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

“DESARROLLO DE UN MODELO PREDICTIVO PARA MEJORAR LA ESTIMACIÓN DE ENERGÍA POR LEER EN LOS MEDIDORES (ELM) CON EL USO DE DATOS GENERADOS POR MEDIDORES DEL SEGMENTO MASIVO”

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

GUILLERMO RENÉ MORALES YÉVENES

PROFESORA GUÍA:
CAROLINA SEGOVIA RIQUELME

MIEMBROS DE LA COMISIÓN:
ALEJANDRA PUENTE CHANDÍA
PEDRO URZÚA SALINAS

SANTIAGO DE CHILE
2021

RESUMEN DE LA MEMORIA PARA OPTAR AL TÍTULO DE: Ingeniero Civil Industrial
POR: Guillermo René Morales Yévenes
FECHA: 30/06/2021
PROF. GUÍA: Carolina Segovia

DESARROLLO DE UN MODELO PREDICTIVO PARA MEJORAR LA ESTIMACIÓN DE ENERGÍA POR LEER EN LOS MEDIDORES (ELM) CON EL USO DE DATOS GENERADOS POR MEDIDORES DEL SEGMENTO MASIVO

Una de las grandes problemáticas en el sector de la distribución de energía eléctrica es la dificultad para conocer el detalle del comportamiento del consumo energético de sus clientes en el tiempo. Para ello, se han incorporado nuevas tecnologías, como los medidores inteligentes, que permiten medir y monitorear de forma remota y constante las fluctuaciones del consumo. Con esto, una de las decisiones más importantes de parte de las distribuidoras eléctricas es saber cuánta energía comprar mes a mes a las generadoras para distribuirla a sus clientes durante el transcurso de éstos, y no quedarse sin abastecimiento con esta compra, lo que se traduciría en pérdidas monetarias para la empresa causadas por subestimación de compra.

Dado que esta orden de compra se realiza los días previos a los últimos días del mes y que no existe la capacidad productiva para realizar la lectura de todos los medidores tradicionales al finalizar el mes, es que se realizó una estimación de la energía por leer de estos medidores (ELM). Con esta estimación, se realiza la orden de compra del mes posterior.

El siguiente trabajo tiene como objetivo diseñar un modelo predictivo que permita estimar la lectura de los medidores cuyo consumo no se alcanza a facturar al finalizar el mes. Para ello, se utilizó la facturación histórica de los clientes para entrenar modelos tradicionales y modelos de aprendizaje automático para la estimación de series de tiempo. Se implementó la metodología KDD para realizar los pronósticos, que fueron posteriormente evaluados.

De los modelos implementados, el que logra obtener mejores resultados fue el modelo autorregresivo ARIMA(2,1,0). Respecto al error actual en la determinación de la ELM, ARIMA(2,1,0) alcanza una mejora del 30%; por otro lado, el modelo Prophet logra obtener una mejora del 13% en el cálculo del error.

Se propone como mejoras futuras agregar nuevos periodos de información a la data ya entregada, ya que ésta es una de las razones principales del bajo desempeño obtenido por los modelos más complejos. Además, se sugiere agregar nuevos regresores como el clima, que aporten información oportuna para mejorar la toma de decisiones que conlleva el cálculo del ELM.

TABLA DE CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE ECUACIONES

ÍNDICE DE TABLAS

1. INTRODUCCIÓN

La energía eléctrica es un recurso que ha cobrado vital importancia en los últimos años; el avance de la tecnología y su llegada masiva a las ciudades y hogares ha generado una alta demanda que va en aumento cada día; esto se suma al contexto relacionado a la pandemia y los nuevos hábitos de consumo, como teletrabajo y en general el paso de más tiempo de las personas en los hogares.

Esta fuerte demanda ha llevado a concentrar los esfuerzos de investigación en mejorar la eficiencia de la generación, distribución y consumo de energía eléctrica en hogares e industrias. Con este objetivo, las empresas distribuidoras de energía eléctrica han estado incorporando nuevas tecnologías para obtener información completa y en tiempo real del consumo de sus clientes, como lo es el uso de medidores inteligentes.

En este contexto, la minería de datos ha sido un campo importante de investigación que ha permitido mejorar el estudio de los datos, realizar mejores pronósticos y obtener información más certera para la toma de decisiones.

En este proyecto de tesis se estudia el pronóstico de series de tiempo para mejorar la estimación de energía por leer en los medidores (ELM) con el uso del consumo de energía de los medidores del segmento masivo, incluyendo los medidores inteligentes; esto es para obtener valores más precisos que permitan hacer más eficiente la compra de energía aguas arribas a las generadoras, y así ajustar los costos.

Se utilizó la metodología KDD para implementar diversas técnicas de minería de datos que permiten su análisis para obtener conclusiones sobre el consumo eléctrico. La empresa distribuidora de energía Enel ha proporcionado información del consumo de sus clientes pertenecientes al segmento masivo por un periodo aproximado de un año para el estudio en cuestión.

2. CONTEXTUALIZACIÓN Y PROBLEMA

En el sector de la energía eléctrica existen distintos tipos de actividades desde las cuales puede abordarse el problema del consumo energético. En las secciones siguientes, se entrega un breve repaso de las actividades del sector eléctrico, definiendo la importancia del perfil de demanda y presentando la problemática desde el enfoque del presente proyecto. Finalmente, se mencionan los objetivos y los alcances propuestos para este trabajo.

2.1 CARACTERÍSITICAS DE LA ORGANIZACIÓN

El mercado eléctrico en Chile, desde el lado de la oferta de energía, está compuesto por tres sectores, cuyas actividades hacen posible la disposición de la energía eléctrica en los distintos puntos del mercado. La interconexión física de los componentes de cada uno de estos sectores se denomina sistema eléctrico:

Generación: sector que tiene como función la producción de la energía eléctrica a través de distintas tecnologías tales como la hidroeléctrica, termoeléctrica, eólica, solar, entre otras.

Transmisión: sector que tiene como función la transmisión, en niveles altos de voltaje, la energía producida a todos los puntos del sistema eléctrico.

Distribución: sector que tiene como función el distribuir, en niveles de voltaje más reducidos que los de transmisión, la energía desde un cierto punto del sistema eléctrico a los consumidores regulados que este sector atiende.

Estas actividades son desarrolladas por completo por empresas privadas, las que realizan las inversiones necesarias dentro de la normativa específica que rige para cada uno de estos sectores. Así, los sectores de transmisión y distribución se desarrollan dentro de un esquema de sectores regulados por la característica de monopolio que tienen ambos sectores, mientras que generación lo hace bajo reglas de libre competencia.

La misión de la compañía se centra en permitir el acceso de la energía a más personas, implementado nuevos usos de la energía, como lo son, por ejemplo, la electromovilidad y el uso de medidores inteligentes. Enel distribución se abre a nuevas formas de gestionar la energía, permitiendo con ello generar nuevas alianzas con diferentes *partners*, siendo una compañía líder en implementar un modelo de negocio sostenible, innovador y circular que crea valor a largo plazo para todas las partes interesadas.

El negocio de Enel Distribución se centra en la compra de energía a las generadoras eléctricas para luego distribuirla (venderla) a sus clientes finales dentro de toda el área de concesión, que comprende 33 comunas de Santiago.

El desarrollo sostenible del negocio está esencialmente centrado en las personas, y contempla como eje de trabajo la transición energética, considerando con ello la generación de energía en el futuro, la electrificación, la innovación digital y sus plataformas en cuanto a la cyber seguridad y los soportes digitales. Contempla una cadena de suministro sostenible en el largo plazo y en toda su cadena de valor, preocupada de la salud y seguridad laboral, al igual que de la sostenibilidad ambiental.

Dentro de los objetivos finales del plan de sostenibilidad está la creación de valor para todos los grupos de interés de la compañía en el largo plazo. Durante 2020,

Enel Distribución Chile distribuyó entre sus *stakeholders* todo el valor generado mediante pagos por los bienes y servicios adquiridos a proveedores y contratistas, a los colaborares y a los dueños de capital a través de pago de interés y dividendo, resultando con un Valor Económico Retenido de $ 38.416 millones.

La valorización económica de la energía presenta grandes desafíos. Enel distribución compra energía y vende a sus clientes finales; en este proceso de compra-venta, se generan pérdidas que se deben monitorear y controlar. Es por esto que otro de los objetivos importantes de la compañía es la disminución de las pérdidas totales de energía, que para el 2020 fue de un 5,2%. Para esto, se establecen metas anuales para el mediano y corto plazo.

El área encargada de valorizar todo lo que se ha comprado y vendido es el Área de Planificación, Balance y Recuperación de Energía, que es donde se desarrolló el trabajo de investigación. La valorización económica de la energía se verá reflejada en los balances de energía, que se abarcan desde la perspectiva económica y energética (KwH).

El trabajo desarrollado contempla los 1.959.586 clientes correspondientes al segmento masivo que Enel Distribución tiene a la fecha; dentro de este segmento, se encuentran clientes del tipo residencial (89,7%), comercial (7,7%) y clientes industriales (2,6%). De todo el segmento masivo, el 15% cuenta con medidores inteligentes incorporados que permiten la medición de consumo cada 15 minutos, accediendo con ello a su facturación en casi cualquier momento del mes.

2.2 PERFIL DE DEMANDA

2.2.1 CARACTERÍSTICAS GENERALES

La demanda eléctrica es la potencia presente en los terminales de un sistema promediada en un intervalo corto y específico de tiempo por ejemplo 25, 30 minutos o 1 hora. Se define como la cantidad, de energía que pueden ser adquiridos en los diferentes precios del mercado por un consumidor o por el conjunto de consumidores en un momento determinado. [1] La potencia es la cantidad de trabajo realizado o transferencia de energía por unidad de tiempo y se mide en Watts (W).

Un perfil de carga o demanda eléctrica es una curva que caracteriza el consumo de potencia a lo largo de un periodo de tiempo. Este periodo puede ser diario, semanal, mensual o anual dependiendo del objetivo de estudio. El lapso total se divide en intervalos más pequeños de tiempo y la curva describe para cada uno de ellos un valor de potencia consumida [2].

El estudio de la demanda de potencia horaria es más reciente que el de energía consumida, sin embargo, es un campo que ha sido investigado desde los años 40. La demanda de potencia horaria es más difícil de predecir que la demanda de energía por bloques de tiempo, debido a su naturaleza azarosa y su aspecto

predominantemente fluctuante. En la figura siguiente se observa un ejemplo ilustrativo de curva de demanda diaria [2].

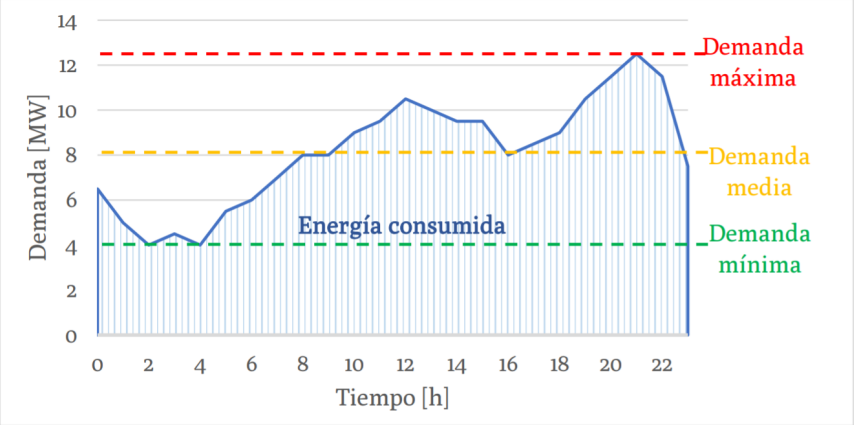


Figura 1: Ejemplo de perfil de demanda diario [2].

De un perfil de demanda es posible extraer variada información acerca del comportamiento del sistema del que se ha obtenido la curva, como, por ejemplo:

- Energía consumida: energía en el periodo o en parte de él que se calcula con el área bajo la curva.
- Demanda máxima: Corresponde a la magnitud más alta de consumo en el periodo estudiado.
- Demanda mínima: es el valor más bajo medido en el periodo representado.
- Demanda media: es el promedio de todos los valores de demanda en el periodo.
- Factor de demanda: razón entre la demanda máxima y la carga total conectada. Esta carga total conectada es la suma de las potencias nominales de los componentes de la instalación.
- Factor de carga: razón entre la demanda media y la demanda máxima del sistema en un periodo dado.
- Factor de simultaneidad: también llamado factor de coincidencia es el cociente entre la potencia eléctrica máxima de un conjunto de cargas y la suma de las potencias máximas de las cargas individuales.

Otros análisis posibles de hacer son: la rapidez de variación de los consumos, que se obtiene por las pendientes de la curva, y los horarios o fechas en que ocurren todos estos eventos.

2.2.2 CARACTERÍSTICAS DEL PERFIL DE DEMANDA

Las formas de las curvas de carga dependen de si corresponden a sectores residenciales, comerciales o industriales. Además, varían fuertemente si corresponden a unidades más pequeñas dentro de cada categoría.

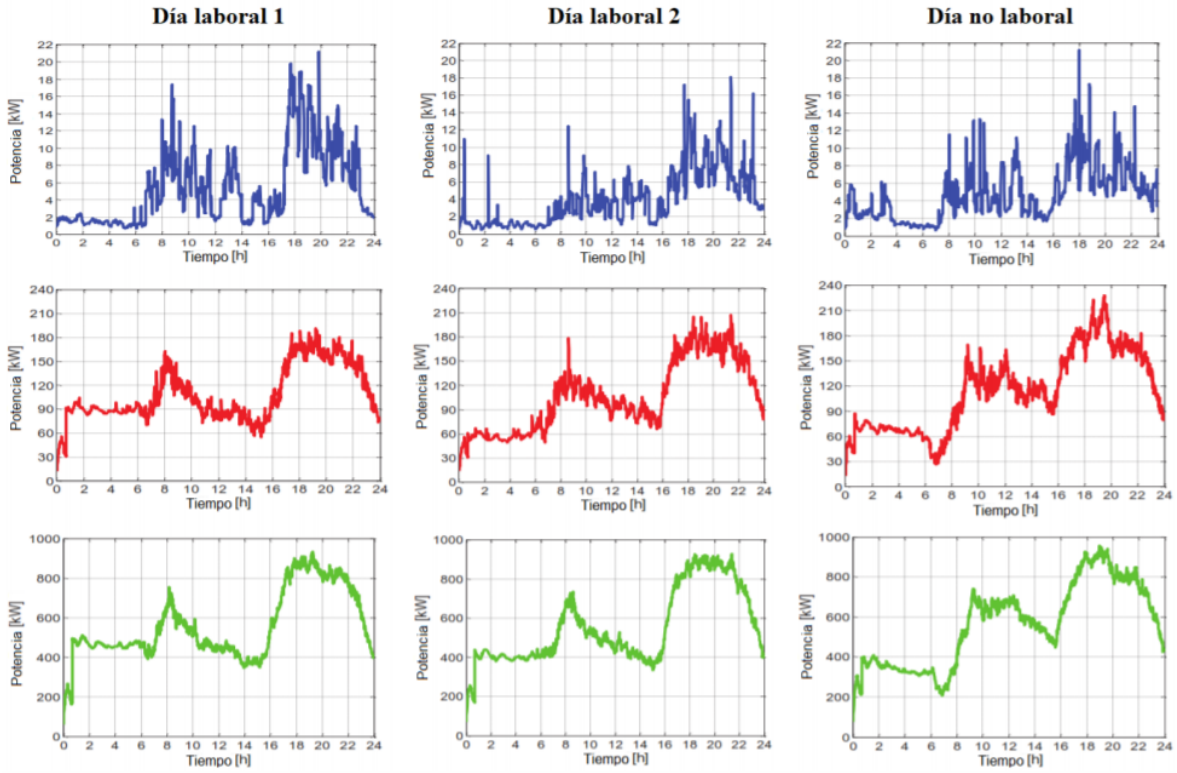


Figura 2: Curvas de demanda en función del nivel de agregación: 10 casas (azul), 200 casas (rojo) y 1000 casas (verde) [2].

Como ejemplo, si se compara las curvas de un día dado para casas de las mismas características y con la misma cantidad de habitantes, el comportamiento tendrá las mismas características de alta variabilidad, pero las curvas no coincidirán. Esto es debido a que los consumos están bastante condicionados por el comportamiento de los usuarios. Si se avanza aguas arriba de un circuito, incorporando un conjunto de consumidores, la curva se va regularizando y tomando un comportamiento muy definido [2].

En la Figura 2 se puede observar este fenómeno, en que para un conjunto pequeño de 10 casas (en azul) los perfiles tienen las mismas características de irregularidad para los distintos días (dos días laborales a la izquierda y centro y un día no laboral a la derecha).

Mientras que si se aumenta la cantidad de casas (a 200 en rojo y 1000 en verde) la curva se va suavizando y definiendo su silueta hacia una forma bien determinada [2].

En general, para un caso en que no existan incentivos tarifarios que los afecten, se da que los consumos residenciales tienen su máximo en las horas de la tarde, entre 18:00 y 22:00 horas aproximadamente (dependiendo del país y la luz natural disponible), y una gran depresión en las horas de la madrugada. Por ende, como la demanda de este tipo varía bastante a lo largo del día, su factor de carga es bastante menor a uno [2].
Por su parte los consumos comerciales tienden a ser más regulares en las horas con luz solar marcando su demanda máxima en ese período, y disminuyendo en las horas de la tarde. A medida que crecen en tamaño, estos consumos comienzan a asemejarse a los industriales disminuyendo la diferencia entre valor máximo y mínimo de la demanda aplanando la curva y acercando su factor de carga al valor uno [2].

Los consumos asociados a lugares de servicios públicos y oficinas siguen el comportamiento de los pequeños comerciales, gracias a que aumentan su demanda en horas laborales (8:00 a 18:00 horas aproximadamente). Es decir, la demanda de éstos disminuye en el mismo horario en que aumenta la residencial [2].

Algunos factores que influyen en la forma del perfil son los relacionados con el momento al que hacen referencia, como: la estación meteorológica, el estado del clima (en especial la temperatura ambiental), la luz solar disponible, el día de la semana, entre otros. Adicionalmente a estos, otros más indirectos son el nivel de electrificación de las viviendas (que indica que tanto se usa la energía eléctrica frente a otras fuentes), los hábitos de la población referentes al uso eficiente de la energía y las señales de precio dadas a través de las tarifas aplicadas a los consumidores [2].

2.2.3 IMPORTANCIA DEL PERFIL DE DEMANDA

En un sistema eléctrico resulta de gran importancia poder predecir el comportamiento de la demanda eléctrica por variadas razones. Una de ellas es poder cuantificar y planear el desarrollo de centrales de generación para determinar el parque de generación necesario y su despacho operacional con tal de cumplir con las restricciones de demanda y seguridad del sistema. También permite la planificación del sistema de transmisión de la energía y con más razón la planificación, control y operación correcta del sistema a nivel distribución [2].

Los perfiles ayudan a la optimización económica del sistema ya que criterios técnicos como las caídas de tensión esperadas, la capacidad de carga de los equipos y las corrientes de cortocircuito a soportar son importantes en el diseño, y si se conoce el comportamiento futuro del consumo es posible acercarse más a encontrar el mejor diseño del sistema al menor costo posible [2].

En la actualidad está siendo relevante el estudio de la predicción de los perfiles ya que se espera que estos comiencen a cambiar de forma notoria, debido a razones de mayor eficiencia en los artefactos eléctricos, la aparición de nuevos consumos,

como autos eléctricos o climatización eléctrica, y el creciente despliegue de la generación distribuida [2].

Otros motivos para el análisis son los incipientes métodos de control del sistema que hacen partícipe al consumidor, como la gestión activa de la demanda, y la inminente aparición de las redes inteligentes que facilitarían estos métodos de control [2].

2.2.4 ENERGÍA POR LEER EN MEDIDORES (ELM)

La estimación de la ELM es utilizada para el proceso de margen compra/venta de energía. Dicho proceso está constituido por una mesa de trabajo multidisciplinario, en la que participa el área de Planificación, Balance y Recuperación de Energía. En esta mesa de trabajo se define mensualmente el margen de compra/venta de la compañía, el que finalmente se ve reflejado en los estados financieros.

El cálculo de la ELM considera el consumo entre la fecha de lectura para facturación y el cierre de mes. Es decir, si se factura a un cliente el día 15 de diciembre, se tendrán 15 días de consumo y el resto (16 días restante de diciembre) serán estimados.

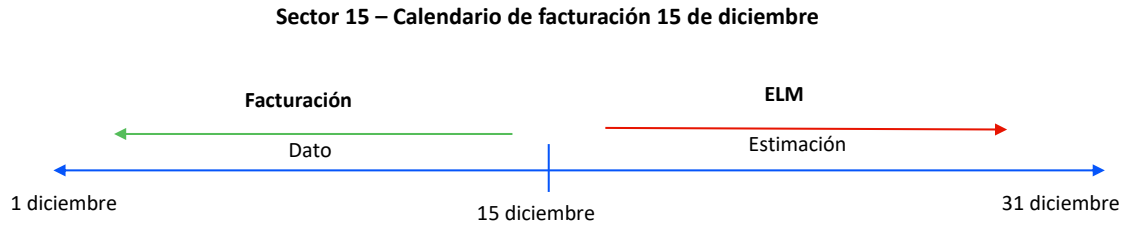


Figura 3: Ejemplo de facturación y cálculo de ELM para un sector en particular (elaboración propia)

La fórmula que permite calcular la ELM, nace de la siguiente relación:

$$\frac{F_{Sn(1:m)}}{C_{Sn(1:m)}} = \frac{ELM_{Sn(m+1:31)}}{C_{Sn(m+1:31)}}$$

$$ELM_{Sn(m+1:31)} = \frac{F_{Sn(1:m)}}{C_{Sn(1:m)}} * C_{Sn(m+1:31)}$$

Ecuación 1: Fórmula de cálculo actual ELM utilizada por ENEL Distribución

La facturación del sector n entre los días 1 y m ($F_{Sn(1:m)}$), es a la Compra asociada al sector n entre los días 1 y m ($C_{Sn(1:m)}$), como la ELM del sector n entre los días m+1 y fin de mes ($ELM_{Sn(m+1:31)}$) es a la compra asociada entre los días m+1 y fin de mes ($C_{Sn(m+1:31)}$).

Los inputs para dicho cálculo son los siguientes:

- Mes de Cálculo
- Calendario de Facturación
- Facturación de cada sector
- Curva de compra diaria (dato calculado)

Por un tema de recursos, la facturación se realiza por sectores (20 en total), un sector por cada día del mes; es decir, el día 1 del mes se factura al sector 1, el día 2 se factura al segundo sector y así sucesivamente hasta el sector 20. Los siguientes 10 días restantes del mes se comienza a facturar nuevamente el sector 1 y así sucesivamente, alcanzando a facturar por segunda vez en el mes los primeros 10 sectores. Con esta segunda facturación de los 10 primeros sectores, se realiza un ajuste con respecto al cálculo ya realizado de la ELM. Esto es, se calcula un error relativo a la ELM estimada con respecto a la facturación real (para estos 10 sectores) y se calcula un promedio de estos errores. Finalmente se realiza un prorrateo a este cálculo realizado.

El error de los cálculos actuales para estimar el ELM, en promedio varía en el orden del 20% al 40%. Por ejemplo, si el error promedio es del 40%, el cálculo de ELM se prorratea al 60%, obteniendo así un mejor ajuste del cálculo total de la ELM. De no realizarse dicho ajuste, la sobreestimación sería muy alta, lo cual se ve reflejado en el balance.

En la práctica, de no corregirse la ELM inicial, el balance arroja que no existen pérdidas de energía; lo cual es físicamente imposible. El balance sin el ajuste arroja valores que se interpretan como una “ganancia” de energía.

Con el ajuste, la pérdida de energía (compra energía – venta energía) se mantiene dentro de los rangos históricos que se consideran “aceptables”. Si se lograra mayor precisión en el cálculo de ELM, la estimación de esta pérdida final de energía sería también más precisa.

3. INFORMACIÓN DEL ÁREA DE LA EMPRESA

La compañía es la encargada de distribuir la energía sus clientes dentro de toda su área de concesión, que comprende 33 comunas de Santiago. A grandes rasgos, el *core* del negocio es la compra de energía a las generadoras y su venta (distribución) a los clientes finales.

Las distribuidoras eléctricas deben asegurar la calidad del servicio entregado, cumpliendo con un conjunto de propiedades y estándares. De no cumplir con los estándares establecidos, la SEC está en el derecho absoluto de amonestar, multar

o sancionar a la empresa distribuidora, lo cual puede ser causado por compraventas de cantidades inadecuadas de energía, según la demanda.

En Enel Distribución, el área encargada de valorizar todo lo que se ha comprado y lo que se ha vendido, es el Área de Planificación, Balance y Recuperación de Energía, que es donde se desarrollará el trabajo de investigación. La valorización se ve reflejada en los balances de energía, que se abarcan desde dos perspectivas: una de ellas es la valorización de la energía como tal (KwH) y la otra perspectiva es la valorización económica de esta energía.

La valorización económica de la energía presenta una variedad de desafíos. Enel distribución compra energía a generadoras y vende a clientes finales. En este proceso de compra-venta existe una pérdida asociada, que es necesario monitorear. Estas pérdidas se dividen en pérdidas técnicas, ocasionadas por las características inherentes de los materiales, cables y elementos utilizados; y las no-técnicas, que ocurren por fraudes o anormalidades, conexiones ilegales, pérdidas comerciales/administrativas y por irregularidades operacionales.

Para decidir la cantidad de energía que se compra cada mes, Enel Distribución debe generar mensualmente el balance de energía de la compra-venta realizada, el cual, por regulación del coordinador eléctrico, debe ser entregado al final de cada mes.

Esta decisión de compra es de suma importancia porque con ella y con las del resto de distribuidoras, se encargará de que las empresas generadoras efectivamente produzcan tal energía durante el mes para cumplir con la demanda de las empresas distribuidoras, y por ende, poder abastecer de energía a los clientes finales. De no hacer una correcta estimación de compra, se cursarán multas por sobrecomprar o subcomprar dependiendo del caso, provocando una pérdida adicional a las mencionadas, que afectará los resultados de la compañía, reflejados en los balances.

3.1 PROBLEMÁTICA – OPORTUNIDAD DE INTERVENCIÓN Y MEJORA

Tal como se mencionó anteriormente, mes a mes se realiza el proceso de margen compra-venta, donde se valoriza el balance. Dicho proceso lo constituye una mesa de trabajo multidisciplinaria en la que participan diferentes actores, y se define mes a mes el margen de compra-venta de la compañía, el que finalmente se ve reflejado en los estados financieros.

El balance debe ser creado y cuantificado considerando el mes completo terminado. La estructura del balance que se realiza actualmente se encuentra disponible en el anexo 1. Con el balance, se puede determinar las pérdidas de energía, las cuales por definición serán la diferencia entre la compra y la venta, medidas en el mismo periodo.

$$P_t = C_t - V_t$$

Ecuación 2: Cálculo de las pérdidas totales

Para este cálculo, se necesita cuantificar tanto las compras como las ventas del periodo. Las compras corresponden a la energía total que ingresa a la zona de concesión, comprada a diferentes generadoras, principalmente Enel Generación. Esta energía que ingresa gracias a la tecnología de los equipos de medición se puede telemedir en cualquier momento, es decir, a fin de mes se puede tener el valor total de la energía que ingresó durante todo el mes (que se compró). Para el cálculo de la venta, se necesitará la facturación de todos los clientes del mes completo. Dentro de estos clientes se encuentran los Grandes Clientes (GG.CC.) y los clientes masivos.

Los GG.CC. son alrededor de 3.000 a la fecha y al igual que los clientes con SMT, se puede telemedir su consumo en cualquier momento del mes, y por ende, se puede tener la facturación completa del mes. Sin embargo, del resto de clientes (masivos) – que son 1.958.954 dentro de la zona de concesión –, el 85% cuenta con medidores tradicionales, que son aquellos que requieren de una persona para registrar su consumo (el personal que visita domicilios para medir el consumo eléctrico) (Figura 4); es por esto que se obtiene la facturación del cliente una vez, en un momento determinado dentro del mes.

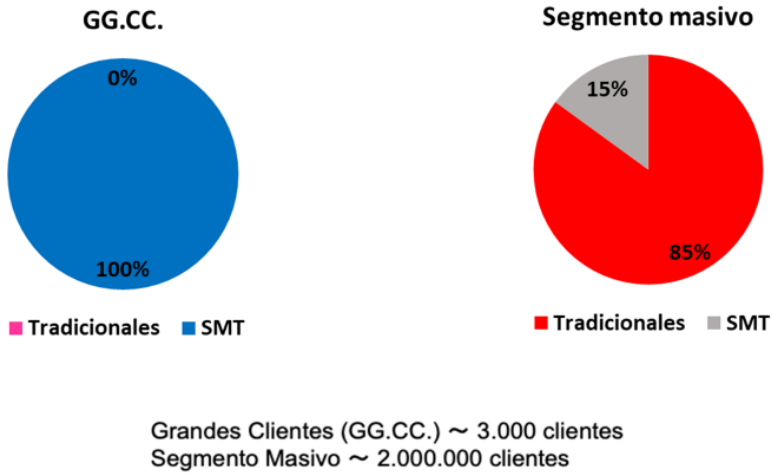


Figura 4: Distribución de clientes de la zona de concesión (elaboración propia).

El 15% del total de clientes masivos corresponde a clientes con medidores inteligentes (SMT). Gracias a esta tecnología, es posible de telemedir el consumo del cliente. Por regulación, se solicita medir el consumo del cliente cada 15 minutos, teniendo a fin de mes un total de 2880 lecturas por cliente. En particular, se puede tener el consumo completo del cliente durante un mes, a diferencia de los clientes con equipos tradicionales.

La forma de facturar de los medidores tradicionales es medir de forma pedestre. Además, por regulación, se solicita facturar dentro de los primeros 20 días hábiles del mes a todos los clientes, cuyo periodo de facturación corresponda a 27-33 días de consumo. Por asuntos logísticos y de recursos para lograr esto, la empresa divide sus clientes en 20 grupos, (en adelante sectores), y factura cada sector por cada día del mes: esto es, el primer día hábil del mes se factura el sector 1, el segundo día se factura el sector 2 y así sucesivamente, porque no existe la dotación suficiente de personal para facturar a todos los clientes el último día del mes.

Entonces, como el periodo de la energía de compra de un mes en particular no coincide necesariamente con los periodos de facturación de los clientes por lo explicado anteriormente, se produce un desfase que se corrige a través del cálculo de la energía por leer en medidores o ELM. Por ejemplo, para un sector en particular, si se factura al quinto día se tendrá una parte de la factura (5 días) que corresponderán al mes en curso y el resto de los días corresponderá a facturación de los días del mes anterior. Como se muestra en la siguiente imagen:

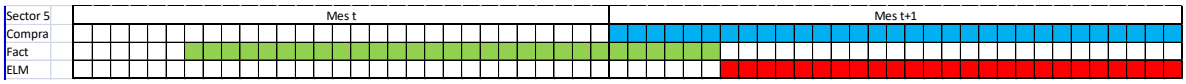


Figura 5: Esquema de facturación del sector 5. Se aprecia en rojo la ELM estimado para el mes t+1, en verde el periodo facturado, y en azul la demanda de energía por comprar para el mes t+1.

La ELM corresponde a la energía (KwH) del periodo comprendido entre la fecha de última lectura y fin de mes. Para los diferentes sectores se tendrá lo siguiente:

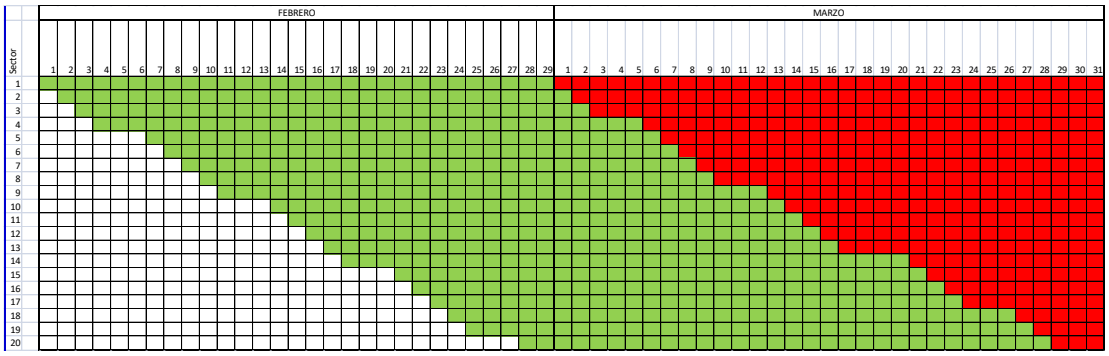


Figura 6: Esquema de facturación de todos los sectores. En verde se muestran los días medidos que se facturan, y en rojo el cálculo de ELM para definir la compraventa, en este caso, con datos del mes de marzo.

El triángulo rojo formado es la ELM a calcular y se observa que los primeros sectores tienen más días calculados con ELM que el resto de los sectores, mientras que los últimos sectores poseen más días medidos por método pedestre (en verde,
Figura 6).

El proceso actual de cálculo de ELM es realizado para cada sector de facturación, estimando las ELM de masivos y de GGCC por separado. Además, se calcula una proporción de crecimiento por sector, considerando así un consumo lineal e igual para todos los clientes de un mismo sector. Sin embargo, los clientes de cada sector no comparten necesariamente características de consumo entre sí ni tampoco proporciones de crecimiento, ya que, como los sectores son definidos por la logística de facturación y características físicas de las conexiones, dentro de un sector puede haber diferentes tipos de clientes, por ejemplo: un mall, un hospital, una casa residencial, un hotel, comercio y una industria. Es claro que sus patrones de consumo energético son muy distintos.

Existen algunas consideraciones para el cálculo del ELM al ser función de la facturación, del calendario y de la compra: se ve muy afectado por estacionalidades, producidas por los cambios de estación lo que generan cambios en el comportamiento de consumo, principalmente en los meses de septiembre y febrero. Como el cálculo de la ELM es crítico para definir la compra a los generadores, y dado que las aproximaciones actuales no consideran estacionalidades, esta investigación propone una alternativa robusta y de alta fidelidad para el cálculo de la ELM, a partir a la explotación de datos reales y proyecciones de crecimiento en base a tendencias encontradas en tales datos.

3.2 HIPÓTESIS Y ALTERNATIVAS DE SOLUCIÓN

Luego de haber analizado la oportunidad de intervención, se propone mejorar la estimación de la ELM con ayuda de los datos generados por los medidores correspondientes al segmento masivo, donde se encuentran incluidos los de medidores inteligentes.

Pensando en los datos disponibles para crear modelos de pronóstico, se propone realizar una estimación del consumo diario, considerando que se tiene la mitad de la facturación del mes y de los meses anteriores a nivel diario. Específicamente, se propone agregar los consumos de todos los clientes a nivel diario con el objetivo de pronosticar sólo los días faltantes del mes, logrando con ello reducir el ruido de la predicción respecto a la estrategia de predecir cliente a cliente.

Así, se plantea como hipótesis que es posible estimar mejor el consumo mensual real de los días que restan del mes para los clientes del segmento masivo, mediante la utilización de los datos existentes históricos y herramientas de inteligencia artificial.

Para modelar y predecir la ELM, se considera como supuesto que la instalación realizada de SMT dentro de la zona de concesión es estadísticamente representativa del parque completo de clientes; es decir, la compañía ha realizado una instalación progresiva de modo tal de poder representar estadísticamente a todos sus clientes a partir de los SMT que se han instalado.

Para la creación del modelo predictivo, se toma como base de datos los de los clientes con SMT, representativos del parque completo de clientes. Se espera obtener una predicción que se ajuste más a la realidad del consumo de cada tipo de cliente (gráfico a la derecha de la Figura 7), debido a que se tiene más información para entrenar mejor el modelo.

Figura 7: Curvas de consumo energéticos de un cliente en distintos escenarios (elaboración propia).

3.3 PROPUESTA DE VALOR

Para cada cliente se tiene una curva de consumo mensual como se muestra en el gráfico a la izquierda de la Figura 7. Hoy, para calcular el ELM, se considera para cada sector que el comportamiento de consumo de los clientes pertenecientes a este sector será lineal, proporcional y el mismo para todos por igual (gráfico central, Figura 7), lo cual no es necesariamente cierto, porque como se mencionó antes, los clientes pertenecientes al mismo sector no comparten las mismas características de consumo, debido a que no son clientes similares entre sí.

Hoy en día en la zona de concesión, hay clientes tradicionales y los SMT. Lamentablemente para el cliente tradicional se tiene una medida al mes, pero para SMT se tiene mucho mayor cantidad de información, debido a la granularidad de las mediciones. Considerando como supuesto que los clientes con SMT son representativos de los clientes masivos de toda la zona de concesión, se creará un modelo de predicción entrenado con información real de alta resolución, por lo que las predicciones de ELM mediante esta herramienta serían considerablemente más similares a la realidad, permitiendo estimar con mucho mayor precisión la compra mensual de energía que se debe realizar.

Con el modelo de predicción entrenado con una base de datos robusta y su predicción de ELM asociada, será posible realizar con mucho mayor precisión el balance de energía, lo que permitirá tomar una decisión de compra informada, precisa y responsable, disminuyendo la existencia de multas por comprar una cantidad de energía distinta a la utilizada.

4. OBJETIVOS

4.1 OBJETIVO GENERAL

Desarrollar un modelo predictivo que permita mejorar la estimación de energía por leer en los medidores (ELM) con el uso de la facturación del consumo de energía de los medidores pertenecientes al segmento masivo, y así obtener valores más precisos para hacer más eficiente la compra de energía aguas arriba a las generadoras.

4.2 OBJETIVOS ESPECÍFICOS

1. Diseñar y estimar los parámetros de modelos utilizados para la predicción de series de tiempo.
2. Realizar la predicción de la ELM para clientes de la zona de concesión, con datos de la facturación de energía de clientes pertenecientes al segmento masivo.
3. Evaluar el desempeño de los modelos propuestos usando indicadores de bondad.

5. METODOLOGÍA

La metodología KDD es el resultado de la estandarización de los procesos relacionados con la transformación de grandes volúmenes de datos en conocimiento útil en cualquier área del saber que necesite de modelos matemáticos (generalmente estadísticos) para interpretar relaciones no triviales en bases de datos [3] [4]. Esquemáticamente, el proceso se puede observar en la Figura 8.

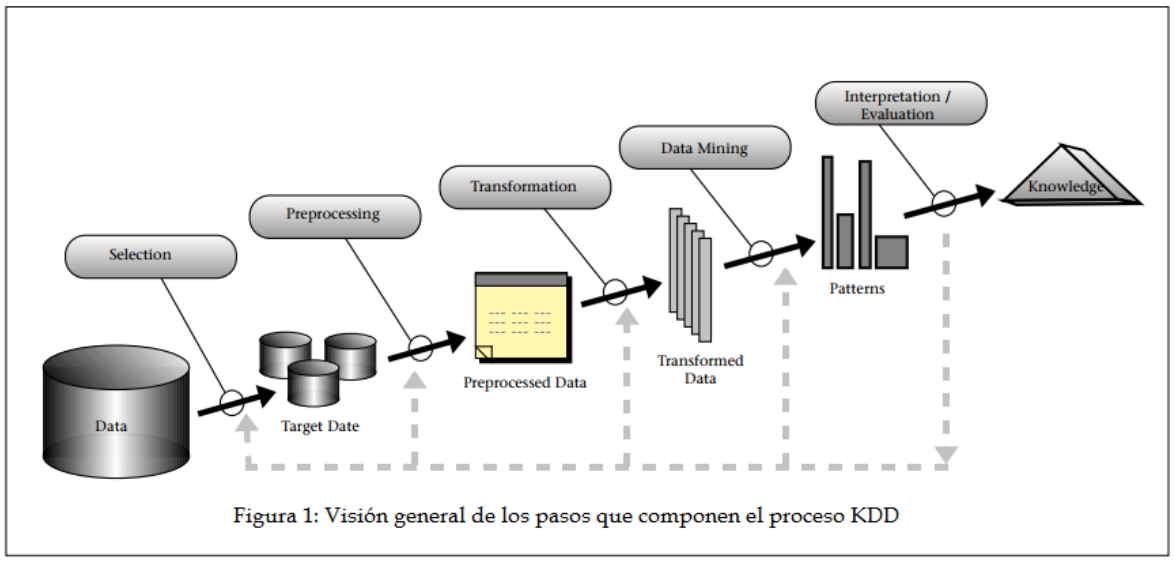


Figura 8: Diagrama del proceso KDD [3].

Otra metodología de trabajo conocida es la llamada CRISP-DM (*Cross Industry Standard Process for Data Mining*) [5], que pretende estandarizar los proyectos de minería de datos, y cuyo enfoque es obtener el mejor provecho del uso de Data Mining al entender de la manera más completa posible el negocio y el problema que se desea resolver.

Lo anterior permite hacer una correcta recolección de datos e interpretar bien los resultados de los análisis, alcanzando los objetivos que se hayan propuesto. CRISP-DM organiza el desarrollo de un proyecto de Data Mining en una serie de fases o etapas, con tareas generales y específicas que permitan cumplir con los objetivos del proyecto. Estas fases funcionan de manera cíclica e iterativa, pudiendo regresar desde alguna fase a otra anterior.

Los pasos a seguir en la metodología CRISP-DM se resumen en los siguientes cuatro puntos [5]: comprensión del negocio, comprensión y preparación de los datos, modelamiento y evaluación, y despliegue del proyecto.

Se puede decir que las etapas de esta metodología están relacionadas de alguna manera con las del proceso KDD; incluso se puede llegar a considerar CRISP-DM

como una implementación del proceso KDD. La analogía entre los pasos que se presentan en ambos casos se puede ver en la siguiente tabla:

KDD	**CRISP-DM**
Identificación del problema en estudio	Comprensión del negocio
Selección e integración de los datos	Comprensión de los datos
Limpieza y pre-procesamiento de los datos	
Transformación de los datos	Preparación de los datos
Selección y aplicación de Data Mining	Modelamiento y evaluación
Interpretación y evaluación	
Post KDD	Despliegue del proyecto

Tabla 1: Cuadro comparativo de las metodologías utilizadas en minería de datos [6]

El proceso KDD se implementa en este trabajo al tratarse de una metodología más completa en relación a sus semejantes, porque cuenta con más etapas donde se analiza paso a paso la información hasta llegar al resultado. Se puede indicar que incluso las otras metodologías parten sus bases sobre el proceso KDD con ciertas variantes, según sus necesidades.

A partir de lo ya descrito, el foco principal de esta investigación se relaciona con la predicción del consumo de todos los clientes pertenecientes al segmento masivo para poder determinar la ELM total del mes, utilizando la metodología KDD mencionada. A continuación, se detalla cada una de las etapas de la metodología escogida a utilizar, según lo expresado por el autor K.M. Han J. [7]:

- **Selección e integración de los datos:** en esta primera etapa del proceso se realiza una selección de las fuentes de datos; éstas pueden ser bases de datos y/o archivos. Además, se eliminan los datos inconsistentes y se combinan las diferentes fuentes de datos que fueron seleccionadas.

- **Preprocesamiento y transformación de los datos**: se seleccionan los atributos que serán utilizados para el análisis y son transformados en un formato apropiado para el análisis que será realizado posteriormente con la minería de datos. Las dos primeras etapas del proceso KDD son las etapas en las que se consume más tiempo dado que es aquí donde se debe tener especial cuidado en la “limpieza” que haya en los datos, ya que sin calidad en ellos no habrá calidad en los resultados obtenidos a través de la minería de datos.

- **Minería de datos**: la minería de datos es la parte medular del proceso KDD y su objetivo, como se mencionó anteriormente, es identificar y extraer patrones de comportamiento descriptivo y predictivos de grandes almacenes de datos.

- **Evaluación de patrones y presentación del nuevo conocimiento**: se aplican distintas medidas, principalmente estadísticas, para identificar los patrones más interesantes. También se utilizan técnicas para visualizar los patrones descubiertos y así facilitar la interacción del usuario con el sistema.

5.1 SELECCIÓN DE DATOS

Los datos utilizados para realizar el presente estudio corresponden a la facturación de los clientes pertenecientes al segmento masivo (1.959.586 clientes). Son bases de datos en formato Access, que entregan la facturación mensual de los clientes entre agosto 2019 y agosto 2020 (13 archivos, 850 MB cada uno), desagregada por sectores (20 sectores distribuidos en 20 tablas en cada archivo) y también la información unificada para todos los sectores (una última tabla en cada uno de los 13 archivos). Cada base de datos contiene 1.959.586 registros (filas) y con 31 columnas de información, dentro de la cual se encuentra información del cliente e información del consumo del cliente (KwH) (ver glosario en anexo 2).

Por otra parte, se tiene a disposición una muestra de la base de clientes SMT (1.000 clientes), para los cuales se tienen dos archivos (.csv): uno contiene la información relativa al cliente, y el otro, la información de su consumo (facturación). Este último archivo se desagrega a nivel mensual y a nivel diario, con una ventana de tiempo entre enero 2019 y agosto 2020. En promedio, se tiene una muestra de 500 facturaciones de clientes por mes.

De las bases de datos de facturación del segmento masivo, de un total de 31 variables (columnas) iniciales, se eliminaron 8 variables relacionadas al cargo monetario, debido a que éstas se ven representadas en una sola variable llamada “TOTAL_CARGOS_AFECTOS” que queda incluida en el set de variables a analizar. La variable “SECTOR-ZONA” también fue eliminada debido a que es una concatenación de otras dos variables, “SECTOR” y “ZONA”, que quedarán a disposición para ser analizadas.

Las variables utilizadas para la predicción de las series de tiempo fueron la variable fecha y la variable creada “Consumo_total”, que se obtiene a partir de 8 variables entregadas por Enel. El rango de fechas utilizados corresponde a la totalidad de la información entregada que contempla desde el 29 de julio del 2019 al 12 de agosto del 2020, teniendo un total de 269 mediciones que fueron utilizadas para entrenar el modelo. Dicha información no contempla los consumos del fin de semana dado

que no se tienen lecturas. El resto de las variables fueron utilizadas para caracterizar el segmento en estudio.

5.2 PREPROCESAMIENTO

Esta etapa consiste en la preparación y limpieza de los datos extraídos desde las distintas fuentes de datos para tenerlos en un formato manejable y necesario para las fases posteriores. Se utilizaron diversas estrategias para manejar datos faltantes o en blanco, datos inconsistentes y datos que están fuera de rango, obteniéndose al final una estructura de datos adecuada para su posterior transformación.

Primero se implementó un código en Python que permite la conexión con Access, dejando las bases cargadas a disposición para ser explotadas y analizadas. En cuanto a los valores nulos, las bases no presentan “NULL VALUES”. En el caso de los valores atípicos, de las 1.952.766 facturaciones se eliminaron 11 debido a que su consumo mensual era mayor a 300.000 KwH (3 desviaciones estándar por sobre la media de la variable “CONSUMO_TOTAL”).

5.3 TRANSFORMACIÓN

Consiste en el tratamiento preparatorio de los datos para la generación de nuevas variables a partir de las ya existentes, con una estructura de datos apropiada. Se realizaron operaciones de agregación o normalización, consolidando los datos de una forma necesaria para la fase siguiente.

En primer lugar, se creó una nueva variable para obtener el consumo total de energía en un mes para cada cliente. Esta variable se obtiene a partir de una fórmula entregada por la contraparte, la cual se utiliza para facturar. Dicha variable nace de una serie de condiciones, que dependen de las siguientes variables:

[CONSPROVIS, ENERGIA, ENERGIA BASE, CONSADIC, CONSMEDDIA, CONSMEDNOC, CONSMEDPTA]

Esto implica que la nueva variable “CONSUMO_TOTAL” depende de las variables anteriores y de su magnitud.

Por otra parte, para el desarrollo del modelo, fue necesario unificar todas las bases con la facturación mensual de los clientes en una “base madre”, que incluye todos los meses y todos los clientes que se tienen a disposición para el análisis. El resultado de esto es una base unificada con los 13 meses de facturación, para los casi 2 millones de clientes.

Para agilizar la carga de las bases de datos se implementó un código que las transforma a formato (.csv). Finalmente, para la implementación de los modelos de

data mining, se agrupó la información entregada, obteniendo como data final el consumo diario (a nivel agregado) de los clientes en KwH.

5.4 DATA MINING

Es la fase de modelamiento propiamente tal, en donde métodos inteligentes son aplicados con el objetivo de extraer patrones previamente desconocidos, válidos, nuevos, potencialmente útiles y comprensibles, y que están contenidos u ocultos en los datos.

La minería de datos ha se ha vuelto una de las áreas de estudio de mayor desarrollo e impacto en las distintas ciencias, principalmente porque presenta una oportunidad muy importante en la industria en general. Acorde a Myatt [8]: “el volumen de datos generado ha llevado a una sobrecarga de información y la habilidad de obtener algún sentido de ella se ha vuelto cada vez más importante”. En este contexto se define la minería de datos como “la extracción de información previamente desconocida de grandes bases de datos que pueden ruidosas y tener datos perdidos” [9].

6. MARCO TEÓRICO

En este capítulo se presenta una revisión del estado del arte relativo al reconocimiento de patrones y pronóstico de consumo eléctrico. Se comienza repasando las series temporales, indicando las diferentes técnicas que se utilizan en el desarrollo del trabajo y el proceso general que se sigue para obtener resultados. Finalmente, se muestran las métricas más comunes utilizadas para medir el desempeño del pronóstico.

6.1 SERIES TEMPORALES

Una serie temporal es un conjunto de muestras tomadas a intervalos de tiempos regulares. Es interesante el análisis de su comportamiento porque permite la detección de patrones y la confección pronósticos de su comportamiento. Una característica intrínseca de las series de tiempo es que las observaciones adyacentes pueden además ser dependientes entre sí. La naturaleza de dicho fenómeno es de considerable interés práctico. El análisis de series de tiempo se refiere a las técnicas para llevar a cabo el estudio de dicha dependencia, desarrollando modelos matemáticos que proveen descripciones plausibles para los datos de una muestra.

Con el objetivo de describir el carácter de datos que pareciesen fluctuar de manera aleatoria en el tiempo, se asume que una serie de tiempo puede ser definida como

una colección de variables aleatorias indexadas de acuerdo al orden en que son obtenidas en el tiempo, x_1, x_2, x_3, . . . , donde la variable aleatoria x_1 denota el valor obtenido de la serie en el primer instante de tiempo, la variable x_2 denota el valor para el segundo período de tiempo y así sucesivamente. En general, una colección de variables aleatorias, $\{x_t\}$, indexada por t, es usualmente referida como un proceso estocástico. Los valores observados de un proceso estocástico son entendidos como realizaciones de un proceso estocástico. Usualmente es posible inferir el uso del término serie de tiempo dependiendo del contexto en el que es usado, pudiendo referirse de manera genérica al proceso, o a una realización particular, sin necesidad de hacer una distinción en la notación de ambos conceptos. [10]

A continuación se presentan algunas de las técnicas utilizadas en el desarrollo del trabajo para el pronóstico de la demanda eléctrica.

6.1.1 PROMEDIO MÓVIL

Suponiendo que la demanda de un periodo determinado varía lentamente, el modelo más simple posible a utilizar es el modelo constante. Este modelo supone que las demandas en diferentes periodos están representadas por desviaciones estables que tienden a un promedio y que son completamente independientes de las desviaciones aleatorias en el tiempo.

Introduciendo la notación del modelo:

$$X_t : Demanda\ en\ el\ período\ t$$
$$A : Demanda\ media\ por\ período$$
$$\varepsilon_t : Desviación\ aleatoria\ independiente\ con\ media\ cero$$

Un modelo constante asume que la demanda en el periodo t puede ser representada como:

$$X_t = A + \varepsilon_t$$

Ecuación 3:Notación modelo constante promedio móvil

Sin embargo, bajo el supuesto que la demanda en el tiempo varía lentamente, razonablemente es mejor utilizar un modelo que considere sólo el período de las últimas demandas. A este modelo se le denomina modelo de promedio móvil y la idea es tomar el promedio sobre los N valores más recientes, como sigue,

$\hat{a_t} : Estimación\ de\ A\ después\ de\ observar\ la\ demanda\ en\ el\ período\ t$
$\hat{x_{t,\tau}} : Estimación\ del\ período\ \tau > t\ al\ observar\ la\ demanda\ en\ el\ período\ t$

Se obtiene

$$\hat{x_{t,\tau}} = \hat{a_t} = \frac{(x_t + x_{t-1} + x_{t-2} + \cdots + x_{t-N+1})}{N}$$

Ecuación 4: Estimación modelo promedio móvil

El valor de N depende de qué tan lento varía la demanda en un período; razonablemente para una demanda que varía muy lentamente puede ser mejor un N grande, sin embargo, si esta demanda varía no tan lentamente, entonces se vuelve más ideal un N pequeño [11].

6.1.2 SUAVIZAMIENTO EXPONENCIAL

El modelo de suavizamiento exponencial es en muchos aspectos similar al modelo de promedio móvil, sin embargo, la diferencia radica en que se utiliza una combinación lineal entre el pronóstico del período anterior y la demanda más reciente. El modelo entonces es,

$$\hat{x_{t,\tau}} = \hat{a_t} = (1 - \alpha)\hat{a_t}{-1} + \alpha x_t$$

Ecuación 5: Estimación modelo suavizamiento exponencial

Donde, $\alpha : Constante\ de\ suavizamiento$ ($0 < \alpha < 1$)

Si elegimos un $\alpha = 0$ significa que no se está modificando el pronóstico, si elegimos un $\alpha = 1$ significa que se está eligiendo la demanda más reciente como nuestro pronóstico [12].

6.1.3 MODELOS AUTOREGRESIVOS

Las variaciones estocásticas en los modelos de demanda que se han considerado en las ecuaciones de los modelos anteriores se asumen que son independientes. Sin embargo, es evidente que en situaciones esto no es cierto, ya que, por ejemplo, a veces se puede esperar que la demanda en períodos consecutivos este negativa o positivamente correlacionada. Una de las técnicas de pronóstico que pueda manejar variaciones estocásticas correlacionadas es la desarrollada por (Jenkins, 2013). La técnica desarrollada es conocida como modelo 'AutoRegressive Integrated Moving Average' (ARIMA). Existe una gran variedad de estos modelos, y es común utilizar la notación ARIMA (p, d, q) donde,

$$AR: p = Orden\ de\ la\ parte\ autoregresiva\ del\ modelo.$$
$$I : d = Grado\ de\ primera\ diferencia$$
$$MA: p = Orden\ del\ promedio\ móvil$$

Si se considera el caso donde el orden integrado (o de primera diferencia) es cero, denotando ARIMA (p, 0, q) se tiene el siguiente modelo general,

$$x_t = a + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \beta_2 x_{t-p} \ldots + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_1 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

Ecuación 6: Estimación modelo ARIMA

En la ecuación anterior, ε_t son errores que son considerados como variables exógenas para el modelo que tienen por objetivo predecir el comportamiento estocástico de la demanda a través de la correlación entre los errores de periodos consecutivos. [13]

6.1.4 RANDOM FOREST

Los modelos tradicionales de predicción son bastantes flexibles; en particular, el modelo ARIMA lo es, pero están limitados por el supuesto de dependencias intertemporales lineales. Por consiguiente, no son capaces de capturar efectos no lineales, que a menudo están presentes en el mundo real (Zhang et al., 1998, 2001; Gooijer y Hyndman. 2006).

Una variedad de modelos ha surgido buscando capturar los efectos no lineales. Unos ejemplos son el modelo bilineal (Granger y Anderson, 1978), modelo STAR (Chan y Tong, 1986), modelo ARCH (Engle, 1982) y el modelo GARCH (Bollerslev, 1986; Taylor, 1987), entre otros. Lamentablemente, la mayoría de estos modelos funcionan bien para problemas específicos, pero no logran generalizar para otras series temporales no lineales.

La insuficiencia predictiva y el crecimiento de los procesadores computacionales han impulsado el crecimiento de los modelos de *machine learning*. La gran ventaja de estos modelos, sobre los modelos clásicos no lineales es que no requieren supuestos sobre la estructura preliminar de los datos (Ej. estacionalidad).

Árboles aleatorios [14], también conocido como *random forests* en inglés, es una técnica que promedia las predicciones de una gran cantidad de árboles de decisión no correlacionados entre ellos. Usualmente se obtiene un buen rendimiento con mejores propiedades de generalización que los árboles de decisión, son relativamente robustos a los *outliers* y prácticamente no requieren ajuste de parámetros (Raschka, 2015).

Random forest es uno de los modelos de predicción más robustos que existe actualmente. Presenta una gran capacidad de procesamiento en cuanto a la cantidad de datos y variables utilizadas. Además, presenta buen desempeño respecto de los valores fuera de rango (*outliers*), funcionando bien inclusive con varios *missing values*. A pesar de tener estas ventajas, y debido a hecho de tratar con muchos datos, conlleva a un gran costo computacional y tiempo de procesamiento, pudiendo incluso sobreajustar el modelo, generando pronósticos ruidosos.

Random Forest es una modificación sustancial de *bagging* que construye una larga colección de árboles no correlacionados y luego los promedia. La idea esencial del *bagging* es promediar muchos modelos ruidosos, pero aproximadamente imparciales, y por tanto, reducir la variación. Los árboles son los candidatos ideales para el *bagging*, dado que ellos pueden registrar estructuras de interacción compleja en los datos, y si crecen lo suficientemente profundo, tienen relativamente baja parcialidad. Producto de que los árboles son notoriamente ruidosos, ellos se benefician grandemente al promediar [14].

Cada árbol es construido usando el siguiente algoritmo:

i. Sea N el número de casos de prueba, M es el número de variables en el clasificador
ii. Sea m el número de variables de entrada a ser usado para determinar la decisión en un nodo dado; m debe ser mucho menor que M
iii. Elegir un conjunto de entrenamiento para este árbol y usar el resto de los casos de prueba para estimar el error.
iv. Para cada nodo del árbol, elegir aleatoriamente m variables en las cuales basar la decisión. Calcular la mejor partición a partir de las m variables del conjunto de entrenamiento.

Para la predicción, un nuevo caso es empujado hacia abajo por el árbol. Luego, se le asigna la etiqueta del nodo terminal, donde finaliza. Este proceso es iterado por todos los árboles en el ensamblado, y la etiqueta que obtenga la mayor cantidad de incidencias es reportada como la predicción [15].

6.1.5 REDES NEURONALES (KNN)

Una red neuronal se puede describir como un modelo de regresión no lineal cuya estructura se inspira en el funcionamiento del sistema nervioso. En términos generales, una red neuronal consiste en un gran número de unidades simples de proceso, denominadas neuronas, que actúan en paralelo. Además, estas neuronas están agregadas en capas y conectadas mediante vínculos ponderados [13]. En la Figura 9 se representa gráficamente el concepto de una red neuronal.

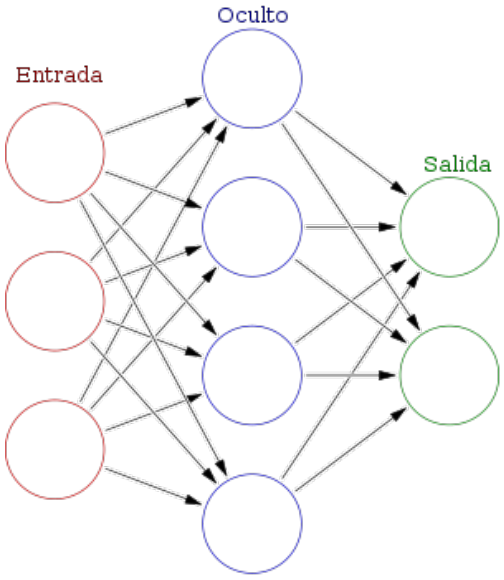


Figura 9: Diagrama de una red neuronal artificial [16]

Al igual que en el cerebro, cada neurona recibe distintos *inputs* desde otras neuronas y genera un resultado que depende solo de la información localmente disponible, ya sea almacenada internamente o plasmada en los ponderadores de las conexiones. Finalmente, el *output* generado por la neurona servirá de *input* para otras, formándose una red de neuronas.

A continuación se presenta una representación gráfica del funcionamiento de la red neuronal:

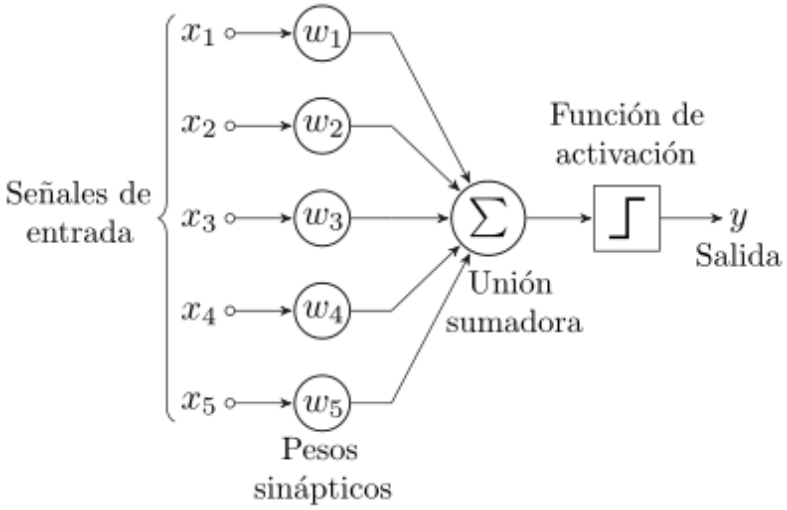


Figura 10: Funcionamiento de una red neuronal [17]

La función de activación es una función que emula el umbral presente en el sistema nervioso. Esto significa, si la respuesta de una neurona no es lo suficientemente grande, entonces esta no afecta en las siguientes neuronas. Las funciones más usadas son la función de escalón, signo, sigmoidal, gaussiana y lineal.

Al igual que el modelo de *random forest*, los algoritmos de redes neuronales pueden aprender por si mismos a medida que se le proporciona nueva información, el cual adapta las ponderaciones de las conexiones entre nodos [18].

Es importante destacar que, a mayor cantidad de capas, la red neuronal puede incorporar más información. En particular, un modelo sin capas ocultas es equivalente a un modelo de regresión lineal. Lo anterior se debe a que los valores de salida corresponden a una combinación lineal de los valores de entrada.

Una de las principales ventajas de la red neuronal es su flexibilidad, ya que puede manejar cambios leves en la información de entrada sin modificar su resultado final. Sin embargo, esto conlleva a que no se pueda interpretar la información aprendida; la red sólo proporciona un valor de salida, pero los pasos intermedios son una "caja negra" [19].

Al tratar con series de tiempo en redes neuronales, los valores en las entradas se corresponden con el nivel de la función en los periodos pasados, de igual manera que se vería en un modelo autorregresivo. Se suele utilizar la siguiente notación para especificar estos modelos:

$$NNAR\ (p,\ k)$$

Ecuación 7: Notación general modelo redes neuronales

Donde *p* corresponde al número de observaciones previas a considerar (y_{t-1}, y_{t-2}, y_{t-3}, ..., y_{t-p}) y *k*, el número de neuronas en la capa intermedia. De esta manera, un modelo NNAR(p, 0) es equivalente a un ARIMA (p, 0, 0).

Luego, se puede hacer una generalización adicional en el caso de tratar con series estacionales, de manera de poder incorporar no sólo una determinada cantidad de periodos anteriores a y_t, sino también el valor correspondiente al mismo elemento del periodo anterior. Es decir, si se tratara de una serie con estacionalidad anual, es esperable que el valor para un determinado mes guarde estrecha relación con aquel de igual periodo del año previo. Así, la notación se extiende como sigue:

$$NNAR(p,\ P,\ k)_m$$

Ecuación 8: Notación modelo red neuronal con estacionalidad anual

Donde p hace referencia, justamente, al número de elementos de periodos previos a considerar para el mismo elemento. Es decir, que las entradas son (y_{t-1}, y_{t-2}, y_{t-3}, ..., y_{t-p}, y_{t-m}, y_{t-2m}, ..., y_{t-Pm}), siendo *m* la longitud del periodo (12, si es anual). Luego, un modelo NNAR(p, P, 0)$_m$ es equivalente a un ARIMA (p, 0, 0)(P, 0, 0)$_m$.

6.1.6 PROPHET

En febrero 2017, Facebook liberó una herramienta (de uso público) llamada Prophet. Dicha herramienta es una librería que permite construir modelos de ajuste y pronóstico de series. Para ello no utiliza los métodos más tradicionales mencionados anteriormente como ARIMA, sino que lo que Facebook denomina *curve fitting.*

Prophet permite manejar sets de datos donde existen observaciones nulas, faltantes, *outliers*, cambios de tendencia importantes (por ejemplo, las semanas previas a Navidad) y tendencias no lineales. Es decir, resulta considerablemente versátil para utilizarlo en series de tiempo sobre las cuales se quieran realizar predicciones [18]. Dicha herramienta es un modelo regresivo aditivo con las tres componentes de una serie temporal:

1. **Tendencia**: Detecta automáticamente cambios de tendencia seleccionando los distintos quiebres de tendencia dentro del set de datos. Así, arma una función (la cual está definida por partes) de tendencia lineal o de crecimiento logístico (que alcanza nivel de saturación).

$$g(t) = kt \qquad g(t) = \frac{C}{1 + e^{k(m-t)}}$$

Ecuación 9: Función de tendencia utilizada por el modelo Prophet

Donde C es el nivel de saturación, k es el *ratio* de crecimiento y m es un parámetro *offset*.

2. **Estacionalidad**: Se modela utilizando series de Fourier con el fin de obtener un modelo más flexible. Los efectos estacionales se determinan a través de la siguiente función:

$$s(t) = \sum_{n=1}^{N} a_n \cos\left(\frac{2\pi nt}{P}\right) + b_n \sin\left(\frac{2\pi nt}{P}\right)$$

Ecuación 10: Función de estacionalidad utilizada por el modelo Prophet

3. **Ruido**: El término de error ε_t representa cualquier cambio idiosincrático que no puede ser acomodado por el modelo. Se asume que dicho termino distribuye normalmente.

Además de los tres componentes mencionados, Prophet permite un cuarto para días feriados o eventos, cuya finalidad es señalar hechos concretos que pueden alterar los valores de la serie.

En consecuencia, Prophet puede ser descrito por la siguiente suma aditiva de los componentes.

$$x_t = T_t + E_t + H_t + \varepsilon_t$$

Ecuación 11: Función general estimación modelo Prophet

Prophet actúa como una caja negra a la cual se le proporciona una cantidad de *inputs* (serie de tiempo, variables independientes, estacionalidades, feriados o fechas especiales entre otros), procesa la información y entrega el pronóstico. Además, indica los *outliers*, cambios de tendencia, impacto marginal de los feriados y recomendaciones para mejorar el modelo, entre otros.

El modelo Prophet es de fácil uso para principiantes y tiene la ventaja de la fácil interpretabilidad de sus componentes. El modelo se ajusta rápidamente, lo que permite al analista explorar variaciones dentro del modelo [20].

6.2 INTERPRETACIÓN Y EVALUACIÓN

Se identifican los patrones obtenidos y que son realmente interesantes, basándose en algunas medidas y se realiza una evaluación de los resultados obtenidos.

6.2.1 MEDIDA DE ERROR

Para poder analizar la precisión de las estimaciones, es posible hacer uso de varias métricas que miden el error de la estimación respecto del dato real. En adelante se describen algunas de las métricas más usadas.

Error absoluto medio

Es una medida de error que compara directamente el valor estimado de una variable con el valor real. Para ello se obtiene el valor absoluto de la diferencia entre estos valores. Para variables que son de más de una dimensión se suele obtener el promedio de estas diferencias absolutas por cada dimensión. Por lo tanto, si se considera una variable X de dimensión n y su valor estimado X', el error absoluto será:

$$MAE = \frac{1}{N}\sum_{i=i}^{N} |x_i - x_i'|$$

Ecuación 12: Cálculo del error absoluto

En que x_i es el valor de la coordenada i del vector X. El error absoluto tiene unidades iguales a la unidad de la variable evaluada [2].

Error relativo

El error relativo se define a partir de la razón entre el error absoluto normalizado y el valor real de la variable. De esta manera se obtiene un resultado adimensional del error, que puede ser expresado en valor por unidad o porcentaje respecto del real. El cálculo de este error queda entonces dado por:

$$e_{relativo} = \frac{1}{N}\sum_{i=i}^{N}\frac{|x_i - x_i'|}{x_i}$$

Ecuación 13: Cálculo del error relativo

Error cuadrático medio

El error cuadrático es una variación del error absoluto, midiendo el promedio de los errores absolutos al cuadrado. Este hecho penaliza los errores a medida que crecen ya que eleva al cuadrado la diferencia obtenida entre los datos. Se obtiene con esto una medida positiva de los errores sin discriminar si el error fue por sub o sobredimensionamiento, al igual que en los casos anteriores. El cálculo se realiza mediante lo siguiente [2]:

$$ECM = \frac{1}{N}\sum_{i=1}^{N}(x_i - x_i')^2$$

Ecuación 14: Cálculo del error cuadrático medio

Error porcentual absoluto medio

Expresa la exactitud como un porcentaje del error. Como este número es un porcentaje, puede ser más fácil de entender que los otros indicadores. Por ejemplo, si el MAPE es 5, en promedio, el pronóstico está errado en un 5% [21]. La ecuación es:

$$MAPE = \frac{\sum_{t=1}^{N} \left| {(x_t - \hat{x}_t)}/{x_t} \right|}{n} * 100$$

$$x_t \neq 0$$

Ecuación 15: Cálculo del MAPE

Coeficiente de determinación

El coeficiente de determinación es una medida estadística de la bondad del ajuste o fiabilidad del modelo estimado a los datos. Se representa por R2 e indica cuál es la proporción de la variación total en la variable dependiente (Y), que es explicada por el modelo de regresión estimado, es decir, mide la capacidad explicativa del modelo estimado [22]:

$$R^2 = \frac{S_{xy}^2}{S_x^2 S_y^2}$$

Ecuación 16: Cálculo coeficiente de determinación

7. RESULTADOS

En este capítulo se muestran todos los análisis sobre los datos para caracterizar el consumo actual de los clientes. Se comienza con la caracterización y estudio de cada base de datos. Se entrega también un estudio de estadística descriptiva de los atributos entregados por las bases de datos, complementando este análisis con gráficos que enseñen su comportamiento. Finalmente, se explica el preprocesamiento de los datos utilizados para la implementación de los modelos predictivos.

7.1 ANÁLISIS DESCRIPTIVO DE LOS DATOS

Los datos utilizados en el experimento fueron obtenidos a partir de la facturación mensual del consumo eléctrico de los clientes. La base de datos consiste en 13 archivos (.accdb), que contiene mediciones registradas entre julio 2019 y agosto 2020.

Las bases cuentan con 30 campos en total, de los cuales un grupo corresponde a variables demográficas que caracterizan al cliente, como la comuna a la cual pertenece, la tarifa cobrada, el tipo de documento tributario (boleta o factura), el sector de facturación al cual pertenece, entre otras. También posee variables cuantitativas que registran diferentes consumos, sin tener explícito el consumo total por cliente. A partir de estas variables y la formulación entregada por la compañía, se realizó el cálculo para tener el consumo total mensual en KwH por cliente.

A partir de la información anterior se realizó el análisis para un mes en particular para tener una primera noción en cuanto a los órdenes de magnitud de los datos y de *insights* que éstos puedan arrojar. Se escogió el mes de abril, dado que es un mes promedio en cuanto al consumo histórico registrado.

Tabla 2: Estadística descriptiva de la serie temporal original

Nro clientes	1.952.766
Consumo promedio (KwH)	333,59
Desv est (KwH)	2.474,94
Valor mínimo	0
Percentil 25 (KwH)	91
Mediana (KwH)	152
Percentil 75 (KwH)	244
Valor máximo (KwH)	968.714

La figura siguiente evidencia una gran diferencia con respecto a los consumos de los clientes de la compañía. Esto se debe principalmente a que existen diversos tipos de clientes; acá se puede encontrar clientes del tipo residencial, que en proporción son la mayor cantidad de clientes, pero en promedio su consumo es menor a la categoría de clientes industriales, comerciales, municipales y fiscal (ordenados en forma decreciente con respecto al consumo). El valor 0 registrado como el valor mínimo se debe a clientes sin registro de consumo para ese mes en particular.

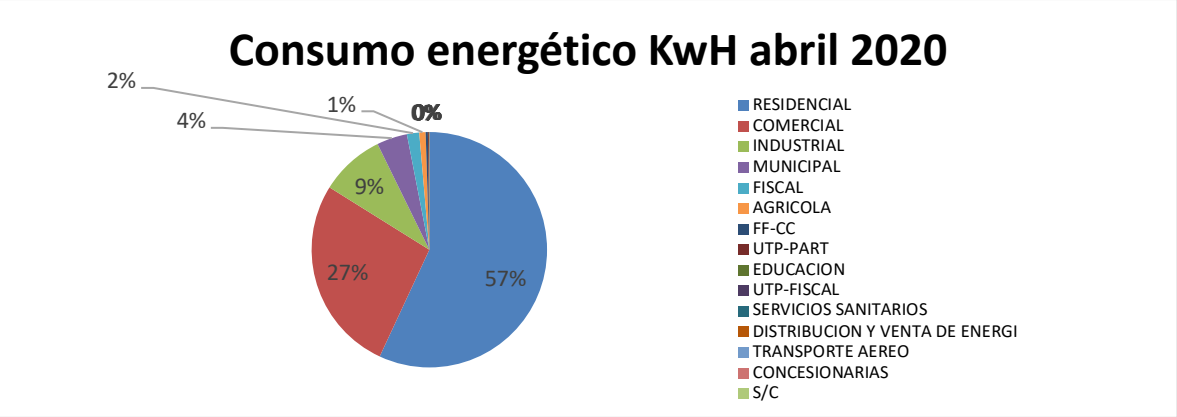


Figura 11: Gráfico de consumo energético por categoría de cliente

La Figura 11 evidencia que el consumo residencial representa más del 50% del consumo total del mes, seguido del consumo comercial, que representa un 27% del consumo total. Con respecto a la distribución consumo por sectores, no existe una diferencia considerable entre sector; el consumo por cada uno representa del 4% al 6% del consumo total, evidenciando que no existen diferencias significativas, como se muestra en la Figura 12.

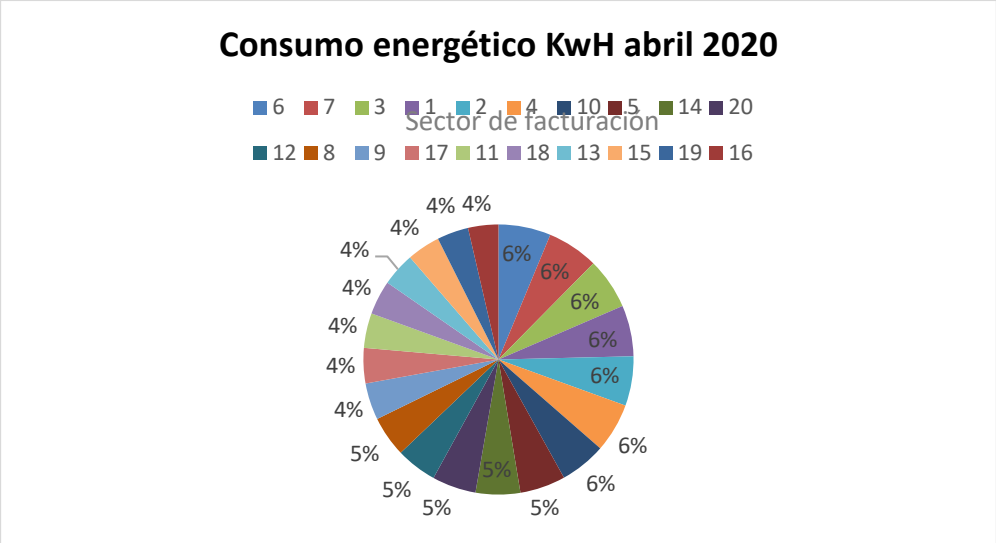


Figura 12: Gráfico de consumo energético por sector de facturación

De las comunas que representan mayor consumo, en primer lugar, se encuentra la comuna de Santiago (14%), seguido de la comuna de Las Condes (9%), Maipú (6%) y Providencia (6%); el resto de las comunas su consumo es menor al 5%; como se puede evidenciar en la Figura 13.

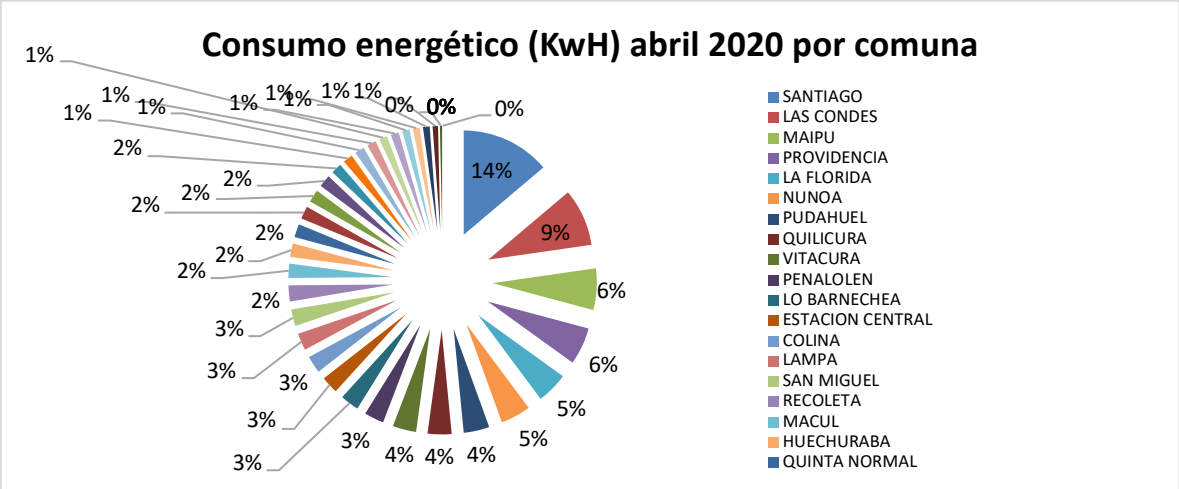


Figura 13: Gráfico de consumo energético por comuna

Al analizar las tarifas que tienen los clientes en sus contratos, es posible determinar que principalmente son tarifas del tipo BT1, que son las que generalmente tienen los clientes del tipo residencial que, como ya se ha mencionado, es el segmento con mayor cantidad de clientes.

El consumo energético total por sector y su variación a lo largo del año 2020 se puede ver en la Figura 14. El sector residencial es el que presenta mayores cambios, especialmente en los meses de invierno; se atribuye a esta estacionalidad el efecto de la pandemia y el aumento del consumo energético debido al mayor uso de calefacción electrificada. Sin embargo, para el resto de los sectores, el consumo se mantiene estable durante los meses analizados.

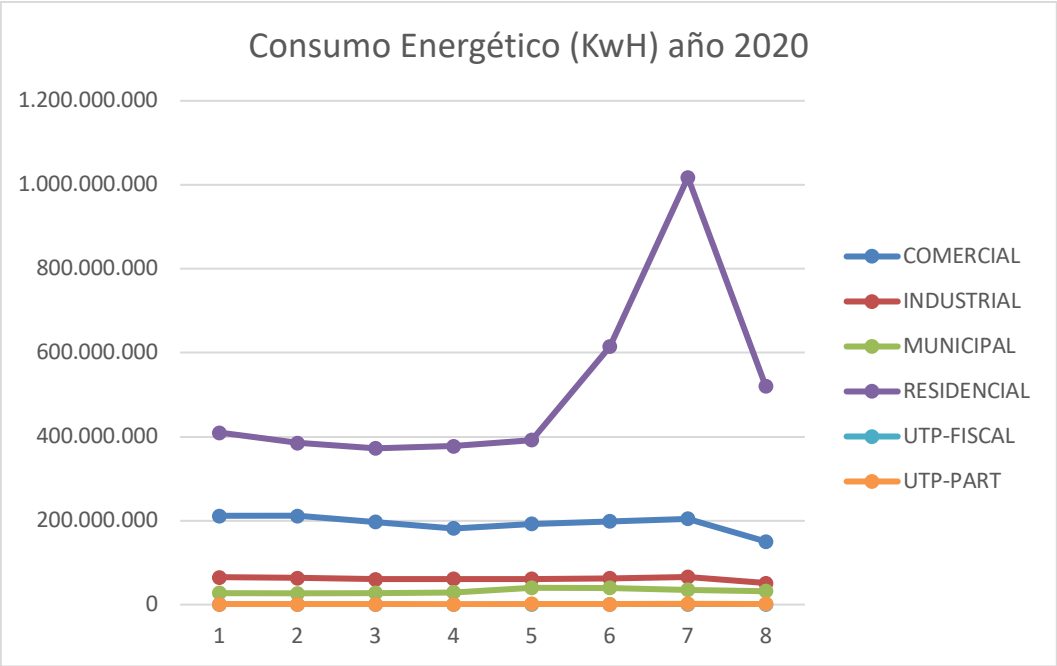


Figura 14: Gráfico de consumo energético en el año 2020 por categoría de cliente

7.2 PREPROCESADO DE DATOS

Antes de realizar los experimentos necesarios para el desarrollo del trabajo propuesto, y luego de haber analizado las bases de datos a utilizar, se realizaron algunos ajustes que permitan el desarrollo correcto de la investigación.

En primer lugar, se verificó la existencia de valores nulos dentro de los registros de consumo. De haberse presentado esta condición, se eliminaron dichos registros. Luego, se eliminaron los registros duplicados, aunque esta condición no aplica a las bases entregadas, dado que cada registro tiene un sólo valor de consumo registrado.

Para analizar las series de tiempo, se agruparon los consumos individuales en consumos diarios totales. Con ello, la dimensión de la data se reduce de casi 2 millones de registros a 275 registros en total; esto contempla un dato por día desde el 29 de julio del 2019 al 25 de agosto del 2020. Dicha información no considera datos de los fines de semana, debido a que las lecturas se realizan en general sólo los días de semana. Sin embargo, existe un par de registros descubiertos para fin de semana, que corresponden a un reemplazo para algunos días que no hubo lectura semanal. Se muestra a continuación en la tabla 11 la estadística descriptiva de los valores diarios obtenidos a partir de la transformación realizada:

Tabla 3:Estadística descriptiva serie temporal consumo energético diario

Cantidad observaciones	275
Media (KwH)	38.557.781
Desviación estándar (KwH)	14.100.449
Valor mínimo (KwH)	0
25% (KwH)	30.886.951
50% (KwH)	37.615.276
75% (KwH)	42.331.680
Valor máximo (KwH)	93.703.490

Se analiza la nueva base para detectar valores atípicos o valores muy alejados de la media de la muestra, obteniendo como resultado la Figura 15:

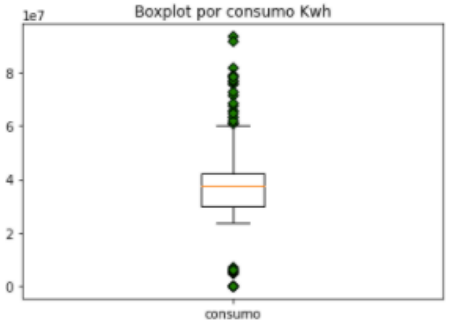


Figura 15: Diagrama de cajas para el consumo diario (Kwh)

Se puede evidenciar que existen valores fuera del rango normal. Dichos valores fueron estudiados y los que están en el extremo inferior fueron eliminados, dado que se concluye que algunas fueron mediciones de prueba y otras, lecturas en 0 que se registran en particular para fechas donde no existe lectura efectiva. Con respecto a los valores superiores, se eliminan aquellos valores superiores a $6x10^7$ para trabajar con valores más normalizados. Esto se refuerza con la siguiente gráfica donde se muestra la variación de los valores de consumo diarios, reflejando variaciones muy altas que principalmente vienen dadas por las mismas razones mencionadas:

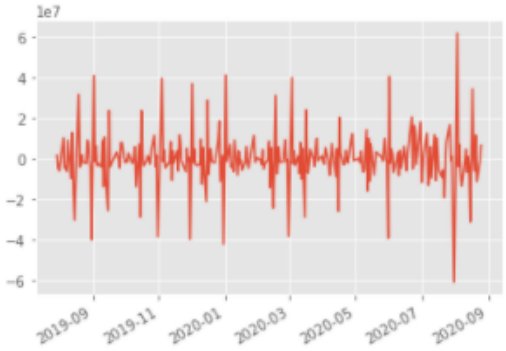


Figura 16: Gráfico de variación diaria del consumo (KwH)

Siguiendo con el análisis descriptivo de los datos finales a utilizar en los modelos, se presentan en las siguientes figuras el comportamiento del consumo energético (Kwh) a nivel diario (Figura 17) y a nivel semanal (Figura 18).

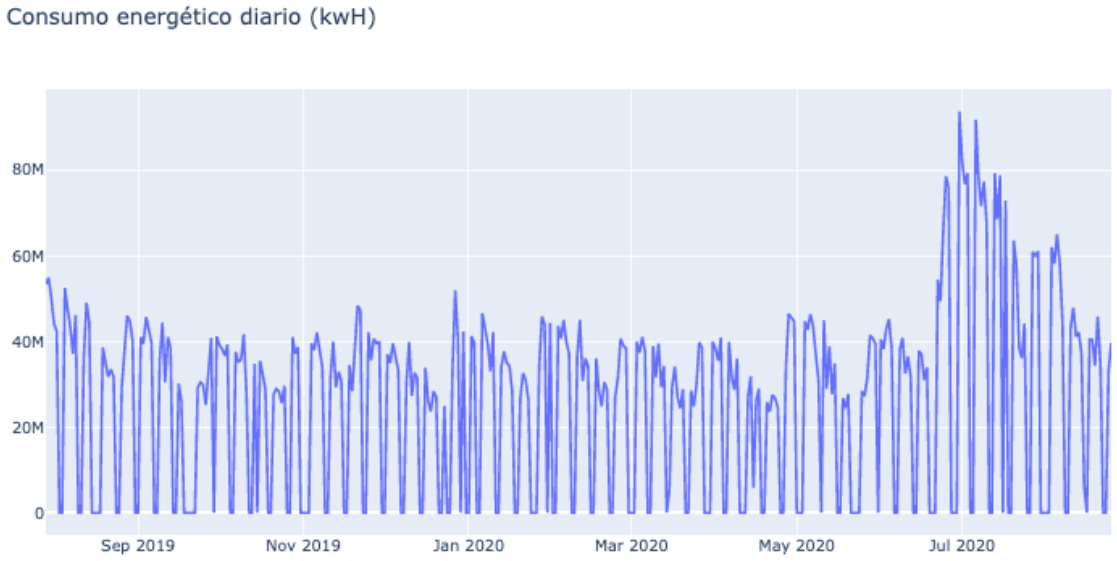


Figura 17: Curvas de consumo energético diario de la serie agregada

De la Figura 17 se puede apreciar que los consumos más altos corresponden a los primeros días de la semana, siendo el día lunes el día que presenta un mayor consumo. Por otra parte, los valores 0 que se muestran en la gráfica corresponden a los consumos del fin de semana, ya que no existen registros para estos días.

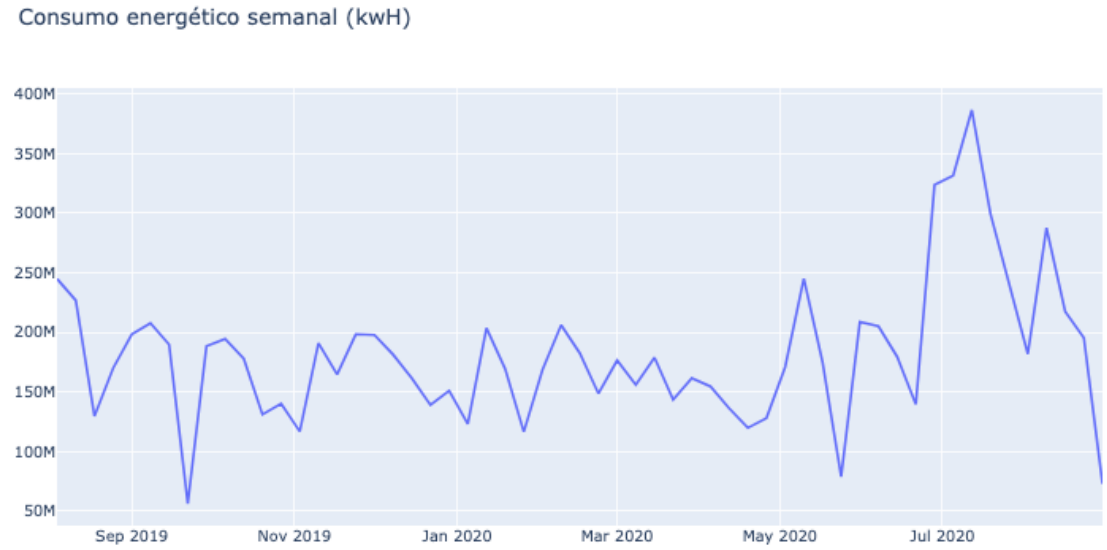


Figura 18: Curvas de consumo energético semanal de la serie agregada

La Figura 18 muestra el consumo a nivel semanal de agregación. Con esto, es más claro ver que los *peaks* se encuentran cercanos a los meses de invierno, debido al aumento de consumo de energía por el uso de sistemas de calefacción eléctrico, y en particular para el 2020 por la pandemia, dado que los usuarios pasaron la mayor parte del tiempo dentro del hogar. Por otro lado, los declives de consumo que se muestran en el mes de septiembre 2019 se deben a que existen 3 días feriados donde no se realizaron lecturas, las cuales se acumularon para la semana siguiente. Los mismo ocurre en el mes de mayo, donde el feriado corresponde a el día jueves 21 de mayo, pero tampoco se realizaron lecturas el día 22 por motivos de feriado a convenio con la empresa. Esto implica que el consumo de la semana se vio afectado y reducido por tener menos días de lectura.

En general el consumo es más bien estable y no presenta mayores variaciones para los meses que no son considerados periodo de invierno (junio a agosto).

7.3 RESULTADO EXPERIMENTAL

En esta sección se muestran los resultados obtenidos para los pronósticos realizados. Se apoyó la evaluación de los pronósticos con gráficos que muestran el valor real versus el pronóstico; además, lo anterior se complementó con tablas comparativas de la evaluación de los modelos con mejores resultados. Se comenzó con la implementación de modelos más simples para culminar con los más complejos.

De las siguientes gráficas presentadas en esta sección, que muestran los resultados de los valores reales versus los pronósticos de los modelos, en el eje X se encuentran los días estimados y en el eje Y el consumo energético en KwH.

7.3.1 PROMEDIOS MÓVILES

Se comenzó con la implementación de un modelo con media móvil, con una ventana rodante de 5 días obteniendo, lo siguiente:

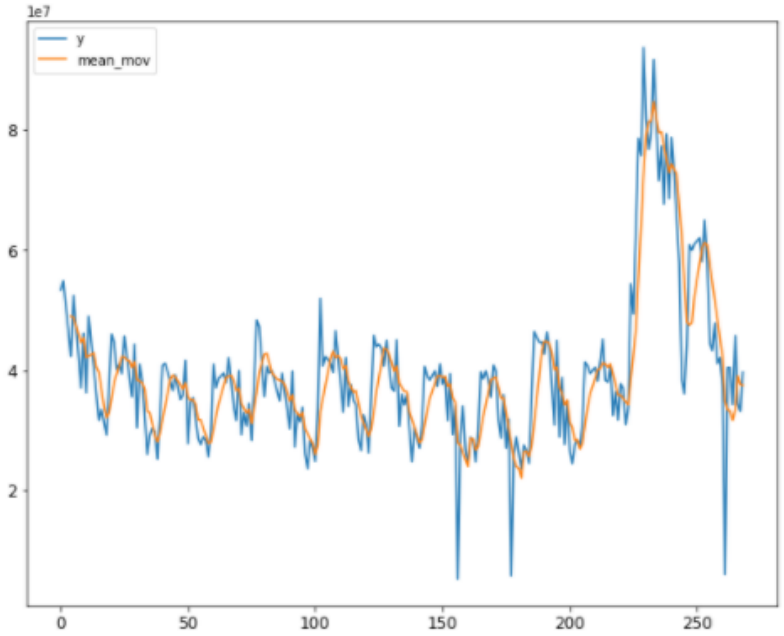


Figura 19: Curva de consumo energético diario versus pronóstico promedio móvil

En la figura anterior, la línea azul representa el consumo real diario en KwH, mientras que la línea de color anaranjado representa el promedio móvil estimado. En la Tabla 4 se muestra el desempeño obtenido por el modelo:

Tabla 4: Resultados ajuste modelo promedios móviles

RMSE	19.285.069,22
R2	0,64
MAE	5.695.351.27
MAPE	0,167

Se evidencia un R^2 mayor a 0,5, que es considerado aceptable para la ventana de información. También se observa un MAE relativamente pequeño acorde al orden de los valores pronosticados; sin embargo, el modelo de promedios móviles, al pronosticar sólo un periodo más hacia adelante, no es suficiente para lo que se desea pronosticar en este trabajo (15 periodos [días] futuros). Sin embargo, es útil para tener como punto de referencia estos errores al ser el modelo más simple a implementar.

En la Figura 19, se puede ver una suave representación de la tendencia del consumo para los periodos señalados; haciendo una descomposición de esta serie, se obtiene lo siguiente:

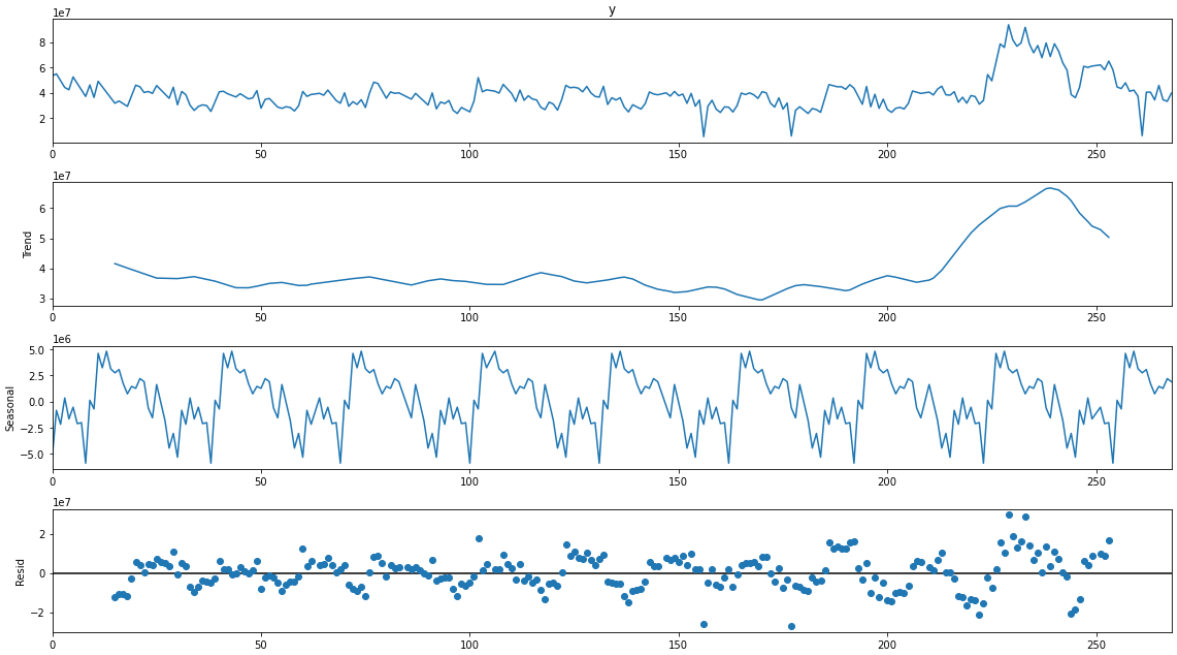


Figura 20: Descomposición serie temporal del consumo energético diario

Se evidencia una tendencia (Trend) al alza para los meses finales de información, que corresponde a los meses entre junio y agosto (desde el valor 200 en adelante del eje X en la Figura 20). Con la descomposición de la serie se aprecia además la variación estacional (Seasonal) que se había mencionado en las secciones anteriores; aquí se muestra un patrón en el comportamiento del consumo diario, evidenciando un mayor consumo para los primeros días de cada mes, decayendo hacia los días finales. Finalmente, con respecto a las irregularidades (Resid) se puede ver que, a partir del mes de mayo (a partir del valor 150 del eje X de la Figura 20), aparecen eventos aleatorios que corresponden a consumos irregulares producto de la pandemia y los cambios en los hábitos del consumo, ligado a que estos meses coinciden con la época más fría del año.

7.3.2 ARIMA

Para la implementación de este modelo se debe cumplir con que la serie sea estacionaria [23].

Con la ayuda de la gráfica de autocorrelación (dispuesta en el anexo 3), se identificó que la serie presenta un rápido decaimiento de los valores de autocorrelación cercanos a 0, con lo que se evidencia la no estacionariedad de ésta. Para cumplir con la estacionariedad de la serie, se aplicó una transformación logarítmica, obteniéndose los siguientes resultados para la prueba Dickey-Fuller (DF):

Tabla 5: Resultado prueba Dickey-Fuller de estacionariedad

```
Results of Dickey-Fuller Test:
Test Statistic                -1.368766e+01
p-value                        1.362690e-25
#Lags Used                     0.000000e+00
Number of Observations Used    2.580000e+02
Critical Value (1%)           -3.455953e+00
Critical Value (5%)           -2.872809e+00
Critical Value (10%)          -2.572775e+00
dtype: float64
```

Para cumplir con la estacionarierdad a partir de la prueba DF se debe verificar que el estadístico sea menor a los valores críticos, lo cual se cumple para los distintos niveles de confianza (99%, 95% y 90%); también debe cumplirse que el P-valor sea lo más pequeño posible [24].

Para la implementación del modelo, la serie se dividió en datos para entrenamiento (80%) y datos para testeo (20%). Con esta división se entrenó el modelo y se ajustaron los parámetros hasta obtener el mejor rendimiento. Luego de realizar diferentes variaciones en los parámetros, el modelo con mejor pronóstico obtenido es el ARIMA(2,1,0).

Finalmente, para hacer la evaluación del modelo, se aplicó la transformación exponencial, regresando los valores a su escala inicial. El resultado gráfico se puede apreciar en la siguiente figura, donde la línea azul representa los valores originales y la línea roja los valores de la predicción.

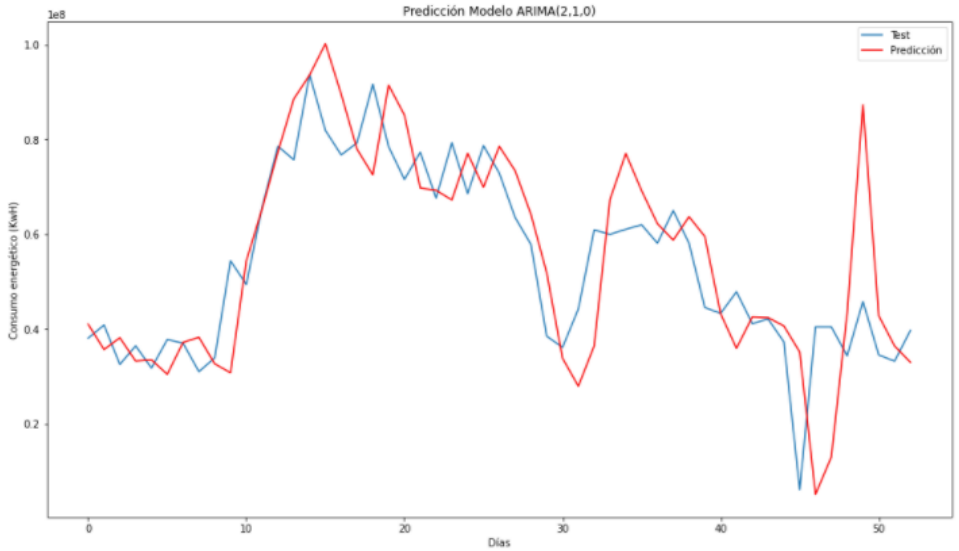


Figura 21: Gráfica de la serie temporal de consumo energético diario versus pronóstico ARIMA(2,1,0)

En tabla siguiente se muestra la evaluación del modelo ARIMA(2,1,0), el cual presenta mejor desempeño que el primer modelo implementado, pronosticando 15 periodos hacia adelante. Se puede observar un MAPE cercano a 0, indicando una mayor precisión por parte de este modelo. El error promedio porcentual del pronóstico es de un 10,25%, obteniendo con esto una mejor precisión con respecto

a los cálculos actuales realizados por la compañía, cuyo error oscila entre el 20 y el 40%; dado esto, la mejora en la predicción con respecto al modelo actual es en promedio un 20%.

Tabla 6: Resultados ajuste del modelo ARIMA(2,1,0)

RMSE	13.281.773,69
R^2	0,505
MAE	9.662.684,8
MAPE	0,269

7.3.3 SARIMA

Se implementó un nuevo modelo bajo la misma lógica del modelo anterior: se dividió la serie en 80/20 y se realizó una normalización de los datos mediante transformación logarítmica. esta vez, se agregaron parámetros para la componente estacional. Se corrieron varios modelos probando diferentes combinaciones de parámetros, para quedarse con el modelo que presenta menor AIC. El resultado es un modelo SARIMA(2,0,2) con componente de orden estacional (0,1,1,5). Se obtuvieron los siguientes resultados para el conjunto fuera de la muestra de entrenamiento:

Tabla 7: Resultados ajuste modelo SARIMA(2,0,2)

RMSE	19.285.069,22
R2	0,35
MAE	16.794.815
MAPE	0,878

De los resultados obtenidos se observa un MAPE alto (cercano a 1), lo que demuestra un error mayor al esperado con respecto al modelo anteriormente implementado. El orden del pronóstico aumenta con respecto a los valores reales, dispersando la precisión del modelo, lo cual se ve reflejado en el MAE.

7.3.4 TBATS

TBATS es una extensión de los métodos de suavizado exponencial, que agrega errores ARMA, términos trigonométricos y componentes estacionales. El modelo se ha descrito en De Livera, Hyndman & Snyder (2011).

Para la serie analizada que presenta dos estacionalidades: una mensual y otra semanal, se decide implementar el modelo TBATS, que pronostica series de tiempo con múltiples estacionalidades para ver si ayuda a mejorar la predicción.

El modelo TBATS tiene varias opciones de modelado; los ajustes realizados para la serie analizada fueron las siguientes:

- Se utilizó transformación box-cox para convertir modelos no lineales en modelos lineales, convirtiendo variables no gaussianas en variables gaussianas.
- Se utilizó una tendencia semanal
- Se utilizó el modelado ARMA en el error

Se ajustó el modelo utilizando la división 80/20 datos de entrenamiento vs datos de testeo; luego, se probaron diferentes parámetros y se determinó el mejor modelo usando AIC, entregando el modelo con mejor ajuste. Se obtuvieron los siguientes resultados:

Tabla 8: Resultados ajuste del modelo TBATS

AIC	9.371,98
RMSE	19.103.735,65
R^2	0,418
MAE	14.239.597,2

Si bien no se puede tomar una decisión apresurada con respecto al ajuste del modelo sólo en base a R^2, sí es posible observar su valor, que entrega una primera noción del comportamiento de esta predicción; en este caso particular, el modelo explica menos del 50% de la variabilidad de los datos en torno a su media. Complementando esto, la Tabla 8 muestra que el orden de magnitud de los pronósticos realizados se aleja de los pronósticos anteriores, según lo observado por el MAE, concluyendo que este modelo no es mejor que los realizados inicialmente.

7.3.5 RED NEURONAL UNIVARIADA

El siguiente modelo es creado gracias a la utilización de la herramienta Tensorflow, que es una plataforma de código abierto que facilita el desarrollo de modelos de aprendizaje automático. Junto con ello, se utilizó Keras, que permite el entrenamiento de hiperparámetros y la implemntación de la solución de los modelos con mayor facilidad.

Se utilizó una arquitectura sencilla de red neuronal llamada FeedForward (concida como MLP o *Multi-Layered Perceptron*), con pocas neuronas, cuyo método de activación es tangente hiperbólica, pues entrega valores entre -1 y 1.

Es necesario en esta parte tener en consideración que la data a trabajar es una serie de tiempo cuyo índice es la variable de tiempo y otra columna que contiene los consumos diarios totales de la compañía para el segmento masivo (KwH). Para preparar la red neuronal, se convirtió el problema en uno de tipo supervisado, transformando la única columna de datos que se tiene en 7 columnas extras que contienen los valores registrados para los siete días anteriores. Este cambio es realizado para entrenar el modelo con *backpropagation*; teniendo entonces como entrada los 7 días previos de consumo energético y como salida el consumo para el octavo día (KwH).

La transformación de los valores (entre -1 y 1) se realiza con la ayuda de MinMaxScaler, favoreciendo el desempeño de la red neuronal. Antes de la creación de la red, se dividió el set de datos en entrenamiento y testeo, al igual que los modelos anteriores (80/20), manteniendo el orden cronológico de los datos, quedando entonces un modelo con 238 entradas (cantidad total de medidas en el periodo estudiado) con vectores de 1x7, teniendo la siguiente configuración:

- 7 inputs de entrada
- 1 capa oculta con 7 neuronas
- Salida de 1 neurona
- Función de activación tangente hiperbólica
- Optimizador utilizado Adam

En primer lugar, se visualiza el conjunto de validación, que en una primera iteración se dejó en 30 días:

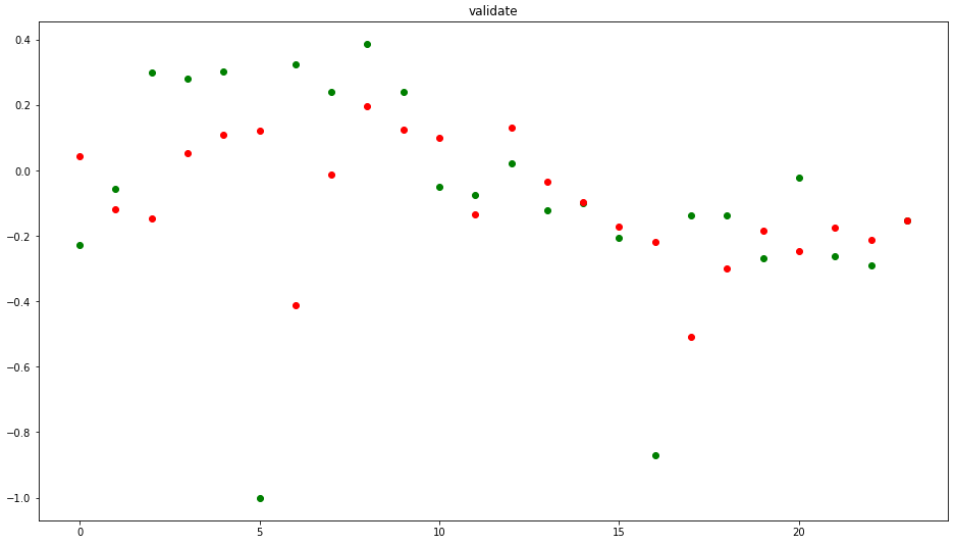


Figura 22: Gráfica conjunto de validación modelo red neuronal. Se puede ver que los puntos verdes (validación) intentan acercarse a los puntos rojos (entrenamiento). Cuanto más cerca estén mejor será el modelo.

Se puede observar en la siguiente figura la función de pérdida del entrenamiento de la red neuronal tanto del conjunto de entrenamiento (línea azul) como el del conjunto de validación (línea naranja). Dicha función evalúa la desviación entre las predicciones realizadas por la red y los valores reales de las observaciones utilizadas durante el aprendizaje. Cuanto menor es el resultado de esta función, más eficiente será la red neuronal. En la siguiente figura, se muestra que la función de pérdida va disminuyendo a medida que la red aprende. Además, pareciera no haber sobreajuste, pues las curvas de validación y entrenamiento son distintas.

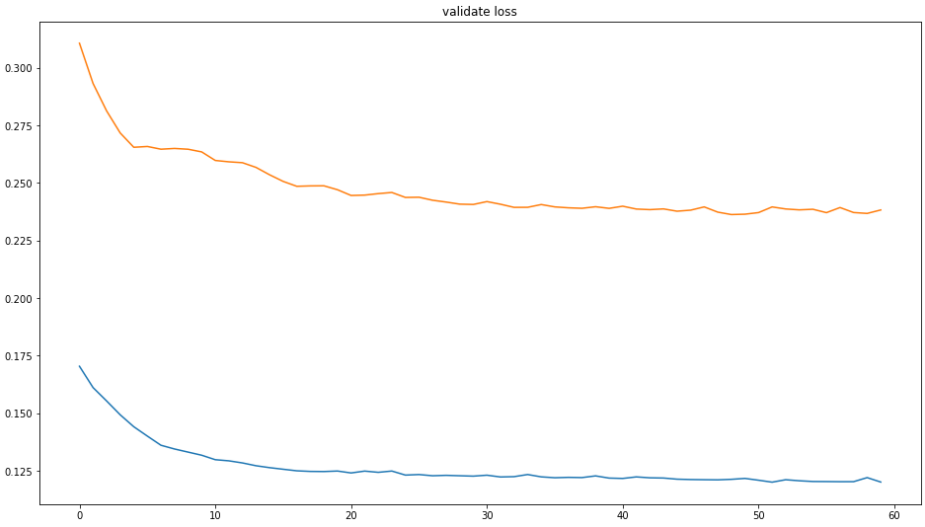


Figura 23: Gráfico de validación de la función de pérdida del modelo red neuronal

Finalmente, los valores obtenidos por la predicción son transformados para obtenerlos en la escala original y se logran los siguientes resultados para el pronóstico realizado para 30 días:

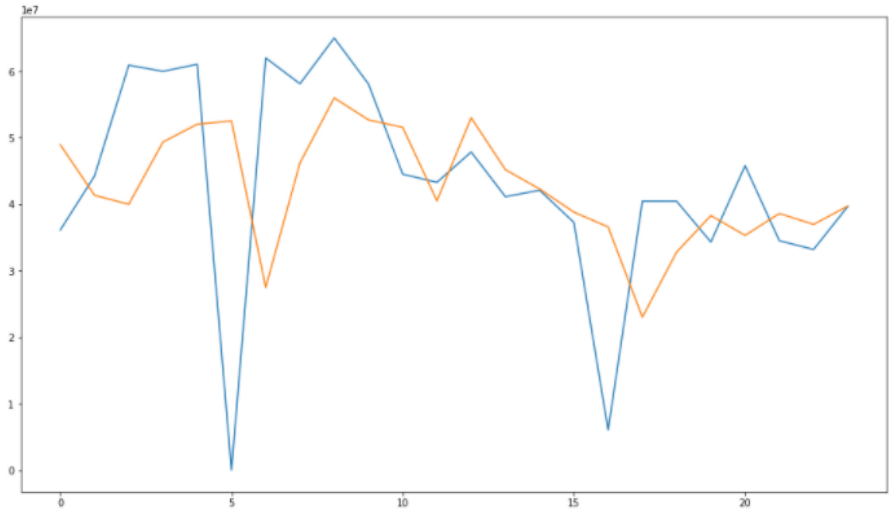


Figura 24: Gráfica de la predicción realizada por la red neuronal univariada entrenada versus valor real. En color azul se observan los valores reales y en color naranjo los valores predichos por la red neuronal. El ajuste de dicho modelo se presenta en la tabla 8:

Tabla 9: Resultados testeo modelo red neuronal univariada

RMSE	16.532.489,10
R^2	0,45
MAE	11.182.209,0

En los resultados obtenidos (Tabla 9) se observa un fenómeno similar al modelo TBATS. Se considera que no necesariamente utilizar un modelo más complejo mejora la calidad de la predicción, ya que esto depende en parte de la calidad de la data utilizada y del consecuente comportamiento del modelo. Finalmente es importante observar que el error se debe en parte considerable a la cantidad de días que se estima, siendo esta mayor a la estimación realizada con anterioridad (30 días vs 15 días).

7.3.6 RED NEURONAL MULTIVARIADA

Se entrenó un modelo similar al anteriormente descrito pero esta vez se agregaron nuevas variables al modelo, lo cual supone mejoras en el desempeño. Dado que el único dato que se tiene además del consumo es la fecha de facturación, se utilizó esta última para agregar nuevas variables. En otras palabras, se transformó la variable fecha en dos nuevas variables: una variable categórica que indica el día de la semana, y otra que indica el mes. A priori, se esperaba que la red "entendiera" las estacionalidades dadas entre semana y meses.

Dado lo anterior, al existir la variable día que toma valores de 0 a 6, y la variable mes que toma valores de 0 a 12, fue necesario trabajarlas para que el modelo no entendiera que una variable que toma un valor mayor (para día o para mes) es necesariamente mejor que la otra. Para ello, se utilizó *embeddings* que sirven para dar valoración útil a datos categóricos. Se asignó una profundidad a cada *embed*, esto quiere decir, un vector con valores continuos inicialmente aleatorios. Estos valores son ajustados con el "*backpropagation*", al igual que la red neuronal.

En resumen, al asignarle un vector con valor numérico continuo a entradas categóricas, los *embeddings* terminan funcionando como una pequeña red neuronal dentro de la red principal, la cual aprende con el mismo mecanismo de *backpropagation*, y resuelve como valores continuos la categoría de las variables (0-6 o 1-12), acentuando su valor intrínseco.

Para el entrenamiento de la red se utilizó la misma configuración mencionada antes, obteniendo como resultado lo siguiente:

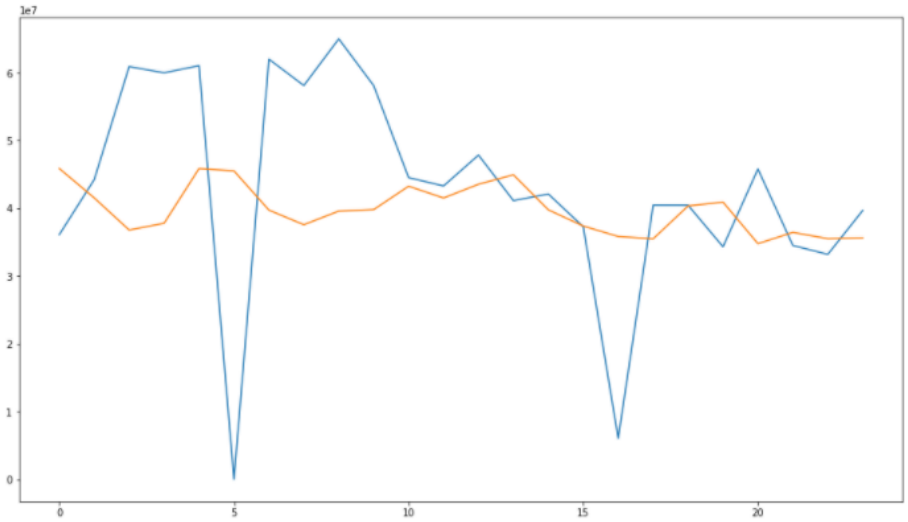


Figura 25: Gráfica de la predicción realizada por la red neuronal multivariada entrenada versus valor real

En la figura anterior se observa que el modelo no mejora el ajuste. La evaluación del modelo para los valores fuera del set de entrenamiento se muestra a continuación:

Tabla 10: Resultados testeo modelo red neuronal multivariada

RMSE	16.484.179,19
R^2	0,34
MAE	11.680.279,5

Se observa que el error de la predicción aumenta con respecto al segundo modelo calculado. Presenta un R^2 menor al modelo anterior y el orden del MAE es similar al de los valores calculados, sin embargo, se prefiere la utilización de ARIMA dado que presenta un mejor pronóstico y es menos complejo de implementar.

7.3.7 RANDOM FOREST

Para la ejecución del siguiente modelo, se empleó la estrategia recursiva para predecir múltiples valores a futuro. Dado que para predecir el momento t+2 se necesita el valor de t+1, que desconoce, fue necesario hacer esta recursividad en las predicciones, en que cada nueva predicción se basa en la predicción del valor anterior. Cada valor estará asociado a una ventana temporal (*lags*) que lo preceden.

Para este modelo se utilizó sólo una parte de la serie, la cual contempla desde agosto 2019 a mayo 2020; esto porque, realizando varias pruebas de

entrenamiento; esta ventana fue la ventana móvil que mejor resultados entregó. Además, considerando el tamaño de la muestra inicial entregada, se decidió priorizar un mejor rendimiento para, cuando se tenga mayor historial, poder reforzar el modelo.

Seguido de lo anterior, se dividió la serie temporal, dejando los 36 últimos días de información para la validación de la predicción, dejando así el 80% para el entrenamiento del modelo y el 20% para evaluar la capacidad predictiva del modelo.

Se creó y entrenó un modelo predictivo autorregresivo a partir del regresor *Random Forest* y una ventana temporal de 12 l*ags*. Esto significa que, para predicción, se utilizaron los 12 días anteriores como predictores.

Una vez entrenado el modelo, se predijeron los datos de prueba, obteniendo lo siguiente:

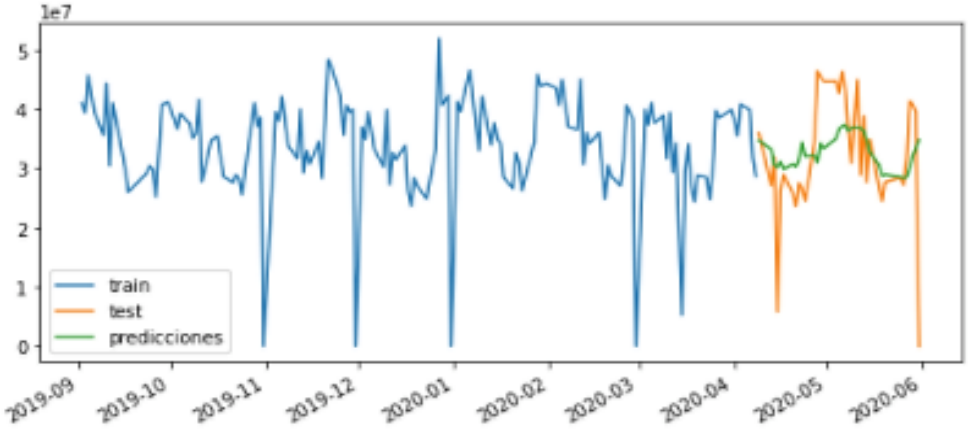


Figura 26: Gráfica de la predicción realizada por el bosque aleatorio entrenado versus valor real

Donde la línea azul representa los datos de entrenamiento, la naranja los datos de testeo y la verde las predicciones realizadas. Una vez obtenida esta información, se procede a evaluar este modelo inicial:

Tabla 11: Resultados testeo modelo Random Forest

RMSE	9.620.031,30
R^2	0,126
MAE	6.938.456,55

El modelo no ajusta perfectamente porque se comenzó iterando con hiperparámetros por defecto. Para identificar la mejor configuración de lags e hiperparámetros, se realizó la técnica de validación cruzada temporal y *backtesting*; que permite al modelo hacer la validación con valores históricos anteriores y así encontrar el mejor ajuste para el modelo.

Luego de realizar la validación cruzada, se determinó que los mejores resultados se obtienen si se utiliza una ventana temporal de 20 *lags* y una configuración *Random Forest* {'max_depth':5.'n_stimators':1000}.

Finalmente, se entrenó el nuevo modelo con la configuración óptima encontrada mediante la validación, reduciendo el error anterior obtenido, como se muestra a continuación:

Tabla 12: Resultados testeo modelo random forest ajustado

RMSE	9.130.645,90
R^2	0,24
MAE	6.540.238,3

7.3.8 PROPHET

Se realizó la implementación del modelo propuesto inicialmente cambiando la forma de la base de datos. Se cambió el nombre de la columna fecha (con formato de fecha 'año-mes-día') a 'ds' y los valores registrados de consumo (KwH) con el nombre de variable 'y'; que es la que se quiere pronosticar.

Siguiendo con la lógica de los modelos anteriores, para hacer la posterior evaluación, se decidió trabajar con el 80% de la base de datos para entrenar el modelo y 20% para evaluar la calidad de este.

El modelo Prophet es un modelo predeterminado y ya desarrollado, por lo que cuenta con todas las instancias pre hechas para poder ser utilizado si es que se le entrega la forma correcta de la base para trabajarlo. Como ya se transformó la base a la forma necesaria para trabajarla, se comenzó ajustando el modelo mediante la creación de una instancia de objeto Prophet. Luego, se realizó el ajuste y se predijeron las fechas futuras para la posterior evaluación del modelo.

El modelo arrojó como resultado el valor predicho, además de un intervalo de confianza que muestra el valor mínimo y máximo que pudiese tomar cada predicción. El modelo fue entrenado con un regresor extra llamado "vacaciones", el cual considera todos los días feriados durante el año, lo que permite tener un mejor ajuste para los días previos y posteriores a éstos (los días incluidos se muestran en el anexo 4).

Por otra parte el modelo, se ajustó automáticamente a las estacionalidades semanales y anuales, dependiendo sea el caso. Dado que la serie analizada solo cuenta con la historia de un año, se ajustó a la estacionalidad semanal y se incluyó una estacionalidad extra mensual.

Prophet incluye la funcionalidad de validación cruzada de series de tiempo para medir el error del pronóstico, utilizando datos históricos. Se utilizó para medir la performance del modelo y para ajustar los hiperparámetros utilizado por éste.

Se presenta a continuación los resultados obtenidos para el modelo inicial sin regresores extras; el segundo modelo contiene regresores que diferencia los días de semana de los fines de semana, el tercero contiene un regresor adicional que muestra si la fecha corresponde a un periodo de invierno o verano y finalmente el ultimo modelo contiene los regresores para los días de vacaciones:

Tabla 13: Resultados testeo modelo prophet para diferentes ajustes

MAE (sin regresores)	13.622.712,15
MAE (días de semana)	9.888.658,4
MAE (invierno-verano)	12.574.698,32
MAE (vacaciones)	10.569.376,21

De los 4 modelos analizados, el que mejor ajuste de predicción presentó fue el que incluye regresores que diferencian los días de la semana frente los días de fin de semana; así, es éste el que queda como modelo a utilizar. Dicho modelo presenta un error promedio porcentual del 16,67% para los días calculados, lo cual mejora la predicción con respecto al cálculo actual utilizado por Enel en un 13% aproximadamente.

8 CONCLUSIONES

Se desarrollaron diferentes modelos predictivos para estimar la energía por leer en medidores. Se realizó un resumen de los modelos estudiados y luego se presentan los resultados para los modelos entrenados; se utilizaron modelos simples y modelos de aprendizaje supervisado más complejos.

De los modelos estudiados, haciendo la evaluación de la predicción fuera de la muestra de testeo, se concluye que el que tiene mejor rendimiento evaluando en términos de MAE es el modelo ARIMA(2,1,0). Se puede ver en la tabla siguiente una comparativa de los desempeños obtenidos ordenados de menor a mayor desempeño.

Tabla 14: Comparación desempeño de modelos de predicción utilizados

	ARIMA (2,1,0)	**PROPHET (días de semana)**	**RED NEURONAL (univariada)**	**RED NEURONAL (multivariada)**	**TBATS**	**SARIMA (2,0,2)**
MAE	9.662.684,8	9.888.658,4	11.182.209	11.680.279,5	14.239.597,2	16.794.815

Se utiliza el MAE porque es de fácil interpretación, arrojando un número en las mismas unidades que la variable de salida, es decir, los valores se muestran en KwH. De los dos mejores modelos implementados se tiene que ARIMA, en promedio muestra un error porcentual de un 10,25% y el modelo Prophet un 16,67%; la mejora porcentual de ambos modelos es de 30% y 13% respectivamente, con respecto al modelo utilizado por Enel Distribución.

Los modelos implementados a lo largo del estudio presentan un horizonte de estimación que, generalmente, depende de la estacionalidad y el orden de la parte autoregresiva para obtener buenos resultados, lo que obliga a realimentar de forma permanente el modelo. Por tanto, para mejorar la predicción, se requiere de una historia más completa que permita estudiar las diferentes estacionalidades que se pueden dar a través del tiempo, y así tener puntos de referencia del consumo con más años de información.

En términos generales, con sólo un año de información, el modelo y su desempeño se ven afectados por el particular periodo de la historia de Chile. Por una parte, existen variaciones en el consumo partiendo por las fechas cercanas al estallido social (octubre 2019). Seguido de esto, la serie estudiada presenta un desajuste en el consumo habitual, que se acentuó en mayor medida en el periodo de invierno (junio -julio 2020). Se ve un claro aumento por sobre la media, para estos valores lo cual se ve sustentado con el efecto de la pandemia; lo cual obliga a las familias a pasar mayor tiempo en el hogar, afectando el consumo residencial; que es el que mayor consumo presenta dentro del segmento masivo en estudio.

En particular, la intuición inicial esperaba mejores resultados para los modelos más complejos (de aprendizaje automático). Sin embargo, al tener los resultados, se comprueba que el modelo más simple, ARIMA, fue el que tuvo mejor rendimiento, mejorando la estimación actual utilizada en casi un 20%. Con ello, se comprueba que los modelos más complejos necesitan una serie con más puntos de referencia para reforzar el entrenamiento y mejorar el posterior desempeño de éste.

Finalmente, para tener la predicción completa del cálculo de la ELM se necesitará mejorar la predicción y hacer la estimación para los últimos 15 días del mes; que son los que no se alcanzan a facturar; previo a entregar la orden de compra de energía al coordinador eléctrico. Con esta predicción más la facturación de los primeros 15 días de cada mes, se tendrá el valor completo del cálculo en cuestión.

9 TRABAJOS FUTUROS

Para posteriores mejoras del modelo, se propone complementar con otras variables exógenas que puedan afectar a la serie temporal. Se considera plausible agregar variables meteorológicas, (temperatura, precipitaciones, etc.). Además, incluso se podrían incluir variables asociadas a indicadores de la economía local, incluyendo el precio del dólar. Teniendo en cuenta que la suma de regresores podría implicar un sobreajuste en el modelo, se debe realizar un *trade-off* entre la ganancia de información capturada por el modelo y el sobreajuste que esto pudiese implicar, siempre considerando que la complejidad del modelo influye en su tiempo de respuesta de entrenamiento.

Por otra parte, se considera que es posible aumentar la capacidad predictiva de los modelos incorporando más años completos de datos. Se podría realizar la predicción a un menor nivel de granularidad, como lo es a nivel mensual; con esto se esperaría disminuir el error de predicción.

Se propone también como mejora estudiar las proyecciones de consumo por grupo, identificando segmentos de clientes y realizando proyecciones diferenciadas para cada grupo. Finalmente, se propone probar con otro tipo de normalización la serie de los datos, ya que esto podría mejorar el rendimiento de los pronósticos totales.

En esta investigación se decidió, debido a criterios corporativos, que el MAE sería la métrica que decidiría qué modelo era el mejor, pero este tiene ciertas desventajas como ser circunstancial, es decir, varía a través del tiempo. En este sentido sería de interés validar mediante métricas diferentes para decidir qué modelo es el mejor para cada serie.

10. BIBLIOGRAFÍA

[1] R. P. L. V. W. Brokering, «“Los Sistemas Eléctricos de Potencia”,» Primera Edición 2008, Editorial Pearson.

[2] M. Rojas, «ABORACIÓN DE PERFILES DE DEMANDA A NIVEL DISTRIBUCIÓN. Memoria para optar al título de Ingeniero Civil Electricista,» Santiago, 2019.

[3] U. P.-S. G. y. S. P. Fayyad, «From Data Mining to Knowledge Discovery in Databases,» *ommunications of the ACM,* vol. 39, nº 11, pp. 24-26, 1996.

[4] M. y. G. N. Mackinson, «Data Mining and Knowledge Discovery in Databases - An Overview,» *Australian & New Zealand Journal of Statistics,* vol. 41, nº 3, pp. 255-275, 1999.

[5] P. CHAPMAN, «CRISP-DM 1.0: Step-by-step Data Mining Guide.,» 2000. [En línea].

[6] J. J. Pacherres Gutiérrez, «Propuesta de una metodología de extracción de conocimientos a partir de datos de las prestaciones del seguro integral de salud en la región Piura en el año 2016,» Universidad Católica Los Ángeles de Chimbote, 2018.

[7] K. M. HAN J, Data mining: concepts and techniques, United States of America: Morgan Kaufmann Publishers, 2001.

[8] G. Myatt, Making Sense of Data, New Jersey: ohn Wiley & Sons Inc., 2007.

[9] C. Chatfield, «Model Uncertainity, Data Mining and Statistical Inference,» *Journal of the Royal Statistical Society,* 1995.

[10] R. Shunway, «Time Series Analysis and Its Aplications whit R Examples,» Springer Cham , Switzerland, 2017.

[11] S. K. AHMAD, «MODELOS DE ESTIMACIÓN DE LA DEMANDA PARA EMBONOR COCA-COLA S.A,» 2017.

[12] M. A. C. Fuentes, «Proyección de series de tiempo para el consumo de la energía eléctrica a clientes residenciales en Ecuado,» 2016.

[13] G. Ríos, «Series de tiempo,» Chile , 2008.

[14] Breiman, «"Random Forests". Machine Learning,» 2001.

[15] M. R. Segal, «Machine Learning Benchmarks and Random Forest Regression,» Division of Biostatics, University of California , Abril 2003 .

[16] D. Calvo, «Diegocalvo,» 2017. [En línea]. Available: https://www.diegocalvo.es/definicion-de-red-neuronal/.

[17] R. Mendoza, «Medium,» 2019. [En línea]. Available: https://medium.com/@ricardojmv85/aplicaci%C3%B3n-del-gradiente-en-redes-neuronales-78bff0d802d5.

[18] J. López, « Análisis de Series de Tiempo. Pronóstico de demanda de uso de aeropuertos en Argentina al 2022.,» Universidad Tecnológica de Buenos Aires., Buenos Aires., 2018.

[19] Z. A. C. y. F. P. Tang, «Time series forecasting using neural networks vs Box-Jenkins methodology,» 1991.

[20] F. developer, «Facebook research,» 2021. [En línea]. Available: https://research.fb.com/prophet-forecasting-at-scale/.

[21] Minitab, «¿Qué es MAPE, MAD y MSD?,» 2018. [En línea]. Available: http://support.minitab.com/es- mx/minitab/17/topic-library/modeling-statistics/time-series/time-series-models/what-are- mape-mad-and-msd/. .

[22] M. d. C. G. Centeno, «Coeficiente de determinación,» 2020. [En línea]. Available: http://www.expansion.com/diccionario-economico/coeficiente-de-determinacion.html..

[23] J. Browniee, «Machine Learning Mastery,» Enero 2017. [En línea]. Available: https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/.

[24] O. Arzamendia, «Ciencia y datos,» mayo 2019. [En línea]. Available: https://medium.com/datos-y-ciencia/introducci%C3%B3n-al-an%C3%A1lisis-de-series-cronol%C3%B3gicas-con-python-y-pandas-99fc8d4bb56d.

[25] ENEL Distribución Chile, «ENEL Distribución Chile,» [En línea]. Available: https://www.enel.cl/es/conoce-enel/enel-distribucion-chile.html .

[26] ENEL Distribución Chile, «Memoria anual 2018, Enel Chile,» 2018.

[27] ENEL Distribución Chile, «Productos y servicios,» 2019. [En línea]. Available: https://www.enel.cl/es/empresas/productos-y-servicios.html . [Último acceso: 29 junio 2019].

[28] Generadoras de Chile, «Empresas asociadas,» 2017. [En línea]. Available: http://generadoras.cl/empresas-asociadas/enel . [Último acceso: 30 junio 2019].

[29] H. Rotther, Interviewee, *Generalidades ENEL Distribución Chile.* [Entrevista]. 20 junio 2019.

[30] J. R. Abbad, «Calidad del Servicio, Regulación y optimización de inversiones. Tesis para optar al grado de Doctor,» 1999.

[31] R. Bilinton, Distibution system reliability indexes", 1989.

[32] G. y. L. A. Seber, Linear Regression: Estimation and Distribution Theory. En: Linear Regression Analysis, New Jersey: ohn Wiley & Sons Inc., 2003.

[33] V. S. A. K. L. B. C. y. D. H. Vapnik, «Advances in Neural Information Processing Systems,» *Support Vector Regression Machines,* 1996.

[34] «Coordinador,» 20 11 2020. [En línea]. Available: https://www.coordinador.cl/nosotros/objetivos-y-funciones/.

[35] P. Development, «Prophet Forecasting at Scale,» [En línea]. Available: <https://research.fb.com/prophet-forecasting-at-scale/> .

[36] [En línea]. Available: http://support.minitab.com/es- mx/minitab/17/topic-library/modeling-statistics/time-series/time-series-models/what-are- mape-mad-and-msd/. .

11. ANEXOS

Anexo 1:

Detalle de Balance
Ventas
Facturación masivos
Refacturaciones
Devoluciones por provisión
Facturación GGCC
Venta Río Maipo (Libres de CGED)
Fact.Manual
CNF (propios + SSEE + bomberos)
Ventas Peajes
Mes 13 Actual Masivos
Mes 13 Anterior Masivos
ELM mes masivos
Mes 13 Actual GGCC
Mes 13 Anterior GGCC
ELM mes GGCC
ELM Actual - ELM Anterior
CNR´s
TOTAL VENTAS
TOTAL COMPRAS
Pérdida Física Mes Chilectra PEOD Mes
Perdidas Energía

Figura 27: Detalle del balance

Anexo 2:

Glosario variables data:

1. **Facturación clientes masivos:**

Las bases se encuentran desagregadas a nivel mensual y por sector, en formato (.accdb).

1. 'SECTOR': segmentación de acuerdo al calendario de facturación Define el día en el cual debe leerse y facturarse
2. 'NSUMINISTRO': el número del suministro, del cliente su identificador
3. 'FECULTLEC': la fecha de última vez que se leyó/facturó
4. 'TARIFA': tarifa del cliente, define el valor del KWh, entre otros
5. 'TIENEPRESPTA': indica si tiene presencia en hora de punta
6. 'CLTARIFA': clave de tarifa, una subcategorización de la tarifa
7. 'TIPDOC': tipo de documento, boleta, factura, etc
8. 'COMUNA': la comuna donde se ubica el suministro
9. 'CATEGORIA': información del "rubro" (residencial, comercial, etc)
10. 'AREA TIPICA': muy técnico, define areas similares en todo chile para unificar tarifas
11. 'CONSPROVIS': el consumo provisionado
12. 'CUOTA REEMBOLSO': si tiene reenbolsos, indica su valor
13. 'ENERGIA': energía en KWH para clientes en BT
14. 'ENERGIA BASE': energía en KWH para clientes No BT
15. 'CONSREAC'_ consumo reactivo
16. 'CONSADIC': consumo adicional (por sobre el límite de invierno)
17. 'CONSMEDDIA': consumo tramo día
18. 'CONSMEDNOC': consumo tramo noche
19. 'CONSMEDPTA': consumo tramo punta
20. 'DEMFP': demanda en tramo fuera de hora punta
21. 'DEMHP': demanda en tramo en hora punta

Todos los campos siguientes son cargos (plata) relacionado a las medidas físicas de los campos anteriores Depende de la tarifa y el área típica, entre otros

22. 'CARGO ENERGIA'
23. 'CARGO ENERGIA BASE'
24. 'CARGO ENERGADIC'
25. 'CARGO CONSDIA'
26. 'CARGO CONSNOC'
27. 'CARGO CONSPTA'
28. 'CARGO DEMFP'

29. 'CARGO DEMHP'
30. 'TOTAL CARGOS AFECTOS'
31. 'SECTOR-ZONA’: concatenación entre el sector que vimos más arriba y la zona, un lugar geográfico segmentado internamente para optimización de rutas de lectura

2. **Base de información de clientes (muestra) SMT:**

Base en formato (xlsx), cuenta la información de una muestra de 1000 clientes que cuentan con medidores inteligentes La base cuenta con las siguientes variables:

1. nro_suministro: el número del suministro, del cliente su identificador
2. nro_medidor: el ID del medidor
3. fecha_instalacion: fecha de instalación del medidor inteligente
4. comuna: comuna donde se encuentra instalado el medidor
5. tarifa: tarifa asociada al cliente
6. procedimiento_lectura: si corresponde a telemedida (SMT) o pedestre
7. tip_documento: si corresponde a un pago con boleta o factura

3. **Base de consumo de clientes (muestra) SMT:**

Base en formato (csv), cuenta con la información asociada al consumo (Kw) del cliente Se presenta a dos niveles de granularidad, mensual y diario

1. 'nro_aparato': el ID del medidor
2. 'fecha_consumo’: fecha y hora del registro de consumo
3. 'tip_medida': que tipo de medida es (energía, energía por tramos, etc)
4. 'consumo': consumo medido en dicha fecha y hora

Anexo 3:

Gráfica de autocorrelación para determinar estacionariedad en la serie.

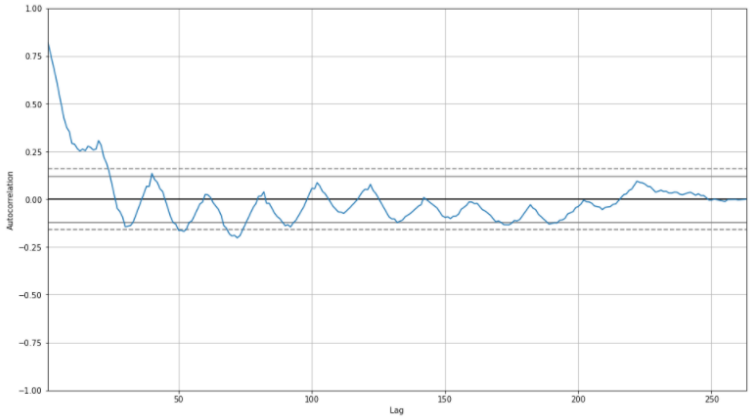
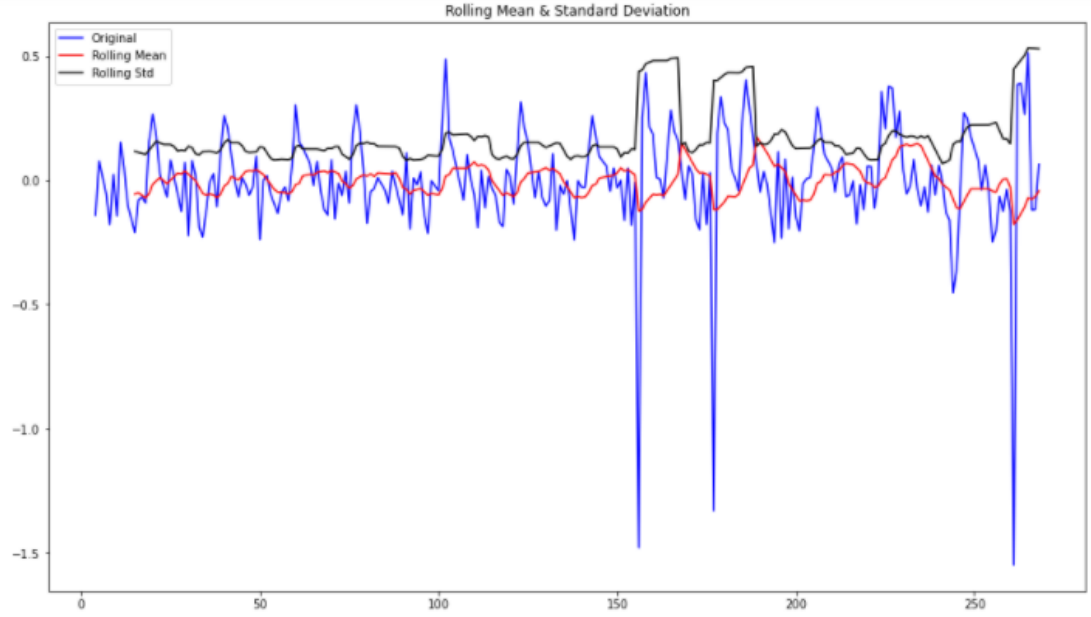


Figura 28: Gráfica de autocorrelación para determinar estacionariedad en la serie

Prueba estacionariedad de la serie




```
Results of Dickey-Fuller Test:
Test Statistic                -1.368766e+01
p-value                        1.362690e-25
#Lags Used                     0.000000e+00
Number of Observations Used    2.580000e+02
Critical Value (1%)           -3.455953e+00
Critical Value (5%)           -2.872809e+00
Critical Value (10%)          -2.572775e+00
dtype: float64
```

Figura 29: Prueba estacionariedad de la serie

Anexo 4:

Tabla 15: Feriados utilizados modelo Prophet

Año Nuevo
Semana Santa (Viernes Santo)
Semana Santa (Sábado Santo)
Día de Pascuas
Día Nacional del Trabajo
Día de las Glorias Navales
San Pedro y San Pablo
Virgen del Carmen
Asunción de la Virgen
Día de la Independencia
Día de las Glorias del Ejército
Fiestas Patrias
Día del Respeto a la Diversidad
Día Nacional de las Iglesias Evangélicas
Día de Todos los Santos
La Inmaculada Concepción
Navidad
Día del Descubrimiento de dos Mundos