de la siguiente forma:

$$F_m(x) < -F_{m-1}(x) + h_m(x)$$

Las principales ventajas del algoritmo XGBoost son:

a) Puede manejar grandes bases de datos con múltiples variables.
b) Puede manejar valores perdidos.
c) Sus resultados son muy precisos.
d) Excelente velocidad de ejecución.

Por otra parte, sus principales desventajas son:

a) Puede consumir muchos recursos computacionales en grandes bases de datos, por lo que se recomienda antes de aplicar esta técnica en bases de este tipo, determinar cuáles son las variables que aportarán más información a fin de considerar solo dichas variables en la obtención del modelo.
b) Se deben ajustar correctamente los parámetros del algoritmo a fin de minimizar el error de precisión y evitar sobreajuste del modelo (lo que puede darse si se maneja un número muy grande de árboles).
c) Solo trabaja con vectores numéricos, por lo que se requieren convertir previamente los tipos de datos no numéricos a numéricos. Las ventajas de este algoritmo hacen que se aplique en campos como: identificación de huellas digitales (Luckner et al., 2017), seguridad vial (Bahador et al., 2020) y análisis de mercados financieros (Nobre y Ferreira, 2019), entre otros.

1.4.2 Algortimos de aprendizaje no supervisado

Corresponden a algoritmos donde la salida es desconocida. El algoritmo solo tiene de entrada los datos y se entrena en base a ciertos patrones implícitos para generar conocimiento a partir de los datos. La evaluación de este tipo de algoritmos es notablemente más inexacta.

El objetivo es encontrar grupos de objetos, en donde los objetos en un grupo seas similares o relacionales entre sí, y, además, que sean diferentes (o no relacionados) a los objetos de otros grupos. Por lo anterior, se utilizan cuando se desean dividir los datos en grupos que sean significativos o útiles o se desea capturar la estructura natural de los datos. En algunas ocasiones puede corresponder solo a un análisis exploratorio para obtener mayor conocimiento mediante el uso de otras técnicas.

Para el trabajo de investigación en particular se utilizará el algoritmo OPTICS (Ordering Points To Identify the Clustering Structure), que consiste en, dado un numero de muestras en cierta vecindad, que el usuario debe especificar, encontrar los núcleos de alta densidad presente.

1.4.3 Aplicaciones de psicología de la percepción en visualización de datos

La percepción corresponde a la organización e interpretación de la información que provee el ambiente, interpretación del estímulo como objeto significativo. Los hechos que dan origen a la percepción no están fuera de cada persona, sino en su sistema nervioso. (Dörr et al., 2008)

En este trabajo la psicología de la percepción juega un rol importante en la elaboración de tableros de visualización, donde se busca el enfoque de comunicar la mayor cantidad de información posible, en el menor tiempo posible y optimizando el espacio. Para esto se estudian dos aplicaciones específicas: el procesamiento pre-atencional (pre-attentive processing) y las leyes de la Gestalt.

1.4.3.1 Pre-attentive processing

Un primer resultado importante acerca de como se forma la percepción de los elementos visualizados es el descubrimiento de un conjunto limitado de propiedades visuales que son detectadas de forma muy rápida y exacta por el sistema visual de bajo nivel. Estas propiedades fueron inicialmente llamadas ‘preattentive’ dado que su detección parecía que precediera a la atención focalizada. (Ward et al., 2018)

Generalmente, estos preattentive attributes, son las mejores maneras de presentar datos debido a que se pueden visualizar patrones sin pensar o procesar. De hecho, estos atributos evolucionaron en los seres humanos como formas de evaluar rápidamente una situación, detectar un patrón y decidir rápidamente como reaccionar. (Tableau, 2021)

La implicancia de conocer estos elementos es que pueden ser utilizados en la elaboración de tableros de monitoreo para que el usuario pueda captar la información de forma más inmediata y tener la posibilidad de dirigir su atención donde se encuentra la información más importante para el contexto de su negocio.

Estos preattentive attributes son diez: largo, ancho, orientación, tamaño, forma, cerradura, posición, agrupación, matiz e intensidad del color. Se pueden apreciar en la siguiente ilustración.

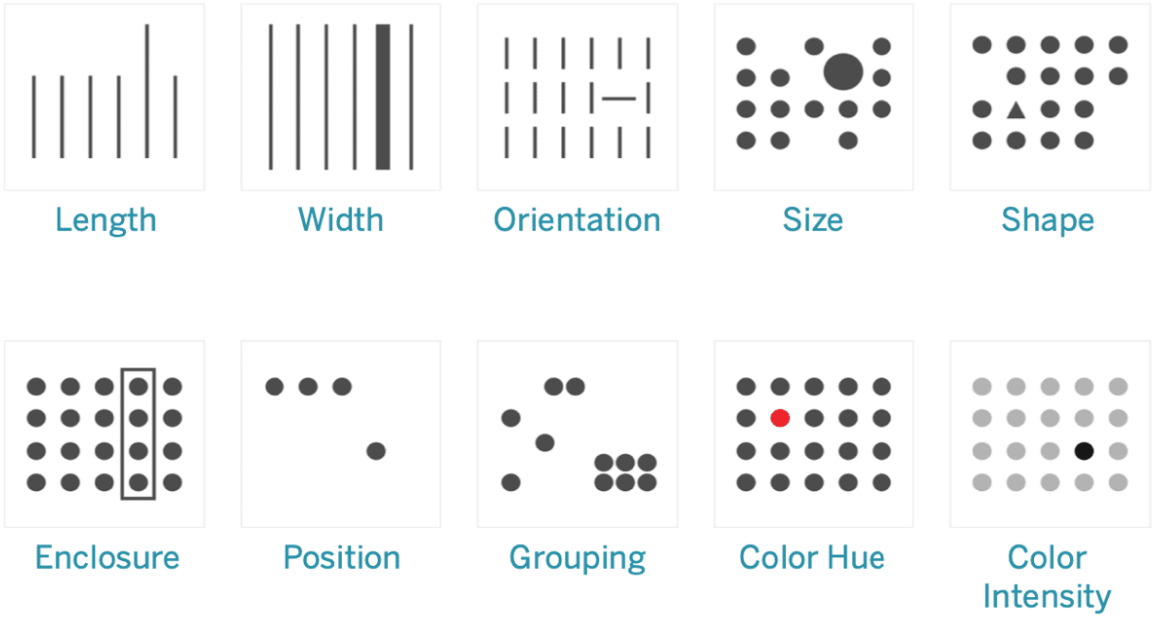

Ilustración 5: Preattentive attributes: largo (Length), ancho (Width), orientación (Orientation), tamaño (Size), forma (Shape), cerradura (Enclosure), posición (Position), agrupación (Grouping), matiz (Color Hue) e intensidad del color (Color Intensity). Fuente: Tableau e-learning, 2021

1.4.3.2 Leyes de la Gestalt

Las percepciones poseen un carácter integral de modo que no se las puede explicar como producto de una mera sucesión y yuxtaposición de simples sensaciones, los hechos son más complejos y en el conjunto de lo que se llama percepción también interviene de un modo más decisivo un factor más elevado que integra la heterogénea pluralidad especial y temporal de las distintas sensaciones delimitadas. A esto, la Teoría de la Gestalt (cuyos representantes más connotados son Wertheimer, Köhler y Koffka) lo denomina factor de forma o de la Gestalt. (Dörr et al., 2008)

En la siguiente ilustración se pueden apreciar las distintas aplicaciones de estas leyes. Para efectos de este trabajo, se utilizarán principalmente las leyes de proximidad y similitud.

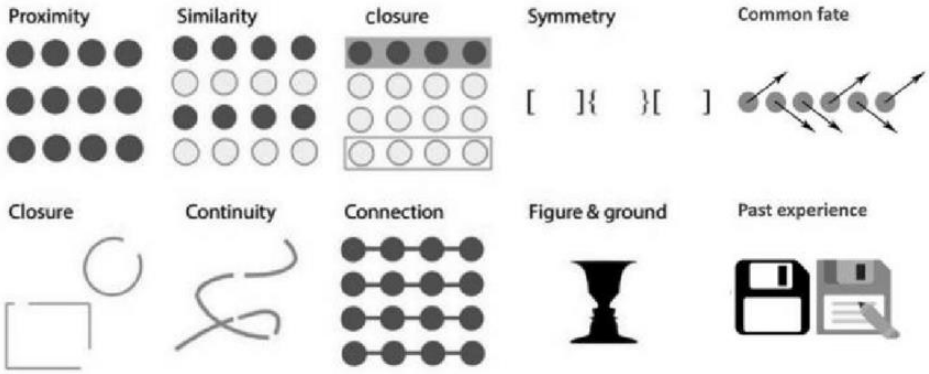

Ilustración 6: Resumen leyes de la Gestalt. Fuente: Yalcinkaya, 2018

La ley de la Gestalt de proximidad implica que cuando los objetos se despliegan uno cerca de otro, especialmente en la comparación con otros objetos, se tenderá a suponer que los objetos más cercanos son parte de un grupo. Esto incluye a las marcas en una visualización como las líneas en

un gráfico de línea o puntos en un gráfico de dispersión. También aplica para otros elementos de una visualización como las leyendas o etiquetas que aparecen cerca de las marcas.

Por otro lado, la ley de similaridad implica que la gente supone que una relación existe dentro de los objetos que tienen una apariencia similar, como objetos de una forma o color parecido. Por ejemplo, en un gráfico de dispersión que codifica las marcas con una forma o un matiz de color específico, los espectadores supondrán que las marcas que se ven similar a otras están relacionadas. (Tableau, 2021)

1.5 METODOLOGÍA

Se utiliza la metodología CRISP-DM, principalmente debido a su capacidad de distinguir entre los objetivos del negocio y de la minería de datos. Esta metodología aborda todo el flujo de trabajo dentro del ejercicio de la minería de datos, por lo que se utilizará durante todo el trabajo de memoria de título, desde la obtención y recolección de datos, hasta la fase de modelamiento. No se considera abordar la fase de evaluación ni implementación en este trabajo.

Adicionalmente, se incorporará una metodología de construcción de dashboards para generar este tipo de entregables a la empresa, que se incluirán en las fases de entendimiento del negocio e implementación respectivamente de los capítulos 3 y 4.

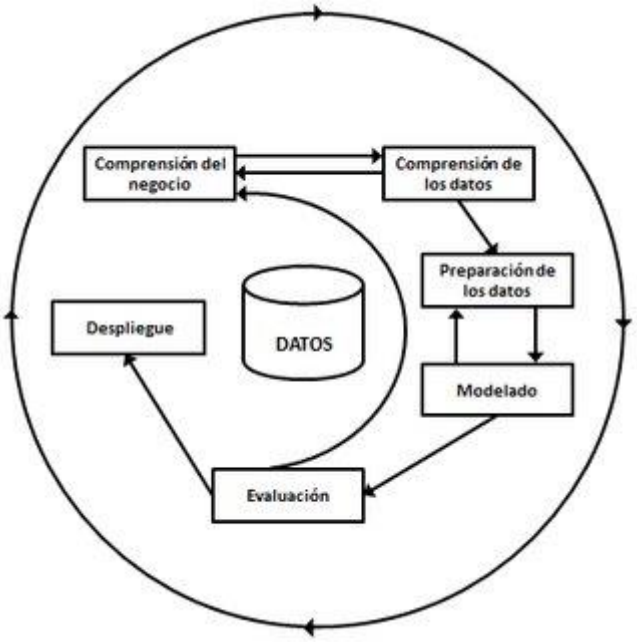

Ilustración 7 Flujo CRISP-DM

A continuación, se detallan todas las fases que tendrá la metodología.

1.5.1 Entendimiento del negocio

Para el capítulo dos, como se actuará en la lógica de negocio del proceso de cobro, se estudia la forma en que se formulan los cobros de energía eléctrica. Adicionalmente, se entienden los patrones de reclamos por parte de los clientes, considerando motivos y espacios temporales en los que son realizados. Se verifica que tan relacionados están estos patrones en la empresa Enel Distribución

S.A. y otras del rubro a nivel nacional, para entender el posible impacto del proyecto. Se clarifica la(s) métrica(s) objetivo a disminuir.

Para el trabajo a realizar en los capítulos tres y cuatro, se debe entender el proceso de lectura que realiza la empresa, lo que se realiza mediante entrevistas a trabajadores. Por otro lado, es necesario entender el contexto en que se abordan los problemas y las posibles acciones preventivas y correctivas que puede tomar la empresa en cada caso, para determinar las soluciones a generar. Finalmente, se debe identificar la métrica objetivo a maximizar con respecto a la minería de datos.

El trabajo realizado en el capítulo cinco incluye el entendimiento que se genera al trabajar en las partes anteriores. Además, se debe clarificar los objetivos del negocio para disminuir los reclamos y un segmento objetivo para el ajuste de los modelos.
Posteriormente, se definirán criterios de éxito asociados a la minería de datos, vale decir, los requisitos que deberán tener los resultados asociados a los modelos predictivos y los criterios de éxito a utilizar.

Finalmente, se debe reevaluar el plan de trabajo, utilizando los hallazgos encontrados en el desarrollo de esta fase.

1.5.2 Entendimiento de los datos

Para abordar el trabajo realizado en el capítulo dos, se deben comprender los distintos patrones de consumo de los clientes y buscar las relaciones con los reclamos por facturación recibidos. Se debe entender el origen de los valores nulos y duplicados dentro de los consumos de los clientes.

Para los capítulos tres, cuatro y cinco, se realizan análisis descriptivos para entender las distintas relaciones entre las variables a incluir en los modelos. Esto incluye estudios de correlación y test chi-cuadrado. Adicionalmente, en esta fase se construyen tableros de visualización utilizando la metodología expuesta en la sección 1.5.5, los que, además de ser entregados a la empresa para mejorar su operación, son utilizados para generar un adecuado entendimiento del problema.

1.5.3 Preparación de los datos

Para abordar el trabajo del capítulo dos, en esta fase se forma la matriz de consumos históricos de los clientes, para lo cual se deben tratar consumos duplicados y nulos, además de realizar los filtros para cada segmento objetivo, por ejemplo, no incluir clientes comerciales.

Para los capítulos tres, cuatro y cinco, a partir del entendimiento generado en la fase anterior, se construyen los modelos relacionales a partir de los datos tabulares extraídos, para poder facilitar su posterior modelamiento. Se procede a la limpieza de los datos, incluyendo valores, nulos, duplicados y errores de formato en los distintos atributos. En particular, para los modelos de lectura, se deberá asegurar la disponibilidad de coordenadas geográficas decimales, para lo cual se pueden utilizar un conjunto de métodos para su obtención.

1.5.4 Modelamiento

Para esta fase se podrá realizar más de una iteración, para así, poder refinar los resultados, debido a la naturaleza de la data y los análisis a realizar.

Se trabaja, primero, sobre el modelamiento asociado a los cobros utilizando los datos de la matriz de consumos formulada en la parte anterior. El foco es determinar un mejor modelo de cobro, vale decir, que el cobro estimado para cuando la lectura del consumo no sea posible se asemeje lo más posible al correspondiente al consumo real del cliente, para disminuir el número de reclamos asociado a ese tópico. Se deben seleccionar los criterios y modelos correctos para cumplir con esta tarea, ejecutar los distintos modelos y evaluarlos utilizando las métricas seleccionadas en la primera fase de la metodología.

Posteriormente, con respecto a los modelos asociados al proceso de medición, se utilizan lógicas no supervisadas para utilizarlas como input en los modelos de clasificación si corresponde. Finalmente, se deben seleccionar las variables y definir correctamente los modelos a utilizar en base a la dimensionalidad de los datos y su tipo. Finalmente, se optimizan los hiperparámetros utilizando técnicas a explicitar y se ejecutan los modelos para proceder a evaluarlos.

1.5.5 Contrucción de dasboards

A partir del conocimiento obtenido, se determinan los cambios en los flujos de ambos procesos asociados. Se construyen indicadores para la mejora continua de los procesos y, además, se construyen un panel de visualización siguiendo la metodología propuesta por el curso "Dashboard Design: Visual Best Practices" de Tableau E-Learning que se compone de las siguientes fases:

1. **Determinar audiencia y propósito:** se determinará la audiencia objetivo y el propósito del panel en consecuencia, lo que definirá los atributos y visualizaciones a incluir.
2. **Planificar el dashboard:** se explicita el mapeo visual y las vistas asociadas. Se realiza un bosquejo del dashboard indicando todos los elementos que contendrán y los filtros asociados.
3. **Construir el dashboard siguiendo las buenas prácticas:** con lo obtenido en la parte 2, se construye el panel de visualización utilizando un software BI que lo permita y se adapte correctamente a la empresa y su flujo de trabajo. Se siguen buenas prácticas relativas a los fundamentos de la visualización humana (leyes de la Gestalt)
4. **Testear:** se utiliza la regla de los 5 segundos directamente en la audiencia, vale decir, se muestra el panel a integrantes del grupo objetivo y se verifica que en cinco segundos puedan captar el propósito del dashboard y percibir su funcionalidad. Si es que no se cumple esta regla, se vuelve nuevamente a la primera fase.

1.6 ALCANCES

Con respecto al tipo de medidores a abordar, se define como alcance a los medidores convencionales. Vale decir, el impacto del proyecto puede alcanzar a alrededor 1.450.000 clientes de la empresa que utilizan este tipo de medidores. Este alcance se fija debido a la misma naturaleza de los problemas, que involucran lectura en terreno.

El trabajo de memoria no abordará el seguimiento de los resultados en la empresa y el cumplimiento de los criterios de negocio, debido al tiempo disponible para realizar el trabajo.

Se abordarán todos los clientes donde se tengan datos de consumo por al menos tres años, para poder obtener patrones de consumo asociados.

Se abordarán solo clientes residenciales para los modelos asociados al proceso de cobro, en acuerdo con la contraparte. Los modelos asociados al proceso de lectura abordarán también clientes comerciales.

1.7 RESULTADOS ESPERADOS

A partir del logro de los objetivos específicos, se planea generar:

- Una propuesta clara y fundada de un modelo de cobro a aplicar en la empresa, acompañado de una simulación de como mejoría los cobros realizados

- Tableros de visualización que permitan apoyar la toma de decisiones con respecto a la lectura de viviendas sin registro de consumo producto de la pandemia COVID-19. Además, planillas de cálculo que indiquen la probabilidad de lectura de cada cliente afectado

- Una herramienta de visualización que integre lógicas descriptivas, supervisadas y no supervisadas, para utilizarla como apoyo a la toma de decisiones para mejorar la productividad y lectura de consumos durante abril de 2021

- Una planilla de cálculo donde se indiquen los clientes de un segmento vulnerable socioeconómicamente y la probabilidad de reclamo de cada uno

CAPÍTULO 2: NUEVO MODELO DE COBRO

2.1 ENTENDIMIENTO DEL PROBLEMA

El problema por solucionar en este capítulo es el de la imprecisión de los cobros en casos de que no se pueda registrar el consumo a los clientes con medidores convencionales, entendiendo por precisión a la magnitud con que el consumo predicho se asemeja al real. En este caso el problema se produce debido a dos causas:

- La imposibilidad de lectura en diversas situaciones por parte de los contratistas asociados. Esta situación ha aumentado de forma considerable con la pandemia COVID-19.
- El artículo 129 del decreto 327 del Ministerio de Minería (1998), que indica directamente: "Si por cualquier causa no imputable al concesionario no pudiere efectuarse la lectura correspondiente, el concesionario dejará una constancia de esta situación en un lugar visible del inmueble y podrá facturar provisoriamente, hasta por dos períodos consecutivos, una cantidad equivalente al promedio facturado en los seis meses anteriores"

Lo anterior implica que si a un cliente con medidor convencional, no se pudiera registrar su lectura por parte de los técnicos lectores, se le aplicaría un cobro considerando un promedio móvil de los últimos 6 meses. Este cobro tiende a ser poco preciso según la contraparte, lo que se verificará a partir del análisis de datos a realizar, justificando el problema antes de proponer nuevos modelos.

Los efectos de que el cobro sea poco preciso son los que siguen:

- Reclamos por parte de clientes hacia la compañía y la SEC
- Desconfianza con respecto al actuar de la compañía y desprestigio de esta en RRSS
- Se afecta la planificación financiera de las distintas familias

Es necesario conocer también la información que se les da a los clientes al momento de cobrar. Por esto, en las siguientes dos ilustraciones se muestra un ejemplo de boleta de cobro de Enel Distribución Chile.

Ilustración 8: Anverso de boleta de cobro de Enel Distribución. Fuente: Reclamos.cl

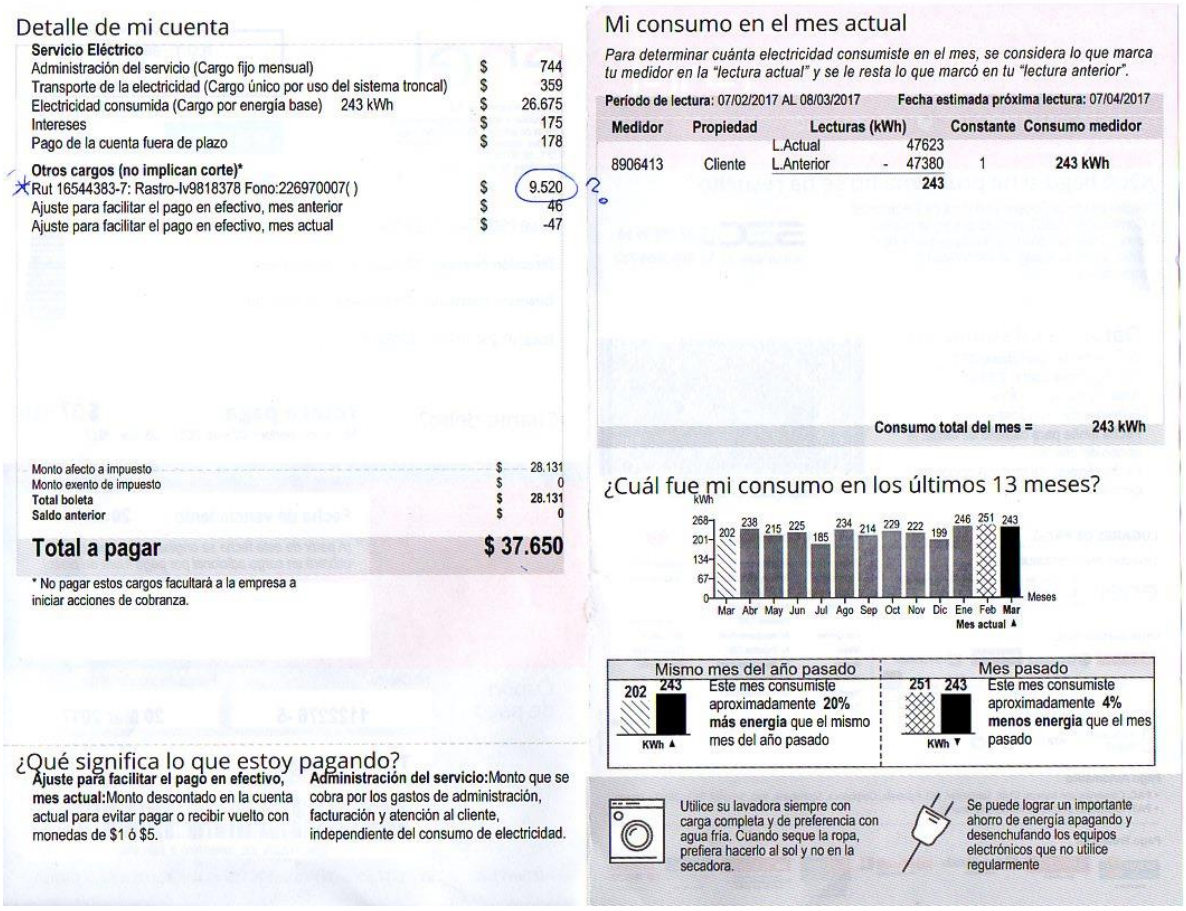

Ilustración 9: Reverso de boleta de cobro Enel. Fuente: Reclamos.cl

En las ilustraciones se puede apreciar que el cliente tiene información sobre el consumo del mes de cobro, el desglose del cobro, su patrón de consumo de los trece meses anteriores, donde puede realizar consultas y reclamos (donde se señala a la SEC como última opción), datos del suministro e instrucciones para pagar.

La información desplegada para el cliente puede llevar a confusiones y generar desconfianza. Esto dado que se muestra cierto consumo histórico que figura, pero si se registra un consumo mayor puede llevar a que el cliente tenga dudas, ya sea con respecto al cobro y la lectura, pudiendo pedir revisiones, cambios de medidores, entre otras medidas.

Por ejemplo, un cliente, cuyo consumo no ha sido registrado en junio, pero si ha podido ser registrado en julio, cancela un promedio móvil de 6 meses que incluye meses de bajo consumo en junio, pero, como el consumo de junio y julio se registran posteriormente, y son más altos, cancela en julio lo que no pagó en junio (producido por la imprecisión generada por el modelo de cobro actual analizado en la siguiente sección). Así, al cliente se le generará la percepción de que la empresa está realizando cobros excesivos o que hubo errores de lectura, dado el cobro anormal de su boleta, que es inconsistente con los valores históricos que se registran en la misma.

En conversaciones con la contraparte, se concluye que el segmento prioritario, según nivel socioeconómico a abordar, deben ser los de menores ingresos, por lo tanto, se utiliza la

categorización del SEREMI de desarrollo social del año, en base al índice de prioridad social (IPS) del año 2019.

El IPS es un indicador compuesto que integra aspectos relevantes del desarrollo social comunal, esto es, las dimensiones de: ingresos, educación y salud. Se trata de un índice sintético cuyo valor numérico permite dimensionar el nivel de vida relativo alcanzado por la población de una comuna. Así, el valor del IPS obtenido por cada comuna sólo se entiende en relación con los valores de dicho índice en las restantes comunas.

En la figura que sigue se exponen las distintas comunas de Santiago, con su índice de prioridad social y clasificación.

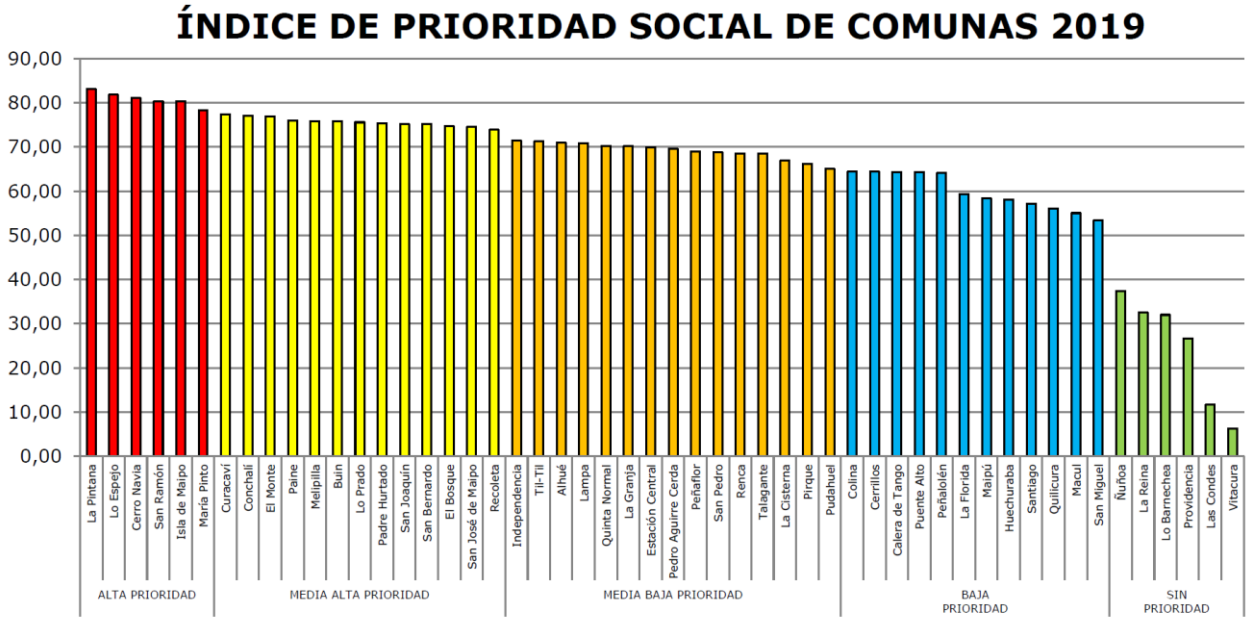

Ilustración 10: Índice de prioridad social de las distintas comunas de la región metropolitana. Fuente: Área de Estudios e Inversiones, Seremi de Desarrollo Social y Familia Metropolitana.

Para este trabajo, se prioriza a los clientes habitantes de segmentos de comunas con alta y baja prioridad, debido a tanto el impacto social como sugerencia de la contraparte. También se incluirán en el análisis las comunas sin prioridad social, como un segmento específico cuyos habitantes en promedio poseen ingresos particularmente altos según la encuesta de indicadores comunales CASEN 2013, cuyos ingresos promedio mínimo los posee la comuna de Ñuñoa, con $2.295.026. Este comportamiento particular de estas comunas también se puede apreciar en la misma figura 8, cuyos, IPS son particularmente alto. Por lo anterior, se tiene como hipótesis que este segmento de clientes presentará patrones de consumo totalmente distintos a los anteriores, lo que se verificará a partir de los análisis posteriores.

Se especifican las comunas de todos los segmentos en la siguiente tabla.

Tabla 2: Comunas según segmento que tienen concesión con Enel Distribución Chile

Prioridad Alta	Prioridad Media Alta	Prioridad Media Baja	Prioridad Baja	Sin Prioridad
La Pintana	Conchalí	Independencia	Colina	Ñuñoa
Lo Espejo	Lo Prado	Til-Til	Cerrillos	La Reina
Cerro Navia	San Joaquín	Lampa	Puente Alto	Lo Barnechea
San Ramón	El Bosque	Quinta Normal	Peñalolén	Providencia
	Recoleta	La Granja	La Florida	Las Condes
		Estación Central	Maipú	Vitacura
		Pedro Aguirre Cerda	San Miguel	
		Renca		
		La Cisterna		
		Pudahuel		

2.2 ADQUISICIÓN, DESCRIPCIÓN Y PRE-PROCESAMIENTO DE LOS DATOS

2.2.1 Adquisición de los datos

Primero, se extraen los datos de consumos de todos los clientes de la empresa desde el repositorio centralizado de datos que posee. Este corresponde a un Data Lake que posee datos sin procesar, almacenado en Apache Hive, una infraestructura diseñada almacenamiento masivo de datos basada en Apache Hadoop, cuyas consultas se realizan utilizando Cloudera Impala, un lenguaje SQL open source diseñado para el procesamiento masivo de datos en paralelo.

Al realizar las consultas se procura convertir las fechas UNIX en formato de fecha internacional ISO (DD/MM/AAAA).

Posteriormente, se procede a solicitar a la organización, la base de datos de todos los clientes actualizada a 2021. Se accede a la base de medidores de la compañía que posee atributos de los distintos clientes.

En total, se generan 51 archivos de texto separado por comas (.csv) con los datos de consumo desde enero de 2017 a marzo de 2021. La base de medidores corresponde a un archivo de texto separado por espacios (.txt).

2.2.2 Descripción de los datos

Cada planilla de los datos de consumo, sin procesar, se componen de una columna destinada al número de suministro (o de cliente) de cada suministrado, que funciona como identificador único. Además, se posee la fecha correspondiente y el consumo mensual en kilovatios hora (kW-h). Se poseen los datos de consumo de todos los clientes en esta base de datos. Cada planilla posee el total de los clientes en la fecha correspondiente, que en marzo de 2021 son 2.003.812 y en enero de 2017 son 1.835.671.

Sobre estos últimos datos, cabe destacar que cuando un cliente posee un consumo de 0 kW-h, se podrán dar dos situaciones asociadas. La primera es que el cliente efectivamente no haya consumido energía eléctrica durante todo el mes estudiado. La otra situación es que no se haya podido registrar el consumo por parte de la empresa. Para efectos de este trabajo, se opera bajo el supuesto de que siempre ocurre la segunda opción, ya que las viviendas deshabitadas corresponden a un segmento minoritario y específico estudiado en la tercer capítulo de este documento.

La base de los medidores de la empresa se compone de su número de suministro, funcionando como identificador único, atributos relacionados al medidor correspondiente y sus características (que se utilizan al trabajar en los otros problemas abordados en este trabajo, por lo que ahí se ahonda en ellos), la forma de medición para cada caso, datos de cada cliente (comuna donde reside y dirección), entre otros.

2.2.3 Pre-procesamiento inicial

Para realizar el análisis exploratorio completo, se requiere formar un modelo de datos con las planillas señaladas, utilizando lenguaje Python, con su librería Pandas.

Para lo anterior, se concatenan las 51 planillas correspondientes, utilizando como identificador único el número de cliente en cada registro. Así, se forma un DataFrame único donde se puede manipular la data de todos los consumos de todos los clientes de la empresa desde enero de 2017 a marzo de 2021. Se verifica que no existen valores nulos.

Por otro lado, se utiliza la base de medidores para segmentar a los clientes. Como esta base posee valores duplicados, se trabajan manteniendo el primer registro para cada cliente, pues corresponde a la última instalación de cada medidor.

2.3 ANÁLISIS EXPLORATORIO Y JUSTIFICACIÓN DEL PROBLEMA

El flujo de trabajo a seguir para justificar este problema será primero, mostrar que existe insatisfacción por parte de los clientes, asociada a problemas de facturación, tanto a nivel de Enel Distribución Chile S.A., como del resto de las empresas. Posteriormente, se buscarán los espacios de tiempo donde más se centran estos reclamos.

Luego, a través de la manipulación y análisis del modelo de datos generados en la sección en la sección 2.2.3, se buscará ver patrones generales de consumo y como tienen relación con los patrones de reclamos por parte de los clientes. Así, a partir de estas coincidencias, se podrán establecer juicios fundados sobre el origen de los problemas de facturación.

Posteriormente, se mostrará de forma empírica que la fórmula de cálculo utilizada actualmente según la regulación potencialmente produce insatisfacción en los clientes, al poseer un error considerable.

2.3.1 Patrones de reclamos

Antes de manipular los datos de la empresa, se demuestra la existencia y se explora la magnitud del problema. Esta sección es de vital importancia debido a que se planea influir en las políticas públicas regulatorias del ministerio de energía con lo analizado en este capítulo.

Para lo anterior, se solicita a la Subsecretaría de Electricidad y Combustibles, vía transparencia de datos, datos agregados sobre el número de reclamos por motivo de los distintos clientes de todas las empresas distribuidoras de energía eléctrica del país, desagregada por fechas y años. El horizonte de tiempo del que se dispone es desde enero de 2016 a mayo de 2021.

Cabe destacar que los reclamos realizados a la SEC corresponden a una última instancia para los clientes que posean insatisfacción por uno o más factores. Como se mostró anteriormente, en la boleta de pago se índica de forma implícita que el cliente puede consultar a la empresa directamente, y si no se siente satisfecho con la atención, puede acudir al órgano regulador. Por lo mismo, se utilizan estos datos en vez de los reclamos directamente a la empresa, ya que, corresponden a clientes que se encuentran con mayor insatisfacción y pueden generar mayor daño a la empresa.

Primero, se verifica la distribución de los motivos de reclamos, transversales a todas las empresas.

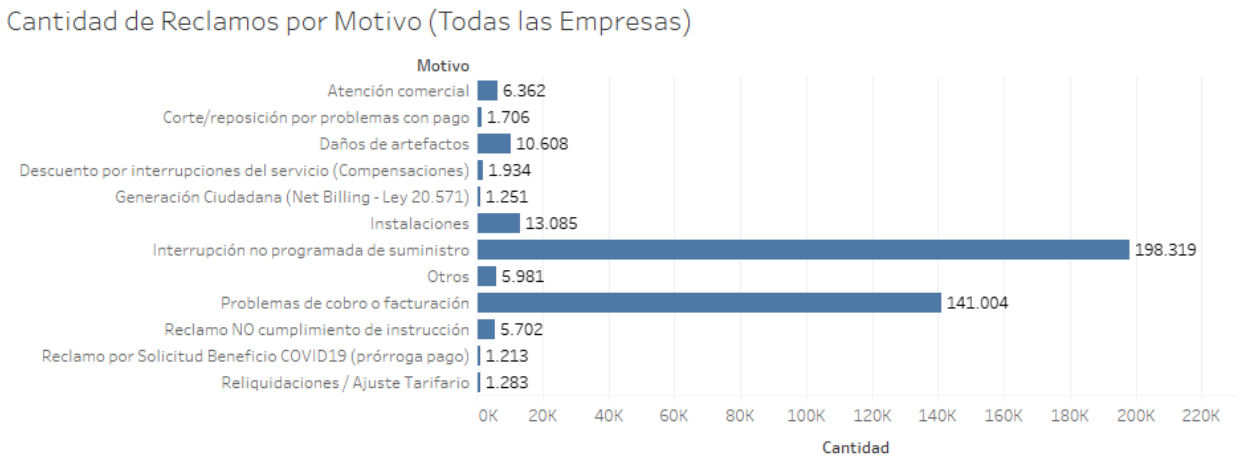

Ilustración 11: Cantidad de reclamos por motivo en todas las empresas

Se puede visualizar directamente que los motivos más frecuentes son interrupción no programada de suministro y problemas de cobro o facturación. Estos dos motivos en conjuntos conforman un 87,35% de los reclamos (51,05% y 36,30% respectivamente), lo que indica que son las principales causas históricamente. Lo anterior se verifica considerando que la causa siguiente en magnitud, instalaciones, representa solo un 3,37% de los reclamos históricos.

Notar que se realizan análisis incluyendo todas las empresas distribuidoras de energía eléctrica, ya que se espera que a partir de este trabajo se realicen cambios a nivel de regulación a nivel nacional, por lo que es necesario mostrar que el problema es transversal. Si se observa solamente Enel Distribución Chile, la distribución es similar y se pueden obtener las mismas conclusiones. En el anexo D se encuentra el análisis solo para la empresa.

Luego, se debe ver cómo ha evolucionado la frecuencia de los reclamos por motivo, para verificar su importancia actual con respecto al comportamiento pasado.

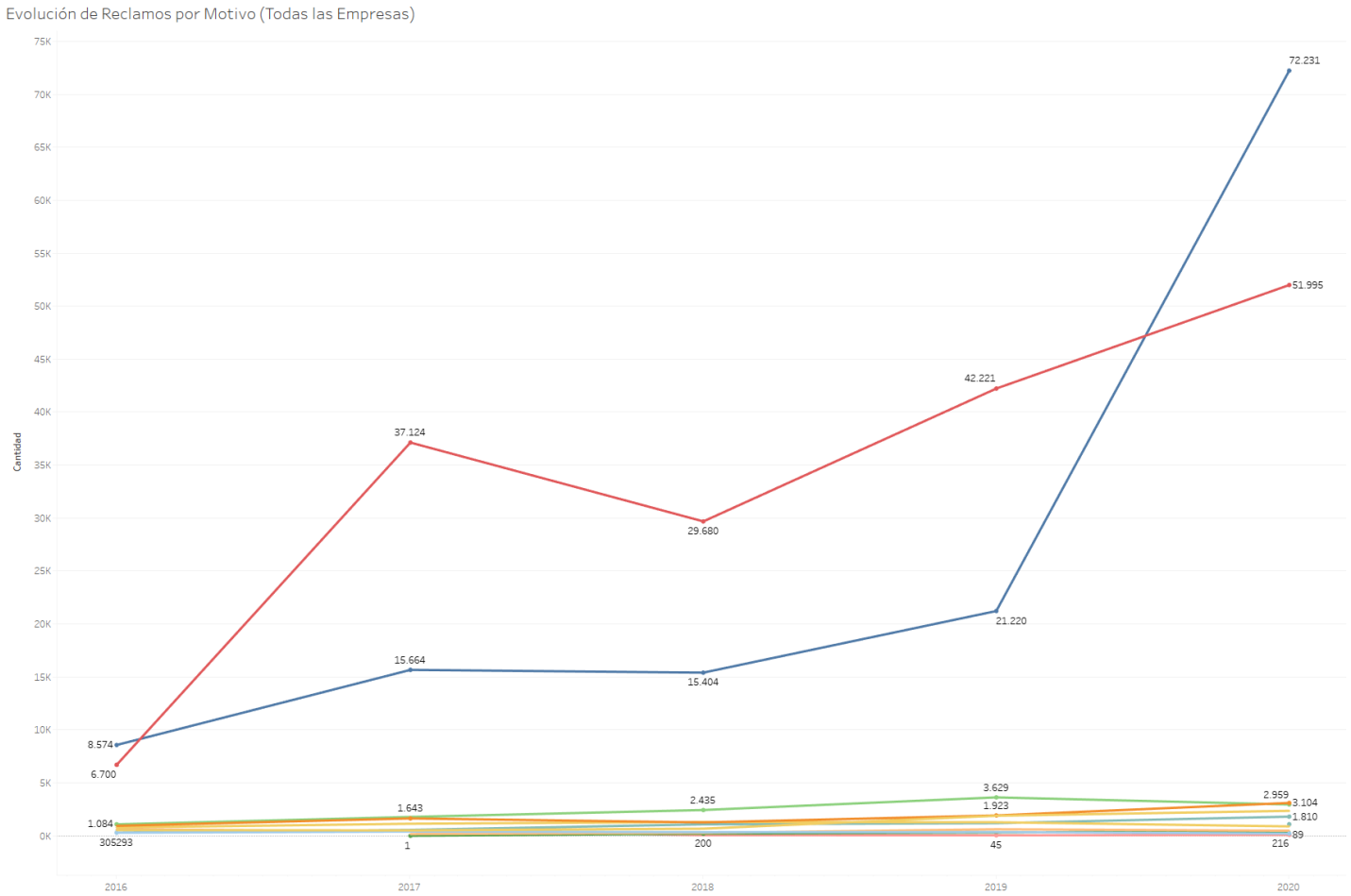

Ilustración 12: Evolución del número de reclamos por motivo a través de los años en todas las empresas. El color rojo representa los reclamos por corte de suministro y el azul los reclamos por facturación. Fuente: Elaboración propia

Primero, se puede apreciar que la causa de facturación históricamente ha estado entre las dos con mayor número de reclamos, lo que indica que es un problema consistente.

Además, se puede apreciar un aumento de un 240,39% de la cantidad de reclamos por facturación en el año 2020 con respecto al año 2019. Considerando la causa de interrupción de suministro, esta aumenta solo un 23,14%. Otro aumento considerable en el año 2020 es el de daño de artefactos, con un aumento de un 61,41%, pero el número continúa siendo bajo. Así, se concluye que la causa de facturación que más se incrementa con la situación COVID-19 es la de facturación y es la de mayor magnitud.

Cabe destacar que en el análisis anterior se excluye al año 2021 debido a que no se posee información de todos los meses y podría sesgar el análisis que se realiza.

Para Enel Distribución Chile en particular, se llegan a conclusiones similares, llegando a un aumento de 304,52% de los reclamos a la SEC por facturación. La visualización asociada se encuentra en el anexo E.

Adicionalmente, para verificar la transversalidad a nivel nacional problema de facturación, se analiza, para todas las empresas, la proporción de reclamos por facturación. Esto se puede apreciar la figura que sigue.

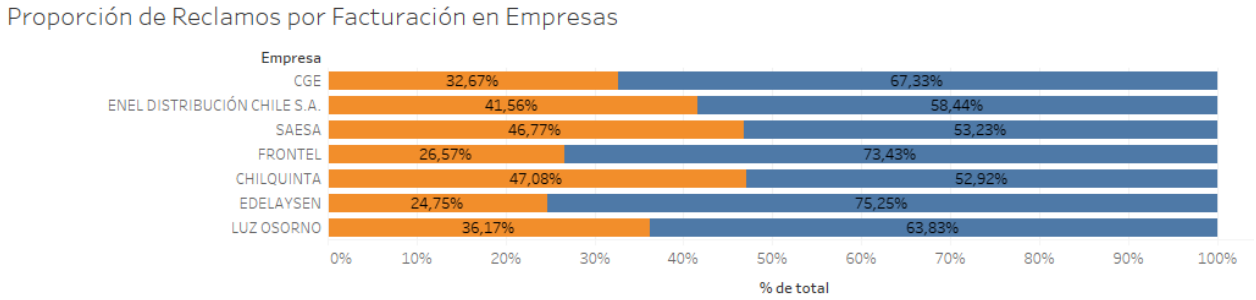

Ilustración 13: Proporción de reclamos por facturación con respecto al total en las distintas empresas distribuidoras. El color naranjo representa los reclamos por facturación, y el azul, los reclamos por cualquier otra causa. Fuente: Elaboración propia

Se observa que, para todas las empresas distribuidoras los reclamos por facturación tienen una incidencia importante.

Analizando los reclamos por facturación con respecto a los meses en que se producen, se tiene que el mayor valor se da en los meses de invierno de forma consistente, notando que esta concentración de reclamos se incrementó durante el año 2020, como efecto de la pandemia COVID-19. Esto se puede apreciar de forma resumida en la siguiente tabla.

Proporción de Reclamos por Fecha de Creación (Facturación Todas las Empresas)

	Mes creación											
Año de Creación	01	02	03	04	05	06	07	08	09	10	11	12
2016	6,72%	7,10%	7,03%	7,30%	7,77%	8,64%	8,56%	9,87%	12,81%	8,35%	8,15%	7,70%
2017	4,19%	3,77%	4,27%	3,77%	4,28%	4,39%	25,34%	18,64%	8,78%	8,92%	7,89%	5,74%
2018	6,44%	5,28%	6,15%	6,22%	5,98%	6,40%	8,17%	13,89%	10,91%	13,57%	9,26%	7,73%
2019	5,94%	5,51%	7,31%	6,76%	6,64%	6,67%	8,18%	12,35%	12,23%	9,60%	9,69%	9,11%
2020	2,69%	2,23%	1,89%	1,74%	18,61%	14,84%	24,71%	14,93%	7,26%	4,94%	3,58%	2,57%

Ilustración 14: Proporción de reclamos por facturación en cada año por mes, considerando todas las empresas. Los colores más saturados representan valores más altos. Fuente: Elaboración propia

Finalmente, se realiza el mismo análisis solo considerando Enel Distribución Chile, para así, comparar posteriormente los patrones de consumo con los reclamos generados.

Proporción de Reclamos Por Fecha de Creación (Facturación Enel)

	Mes creación											
Año de Creación	01	02	03	04	05	06	07	08	09	10	11	12
2016	7,07%	6,16%	6,06%	7,02%	8,30%	8,85%	10,12%	10,81%	11,31%	8,98%	7,43%	7,89%
2017	2,18%	2,22%	2,82%	2,27%	2,62%	2,89%	35,53%	23,88%	8,33%	6,91%	5,88%	4,48%
2018	7,52%	5,89%	6,73%	6,68%	6,60%	6,07%	9,29%	12,65%	9,82%	13,02%	8,33%	7,40%
2019	6,50%	5,26%	7,18%	6,42%	6,00%	6,81%	8,72%	13,68%	12,49%	9,75%	9,02%	8,18%
2020	2,14%	1,87%	1,56%	1,13%	2,32%	3,81%	39,72%	24,07%	9,93%	6,30%	4,25%	2,92%

Ilustración 15 Proporción de reclamos por facturación en cada año por mes, considerando solo Enel Distribución Chile. Los colores más saturados representan valores más altos. Fuente: Elaboración propia

Para este caso se encuentran patrones similares, concentrándose los reclamos en los meses de invierno. En la siguiente sección se intentarán establecer relaciones con el consumo de los clientes.

2.3.2 Patrones de consumo

Para analizar los patrones de consumo se debe realizar un pre-procesamiento adicional al DataFrame generado anteriormente.

Solo se trabajará con clientes cuyos consumos hayan sido registrados de forma completa en el horizonte de tiempo a estudiar. Así, se trabaja con un menor número de registros para realizar el análisis, pero se evita realizar procesos de imputación de datos al resto de clientes.

Adicionalmente, se detecta que en el año 2020 hubo un aumento considerable de consumos cero, concentrándose mayoritariamente desde marzo hasta julio, lo que produce que se trabaje solo con aproximadamente un 3% del total de los datos. Si bien, durante el año 2020 existe un gran número de casos de interés, existe una imposibilidad de conocer el valor real del consumo en un mes determinado, lo que imposibilitaría la evaluación de los modelos de cobro a proponer. Por lo tanto, se elimina este año completo del análisis, y solo de trabaja con consumos desde enero de 2017 a diciembre de 2019.

Los números respectivos por segmento de clientes, sin la eliminación de clientes con consumos nulos y considerándolos, se puede visualizar en la tabla que sigue.

Tabla 3: Disponibilidad de registros con y sin tratamiento propuesto aplicado. Fuente: Elaboración propia

Segmento	N° de Registros sin Tratamiento	N° de Registros con tratamiento
IPS Alto o Medio Alto	80.134	31.684
IPS Medio Bajo o Bajo	623.314	302.974
Sin Prioridad	224.807	110.578
Total	**928.255**	**445.236**

Dado el análisis realizado, al realizar el tratamiento, se trabaja con el 47,96% de los datos disponibles.

Se estudian las diferencias de la composición de los datos de la muestra con tratamiento en contraste con el conjunto que resulta fuera del tratamiento. Los resultados se resumen en la tabla que sigue.

Tabla 4: Comparación de estadísticos en las distintas muestras asociadas al consumo. Fuente: elaboración propia

Estadístico por Segmento/Conjunto	Datos Dentro de la Muestra	Datos Fuera de la Muestra
Media IPS Alto o Medio Alto	212,91	199,76
Desv. Estándar IPS Alto o Medio Alto	349,7	370,2
Media IPS Bajo o Medio Bajo	215,58	204,56
Desv. Estándar IPS Bajo o Medio Bajo	504,79	564,13
Media Segmento Sin Prioridad	344,23	330,51
Desv. Estándar Segmento Sin Prioridad	1309,43	1491,9

Teniendo en cuenta que la media considera todos los meses en que el consumo no ha podido ser registrado, pero dicho consumo se registra posteriormente. Las diferencias en los consumos promedio pueden deberse a las viviendas que están cerradas por un número considerable de meses. Debido a esta misma acumulación, es natural que la desviación estándar aumente al tener más registros lejanos a la media aritmética.

Por otro lado, se deben estudiar los valores extremos en la muestra, que pueden sesgar considerablemente el cálculo de las métricas de evaluación. Este análisis se realiza utilizando el consumo promedio de cada cliente. Si es que hay clientes con un consumo promedio muy grande a través del horizonte del tiempo, puede deberse a clientes industriales que están codificados como residenciales, o a muchos clientes que utilizan un mismo medidor, por ejemplo, más aún considerando el segmento más vulnerable.

Las visualizaciones asociadas para la distribución de consumos para cada segmento de clientes se encuentran en el anexo F, G y H. Todas solo muestran a grandes rasgos lo tendientes que son las distribuciones hacia valores muy altos de consumo. En todas se observa una gran concentración de los datos en el extremo y una distribución uniforme de los datos hacia el valor máximo. Lo anterior se solucionará mediante el tratamiento de los outliers.

Es necesario realizar un análisis numérico más exacto para tomar decisiones acerca del tratamiento de outliers, que se ve reflejado en la tabla que sigue.

Tabla 5: Análisis de la distribución de los consumos para distintos segmentos

	IPS Alto o Medio Alto	IPS Medio bajo o Bajo	Sin Prioridad
Cuenta	31.684	302.974	110.578
Media	212,91	215,58	344,23
Desviación Estándar	349,7	504,79	1309,43
Mínimo	3,33	1,33	1,52
25%	117,5	117,77	133,02
50%	171,97	168,13	200,27
75%	248,77	239,19	309,49
90%	335,05	345,25	499,67
95%	450,37	447.278	690,64
97%	531,72	539,7725	875,02
98%	615,56	632,03	1.116,61
99%	798,43	842,01	2.441,03
99.5%	1.032,70	1.160,50	5.828,15
Máximo	35.549	77.193	107.880

Considerando este análisis, se considera eliminar solo el último cuántil de la muestra para todos los segmentos, bajo consideración de que los consumos promedio por año son razonables al estar solo cuatro veces sobre el promedio. Explorando los valores inferiores, se encuentran consumos razonables en el primer cuantil, por lo que se mantiene.

Con este tratamiento se procede a analizar los distintos patrones de consumo para los distintos segmentos.

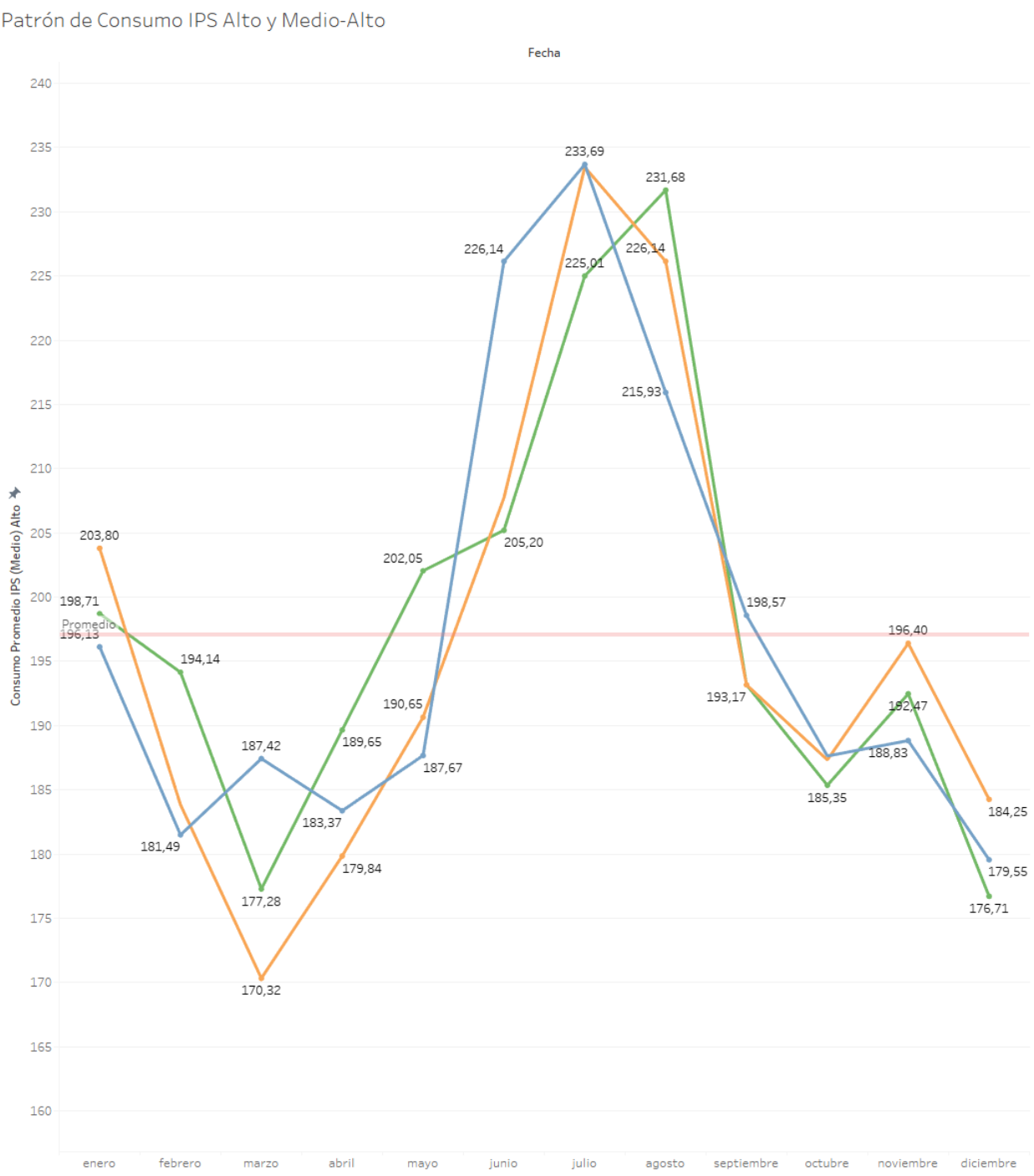

Ilustración 16: Patrón de consumo promedio del segmento IPS alto o medio alto. Las líneas representan el consumo promedio del segmento. El color azul representa el año 2017, el naranjo el año 2018 y el verde el año 2019. Fuente: Elaboración propia.

Para el segmento más vulnerable, se observa un patrón de consumo similar al promedio de todos los segmentos, durante todos los meses excepto en los meses de invierno, donde su consumo es considerablemente menor.

A continuación, se analiza el segmento de vulnerabilidad media.

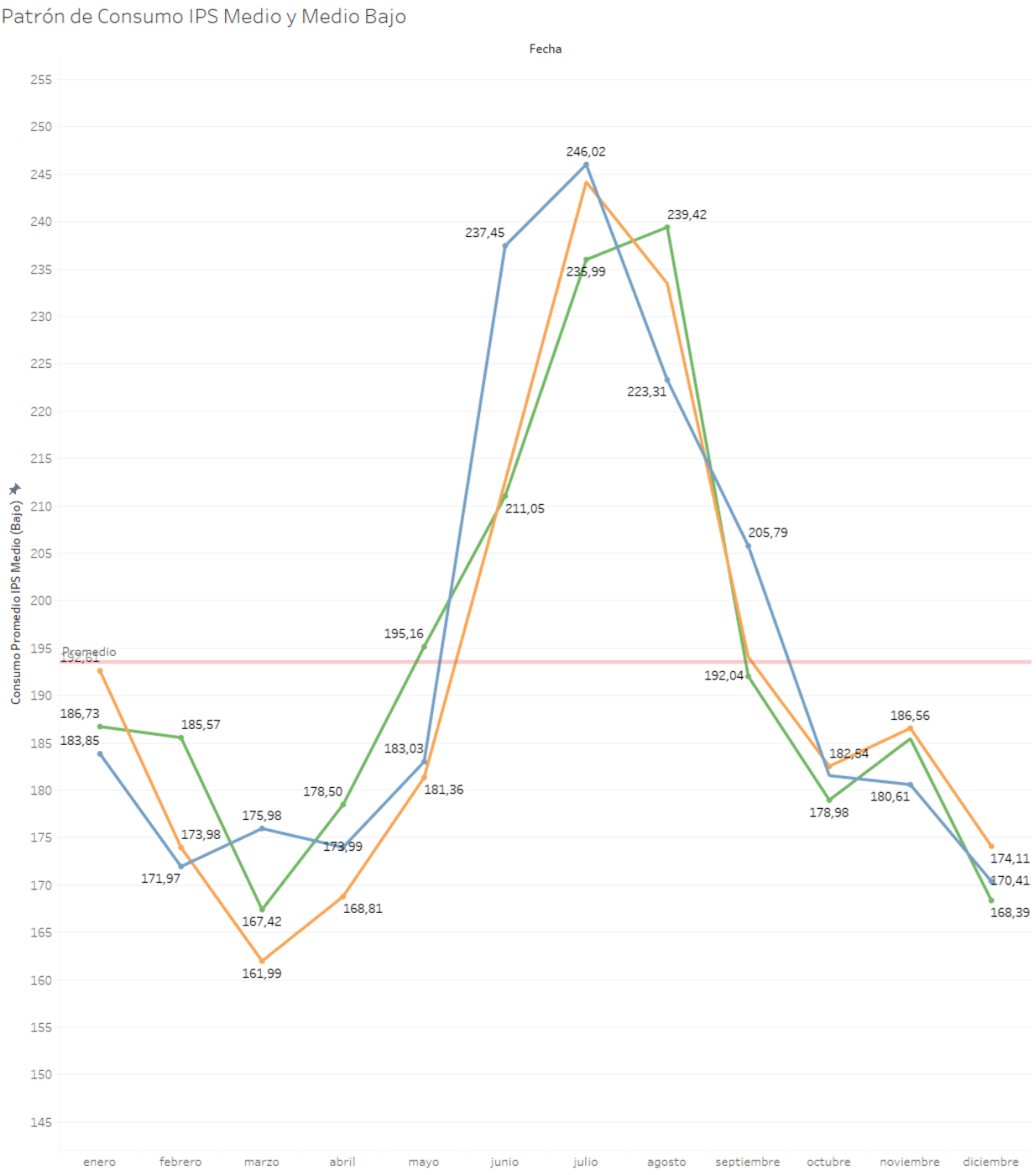

Ilustración 17: Patrón de consumo promedio del segmento IPS bajo o Medio bajo. Las líneas representan el consumo promedio del segment. El color azul representa el año 2017, el naranjo el año 2018 y el verde el año 2019. Fuente: Elaboración propia.

Para este segmento, se observa un consumo ligeramente menor al del segmento más vulnerable, pero solo en los meses que no son de invierno, dado que en estos meses el consumo es mayor que en dicho segmento. Aun así, el consumo promedio de este segmento es menor al más vulnerable, lo que se podría atribuir al mayor capital cultural que podría poseer o el poder adquisitivo para comprar electrodomésticos con mayor eficiencia energética. Se plantearán estas interrogantes a la organización para estudiar posibles acciones correctivas al respecto.

Finalmente, se analiza el segmento sin vulnerabilidad.

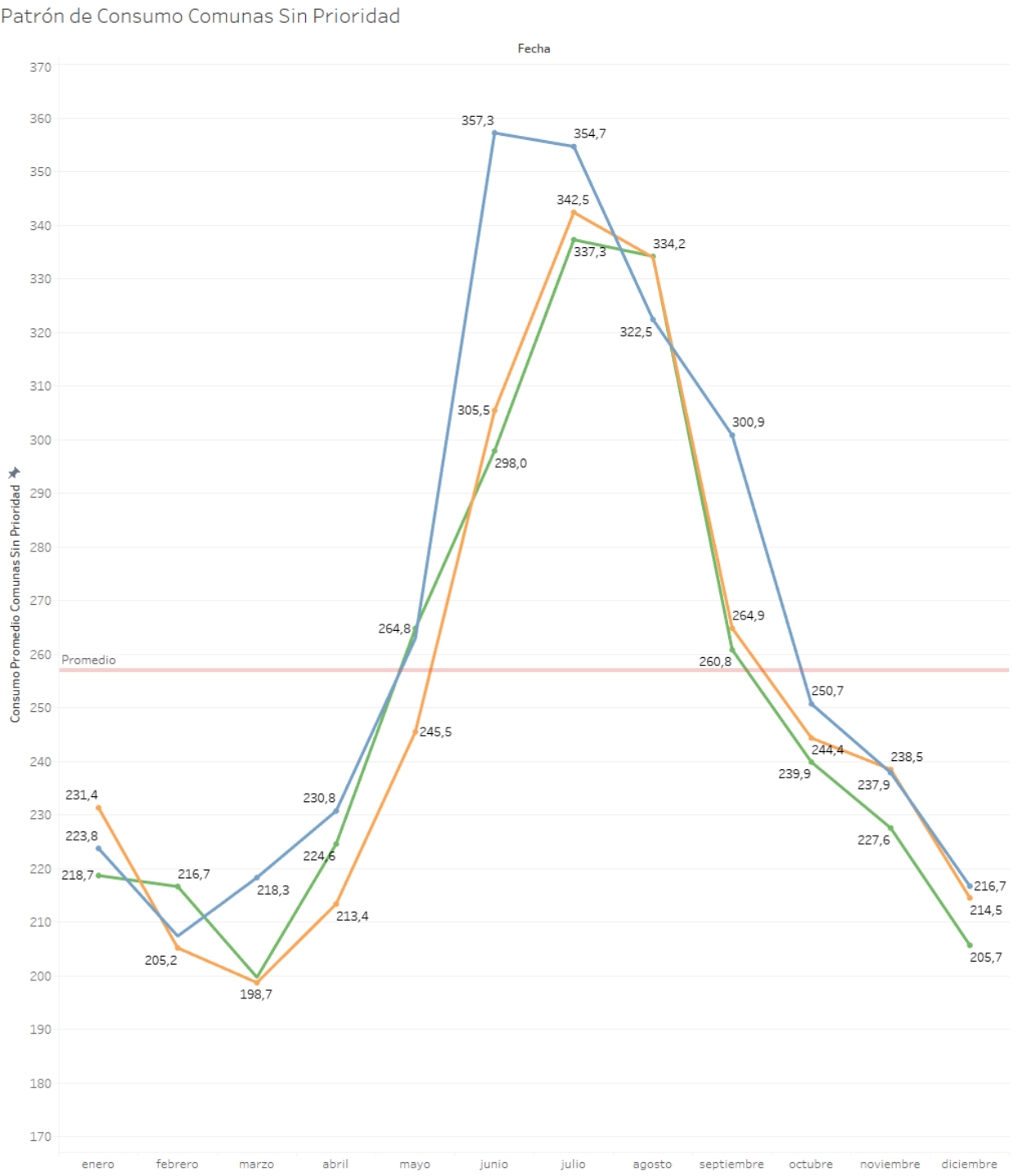

Ilustración 18: Patrón de consumo promedio del segmento sin prioridad. Las líneas representan el consumo promedio del segmento. El color azul representa el año 2017, el naranjo el año 2018 y el verde el año 2019. Fuente: Elaboración propia.

En este segmento se produce un consumo notablemente mayor que todos los otros segmentos, en todos los meses, pero se acentúa más aún en los meses de invierno, llegando a valores particularmente altos. Por lo anterior, se tiene la hipótesis de que a este segmento le puede afectar más que a los anteriores las imprecisiones en los cobros.

Curiosamente, el gasto en meses de verano es ligeramente mayor al del segmento más vulnerable, pero como existe mayor variabilidad, no pueden establecerse hipótesis al respecto.

Las desviaciones estándar a través del tiempo se encuentran en el anexo I, J y K, y representan un comportamiento similar al estudiado con los promedios, con mayores valores en los meses de invierno. Cabe destacar que la variabilidad en el segmento de menor vulnerabilidad es considerablemente mayor en todos los meses.

Del análisis general se puede concluir que, como los consumos son tan altos durante los meses de invierno, en todos los segmentos, se ve cierta relación entre los reclamos analizados anteriormente, que comienzan desde julio con sus valores más altos.

Si la situación de no lectura se repite en varios meses, la facturación en el mes en que el consumo si puede ser registrado puede ser muy excesiva cuando se realice. Por esto es posible que se pueda visualizar un alto número de reclamos en octubre, inclusive, dado que se genera un arrastre de los consumos de invierno a facturarse en octubre.

Se concluye que hay una insatisfacción considerable asociada a las imprecisiones en la facturación, que se ven relacionadas directamente con los meses con mayor aumento de consumo. Por lo mismo, los meses a analizar serán junio, julio, agosto, septiembre y octubre. Además, existen diferencias considerables en los patrones de consumo de los distintos segmentos, por lo que es necesario hacer un análisis específico para cada uno de ellos y mostrar que este es un problema transversal en la población, para así, mostrar que un cambio en la regulación.

2.3.3 Efectivad del modelo de cobro actual

Como métricas a utilizar para este problema, se propone el Error Absoluto Medio (MAE), el Error Porcentual Absoluto Medio (MAPE), además de la desviación estándar de los errores de predicción. El cálculo de los dos errores es como sigue.

$$MAE = \frac{\sum_{t=1}^{n} |A_t - F_t|}{n}$$

$$MAPE = \frac{\sum_{t=1}^{n} \frac{|A_t - F_t|}{|A_t|}}{n}$$

Donde A_t corresponde al valor real de los datos, F_t corresponde al valor predicho y n corresponde al número de registros.

A continuación, se muestran los resultados por segmento para estas métricas, considerando el modelo de cobro actual.

Tabla 6: Métricas obtenidas para el segmento IPS alto o medio alto. Fuente: Elaboración propia.

	Año	MAE Promedio	MAPE Promedio	Desv. Estándar Errores
IPS Alto o Medio Alto	Año 2018	48,95	31,88%	85,69
	Año 2019	47,83	32,22%	81,47
IPS Medio o Medio Bajo	Año 2018	54,57	35,6%	95,21
	Año 2019	53,83	36,2%	85,11
Sin Prioridad	Año 2018	89,41	49,04%	155,34
	Año 2019	86,4	48,79%	145,2

Se observa que, el segmento con mayores discordancias es el menos vulnerable, ya que se obtiene un MAE y MAPE considerablemente mayor además de que los errores poseen mayor variabilidad. En ese mismo sentido el segmento de vulnerabilidad media es el segundo lugar considerando estas mismas métricas. Finalmente, el segmento más vulnerable es el que posee métricas más favorables.

Aun así, se debe considerar la diferencia de ingresos que posee cada segmento al elaborar conjeturas sobre el impacto que puedan tener estas métricas. Por ejemplo, si hay un error de predicción de un 30% en una familia de segmento vulnerable, que consume 250 Kw-h en un mes de invierno, y se considera el valor de 110 pesos por Kw-h, deberá cancelar $8.250 adicional en la cuenta de su próximo mes. Si la situación de no lectura perdura por dos meses más, los cobros pueden elevarse, afectando directamente su planificación financiera.

Se sospecha que este modelo de cobro es ineficaz, dadas las métricas obtenidas. Por esto, se propondrán y estudiarán nuevos posibles modelos en las siguientes secciones, para compararlo con este de forma más detallada.

2.4 CRITERIOS INICIALES PROPUESTOS

Dado que se planea promover una política pública con este trabajo, se considerarán, primordialmente, formas de cálculo que sean fáciles de entender. Entonces, los meses en que no se puedan registrar consumos de un cliente, se aplicarían las siguientes políticas a testear.

- Criterio 1: Estimación en base a consumo del mismo mes del año recién pasado. Si dicho consumo no está disponible, se utiliza el primer mes anterior que esté disponible.
- Criterio 2: Estimación en base a media móvil de 12 meses. Si algún consumo no está disponible, se calcula la media móvil con menos meses.
- Criterio 3: Estimación en base a consumo promedio en mismo mes, anterior y posterior del año recién pasado. Si alguno de los consumos no está disponible, se utiliza los de los meses contiguos más cercanos. Si estos últimos no están disponibles, se utiliza el modelo 2.

- Criterio 4: Estimación en base a media móvil de 6 meses (Weighted Moving Average), considerando distintas ponderaciones en los meses como sigue:

$$Consumo\ Mes\ t = 0{,}1 * \left(\frac{\sum_{t=0}^{1} Consumo\ Mes\ t-6}{2}\right) + 0{,}3 * \left(\frac{\sum_{t=2}^{3} Consumo\ Mes\ t-6}{2}\right) + 0{,}6 * \left(\frac{\sum_{t=4}^{5} Consumo\ Mes\ t-6}{2}\right)$$

Los ponderadores dentro de los meses de consumo se determinan de forma empírica, utilizando distintos valores para los modelos y procurando los valores que maximicen las métricas objetivo.

- Criterio 5: Estimación en base a media móvil de 12 meses (Weighted Moving Average), considerando distintas ponderaciones en los meses como sigue:

$$Consumo\ Mes\ t = 0{,}6 * \left(\frac{\sum_{i=0}^{1} Consumo\ Mes\ t-12}{2}\right) + 0{,}2 * \left(\frac{\sum_{i=2}^{9} Consumo\ Mes\ t-12}{8}\right) + 0{,}2 * \left(\frac{\sum_{i=10}^{11} Consumo\ Mes\ t-12}{2}\right)$$

Los ponderadores utilizados se determinan según la misma manera utilizada en el criterio anterior.

Las métricas propuestas serán las mismas con las que se evaluó el modelo actual, vale decir, MAE y MAPE

El modelo final se verá definido como:

$$Cobro\ por\ Consumo = \begin{cases} Consumo\ Registrado & Si\ el\ consumo\ es\ registrado \\ Criterio\ Seleccionado & Si\ el\ consumo\ no\ puede\ ser\ registrado \end{cases}$$

2.5 RESULTADOS Y DISCUSIÓN

En esta sección, se detallarán los resultados para cada segmento con las métricas propuestas utilizando las mismas métricas utilizadas para el estudio del rendimiento del modelo actual (MAE y MAPE).

Dado el análisis exploratorio realizado, se sostiene la hipótesis de que la aplicación de los criterios 1 y 3 podrían comportarse de mejor forma al considerar un consumo cuyo promedio tiene una periodicidad de doce meses. Aun así, como la desviación estándar de los consumos es alta, podrían refutarse esta hipótesis con modelos que consideren más datos de consumo para su cálculo.

2.5.1 Resultados segmento IPS alto o medio alto

En la tabla 7 se presentan los resultados considerando todos los criterios ejecutados para el año 2018.

Tabla 7: Resultados para el segmento IPS alto o medio alto para el año 2018. Fuente: Elaboración propia

Métricas/ Criterios	Criterio Actual	Criterio 1	Criterio 2	Criterio 3	Criterio 4	Criterio 5
MAPE Junio 2018	31,82%	47,85%	32,08%	42,64%	31,42%	36,85%
MAPE Julio 2018	36,68%	50,63%	37,25%	45,66%	35,88%	38,52%
MAPE Agosto 2018	29,45%	43,68%	32,95%	42,01%	28,65%	33,71%
MAPE Septiembre 2018	28,2%	45,53%	31,64%	43,17%	27,93%	34,51%
MAPE Octubre 2018	33,24%	42,07%	32,81%	42,54%	30,36%	34,8%
MAE Promedio	48,95	60,76	47,03	54,92	46,66	46,61
MAPE Promedio	31.88%	45.95%	33,34%	43,2%	30,84%	35,67%
Desv. Est. Errores	85,69	108,96	78,96	93,35	82,95	80,32

Se observa que solo el criterio 4 posee un mejor rendimiento que el actual. Por otro lado, se rechaza la hipótesis de que los modelos 1 o 3 generarían un mejor resultado, pues poseen, un MAE promedio, MAPE promedio y desviación estándar de los errores notablemente mayor que el criterio actual y otras alternativas.

Estos resultados, la tabla 8, se contrastan con las del año 2019 para verificar su consistencia a través del tiempo.

Tabla 8: Resultados para el segmento IPS alto o medio alto para el año 2019. Fuente: Elaboración propia

Métricas/ Criterios	Criterio Actual	Criterio 1	Criterio 2	Criterio 3	Criterio 4	Criterio 5
MAPE Junio 2019	32,66%	46,11%	34,19%	44,26%	31,78%	38,14%
MAPE Julio 2019	35,61%	52,31%	37,49%	47,32%	35,13%	40,46%
MAPE Agosto 2019	29,08%	43,31%	31,86%	40,37%	28,58%	33,01%
MAPE Septiembre 2019	28,41%	41,76%	30,81%	42,49%	28,18%	32,69%
MAPE Octubre 2019	35,36%	41,65%	33,67%	41,28%	32,55%	34,99%
MAE Promedio	47,83	57,88	46,01	54,92	46,03	45,45
MAPE Promedio	32,22%	45,28%	33,6%	43,14%	31,24%	35,86%
Desv. Est. Errores	81,47	104,67	79	93,48	78,91	79,67

Para el año 2019 se observan métricas similares a las del año anterior, por lo que se concluye, de forma consistente, que el criterio 4 es el más adecuado para aplicar.

2.5.2 Resultados segmento IPS medio bajo o bajo

En la tabla 9 se observa la ejecución para el segmento de media vulnerabilidad.

Tabla 9: Resultados para el segmento IPS medio bajo o bajo para el año 2018. Fuente: Elaboración propia

Métricas/ Criterios	Criterio Actual	Criterio 1	Criterio 2	Criterio 3	Criterio 4	Criterio 5
MAPE Junio 2018	37,32%	60,01%	38,96%	52,34%	36,28%	45,52%
MAPE Julio 2018	38,69%	54,71%	39,72%	49,15%	37,51%	41,07%
MAPE Agosto 2018	32,47%	46,54%	35,37%	45,15%	31,43%	36,22%
MAPE Septiembre 2018	31,52%	52,8%	34,24%	50,21%	31,17%	38,93%
MAPE Octubre 2018	38%	42,14%	33,98%	43,21%	33,10%	35,08%
MAE Promedio	54,57	62,75	50,63	56,65	51,39	47,87
MAPE Promedio	35,6%	51,36%	36,45%	48,01%	33,89%	39,36%
Desv. Est. Errores	95,21	109,76	86,4	98,87	91,46	84,72

Para este segmento se observan mayores valores que para el anterior y una variabilidad mayor, por lo que ciertos sub-segmentos de esta clase podrían verse considerablemente afectados. Considerando las desviaciones estándar de los consumos, que se encuentra en el anexo J, que son altas, este valor podría afectar de sobremanera a un número relevante de familias.

En la tabla 10 se verifica consistencia de los resultados para el siguiente año.

Tabla 10: Resultados para el segmento IPS medio bajo o bajo para el año 2019. Fuente: Elaboración propia

Métricas/ Criterios	Criterio Actual	Criterio 1	Criterio 2	Criterio 3	Criterio 4	Criterio 5
MAPE Junio 2019	35,18%	50,36%	36,93%	49,56%	34,17%	42,2%
MAPE Julio 2019	39,46%	62,74%	42,52%	55,01%	38,85%	46,5%
MAPE Agosto 2019	32,13%	53,46%	36,76%	49,43%	31,76%	39,02%
MAPE Septiembre 2019	31,13%	45,73%	32,80%	47,52%	30,73%	35,78%
MAPE Octubre 2019	43,13%	49,73%	39,56%	48,64%	38,63%	41,2%
MAE Promedio	53,83	59,73	50,34	55,55	51,08	47,36
MAPE Promedio	36,2%	52,40%	37,71%	50,03%	34,83%	40,94%
Desv. Est. Errores	85,11	107,6	81,79	97,23	81,41	81,13

Se obtienen métricas similares que el año anterior, por lo que se confirma consistencia de los modelos a través de los años.

Por otro lado, se observa un aumento general de las métricas con respecto al segmento de mayor vulnerabilidad. Se tiene la percepción de que este aumento es condicionado por el aumento de la desviación estándar del segmento.

2.5.3 Resultados segmento de comunas sin prioridad

Finalmente, para el segmento sin prioridad, se tienen los siguientes resultados para el año 2018.

Tabla 11: Resultados para el segmento de comunas sin prioridad para el año 2018. Fuente: Elaboración propia

Métricas/ Criterios	Criterio Actual	Criterio 1	Criterio 2	Criterio 3	Criterio 4	Criterio 5
MAPE Junio 2018	49,3%	95,23%	55,96%	81,68%	48,01%	69,53%
MAPE Julio 2018	47,93%	92,41%	54,91%	84,54%	46,49%	63,40%
MAPE Agosto 2018	45,37%	73,17%	50,93%	72,48%	43,74%	56,50%
MAPE Septiembre 2018	42,57%	78,75%	45,57%	72,56%	40,86%	55,04%
MAPE Octubre 2018	60,05%	67,59%	52,65%	72,53%	51,04%	55,02%
MAE Promedio	89,41	94,46	78,98	84,46	83.37	71,38
MAPE Promedio	49,04%	81,43%	52,00%	76,76%	46,03%	59,90%
Desv. Est. Errores	155,34	163,7	136,55	154,61	147,98	131,58

Para este segmento en particular se obtienen errores más grandes, a la vez que desviaciones más grandes, por lo que pueden verse afectadas de igual forma, a pesar de sus mayores ingresos. Como la variabilidad es particularmente alta con respecto a los otros segmentos, los cobros pueden aumentar de forma significativa, generando

Tal como para el segmento de vulnerabilidad media analizado anteriormente, se tiene que para este año el criterio 4 se comporta de mejor forma que el criterio actual. Aun así, se debe analizar el siguiente año para proceder a proponer el criterio a utilizar en el modelo de cobro.

Tabla 12: Resultados para el segmento de comunas sin prioridad para el año 2019. Fuente: Elaboración propia

Métricas/ Criterios	Criterio Actual	Criterio 1	Criterio 2	Criterio 3	Criterio 4	Criterio 5
MAPE Junio 2019	46,28%	74,97%	51,13%	69,17%	44,34%	59,95%
MAPE Julio 2019	47,15%	73,77%	51,14%	69,61%	45,79%	56,34%
MAPE Agosto 2019	39,89%	68,54%	45,51%	62,09%	38,86%	48,79%
MAPE Septiembre 2019	46,71%	67,06%	48,19%	71,86%	45,26%	53,59%
MAPE Octubre 2019	63,93%	64,13%	53,29%	63,41%	55,11%	54,32%
MAE Promedio	86,4	84,05	77,07	84,46	83,37	67,75
MAPE Promedio	48,79%	69,69%	49,85%	67,23%	45,87%	54,60%
Desv. Est. Errores	145,2	173,49	137,14	149,4	138,54	128,32

Se observan métricas con valores ligeramente menores a los del año anterior, pero la tendencia con respecto a la conveniencia de los criterios a utilizar es similar y el segmento continua teniendo los valores menos convenientes con respecto a las métricas utilizadas.

2.5.4 Discusión general

Para apoyar la discusión se presenta la tabla 13, que contiene las medias ponderadas para cada segmento, considerando el número de clientes en cada uno.

Tabla 13: Resultados ponderados totales para el año 2019. Fuente: Elaboración propia

Métricas/ Criterios	Criterio Actual	Criterio 1	Criterio 2	Criterio 3	Criterio 4	Criterio 5
MAE Promedio	61,49	65,64	56,67	62,69	58,74	52,29
MAPE Promedio	39,04%	56,19%	40,43%	53,81%	37,32%	43,97%
Desv. Est. Errores	99,77	123,76	95,34	109,92	95,42	92,75

Consistente con el análisis anterior, se observa un rendimiento superior del criterio cuatro por sobre los demás. La hipótesis se ve rechazada debido a que los criterios 1 y 3 obtienen valores muy por sobre los demás, vale decir, se produjo un efecto contrario al deseado. Esto puede deberse debido a que se utilizan muy pocos datos para el cálculo, por lo tanto, su aplicación puede ser muy sensible a la alta variabilidad presente en los consumos. También se debe notar que los criterios que utilizan una mayor cantidad de datos para realizar la predicción poseen una desviación estándar menor, por lo que se debe considerar este factor en el trabajo futuro.

Entonces, la utilización de estos criterios provoca una mejora de un 1,62% del MAPE. Dados estos resultados, se procede a comparar la aplicación del criterio con mayor rendimiento con la utilización de modelos de aprendizaje automático supervisado en el siguiente capítulo.

2.6 APLICACIÓN DE MODELOS DE APRENDIZAJE AUTOMÁTICO

Con el fin de comparar el rendimiento del criterio de mayor rendimiento con modelos de aprendizaje automático, se aplicarán los siguientes para cada segmento, en sus versiones para problemas de regresión: Decision Tree, Regresión Lineal (OLS), K-Nearest Neighbors, SVM, Random Forest y XGBoost.

Para este apartado se utiliza solo el año 2019 debido a que se dispone datos desde el 2017 asociado a los consumos, por lo cual, se utilizarán dos años en el entrenamiento del modelo, para que se puedan capturar los patrones de consumo de forma correcta.

Se utiliza para la ejecución una partición del 70% de los datos para el entrenamiento y un 30% para la validación, utilizando los mismos registros que para la parte anterior. Cabe destacar que para el modelo SVM se utilizó una submuestra del 40% del total de la muestra utilizada en la sección anterior, para cuidar la buena optimización de hiperparámetros y disminuir los tiempos de ejecución, debido a la lentitud de cómputo de la aplicación de este algoritmo con respecto a los otros.

La configuración de hiperparámetros de realiza por búsqueda exhaustiva de los mejores resultados cuidando de que los modelos entreguen una desviación estándar similar a los demás y menor que la obtenida en la aplicación de los criterios anteriores. Dado el alto volumen de los datos, se consideraron algunos hiperparámetros clave para cada modelo, como se lista a continuación tal como se utilizan en la librería sklearn (y xgboost respectivamente para el algoritmo del mismo nombre) de Python.

a) Decisión Tree: max_depth, min_samples_leaf y min_samples_split
b) Regresión Lineal (OLS): parámetros por defecto
c) K-Nearest Neighbors: n_neighbors
d) SVM: kernel, C y epsilon
e) Random Forest: n_estimators, max_depth, min_samples_leaf y min_samples_split
f) XGBoost: n_esimators, learning_rate, max_depth, min_split_loss, min_child_weight

2.6.1 Resultados segmento IPS alto o medio alto

Una vez encontrada la mejor configuración de parámetros para este segmento para cada modelo, se procede a la ejecución, cuyos resultados se muestran en la siguiente tabla.

Tabla 14: Resultados de aplicación de algoritmos de aprendizaje automático para el segmento de comunas IPS alto o medio alto para el año 2019. Fuente: Elaboración propia

Métricas/ Modelos	Criterio 4	Decision Tree	OLS	KNN	SVM	Random Forest	XGBoost
MAPE Junio 2019	32,66%	25,78%	26,98%	28,94%	24,03%	24,45%	24,26%
MAPE Julio 2019	35,61%	31,09%	32,57%	36,84%	30,87%	30,81%	30,14%
MAPE Agosto 2019	29,08%	28,35%	29,26%	32,14%	28,31%	27,51%	27,75%
MAPE Septiembre 2019	28,41%	23,87%	25,53%	29,15%	26,31%	22,76%	22,26%
MAPE Octubre 2019	35,36%	26,08%	27,36%	33,75%	28,38%	25,34%	24,89%
MAE Promedio	47,83	35,43	34,78	38,51	35,69	33,5	32,1
MAPE Promedio	32,22%	27,03%	28,34%	32,16%	27,58%	26,17%	25,36%
Desv. Est. Errores	78,91	67,6	68,81	70,62	73,85	65,81	75,32

Se observa, de forma consistente en el espacio temporal estudiado, que todos los modelos de aprendizaje automático se comportan mejor que el criterio de mejor comportamiento seleccionado. El modelo que posee un mejor comportamiento en relación con su MAPE es el XGBoost seguido del Random Forest. Aun así, la desviación de los errores de XGBoost es mayor que el del Random Forest.

Por otro lado, algoritmos de mayor tradición y que requieren menor capacidad de procesamiento computacional, como la regresión lineal y el decision tree, poseen un buen comportamiento en comparación a los anteriores, por lo que también se pueden considerar, ya que poseen la ventaja de una mejor posibilidad de interpretación de su funcionamiento.

Se observa una efectividad menor en el mes de julio, donde existe un mayor consumo por parte de todos los usuarios, lo que puede estar directamente correlacionado con la mayor desviación estándar de los consumos que se registra en ese mes.

2.6.2 Resultados segmento IPS medio bajo o bajo

Se reportan los resultados para el segmento de media vulnerabilidad en la siguiente tabla.

Tabla 15: Resultados de aplicación de algoritmos de aprendizaje automático para el segmento de comunas IPS bajo o medio bajo para el año 2019. Fuente: Elaboración propia

Métricas/ Modelos	Criterio 4	Decision Tree	OLS	KNN	SVM	Random Forest	XGBoost
MAPE Junio 2019	34,17%	33,59%	31,02%	38,39%	32,21%	32,55%	30,25%
MAPE Julio 2019	38,85%	34,83%	33,08%	44,36%	36%	35,05%	32,91%
MAPE Agosto 2019	31,76%	30,12%	30,18%	35,21%	31,49%	28,86%	27,04%
MAPE Septiembre 2019	30,73%	24,24%	25,71%	30,56%	27,6%	23,43%	24,65%
MAPE Octubre 2019	38,63%	25,77%	25,65%	32,6%	28,25%	25,31%	25,1%
MAE Promedio	51,08	36,5	34,85	39,87	34,1	34,58	32,83
MAPE Promedio	34,83%	29,71%	29,13%	36,22%	31,11%	29,04%	27,99%
Desv. Est. Errores	81,41	66,01	64,93	70,31	72,3	63,99	73,42

Se observa que para este segmento se repiten las tendencias del anterior con respecto a la conveniencia de los modelos a utilizar, pero son menos acertadas en general. Finalmente se estudiará el último segmento para generar conclusiones generales.

2.6.3 Resultados segmento de comunas sin prioridad

En la tabla que sigue se muestran los resultados para el último segmento.

Tabla 16: Resultados de aplicación de algoritmos de aprendizaje automático para el segmento de comunas sin prioridad para el año 2019. Fuente: Elaboración propia

Métricas/ Modelos	Criterio 4	Decision Tree	OLS	KNN	SVM	Random Forest	XGBoost
MAPE Junio 2019	44,34%	34,88%	35,68%	43,18%	37,03%	33,75%	33,58%
MAPE Julio 2019	45,79%	36,31%	35,29%	46,98%	38,95%	35,03%	35,69%
MAPE Agosto 2019	38,86%	31,8%	31,37%	37,18%	35,24%	30,16%	30,05%
MAPE Septiembre 2019	45,26%	38,24%	37,91%	45,81%	42,82%	37,29%	36,71%
MAPE Octubre 2019	55,11%	27,39%	27,63%	37,81%	31,59%	26,51%	26,48%
MAE Promedio	83,37	55,82	50,9	72,46	60,42	52,44	50,4
MAPE Promedio	45,87%	33,72%	33,58%	42,19%	37,13%	32,55%	32,50%
Desv. Est. Errores	138,54	104,3	95,8	108,41	137,02	99,85	124,65

Se observa que, las tendencias varían, ya que, para este caso en específico, sería más acertado utilizar un modelo Random Forest a un XGBoost, debido a que posee un nivel de acierto muy similar, pero una desviación estándar notablemente menor. Modelos más tradicionales como decision tree y regresión lineal, a pesar de que poseen un MAPE ligeramente mayor al de los algoritmos de mayor rendimiento, pueden ser una alternativa viable, ya que la desviación estándar es comparable al de estos modelos y permiten un mayor nivel de interpretabilidad.

Los algoritmos SVM y KNN dejan de ser una opción para aplicar en este segmento ya que tanto su desviación estándar como MAPE poseen valores consistentemente más altos al de otras alternativas que requieren menor velocidad de ejecución.

Cabe destacar que, en este segmento, en general, se observa una consistente mayor variabilidad en los resultados y una tendencia a mostrar valores menos favorables de las métricas utilizadas no solo en el mes de julio, sino que también en septiembre.

2.7 DISCUSIÓN GENERAL Y CONCLUSIONES

Para apoyar la discusión general se presentan los resultados ponderados por número de habitantes para cada segmento en la siguiente tabla.

Tabla 17: Resultados generales de la aplicación de modelos de aprendizaje automático. Fuente: Elaboración propia.

Métricas/ Modelos	Criterio 4	Decision Tree	OLS	KNN	SVM	Random Forest	XGBoost
MAE Promedio	58,87	41,22	38,83	47,87	40,75	38,94	37,14
MAPE Promedio	37,39%	30,52%	30,18%	37,41%	32,35%	29,71%	28,92%
Desv. Est. Errores	95,42	75,63	72,87	79,79	88,48	73,03	86,28

En general, se concluye que los modelos KNN y SVM deben ser descartados de forma directa, ya que involucran los tiempos de ejecución más altos y sus métricas no son proporcionalmente favorables a este costo de tiempo, por lo que no son eficientes ni eficaces.

En general, se observa una clara tendencia a obtener resultados menos favorables en meses en donde la desviación estándar es mayor, y estas situaciones se dan mayoritariamente durante los meses en estudio de este trabajo. Se verifica exploratoriamente que los modelos obtendrían mejores resultados en los meses fuera del estudio, por lo que la aplicación del modelo seleccionado, teóricamente, debería poseer mejores resultados durante todos los meses.

Para la elección del mejor modelo, se considera el paradigma de que el fin de las políticas públicas es lograr el bien común, por lo tanto, se deben privilegiar los segmentos con mayor cantidad de habitantes, en este caso el segmento de media vulnerabilidad. Pero, por otro lado, se debe tener en cuenta los segmentos más afectados con esta política pública, que son los de menores ingresos. Adicionalmente, en conversaciones con la contraparte, se concluye que los sectores de menores ingresos son los que pueden dañar de mayor forma la imagen de la empresa. Considerando esto, se privilegiarán los resultados en los segmentos de media y alta vulnerabilidad social.

Bajo ese esquema, el modelo seleccionado es el XGBoost Regressor, ya que posee un mejor comportamiento en los segmentos objetivos y un mejor rendimiento general.

Aun así, al implementar una política pública, es necesario tener en cuenta ciertas limitantes éticas que se puedan tener. Una muy importante es el de la interpretabilidad de los algoritmos utilizados, a lo que César Buenadicha y otros autores en el artículo 'La Gestión Ética de los Datos' definen como 'opacidad'. En la implementación de una política pública que implique uso de algoritmos predictivos, una de las exigencias más generalizadas es la transparencia frente a los errores y riesgos de los sistemas (Buenadicha et al., 2019). Un algoritmo sofisticado como XGBoost Regressor puede ser interpretado por la población como una 'caja negra', es decir, como un mecanismo incomprensible, difícil de interpretar. A pesar de esto, esta situación puede entenderse más bien como una opacidad intrínseca del algoritmo, no como una opacidad intencionada o analfabeta, como definen los autores del artículo recién nombrado. Así, se concluye que es una situación con la que es muy difícil de lidiar, por lo que se debe juzgar posible utilización del algoritmo bajo su desempeño en vez de su funcionamiento.

Como se muestra en la siguiente ilustración, existe cierta relación con la exactitud que entrega la ejecución de un modelo con su interpretabilidad.

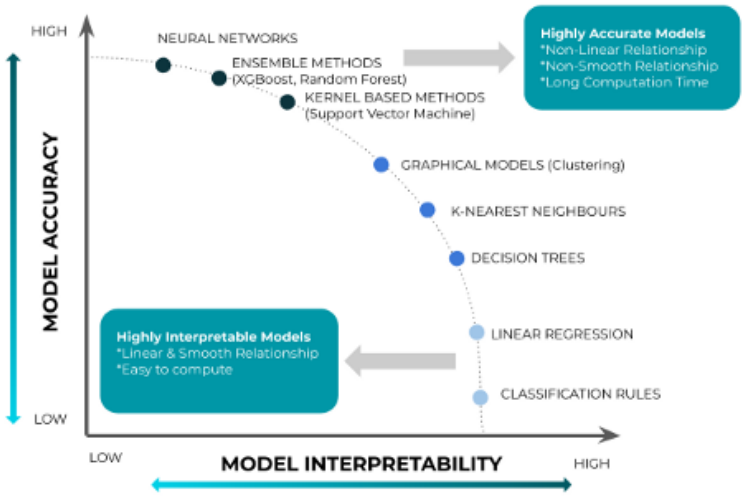

Ilustración 19: Precisión frente a interpretabilidad. Fuente: ExplainX.ai Internal

Lo expresado en la imagen se cumple parcialmente para este caso, ya que un modelo altamente interpretable, como la regresión lineal, posee mejor rendimiento que un SVM en este caso. Así, en caso de que no se pueda impulsar una política pública debido a la poca interpretabilidad del algoritmo, podría proponerse el uso de una regresión lineal, que sacrifica poco poder predictivo por una notable mayor interpretabilidad, como conocer de forma fiable que consumos se utilizan para el cálculo y su peso en la predicción.

Utilizando XGBoost Regressor en promedio se rebajaría un MAPE desde un 39,04% a un 28,92%, considerando el último año analizado y ponderando los resultados por el número de integrantes de cada segmento.

Dado que no se requieren nuevos procesos, tecnología, infraestructura o personas para cambiar esta lógica de negocios y utilizar un modelo de cobro más conveniente, tanto para la empresa, como para sus clientes y el órgano regulador, se sugiere al ministerio de energía y órganos reguladores estudiar la sustitución de dicho artículo para otorgar una mayor libertad a las empresas distribuidoras para poder utilizar modelos de cálculo más sofisticados y precisos.

2.8 TRABAJO FUTURO

Para realizar un juicio más certero sobre qué modelo utilizar se podrían utilizar dos años anteriores más para respaldar la investigación. Por disponibilidad de datos y tiempo para realizar este trabajo no se realizó dicha tarea, sin embargo, es difícil saber a ciencia cierta el valor que puede generar, ya que se actuaría con patrones de consumo de hace 5 años, lo que podría sesgar los resultados, dado que los distintos avances tecnológicos pudieron variar la eficiencia energética de los distintos electrodomésticos, o bien, cambios naturales en la población pudieron cambiar sus hábitos de consumo de electricidad.

CAPÍTULO 3: AUMENTO DE LA LECTURA EN VIVIENDAS CERRADAS PRODUCTO DEL COVID-19

3.1 ENTENDIMIENTO DEL NEGOCIO

La lectura de los medidores convencionales se realiza mediante dos empresas subcontratadas a base de licitaciones. A fecha de octubre de 2020, cuando se empieza a trabajar sobre este problema, las empresas contratistas son HelpBank S.A. y Provider. Hoy en día, debido a un cambio en la licitación, dichas empresas son Provider y Cobra.

El proceso opera considerando 20 sectores, cuyas lecturas inician con el número uno a principio del mes y terminan con el número 20 a fin de mes. Se hace seguimiento y recolección de datos acerca de las mediciones cuando se termina de leer un sector. Dada la normativa vigente del ministerio de minería, la empresa debe acudir a medir a todos sus clientes que posean medidores convencionales de forma mensual, en espacios que sean de entre 27 a 33 días.

La empresa realiza seguimiento de todos los clientes que no se han podido leer en, al menos, 3 meses (estos reciben la denominación de 'cerrados'). La problemática para abordar corresponde al alto número de clientes cerrados debido a la pandemia COVID-19. La intención de la empresa es lograr registrar el consumo a los clientes cerrados desde los meses de abril hasta junio de 2020, cuyo número a fecha del 23 de septiembre asciende a los 26.095.

Para realizar lo anterior, se proponen dos enfoques de solución: uno descriptivo, que permita a los analistas realizar seguimiento de este proceso de manera detallada y permitir toma de decisiones para cada caso, y posteriormente se implementa un enfoque predictivo, que permitirá enfocar mejor las medidas preventivas a tomar por parte de la empresa con respecto a los distintos segmentos de clientes.

3.2 ENTENDIMIENTO DE LOS DATOS

3.2.1 Gestión de clientes cerrados

La información asociada a la identificación de clientes cerrados la gestiona mediante planillas de cálculo que los analistas envían mediante e-mails de forma periódica al área para poder realizar seguimiento. En esta planilla cada registro corresponde a un cliente definido como cerrado. Los siguientes son los principales atributos de dicha tabla:

- **Número de Cliente:** Identificador único de cada cliente
- **Sector:** Variable categórica que toma valores de uno a veinte. Cada sector corresponde a un conjunto de clientes a medir, que pueden proceder de distintos lugares geográficos. Cada sector se divide en zonas, representado en otra variable categórica, con 594 valores entre todos los sectores.
- **Ruta:** Variable categórica que indica las rutas que pueden seguir los técnicos lectores. Son 5.075 en total
- **Tarifa:** Variable categórica que indica la tarifa con la que consume energía el cliente. En total son 24, siendo las más populares la BT1 y BT3. Los clientes que no poseen tarifa BT1 son conocidos como no residenciales en la empresa
- **Marca:** Variable categórica que indica la marca del medidor del cliente, son 84 en total

- **Dirección:** Variable categórica que indica la dirección del cliente, con su nombre de calle y número. Cada dirección tiene asociada una comuna representada en otro atributo
- **Meses Cerrado:** Variable numérica con valores mayores a tres, que indica el número de meses que lleva la vivienda sin poder ser leída
- **Irregularidad:** Código interno de la empresa que indica el tipo de causa específica de no lectura. Se ve complementada con la variable categórica 'Clave de Lectura' que indica una causa general por la que una vivienda no pudo ser leída

3.2.2 Situación de pago

La empresa de manera ocasional estudia la situación de morosidad de los clientes en estudio, para así realizar los análisis necesarios. Se dispone de la situación de pago al 23 de septiembre de 2020, lo que condicionó los análisis a esa fecha de inicio. Se compone de los siguientes atributos:

- **Cliente:** Indica el número de cliente en formato similar al de las planillas de gestión de clientes cerrados, se utiliza para realizar las uniones de tablas posteriormente
- **Fecha pago:** Indica la fecha en que se realizó el último pago por parte del cliente
- **Monto pago:** Indica el monto del pago que realizó el cliente

3.2.3 Reclamos de clientes

Quincenalmente, la empresa registra los reclamos realizados por los clientes por distintas vías. Concatenando las distintas fuentes de datos, se obtiene un dataframe donde cada registro contiene un reclamo realizado por un cliente único. Los datos se componen principalmente de los siguientes atributos:

- **Número del Caso:** Corresponde a un número único que identifica el reclamo realizado por el cliente
- **Número de Suministro:** Identificador único para el cliente. Se debe transformar esta variable para poder realizar la unión de datos con las otras entidades
- **Reclamo Formal:** Variable binaria que indica si el reclamo corresponde a uno formal, según políticas de la empresa
- **Fecha/Hora de Apertura:** Datetime en que se realizó el reclamo
- **Fecha de Cierre:** Fecha en que se cierra la solicitud/reclamo
- **Motivo:** Razón del reclamo, que pueden ser 46 distintas. Los motivos más frecuentes son: facturación, emergencia, recaudación, entrega de boletas y facturas, y lectura. Adicionalmente, un motivo puede tener submotivos asociados, que pueden ser 267 distintos
- **Canal de Origen:** Variable categórica que indica la vía por donde se capturó la data del reclamo. Son 27 distintas y las más frecuentes son a través IVR (Interactive Voice Response), call center, web y oficinas comerciales.
- **Tipo de Atención SEC/SENAC:** Indica si el reclamo se cataloga como una consulta, solicitud o reclamo según los estándares de los organismos gubernamentales

3.3 ENFOQUE DESCRIPTIVO

Como primera propuesta para abordar la problemática, se propone construir un tablero de visualización de monitoreo (o dashboard) funcional, que pudiera ser directamente utilizado por los

analistas de lectura en la organización y de las empresas contratistas externas. Este dashboard se concibe con la necesidad de actualizarse de forma mensual.

Para elegir la herramienta se utilizó primero el cuadránte mágico de Gartner que categoriza las distintas soluciones BI del mercado. Se muestra en la ilustración que sigue. Se opta por analizar las plataformas líderes de mayor rendimiento como lo son Micrososft Power BI y Tableau.

Se determina que, si bien ambos softwares tienen funcionalidades similares, y la primera alternativa es más económica, Tableau ofrece software desktop multiplataforma y la posibilidad de visualizar la herramienta de forma gratuita utilizando Tableau Reader, por lo tanto, se opta por esta herramienta, para que el sistema operativo de cada analista no sea una limitante.

3.3.1 Audiencia y propósito

La audiencia del tablero corresponde a analistas del área implicada, así como analistas pertenecientes a las dos empresas subcontratadas para la lectura: Helpbank S.A. y Provider.

El propósito del dashboard es que se realice seguimiento por sector de los clientes cerrados y se puedan visualizar espacialmente con facilidad su composición por tipo de cliente e irregularidades presentadas, además de su situación de morosidad y su cantidad de reclamos asociada, entre otros atributos relevantes para cada cliente.

3.3.2 Planificación de la herramienta

Se procede a entrevistar a la contraparte para el diseño de la herramienta, presentándole un piloto de esta para recoger sus impresiones.

Con estos antecedentes, se planifica la herramienta tal que pueda ser visualizada en un computador portátil, en un entorno donde se pueda mostrar en una reunión de área de forma gratuita. Se selecciona el software BI Tableau para trabajar debido a la disponibilidad de Tableau Reader para su distribución gratuita para los analistas.

El dashbords debe contener filtros claros para los clientes, pudiendo realizarse de manera espacial de forma flexible, por lo que la vista principal se verá cubierta por un mapa de la ciudad. Se deben utilizar elementos pre-atencionales como el color para visualizar con facilidad a los clientes con condiciones críticas.

3.3.3 Construcción de la herramienta

3.3.3.1 Tratamiento de los datos

A partir de las distintas entidades mencionadas en la sección 2.2 se procede a confeccionar una estructura capaz de incorporar todos los cruces de información necesarios para completar el panel de control, utilizando lenguaje de programación Python, con sus librerías Pandas y Numpy.

Posteriormente, para poder hacer un correcto seguimiento de los clientes de forma espacial, es necesario definir la latitud y longitud de cada cliente. Para realizar lo anterior se prueba con varias herramientas de geocodificación y finalmente se procede a seleccionar el servicio pagado Google Maps API debido a su exactitud al momento de geocodificar, lo que es de vital importancia para la audiencia objetivo.

Se procede a transformar la base de datos para realizar web scrapping con la herramienta anterior, utilizando lenguaje de programación Python. Se optimiza el texto libre del atributo 'Dirección' del dataframe, de modo que se obtenga la mayor exactitud posible desde Google Maps API.

Finalmente, se procede a arreglar de forma manual las imprecisiones que resultaron del uso de la herramienta geocodificadora.

3.3.3.2 Requerimientos funcionales de la solución

Los siguientes corresponden a los requerimientos funcionales mínimos de la solución BI:

- La herramienta debe poder mostrar los principales indicadores y métricas definidas en conjunto con el cliente
- La herramienta debe permitir navegar por los datos a través de los filtros, que pueden ser por comuna, sector, tipo de irregularidad, entre otros.
- La herramienta debe permitir visualizar en un mapa las viviendas cerradas y visualizar sus indicadores al seleccionar alguna.
- La herramienta debe permitir seleccionar varias viviendas cerradas y visualizar sus indicadores utilizando un promedio de ellos
- La herramienta debe incluir un selector radial que permita indicar en kilómetros y metros la distancia asociada en el mapa
- La herramienta debe permitir realizar análisis acerca de las variaciones que han tenido las viviendas a través del tiempo

3.3.3.3 Construcción en software BI

Con la base de datos ya configurada en Python, se procede a la carga de datos en Tableau. Como la empresa actualmente no posee una descripción de los datos estructuradas, se procede a averiguar mediante entrevistas el significado de las combinaciones de las distintas claves de lectura e irregularidades, y con eso se definieron seis tipos de clientes cerrados: habitados, deshabitados, por analista, sin medidor, con medidor cambiado y otros, donde esta última categoría se reserva para combinaciones de claves cuya significancia se desconoce. Se utiliza la variable visual color para indicar los distintos tipos de clientes cerrados. Los colores que representan atributos pre-atencionales frente a otros, como el rojo y el naranjo, representan viviendas en situaciones más complejas, como cerradas deshabitadas y cerradas por analista.

Utilizando las tarifas de los clientes se procede a separarlos según su actividad económica en residenciales y no residenciales. Se utiliza la variable visual forma para indicar cada uno en el tablero.

Luego de activados los filtros solicitados por el cliente en el software se obtiene un dashboard como se muestra en la ilustración 8, listo para validar.

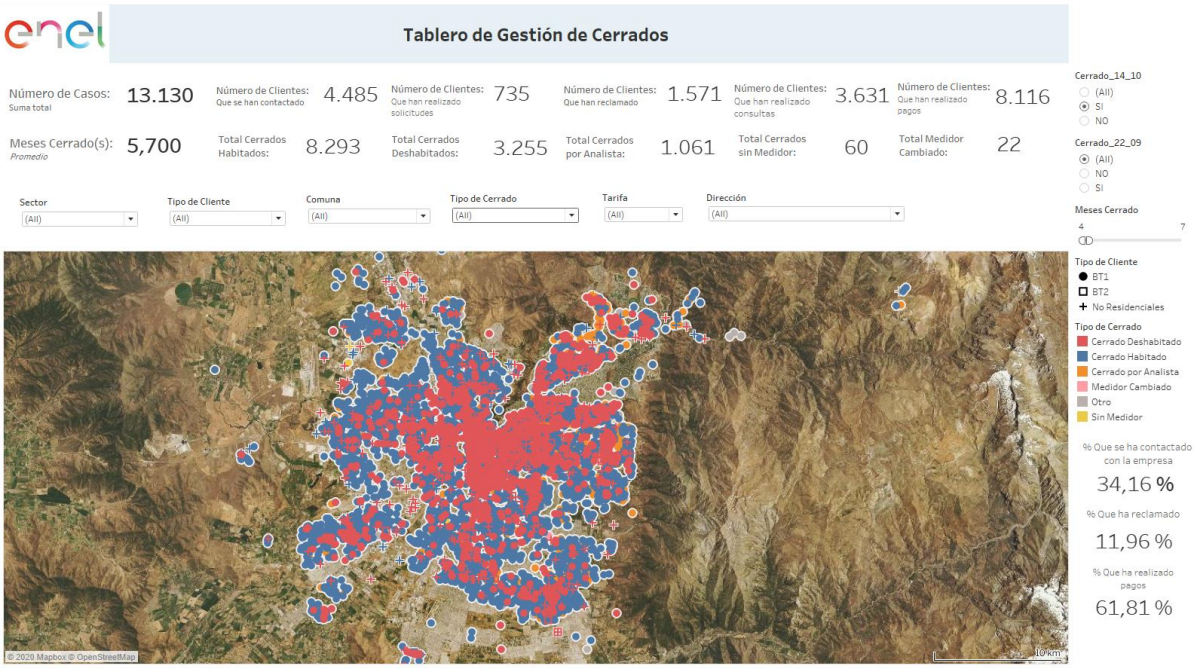

Ilustración 20 Tablero de Gestión de Cerrados, Primera Versión

3.3.3.3 Casos de uso

El caso de uso más básico del tablero es visualizar indicadores utilizando los filtros de listas desplegables. Por ejemplo, en la siguiente imagen se visualizan los indicadores y viviendas solo para el sector 5, que son tipo de clientes BT1 (residenciales) y están en la comuna de la florida, además de cumplir con cualquier irregularidad (tipo de cerrado), ser de cualquier tarifa y se incluyan todas las direcciones.

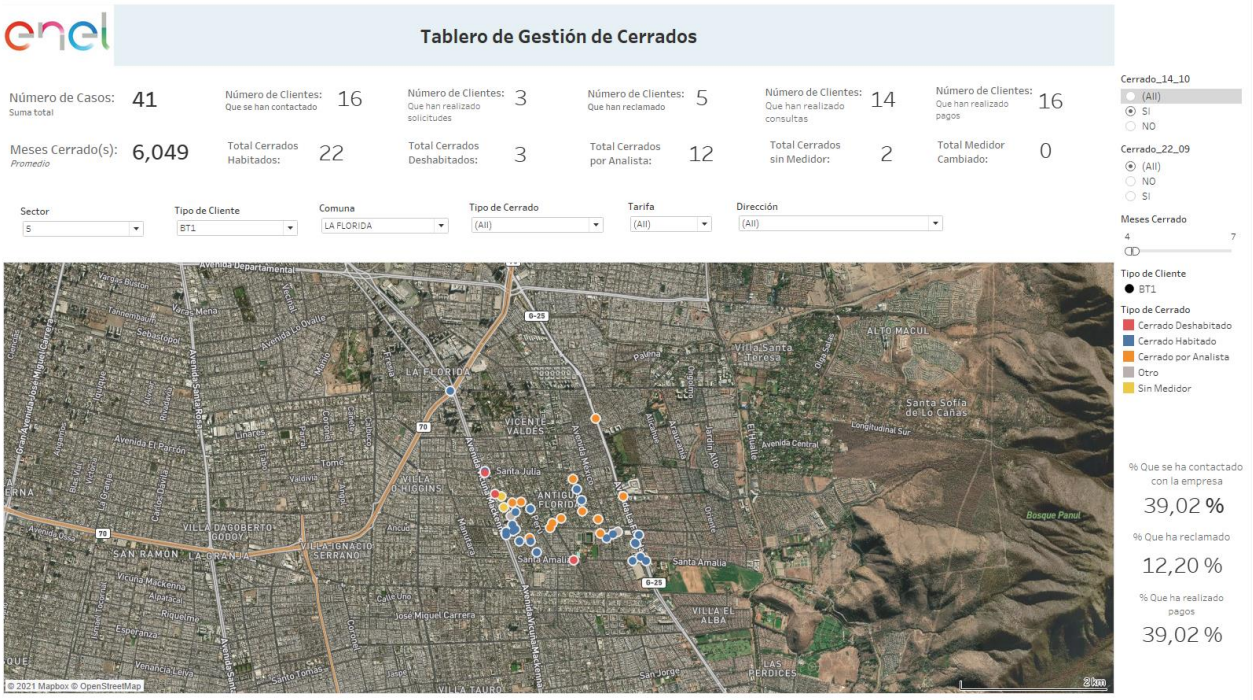

Ilustración 21: Caso de uso asociado al uso de filtros básicos

Esta funcionalidad puede ser utilizada por los analistas para saber donde se deben concentrar los mayores esfuerzos de lectura, o utilizar acciones previas a la lectura al principio de cada mes en ciertas localizaciones donde se encuentre una gran localización de viviendas sin lectura.

Adicionalmente permite buscar indicadores asociados a direcciones específicas. Por ejemplo, en este caso se observan los datos de la vivienda ubicada en la dirección Blaise Cendrars 6591, que se escribe en el campo 'dirección' y la herramienta propone posibles coincidencias, como funcionaría cualquier buscador moderno.

Ilustración 22: Caso de uso asociado al uso de filtro por dirección

Con esta funcionalidad se pueden buscar direcciones específicas de las cuales han solicitado consultas o revisiones, para que los técnicos lectores correspondientes tengan clara dicha información.

Por otro lado, si no se desea navegar por una dirección en específico, se puede buscar en el mapa alguna vivienda en particular de la que se quiera obtener más información, como se muestra en la siguiente imagen, donde se obtiene más información acerca de una dirección en particular, que se destaca ante las otras en el mapa, además de que se descubre que ya ha habido reclamos desde este cliente.

Ilustración 23: Caso de uso de selección individual por mapa

Adicionalmente, la herramienta permite, de forma simple y directa (solo una acción con el mouse), buscar información en Google acerca de una dirección encontrada en el mapa, como la utilizada anteriormente. Para esto se utiliza un menú desplegable simple como se muestra y un enlace. Esta funcionalidad permite buscar de forma ágil información acerca de las viviendas cuyas particularidades llamen la atención de los analistas.

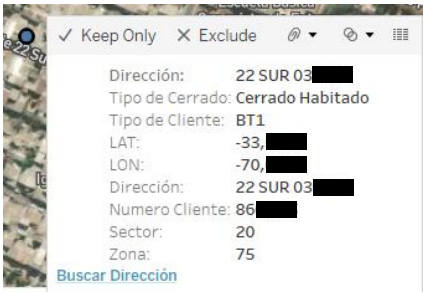

Ilustración 24: Menú desplegable para buscar dirección en la web

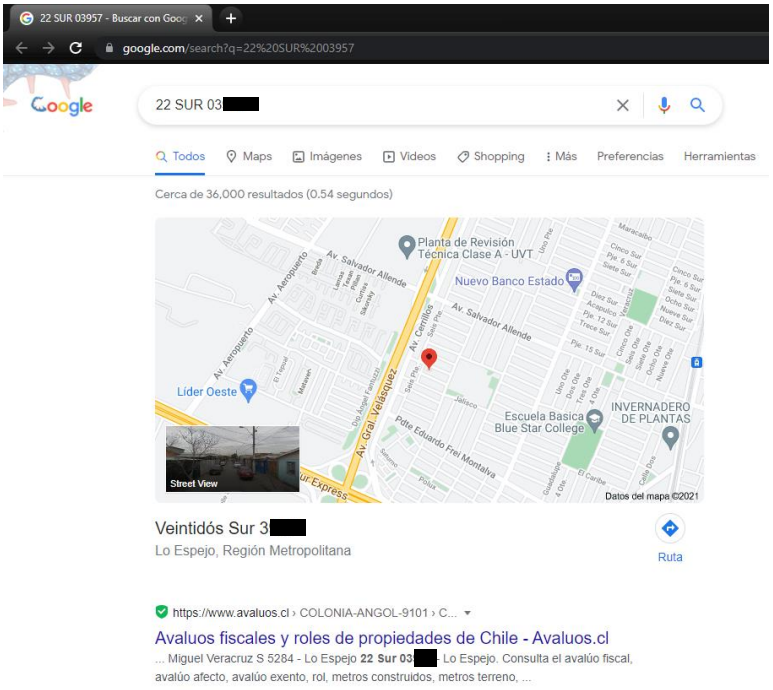

Ilustración 25: Búsqueda inmediata de información en la web acerca de la dirección

Por otro lado, se podrá buscar información acerca de varias viviendas y sus indicadores agregados utilizando distintos tipos de selectores, como la imagen que sigue donde se utiliza, por ejemplo, el selector rectangular.

Ilustración 26: Uso del selector rectangular en herramienta BI

Así, los analistas podrán visualizar indicadores agregados acerca de un conjunto de viviendas de interés y encontrar a los clientes que cumplan con ciertos valores que indiquen una mayor posibilidad de lectura según su criterio experto.

Adicionalmente, con el selector radial los analistas podrán buscar información considerando las distancias abarcadas, para así, asignar mejor a los técnicos lectores.

Ilustración 27: Uso de selector radial para visualizar viviendas ubicadas en un radio de un kilometro.

Finalmente, la herramienta permite realizar análisis sobre la evolución que ha tenido la lectura en una fecha anterior, para ver los distintos patrones asociados a las viviendas que pueden ser leídas. Por ejemplo, en la siguiente imagen se pueden visualizar las viviendas que estaban cerradas el día 22 de septiembre, pero que ya no están cerradas al 14 de octubre en el sector 1, en conjunto con sus indicadores.

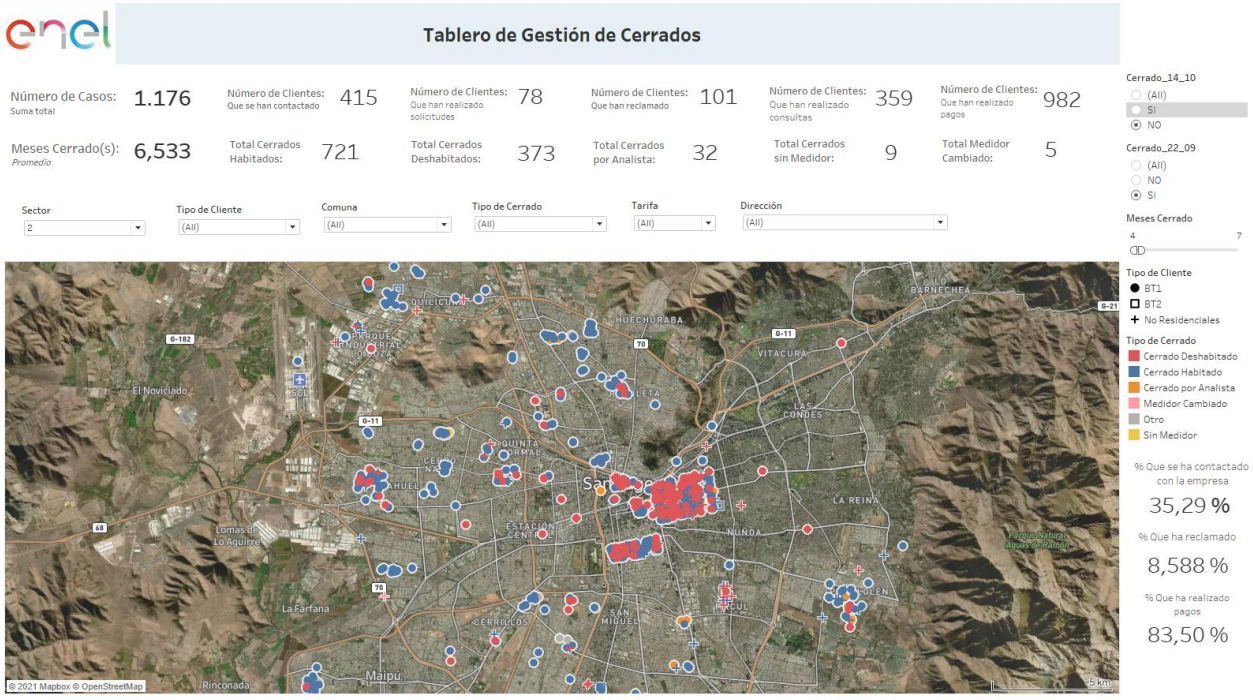

Ilustración 28: Análisis de la evolución de la lectura

Esto último puede ser utilizado por los analistas para obtener conclusiones acerca de los factores que pueden incidir en la lectura en distintos sectores, zonas u otros segmentos específicos de la población.

3.3.4 Validación de la herramienta

Se utiliza la regla de los cinco segundos expuesta en la metodología, para validar el instrumento y su futura implementación.

Primero, se presenta el tablero a la contraparte, para que pueda interactuar con él y dar sus impresiones. El resultado es que reconoce rápidamente el propósito del dashboard y su utilidad, y solicita algunos cambios menores que se realizaron.

Posteriormente, la contraparte muestra la herramienta en una reunión de sub-gerencia, para estudiar la posibilidad de una implementación en la organización, lo que resulta en una respuesta positiva.

Luego, se procede a realizar una reunión con los analistas del área, donde se utiliza nuevamente la regla de los cinco segundos obteniendo resultados positivos. Se sugieren nuevos cambios al tablero, para que se incluyeran a los clientes que la empresa había contactado por SMS o vía telefónica, para así, observar efectos causales en cada uno de los sectores y fortalecer los análisis. El tablero resultante es el mostrado en siguiente ilustración. Se entrega a la organización el tablero con vista de mapa de satélite y streetmap.

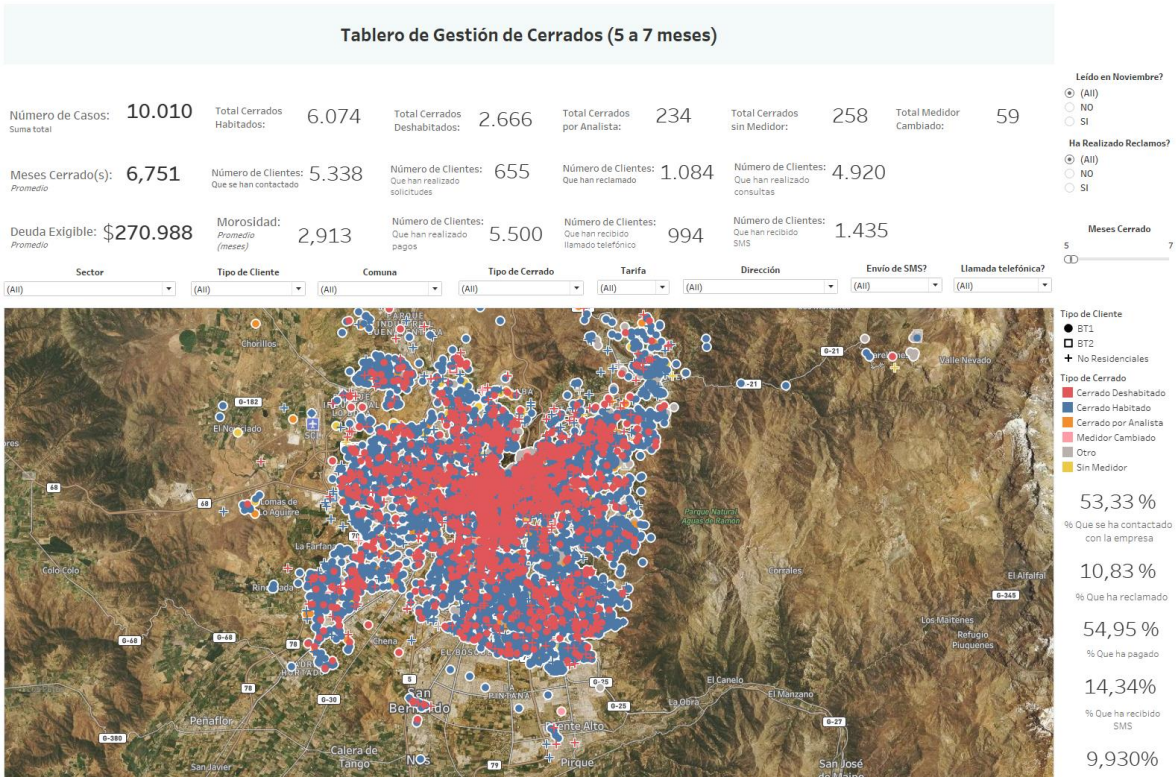

Ilustración 29 Versión Final del Dashboard

Esta versión final no solo es útil para monitoreo, sino que también para estudiar la relación que tienen las distintas variables y los efectos de las acciones de la empresa sobre los clientes, por lo que se utiliza como soporte de análisis exploratorio para los modelos predictivos descritos en la sección subsiguiente.

3.3.5 Trabajo futuro

Dado que este problema es permanente de la empresa, considerando no solo las viviendas cerradas de cuatro a siete meses, sino que las que poseen mayor tiempo en no ser registradas sus consumos, se planea construir una versión definitiva para el control de gestión de la empresa, para lo anterior, se debe diseñar un repositorio único que se actualice de forma periódica, que esté directamente conectada al tablero.

3.4 ENFOQUE PREDICTIVO

El modelo por describir en esta sección será implementado al inicio de cada mes para etiquetar a los clientes que no han sido leídos sus medidores con su probabilidad de que puedan ser leídos.

El esquema de uso es por parte de los analistas, ya que podrán enfocar un mayor número de medidas de contacto con clientes hacia los que tienen mayor probabilidad de lectura. Lo métodos que maneja la empresa actualmente corresponden al envío de mails, mensajes de texto y llamados telefónicos, ordenados en orden creciente de costos de utilizar cada uno. Estos se podrían asignar de mejor manera a cada cliente con la utilización del modelo, ya que actualmente se realizan sin criterios definidos según declaran.

Por lo anterior, se tiene que el baseline sobre el cual se actúa es la toma de decisiones según el azar. Así, la meta mínima a la que deben llegar los modelos es una métrica objetivo mayor a 0.5 para ambas clases y una curva ROC que esté sobre la línea de referencia.

Se comienza a trabajar con la misma base de datos resultantes de realizar la herramienta anterior, así como el entendimiento del negocio y los datos resultantes.

3.4.1 Preprocesamiento de los datos

Cabe destacar que la mayor parte de la limpieza de datos se realizó para construir la herramienta expuesta en la sección anterior, por lo que se procede directamente a la creación de variables.

La variable target en este caso se crea con el nombre de ‘Cambio Estado’ como sigue:

$$\boldsymbol{Cambio\ Estado} = \begin{cases} \mathbf{1} & \boldsymbol{Si\ es\ leído\ en\ una\ fecha\ posterior} \\ \mathbf{0} & \boldsymbol{Si\ no\ es\ leído\ en\ una\ fecha\ posterior} \end{cases}$$

Se crean variables relacionadas a distintas combinaciones de la base de reclamos, probando su efectividad en los modelos de forma empírica y seleccionando las más convenientes. Se utiliza el mismo método para las variables relacionadas a la morosidad de los clientes.

Con el método anterior, se elige generar variables dummy si es que los clientes han realizado solicitudes, reclamado o consultado a la empresa en los dos meses anteriores. Por otro lado, se consideran clientes que han realizado pago del servicio en los dos meses anteriores como máximo, representando esto en una variable dummy.

Finalmente, se crean variables dummy para todas las variables categóricas utilizando el enfoque dummy encoding, para evitar la multicolinealidad.

3.4.2 Modelamiento y validación

Se realizan dos iteraciones de modelamiento y validación, las cuales se ven acompañadas por fases de pre-procesamiento distintos cada una y van mejorando incrementalmente el rendimiento de los modelos.

Cabe destacar que, dado el problema de negocios, se seleccionan la métrica f1, que se define como sigue.

$$f1 = 2 * \frac{precision * recall}{precision * recall}$$

La justificación es que para la empresa es tan importante mantener tanto una exhaustividad como una precisión alta para la lectura. Debido a que intenta rebajar el número de viviendas cerradas lo más pronto posible, lo que sugiere necesario controlar el valor del recall. Por otro lado, la precisión cobra alta relevancia ya que el cliente utilizará la probabilidad de lectura, que será el output del modelo, en parte para tomar acciones focalizadas como llamadas por teléfono e-mails, que deben ser realizadas sobre clientes que realmente tengan una probabilidad alta de lectura. Por lo anterior se selecciona la métrica f1, debido a que un valor alto de está métrica tiene como requisito que ambos valores de interés posean un buen nivel.

Por otro lado, se privilegiará la capacidad de generalización del modelo para este caso, ya que el contexto puede ser variable, por lo que procurará mantener métricas asociadas a la validación cruzada y pruebas del modelo dentro de los conjuntos de validación y testeo dentro de valores aceptables.

3.4.2.1 Primera iteración

Se procede a realizar una primera iteración utilizando los modelos SVC, Random Forest y XGBoost.

Los features utilizados en el modelo corresponden a transformaciones realizadas con respecto a la morosidad, la marca del medidor de cada cliente, el tipo de irregularidad y la existencia de reclamos, solicitudes o consultas anteriores.

Los parámetros del modelo Random Forest se optimizan mediante el uso del método GridSearch CV (búsqueda exhaustiva en la grilla) de la librería Sklearn de Python, mientras que el modelo SVC utiliza el método RandomizedSearchCV de la misma librería, que selecciona combinaciones aleatorias de una serie dada para maximizar la métrica objetivo.

Considerando 26.095 registros, se procede a realizar una evaluación de los modelos en base a una división de conjuntos de entrenamiento (30%) y validación (70%), considerando los doce atributos generados al pre-procesar los datos.

Tabla 178: Resultados de la Primera Iteración

	Random Forest	SVC	XgBoost
Accuracy	**0.58**	**0.58**	**0.57**
Recall (Clase Positiva)	**0.58**	**0.57**	**0.58**
Recall (Clase Negativa)	**0.58**	**0.58**	**0.55**
Precision (Clase Positiva)	**0.58**	**0.58**	**0.56**
Precision (Clase Negativa)	**0.58**	**0.58**	**0.57**
F1 (Clase Positiva)	**0.58**	**0.58**	**0.57**
F1 (Clase Negativa)	**0.58**	**0.57**	**0.56**
Mean (5 fold - Cross Val)	**0.55**	**0.546**	**0.544**
Std (5 fold - Cross Val)	**0.035**	**0.033**	**0.033**
F1 en train set	**0.579**	**0.586**	**0.55**
F1 en test set	**0.576**	**0.580**	**0.56**
AUC	**0.603**	**0.583**	**0.562**

Adicionalmente, se obtienen los pesos de las features en el modelo para estudiar futuros posibles cambios.

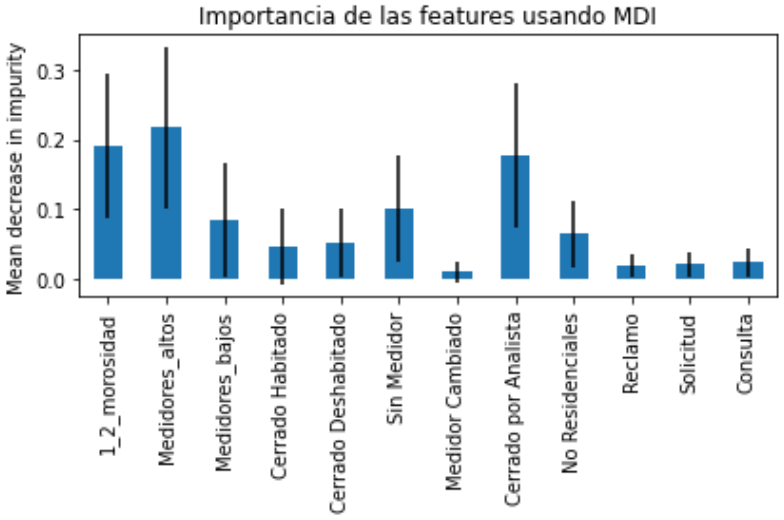

Ilustración 30: Importancia de las features en el modelo Random Forest

Se observan que las features más importantes para el modelo Random Forest son la morosidad del cliente (que en este caso se determina que la mejor forma de tratarla es si el cliente ha cancelado en al menos los dos meses anteriores), la marca del medidor y algunas irregularidades que se puedan dar.

En general, se observan rendimientos similares para los modelos Random Forest y SVC, pero se prefiere el primero debido a que el tiempo de ejecución es considerablemente menor.

En general, todos los modelos se comportan de mejor forma que el baseline declarado anteriormente, pero se continua a la siguiente iteración para obtener mayor poder predictivo.

3.4.2.2 Segunda iteración

La tarea pendiente en la iteración anterior es el incluir la ubicación dentro de los modelos, lo que intuitivamente resulta de gran importancia para predecir la probabilidad de lectura.

Se tienen datos espaciales como las rutas, zonas y sectores, que ya están procesados y son potencialmente útiles. Las rutas son 5075, las zonas son 594 y los sectores son 20.

El problema surgió al probarlos de forma empírica en los modelos señalados, ya que producían un notorio sobreajuste por separado. La razón de lo anterior se presume que es en base al número de valores únicos que tiene cada uno, lo que produciría un exceso de atributos que provocarían un sobreajuste, lo que se comprueba al comparar los rendimientos en los conjuntos de entrenamiento y prueba, aumentando la complejidad de los modelos utilizando los parámetros correspondientes.

Entonces la tarea principal en esta iteración es determinar una forma de incluir la ubicación en los modelos sin producir sobreajuste y mejorando su métrica objetivo.

Como también se tienen las latitudes y longitudes, generadas meticulosamente para construir los tableros de visualización de la sección anterior, se procede a investigar métodos de aprendizaje automático no supervisado que permitan clusterizarlas, para así generar un features para integrar a los modelos de clasificación.

Se evalúan dos métodos eficientes para realizar la tarea, que corresponden a los algoritmos DBSCAN y OPTICS.

Como el algoritmo DBSCAN entrega resultados bajo el supuesto de que la densidad es constante, es muy probable que no tenga buenos resultados, dado que, en una ciudad de la magnitud de Santiago de Chile, no se cumple dicho supuesto en todas las ubicaciones espaciales. Por lo anterior se utiliza el algoritmo OPTICS, que garantiza una formación de clústeres más precisa, a pesar de que no se pueda configurar de forma más flexible como DBSCAN, al no poder fijar un radio para encontrar los neighbors.

Por otro lado, se debe tomar la decisión de si realizar un solo modelo general o realizar un número mayor de modelos donde se entrene cada uno en un segmento específico de clientes.

La decisión anterior se tomará utilizando la experimentación, para lo cual primero se utiliza un modelo OPTICS sobre toda la región metropolitana, cuya clusterización resulta como sigue.

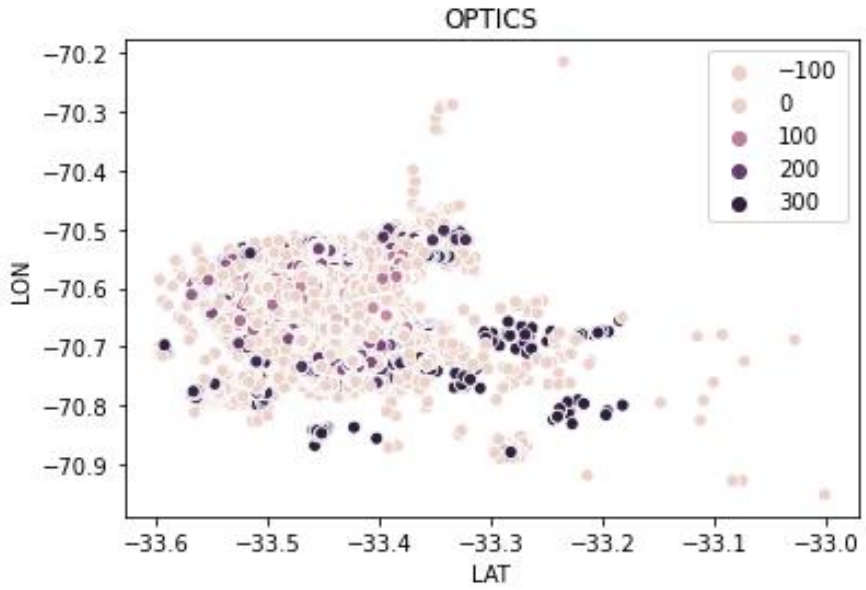

Ilustración 31: Resultados de clusterización sobre toda la región metropolitana

Cabe destacar que el parámetro asociado al modelo (número de vecinos) se ajusta de forma iterativa mediante búsqueda exhaustiva, procurando maximizar la métrica objetivo (F1 de la clase positiva). Así, el número de vecinos utilizados para la clusterización finalmente es de 20.

Cabe destacar que experimentalmente, en concordancia con expuesto en la subsección anterior, se muestra que el modelo Random Forest posee una mejor efectividad, al ser más rápido de procesar y entregar mejores resultados. Lo anterior debido a que permite la optimización de sus hiperparámetros utilizando la búsqueda exhaustiva por grilla de sus parámetros en un tiempo razonable (desde 2 a 3 horas), en comparación como modelos como el SVC, cuyos tiempos de ejecución se observa que son considerablemente más altos. Adicionalmente el modelo de bosque aleatorio permite conocer los pesos de los features por parte del modelo para la toma de decisiones iterativas. Por lo anterior, que desde esta iteración se utiliza solo dicho algoritmo.

Los resultados obtenidos al respecto son los que siguen.

Tabla 189: Resultados modelo inicial

	Random Forest
Accuracy	**0,59**
Recall (Clase Positiva)	**0,65**
Recall (Clase Negativa)	**0,53**
Precision (Clase Positiva)	**0,57**
Precision (Clase Negativa)	**0,61**
F1 (Clase Positiva)	**0,61**
F1 (Clase Negativa)	**0,57**
Mean (5 fold - Cross Val)	**0,561**
Std (5 fold - Cross Val)	**0,02**
AUC	**0,62**

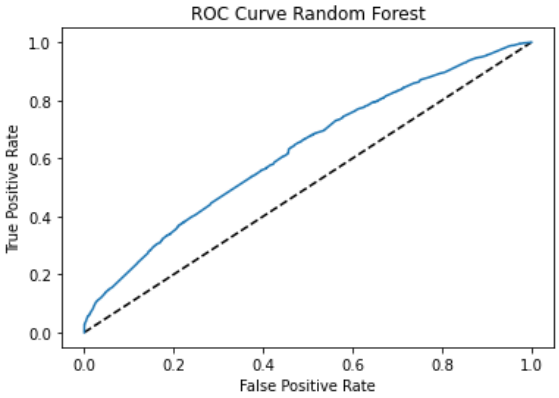

Ilustración 32: Curva ROC para modelo inicial

Se observan mejores resultados con respecto al modelo utilizado en la subsección anterior, aunque la capacidad de generalización disminuye ligeramente, además el desempeño en la clase positiva.

Por lo anterior, se planeará entrenar distintos modelos para los distintos segmentos posibles de clientes con los siguientes enfoques:

1. Según distritos presentes en la ciudad
2. Según el IPS utilizado en el capítulo anterior

Para el primer enfoque se utiliza la siguiente distribución.

- Distrito Norte (4.011 registros): Renca, Conchalí, Huechuraba, Recoleta, Til-Til, Colina, Independencia, Quilicura
- Distrito Oriente (4.893 registros): Vitacura, Lo Barnechea, Las Condes, La Reina, Peñalolén
- Distrito Poniente (4.984 registros): Pudahuel, Maipú, Estación Central, Cerrillos, Quinta Normal, Cerro Navia, Lo Prado, Pedro Aguirre Cerda, Lo Espejo
- Distrito Sur (3.590 registros): La Florida, La Granja, La Cisterna, San Ramón
- Distrito Centro (9.618 registros): Santiago, Providencia, Ñuñoa, San Miguel, San Joaquín, Macul

Se procede generando subconjunto de los datos para cada distrito y determinando el número óptimo de clústeres a través de la experimentación para cada distrito.

Los resultados son del algoritmo OPTICS para el distrito centro se muestran en la ilustración que sigue y los otros distritos se detallan en el anexo L.

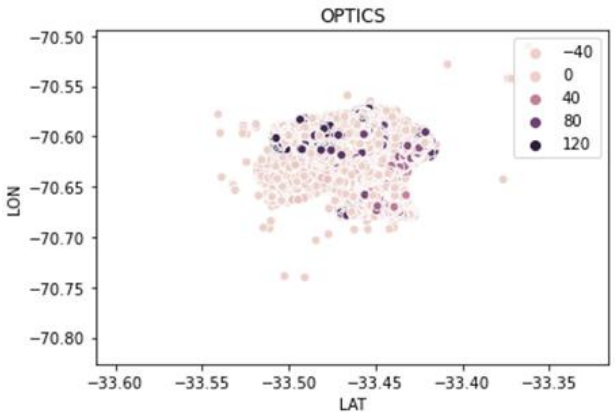

Ilustración 33: Aplicación de OPTICS para el distrito centro

Los resultados de la ejecución del modelo para todos los distritos son los que siguen.

Tabla 20: Resultados para modelos con enfoque de distritos

	Distrito Centro (N Neighbors: 12)	Distrito Norte (N Neighbors: 35)	Distrito Sur (N Neighbors: 20)	Distrito Oriente (N Neighbors: 30)	Distrito Poniente (N Neighbors: 15)
Recall (Clase Positiva)	**0,35**	**0,79**	**0,34**	**0,30**	**0,59**
Recall (Clase Negativa)	**0,84**	**0,38**	**0,82|**	**0,86**	**0,64**
Precision (Clase Positiva)	**0,65**	**0,6**	**0,58**	**0,66**	**0,65**
Precision (Clase Negativa)	**0,59**	**0,61**	**0,63**	**0,56**	**0,58**
F1 (Clase Positiva)	**0,45**	**0,68**	**0,68**	**0,41**	**0,62**
F1 (Clase Negativa)	**0,69**	**0,47**	**0,44**	**0,68**	**0,61**
Mean (5 fold - Cross Val)	**0,53**	**0,566**	**0,57**	**0,54**	**0,53**
Std (5 fold - Cross Val)	**0,03**	**0,04**	**0,07**	**0,03**	**0,08**
AUC	**0,61**	**0,62**	**0,62**	**0,64**	**0,67**

Los modelos asociados a los segmentos más vulnerables, como lo son el distrito norte y el sur no entregan resultados favorables en relación al caso anterior.

Cabe destacar que las curvas ROC asociadas a cada modelo se encuentran en el anexo L.

Por otro lado, se aplica el enfoque basado en la prioridad social estudiado en capítulo anterior. Los resultados son del algoritmo OPTICS para el segmento de IPS alto y los otros distritos se detallan en el anexo M.

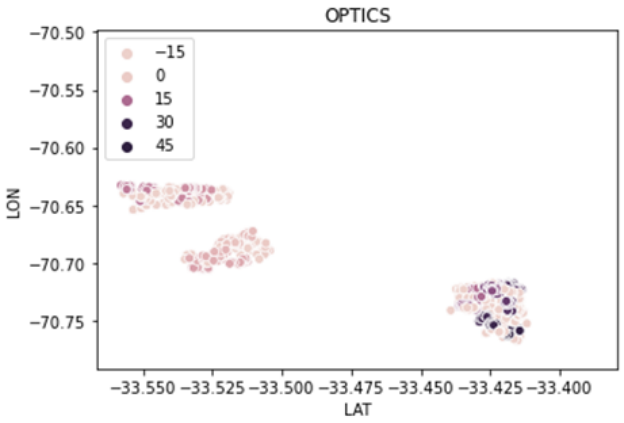

Ilustración 34: Aplicación de OPTICS para segmento IPS alto

Los resultados de la ejecución del modelo son los que siguen.

Tabla 21: Resultados de modelos con enfoque de segmentos según IPS

	IPS Alto (N Neighbors: 8)	**IPS Medio Alto** (N Neighbors: 12)	**IPS Medio Bajo** (N Neighbors: 20)	**IPS Bajo** (N Neighbors: 28)	**Sin Prioridad** (N Neighbors: 30)
Recall (Clase Positiva)	**0,92**	**0,57**	**0,83**	**0,61**	**0,38**
Recall (Clase Negativa)	**0,33**	**0,60**	**0,35**	**0,57**	**0,80**
Precision (Clase Positiva)	**0,6**	**0,64**	**0,65**	**0,58**	**0,60**
Precision (Clase Negativa)	**0,79**	**0,67**	**0,58**	**0,59**	**0,62**
F1 (Clase Positiva)	**0,74**	**0,60**	**0,68**	**0,58**	**0,47**
F1 (Clase Negativa)	**0,53**	**0,63**	**0,45**	**0,59**	**0,70**
Mean (5 fold - Cross Val)	**0,57**	**0,57**	**0,59**	**0,51**	**0,55**
Std (5 fold - Cross Val)	**0,06**	**0,07**	**0,06**	**0,06**	**0,04**
AUC	**0,63**	**0,61**	**0,66**	**0,63**	**0,63**

Las curvas ROC asociadas a estos modelos se encuentran en el anexo M.
Este modelo presenta un mejor rendimiento promedio para la clase positiva, que es la de interés. Además, este mejor rendimiento se centra en las clases más vulnerables. Aun así, el modelo se comporta de mala forma en los segmentos menos vulnerables, lo que puede limitar su uso.

Por lo anterior, se concluye que la utilización de este modelo es más beneficiosa para la empresa, por lo que se le recomienda guiarse por la probabilidad de reclamo que entrega, que constituye una mejor herramienta que la utilizada actualmente en la empresa, sin criterios claros hacia donde focalizar sus acciones preventivas

Cabe destacar que un ejemplo de entregable a la empresa a partir de estos resultados es el que se muestra en la siguiente ilustración.

	Numero Cliente	Sector	Probabilidad Lectura	lectura	Zona	Ruta	Tarifa	Medidor	Marca	Corr	Dirección	Comuna	Meses Cerrad	regularidad
0	222248	1	44,489096%	0	155	540	BT3	19528148	VEC	1	LA HABANA	LO ESPEJO	6	5
1	1158593	1	54,448377%	1	167	3582	BT1	5248100	GAN	1	FRENTE A AV	INDEPENDE	6	6
2	2879551	1	55,845879%	1	182	6136	BT1	23267312	CP4	1	CERRO TROI	QUILICURA	6	6
3	1003739	1	44,489096%	0	197	4370	AT43	299205	ABB	1	TOTORAL BA	RENCA	6	5
4	3103143	1	44,815399%	0	215	5801	BT43	7385443	EMH	1	LAS REJAS N	LO PRADO	6	5
5	3186805	1	47,237791%	0	242	3975	BTEP	4719688	EMH	1	JOSE JOAQL	PUDAHUEL	6	6
6	2695323	1	49,946666%	0	264	2030	BT3	35424	ACT	1	TENIENTE C(	PUDAHUEL	6	5
7	1087889	1	53,323843%	1	264	5010	BT1	8842456	CCM	1	SERRANO 8(	PUDAHUEL	6	6
8	1266458	1	53,041262%	1	276	1700	BT1	7,01E+08	ENL	1	LO CASTILL(	PUDAHUEL	6	6
9	10735	1	50,635374%	1	280	1860	BT43	6,13E+08	STR	1	LO LOPEZ 1(	CERRO NAV	6	6
10	16026	1	46,925775%	0	371	4585	BT1	97229965	CCM	1	5 NORTE 54	LA GRANJA	6	6
11	1229618	1	49,958558%	0	372	4810	BT1	157086	OSK	1	PADRE JUAI	LA GRANJA	6	5
12	1091396	1	55,991089%	1	481	1060	BT1	1334377	CPL	1	CATAPILCO	MAIPU	6	6
13	6632	1	53,041262%	1	482	3883	BT1	2520106	FAE	1	TUTUQUEN	MAIPU	6	6
14	1511989	1	46,925775%	0	503	2020	BT1	98058815	CCM	1	OBISPO RAI	MAIPU	6	6
15	3501892	1	50,696427%	1	542	1875	BT1	7,01E+08	ENL	1	SERVIDUME	COLINA	6	6
16	3012189	1	53,097658%	1	623	580	AT43	96973487	LAG	1	RECOLETA 2	RECOLETA	6	6
17	219337	1	53,660844%	1	869	1160	BT1	166907	SIN	2	NSTRA SRA I	VITACURA	6	5
18	310077	1	49,221479%	0	871	2650	BT1	95169123	CCM	1	ARIZONA 11	VITACURA	6	6
19	486276	1	50,635374%	1	873	540	BT3	7778	ACT	1	P HURTADO	VITACURA	6	6
20	800846	1	51,001317%	1	886	2780	BT3	11434428	VEC	1	VITACURA 9	VITACURA	6	5
21	530139	1	55,845879%	1	892	2120	BT1	5057270	GEN	1	ISLAS FIDJI	VITACURA	6	6
22	3366261	1	28,211127%	0	897	1571	BT43	6049739	EMH	1	CAMOENS 6	VITACURA	6	3
23	754326	1	45,602483%	0	903	1890	BT3	488678	EMH	1	MANQUEHL	VITACURA	6	5
24	754713	1	45,602483%	0	913	562	BT3	70488	ACT	1	VITACURA 7	VITACURA	6	5
25	312312	1	49,185462%	0	942	1220	BT43	6,13E+08	STR	1	VIA GRIS 98	VITACURA	6	6
26	2616832	1	54,731601%	1	167	3610	BT1	1234507	CPL	2	FRENTE A AV	INDEPENDE	7	6
27	372789	1	43,647619%	0	248	970	BT1	32036631	SAN	1	BRAVO LUC	PUDAHUEL	7	5
28	503830	1	49,958558%	0	248	3251	BT1	2364181	CPL	1	SERRANO 1	PUDAHUEL	7	5
29	1248014	1	53,268834%	1	264	2100	BT1	2520549	FAE	1	TENIENTE C(	PUDAHUEL	7	5
30	1055031	1	54,448377%	1	408	2090	BT1	5035463	GAN	1	PASAJE 8 72	LA FLORIDA	7	6

Ilustración 35: Ejemplo de entregable para la empresa

3.4.3 Conclusiones

Introducir estos dos enfoques permite agregar mayor flexibilidad al aplicar los modelos y mejorar el poder predictivo.

A partir de la comparación de estos dos enfoques, se puede observar que un enfoque por IPS funciona mejor que uno por distritos, lo que se puede producir dado que las comunas contiguas no necesariamente van a compartir características similares. Por ejemplo, podría compararse Macul y Providencia, que pertenecen al mismo distrito, pero sus realidades y densidades pueden ser muy distintas. El IPS permite agrupar clientes formando segmentos en base a un estudio complejo ya realizado, lo que permite que el algoritmo OPTICS genere mejores resultados al formar los clústeres en ambientes más similares.

A pesar de lo anterior, los resultados pueden verse afectados, debido a que hay ciertas variables relevantes omitidas que se dan como ciertas al utilizar esta forma de proceder. Por ejemplo, puede haber comunas que se compongan de segmentos de clientes muy distintos, y pueden tener poblaciones con mucha densidad y otras con muy poca, lo que no se puede incluir en el modelo, por lo tanto, es menester que la empresa tenga mejor caracterizados a los clientes para mejorar la capacidad predictiva.

Si bien el poder predictivo logrado es superior al baseline con el que se actúa en la empresa, podría mejorarse a partir de la inclusión de consumo y de pago de los clientes en los modelos, pero a la fecha de entrega de este trabajo no se encontraron relaciones claras.

CAPÍTULO 4: SOPORTE A LAS MEDIDAS CORRECTIVAS DEBIDO A CAMBIO DE LICITACIÓN DE EMPESAS ENCARGADAS DE LA LECTURA

4.1 ENTENDIMIENTO DEL PROBLEMA

Antes del mes de diciembre, el trabajo de lectura en terreno era realizada por dos empresas externas a Enel Distribución Chile, Helpbank y Provider. Cada una poseía alrededor de ochenta técnicos lectores cada una, para los alrededor de 1.450.000 clientes cuyos medidores eran convencionales. Dichos técnicos lectores poseían cierta expertiz en la tarea de registrar el consumo en los medidores, ya que, gran parte poseían tiempo considerable a la empresa y acostumbramiento a las rutas y las ubicaciones espaciales de los medidores de cada cliente en su vivienda o local comercial respectivo.

Por normativa del regulador, la lectura debe ser realizada todos los meses para todos los clientes. En base a esto, se define un indicador de productividad mensual por sector como sigue:

$$Productividad_{Sector_i} = \frac{Número\ de\ Clientes\ Visitados_{Sector_i}}{Clientes\ Totales_{Sector_i}} * 100$$

Este indicador mide la proporción de medidores cuyos registros se midieron, o bien, se intentaron medir por parte de la empresa. Como este indicador debe ser de un 100% para cumplir la regulación, la empresa le realiza seguimiento periódico. La lectura realizada con dichas empresas externas permitía mantener niveles altos durante la pandemia COVID-19.

Durante el mes de diciembre se anuncia el cambio de licitación, donde la empresa Helpbank se ve desvinculada, lo que genera un cambio en la dotación de personal de técnicos lectores, de la cual gran parte poseía gran experiencia en el cargo.

En lugar de Helpbank, la empresa externa Cobra asume su rol y debió contratar nuevos técnicos lectores, los que no poseían la experiencia de los anteriores. Cabe destacar que existe un número fijo de técnicos lectores para cada empresa, por lo que se deben optimizar los recursos para poder cumplir con una productividad cercana al 100%

A fines de marzo, esta productividad era de alrededor del 70% para los distintos sectores según la contraparte, vale decir, no se estaba cumpliendo con la regulación debido a esta situación. Adicionalmente, esta situación generaba viviendas ‘cerradas’, considerándolo como una situación similar a la definida en el capítulo anterior, pero se consideraban cerradas desde 2 a 4 meses al mes de abril de 2021. Lo anterior implica que su consumo no había podido ser registrado en ningún mes posterior diciembre de 2020, enero de 2021 o febrero de 2021.

Debido a lo anterior el problema del aumento del número de clientes con medidores convencionales cuyos consumos no han podido ser registrados desde dos a cuatro meses anteriores a abril de 2021. Estas viviendas son 57.035.

Los efectos de este problema son directamente que se aumenta el número de reclamos debido a la imposibilidad de lectura y el actual modelo de cobro en estos casos, demostrado como ineficaz en el capítulo 2.

Para solucionar esta problemática se construye una herramienta que integra un enfoque descriptivo de datos, además de lógicas supervisadas y no supervisadas.

Cabe destacar que, el contexto en el que se desarrolló en este capítulo es en uno de urgencia, ya que se solicita presentar una solución en un plazo limitado. Por lo tanto, lo formulado en este apartado se realiza en cuatro días hábiles y tres días no hábiles.

4.2 ADQUISICIÓN, DESCRIPCIÓN Y PRE-PROCESAMIENTO DE LOS DATOS

Se solicita a la empresa datos asociados a los clientes con esta situación, en un formato similar al que se tenía en el capítulo anterior.

En este caso los datos se adquieren directamente utilizando una planilla que posee la data de estos clientes cerrados.

4.3 MODELAMIENTO

En este caso, se considera un modelo simplificado al obtenido en el capítulo 3, debido a, tanto la disponibilidad de datos, como a la intención de obtener una mejor generalización del modelo, eliminando ciertos atributos que se vislumbra que podrían variar considerablemente considerando esta situación.

Se trabajará bajo el supuesto de que las variables que influyen en la lectura de una vivienda son las mismas para las variables que se seleccionen, por lo que, mediante heurísticas, se incluirán en el modelo solo las variables que se sospeche que puedan proveer una buena capacidad de generalización.

Se desecha utilizar como atributo a los clústeres generados por el modelo OPTICS, ya que, el modelo inicial se confeccionó en octubre de 2020 y este modelo se entrena a inicios de 2021. Por lo anterior, como precaución no se actúa bajo el supuesto de que dichos clústeres se mantienen en el tiempo, debido a que la causa de que los medidores de esas viviendas no hayan podido ser leídos es completamente distinta. Este enfoque se utilizará en la siguiente sección.

Por otro lado, tampoco se utilizan datos asociados a los reclamos, debido a que los datos no alcanzaron a estar disponibles en la fecha requerida.

Se agregan los datos pertenecientes a la base de medidores descrita en el capítulo 2, debido a que a la fecha de realización de este capítulo se tiene acceso a ella. Esto permite utilizar data más adecuada con respecto a los medidores, que solo la marca utilizada en el capítulo anterior.

Con lo anterior el modelo posee dos atributos relacionados a la caracterización del medidor, que es su antigüedad en años y si este es electrónico o electromecánico. Estos atributos se determinan utilizando test chi-cuadrado con respecto a la variable target. Por otro lado, datos asociados a las irregularidades presentes y la morosidad de cada cliente.

Se utiliza el mismo enfoque de entrenamiento de cada modelo según cada segmento por IPS, debido a que las features afectan de distinta forma a los distintos segmentos. Cabe destacar que no se trabaja con el segmento sin prioridad debido a que la data está incompleta para las comunas asociadas.
Con esto se obtienen los siguientes resultados para cada segmento, utilizando un modelo Random Forest cuyos hiperparámetros han sido optimizados utilizando la búsqueda exhaustiva por grilla.

Tabla 19: Resultados de modelos de lectura.

	IPS Alto	IPS Medio Alto	IPS Medio Bajo	IPS Bajo
Recall (Clase Positiva)	**0,67**	**0,58**	**0,68**	**0,59**
Recall (Clase Negativa)	**0,40**	**0,56**	**0,49**	**0,59**
Precision (Clase Positiva)	**0,52**	**0,55**	**0,6**	**0,59**
Precision (Clase Negativa)	**0,56**	**0,59**	**0,57**	**0,58**
F1 (Clase Positiva)	**0,59**	**0,57**	**0,64**	**0,59**
F1 (Clase Negativa)	**0,47**	**0,57**	**0,53**	**0,58**
Mean (10 fold - Cross Val)	**0,61**	**0,56**	**0,61**	**0,53**
Std (10 fold - Cross Val)	**0,04**	**0,05**	**0,06**	**0,07**
AUC	**0,61**	**0,58**	**0,63**	**0,62**

Las curvas ROC asociadas se encuentran en el anexo N, y los pesos de las features en los modelos en el anexo Ñ.

Si bien los modelos fueron entrenados con una cantidad muy limitada de atributos debido a la disponibilidad de información y urgencia del proyecto, se obtiene un poder predictivo que es notablemente que un modelo aleatorio. Además, no se observa gran variabilidad en los modelos.

A pesar de lo anterior, es necesario reforzar la toma de decisiones con otros métodos analíticos, como lo son el uso de herramientas descriptivas y algoritmos de aprendizaje no supervisado.

Un aspecto que llama la atención en este caso es el gran peso que posee la antigüedad del medidor dentro de los modelos, lo que, en conversaciones con la contraparte, se conjetura que pueden deberse a que, a partir de cierto rango, los medidores pueden facilitar la lectura. Este rango se produce en los medidores más antiguos según las correlaciones calculadas con la variable target.

Cabe destacar que un entregable realizado a la empresa dados los resultados de este modelo es el que se muestra en la siguiente ilustración.

Numero Clien	Prob Lectu	Predicci	Sect	Zona	Ruta	Tarifa	Medid	Marca	Corr	Constar	Direcci	Comu	es Factu	Clave	Irreg	Coordenadas GPS	Coordenadas GPS
284943	62,2%	1	1	73	542	BT1	7747277	MIT	1	1	SALTILLO	LO ESPEJO	3	J	6	343737,56374	6289901,2854
285012	58,7%	1	1	73	1292	BT1	97219033	CCM	1	1	MEXICO 0	LO ESPEJO	3	J	6	343943,09778	6289716,058
285013	58,7%	1	1	73	1302	BT1	97202120	CCM	1	1	MEXICO 0	LO ESPEJO	4	J	6	349936,11386	6287712,4272
285014	52,6%	1	1	73	1312	BT1	200363	SIN	1	1	MEXICO 0	LO ESPEJO	4	J	6	349935,81298	6287711,8679
285075	41,8%	0	1	73	1942	BT1	8854083	CCM	1	1	PATAMBA	LO ESPEJO	3	J	97	349993,47116	6287409,232
285106	58,9%	1	1	73	2272	BT1	5046474	GAN	1	1	PATAMBA	LO ESPEJO	3	J	6	350031,84239	6287181,9016
1227821	50,6%	1	1	73	2532	BT1	98135765	CCM	1	1	PRESIDEN	LO ESPEJO	4	J	6	350040,92584	6287045,9886
591168	65,0%	1	1	73	3432	BT1	7780083	MIT	1	1	PRESIDEN	LO ESPEJO	4	J	6	350180,44976	6286535,1844
591169	59,3%	1	1	73	3442	BT1	52937632	LAN	1	10	PRESIDEN	LO ESPEJO	4	J	6	350197,68475	6286472,0469
591174	63,5%	1	1	73	3492	BT1	33043806	SAN	1	10	PRESIDEN	LO ESPEJO	4	J	6	350252,82851	6286260,5049
591225	53,4%	1	1	73	3982	BT1	24098470	CP4	1	1	PRESIDEN	LO ESPEJO	4	J	6	350396,34387	6286118,7342
591302	53,4%	1	1	73	4742	BT1	24098038	CP4	1	1	PRESIDEN	LO ESPEJO	4	J	6	350432,70984	6286119,1148
2957426	52,6%	1	1	73	5032	BT1	24874859	CPL	1	1	HERMOSIL	LO ESPEJO	4	J	6	350545,79928	6286122,3512
591564	57,5%	1	1	73	5822	BT1	96121174	CCM	1	1	MATAMO	LO ESPEJO	3	J	6	350723,77783	6286123,081
12947	50,3%	1	1	74	90	BT1	25647989	CPL	1	1	BUENAVE	LO ESPEJO	4	J	6	346543,73964	6296198,6809
12950	62,0%	1	1	74	180	BT1	97128249	CCM	1	1	BUENAVE	LO ESPEJO	2	J	6	343445,56045	6291071,7144
12991	54,6%	1	1	74	1470	BT1	4893096	SAN	1	1	BUENAVE	LO ESPEJO	3	J	99	343305,29352	6291107,5164
12994	55,3%	1	1	74	1560	BT1	125055	OSK	1	1	BUENAVE	LO ESPEJO	2	J	99	343283,49838	6291114,5564
13024	49,6%	0	1	74	2520	BT1	8060828	ABB	1	1	5 DE MAY	LO ESPEJO	4	J	6	343500,63174	6291142,1171
13025	62,0%	1	1	74	2550	BT1	97180922	CCM	1	1	5 DE MAY	LO ESPEJO	2	J	6	343502,99008	6291139,937
13043	46,2%	0	1	74	3150	BT1	4880164	SAN	1	1	5 DE MAY	LO ESPEJO	3	J	6	343374,03946	6291174,9994
13060	62,0%	1	1	74	3720	BT1	97128097	CCM	1	1	5 DE MAY	LO ESPEJO	3	J	6	343209,87536	6291242,208
13087	46,2%	0	1	74	4590	BT1	4886803	SAN	1	1	5 DE ABRI	LO ESPEJO	4	J	6	346543,73964	6296198,6809
13110	43,5%	0	1	74	5230	BT3	21176635	VEC	1	1	MAIPU 64	LO ESPEJO	3	J	6	343463,53483	6291271,2878
13135	50,6%	1	1	74	6000	BT1	97081190	CCM	1	1	MAIPU 64	LO ESPEJO	3	J	6	343461,96039	6291244,2724
13150	46,2%	0	1	74	6450	BT1	3207013	SAN	1	1	MAIPU 66	LO ESPEJO	3	J	6	343400,13433	6291122,5529
13228	62,0%	1	1	74	8820	BT1	97128939	CCM	1	1	SANTIAGO	LO ESPEJO	4	J	6	343497,28053	6291224,5111
2566647	50,3%	1	1	75	672	BT1	106221	OSK	1	1	FRENTE A	LO ESPEJO	4	J	6	348276,47576	6287517,0803
548195	48,3%	0	1	75	982	BT1	10078234	COS	1	1	PJE CADE	LO ESPEJO	4	J	6	348261,02129	6287711,8622
3010376	50,3%	1	1	75	1122	BT1	25675873	CPL	1	1	CADEPINO	LO ESPEJO	4	J	6	348253,41845	6287820,9936
548253	40,4%	0	1	75	1562	BT1	24096227	CP4	1	1	PJ.CODEH	LO ESPEJO	3	J	6	348500,50887	6287927,6667
548261	40,4%	0	1	75	1642	BT1	24096230	CP4	1	1	PJE CODE	LO ESPEJO	3	J	6	348576,09609	6287924,9739
548262	40,4%	0	1	75	1652	BT1	24096228	CP4	1	1	PJE CODE	LO ESPEJO	3	J	6	348583,98337	6287925,4677
1060209	40,4%	0	1	75	1792	BT1	24097935	CP4	1	1	CADEHUA	LO ESPEJO	3	J	6	348664,01227	6287925,8019

Ilustración 36: Muestra de planilla entregada a la empresa

4.4 SOLUCIÓN DESCRIPTIVA E INTEGRACIÓN

Es de vital importancia para la empresa conocer donde está la mayor concentración de viviendas que no han podido ser leídas, para tomar acciones preventivas como asignar mejor la cantidad de técnicos lectores a trabajar en dicha zona.

Por lo anterior, y para reforzar el scoring obtenido, se plantea generar un tablero de visualización para este problema en específico, que incluya las salidas provenientes, tanto del modelo anterior, como de un modelo de aprendizaje no supervisado que permita detectar mediante criterios objetivos las localizaciones donde se encuentran los centros de densidad.

4.4.1 Propósito y audiencia

La audiencia directa son los analistas tanto de la empresa como subcontratados que puedan influir en la asignación de los lectores.

El propósito del tablero es permitir una asignación mediante criterios objetivos que tengan real influencia en la probabilidad de lectura de una vivienda. Así, se podrá acelerar la disminución de los casos problemáticos de forma más rápida.

4.4.2 Planificación de la herramienta

El tablero será diseñado para ser utilizado en un computador, por lo que tendrá las dimensiones estándar, de 1366 x 1024 pixeles.

Sus requerimientos funcionales principales serán los que siguen:

- La herramienta permitirá observar las principales métricas de negocio según distintos filtros que sean seleccionados por el usuario
- La herramienta mostrará centros de densidad y seleccionarlos. Mediante un selector puede ver distancias reales en cada centro.
- La herramienta permitirá mostrar la probabilidad de lectura para las viviendas que seleccione el usuario

4.4.3 Construcción de la herramienta

Para construir esta herramienta, como en su lógica de negocios incluye outputs de distintos modelos.

Para lo anterior se concatenan los resultados de los cuatro modelos anteriores con los datos necesarios para los indicadores a desplegar.

Adicionalmente, para poder visualizar en un mapa en el software Tableau se requieren las coordenadas geográficas decimales de cada vivienda. En esta ocasión no se puede utilizar la API de Google Maps mediante web-scrapping utilizando lenguaje Python como se realiza en el capítulo 3, debido al poco tiempo disponible y los costos monetarios que puede implicar convertir datos de esta dimensionalidad en poco tiempo. Por lo anterior, se solicitan las coordenadas a la empresa con anticipación, cuyo formato es UTM. Para realizar la conversión se utiliza el módulo pp de la librería Proj de Python, que se configura con la zona 19S, que es utilizada en Chile. Con esto, se obtienen coordenadas en sistema WGS84, que corresponde al sistema geodésico de coordenadas utilizado en el capítulo 3.

Para aplicar el algoritmo de aprendizaje no supervisado OPTICS, se utiliza una configuración del parámetro que permitiera que solo un tercio de las viviendas pudiera ser identificada como centro de densidad, cantidad que se consideró óptima en dicho escenario, ya que la empresa podría enfocarse correctamente. Un número mayor podría dejar la herramienta inutilizada ya que no separaría las localizaciones críticas correctamente. Así, se utilizó un número de neighbors igual a 75.

A continuación, se muestra la herramienta terminada con los principales casos de uso. Cabe destacar que la herramienta permite desplegar 30.076 casos del total debido a que de ellos se poseía la totalidad de la data requerida, como resultado no se incluyen las comunas del sector oriente y algunas poblaciones del sector sur de Santiago.

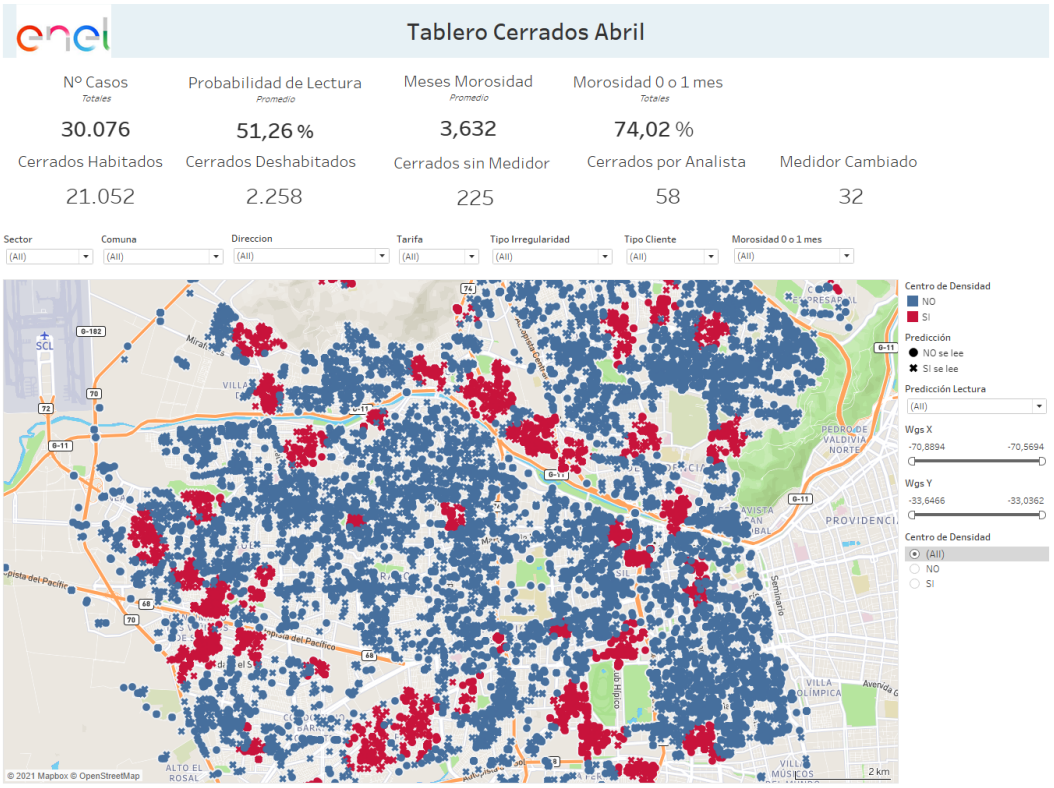

Ilustración 37: Vista general de la herramienta

El caso de uso principal corresponde a la navegación por filtros, que puede seleccionar el usuario. Por ejemplo, en la siguiente imagen se utiliza el sector 5 y la comuna de Pudahuel. Esto puede ser útil al momento de asignar personal para requerimientos específicos

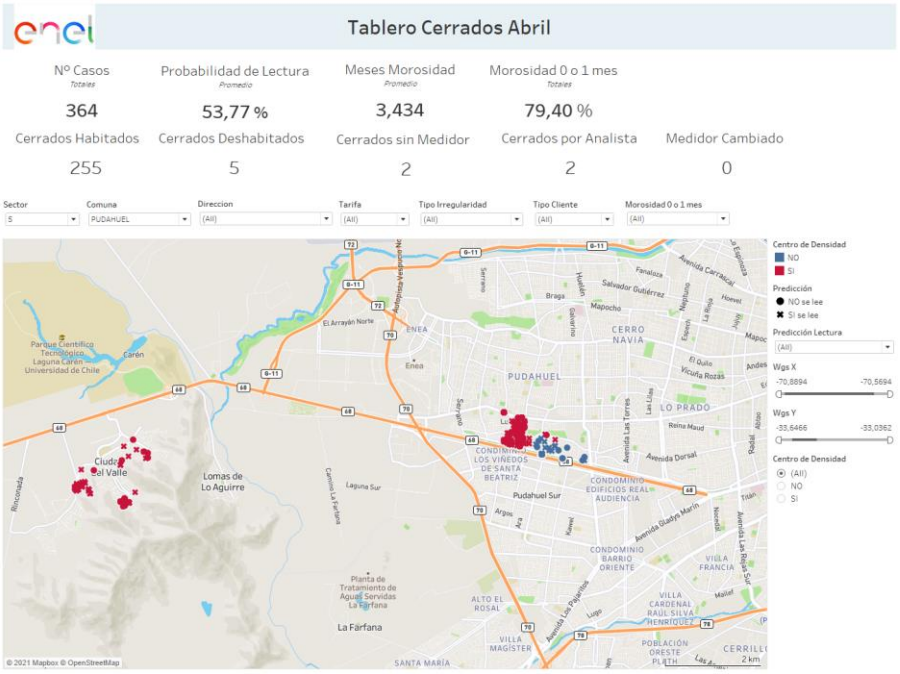

Ilustración 38: Filtros por categorías

Otro caso de uso puede ser la navegación solo por los centros de densidad, para lo cual hay una lista desplegable. La utilidad de esto es estudiar las características que comparten los centros de densidad y el asignar directamente a los lectores a cada uno, utilizando el selector radial mostrado en el capítulo 3

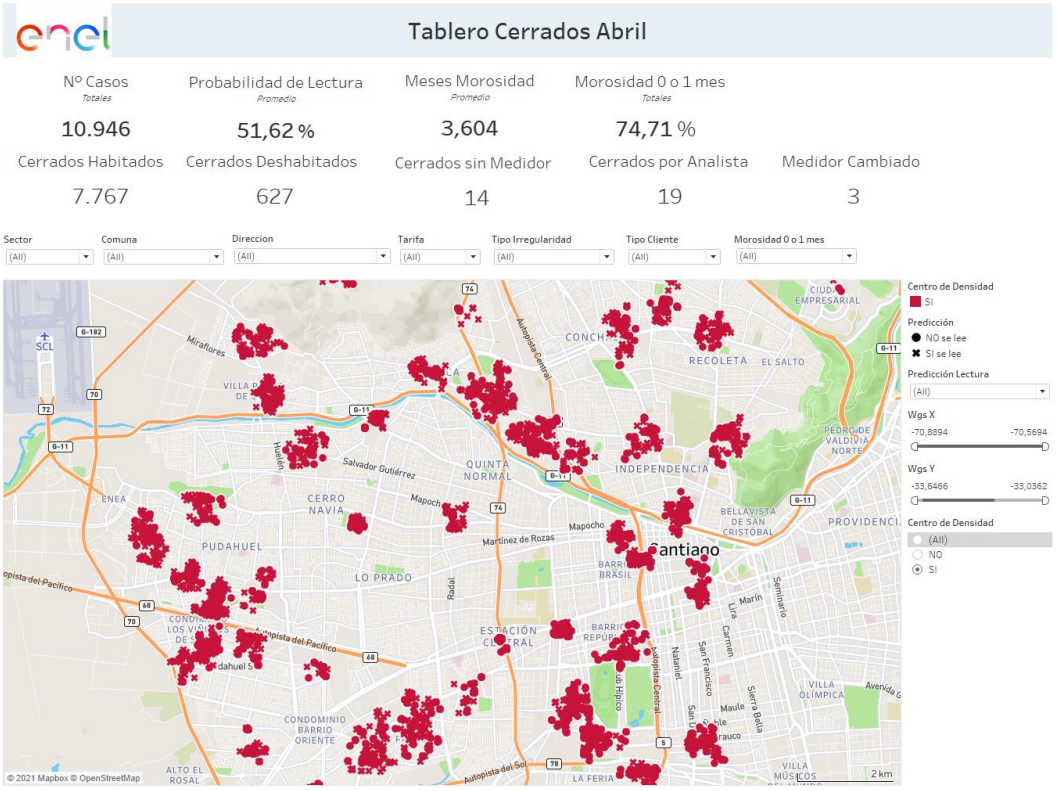

Ilustración 39: Filtros por centros de densidad

Finalmente, el tablero podrá desplegar las viviendas para las cuales el modelo indica si podrán ser leídas o no, lo que podría utilizarse para obtener conclusiones sobre los factores que influyen en la lectura general o de cada sector. Por ejemplo, utilizando el filtro es posible darse cuenta de que la morosidad es crucial, y que cuando los clientes tienen cero o un mes de morosidad, es considerablemente más probable que sus consumos sean registrados por los técnicos.

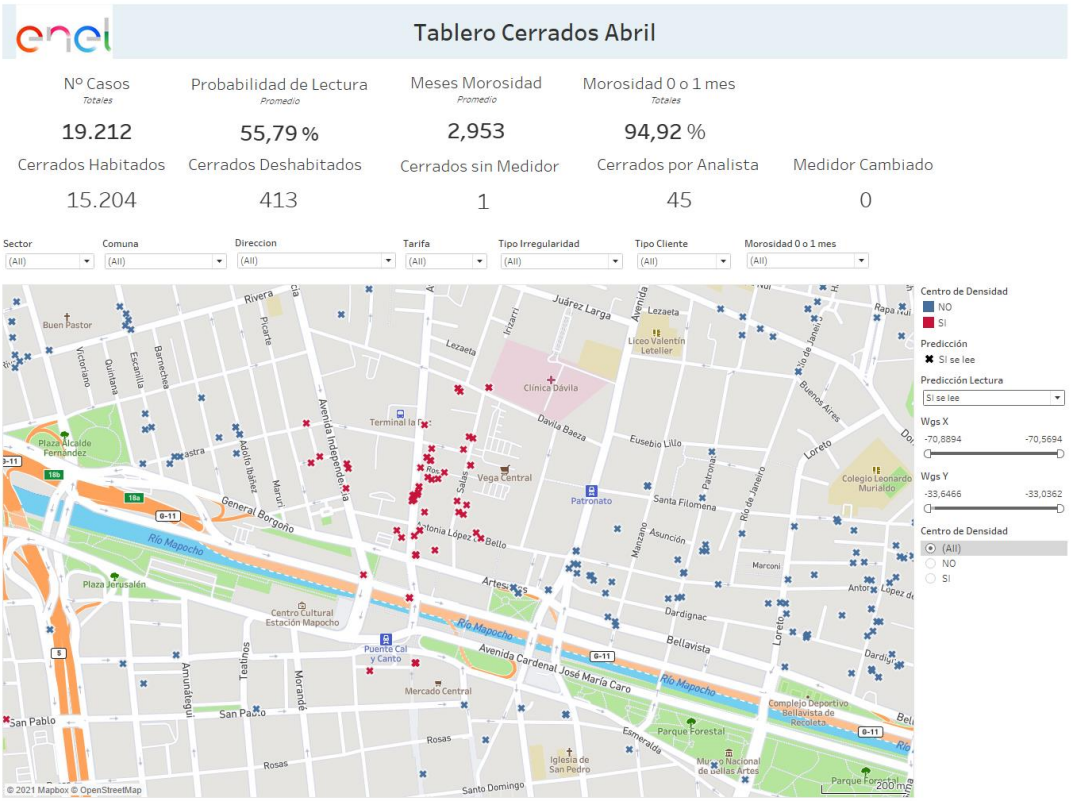

Ilustración 40: Filtros por predicción de lectura

Cabe destacar que, lamentablemente, la geocodificación que poseía la empresa no era completamente exacta, por lo que un número considerable de viviendas se geocodificaron desde 10 a 50 metros de diferencia de su ubicación original. Esto se hizo notar a los usuarios.

4.5 DISCUSIÓN Y CONCLUSIONES

Dada la madurez analítica de la organización, resulta difícil conseguir los datos de forma ágil para hacer un desarrollo rápido en ciencia de datos. En este caso, como no hay un repositorio único centralizado, se dificultó, por ejemplo, obtener datos de los reclamos asociados a los clientes de forma rápida. Estos debieron pedirse al área de marketing que no necesariamente va a responder de forma rápida a la entrega de los datos.

Por lo anterior, esta problemática debió ser abordada con límite de recursos en un tiempo acotado. Aun así, mediante la sinergia de una lógica supervisada, otra no supervisada y un enfoque descriptivo se pudo generar una herramienta que puede aportar valor a la organización de forma rápida para abordar el problema.

CAPÍTULO 5: PROPUESTA PARA CALCULAR LA PROBABILIDAD DE RECLAMOS DE CLIENTES

5.1 ENTENDIMIENTO DEL PROBLEMA

Debido a la ineficacia del modelo de cobro expuesto en el capítulo 2, se producen reclamos asociados en los meses de invierno, además de octubre. En este capítulo se aborda una forma preventiva para lidiar con esos reclamos, eliminando las causas para que no aparezcan.

Se abordan los reclamos producidos por facturación, ya que, generan la percepción de que la empresa no es honesta con sus cobros y se puede generar una mala imagen de esta.

Así es problema se define como 'alto número de reclamos debido a problemas de facturación durante los meses de invierno'. Cabe justificar que, por petición de la contraparte, el modelo también incluye a los meses de abril y mayo, debido a que en el año 2021 se corría el peligro de obtener reclamos desde esos meses dado el problema expuesto en el capítulo anterior.

La justificación está relacionada directamente con la realizada en el capítulo 2, aunque, como ahora se deben considerar los reclamos realizados directamente a Enel Distribución Chile S.A., dado que sobre estos se puede definir la variable dependiente correspondiente a si un cliente reclamó o no por motivos de facturación.

5.2 ADQUISICIÓN, DESCRIPCIÓN Y PRE-PROCESAMIENTO DE LOS DATOS

Como para abordar esta problemática, se requiere utilizar datos de todos los clientes de la empresa, los atributos disponibles son limitados. Estos son:

- Matriz de consumos de todos los clientes desde enero de 2017 a diciembre de 2020. Similar a la utilizada en el capítulo 2

- Base de medidores utilizada en los capítulos anteriores, actualizada a enero de 2021. Descritos en el capítulo 2

- Reclamos realizados desde enero a 2020 a diciembre de 2021. Descritos en el capítulo 3

Debido a esta limitación en los datos, se debe realizar ingeniería de variables sobre la matriz de consumo. Así, se podrán predecir los reclamos en base al patrón de consumo, características de los medidores y el historial de reclamo de cada cliente.

5.2.1 Preprocesamiento inicial

Tanto la variable objetivo como el modelo de datos asociados a los atributos se debe construir de forma manual.

Para lo anterior, primero se carga la base de medidores, y con el identificador único, correspondiente al número de suministro de cada cliente, se realizará el cruce de tablas con las otras dos. Con la base de medidores la primera acción a desarrollar es formar los segmentos a estudiar, que son tres en base al IPS definido en los capítulos anteriores. Se forma un segmento para los

clientes de comunas con IPS Alto o medio alto, otro con IPS medio bajo y bajo, y un último con clientes de comunas sin prioridad.
Por otro lado, de forma paralela se carga la base de reclamos y se forma un subconjunto solo con reclamos efectuados por motivo de facturación. Para luego detectar los errores de codificación presentes, como en los números de suministro.

Con estos dos DataFrame, se realiza un left-outer join para generar tanto la variable target, como un atributo que indique si existen reclamos pasados de un cliente asociados a facturación. Ambos en la base de medidores. Finalmente se forman los subconjuntos de interés correspondientes a los clientes residenciales y medidos por contratistas. Cabe destacar que los segmentos de clientes descritos se pueden operar con una entrada introducida por el usuario en el modelo, así, puede utilizar el poder predictivo para los distintos clientes. Aun así, en este trabajo se describe solo el segmento más vulnerable.

Debido a su gran tamaño y lentitud para procesar, la matriz de consumos se trabaja en un archivo distinto. Así, cada vez que se utilice el modelo, los tiempos de ejecución sean considerablemente más cortos, ya que las variables estarán formadas desde otro output.

En este caso el objetivo a perseguir es generar variables con esta matriz, que permitan explicar de buena manera los reclamos realizados por los clientes.

Para lo anterior, primero se procede a configurar el número de suministro de cada cliente como índice único en el DataFrame. Posteriormente se realiza un semi-join utilizando la base de medidores. Así, se obtiene una sub-matriz de consumos que contiene solo los clientes de interés y se disminuye el tiempo de procesamiento.

A diferencia del preprocesamiento realizado en el capítulo 2, en este caso no se eliminan los clientes con consumos cero, debido a que pueden resultar útiles para la creación de variables que describan el proceso. Se eliminan los clientes que tienen valores nulos dentro de la matriz, vale decir, que no fueron clientes en algún período dentro principios del año 2018 y marzo de 2020, para que así, las variables se creen bajo un mismo criterio. Se seleccionan estos años debido a que existe un trade-off entre tomar mayor data histórica y dejar clientes fuera del modelo, entonces evaluando el período histórico el cuál tomar de forma empírica, se determina que este período es el adecuado a estudiar. Lo anterior implica que el alcance de esta solución limita a los clientes que hayan sido clientes de la empresa en todo ese período.

El número de clientes con los que se trabaja es de 1.064.258, los que potencialmente se pueden ver beneficiados con la implementación de la solución.

5.2.2 Definición de variables

Terminado este preprocesamiento y debido a que la magnitud de los datos es particularmente alta, los algoritmos para formar las variables pueden altos, se planifican las siguientes variables de forma meticulosa, utilizando tanto el análisis de los consumos realizado en el capítulo 2 como las conversaciones tenidas con la contraparte:

- Comportamiento de lectura: Representado por el número de consumos cero para cada cliente. Un cliente cuyos consumos no pueden ser registrados a través del tiempo, es más probable que reclame

- Media móvil de los últimos 6 meses a abril de 2020: Representa el cobro que tendría el cliente de no poder ser leído en el mes siguiente

- Diferencia entre mes de mayor y menor consumo en los últimos seis meses: Corresponde a una medida que indica el rango en kW-h en que puede variar el consumo de un cliente

- Diferencia de consumos entre los meses de febrero y marzo: Representa la caída que posee el consumo desde verano a otoño

- Desviación estándar de los últimos seis meses: Indica la variabilidad total en los últimos seis meses

Una vez generadas estas variables para todos los clientes, utilizando los distintos algoritmos, se procede a generar un archivo de texto separado por comas con los resultados y los índices. Mediante un left-outer join se une con la base de medidores.

5.2.3 Selección de variables

Se utilizan las variables asociadas a la antigüedad del medidor y su tecnología, que en conjunto ya se había probado su relación con la lectura, y por lo tanto con los reclamos. Por otro lado, se utiliza la variable binaria creada anteriormente que indica la existencia de reclamos en los meses pasados del mismo año por parte de un cliente. Esta última se verifica su utilidad mediante un test chi cuadrado que es significativo (p-valor = 0).

Para seleccionar las variables creadas asociadas a los consumos, se trabaja de forma empírica, utilizando distintas combinaciones para los modelos y, mediante búsqueda exhaustiva se procura que los valores que maximicen la métrica objetivo para un número limitado de combinaciones de parámetros. Con esto se consideran las variables de comportamiento de lectura, desviación estándar de los últimos seis meses y la diferencia de consumo entre febrero y marzo de 2020

5.3 MODELOS Y MÉTRICAS DE EVALUACIÓN PROPUESTOS

Como en este escenario no hay un problema de productividad, sino de tomar medidas preventivas, y además dichas medidas preventivas deben estar focalizadas, ya que significan un costo para la empresa, como el contacto vía telefónica con el cliente, se maximizará la métrica de precisión.

Como el número de registros es de alrededor de 85.992 para el segmento en estudio, y el segmento medio posee alrededor de 600.000 registros, se determina utilizar el algoritmo Random Forest, tal como en los capítulos anteriores, como una alternativa que posee una buena relación entre poder predictivo y tiempo de ejecución, considerando que se trabaja con datos equilibrados entre variables cuantitativas y categóricas. Utilizar algoritmos más exhaustivos como SVC o XgBoost podría elevar los tiempos de ejecución de sobremanera y entorpecer la optimización de los hiperparámetros.

Considerando que, en este caso, la variable objetivo posee 25.055 valores positivos (cliente reclama) y 60.937 valores negativos, podría mejorar el rendimiento utilizar estrategias de balanceo de clases, lo que se estudiará empíricamente.

Además, la optimización de parámetros, al corresponder a un modelo que tarda un tiempo considerable en ejecutarse, se ejecuta de la siguiente forma:

1. Se ejecuta RandomizedSearchCV con una configuración variada de hiperparámetros varias veces y se registran los resultados

2. Se ejecuta el modelo Random Forest con todas las configuraciones registradas. Se registra las configuraciones que generan mejores valores de la métrica objetivo (precisión)

3. Se ejecuta GridSearchCV con las configuraciones registradas anteriormente

5.4 RESULTADOS

Los resultados obtenidos son los que siguen.

Tabla 23: Resultados Modelo de Predicción de Reclamos

	Random Forest Sin Upsampling	Random Forest Con Upsampling	Random Forest Con Upsampling e hiperparámetros optimizados
Recall (Clase Positiva)	**0,45**	**0,48**	**0,43**
Recall (Clase Negativa)	**0,99**	**0,92**	**0,99**
Precision (Clase Positiva)	**0,80**	**0,70**	**0,93**
Precision (Clase Negativa)	**0,81**	**0,81**	**0,81**
F1 (Clase Positiva)	**0,58**	**0,57**	**0,59**
F1 (Clase Negativa)	**0,88**	**0,86**	**0,89**
Mean (5 fold - Cross Val)	**0,9079**	**0,9075**	**0,9696**
Std (5 fold - Cross Val)	**0,0011**	**0,0017**	**0,0007**
Precisión en Train Set	**0,99**	**0,99**	**0,96**
Precisión en Test Set	**0,79**	**0,70**	**0,93**
AUC	**0,73**	**0,73**	**0,74**

El modelo sin upsampling está muy probablemente sobreajustado lo que se observa fácilmente con la gran diferencia de rendimiento en los conjuntos de entrenamiento y validación. Por esto, es necesario utilizar un método de balanceo de clases como upsampling para que no se produzca este problema.

Por otro lado, el modelo Random Forest con upsampling utilizando los hiperparámetros por defecto no tiene un rendimiento sustancialmente mejor que el modelo sin upsampling, por lo que se hace necesario optimizarlo utilizando el método antes descrito.

El modelo optimizado permite obtener una métrica objetivo favorable, aunque con un valor bajo para la exhaustividad del modelo.

A continuación, se muestran los pesos de cada variable en el modelo final y su curva ROC asociada.

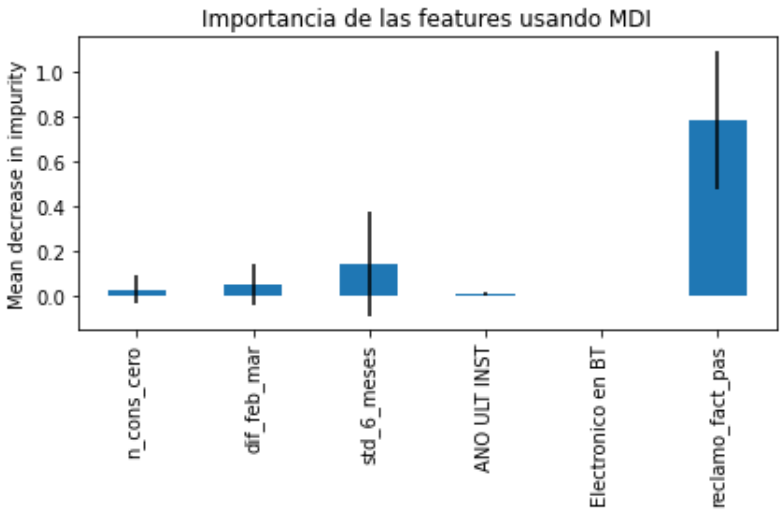

Ilustración 41: Pesos de los atributos en el modelo optimizado

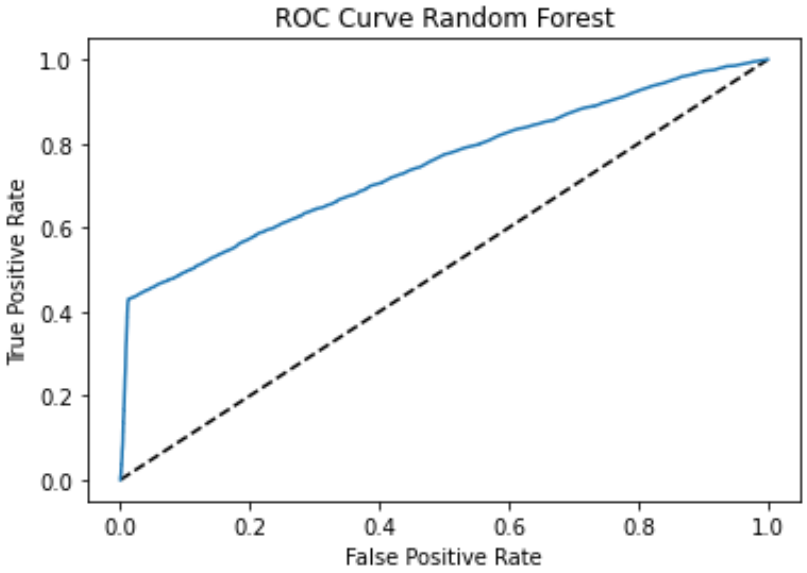

Ilustración 42: Curva ROC en el modelo optimizado

Como se puede apreciar, que un cliente haya reclamado por facturación a la empresa es un factor determinante para que el reclamo se vuelva a producir, pues es la variable que más influye.

5.5 CONCLUSIONES

Si bien el modelo cumple con poseer una buena métrica objetivo, se puede perfeccionar introduciendo nuevos atributos, que podrían ir en dirección del comportamiento de pago de los clientes y la ubicación espacial en la que se encuentran. Desafortunadamente no se pudo acceder de forma íntegra a estos datos desde el repositorio centralizado de la empresa.

Aun así, el modelo podría ser utilizado para utilizar acciones focalizadas debido a su alta precisión, pero, dada la vigencia de la ley 21.340, algunos clientes pudieron acogerse a ella, por lo que su probabilidad de reclamo puede estar disminuida, por lo tanto, es de vital importancia que la empresa tenga identificados a estos clientes. Por esta misma situación, no se recomienda la implementación directa de este modelo, debido a que sus resultados pueden estar fuertemente sesgados debido a los

datos de entrenamiento que se utilizaron y su alta dependencia con los reclamos realizados anteriormente.

Como se da la situación de que los clientes repiten reclamos por facturación, se podrían crear soluciones alternativas a esta, como tableros de visualización que permitan identificar clientes con un mayor número de reclamos, como se estudió entregar en un principio. Aun así, la solución que se propone en este trabajo integra los patrones de consumo de los clientes, por lo que corresponde a una solución más flexible que la toma de decisiones en base a un solo factor.

CAPÍTULO 6: CONCLUSIONES GENERALES

Con lo expuesto en el segundo capítulo de este trabajo se formula una solución para cobrar mejor en caso de que no se pueda realizar lectura. Se demuestra que esta estimación es deficiente y se determina que, utilizando un modelo XGBoost Regressor, se rebajaría el MAPE asociado de un 39,04% a un 28,92%. Este resultado puede potencialmente influenciar políticas públicas que involucren cambios a nivel nacional.

Con lo trabajado en el tercer y cuarto capítulo se podría mejorar la lectura de las distintas viviendas críticas afectadas, logrando una precisión y exhaustividad adecuada. Una combinación de estas medidas es la métrica F1 para la cual se logró un valor de promedio de 61% para las viviendas con problemas de lectura debido a la situación COVID-19 y de un 60% para las que estaban sin lectura debido al cambio de licitación en tiempos críticos.

Con lo trabajado en el quinto capítulo, se construye una solución que permite predecir, potencialmente, con una precisión del 93%, los clientes que reclamarán por motivo de facturación, lo que permitiría, teóricamente, tomar las acciones preventivas para que dichos reclamos no se produzcan.

Con lo anterior señalado, se concluye que el resultado de este trabajo apoyaría a que la empresa realice una mejor lectura en tiempos críticos provocados por la pandemia COVID-19. Si esta lectura, por un motivo exógeno, no puede ser buena, que pueda cobrar mejor el consumo correspondiente, aunque no se pueda leer. Y si finalmente hay errores de facturación por cualquier otro motivo, pueda anticiparse a los reclamos de los clientes, generando una mejor imagen para la empresa y tranquilidad para la población.

Para las cuatro problemáticas abordadas se formularon soluciones, bajo distintos enfoques, potencialmente útiles para la empresa, dado que obtienen un rendimiento superior al baseline indicado en cada situación. Aun así, el poder predictivo podría ser mayor en casos de que hubiera mayor cantidad de información disponible, especialmente enfocada a la caracterización de los clientes, que se podría utilizar de forma transversal.

No obstante, dada la integración de distintas técnicas, se pudo mitigar este factor, dando distintas alternativas a la organización para implementar y apoyar la toma de decisiones al respecto. Dado el grado de madurez analítica presente, se propusieron herramientas de visualización que tuvieron buena aceptación tanto dentro del área como de la empresa, y se espera en el futuro una implementación permanente. Como resultado de este trabajo, de igual forma, se espera aportar a un cambio de cultura en la organización para la utilización de herramientas analíticas de forma activa.

Dado el contexto, se generan dificultades en la implementación, ya que la ley 21.340, puede alterar el comportamiento de pago y reclamo de los clientes, pudiendo alterar los resultados futuros de sobremanera, ya que la data de entrenamiento no consideró este factor.

CAPÍTULO 7: BIBLIOGRAFÍA

1. Memoria anual Enel Distribución 2019 [en línea] https://www.enel.cl/content/dam/enel-cl/inversionistas/enel-distribucion-chile/reportes/memorias/2019/Memoria-Enel-Dx-2019.pdf> [consulta: 18/07/2020]

2. Informe de sostenibilidad 2017 Grupo Enel [en línea] https://www.enel.com/es/nuestra-compania/historias/articles/2018/07/informe-sostenibilidad-2017-enel-modelo-open-power-seeding-energies [consulta: 18/07/2020]

3. Enel Chile descripción [en línea] https://www.enel.cl/es/conoce-enel/enel-chile.html [consulta: 18/07/2020]

4. Descripción del sector eléctrico en Chile [en línea] https://www.cge.cl/sector-electrico/descripcion-general-sector-electrico/ [consulta: 18/07/2020]

5. Reglamento de la ley general de servicios eléctricos [en línea] https://www.cne.cl/wp-content/uploads/2015/06/DOC23_-_reglamento_electrico.pdf> [consulta: 23/07/2020]

6. Enel Distribución suspende reparto de boletas y lectura de consumo ante expansión del coronavirus [en línea] https://www.latercera.com/pulso/noticia/enel-distribucion-suspende-reparto-de-boletas-y-lectura-de-consumo-ante-expansion-del-coronavirus/X3S3WAEEYNFQZPFGQPYKKN6GTM/ [consulta: 26/07/2020]

7. El hoyo negro que alumbraron los medidores inteligentes: las súper ganancias que la ley le asegura a las eléctricas [en línea] <https://ciperchile.cl/2019/03/29/el-hoyo-negro-que-alumbraron-los-medidores-inteligentes-las-super-ganancias-que-la-ley-le-asegura-a-las-electricas/> [consulta: 19/08/2020]

8. Müller A., Guido S., 2016, Introduction to Machine Learning With Python. O'Reilly Media.

9. Ley 21.340 Servicios Básicos [en línea] https://www.enel.cl/es/clientes/informacion-util/ley-servicios-basicos.html [Consulta 21/03/202]

10. Journal of data warehousing [en línea] <https://mineracaodedados.files.wordpress.com/2012/04/the-crisp-dm-model-the-new-blueprint-for-data-mining-shearer-colin.pdf [Consulta: 19/07/2020]

11. Principales empresas distribuidoras en Chile [en línea] https://www.electricas.cl/asociados/ [consulta: 10/11/2020]

12. REGIÓN METROPOLITANA DE SANTIAGO ÍNDICE DE PRIORIDAD SOCIAL DE COMUNAS 2019 Seremi de Desarrollo Social y Familia Metropolitana [en línea]

http://www.desarrollosocialyfamilia.gob.cl/storage/docs/INDICE._DE_PRIORIDAD_SOCIAL_2019.pdf

13. Indicadores comunales. Encuesta CASEN 2013 [en línea] https://www.gobiernosantiago.cl/wp-content/uploads/2014/12/INDICADORES-COMUNALES-CASEN-RMS-2013.xls [Consulta: 31/03/2021]

14. Tableau Desktop I: Aspectos Básicos [en línea] https://elearning.tableau.com/tableau-desktop-i-aspectos-basicos-de-tableau-102018 [Consulta: 15/10/2020]

15. Tableau Desktop II: Nivel Intermedio [en línea] https://elearning.tableau.com/desktop-ii-nivel-intermedio [Consulta: 21/10/2020]

16. Tableau Desktop III: Nivel Avanzado [en línea] https://elearning.tableau.com/tableau-desktop-iii-avanzado [Consulta: 21/11/2020]

17. Dashboard Design: Visual Best Practices [en línea] https://elearning.tableau.com/dashboard-design-visual-best-practices?_gl=1*uy582g*_ga*MTA2ODI5MTE2MC4xNjI0NjgwODcx*_ga_8YLN0SNXVS*MTYyNDY4MDg3MC4xLjAuMTYyNDY4MDg3MC4w [Consulta: 09/09/2020]

18. M. Ward, G. Grinstein, D. Keim, 2010, Interactive Data Visualization: Foundations, Techniques, and Applications, A.K. Peters.

19. EPSG:32719 [en línea] https://epsg.io/32719 [consulta: 31/03/2021]

20. F. Halper, D. Stodder. TDWI Analytics Maturity Model Guide. [en línea] https://tdwi.org/~/media/545E06D7CE184B19B269E929B0903D0C [consulta: 12/11/2020]

21. Breiman, L. (2001). Random Forests. Machine Learning, 45, 5-32 [en línea] https://doi.org/10.1023/A:1010933404324 [consulta: 21/07/2020]

22. Cánovas, F., Alonso, F., Gomariz, F. & Oñate, F. (2017). Modification of the random forest algorithm to avoid statistical dependence problems when classifying remote sensing imagery. Computers & Geosciences, 103, 1-11. [en línea] https://doi.org/10.1016/j._cageo.2017.02.012 [consulta: 21/07/2021]

23. Lizares, M. (2017). Universidad Nacional Mayor de San Marcos. http://cybertesis.unmsm.edu.pe/bitstream/handle/cybertesis/7122/Lizares_cm.pdf?sequence=1&isAllowed=y [consulta:21/07/2021]

24. Luckner, M., Topolski, B. & Mazurek, M. (2017). Application of XGBoost algorithm in fingerprinting localisation task. 16th IFIP TC8 International Conference, CISIM 2017 661-671. Bialystok, Poland: CISIM. [en línea] https://doi.org/10.1007/978-3-319-59105-6_57 [consulta: 21/07/2021]

25. Sandoval, L. L. (2017). Machine Learning algorythms for analysis and data prediction. 2017 IEEE 37th Central America and Panama Convention (CONCAPAN XXXVII), 1-5. Managua,

Nicaragua. IEEE. [en línea] http://doi.org/10.1109/CONCAPAN.2017.8278511 [consulta: 21/07/2021]

26. Nobre, J. & Ferreira, R. (2019). Combining principal component analysis, discrete wavelet transform and XGBoost to trade in the financial markets. Expert Systems with Applications, 125, 181-194. [en línea] https://doi.org/10.1016/j.eswa.2019.01.083 [consulta: 21/07/2021]

27. Tableau: Informing Without Misleading. [en línea] https://elearning.tableau.com/analytics-best-practices/395548/scorm/35ma3yps5iw0n [consulta: 21/07/2021]

28. Dörr, A. et al., 2008, Psicología General y Evolutiva. Ediciones Mediterraneo.

29. Yalcinkaya, M. & Singh, V (2018). VisualCOBie for Facilities Management: A BIM integrated, Visual Search and Information Management Platform for COBie Extension [en línea] https://www.researchgate.net/publication/325988233_VisualCOBie_for_Facilities_Management_A_BIM_integrated_Visual_Search_and_Information_Management_Platform_for_COBie_Extension [consulta: 21/07/2021]

30. Buenadicha C., Goldón G., et al. (2019). La Gestión Ética de los Datos [en línea] https://publications.iadb.org/es/la-gestion-etica-de-los-datos [consulta: 27/07/2021]

CAPÍTULO 8: ANEXOS

ANEXO A: COMUNAS CON CONCESIÓN EN ENEL

La empresa presta servicios en 33 comunas de la Región Metropolitana: Cerrillos, Cerro Navia, Colina, Conchalí, Estación Central, Huechuraba, Independencia, Lampa, La Cisterna, La Florida, La Granja, La Reina, Las Condes, Lo Barnechea, Lo Espejo, Lo Prado, Macul, Maipú, Ñuñoa, Pedro Aguirre Cerda, Peñalolén, Providencia, Pudahuel, Quilicura, Quinta Normal, Recoleta, Renca, San Joaquín, San Miguel, San Ramón, Santiago, Til Til y Vitacura.

Las áreas de concesión específicas de Enel Chile se pueden visualizar en la siguiente figura:

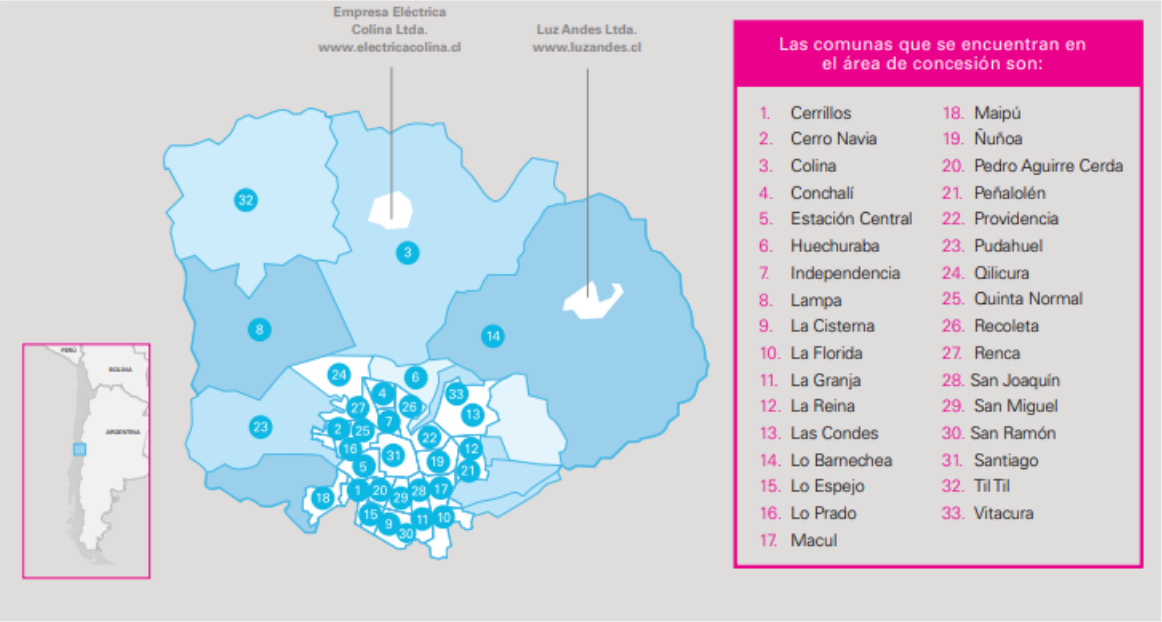

Ilustración 43 Áreas de concesión de la empresa. Fuente: reporte de sostenibilidad 2019

ANEXO B: ARTÍCULOS CORRESPONDIENTES A LA REGULACIÓN DE DISTIBUCIÓN DE ENERGÍA ELÉCTRICA

- Artículo 123: Los concesionarios de servicio público de distribución deberán facturar en base a las cantidades que consten en el equipo que registra los consumos del usuario, exceptuando los casos en que este reglamento autoriza la estimación del consumo

- Artículo 126: La facturación de los consumos, en caso de suministros sometidos a fijación de precios, deberá hacerse en forma mensual o bimestral. Se entenderá por período mensual de facturación de consumos, aquél que no sea inferior a 27 ni superior a 33 días, y por período bimestral, aquél que no sea inferior a 57 ni superior a 63 días.
- Artículo 129: Los usuarios deberán dar facilidades para que los concesionarios puedan tomar lectura de medidores cualquier día del mes, en el horario comprendido entre las 08:00 y 18:00 horas. En casos calificados, la Superintendencia podrá autorizar otros horarios respecto de clientes determinados. Si por cualquier causa no imputable al concesionario no pudiere efectuarse la lectura correspondiente, el concesionario dejará una constancia de esta situación en un lugar visible del inmueble y podrá facturar provisoriamente, hasta por dos períodos consecutivos, una cantidad equivalente al promedio facturado en los seis meses 36 anteriores. En la boleta o factura siguiente que se emita de acuerdo con las lecturas del medidor, se abonarán los pagos referidos, dejándose constancia de esta circunstancia. Para estos efectos, la demanda máxima registrada al momento en que pueda tomarse la lectura se considerará también para el período anterior. Con todo, si se emitieran respecto de un mismo usuario más de cuatro facturaciones estimadas en un período de doce meses, se deberán anotar en un registro que deberá estar disponible para revisión de la Superintendencia.

ANEXO C: LEY 21.249

LEY NÚM. 21.249

DISPONE, DE MANERA EXCEPCIONAL, LAS MEDIDAS QUE INDICA EN FAVOR DE LOS USUARIOS FINALES DE SERVICIOS SANITARIOS, ELECTRICIDAD Y GAS DE RED

Teniendo presente que el H. Congreso Nacional ha dado su aprobación al siguiente proyecto de ley, originado en mociones refundidas:

- De los diputados señoras Joanna Pérez Olea, Sofía Cid Versalovic y Andrea Parra Sauterel, y señores Pablo Kast Sommerhoff, Gonzalo Fuenzalida Figueroa, Andrés Longton Herrera, Pablo Lorenzini Basso, José Miguel Ortiz Novoa, Jorge Sabag Villalobos y Víctor Torres Jeldes, boletín N° 13.329-03.

- De los diputados señores Marcelo Schilling Rodríguez, Luis Rocafull López, Juan Luis Castro González, Jaime Naranjo Ortiz, señoras Jenny Álvarez Vera y Emilia Nuyado Ancapichún, señores Leonardo Soto Ferrada, Marcos Ilabaca Cerda y Gastón Saavedra Chandía, boletín N° 13.342-03.

- De los diputados señores Alexis Sepúlveda Soto, Boris Barrera Moreno, Manuel Monsalve Benavides, Mario Desbordes Jiménez, Francisco Eguiguren Correa, Alejandro Bernales Maldonado, Alejandro Santana Tirachini, Hugo Rey Martínez y Sebastián Torrealba Alvarado y señora Marcela Hernando Pérez, boletín N° 13.347-03.

- De los diputados señores Boris Barrera Moreno, Renato Garín González, Hugo Gutiérrez Gálvez, Daniel Núñez Arancibia, Alexis Sepúlveda Soto, Raúl Soto Mardones y señora Alejandra Sepúlveda Orbenes, boletín N° 13.354-03.

- De los diputados señores Resolución 1, ENERGÍA D.O. 26.08.2020 Andrés Longton Herrera, Francisco Eguiguren Correa, Frank Sauerbaum Muñoz, Pablo Prieto Lorca, Gonzalo Fuenzalida Figueroa, Hugo Rey Martínez, Alejandro Santana Tirachini, Cristóbal Urruticoechea Ríos y señoras Erika Olivera de la Fuente y Ximena Ossandón Irarrázabal, boletín N° 13.355-03.

- De los diputados señores Hugo Rey Martínez, Sebastián Keitel Bianchi, Andrés Celis Montt, Frank Sauerbaum Muñoz y Alexis Sepúlveda Soto, boletín N° 13.356-03.

- De los senadores señor Alejandro Navarro Brain, señora Ximena Rincón González y señores Carlos Bianchi Chelech, Guido Girardi Lavín, y Rabindranath Quinteros Lara, boletín N° 13.315-08.

- De los senadores señora Yasna Provoste Campillay y señores Carlos Bianchi Chelech, Guido Girardi Lavín, Alejandro Guillier Álvarez y Alejandro Navarro Brain, boletín N° 13.417-03.

- De los senadores señores Álvaro Elizalde Soto y Rabindranath Quinteros Lara, boletín N° 13.438-03.

"Proyecto de ley:

Artículo 1.- Ley 21340

Art. único N° 1

D.O. 22.05.2021 Hasta el 31 de diciembre de 2021, las empresas proveedoras de servicios sanitarios, empresas y cooperativas de distribución de electricidad y las empresas de distribución de gas de red no podrán cortar el suministro por mora en el pago a las personas, usuarios y establecimientos, en adelante usuarios, clientes o beneficiarios, que a continuación se indican:

a) Usuarios residenciales o domiciliarios.
b) Hospitales y centros de salud.
c) Cárceles y recintos penitenciarios.
d) Hogares de menores en riesgo social, abandono o compromiso delictual.
e) Hogares y establecimientos de larga estadía de adultos mayores.
f) Bomberos.
g) Organizaciones sin fines de lucro.
h) Microempresas, de acuerdo a lo establecido en la ley N° 20.416, que fija normas especiales para las empresas de menor tamaño.

Se suspende para los clientes señalados en el inciso anterior, por el plazo a que se refiere este artículo, la aplicación de los incisos tercero, cuarto, quinto y sexto del artículo 36 del decreto con fuerza de ley N° 323, de 1931, del Ministerio del Interior, Ley de Servicios de Gas; del artículo 141 y del inciso segundo del literal q) del artículo 225 del decreto con fuerza de ley N° 4, de 2007, del Ministerio de Economía, Fomento y Reconstrucción, que fija el texto refundido, coordinado y sistematizado del decreto con fuerza de ley N° 1, de Minería, de 1982, Ley General de Servicios Eléctricos, en materia de energía eléctrica, y lo establecido en la letra d) del artículo 36 del decreto con fuerza de ley N° 382, de 1988, del Ministerio de Obras Públicas, Ley General de Servicios Sanitarios.

Se excluye de la aplicación de la presente ley a las empresas sanitarias con menos de 12.000 clientes que constituyan una sola unidad económica y no sean filial de otra empresa sanitaria, y a las cooperativas y comités de agua potable rural, sin perjuicio de los convenios, descuentos o facilidades de pago que otorguen a sus clientes.

Artículo 2.- Las deudas contraídas con las empresas de servicios sanitarios, empresas y cooperativas de distribución de electricidad y empresas de gas de red, que se generen entre el 18 de marzo de 2020 y hasta Ley 21340 Art. único N° 2 y N° 3 D.O. 22.05.2021 el 31 de diciembre de 2021, se prorratearán en el número de cuotas mensuales iguales y sucesivas que determine el usuario final a su elección, las que no podrán exceder de cuarenta y ocho, a partir de la facturación siguiente al término de este último plazo, y no podrán incorporar multas, intereses ni gastos asociados.

Adicionalmente, a elección del usuario final, el prorrateo podrá incluir deudas generadas antes de las contraídas según lo señalado en el inciso anterior, hasta el monto de diez unidades de fomento para las empresas distribuidoras y cooperativas de electricidad y hasta el monto de cinco unidades de fomento para las empresas de servicios sanitarios y de distribución de gas de red, en las mismas condiciones.

Artículo 3.- Solo podrán acogerse a lo dispuesto en el artículo 2 los clientes finales que cumplan con, al menos, uno de los siguientes requisitos:

a) Encontrarse dentro del Ley 21340 Art. único N° 4 D.O. 22.05.2021 80 por ciento de vulnerabilidad, de conformidad al Registro Social de Hogares.

b) Tener la calidad de adulto mayor, de acuerdo a la ley N° 19.828, que crea el Servicio Nacional del Adulto Mayor.

c) Estar percibiendo las prestaciones de la ley N° 19.728, que establece un seguro de desempleo.

d) Estar acogido a alguna de las causales de la ley N° 21.227, que faculta el acceso a prestaciones del seguro de desempleo de la ley N° 19.728, en circunstancias excepcionales, ya sea por la suspensión de la relación laboral o por la celebración de un pacto de reducción temporal de jornada.

e) Ser trabajador independiente o informal no comprendido en alguna de las categorías anteriores, y expresar, mediante

declaración jurada simple, que está siendo afectado por una disminución significativa de ingresos que justifica el acceso a los beneficios. La utilización maliciosa de la declaración se sancionará de conformidad al artículo 210 del Código Penal.

Los requisitos señalados en el inciso anterior no serán exigibles a los beneficiarios indicados en los literales b), c), d), e), f), g) y h) del artículo 1.

Artículo 4.- Los usuarios finales no comprendidos en el artículo anterior que acrediten estar imposibilitados de dar cumplimiento a las obligaciones de pago que han contraído con la respectiva empresa o cooperativa prestadora, y así lo expresen mediante declaración jurada simple, podrán solicitar acogerse a la postergación y prorrateo de los pagos, tratándose de las empresas y cooperativas indicadas en el artículo 1. La utilización maliciosa de la declaración se sancionará de conformidad con lo dispuesto en el artículo 210 del Código Penal.

La negativa de la empresa prestadora podrá ser objeto de reclamo ante la subsecretaría, superintendencia u organismo fiscalizador respectivo, y se sujetará a la normativa sectorial que corresponda.

Artículo 5.- Dentro de los cinco días siguientes a la publicación de esta ley, las empresas y cooperativas proveedoras de los servicios señalados en ella, deberán establecer plataformas de atención al cliente, por internet y telefonía, que permitan formular las solicitudes para acceder a los beneficios que la presente ley establece.

En cualquiera de los casos previstos en esta ley, las empresas y cooperativas proveedoras deberán resolver las solicitudes efectuadas por los interesados, dentro de los cinco días hábiles siguientes a su formulación. Respecto de los usuarios que reúnan cualquiera de las condiciones indicadas en el inciso primero del artículo 3, no procederá rechazo alguno, y el beneficio será aplicado, de pleno derecho, por parte de las empresas proveedoras y cooperativas.

La respuesta de la correspondiente empresa o cooperativa deberá ser comunicada al solicitante por medio de correo electrónico o mensaje de texto, dentro del señalado plazo. En caso de que la respuesta fuere negativa, la empresa o cooperativa deberá mencionar y justificar las razones del rechazo.

Del mismo modo, las empresas deberán informar sus resoluciones a la subsecretaría, superintendencia u organismo fiscalizador respectivo y, quincenalmente, deberán publicar en su página web el número y porcentaje de solicitudes aceptadas y rechazadas, conforme a lo establecido en el inciso anterior.

Las empresas y cooperativas deberán elaborar un registro y estadísticas de los solicitantes beneficiarios, en un plazo no mayor a diez días hábiles desde la publicación de esta ley, y deberán actualizarlo quincenalmente.

Las denuncias de infracciones de esta ley deberán ser tratadas, por parte de las subsecretarías, superintendencias u organismos fiscalizadores respectivos, como reclamos, de acuerdo a la normativa vigente.

Artículo 6.- Las infracciones de lo dispuesto en la presente ley serán sancionadas de conformidad a la normativa sectorial respectiva.

Artículo 7.- Ley 21340 Art. único N° 5 a) D.O. 22.05.2021 Hasta el 31 de diciembre de 2021, las empresas generadoras y transmisoras de energía eléctrica, deberán continuar proveyendo con normalidad sus servicios a las empresas distribuidoras domiciliarias de energía y a las cooperativas eléctricas.

Dentro del plazo comprendido entre los treinta días previos a la publicación de esta ley y Ley 21340 Art. único N° 5 b) D.O. 22.05.2021 el 31 de diciembre de 2021, de manera excepcional, el pago de las cooperativas eléctricas a las empresas generadoras y transmisoras podrá ser realizado en cuotas, en el mismo número de meses en que se prorratearán las cuentas de sus beneficiarios, sin multas, intereses ni gastos asociados.

Artículo 8.- Si los beneficiarios de esta ley hubiesen sido objeto de cortes o suspensiones de suministro o servicio, por mora en el pago de cualquiera de los servicios señalados en el artículo 1, la respectiva empresa proveedora o cooperativa deberá proceder a la reposición inmediata del servicio, sin costo alguno para el usuario, una vez publicada la presente ley.".

Artículo 9.- Sin perjuicio del plazo establecido en el artículo 2, los beneficiarios señalados en los artículos 3 y 4 tendrán un plazo de 30 días adicionales para el solo efecto de acogerse a lo dispuesto en el artículo 2.

Quince días antes del vencimiento del plazo establecido en el inciso primero del artículo 1, las empresas deberán remitir a los clientes finales la información correspondiente al monto de su deuda y a los beneficios a los que se pueden acoger de conformidad a esta ley.

Artículo 10.- Las empresas de servicios sanitarios, empresas y cooperativas de distribución de electricidad y empresas de gas de red deberán informar en sus sitios web y en las cuentas, ya sean físicas o virtuales, la deuda que mantiene el usuario por la aplicación de esta normativa, de haberla, y la forma cómo podría prorratearse, de 1 a 48 cuotas.

Artículo 11. Cumplido el plazo indicado en los artículos 1, 2 y 7, si aún se encontrare vigente la declaración de Estado de Excepción Constitucional de Catástrofe por pandemia de Covid 19,

declarado en el decreto 104, de 18 de marzo de 2020, del Ministerio del Interior y Seguridad Pública, y sus prórrogas, dichos plazos se extenderán hasta 60 días desde terminado dicho estado de excepción constitucional.

Y por cuanto he tenido a bien aprobarlo y sancionarlo; por tanto, promúlguese y llévese a efecto como Ley de la República.

Santiago, 5 de agosto de 2020.- SEBASTIÁN PIÑERA ECHENIQUE, Presidente de la República.- Juan Carlos Jobet Eluchans, Ministro de Energía.- Alfredo Moreno Charme, Ministro de Obras Públicas.

Lo que transcribo a Ud. para su conocimiento.- Saluda Atte. a Ud., Francisco López Díaz, Subsecretario de Energía.

ANEXO D: RECLAMOS POR MOTIVO DE ENEL DISTRIBUCIÓN CHILE S.A.

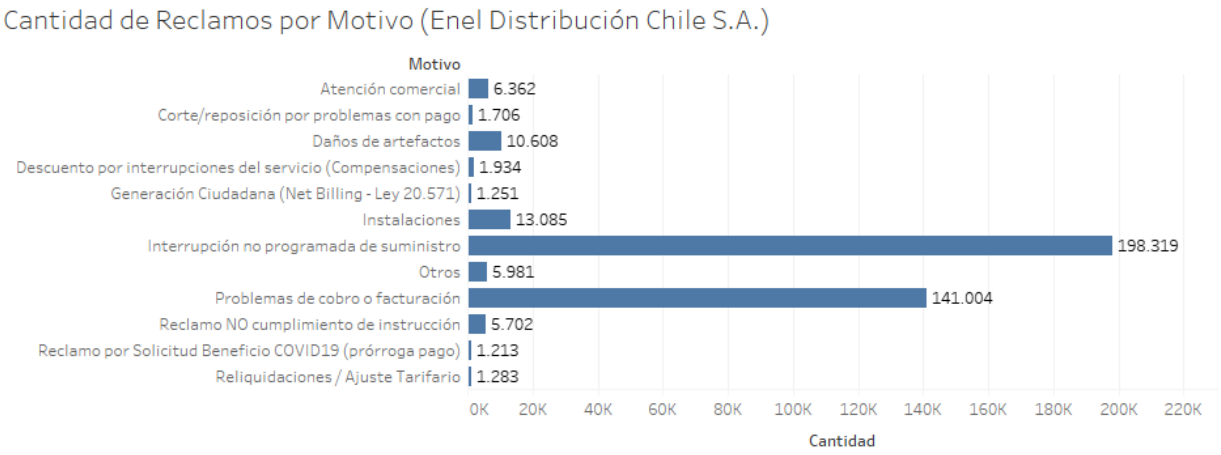

Ilustración 44 Cantidad de Reclamos por Motivo en Enel Distribución Chile S.A.

ANEXO E: EVOLUCIÓN DE LOS RECLAMOS POR MOTIVO DE ENEL DISTRIBUCIÓN CHILE S.A.

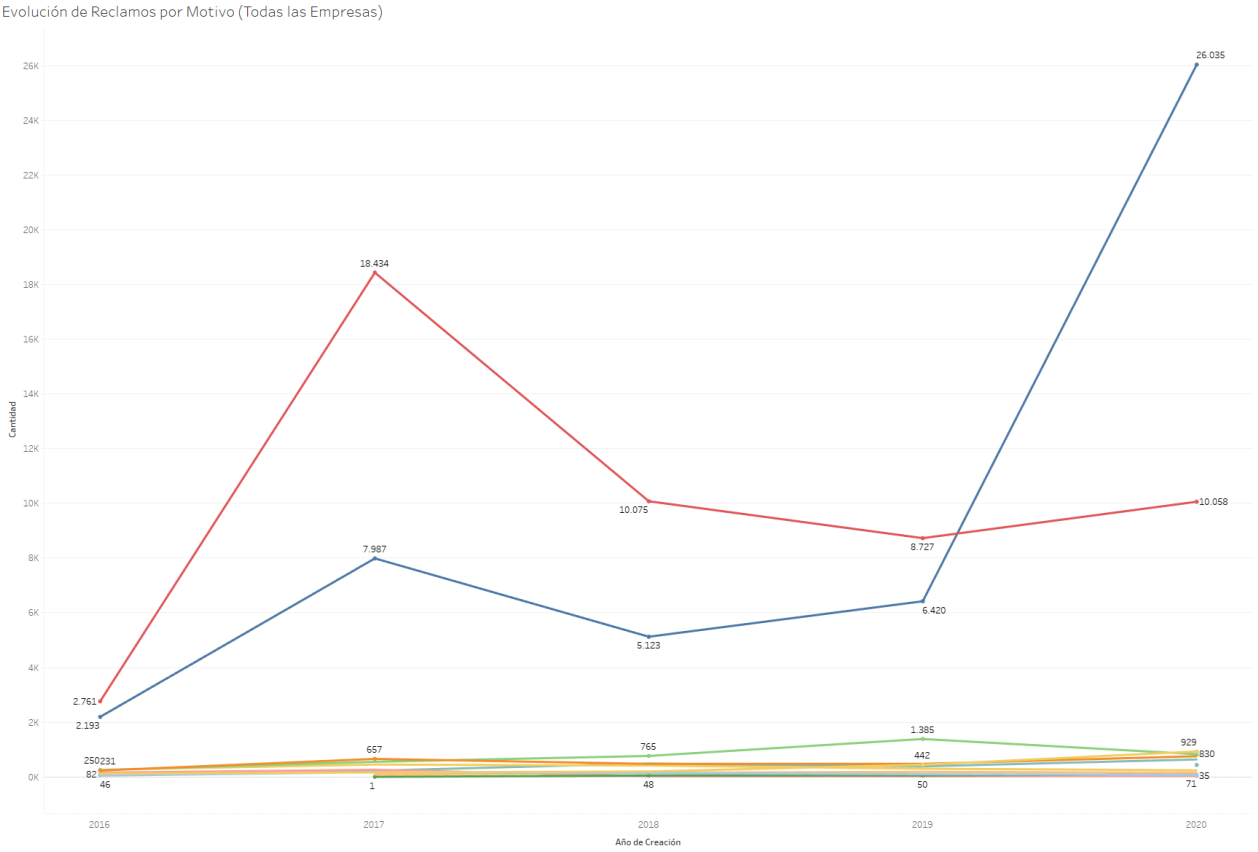

Ilustración 45: Evolución de los reclamos en Enel Distribución Chile. La linea roja representa reclamos por corte de suministro y la azul por facturación

ANEXO F: DISTRIBUCIÓN DE CONSUMOS PROMEDIO EN HORIZONTE DE TRES AÑOS SEGMENTO IPS ALTO O MEDIO ALTO

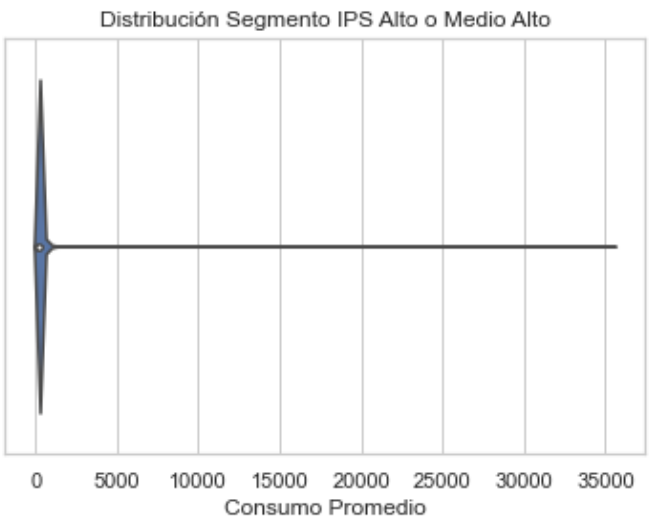

Ilustración 46: Distribución de los datos en segmento IPS alto o medio alto. Fuente: Elaboración propia

ANEXO G: DISTRIBUCIÓN DE CONSUMOS PROMEDIO EN HORIZONTE DE TRES AÑOS SEGMENTO IPS MEDIO BAJO O BAJO

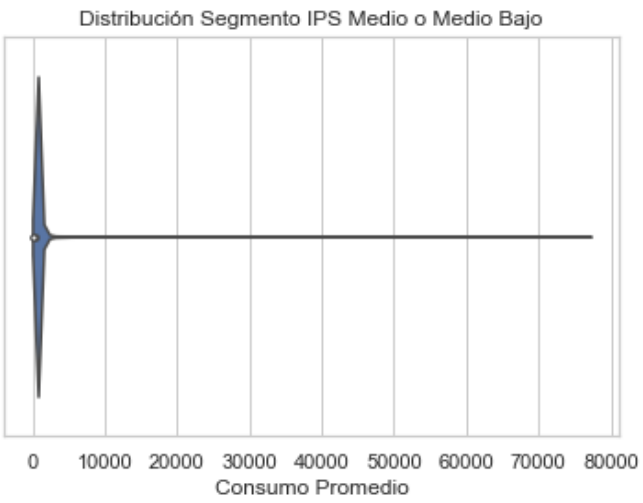

Ilustración 47: Distribución de los datos en segmento IPS medio bajo o bajo. Fuente: Elaboración propia

ANEXO H: DISTRIBUCIÓN DE CONSUMOS PROMEDIO EN HORIZONTE DE TRES AÑOS SEGMENTO DE COMUNAS SIN PRIORIDAD

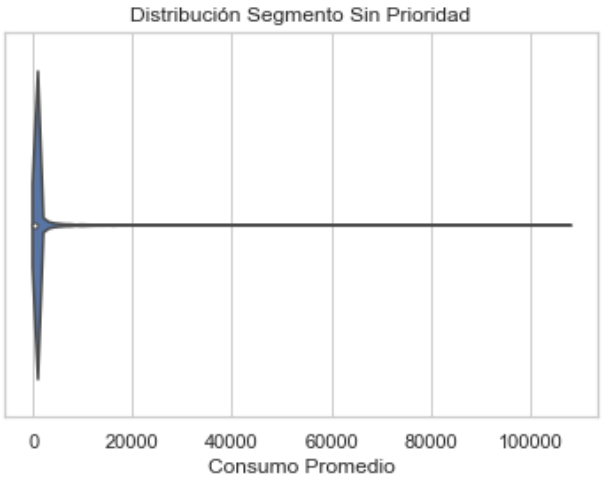

Ilustración 48: Distribución de los datos en segmento de comunas sin prioridad. Fuente: Elaboración propia

ANEXO I: DESVIACIONES ESTÁNDAR SEGMENTO ALTO O MEDIO ALTO

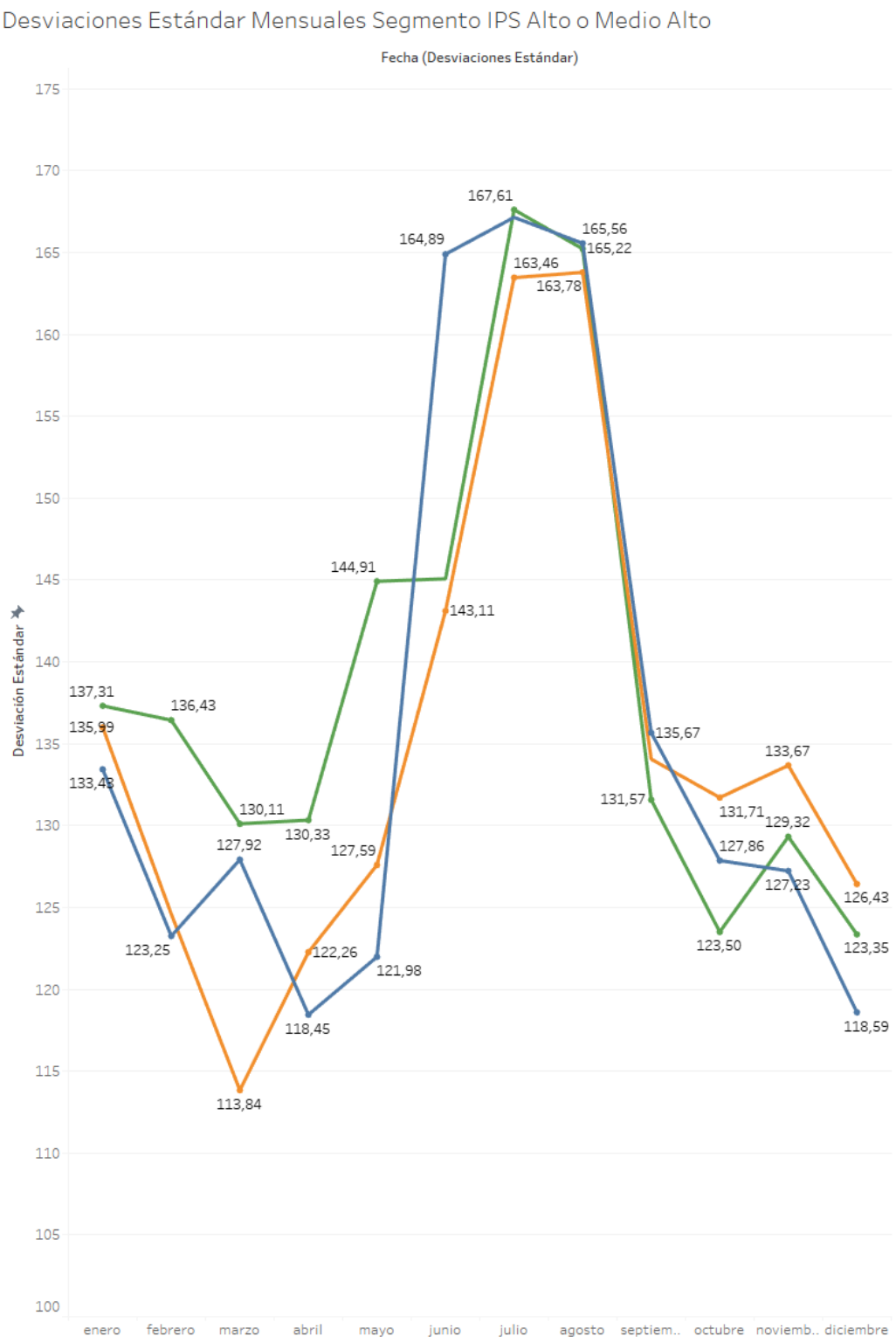

Ilustración 49: Desviaciones en el consumo del segmento IPS Alto o Medio Alto. Las líneas representan el consumo promedio del segmento y los círculos el consumo promedio de todos los segmentos. El color azul representa el año 2017, el naranjo el año 2018 y el verde el año 2019. Fuente: Elaboración propia.

ANEXO J: DESVIACIONES ESTÁNDAR SEGMENTO IPS MEDIO BAJO O BAJO

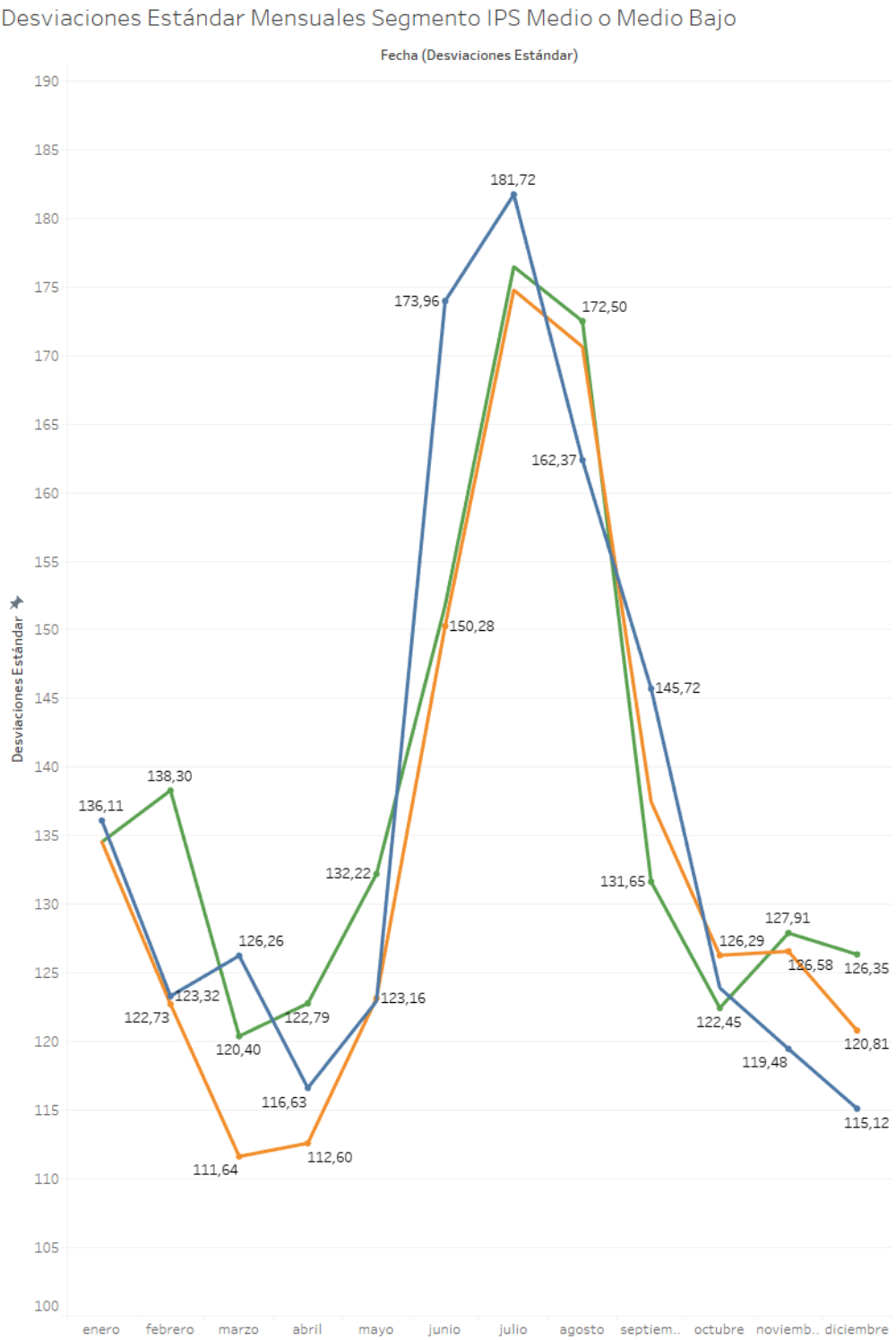

Ilustración 50: Desviaciones en el consumo del segmento IPS Medio o Medio Bajo. Las líneas representan el consumo promedio del segmento y los círculos el consumo promedio de todos los segmentos. El color azul representa el año 2017, el naranjo el año 2018 y el verde el año 2019. Fuente: Elaboración propia.

ANEXO K: DESVIACIONES ESTÁNDAR SEGMENTO DE COMUNAS SIN PRIORIDAD SOCIAL

Ilustración 51: Desviaciones en el consumo del segmento IPS Alto o Medio Alto. Las líneas representan el consumo promedio del segmento y los círculos el consumo promedio de todos los segmentos. El color azul representa el año 2017, el naranjo el año 2018 y el verde el año 2019. Fuente: Elaboración propia.

ANEXO L: EJECUCIÓN DE MODELOS OPTICS Y CURVAS ROC PARA LOS MODELOS DE LECTURA SEGÚN DISTRITO

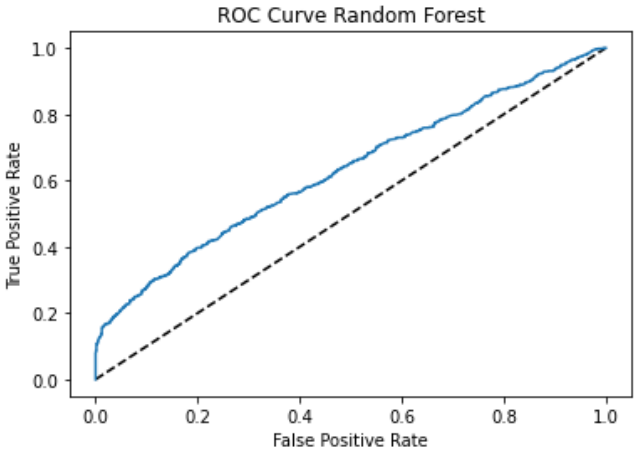

Ilustración 52: Curva ROC para el distrito centro

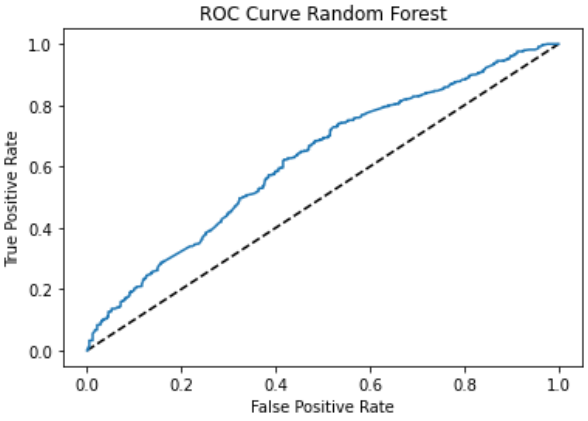

Ilustración 53: Curva ROC para el distrito norte

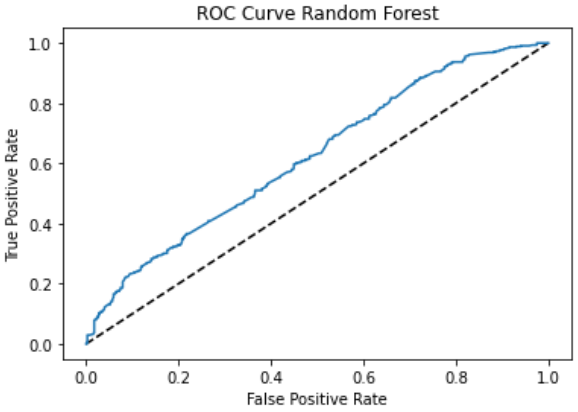

Ilustración 54: Curva ROC para el distrito sur

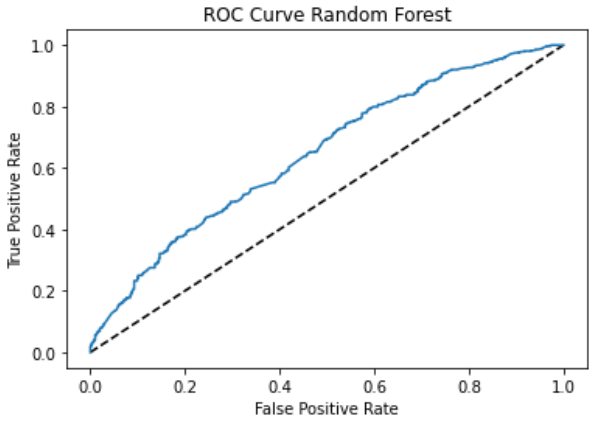

Ilustración 55: Curva ROC para el distrito oriente

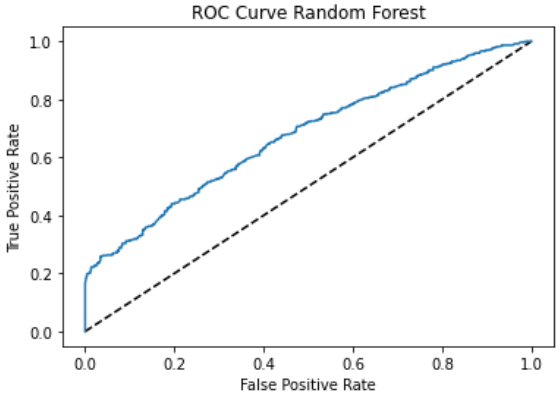

Ilustración 56: Curva ROC para el distrito poniente

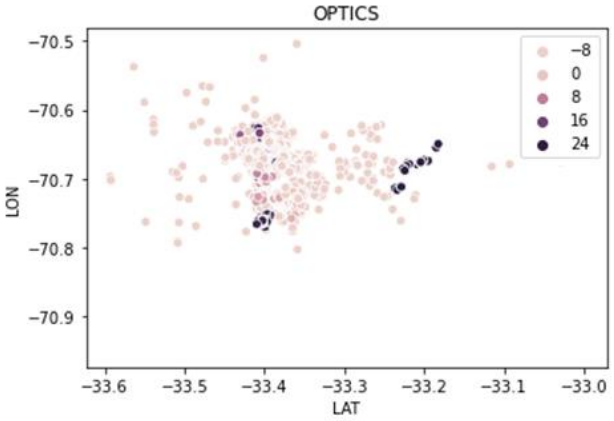

Ilustración 57: Aplicación de OPTICS para el distrito norte

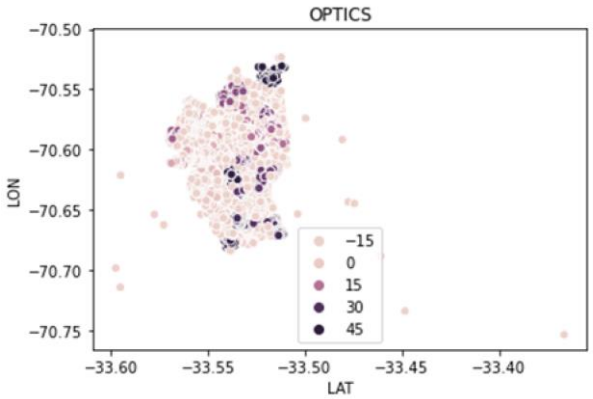

Ilustración 58: Aplicación de OPTICS para el distrito sur

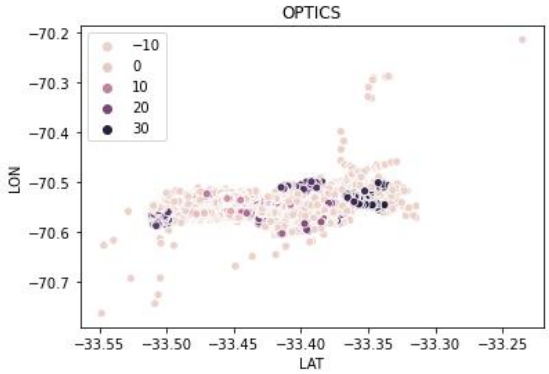

Ilustración 59: Aplicación de OPTICS para el distrito oriente

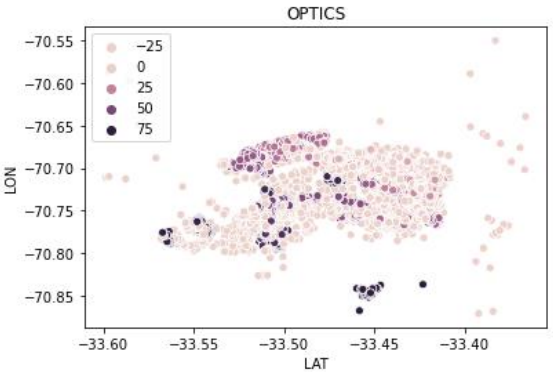

Ilustración 60: Aplicación de OPTICS para el distrito poniente

ANEXO M: EJECUCIÓN DE MODELOS OPTICS Y CURVAS ROC PARA LOS MODELOS DE LECTURA SEGÚN PRIORIDAD SOCIAL

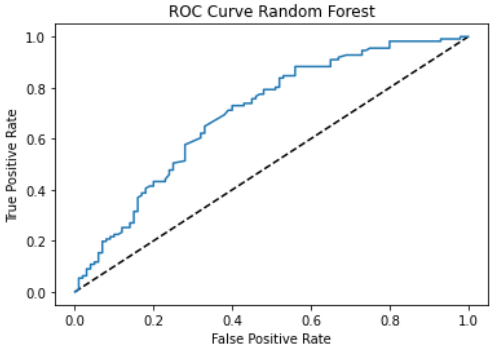

Ilustración 61: Curva ROC para el segmento IPS alto

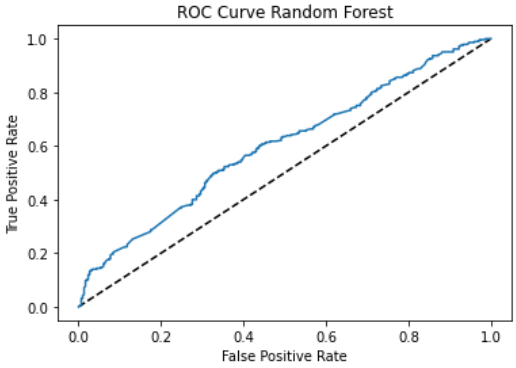

Ilustración 62: Curva ROC para el segmento IPS medio alto

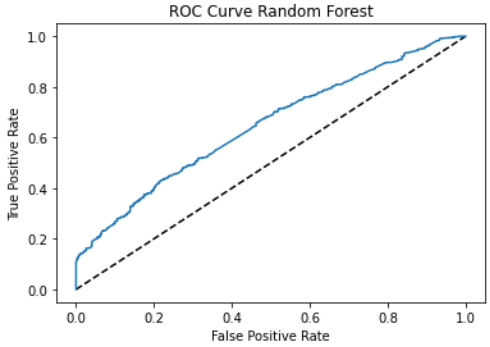

Ilustración 63: Curva ROC para el segmento IPS medio bajo

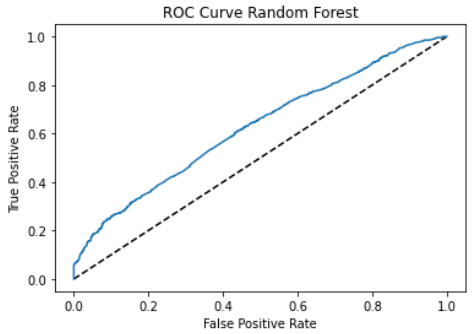

Ilustración 64: Curva ROC para el segmento IPS bajo

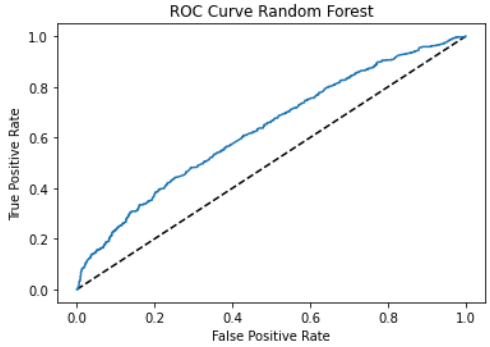

Ilustración 65: Curva ROC para el segmento IPS sin prioridad

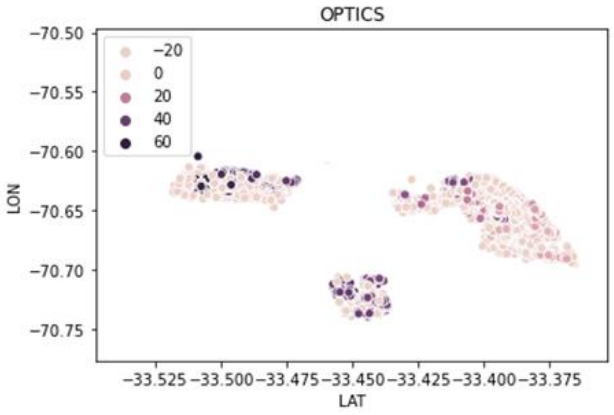

Ilustración 66: Aplicación de OPTICS para segmento IPS medio alto

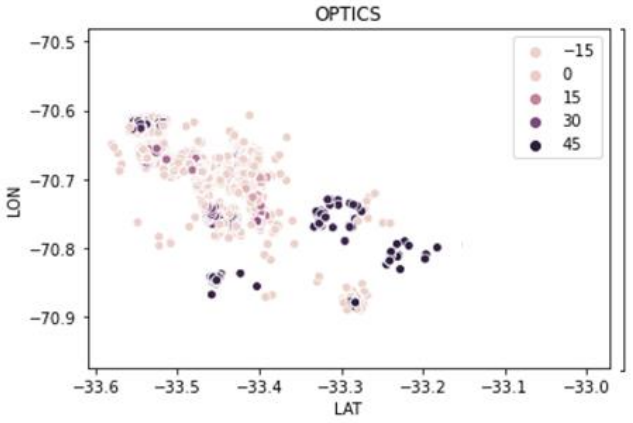

Ilustración 67: Aplicación de OPTICS para segmento IPS bajo

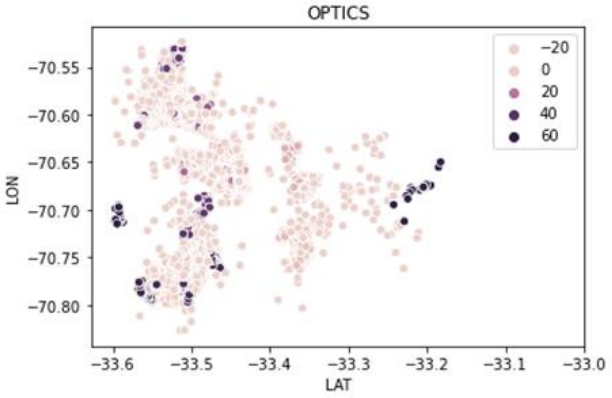

Ilustración 68: Aplicación de OPTICS para segmento IPS bajo

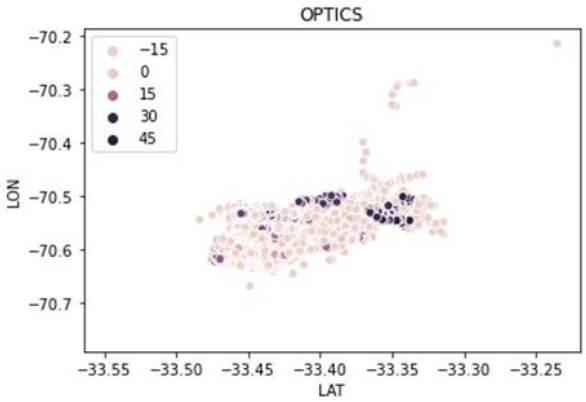

Ilustración 69: Aplicación de OPTICS para segmento de comunas sin prioridad

ANEXO N: CURVAS ROC PARA MODELO DE LECTURA CAMBIO DE LICITACIÓN

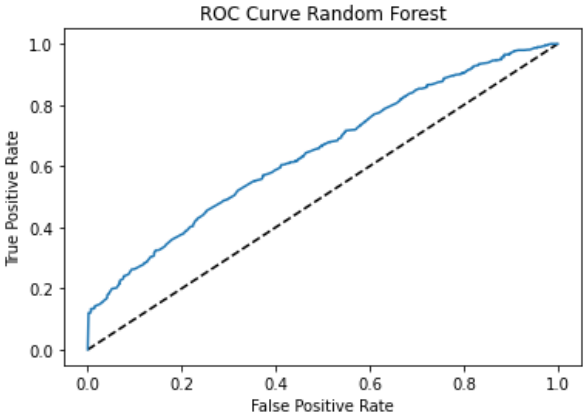

Ilustración 70: Curva ROC en segmento IPS medio bajo

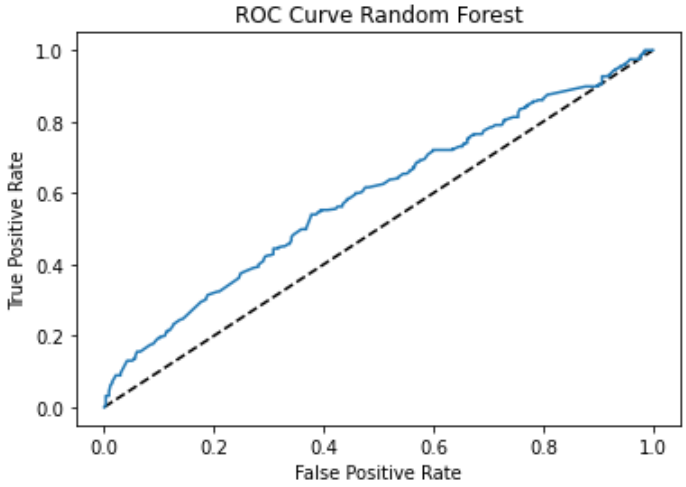

Ilustración 71: Curva ROC en segmento IPS medio alto

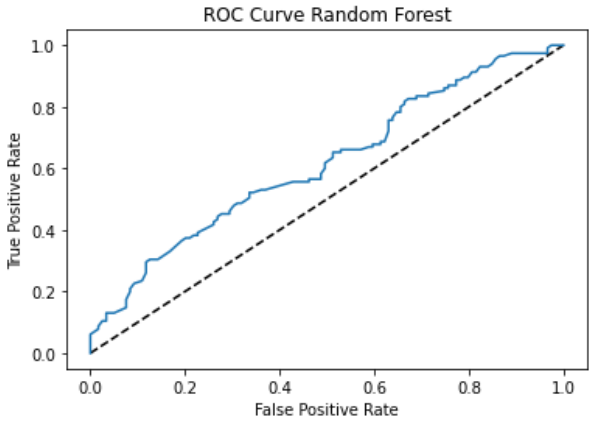

Ilustración 72: Curva ROC en segmento IPS alto

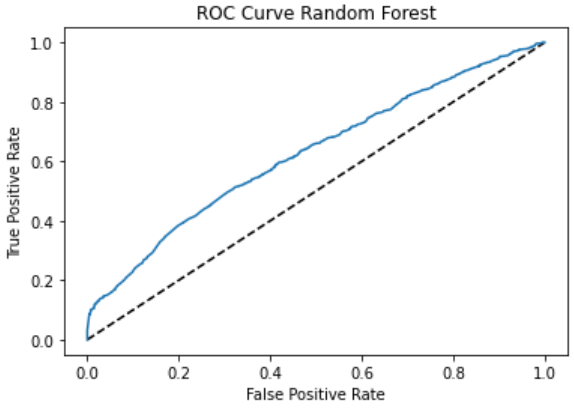

Ilustración 73: Curva ROC en segmento IPS bajo

ANEXO Ñ: FEATURE IMPORTANCES PARA MODELO DE LECTURA CAMBIO DE LICITACIÓN

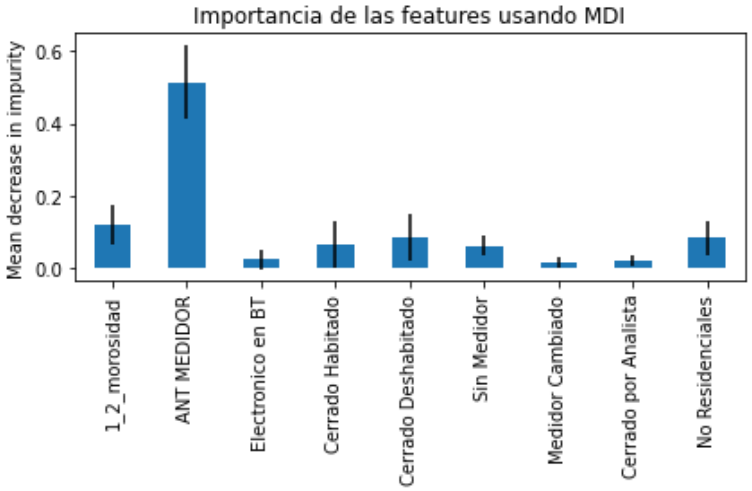

Ilustración 74: Importancia de las features en segmento IPS medio bajo

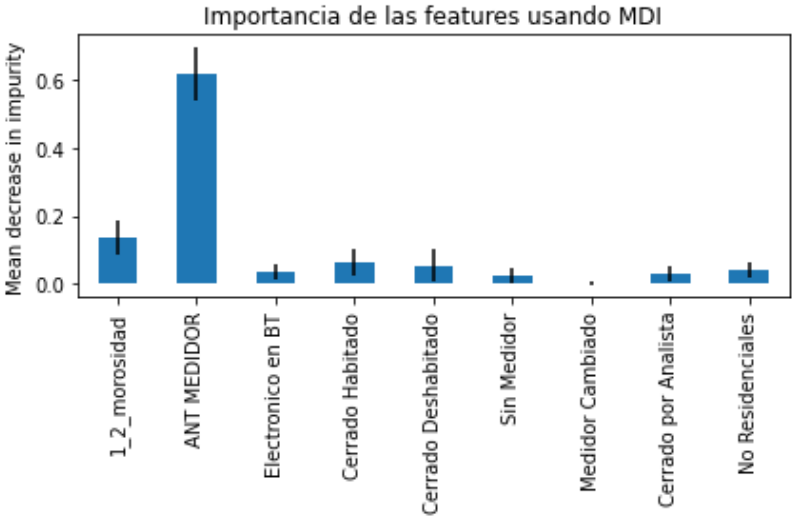

Ilustración 75: Importancia de las features en segmento IPS medio alto

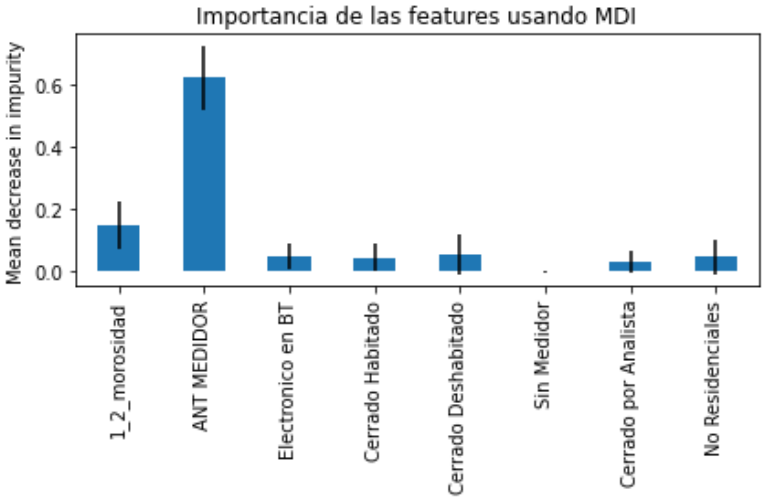

Ilustración 76: Importancia de las features en segmento IPS alto

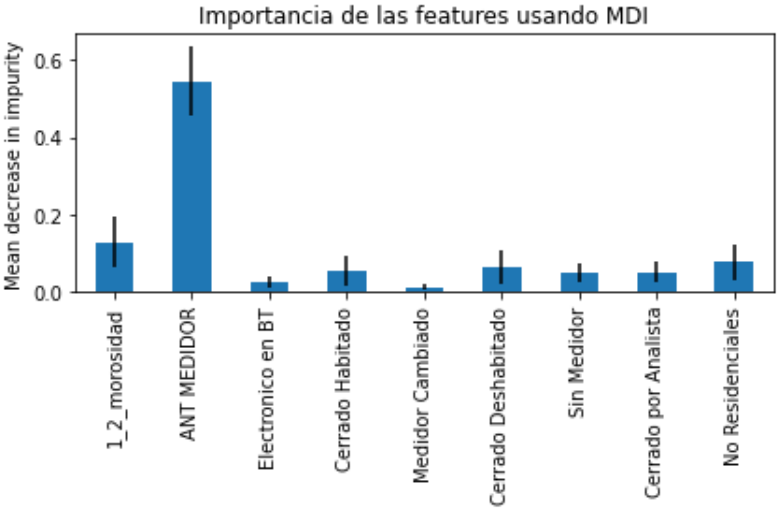

Ilustración 77: Importancia de las features en segmento IPS bajo