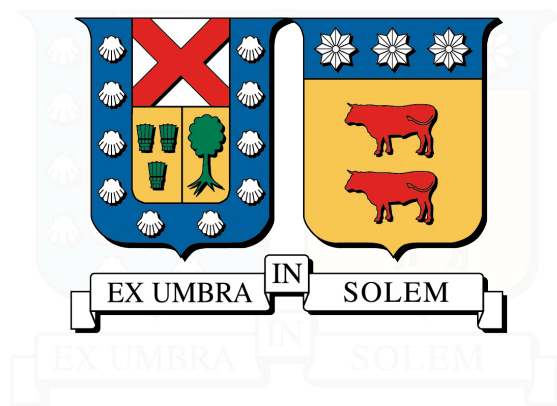


UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INDUSTRIAS
SANTIAGO - CHILE



**ESTUDIO Y ANÁLISIS DEL IMPACTO DEL VIRUS COVID-19 EN LA
DEMANDA ELÉCTRICA DE LA REGIÓN METROPOLITANA USANDO
HERRAMIENTAS DE MACHINE LEARNING**

FELIPE PÁEZ RUIZ

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL INDUSTRIAL

PROFESOR GUÍA : WERNER KRISTJANPOLLER
PROFESORA CORREFERENTE : MARÍA FARÍAS

JULIO 2022

RESUMEN EJECUTIVO

En el presente trabajo se busca realizar un análisis del consumo eléctrico de la Región Metropolitana a través de su modelamiento matemático. Esto surgió debido a la pandemia del virus COVID-19, el cual generó cambios culturales a corto y largo plazo, siendo el consumo eléctrico uno de ellos, por lo que el objetivo principal es estudiar el impacto del mismo sobre 19 comunas.

Para que el estudio fuese factible, se creó una base de datos a partir del repositorio de balances físicos de la entidad Coordinador Eléctrico, junto a la adición de distintas variables que intentarán explicar este consumo. Además, el modelamiento se realizó mayoritariamente usando herramientas y modelos de Machine Learning, ya que estos tienen la característica de ser adaptables a muchas bases de datos debido a su refinada detección de patrones, teniendo cuidado de no caer en sobreajustes ni subajustes.

El análisis final consideró: 7 regresiones de distintas librerías, tanto lineales como no lineales, 11 variables de estudio, y 4 escalamientos distintos para la base de datos inicial, resultando en un total de 532 modelos distintos.

Índice de Contenidos

1. Problema de Investigación	1
2. Objetivos	2
2.1. Objetivo Principal	2
2.2. Objetivos Específicos	2
3. Marco Teórico	3
3.1. COVID-19	3
3.1.1. El virus en Chile	3
3.1.2. La contracción económica	4
3.1.3. Cambios en el consumo eléctrico	4
3.2. Comunas de la región Metropolitana	5
3.3. Generación Eléctrica en Chile	5
3.3.1. Mercado Eléctrico	6
3.3.2. Tipos de Clientes	7
3.4. Machine Learning	7
3.4.1. Supervisión Humana	8
3.4.1.1. Support Vector Machines	9
3.4.1.2. Ridge Regression	10
3.4.1.3. Lasso Regression	11
3.4.1.4. Random Forest	12
3.4.2. Aprendizaje “on the fly”	13
3.4.3. Análisis de Datos y Patrones	14
4. Metodología	15
4.1. Software	15
4.1.1. Librerías	15
4.2. Base de Datos	15
4.2.1. Formación de la BBDD para el consumo eléctrico	16
4.2.2. Elección de variables	17
4.2.3. Preprocesamiento	19
4.3. Elección de Modelo	23
5. Resultados	25
5.1. Entrenamiento	25
5.1.1. Random Forest	25
5.1.2. Support Vector Machine: Radial Basis Function kernel	26
5.1.3. Support Vector Machine: linear kernel	27
5.1.4. OLS (A): biblioteca sklearn	28
5.1.5. OLS (B): biblioteca statsmodel.api	29
5.1.6. Regresión Ridge	30

5.1.7. Regresión Lasso	31
5.1.8. Resultados Entrenamiento	32
5.2. Prueba	33
5.2.1. Resultados Prueba	33
5.3. Comparación Final	33
6. Discusión y Análisis	35
6.1. Alhué	35
6.2. Buin	37
6.3. Cerro Navia	38
6.4. Colina	40
6.5. El Monte	41
6.6. La Pintana	42
6.7. Lo Prado	44
6.8. Melipilla	45
6.9. Paine	47
6.10. Pirque	48
6.11. Pudahuel	49
6.12. Puente Alto	50
6.13. Renca	52
6.14. San Bernardo	53
6.15. San José de Maipo	54
6.16. San Pedro	56
6.17. Santiago	57
6.18. Talagante	59
6.19. Til Til	60
7. Conclusiones	63
Bibliografía	65
A. Anexos	67
A.1. Matrices de Correlación	67
A.2. Gráficos Consumo Eléctrico v/s Renta Mujer y Población	75
A.3. Diferencia entre escalados	84

Índice de Tablas

4.1. Tabla subestaciones Transelec	16
4.2. Consumo eléctrico mensual, en [KWh], por comuna. Años 2018-2020	17
4.3. Valores nulos de Consumo Eléctrico	20
4.4. Comparación R2 y RMSE de los distintos modelos	24
5.1. RMSE de la Regresión Random Forest para cada comuna	26
5.2. RMSE de SVR: “rbf” para cada comuna	27
5.3. RMSE de SVR: “linear” para cada comuna	28
5.4. RMSE de OLS (sklearn) para cada comuna	29
5.5. Diferencia de escalados stdr y minmax, para Alhué	29
5.6. Representación de Data Frame de Alhué utilizado en la biblioteca statsmodel.api	30
5.7. RMSE de OLS (statsmodel.api) para cada comuna	30
5.8. RMSE de regresión Ridge para cada comuna	31
5.9. RMSE de regresión Lasso para cada comuna	32
5.10. RMSE promedios, de cada modelo, para los distintos escalados	32
5.11. RMSE de prueba promedios de cada modelo, para todas las comunas	33
5.12. Comparación final de RMSE promedios, de entrenamiento y prueba.}	33
A.1. Diferencia de escalados MinMax y Robust, para Alhué	84
A.2. Diferencia de escalados Robust y MaxAbs, para Alhué	85

Índice de Figuras

3.1. Avance acumulativo de vacunación para primera dosis	4
3.2. Participación relativa por fuente de generación	6
3.3. Aprendizaje Supervisado	8
3.4. Aprendizaje No Supervisado	9
3.5. Regresión SVM, kernel “linear”	10
3.6. Regresión SVM, kernel “poly”	10
3.7. Regularización Ridge para regresión lineal y polinomial	11
3.8. Regularización Lasso para regresión lineal y polinomial	11
3.9. Árbol de decisión de profundidad 2	12
3.10. Comportamiento del árbol de decisión en función de su profundidad	13
3.11. Automatización de Batch Learning	13
4.1. Matriz de Correlación de Alhué	21
4.2. Matriz de Correlación de Buin	22
4.3. Consumo Alhué v/s Renta M/h y Pob Alhué	22
4.4. Consumo Buin v/s Renta M/h y Pob Buin	23
6.1. Importancia Variables Alhué, del modelo Random Forest	36
6.2. Importancia Variables Buin, del modelo Random Forest	37
6.3. Importancia Variables Cerro Navia, del modelo Random Forest	39
6.4. Importancia Variables Colina, del modelo Random Forest	40
6.5. Importancia Variables El Monte, del modelo Random Forest	42
6.6. Importancia Variables La Pintana, del modelo Random Forest	43
6.7. Importancia Variables Lo Prado, del modelo Random Forest	44
6.8. Importancia Variables Melipilla, del modelo Random Forest	46
6.9. Importancia Variables Paine, del modelo Random Forest	47
6.10. Importancia Variables Pirque, del modelo Random Forest	48
6.11. Importancia Variables Pudahuel, del modelo Random Forest	50
6.12. Importancia Variables Puente Alto, del modelo Random Forest	51
6.13. Importancia Variables Renca, del modelo Random Forest	52
6.14. Importancia Variables San Bernardo, del modelo Random Forest	54
6.15. Importancia Variables San José de Maipo, del modelo Random Forest	55
6.16. Importancia Variables San Pedro, del modelo Random Forest	57
6.17. Importancia Variables Santiago, del modelo Random Forest	58
6.18. Importancia Variables Talagante, del modelo Random Forest	60
6.19. Importancia Variables Til Til, del modelo Random Forest	61
A.1. Matriz de Correlación de Alhué	67
A.2. Matriz de Correlación de Buin	67
A.3. Matriz de Correlación de Cerro Navia	68
A.4. Matriz de Correlación de Colina	68
A.5. Matriz de Correlación de El Monte	68

A.6. Matriz de Correlación de La Pintana	69
A.7. Matriz de Correlación de Lo Prado	69
A.8. Matriz de Correlación de Melipilla	69
A.9. Matriz de Correlación de Paine	70
A.10. Matriz de Correlación de Pirque	70
A.11. Matriz de Correlación de Pudahuel	70
A.12. Matriz de Correlación de Puente Alto	71
A.13. Matriz de Correlación de Renca	71
A.14. Matriz de Correlación de San Bernardo	71
A.15. Matriz de Correlación de San José de Maipo	72
A.16. Matriz de Correlación de San Pedro	72
A.17. Matriz de Correlación de Santiago	72
A.18. Matriz de Correlación de Talagante	73
A.19. Matriz de Correlación de Til Til	73
A.20. Consumo v/s Renta M/h y Pob Alhue	75
A.21. Consumo v/s Renta M/h y Pob Buin	75
A.22. Consumo v/s Renta M/h y Pob Cerro Navia	76
A.23. Consumo v/s Renta M/h y Pob Colina	76
A.24. Consumo v/s Renta M/h y Pob El Monte	77
A.25. Consumo v/s Renta M/h y Pob La Pintana	77
A.26. Consumo v/s Renta M/h y Pob Pirque	78
A.27. Consumo v/s Renta M/h y Pob Buin	78
A.28. Consumo v/s Renta M/h y Pob Alhue	79
A.29. Consumo v/s Renta M/h y Pob Buin	79
A.30. Consumo v/s Renta M/h y Pob Pudahuel	80
A.31. Consumo v/s Renta M/h y Pob Puente Alto	80
A.32. Consumo v/s Renta M/h y Pob Renca	81
A.33. Consumo v/s Renta M/h y Pob San Bernardo	81
A.34. Consumo v/s Renta M/h y Pob San José de Maipo	82
A.35. Consumo v/s Renta M/h y Pob San Pedro	82
A.36. Consumo v/s Renta M/h y Pob Santiago	83
A.37. Consumo v/s Renta M/h y Pob Buin	83
A.38. Consumo v/s Renta M/h y Pob TilTil	84

1 | Problema de Investigación

El flujo constante de energía eléctrica en todas partes del mundo es fundamental y necesario, la demanda no puede superar a la oferta o de otra forma, básicamente el mundo no funcionaría. El virus COVID-19 ha jugado un papel fundamental en muchos aspectos desde que se originó y fue declarado pandemia por la Organización Mundial de la Salud WHO (2020). Esta emergencia sanitaria trajo como consecuencias contracciones en la economía, cuarentenas que reducían la movilidad de las personas y el inminente cierre de muchas PYME.

Debido a esto, se debió poner especial énfasis en la industria energética, ya que la mayoría de las empresas no tenían otra opción que adaptarse a una modalidad de teletrabajo, lo que vendrá de la mano de tener un buen servicio de internet en casa, como también un buen equipo, y claro, una mayor necesidad de estar siempre conectado al sistema eléctrico.

Esta problemática fue estudiada en EE. UU. (Agdas y Barooah, 2020), donde se pudo concluir que el efecto de la pandemia sobre esta demanda no fue una simple reducción, y existían notables diferencias entre los estados. En el caso chileno, la primera cuarentena total fue declarada el 15 de mayo del 2020 (MINSAL, 2020b) y para ese entonces, ya hubo una disminución considerable en el consumo de energía, como en el sector de Administración de Edificios y Condominios, que se redujo en un 40 %, Transporte, Almacenamiento y Comunicaciones disminuyó en un 33 %, entre otros ejemplos más (ISCI et al., 2020). En otros países como en Argentina, que en todo el país cayó en un 5.5 % y 2.2 % en los primeros dos trimestres del año 2020 (Paramo, 2020), o en Perú, que en todo el año 2020, la venta de energía disminuyó un 8.5 % para la distribuidora Enel, y en Luz del Sur un 9.42 % (BNamericas, 2021).

Es necesario dejar en claro que, si bien todos estos datos estadísticos tienen un patrón a la baja, se está considerando a un nivel muy agregado. Por esto, es posible que, si se separa en los clientes residenciales, comerciales e industriales, se vean otras tendencias debido a que, por ejemplo, los residenciales probablemente estén consumiendo más energía ahora que antes debido al confinamiento. Es por esto que surge la necesidad de hacer un análisis del impacto que esta pandemia ha tenido sobre este sector tan necesario en este siglo. Este estudio se centrará en las comunas de la Región metropolitana.

Se debe tener en consideración que existe una brecha socioeconómica bastante alta en algunas comunas de la capital, además de que algunas son consideradas sectores industriales y tienen un número considerable de fábricas, por lo que se tendrá que segregar cuidadosamente el tipo de cliente y comuna. A raíz de esto nacen las preguntas ¿Qué relación existe entre el nivel socioeconómico de una comuna y su consumo de energía eléctrica?, ¿Cuál fue el verdadero impacto de la pandemia en la demanda eléctrica de la Región Metropolitana?, ¿Este patrón volverá a ser como antes una vez se acabe el confinamiento?

Para esto se necesitará realizar un levantamiento de información considerable, con ellos el consumo eléctrico histórico de todas las comunas de Santiago, además de los datos socioeconómicos de cada comuna, localización de empresas, segregación de tipo de cliente, dentro de otras características críticas para finalmente ingresar estos datos a un modelo de Machine Learning para poder establecer un impacto numérico y porcentual de la demanda eléctrica, como también concluir la relación existente entre las características demográficas de cada comuna y su consumo eléctrico.

2 | Objetivos

2.1. Objetivo Principal

Determinar los impactos del COVID-19 en el consumo eléctrico de la Región Metropolitana a través de modelos relacionales de Machine Learning para poder identificar zonas de mayor dependencia eléctrica.

2.2. Objetivos Específicos

Analizar los consumos eléctricos históricos antes de la pandemia de todas las comunas de la RM, para contrastarlos con los de después de COVID-19.

Analizar, identificar y clasificar los distintos tipos de clientes que existen en cada comuna para poder caracterizarlos apropiadamente, focalizándose mayoritariamente en residenciales y comerciales.

Realizar un modelo con Machine Learning que contenga las características de cada comuna para comprobar la relación existente entre consumo y nivel socioeconómico/características demográficas.

Realizar una predicción de cómo se comportará la demanda en un futuro de corto y largo plazo, en función de los datos propuestos, la tendencia entregada desde el modelo y los casos activos de COVID-19 en la RM.

3 | Marco Teórico

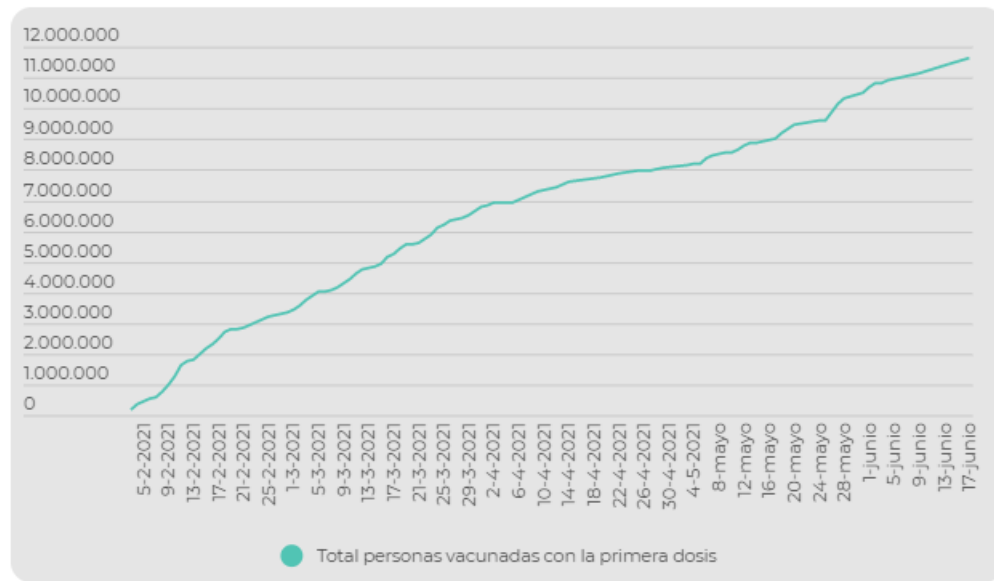
3.1. COVID-19

A mediados de diciembre de 2019 en la ciudad de Wuhan, China, se detectaron los primeros casos del virus respiratorio SARS-CoV-2, propagándose de una manera exponencial a una escala global, declarándose pandemia mundial por la Organización Mundial de la Salud el día 11 de marzo de 2020. Uno de los países más afectados por el virus fue Italia, siendo uno de los países con cifras más altas de casos confirmados y muertes debidas al virus. Con esto, las primeras medidas de distanciamiento social, cuarentena obligatoria, permisos de movilización, entre otras, comenzaron a implementarse, y como es de esperarse, el estilo de vida cambió radicalmente, lo que afectó muy fuerte de manera psicológica a los habitantes del país. En la encuesta realizada por Casagrande et al. (2020) se pudo observar que el 57,1 % de los participantes presentaban problemas para dormir, un 32,1 % declararon tener ansiedad alta, 41,8 % sentían una angustia extrema y otro 7,6 % presentaron síntomas de Trastorno de Estrés Postraumático. Todo esto al final del día, representa el gran shock cultural que significó una pandemia mundial tan repentina, que no demoraría en aterrizar en Chile, generando el mismo cambio e impacto en la población que se vivió en Italia: cuarentenas totales en muchas regiones, teniendo a la Región Metropolitana con todas sus comunas en confinamiento, lo que forzó el teletrabajo y clases en línea. Esto eventualmente generaría un fuerte cambio en el consumo eléctrico general del país.

3.1.1. El virus en Chile

El primer caso en Chile se reportó el 3 de marzo de 2020 (MINSAL, 2020a), tratándose de una persona de 33 años que se encontraba viajando los meses anteriores en países asiáticos. El Ministerio de Salud, muy tardíamente, comenzó a generar planes de retención y posibles cuarentenas para tratar de evitar la inminente expansión del virus. Al día 18 de Junio del año 2021, se han detectado un total de 1.505.001 casos, con un total de 31.259 muertes (MINSAL, 2021).

Dada esta contingencia, fue completamente necesario vacunar a la mayor cantidad de la población posible, partiendo por la más vulnerable, es decir adultos mayores y personas con enfermedades crónicas que generaran conflictos mayores al estar en contacto con el virus. El primer proceso de vacunación comenzó la semana del 5 de febrero de 2021 los efectos de la misma aún no se han visto directamente, teniendo aproximadamente un 60 % de la población ya vacunada con la primera dosis [Ver Figura 3.1] .

Figura 3.1: Avance acumulativo de vacunación para primera dosis

Fuente: Monasterio (2021)

3.1.2. La contracción económica

Debido a todo lo mencionado anteriormente, la economía mundial, y la de cada país, se vio afectada debido a todas las restricciones necesarias para poder abarcar esta pandemia. En el caso de Chile fue peor en este ámbito debido a que meses antes de esta pandemia, se vio afectado por el estallido social, en octubre de 2019, generando un shock económico, sumando después ser afectado por la crisis sanitaria local y finalmente el impacto la crisis sanitaria global. Estos tres shocks, que ocurrieron en un rango de 15 meses, no tenían ningún precedente y fueron completamente inéditos ya que su origen no era económico, y para hacer frente a esto se requirió de una ayuda fiscal y monetaria, por lo que el Banco Central de Chile tomó medidas dentro de estas cuatro áreas: Política Monetaria Convencional, Política Monetaria No Convencional, Recuperación del Espacio de Política y Estabilización de mercados. (Banco Central de Chile, 2020).

3.1.3. Cambios en el consumo eléctrico

Como es de esperarse, además pasar por esta contracción anteriormente mencionada, se establecieron nuevos estilos de vida dentro de la población, y el consumo eléctrico fue uno de estos cambios. La pandemia trajo consigo muchas preocupaciones que antes se veían como un bien no tan esencial, como tener contratado un buen plan de internet, o tener en casa un buen equipo de trabajo. Ahora, éstas dos comodidades pasaron a ser necesidades ya que ninguna empresa se encontraba preparada en términos de infraestructura para poder soportar el aforo exigido por las autoridades sanitarias, por lo que la necesidad de estar siempre conectado a la red se hizo imperante. Es por esto que el consumo eléctrico residencial podría haber aumentado radicalmente, pero éste pudo haber representado solo un cambio marginal dentro del consumo nacional, ya que en Chile existen muchos tipos de clientes y tarifas, que se presentarán más adelante, y que podrían pesar mucho más en términos de Megawatts consumidos, dado su tamaño y rubro. Es por esto que el análisis se debe centrar en el poder de adaptación que cada empresa tuvo a inicios de pandemia, y qué tan crítico para la sociedad es que dejasen de funcionar, ya que depende enormemente de lo que digan las autoridades en relación a que si puede seguir funcionando dentro de las nuevas normas sanitarias. Pero dejando de lado este límite, y asumiendo que la mayoría de las empresas grandes sí pudieron adaptarse a este cambio, es un hecho que el consumo eléctrico de éstas al inicio de la crisis, bajó considerablemente.

3.2. Comunas de la región Metropolitana

La Región Metropolitana cuenta con un total de 52 comunas a lo largo de su territorio, y la mayoría de estas se consideran zonas urbanas (precisamente 32 de ellas), y el resto son rurales. Ahora, para la provincia de Santiago se consideran solo las urbanas y dentro de ellas existe una diferencia considerable en los ingresos municipales. En el informe presentado por el Senado de la República de Chile (2017), se establecen 3 indicadores para poder medir el ingreso comunal. El primero son los Ingresos Propios, que a su vez está compuesto por dos factores:

- **Ingresos Propios Permanentes per Cápita (IPP):** está compuesto por distintas formas que todos los municipios tienen para autofinanciarse, como el impuesto territorial a beneficio municipal, los permisos de circulación de beneficio municipal, las patentes de beneficio municipal, las rentas de la propiedad, etc. El segundo factor corresponde al Fondo Común Municipal, que según Dazarola Leichtle (2020) corresponde un mecanismo de financiamiento de parte del Estado a todas las municipalidades del país, el cual es de suma importancia para las comunas no muy grandes con IPP muy bajos.
- **Otros Ingresos:** Corresponden a las transferencias corrientes, las transferencias para gastos de capital, la venta de activos no financieros y otros ingresos municipales.

De este estudio, se pudo diferenciar cuales son las comunas con más ingresos municipales totales, sumando ambos indicadores. De aquí se obtuvo que las comunas de Vitacura, Lo Barnechea, Las Condes y Providencia, son las que tienen más recursos en este sentido, y las con menos son Pedro Aguirrez Cerda, La Pintana, Peñaflores y Puente Alto.

La diferencia entre la comuna con más y menos ingresos, que vendría siendo Las Condes y Pedro Aguirrez Cerda, es del orden de \$ 664.000.000, lo que deja claramente evidenciado que existe una brecha enorme entre los habitantes de las comunas, donde existen preocupaciones, precios y estilos de vida completamente distintos dentro de una misma región. Todo esto se vio completamente afectado por los shocks mencionados en los apartados anteriores, forzando a cambiar la rutina diaria, independientemente de la comuna en que se haya estado viviendo.

3.3. Generación Eléctrica en Chile

La generación de energía eléctrica de un país es su capacidad para poder transformar un recurso, o energía, como por ejemplo solar, química, cinética, en energía eléctrica que es distribuida a lo largo del territorio. Existen dos tipos de generación relacionadas a su origen, siendo estas renovables y no renovables. El primer término indica que estas fuentes corresponden a recursos naturales que se consideran inagotables o autosustentables hasta cierto punto, como el viento, mar, sol, biomasa, entre otras, y el segundo término indica lo contrario, fuentes limitadas a lo largo del tiempo, de los cuales en su mayoría corresponden a combustibles fósiles como el gas natural, carbón o petróleo.

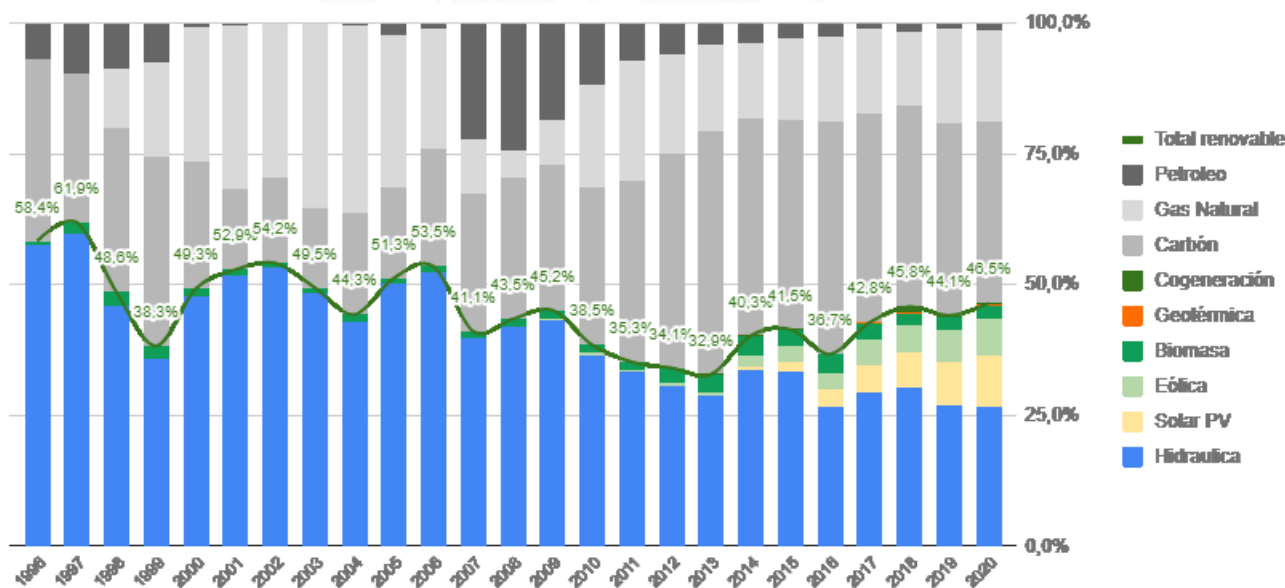
En cada país es esencial que exista una constante generación y distribución de la misma. En general, el suministro pasa por distintos procesos para poder llegar al cliente que necesita de ésta. La demanda es captada por el sistema de generación, y dependiendo de en dónde se encuentre, se generará energía en base de alguna de las fuentes anteriormente mencionadas, pasando por los sistemas de transmisión y distribución antes de que llegue directamente a quien lo haya necesitado. En este extenso y rápido proceso, para que no existan pérdidas mayores, se encuentran distribuidas, a lo largo de la línea, subestaciones donde se eleva la tensión, lo que ayuda a que recorra largas distancias. Habiendo viajado todos estos kilómetros es que el sistema de distribución comienza a tomar las decisiones claves de quién y cuánto está solicitando utilizar, lo que variará dependiendo del tipo de cliente que se encontraba realizando la solicitud.

Gracias al vasto y diverso territorio chileno, existen una cantidad bastante alta de centrales generadoras térmicas y renovables. Al año 2019, el Sistema Eléctrico Nacional, que es el que abastece a la Región Metropolitana, tiene una capacidad instalada de 25.248 MW (refiriéndose a la potencia de generar energía en base a la tecnología disponible). Además, en su matriz energética, cuenta con un 48,3 % de fuentes renovables,

las cuales se descomponen en 27 % hidráulica, 20,8 % solar, 8,6 % eólico, 1,8 % biomasa, y 0,2 % geotérmica, y el resto de fuentes térmicas, las cuales a su vez corresponden a un 21,2 % carbón, 19,2 % gas natural, y 11,4 % petróleo (Generadoras de Chile, 2020)

En la Figura 3.2 se puede apreciar la participación de cada fuente a lo largo de los años en todo el país, desde 1996 hasta 2020, y puede verse claramente la tendencia al alza de las fuentes renovables y la inclusión de Biomasa, Eólica y Solar a la matriz energética, desde el año 2014. Desde hace muchos años se han iniciado procesos para descarbonizar la generación eléctrica en el país, y disminuir así los gases y desechos tóxicos que se producen en su generación por lo que esta tendencia se espera que siga al alza.

Figura 3.2: Participación relativa por fuente de generación



Fuente: Generadoras de Chile (2020)

3.3.1. Mercado Eléctrico

El mercado eléctrico tiene una parte de demanda y otra de oferta. La primera corresponde a los clientes, que serán detallados en una sección más adelante debido a su importancia y diferenciación entre cada uno. Ahora, la oferta está compuesta por empresas privadas en su totalidad, y el rol del Estado solamente recae en regular, fiscalizar y realizar una planificación a modo de recomendación para estas empresas. Dentro de esto, existen tres etapas por las que se deben pasar para que lo generado pueda llegar a mano de los clientes, estos son: Generación, Transmisión y Distribución:

- **Generación:** este segmento produce la energía eléctrica propiamente tal, a través de las centrales instaladas a lo largo de Chile, sin importar de la fuente de la que provengan. Se caracteriza por tener un mercado muy competitivo, donde venderá el que pueda lograr tener un costo marginal menor, logrando tener la mejor economía de escala. Aquí se separa entre 3 grandes sistemas interconectados: Sistema Eléctrico Nacional (SEN), Sistema de Aysén (SEA) y el Sistema de Magallanes (SEM). Cada uno tiene una capacidad instalada distinta y, dado que abastecen a lugares distintos del país, cubren distintas demandas y cada uno de ellos posee una distribución porcentual distintas de las fuentes de generación.
- **Transmisión:** corresponde al conjunto de equipos, subestaciones y líneas que se encuentran destinadas al transporte de la energía desde las estaciones de generación hasta los centros de distribución, o directamente a centros de consumo. En Chile, se considera como Transmisión a todas las líneas o subestaciones que tienen una tensión superior a los 23.000 V (Arias Quezada, 2009). El sistema con el que se transporta la electricidad se categoriza en transmisión troncal, el cual está configurado

únicamente en la oferta del mercado, y subtransmisión de servicio público eléctrico, que permite retirar la energía desde el sistema troncal a los consumidores locales.

- **Distribución:** estos sistemas están constituidos por los mismos elementos que la Transmisión, con la diferencia que es exclusivamente para los clientes finales, lo cuales se ubican en zonas urbanas y rurales específicas, que se encuentran debidamente limitadas.

3.3.2. Tipos de Clientes

Como se mencionó antes, los clientes del mercado eléctrico tienen características en relación a las tarifas que se les cobran, existiendo dos tipos de clientes: libres y regulados. Se entiende por libres a los clientes quienes tienen un consumo eléctrico elevado y la ley les atribuye la capacidad de poder negociar las tarifas que su suministrador, ya sea de generación o distribución, le está cobrando, y los clientes regulados no tienen esta cualidad, por lo que deben acatar las tarifas impuestas por la ley.

Según lo estipulado en el Artículo 147 de la Ley General de Servicios Eléctricos (2007), los clientes regulados, o que están forzados a pagar tarifas, son los que cuentan con una potencia conectada de hasta 5.000 kW, y los que no se encuentran forzados, clientes libres, aquellos que tienen una potencia mayor a la estipulada. No obstante, los clientes que tienen una potencia menor a 5000 kW y mayor a 500kW, pueden optar por pagar una tarifa libre si así lo desean, pero deben permanecer en esta condición por al menos 4 años. Además, este cambio debe ser comunicado con la debida anterioridad de 12 meses a la distribuidora respectiva.

La entidad que se encargará de organizar y planificar las tarifas que se le cobrarán a cada uno de los clientes es la Comisión Nacional de Energía, que está regulada por el Ministerio de Energía. Como se dijo antes, los clientes regulados deben asumir estas tarifas, bajo las condiciones establecidas por la misma ley y que se diferencian mayormente a la línea de tensión en la que se encuentran conectados. Dicho esto, podrán escoger libremente entre cada una de las siguientes tarifas: BT1a, BT1b, TRBT2, TRBT3, TRAT1, TRAT2, TRAT3, BT2, BT3, BT4, BT5 y AT2.

3.4. Machine Learning

Machine Learning es una de las ramas de la inteligencia artificial y método de análisis de datos que automatiza la construcción de modelos. Esta herramienta es sumamente útil al momento de tratar de modelar problemas muy complejos, o de Big Data, que con modelos econométricos tradicionales resultarían en una extensa lista de restricciones que no se podrían escalar al momento de entregarles bases de datos que se comporten de manera muy diferida a la inicial, como también para los escenarios donde el ambiente en el que se encuentran fluctúa con frecuencia, dejando inutilizables modelos que no se puedan adaptar a dichos cambios. Una definición más general podría ser la siguiente:

“Machine Learning es la ciencia y arte de programar computadoras para que puedan aprender de los mismos datos” (Géron, 2019b)

Como es de esperarse, existen muchos tipos de sistemas/estructuras de Machine Learning, por lo que resulta útil clasificarlos en 3 categorías propuestas por Géron (2019b), que no son exclusivas para cada modelo, i.e un mismo sistema puede pertenecer a más de alguna de estas clasificaciones:

- Si se encuentran supervisados por humanos al momento de realizar el entrenamiento de los datos.
- Si pueden aprender a medida que se le están ingresando datos nuevos.
- Si funcionan meramente de manera comparativa o si detecta patrones en los datos de entrenamiento para poder aplicar modelos predictivos.

3.4.1. Supervisión Humana

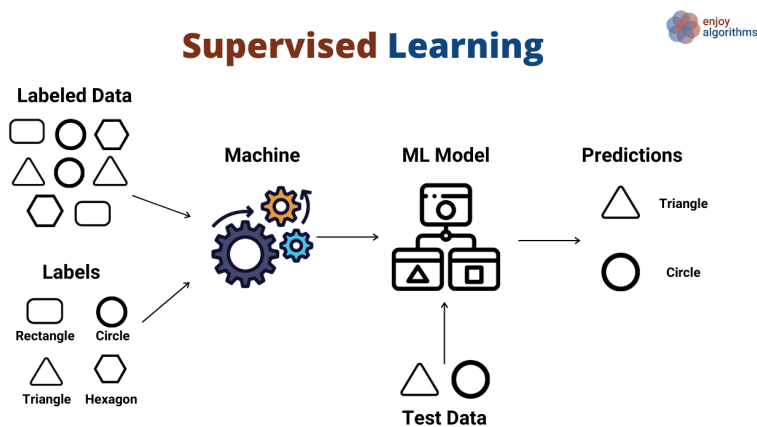
Para la primera clasificación existen 4 tipos de sistemas: Supervisados, No Supervisados, Semisupervisado y Aprendizaje Reforzado.

En el Aprendizaje Supervisado, como lo dice el nombre, está intervenido por humanos y no se le deja el trabajo 100 % a la máquina con la que se está trabajando, esto se ve al momento de entrenar los datos, debido a que una persona es la que le dice al algoritmo qué cosas está clasificando y lo que éstas significan a través de etiquetas [Ver Figura 3.3].

Los algoritmos más conocidos de este tipo de aprendizaje son:

- Vecinos más Cercanos (k-NN)
- Redes Neuronales (NN)
- Máquinas de Vectores de Soporte (SVM)

Figura 3.3: Aprendizaje Supervisado



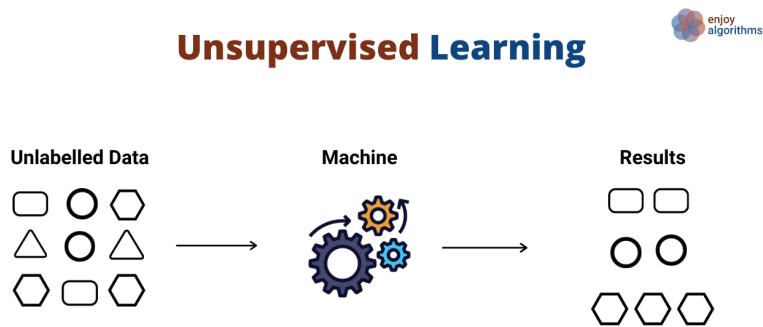
Fuente: Raj (2020)

Ahora, el aprendizaje No Supervisado se caracteriza por la ausencia de las etiquetas anteriormente mencionadas, básicamente intenta aprender de los datos sin ningún tipo de referencia [Ver Figura 3.4]. Este intenta buscar similitudes entre los datos para tratar de clasificarlos.

Los algoritmos más conocidos son:

- K-medias (K-means)
- Análisis de Componentes Principales (PCA)
- Agrupamiento Jerárquico (HCA)

Figura 3.4: Aprendizaje No Supervisado



Fuente: Raj (2020)

Ahora, al aprendizaje Semi Supervisado es una combinación de las dos clasificaciones anteriores, en donde hay datos que están debidamente etiquetados como otros en los que el algoritmo debe aprender a clasificarlos. En general estos modelos utilizan metodologías secuenciales para tratar todos los datos, y cuando ya se entrenó, la gran mayoría de ellos logran tener una etiqueta propia debido a esta naturaleza mixta. Un algoritmo que utiliza este método es la Red de Creencia profunda (DBNs).

Por último, está el Aprendizaje Reforzado, que es bastante distinto a los mencionados previamente debido a que funciona con un sistema de prueba y error, donde por cada error se va actualizando el modelo con “recompensas” (o al contrario, “penalizaciones”) y a raíz de esta práctica, aprenderá por si solo cuales son las mejores estrategias para el correcto procedimiento de lo que se necesita hacer. En general este algoritmo es muy utilizado en la robótica, en la que se le asigna este sistema y a través de muchas iteraciones logra funcionar correctamente.

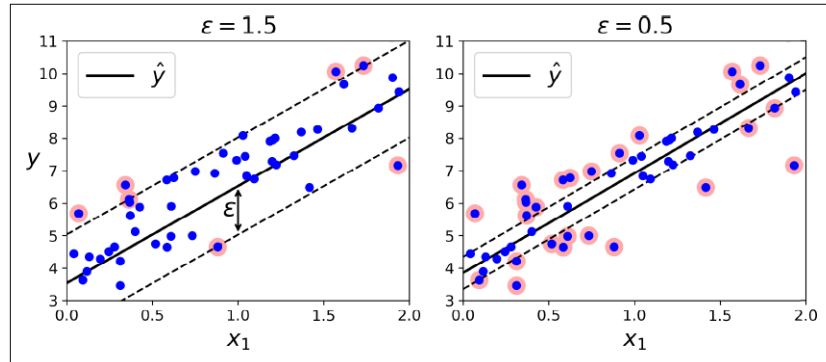
3.4.1.1. Support Vector Machines

Originalmente, los modelos de Support Vector Machines (SVM) son una generalización de un algoritmo de clasificación llamado “maximal margin classifier”, el cual intenta separar la base de datos por métodos meramente lineales, lo que funcionaría en la mitad de los casos ya que las variables de estudio no siempre tendrán comportamiento, o límites, que puedan separarse con líneas o hiperplanos. SVM cubre esta problemática utilizando distintos “kernel”. Este último concepto es muy amplio, por lo que por simplicidad, el kernel es un método que permite realizar separaciones lineales a datos que no se encuentran distribuidos linealmente, gracias al aumento de la dimensionalidad de los datos.

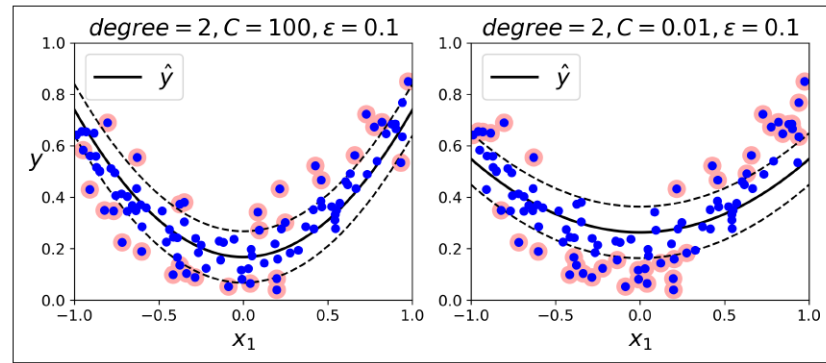
Ahora, su utilidad no se queda en clasificación, sino que también puede ser aplicado a la detección de datos “outlayers” y regresiones. Debido a la naturaleza de este estudio, en este apartado se referirá únicamente a este último tipo de SVM.

Este método de regresión se basa en el principio de intentar realizar la mayor cantidad de ajustes sujetos a lo que se denomina “la calle de SVM”, la cual gráficamente son 3 líneas: una central, que representa la regresión, y dos laterales, generando una especie de margen que será controlado por hiperparámetros. En la Figura 3.5, se puede observar la regresión junto a estos márgenes, generando esta figura en forma de calle. También se puede observar el parámetro ϵ que es proporcional al tamaño del margen.

Ahora, como ya se mencionó, SVM cubre casos en los que los datos no presentan tendencias lineales, y aquí es donde el kernel tiene un rol importante, ya que dependiendo de él es que el modelo se guiará para poder realizar las distintas regresiones. En la Figura 3.6 se puede observar un kernel del tipo “polynomial”, el cual también presenta estos márgenes al momento de realizar la regresión, pero a diferencia del caso anterior, ahora se aprecia claramente su comportamiento no lineal (polinomio de grado 2), y también que ahora depende de más parámetros además de ϵ .

Figura 3.5: Regresión SVM, kernel “linear”

Fuente: Géron (2019a)

Figura 3.6: Regresión SVM, kernel “poly”

Fuente: Géron (2019a)

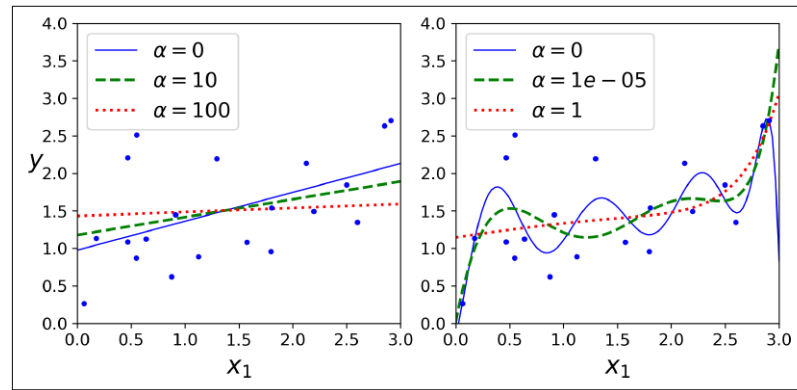
3.4.1.2. Ridge Regression

La regresión Ridge, es una regularización de la regresión lineal clásica. Se le llama regularización debido a que a la función de costos se agrega el término $\alpha \sum_{i=1}^n \theta_i^2$. Como es sabido, la regresión lineal funciona a través de la minimización de esta función, por lo que este proceso, además de realizar el mejor ajuste para todos los puntos, deberá mantener estos pesos lo más pequeño posible.

El hiperparámetro α es el corazón de esta nueva regresión, ya que controla que tanto se quiere “regularizar” el modelo. Por ejemplo, si $\alpha = 0$, todo el término se nulifica y la regresión pasa a ser completamente lineal. En cambio, si $\alpha = \infty$ (un valor muy grande), todos los pesos pasarán a ser valores cercanos a 0 (ya que es una minimización), por lo que el modelo tenderá a ajustarse al promedio de los datos. Considerando estos factores, la regresión Ridge propiamente tal es la siguiente:

$$J(\theta) = MSE(\theta) + \alpha \frac{1}{2} \sum_{i=1}^n \theta_i^2 \quad (3.1)$$

Esta regularización también puede ser utilizada para regresiones no lineales, como un polinomio, el cual se denomina Regresión Polinomial con regularización Ridge. Independientemente de esto, α toma la misma importancia, dictando que tan fuerte será esta regularización. Un ejemplo gráfico de modelos lineales y polinómicos se encuentran representados en la Figura 3.7. En ella se puede observar cómo ambos modelos de comportan frente a una sensibilización de este hiperparámetro.

Figura 3.7: Regularización Ridge para regresión lineal y polinomial

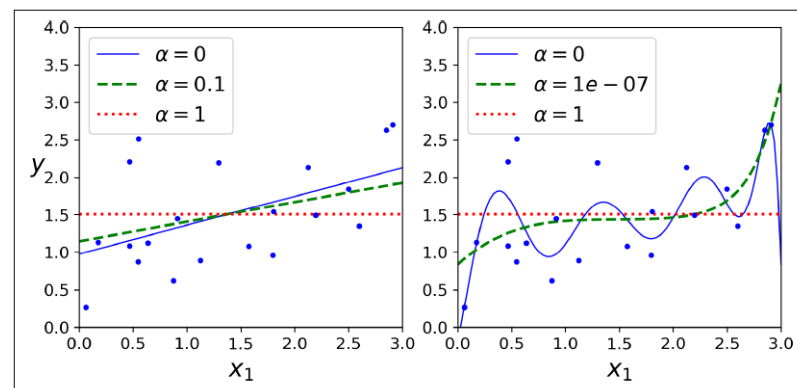
Fuente: Géron (2019a)

3.4.1.3. Lasso Regression

La regresión Lasso (Least absolute shrinkage and selection operator regression) es otro tipo de regularización de la regresión lineal clásica, y al igual que Ridge, se le agrega un término a la función de costos. La lógica es la misma, con la diferencia que ahora los pesos corresponden a la norma del vector de los pesos, quedando la siguiente regresión:

$$J(\theta) = MSE(\theta) + \alpha \sum_{i=1}^n |\theta_i| \quad (3.2)$$

Debido a este cambio, los valores de la minimización entre ambas regularizaciones serán drásticamente distintos debido a que uno tiene un término al cuadrado, lo que eventualmente llevará a que el modelo simplemente elimine algunos atributos, o en otras palabras, la mayor parte de los pesos serán 0, es por esto que esta regularización es mucho más robusta que Ridge, y será más sensible a ciertos comportamientos atípicos de la base de datos. Esto se puede comprobar comparando la Figura 3.7 y 3.8.

Figura 3.8: Regularización Lasso para regresión lineal y polinomial

Fuente: Géron (2019a)

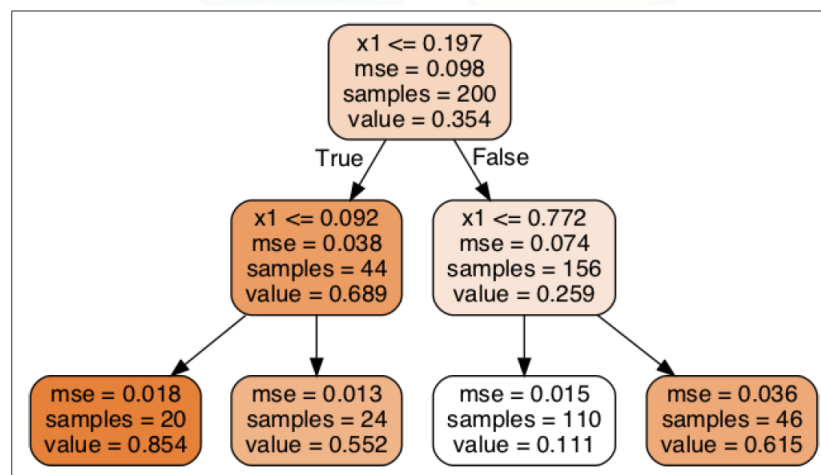
En este nuevo caso de regresión, se puede observar que la línea punteada verde (asociado a un $\alpha = 10^{-7}$) tiene un comportamiento cuadrático, pero al mismo tiempo lineal entre $x_1 = 1,0$ y $x_1 = 2,25$, lo que comprueba lo anteriormente estipulado: Lasso es mucho más estricto a momento de asignarle valores a los pesos.

3.4.1.4. Random Forest

Random Forest es un modelo del tipo “ensemble” de árboles de decisión. Este tipo de modelos tienen la característica de usar más de un tipo de algoritmo de aprendizaje al momento de entrenar el modelo, y es sumamente útil con árboles de decisión ya que su tiempo de cómputo es muy rápido. En general, el modelo es utilizado para clasificación, pero es lo suficientemente versátil como para poder usarse en regresiones más complejas. La diferencia radica netamente en la salida de los datos, ya que en problemas de clasificación, Random Forest predice una clase, y para regresión, un valor.

Como se puede observar en la Figura 3.9, se entrenó un árbol de decisión de profundidad 2. Aquí, por ejemplo, si se quisiera realizar una predicción para la instancia $x_1 = 0,6$, desde la “raíz” se debería primero viajar a la hoja de la derecha, ya que no se cumple con $x \leq 0,197$, seguido de la hoja justamente debajo de la anterior, ya que se cumple que $x \leq 0,772$, entregando una predicción de $\text{value}=0.111$, junto al número de la muestra y el error cuadrático medio (MSE).

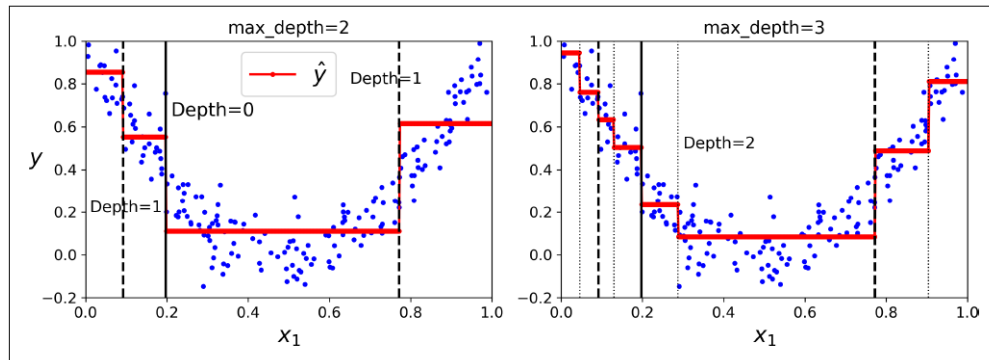
Figura 3.9: Árbol de decisión de profundidad 2



Fuente: Géron (2019a)

Aquí el hiperparámetro que más influye es precisamente la profundidad del árbol, ya que esta es la que dicta que tan precisa será la predicción. Su relevancia se puede ver reflejada en la Figura 3.10, en la que se puede ver que con una profundidad de 3 la regresión será bastante más precisa que con 2. Pero no hay que confiarse en este aspecto ya que siempre se necesita llegar a un equilibrio para no caer en el fenómeno denominado sobreajuste, en el que el modelo aprende tan bien el comportamiento de la base de datos que, al momento de entregarle nuevos datos este no es capaz de realizar una buena predicción.

Ahora, una de las grandes ventajas que tiene usar RF en regresiones, y también clasificación, es que se puede extraer un parámetro llamado “feature importances” o características importantes, la cual mide la relevancia de los nodos iniciales al momento de realizar las predicciones, a lo largo de todos los árboles generados. Esto resulta sumamente útil ya que precisamente cada nodo corresponderá a una variable del modelo, por lo que es muy factible realizar limpieza a la base de datos en función de estos resultados, o un análisis de relevancia para el modelo general.

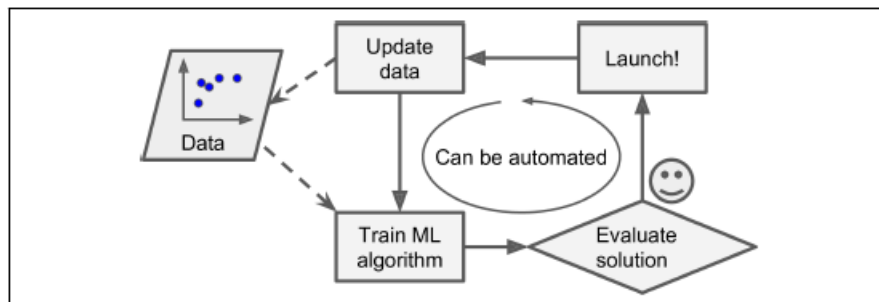
Figura 3.10: Comportamiento del árbol de decisión en función de su profundidad

Fuente: Géron (2019a)

3.4.2. Aprendizaje “on the fly”

Como se mencionó anteriormente, esta clasificación hace referencia al Machine Learning que mientras está operando, puede recibir nuevos datos, y procesarlos. De aquí nacen dos conceptos: Batch y Online Learning.

Batch Learning hace referencia a todos los modelos que son incapaces de aprender de manera paralela y creciente, por esto deben ser entrenados usando toda la base de datos que está disponible en cierto período de tiempo, por lo que en general se toman copias de los datos y se realiza el modelo en algún otro entorno que no influya al proceso en curso, para poder después entregarlo al área de producción. Desde este punto en adelante el modelo en sí no puede seguir aprendiendo, ya que ya fue entrenado y solamente aplicará lo que se le enseñó. No obstante, este proceso sí puede ser automatizado [Ver Figura 3.11], aplicando ingeniería en sistemas y generando un ciclo definiendo la frecuencia en la que se requiera actualizar la información. Lo malo de usar esta metodología, es que mientras más frecuencia de uso tenga, más dependerá en el equipo computacional en el que se esté modelando, debido a que consumirá muchos recursos de éste, como memoria RAM, espacio en el disco, CPU, etc., lo que se traducirá en elevados costos. Además, al tener un proceso que no depende de solo una instancia y siempre estará ocupando una larga lista de datos, realizar el modelo no solo significará un gasto importante, sino que tomará un tiempo bastante largo en entregar resultados, por lo que la frecuencia en la que se desea que el modelo entregue resultados puede que no sea la efectiva. Es por esto que si se tienen bases de datos muy grandes, este modelo no es recomendable.

Figura 3.11: Automatización de Batch Learning

Fuente: Géron (2019b)

Por esto es que existe el otro espectro de esta clasificación: Online Learning. Toman esta denominación los modelos que se entrenan de manera creciente a medida que se les va alimentando de información, lo que permite que se adapte a tiempo real. A diferencia de lo presentado en la Figura 3.11, este nuevo

algoritmo ese entrena en primera instancia, se evalúa, se entrega a producción y mientras opera, va recibiendo nueva información y entrenando al mismo tiempo. Claramente será muy útil en los procesos que reciban información constante y además se cuente con bases de datos extensas. Por otro lado, en términos de recursos computacionales, los recursos ocupados son mínimos debido a que el modelo ya se entrenó con la primera base de datos y aprendió al respecto, por lo que estos no se necesitarán más, al menos que se quieran ocupar para otro propósito, pero eso ya dependerá de cada caso.

En esta clase de modelo, la característica más importante es la tasa de aprendizaje. Este parámetro es clave debido a que dependiendo de su valor, el modelo podrá aprender rápido o lento dependiendo del caso, pero tendrá consecuencias debido a que si se escoge una tasa alta, el modelo aprenderá rápido pero tenderá a olvidar los datos anteriores al mismo ritmo, por lo que no será tan adaptable si los datos nuevos son muy distintos a los anteriores. Ahora, otro factor importante a tomar en cuenta de esta estructura es que, si se le alimenta de manera consistente datos que no corresponden a lo que se le entrenó, puede que se transforme en un modelo muy distinto debido a su constante aprendizaje.

3.4.3. Análisis de Datos y Patrones

La última forma de etiquetar los modelos de Machine Learning es por como “generaliza” los datos. La mayoría de los modelos en esta materia se utilizan para predecir datos inciertos a partir de datos históricos, lo que quiere decir que debe tener la capacidad para poder tener una idea general de los datos al momento de realizar las inferencias para llegar a los resultados, por esto es menester que tengan buen rendimiento al momento de entrenar los datos, pero más importante, poder adaptarse a las nuevas instancias que se les entregarán. Se definen dos acercamientos a estas generalizaciones: Instance-based Learning y Model-Based Learning.

Los modelos Instance-based hacen referencia a aprender por memoria, usualmente se les refiere como “lazy-learning”. Básicamente se le entregan ejemplos de datos que debe modelar y después, ya habiéndolos aprendido de memoria, los compara con las nuevas instancias utilizando medidas de similitud con el ejemplo base. Por ejemplo, si se quisiera realizar un algoritmo de este tipo para poder clasificar spam, éste compararía los mails con el de ejemplo en relación a alguna medida, como podría ser si el mail ocupa ciertas palabras clave, o si tiene cierta estructura, etc. Es una buena opción cuando se quieren realizar clasificaciones rápidas, y es muy adaptable a nuevos datos ya que se pueden recolectar en la medida que se van aplicando. No obstante, tiene un costo computacional alto debido a que por cada comprobación se debe revisar toda la base que se usó para el entrenamiento.

Por otro lado, se tienen los modelos del estilo Model-Based, los cuales toman la base de datos y realizan modelos, para poder finalmente realizar predicciones. En general, en esta metodología se siguen los mismos pasos: estudiar la base de datos para poder detectar alguna especie de comportamiento o tendencia que pueda dar luces de que modelo podría tener un mejor rendimiento, para después seleccionar el deseado. Hasta este punto es donde se diferencia de un modelo econométrico estadístico, ya que se le agrega la característica de aprendizaje, en donde el modelo seleccionado se comienza a entrenar con la base de datos con algún algoritmo, como lo podría ser minimización de costos, y probar con modelos parecidos al que se está trabajando, hasta que pueda llegar a la conclusión de cuál de todos los modelos propuestos es el que encaja mejor. Y finalmente se tiene la fase de prueba, donde se espera que el modelo final pueda realizar predicciones lo más acercadas a la realidad y que esté lo más generalizado posible.

4 | Metodología

4.1. Software

El modelamiento fue realizado con el editor de texto Google Colaboratory, el cual permite guardar el progreso en una nube, y programar en lenguaje Python, con la ventaja de tener integradas muchas de sus librerías para Machine Learning. Si bien no se aprovecha al 100 % el equipo local, ya que ocupa servidores de Google, sirve enormemente para trabajos de análisis y modelamiento con bases de dato pequeñas.

4.1.1. Librerías

Las librerías usadas para el correcto modelamiento, elección de modelos y herramientas de análisis estadístico fueron las siguientes:

```

1 import pandas as pd
2 import matplotlib.pyplot as plt
3 import numpy as np
4 import seaborn as sns
5 import statsmodels.api as sm
6 import inspect
7
8 from sklearn.model_selection import train_test_split
9 from sklearn.svm import SVR
10 from sklearn.metrics import r2_score, mean_squared_error, mean_absolute_error
11 from sklearn.preprocessing import StandardScaler, MinMaxScaler, RobustScaler,
    MaxAbsScaler
12 from sklearn.model_selection import GridSearchCV
13 from sklearn import ensemble
14 from sklearn.ensemble import RandomForestRegressor as RFR
15 from IPython.display import display
16 from sklearn import linear_model
17 from scipy import stats

```

Listing 4.1: Librerías utilizadas

4.2. Base de Datos

En la presente sección se procederá a explicar a fondo el cómo se formó la base de datos utilizada para modelar el consumo eléctrico de las comunas de la RM, el cual se dividirá en dos partes: la investigación y revisión bibliográfica previa, y también el proceso de limpieza de datos para poder trabajarlos de manera correcta.

4.2.1. Formación de la BBDD para el consumo eléctrico

Para comenzar, es menester mencionar que los datos de consumo eléctrico, según la Ley 19.628 Sobre Protección de la Vida Privada (1999) se encuentran protegidos, por lo que ninguno de los datos obtenidos están identificados, o etiquetados, personalmente. Solo figuran las empresas del rubro y la comuna en cuestión. Esto último se desarrollará más adelante.

Teniendo esto en cuenta, la base de datos se obtuvo gracias a la corporación autónoma sin fines de lucro Coordinador Eléctrico, la cual no forma parte de la Administración del Estado. Después de distintas instancias de comunicación con la corporación, se comunicó que ellos guardan datos únicamente de los balances eléctricos mensuales de las subestaciones que cada comuna posee, por lo que si se quisiera estimar el consumo total de alguna, se debería tomar el valor neto de cada subestación y sumar estos valores. Es necesario también aclarar dos cosas: esto es una aproximación al verdadero consumo debido a que no toda la energía que pasa por una subestación se ocupará en la comuna en la que ésta se encuentra, ya que existen muchas interconectadas entre comunas, pero según Coordinador Eléctrico y opiniones de expertos, sería una buena aproximación. Junto a esto, la corporación también facilitó, nuevamente desde su página, un documento el cual contenía todas las subestaciones de la RM, junto a qué comuna se encuentran instaladas, por lo que con estos dos archivos se realizó un cruce para poder identificar, a grandes rasgos, el consumo eléctrico comunal.

Los archivos utilizados consistieron en 36 planillas Excel, una por cada mes del período de estudio, las cuales contaban con una hoja por cada compañía eléctrica, ya fuera de la industria de distribución, transmisión o generación, no obstante, no todas contaban con subestaciones propias precisamente debido al rubro. Ahora, por esta gran dispersión de datos, la creación de la base utilizable tomó bastante tiempo y de externos profesionales en el rubro, que ayudaran a leerla, e interpretarla, de mejor manera. Como se puede observar en la **Tabla 4.1**, la información se presenta ya trabajada, y a modo de ejemplo se utilizará la empresa Transelec. La columna “Nombre” indica la subestación, “clave” es un nemotécnico que representa desde que “barra” está inyectando o retirando la energía, es por esto que la clave para el flujo entre Alto Jahuel y Cerro Chena es CHEJA. La numeración hace referencia a distintos puntos del flujo donde se registra nuevamente esta energía para distintos fines, por lo que se debía tener cuidado en considerar solo uno del mismo nemotécnico, en general el que tuviese el valor más alto. Finalmente la columna “Físico” representa el flujo registrado en el punto en cuestión, medido en kilowatts por hora, y “Monetario” representa la conversión de este último valor a pesos chilenos. Como ya se mencionó, estos datos fueron extraídos de los distintos archivos para poder ser presentados de esa forma, agregando también una nueva columna de “Mes”, en donde se diferencia de qué archivo se extrajo el registro.

Tabla 4.1: Tabla subestaciones Transelec

Mes	Nombre	clave	Barra	Físico [KWh]	Monetario [\$]
Enero	de_A.JAHUEL_220	CHEJA4	C.CHENA_220	104.676.624	3.317.747.066
Febrero	de_A.JAHUEL_220	CHEJA4	C.CHENA_220	79.827.685	2.676.713.191
Marzo	de_A.JAHUEL_220	CHEJA4	C.CHENA_220	72.487.003	3.585.488.722
Abril	de_A.JAHUEL_220	CHEJA4	C.CHENA_220	61.897.454	2.500.205.113
Mayo	de_A.JAHUEL_220	CHEJA4	C.CHENA_220	61.954.010	3.380.263.119
Junio	de_A.JAHUEL_220	CHEJA4	C.CHENA_220	63.269.195	3.394.754.926
Julio	de_A.JAHUEL_220	CHEJA4	C.CHENA_220	91.597.838	4.175.104.093
Agosto	de_A.JAHUEL_220	CHEJA4	C.CHENA_220	72.791.960	4.312.962.888
Septiembre	de_A.JAHUEL_220	CHEJA4	C.CHENA_220	72.505.642	2.947.761.416
Octubre	de_A.JAHUEL_220	CHEJA4	C.CHENA_220	125.658.815	4.935.191.759
Noviembre	de_A.JAHUEL_220	CHEJA4	C.CHENA_220	108.933.358	3.772.102.634
Diciembre	de_A.JAHUEL_220	CHEJA4	C.CHENA_220	73.067.798	2.723.886.317

Fuente: Elaboración propia

Otro punto importante a destacar, que limitó bastante el análisis esperado, fue que no todas las

empresas eléctricas de Santiago se encontraban en esta base, y debido a esto hay muchas comunas que simplemente figuraban sin ninguna subestación. De aquí se lograron obtener datos de consumo únicamente para las siguientes comunas: Alhué, Buin, Cerro Navia, Colina, El Monte, La Pintana, Lo Prado, Melipilla, Paine, Pirque, Pudahuel, Puente alto, Renca, San Bernardo, San José de Maipo, San Pedro, Santiago Centro, Talagante y Tiltill.

Teniendo todo esto en cuenta, se realizó una tabla resumen la cual contenía todas las subestaciones de la RM, junto a qué comuna ésta pertenecía, para poder finalmente generar una tabla dinámica que pudiera organizar todo de mejor manera. El resultado final de este ordenamiento está representado en la **Tabla 4.2**, la cual se utilizó posteriormente para el análisis.

Tabla 4.2: Consumo eléctrico mensual, en [KWh], por comuna. Años 2018-2020

Fecha	ALHUÉ	BUIN	CERRO NAVIA	...	TILTIL
ene-18	-997860,472	-277778856,7	144466374,5	...	298116784,6
feb-18	-917009,119	-51947505,99	125997100	...	179284977,1
mar-18	-936687,774	2027586,474	171038485,7	...	199154391,6
abr-18	-653309,733	-30070392,35	140830225,2	...	209106221,9
may-18	-591880,13	37848581,59	103095382,3	...	212486303,2
⋮	⋮	⋮	⋮	⋮	⋮
jun-19	-425819,43	92645522,14	133496864,7	...	-116391815,8
jul-19	-708498,47	148051208,6	133790713,4	...	-152362788
ago-19	-734864	25967025,98	163032153,5	...	-50960142,5
sept-19	-718972,168	-31615810,16	120387843,6	...	-48648311,34
⋮	⋮	⋮	⋮	⋮	⋮
oct-20	-787165,222	-436499885,3	162570132,9	...	165620944,8
nov-20	-1023905,396	-442499824,2	176386050	...	87168684,3
dic-20	-1108275,639	-315472868,7	166049087,8	...	18571042,77

Fuente: Elaboración Propia

Debido a que dentro de una misma subestación podían existir inyecciones o retiros de energía, hay períodos consecutivos donde el signo del consumo cambia. Esto no afectará al análisis debido a que el signo solo representa la dirección de la energía en la que se midió, en ese período en específico. Por esto, como se abordará más adelante, se transformarán todos los valores a absolutos.

4.2.2. Elección de variables

Como se podrá entender, el fin del análisis que se intentará hacer, es tratar de explicar el consumo anteriormente calculado. Por esto, ahora se hace necesaria la elección de variables independientes, que en conjunto o por sí solas (se debe comprobar), puedan estar relacionadas con el consumo. Las variables de estudio, junto a una explicación de por qué se cree relevante, son las siguientes:

- Casos Covid-19 (CCov-19):

La primera variable a escoger para el modelo, son la cantidad de casos registrados de Covid-19 por comuna desde que se registró el primero en Chile, esto es desde marzo del 2020. Se quiere estudiar en qué medida los efectos del virus, y eventualmente las cuarentenas, afectaron al consumo eléctrico residencial y empresarial. La información, que se encuentra abierta a todo público, se puede encontrar en los informes epidemiológicos del Ministerio de Salud de Chile (2020).

- Renta Hombre por hora (Renta H/h)

Se quiere verificar si el sueldo mensual de los habitantes de las distintas comunas explica de alguna u otra forma el consumo mensual. Aquí, debido a limitaciones de los datos disponibles a todo público, no se lograron encontrar los sueldos promedios específicos por comuna. Por esto, los datos a estudiar

serán el sueldo promedio nacional por hora del período en específico. Los datos, que se encuentran mensualizados y sueldo por hora, están disponibles en la base de datos del Instituto Nacional de Estadísticas (2021).

- **Renta Mujer por hora (Renta M/h)**

Aquí el análisis es análogo a la variable anterior, pero con la diferencia del género que en este caso es el femenino. Se espera, de todas formas, que haya una correlación fuerte entre ellas debido a su naturaleza, pero de todas formas se analizará cual tiene más peso frente al modelo y las demás variables de estudio. Al igual que la variable anterior, los datos se extrajeron desde a base de datos del Instituto Nacional de Estadísticas (2021).

- **Temperatura Media Santiago (TMe)**

Aquí el nombre de la variable habla por sí sola. Esta se obtuvo de un promedio mensual, el cual es el mismo para todos los años en sus respectivos meses debido a que no se pudo encontrar información más específica de manera gratuita. No obstante, se espera encontrar alguna relación debido a que, por ejemplo, en invierno es común gastar más electricidad debido a calefactores eléctricos y otros electrodomésticos. Si bien debido al calentamiento global, es de esperar que las temperaturas promedio de las estaciones del año varíen año a año, no escapa de la lógica asumir la misma temperatura mensual para los 3 años de estudio. Los datos se extrajeron del informe climático hecho por Weather Spark (2022).

- **Horas de sol durante el día (HSol)**

Con esta variable se espera que resulte en análisis parecidos a la TMe, debido a que el consumo eléctrico residencial diario es mayor durante la noche, por lo que se pretende capturar este efecto añadiendo al modelo la cantidad de horas de sol diarias cada mes. Al igual que con la variable anterior, estos datos se extrajeron del informe generado por Weather Spark (2022).

- **Índice de Precios al Consumidor (IPC)**

Como es de conocimiento público, el indicador IPC mide mensualmente la variación de la canasta de bienes y servicios, de los cuales está contabilizado el precio de la luz, es por esto que de alguna forma esta debe explicar la cantidad de KWh consumidos. Los indicadores usados se extrajeron de los informes mensuales del Servicio de Impuestos Internos (SII) (2022).

- **Precio de luz Medio del Mercado (PMM)**

El precio medio de mercado está determinado por las medias de los contratos informados por las empresas generadoras, el cual es utilizado para la indexación del precio de nudo de la energía (Comisión Nacional de Energía, 2020).

- **Tasa de desempleo (Desem)**

La intención de incluir la tasa de desempleo, al igual que CCov-19, intentará ser un acercamiento a un indicador de la cantidad de personas que se quedaron en sus casas antes y durante la pandemia. Se podría pensar que un mejor indicador para esto es teletrabajo, pero este dato lamentablemente tampoco se encontraba disponible para el público general. Los índices en cuestión fueron extraídos del informe generado por Datosmacro (2022).

- **Habitantes por comuna (Pob Comuna)**

Esta variable mide como se ha comportado el crecimiento de la población de las comunas a lo largo de los 3 años en cuestión. El dato se obtuvo del repositorio de la Biblioteca del Congreso Nacional de Chile / BCN (2017). A su vez, los datos disponibles se encontraban anualizados para cada comuna, por lo que el estudiante realizó regresiones en Excel para poder tener un valor creciente mensualizado. Después de un minucioso análisis, se llegó a la conclusión de que el ajuste más adecuado para los datos era lineal, y una vez establecida la curva a utilizar para cada comuna, se realizó la estimación mensual.

- **Índice de movilidad de la comuna (Índice mov comuna)**

Este indicador fue construido por el Instituto de Data Science de la Universidad del Desarrollo Bravo, Loreto, and Ferres, Leo. (2020), el cual representa la movilidad de cada comuna, normalizado por la población total. La medición del mismo se realizó gracias a distintas antenas que existen en las distintas comunas. Por ejemplo si dentro de una comuna existe la antena A, B y C, se considera como 1 viaje el

paso de una antena a otra: entre A y B es 1 viaje, lo mismo con A y C, pero si el habitante pasó por A y B y después por C, se consideran 2 viajes distintos. Es por eso que si bien no es un indicador exacto, sí capta de manera muy acertada todo el movimiento que existió durante cierto período en esa comuna. Esto intentará ser una variable más que trata de captar la cantidad de personas que permanecieron dentro de sus hogares, lo que eventualmente se traducirá en más, o menos, consumo.

- Variable Dummy (Covid)

Las variables “dummies” son variables binarias que intentan representar efectos sobre el modelo completo. En general son altamente utilizadas para poder representar variables categóricas que no pierden su sentido al ser reemplazados por 0 y 1, como sexo, estado civil, etc. En este caso se usará para poder representar el efecto del Covid-19, por lo que la variable tomará el valor de 0 hasta marzo del 2020, ya después de esta fecha tendrá el valor de 1. Una vez se haya escogido el modelo final, se podrá ver su efecto.

4.2.3. Preprocesamiento

Una vez establecidas las variables, es necesario realizar un pequeño preprocesamiento de los datos para poder ser trabajados de manera propicia. Si bien ya se realizó un trabajo previo con los datos (como se mencionó en la sección 4.1.1), esto no es suficiente ya que, en resumen, sólo se ordenaron los datos de manera mensual y por comuna. Es muy posible que si no se realizara este preprocesamiento, los modelos simplemente entreguen resultados erróneos o que no tienen coherencia con la realidad, como consumos negativos que resultarán en malos ajustes debido a que las variables se encuentran en unidades de medida completamente distintas.

Se comenzará revisando el consumo eléctrico por comuna. Primero, como se mencionó anteriormente, hay columnas que tienen valores negativos, pero lo único que representa este signo es la dirección del flujo eléctrico en el que se mide en la barra de la subestación, por lo que sigue siendo consumo. El problema para el modelo en sí, radica en que se interpretará como un cambio drástico de un período al siguiente, lo que entorpecerá el análisis y el buen ajuste. Por esto, se decidió tomar el valor absoluto de todos los registros. Ahora, habiendo hecho este cambio se procederá a verificar los valores nulos, o “NaN”, los cuales corresponden a registros que no fueron computados. Estos valores se encuentran representados en la siguiente tabla:

Tabla 4.3: Valores nulos de Consumo Eléctrico

ALHUÉ	0
BUIN	0
CERRO NAVIA	0
COLINA	1
EL MONTE	0
LA PINTANA	6
LO PRADO	3
MELIPILLA	0
PAINE	0
PIRQUE	0
PUDAHUEL	0
PUENTE ALTO	10
RENCA	12
SAN BERNARDO	0
SAN JOSÉ DE MAIPO	0
SAN PEDRO	0
SANTIAGO	0
TALAGANTE	0
TILTIL	0

Fuente: Elaboración Propia

Como era de esperarse, existen algunos valores nulos a lo largo de las variables. Para las comunas de Lo Prado, La Pintana y Colina, los valores nulos se reemplazarán por el promedio total de la columna (sin considerar “NaN”), ya que el consumo a lo largo de los 3 años de estudio no cambia tan drásticamente, por lo que el promedio se adaptará bien a estos meses en los que no se tiene información. No obstante, para Puente Alto y Renca, hay casi la misma cantidad de valores nulos que reales, por lo que tomar el promedio no sería un valor representativo para toda la comuna. Por esto, se reemplazarán los valores nulos por 0, y el modelo interpretará este valor como un punto de referencia, y la regresión se realizará sobre el primer valor que sea distinto a 0.

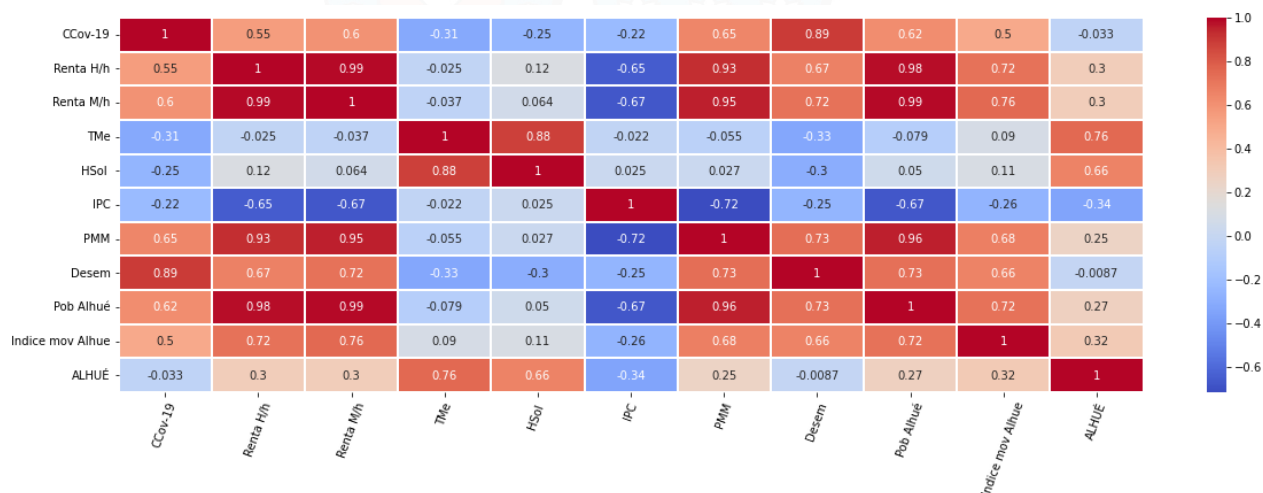
Ahora, se debe realizar un último paso antes de comenzar a escoger y entrenar modelos, que es una matriz de correlación para cada comuna. Esta entregará una visión bastante cercana de la multicolinealidad de las variables de estudio, ya que si ésta entrega valores muy altos, quizás sea mejor para el modelo quitar las variables que explican a otras. La única excepción será la inclusión de la variable dummy para el Covid, ya que es del tipo binaria y una matriz de correlación solo cobra sentido para variables continuas, por lo que únicamente para su análisis se quitará, después se reintegrará al momento de realizar las regresiones. Primero se realizarán las matrices para las comunas de Alhué y Buin, las cuales se pueden ver en la Figura 4.1 y 4.2, respectivamente. Las figuras para las demás comunas se encuentran en la sección de Anexos A.

Para el caso de Alhué se puede ver que existen 3 variables críticas que tienen una correlación bastante alta con las demás, las cuales son: Renta H/h, Renta M/h, Pob Alhué.

Como se puede observar en la Figura 4.1, Renta H/h tiene correlaciones bastante altas con Renta M/h, PMM, Pob Alhué e Índice mov Alhué, siendo todas mayores a 0.7. El hecho que tenga mucha relación con Renta M/h se justifica con que ambos sueldos suben de manera continua, y si bien en promedio la mujer gana menos que el hombre por hora, siguen teniendo un mismo comportamiento en el tiempo de aumentar constantemente, por lo que no es necesario incluir ambas para el modelo final. Así mismo, la alta relación con PMM también cobra sentido debido a que ambas son variables que se encuentran ligadas a una misma naturaleza, y puede que a lo largo de los 3 años ambos hayan presentado un comportamiento creciente, por lo que la matriz lo interpreta con una alta relación. Finalmente, el índice de movimiento comunal, debido a que no tiene una magnitud tan grande como las anteriores, no resulta tan evidente la razón de su valor, pero probablemente esté relacionado a que la gran mayoría de las personas con trabajo formal de esta comuna no podían quedarse en casa trabajando, y por ende se vieron forzados a realizar estos viajes.

Teniendo todo esto en consideración, se llegó a la conclusión de quitar de la base de datos las variables Renta M/h y Pob Comuna, debido a que estas dos presentan mayores correlaciones a lo largo de todas las variables, y la razón de por qué se quitó Renta M/h y no Renta H/h es netamente aleatoria. Si bien Renta M/h presentaba mayores magnitudes que Renta H/h, esta diferencia era lo suficientemente pequeña como para lograr una mejora considerable al momento de realizar los ajustes.

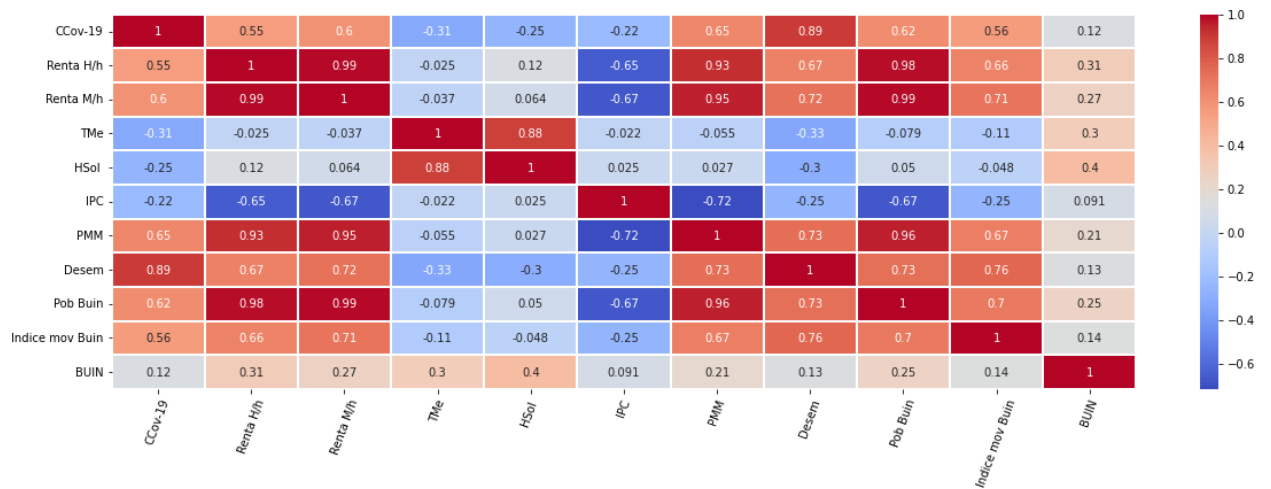
Figura 4.1: Matriz de Correlación de Alhué



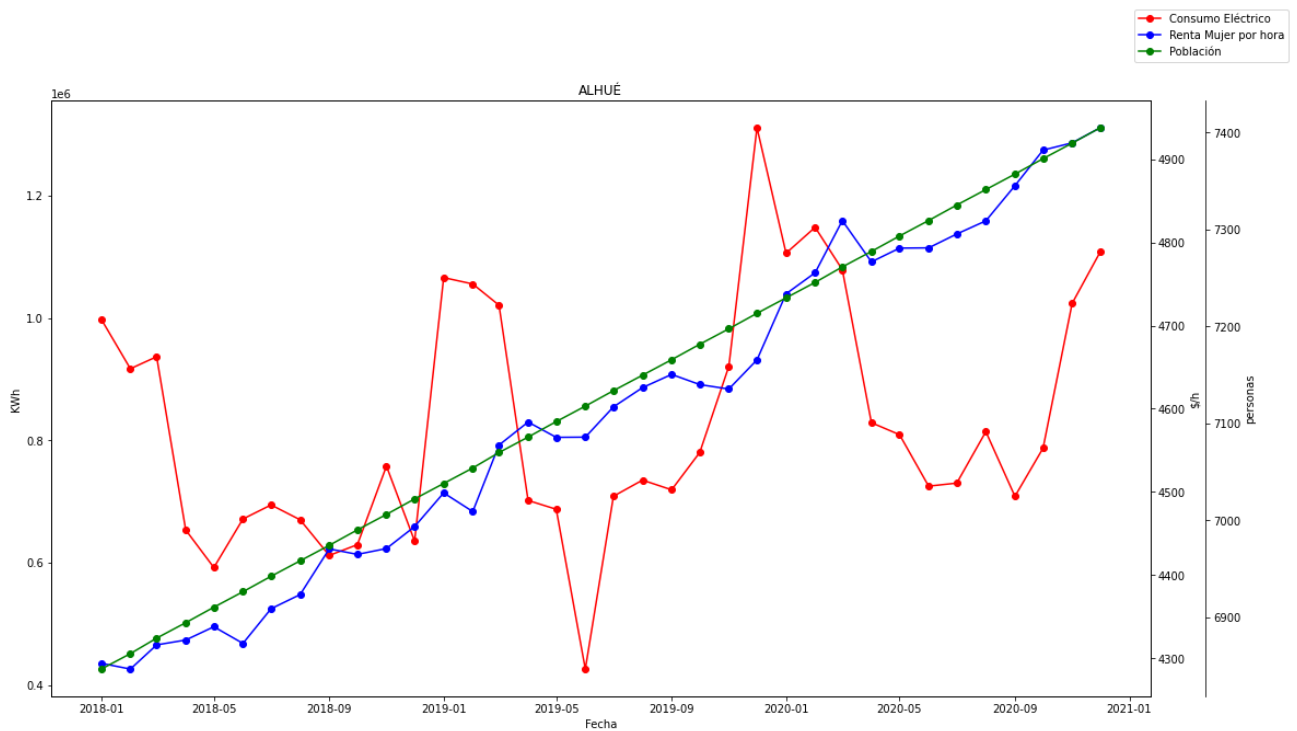
Fuente: Elaboración Propia

Ahora, para la comuna de Buin (Figura 4.2), existen las mismas relaciones entre variables que en Alhué. Esto tiene directa relación con la naturaleza anteriormente mencionada. Además, como se pudo ver en la sección 4.2.2, hay algunas características que son propias de la comuna, como la población o el índice de movilidad, pero con otras se decidió asignar los mismos valores, por falta de información disponible o simplemente porque no pierde sentido, como por ejemplo la temperatura media que entre comunas, que no debería variar significativamente. Esto también significa que la correlación entre estas variables será exactamente la misma para todas las comunas, lo que se puede ver representado en las Figuras 4.3 y 4.4. En ellos se puede ver la relación del consumo eléctrico versus estas dos variables de estudio, donde se aprecia que Renta M/h y Pob Comuna son exactamente iguales en ambos gráficos debido a lo mencionado anteriormente, con la diferencia del consumo, que sí es propio de la comuna.

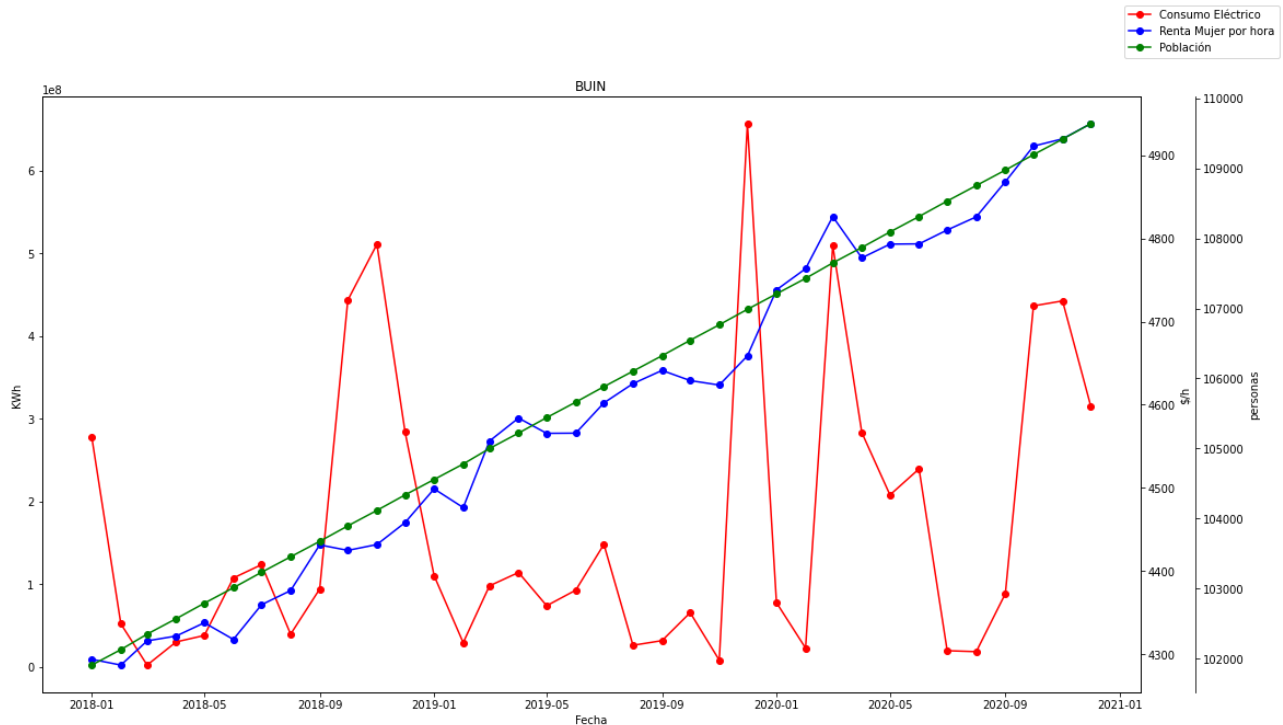
Sobre esto, si bien cada comuna tiene sus propias dimensiones de ingreso medio, cantidad de viviendas/población, etc., que puede llevar a observar distintos patrones en la explicación de su consumo, éste puede que siga tendencias similares en cuando se necesitará energía, al menos que la diferencia organizacional, política, económica y conductual sea muy distinta. Considerando esto, se realizó el análisis multivariado para todas las comunas de estudio, y se llegó a la conclusión final de quitar del estudio estas dos variables mencionadas anteriormente. Las matrices de correlación para las demás comunas, junto a los gráficos de consumo v/s Renta M/h y Población, se pueden encontrar dispuestas en la sección de Anexos A.1 y A.2.

Figura 4.2: Matriz de Correlación de Buin

Fuente: Elaboración Propia

Figura 4.3: Consumo Alhué v/s Renta M/h y Pob Alhué

Fuente: Elaboración Propia

Figura 4.4: Consumo Buin v/s Renta M/h y Pob Buin

Fuente: Elaboración Propia

Habiendo limpiado la base de datos, y teniendo analizadas las correlaciones junto al comportamiento de las variables más problemáticas, la base de datos se encuentra lista para poder ser entrenada en los modelos que en la siguiente sección se presentarán.

4.3. Elección de Modelo

Para poder escoger un modelo que se adapte mejor al comportamiento de los datos, es necesario realizar distintas pruebas, con distintos indicadores, que puedan definir qué es lo que se define como “mejor”. Para esto, hay que fijar el objetivo del estudio, el cual para este caso es encontrar la explicación del consumo eléctrico de las comunas en cuestión, por lo que el factor que más incidencia tendrá a la hora de decidir será el coeficiente de correlación R^2 . No obstante, sería un error guiarse únicamente por él, debido a que un buen ajuste R^2 no necesariamente quiere decir que es un “buen modelo”, si no que podría significar únicamente una fuerte relación lineal entre las variables explicativas con la independiente.

No obstante, existe otro indicador que puede complementarse con R^2 para poder discernir entre un buen y mal modelo, llamado Error Cuadrático Medio (MSE), el cual entregará la suma cuadrada de los errores de predicción del modelo. Además de esto, al momento de entrenar y validar el modelo, el MSE puede ser utilizado para poder medir si el modelo se encuentra sobreajustado o subajustado:

- $MS E^{train} >>> MS E^{test}$: El modelo se encuentra subajustado (underfitted)
- $MS E^{train} <<< MS E^{test}$: El modelo se encuentra sobreajustado (overfitted)

Entonces, considerando todos estos factores, se realizará el siguiente método de resolución para cada comuna:

1. Dividir la base de datos en entrenamiento y prueba

2. Guardar los indicadores de R^2 en la fase de entrenamiento, junto al RMSE
3. Guardar el $RMSE$ de la fase de prueba
4. Escoger un modelo que no presente sobreajuste ni subajuste, junto a un R^2 de prueba a lo menos mayor a 0.7

Habiendo finalizado el método, es necesario escoger modelos ad hoc que sean buenos candidatos en base a lo visto en la sección anterior. Por esto, primero se realizarán 2 regresiones OLS, ya que hay muchas variables con correlaciones positivas, y por la misma naturaleza de las variables, podría ser que existan relaciones lineales, y en este contexto, se compararán regresiones de bibliotecas distintas ya que tienen distintos métodos para calcular la constante del modelo, lo que podría eventualmente llevar a mejores o peores ajustes. Siguiendo la lógica lineal, se utilizará una de las regresiones anteriores, pero con un cambio drástico: se realizarán regularizaciones del tipo Lasso y Ridge. Esto llevará a que el modelo juzgue desde otro punto de vista los coeficientes de las variables, y darle mayor o menor importancia según lo que se estime que tiene relevancia para el modelo.

Por otro lado, también se utilizarán métodos no lineales, siendo estos dos modelos de Support Vector Machines, diferenciados por sus kernel: Linear y Radial Basis Function. El primero no es una regresión lineal convencional, sino que entregará buenos resultados cuando la base de datos presenta patrones que pueden ser separables por líneas rectas, lo que entregará eventualmente una gran recta que intente separarla. El segundo kernel se adaptará bastante a los datos que no se comportan de manera lineal, entregando curvas más complejas, que en general presentan buenos ajustes y realizan buenas predicciones cuando la base de datos no es tan extensa, como en este caso.

Finalmente, se considerará un último modelo: el regresor Random Forest. Este modelo, al igual que SVM, es un acercamiento no lineal a los patrones que toda la base de datos podría entregar. Gracias a esta detección, este modelo puede entregar prematuramente un muy acertado análisis de los coeficientes, al igual que Lasso y Ridge, se puede extraer una matriz con los valores de los coeficientes, calculados según su aporte a la regresión.

Teniendo los modelos y los datos preparados, se obtendrán los grandes promedios a lo largo de todas las comunas, para cada modelo, de los índices R^2 de entrenamiento y $RMSE$ de entrenamiento y prueba. Con estos datos se conformará la siguiente tabla, y será expuesta a modo de resumen con sus respectivos valores en la sección de Resultados:

Tabla 4.4: Comparación R^2 y RMSE de los distintos modelos

	SVR: 'rbf'	SVR: 'linear'	OLS sklearn	OLS stats.model	Ridge	Lasso	Random Forest
R Squared Train							
RMSE Train							
RMSE Test							

5 | Resultados

5.1. Entrenamiento

En esta primera fase, se realizará una división en la base de datos en entrenamiento y prueba, de manera aleatoria, y en este caso, un 80 % será dirigida hacia el entrenamiento, y el restante a prueba. Todos los datos que se trabajarán de ahora en adelante son parte de este mismo conjunto, y se pasará a la fase de prueba únicamente cuando todos los modelos hayan entrenado y estén listos para su comparación. Cabe decir que la división ocurre también para la variable independiente.

Estando en este punto se debe comenzar a escalar la base de datos. Este concepto es fundamental al momento de realizar ajustes, ya que ningún modelo entregará buenos resultados si las unidades de medida de cada variable están en escalas completamente distintas, debido a que la diferencia de magnitud en un mismo eje hará que éste ignore las más pequeñas. Es por esto que se hace necesario escalar la base de datos, y de esta forma todas las variables se encontrarán dentro de una misma escala, o rango, dependiendo del tipo de escalado. Para este caso de estudio, se aplicarán 4 escalados distintos:

- Escalado Estándar (Standard Scaler)
- Escalado Min Max (MinMax Scaler)
- Escalado Robusto (Robust Scaler)
- Escalado Max Abs (MaxAbs Scaler)

Con cada uno de estos escalados, se crearán bases de datos distintas, por lo que para las 19 comunas de estudio, se tendrán 7 modelos distintos (encabezados de Tabla 4.4), y para cada modelo se entrenarán 4 bases de datos originadas de la división anterior (80 % a entrenamiento), generándose un total de 532 modelos.

5.1.1. Random Forest

En este caso, se utilizará el Regresor Random Forest con 20 estimadores, lo que quiere decir que el modelo tendrá 20 “Trees”, y del mejor se extraerá una regresión que estará bien entrenada. Además, de aquí se podrá tener un buen acercamiento a la importancia de cada variable, ya que a su vez entrega valores entrenados de los coeficientes de la regresión, lo que será de mucha utilidad para el análisis.

Para los parámetros de Random Forest, se realizó una Validación Cruzada, o Cross Validation (CV, de aquí en adelante se referirá a este método de esta forma), de 5 iteraciones, lo que quiere decir que la base de datos de dividirá en 5 subconjuntos en los que se realizarán las validaciones, llegando a un modelo óptimo global. Este se utilizará para realizar el ajuste a las 4 bases de datos, para todas las comunas, resultando en la siguiente Tabla:

Tabla 5.1: RMSE de la Regresión Random Forest para cada comuna

Comuna/RMSE	RMSE stdr	RMSE minmax	RMSE robust	RMSE maxabs
ALHUÉ	5.037e+04	5.623e+04	5.607e+04	8.422e+04
BUIN	7.930e+07	7.055e+07	7.925e+07	7.229e+07
CERRO NAVIA	1.298e+07	1.028e+07	1.226e+07	1.074e+07
COLINA	3.698e+06	2.980e+06	2.924e+06	3.452e+06
EL MONTE	6.740e+04	8.090e+04	5.732e+04	7.231e+04
LA PINTANA	3.362e+05	3.686e+05	5.039e+05	4.843e+05
LO PRADO	2.171e+06	1.778e+06	1.899e+06	1.921e+06
MELIPILLA	4.469e+05	4.349e+05	4.839e+05	5.409e+05
PAINE	2.504e+06	2.369e+06	2.068e+06	2.072e+06
PIRQUE	8.541e+05	7.727e+05	7.625e+05	9.246e+05
PUDAHUEL	1.778e+07	2.277e+07	1.620e+07	1.710e+07
PUENTE ALTO	1.029e+06	1.509e+06	1.166e+06	1.216e+06
RENCA	1.457e+07	2.037e+07	1.533e+07	1.639e+07
SAN BERNARDO	1.958e+05	2.301e+05	3.549e+05	3.048e+05
SAN JOSÉ DE MAIPO	5.543e+06	5.202e+06	6.390e+06	5.343e+06
SAN PEDRO	1.456e+06	2.316e+06	2.663e+06	2.545e+06
SANTIAGO	5.599e+05	5.426e+05	7.262e+05	7.440e+05
TALAGANTE	7.806e+05	7.832e+05	6.800e+05	1.155e+06
TILTIL	3.239e+07	2.630e+07	2.410e+07	3.106e+07

Fuente: Elaboración Propia

De esta tabla, se calculará el promedio de los RMSE para cada tipo de escalado, es decir, para las 4 columnas, y será utilizado más adelante para su comparación. Además, se guardarán las magnitudes de los coeficientes para poder analizarlas al momento de la elección de modelos en la sección de Prueba.

5.1.2. Support Vector Machine: Radial Basis Function kernel

Para el siguiente modelo, se usará un regresor de Support Vector Machines (Support Vector Regression, SVR), con el kernel Radial Basis Function. Además, al igual que en el modelo anterior, se realizará una CV de 5 iteraciones para los parámetros de esta regresión, siendo estos dos:

- **C:** el cual corresponde a un parámetro de regularización que controla la tolerancia del modelo hacia errores de predicción, por lo que en este sentido también se debe evitar que se llegue a un sobreajuste (no se puede escoger un valor muy alto, pero la CV lo tendrá en consideración).
- **Gamma (γ):** que corresponde al coeficiente del kernel rbf, el cual influencia la clasificación de la base de datos, y al igual que “C”, influenciará fuertemente en la complejidad del ajuste, y si toma valores muy altos puede ser que el modelo no logre diferenciar patrones y simplemente no entregue un modelo.

Con todo esto en consideración, nuevamente se realizó el ajuste del modelo a lo largo de todas las comunas, para cada base distinta, resultando en la Tabla 5.2. Y, al igual que en el caso anterior, se guardaron los promedios de cada columna para poder ser comparados más adelante.

Tabla 5.2: RMSE de SVR: “rbf” para cada comuna

Comuna/RMSE	RMSE stdr	RMSE minmax	RMSE robust	RMSE maxabs
ALHUÉ	5.988e+04	5.559e+04	3.749e+04	1.074e+05
BUIN	1.235e+08	1.709e+08	1.758e+08	1.087e+08
CERRO NAVIA	1.473e+07	1.368e+07	1.221e+07	2.111e+07
COLINA	5.194e+06	5.044e+06	6.006e+06	6.196e+06
EL MONTE	7.345e+04	7.497e+04	7.502e+04	6.826e+04
LA PINTANA	3.576e+05	1.034e+06	1.135e+06	2.213e+05
LO PRADO	3.975e+06	2.798e+06	1.413e+06	1.482e+06
MELIPILLA	2.510e+00	1.350e+00	1.140e+06	1.214e+06
PAINE	3.664e+06	3.897e+06	1.854e+06	4.302e+06
PIRQUE	6.226e+05	5.187e+05	8.460e+04	5.165e+05
PUDAHUEL	1.050e+00	5.900e-01	3.360e+00	2.298e+07
PUENTE ALTO	2.342e+06	2.340e+06	7.880e+05	1.026e+06
RENCA	7.359e+07	7.359e+07	7.357e+07	7.360e+07
SAN BERNARDO	3.812e+05	3.773e+05	3.718e+05	3.469e+05
SAN JOSÉ DE MAIPO	1.955e+06	1.285e+06	3.405e+06	5.922e+06
SAN PEDRO	5.599e+06	5.109e+06	3.707e+06	6.523e+06
SANTIAGO	7.713e+05	7.024e+05	1.522e+06	4.173e+05
TALAGANTE	1.875e+05	4.769e+05	1.195e+05	4.092e+05
TILTIL	5.498e+07	5.372e+07	5.649e+07	6.051e+07

Fuente: Elaboración Propia

5.1.3. Support Vector Machine: linear kernel

Siguiendo la misma línea del modelo anterior, en este también se usará SVR, pero con un kernel distinto: linear. Como ya se explicó antes, este kernel intentará comprobar si se pueden realizar divisiones con líneas rectas de la base de datos para poder obtener un modelo que se intente adaptar. Para este caso, la CV aplicada también tendrá 5 iteraciones, y solo sobre un parámetro: “C”, el cual cumple la misma función a lo largo de los distintos kernel (regularizador).

Una vez calculado el valor óptimo del parámetro, se realizó el ajuste sobre las bases con distintas escalas, conformándose la Tabla 5.3 y al igual que antes, se guardó el promedio de RMSE para cada escalado, a lo largo de todas las comunas.

Tabla 5.3: RMSE de SVR: “linear” para cada comuna

Comuna/RMSE	RMSE stdr	RMSE minmax	RMSE robust	RMSE maxabs
ALHUÉ	1.175e+05	1.721e+05	1.320e+05	2.049e+05
BUIN	1.883e+08	1.886e+08	1.870e+08	1.886e+08
CERRO NAVIA	3.687e+07	3.697e+07	3.699e+07	3.699e+07
COLINA	1.290e+07	1.342e+07	1.350e+07	1.347e+07
EL MONTE	1.633e+05	3.146e+05	2.302e+05	3.651e+05
LA PINTANA	1.594e+06	1.967e+06	1.375e+06	2.008e+06
LO PRADO	5.631e+06	5.914e+06	5.821e+06	5.961e+06
MELIPILLA	2.493e+06	2.903e+06	2.795e+06	2.946e+06
PAINE	4.543e+06	4.626e+06	4.650e+06	4.635e+06
PIRQUE	2.715e+06	2.988e+06	2.963e+06	3.023e+06
PUDAHUEL	4.546e+07	4.546e+07	4.463e+07	4.546e+07
PUENTE ALTO	6.909e+06	7.470e+06	7.285e+06	7.575e+06
RENCA	7.361e+07	7.359e+07	7.365e+07	7.365e+07
SAN BERNARDO	2.635e+06	3.325e+06	3.182e+06	3.424e+06
SAN JOSÉ DE MAIPO	4.164e+07	4.209e+07	4.150e+07	4.214e+07
SAN PEDRO	8.369e+06	8.510e+06	8.439e+06	8.527e+06
SANTIAGO	2.728e+06	3.203e+06	2.137e+06	3.255e+06
TALAGANTE	4.496e+06	5.162e+06	4.740e+06	5.264e+06
TILTIL	1.102e+08	1.109e+08	1.096e+08	1.110e+08

Fuente: Elaboración Propia

5.1.4. OLS (A): biblioteca sklearn

Como se habló en la sección anterior, se usarán dos regresiones OLS que calculan la constante de la regresión, o el intercepto, por distintos métodos, resultando eventualmente en diferentes ajustes dependiendo de su valor.

En este caso, que para su comparación se denominará OLS (A), no se tiene que realizar nada más que ajustar el modelo a las 4 bases de datos, ya que la constante se calcula intrínsecamente. De aquí se conformó la Tabla 5.4. En ella se puede apreciar que a lo largo de todas las comunas, los errores parecen ser iguales, lo que haría pensar que se calculó de manera errónea, pero esto puede comprobarse de la siguiente manera.

Primero, se escogió aleatoriamente la comuna de Alhué, y los escalados Standard y MinMax para registrar la diferencia entre las predicciones hechas con cada uno. Una muestra de estas diferencias se encuentra expuesta en la Tabla 5.5, y como se puede apreciar, los valores son bastante pequeños, lo que quiere decir que es casi la misma predicción. Esto explica por qué al aproximarlos a los primeros 3 decimales, los valores de las filas presentados en la Tabla 5.4 sean iguales, además demuestra lo bien adaptado que el modelo se encuentra, ya que sigue detectando las mismas relaciones a pesar del escalado aplicado.

Esta prueba se realizó también para los demás escalados, y las respectivas tablas se encuentran en la sección de Anexos A.3. Allí se puede observar que las diferencias son del mismo orden de magnitud, lo que confirma que independientemente del escalado, el modelo entregará predicciones parecidas.

Tabla 5.4: RMSE de OLS (sklearn) para cada comuna

Comuna/RMSE	RMSE stdr	RMSE minmax	RMSE robust	RMSE maxabs
ALHUÉ	9.459e+04	9.459e+04	9.459e+04	9.459e+04
BUIN	1.189e+08	1.189e+08	1.189e+08	1.189e+08
CERRO NAVIA	1.839e+07	1.839e+07	1.839e+07	1.839e+07
COLINA	4.740e+06	4.740e+06	4.740e+06	4.740e+06
EL MONTE	1.135e+05	1.135e+05	1.135e+05	1.135e+05
LA PINTANA	9.513e+05	9.513e+05	9.513e+05	9.513e+05
LO PRADO	2.573e+06	2.573e+06	2.573e+06	2.573e+06
MELIPILLA	1.018e+06	1.018e+06	1.018e+06	1.018e+06
PAINE	3.437e+06	3.437e+06	3.437e+06	3.437e+06
PIRQUE	1.152e+06	1.152e+06	1.152e+06	1.152e+06
PUDAHUEL	2.793e+07	2.793e+07	2.793e+07	2.793e+07
PUENTE ALTO	1.864e+06	1.864e+06	1.864e+06	1.864e+06
RENCA	2.414e+07	2.414e+07	2.414e+07	2.414e+07
SAN BERNARDO	3.140e+05	3.140e+05	3.140e+05	3.140e+05
SAN JOSÉ DE MAIPO	1.201e+07	1.201e+07	1.201e+07	1.201e+07
SAN PEDRO	4.149e+06	4.149e+06	4.149e+06	4.149e+06
SANTIAGO	1.092e+06	1.092e+06	1.092e+06	1.092e+06
TALAGANTE	1.532e+06	1.532e+06	1.532e+06	1.532e+06
TILTIL	5.190e+07	5.190e+07	5.190e+07	5.190e+07

Fuente: Elaboración Propia

Tabla 5.5: Diferencia de escalados stdr y minmax, para Alhué

$y_{pred}^{stdr} - y_{pred}^{minmax}$
1,37E-07
2,40E-07
2,88E-07
4,48E-08
6,92E-07
3,15E-07
3,16E-07
3,38E-07
2,30E-07
1,87E-07
2,54E-07

Fuente: Elaboración Propia

5.1.5. OLS (B): biblioteca statsmodel.api

La diferencia en este caso (B), es que la regresión siempre asume que los dato se encuentran centrados, por lo que siempre la constante, o intercepto, siempre cruzará el origen, al menos que se le agregue una constante, de manera manual, al modelo. Entonces, cuando se realice el ajuste del modelo, la base de datos tendrá una columna extra, con el valor 1 en cada una de sus filas, que representará constante. Un ejemplo de esto se muestra en la Tabla 5.6, que se obtuvo de los datos de la comuna de Alhué, escalada por el método stdr. En ella se puede observar la primera columna de valores unitarios, junto a la fecha y 4 variables presentadas en la sección 4.2.2.

Tabla 5.6: Representación de Data Frame de Alhué utilizado en la biblioteca statsmodel.api

cte	Mes	Año	CCov-19	Renta H/h	TMe
1.0	1.598722	-1.183216	-0.470577	-0.648674	1.103060
1.0	-0.079936	1.183216	3.920149	0.727032	-1.367795
1.0	-0.639489	1.183216	-0.077202	0.720241	0.073537
1.0	1.598722	1.183216	1.223467	1.493357	1.103060
1.0	1.039169	1.183216	0.925349	1.891292	-0.132367

Al realizar esta adicción de una nueva columna a todas las comunas, se realizó el ajuste del modelo, y como es de esperarse se formó la siguiente Tabla 5.7 con sus respectivos RMSE, de la cual se guardaron los promedios para cada escalado.

Tabla 5.7: RMSE de OLS (statsmodel.api) para cada comuna

Comuna/RMSE	RMSE stdr	RMSE minmax	RMSE robust	RMSE maxabs
ALHUÉ	9.459e+04	9.459e+04	9.459e+04	9.459e+04
BUIN	1.189e+08	1.189e+08	1.189e+08	1.189e+08
CERRO NAVIA	1.839e+07	1.839e+07	1.839e+07	1.839e+07
COLINA	4.740e+06	4.740e+06	4.740e+06	4.740e+06
EL MONTE	1.135e+05	1.135e+05	1.135e+05	1.135e+05
LA PINTANA	9.513e+05	9.513e+05	9.513e+05	9.513e+05
LO PRADO	2.573e+06	2.573e+06	2.573e+06	2.573e+06
MELIPILLA	1.018e+06	1.018e+06	1.018e+06	1.018e+06
PAINE	3.437e+06	3.437e+06	3.437e+06	3.437e+06
PIRQUE	1.152e+06	1.152e+06	1.152e+06	1.152e+06
PUDAHUEL	2.793e+07	2.793e+07	2.793e+07	2.793e+07
PUENTE ALTO	1.864e+06	1.864e+06	1.864e+06	1.864e+06
RENCA	2.414e+07	2.414e+07	2.414e+07	2.414e+07
SAN BERNARDO	3.140e+05	3.140e+05	3.140e+05	3.140e+05
SAN JOSÉ DE MAIPO	1.201e+07	1.201e+07	1.201e+07	1.201e+07
SAN PEDRO	4.149e+06	4.149e+06	4.149e+06	4.149e+06
SANTIAGO	1.092e+06	1.092e+06	1.092e+06	1.092e+06
TALAGANTE	1.532e+06	1.532e+06	1.532e+06	1.532e+06
TILTIL	5.190e+07	5.190e+07	5.190e+07	5.190e+07

Fuente: Elaboración Propia

Cabe destacar que para esta librería, sucede la misma situación de la Tabla 5.4, en la que los errores, para cada comuna, a lo largo de los distintos escalados no varía. No obstante, lo que realmente sucede es que la predicción para los distintos escalados es casi la misma, como se pudo comprobar en la Tabla 5.5.

5.1.6. Regresión Ridge

Esta regresión, al igual que las dos anteriores, es lineal pero tiene un gran cambio, ya que se le aplica una regularización, o norma, L2, llamada Ridge. En simples términos es de alguna otra forma lo que aplica Random Forest, cambia los coeficientes de la regresión en función de su importancia para toda la base de datos, asignando una penalización para cada variable. Por esto es posible que algunas variables simplemente no aporten para el problema y su coeficiente sea 0, o al contrario, aporten lo suficiente como para ser bastante alto. De todo esto, se puede obtener mucho análisis pero no valdrá la pena si no se escoge este modelo como el mejor, por lo que se dejará para más adelante de ser necesario.

Al realizar el ajuste de la regresión Ridge, se generó la Tabla 5.8, y para su posterior análisis se guardaron los valores promedio de los RMSE.

Tabla 5.8: RMSE de regresión Ridge para cada comuna

Comuna/RMSE	RMSE stdr	RMSE minmax	RMSE robust	RMSE maxabs
ALHUÉ	1.548e+05	1.988e+05	1.842e+05	2.087e+05
BUIN	1.479e+08	1.644e+08	1.566e+08	1.672e+08
CERRO NAVIA	3.320e+07	3.613e+07	3.433e+07	3.648e+07
COLINA	9.264e+06	1.236e+07	1.117e+07	1.300e+07
EL MONTE	2.439e+05	3.507e+05	2.857e+05	3.737e+05
LA PINTANA	1.383e+06	1.858e+06	1.231e+06	1.953e+06
LO PRADO	4.716e+06	5.629e+06	5.275e+06	5.898e+06
MELIPILLA	1.956e+06	2.482e+06	2.319e+06	2.576e+06
PAINE	4.106e+06	4.293e+06	4.130e+06	4.306e+06
PIRQUE	2.330e+06	2.893e+06	2.728e+06	3.000e+06
PUDAHUEL	4.223e+07	4.461e+07	4.291e+07	4.501e+07
PUENTE ALTO	4.129e+06	5.771e+06	5.201e+06	6.281e+06
RENCA	5.314e+07	6.121e+07	5.798e+07	6.297e+07
SAN BERNARDO	1.597e+06	2.455e+06	2.154e+06	2.707e+06
SAN JOSÉ DE MAIPO	2.722e+07	3.685e+07	3.343e+07	3.855e+07
SAN PEDRO	7.203e+06	8.153e+06	7.846e+06	8.317e+06
SANTIAGO	2.017e+06	2.810e+06	1.989e+06	2.951e+06
TALAGANTE	3.650e+06	4.704e+06	4.340e+06	5.027e+06
TILTIL	7.717e+07	9.894e+07	9.044e+07	1.059e+08

Fuente: Elaboración Propia

5.1.7. Regresión Lasso

Este modelo presenta la misma naturaleza de la regresión anterior, pero se le aplicará otro tipo de regularización, ahora del tipo L1 denominada Lasso. Presenta muchas diferencias pequeñas con la anterior, pero esta considera las penalizaciones “naturales”, en el sentido de que no se les aplica ninguna transformación o escalado al momento de calcularlas, a diferencia de Ridge que toma el cuadrado de las penalizaciones, por lo que restringe mucho más el modelo.

Por lo tanto, el ajuste se realizó sin problemas, generándose la recurrente Tabla 5.9 con los RMSE de cada comuna y cada escalado. En ella se puede observar el efecto de lo comentado en el párrafo anterior, ya que los coeficientes de las regresiones para cada comuna se redujeron a la raíz cuadrada, lo que debería tener un impacto evidente las predicciones.

Tabla 5.9: RMSE de regresión Lasso para cada comuna

Comuna/RMSE	RMSE stdr	RMSE minmax	RMSE robust	RMSE maxabs
ALHUÉ	9.460e+04	9.468e+04	9.462e+04	9.640e+04
BUIN	1.189e+08	1.189e+08	1.189e+08	1.189e+08
CERRO NAVIA	1.839e+07	1.839e+07	1.839e+07	1.849e+07
COLINA	4.740e+06	4.740e+06	4.740e+06	5.148e+06
EL MONTE	1.135e+05	1.137e+05	1.136e+05	1.186e+05
LA PINTANA	9.513e+05	9.513e+05	9.513e+05	9.542e+05
LO PRADO	2.573e+06	2.573e+06	2.573e+06	2.648e+06
MELIPILLA	1.018e+06	1.019e+06	1.018e+06	1.046e+06
PAINE	3.437e+06	3.437e+06	3.437e+06	3.442e+06
PIRQUE	1.152e+06	1.152e+06	1.152e+06	1.312e+06
PUDAHUEL	2.793e+07	2.793e+07	2.793e+07	2.793e+07
PUENTE ALTO	1.864e+06	1.864e+06	1.864e+06	2.023e+06
RENCA	2.414e+07	2.414e+07	2.414e+07	2.423e+07
SAN BERNARDO	3.141e+05	3.142e+05	3.141e+05	3.157e+05
SAN JOSÉ DE MAIPO	1.201e+07	1.201e+07	1.201e+07	1.217e+07
SAN PEDRO	4.149e+06	4.149e+06	4.149e+06	4.477e+06
SANTIAGO	1.092e+06	1.092e+06	1.092e+06	1.145e+06
TALAGANTE	1.532e+06	1.533e+06	1.532e+06	1.555e+06
TILTIL	5.190e+07	5.190e+07	5.190e+07	5.194e+07

Fuente: Elaboración Propia

5.1.8. Resultados Entrenamiento

Habiendo llegado a este punto, con todos los modelos entrenados, es necesario recopilar toda la información para poder ahora verificar qué escalado fue el que entregó menor error. Esto para poder ahorrar trabajo en la siguiente sección, donde se realizará la validación de los modelos con el escalado que se seleccionará ahora. Para esto, y como se mencionó a lo largo de esta sección, se tomarán los RMSE promedio para cada modelo, y se dispondrán en la siguiente Tabla:

Tabla 5.10: RMSE promedios, de cada modelo, para los distintos escalados

Escalado/Modelo	SVR: rbf	SVR: linear	OLS (A)	OLS (B)	Ridge	Lasso	RF
Standard	1.537e+07	2.902e+07	1.454e+07	1.454e+07	2.229e+07	1.454e+07	9.301e+06
MinMax	1.766e+07	2.935e+07	1.454e+07	1.454e+07	2.611e+07	1.454e+07	8.931e+06
Robust	1.788e+07	2.898e+07	1.454e+07	1.454e+07	2.445e+07	1.454e+07	8.835e+06
MaxAbs	1.662e+07	2.939e+07	1.454e+07	1.454e+07	2.698e+07	1.463e+07	8.865e+06

Fuente: Elaboración Propia

Como se puede apreciar, ningún modelo cambia mucho el orden de magnitud del error independientemente del escalado, y efectivamente ocurren variaciones pero no son tan significativas. No obstante el que logra sobresalir como el que generó menos error en promedio, fue el standard, y resulta bastante lógico debido a que en general este escalado es el más utilizado debido a que se adapta muy bien a todo tipo de datos, y en general cualquier otro se usa en caso específicos donde el error disminuye considerablemente, y cuando se tienen diferencias en magnitudes distintas, la respuesta al mejor escalado es clara. Por lo tanto, para poder realizar la comparación en la prueba de modelos, se usará únicamente el escalado standard.

5.2. Prueba

En esta sección se presentarán los resultados de la fase de prueba de los modelos. Como se mencionó al inicio del capítulo de resultados, la base de datos se dividió en entrenamiento y prueba, para poder entrenar el modelo y después comprobar cuál es el mejor con la sección de prueba.

5.2.1. Resultados Prueba

Con la decisión de ocupar el escalado standard, y teniendo los modelos con los coeficientes y pesos ya entrenados, se realizará nuevamente el ajuste pero sobre este otro subconjunto de prueba. El procedimiento será el mismo que en entrenamiento: escalar la base, realizar el ajuste, guardar el valor de sus RMSE promedio y compararlo. Con todo esto en consideración, los resultados de los ajustes de prueba se presentan en la siguiente tabla:

Tabla 5.11: RMSE de prueba promedios de cada modelo, para todas las comunas

Escalado/Modelo	SVR: rbf	SVR: linear	OLS (A)	OLS (B)	Ridge	Lasso	RF
Standard	1.927e+07	2.362e+07	2.303e+07	2.303e+07	2.248e+07	2.303e+07	2.002e+07

Fuente: Elaboración Propia

5.3. Comparación Final

Ahora, para poder decidir cuál será el mejor modelo, se deberán comparar los resultados de las dos fases. Además de esto, como se mencionó en la sección 4.3, también se agregará el coeficiente de determinación para poder tener otra dimensión, básicamente con la premisa de que un modelo que tenga buen RMSE, para los efectos de este estudio, no puede tener un R^2 bajo, ya que se busca encontrar la explicación del consumo eléctrico sobre las distintas comunas de Santiago, no predecirlo.

La tabla final, con todos los índices anteriormente mencionados, se presenta a continuación:

Tabla 5.12: Comparación final de RMSE promedios, de entrenamiento y prueba.}

Índice/Modelo	SVR: rbf	SVR: linear	OLS (A)	OLS (B)	Ridge	Lasso	RF
R Squared Train	7.689e-01	1.056e-01	6.748e-01	7.963e-01	4.193e-01	7.963e-01	9.191e-01
RMSE Train	1.537e+07	2.902e+07	1.454e+07	1.454e+07	2.229e+07	1.454e+07	9.301e+06
RMSE Test	1.927e+07	2.362e+07	2.303e+07	2.303e+07	2.248e+07	2.303e+07	2.002e+07

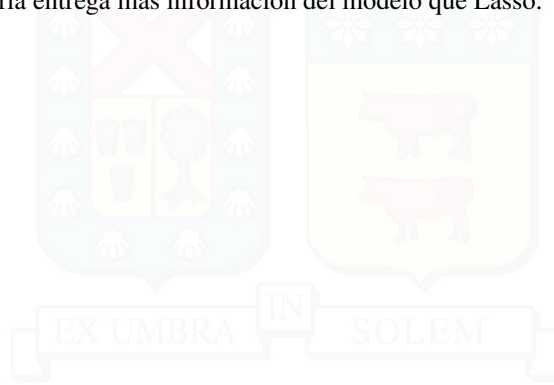
Fuente: Elaboración Propia

De primeras, se puede ver que Random Forest presenta un error de un orden de magnitud más bajo que los demás, lo que querría decir que es un muy buen modelo para predecir, por lo que sería buen candidato hasta que se observa el RMSE de prueba, donde aumenta un orden de magnitud, por lo que tiene un error en raíz cuadrada 10 veces más grande, lo que significa que el modelo presenta sobreajuste, y tiene sentido debido a que tiene un R^2 de 0.9. Ahora, quitando RF de la decisión final, el segundo filtro será el valor de R^2 , que como se estableció en un principio se espera que este sea al menos 0.7, por lo que descarta de inmediato SVR con kernel linear, OLS(A), y Ridge.

Ahora, observando la SVR con kernel rbf, entre estos tres modelos entrega la diferencia más pequeña entre entrenamiento y prueba, por lo que perfectamente podría ser el modelo indicado. No obstante se tiene el factor de R^2 , que es más alto en Lasso y OLS, por lo que se debe hacer la distinción: lo que se busca con el modelo es tener una explicación sólida del consumo eléctrico de la comuna, y que tan bien predice el modelo

no afecta directamente a esta explicación, por esto es que desde un principio se estableció que se debe tomar en cuenta ambos factores, explicación y predicción. Por esto, no se escogerá rbf, pero es menester mencionar que si se quisiera realizar una predicción de consumo, este sería el indicado.

De los modelos restantes, se puede ver en la tabla que son muy parecidos, esto es debido a su naturaleza lineal y forma de cálculo de los coeficientes (recordar que Lasso no es tan restrictivo como Ridge, por lo que hace sentido que los coeficientes sean parecidos con una OLS normal). Debido a esta similitud, no se puede diferenciar a ciencia cierta qué modelo es efectivamente mejor que otro, por lo que se escogerá OLS(B) meramente porque su librería entrega más información del modelo que Lasso.



6 | Discusión y Análisis

Habiendo definido un modelo, el siguiente paso es comenzar a realizar la discusión de los resultados con los parámetros del modelo y la ayuda de pruebas de hipótesis.

A modo de resumen, el modelo en cuestión es una regresión lineal por el método de Ordinary Least Squares, el cual entrega un R^2 promedio a lo largo de todas las comunas de 0.8, y una no muy grade diferencia de los RMSE de entrenamiento y prueba por lo que es un buen modelo explicativo y predictivo. Teniendo esto establecido, el análisis se basará en los resultados de una “Tabla Resumen” de la regresión OLS, que contiene información estadística importante de todo el modelo y sobre todas las variables. Además, se incorporará información que el modelo Random Forest aportó sobre los coeficientes, únicamente para poder agregar más análisis desde un punto de vista no lineal, de un modelo con un alto R^2 .

6.1. Alhué

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

OLS Regression Results

=====

Dep. Variable:

ALHUE

R-squared:

0.802

Model:

OLS

Adj. R-squared:

0.665

Method:

Least Squares

F-statistic:

5.879

Date:

Mon, 11 Jul 2022

Prob (F-statistic):

0.000814

Time:

01:43:55

Log-Likelihood:

-360.54

No. Observations:

28

AIC:

745.1

Df Residuals:

16

BIC:

761.1

Df Model:

11

Covariance Type:

nonrobust

=====

coef

std err

t

P>|t|

[0.025

0.975]

const

8.292e+05

1.19e+05

6.966

0.000

5.77e+05

1.08e+06

x1

3.128e+04

6.51e+04

0.481

0.637

-1.07e+05

1.69e+05

x2

7.563e+04

1.97e+05

0.384

0.706

-3.42e+05

4.93e+05

x3

3.683e+04

6.33e+04

0.581

0.569

-9.75e+04

1.71e+05

x4

6.288e+04

1.58e+05

0.398

0.696

-2.72e+05

3.98e+05

x5

1.766e+05

9.86e+04

1.790

0.092

-3.25e+04

3.86e+05

x6

-1.252e+04

9.9e+04

-0.127

0.901

-2.22e+05

1.97e+05

x7

-1.14e+05

7.5e+04

-1.520

0.148

-2.73e+05

4.5e+04

x8

-2.115e+05

9.96e+04

-2.123

0.050

-4.23e+05

-324.701

x9

82.0436

1.46e+05

0.001

1.000

-3.1e+05

3.11e+05

x10

1.602e+04

9.74e+04

0.164

0.871

-1.9e+05

2.23e+05

x11

3.124e+04

3.63e+05

0.086

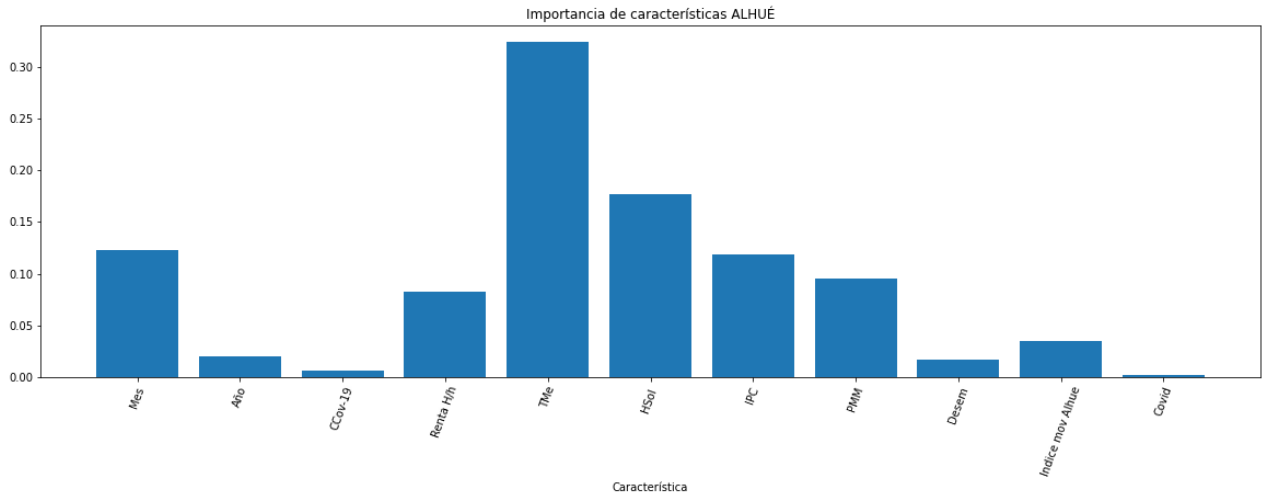
0.932

-7.38e+05

8.01e+05

=====

Listing 6.1: OLS Alhué

Figura 6.1: Importancia Variables Alhué, del modelo Random Forest

Lo primero que se puede destacar es el resultado de la prueba F, el cual tiene un $F^{statistic} = 0,000814$, lo que confirma que se puede rechazar la hipótesis nula, por lo que el modelo en conjunto sí es significativo, lo que se respalda con el valor del $R^2 = 0,802$ y $R^2_{adj} = 0,665$. Todo esto confirma que el modelo puede ser analizado con confianza.

Ahora, para las variables se deben destacar las que son significativas según la prueba t , con su valor $P > |t|$. De las 11 variables del modelo, sin considerar la constante, solamente x_5 y x_8 son significativas, correspondientes a las variables de Temperatura Media y Precio Medio del Mercado. Además, si se observa la Figura 6.1, TMe es la más importante por lo que tiene una sincronía con el análisis anterior. De esto se puede decir que el modelo presenta multicolinealidad porque a pesar de que el modelo el conjunto sea significativo, solo dos variables lo son. Siguiendo en esta última figura, además de TMe, también destaca Mes, Renta H/h, HSol, IPC y PMM. De esto se pueden obtener las siguientes conclusiones:

- La multicolinealidad se puede explicar debido a que el comportamiento de algunas variables es bastante parecido, por lo que se podrían eventualmente representar como combinaciones lineales de otras. Además, como se pudo observar en los resultados de Random Forest, hay más variables que se consideran importantes en un contexto no lineal.
- La TMe y PMM son las variables que más afectan al consumo de Alhué, lo que tiene sentido debido a que es una comuna con viviendas muy antiguas que no tienen muy buen aislamiento, por lo que para temperaturas altas o bajas, es necesario tener un aparato eléctrico que regule la temperatura. Y con la siguiente variable Precio Medio del Mercado, es normal que explique el consumo ya que en general siempre el precio es lo que dicta cuanto se va a consumir, y para esta comuna la luz sí es un bien que se debe controlar.
- La variable x_{11} , que corresponde a la dummy para Covid, presenta un valor $P > |t|$ muy alto, por lo que no se puede afirmar con certeza que la pandemia haya afectado mucho al consumo ya que no generó un cambio estructural. Cabe destacar que el modelo Random Forest tampoco consideró que fuera una variable muy importante para la regresión, por lo que se podría concluir que la pandemia no afectó en gran medida a esta comuna.

6.2. Buin

OLS Regression Results

Dep. Variable:

BUIN

R-squared:

0.510

Model:

OLS

Adj. R-squared:

0.174

Method:

Least Squares

F-statistic:

1.517

Date:

Mon, 11 Jul 2022

Prob (F-statistic):

0.218

Time:

17:26:57

Log-Likelihood:

-560.36

No. Observations:

28

AIC:

1145.

Df Residuals:

16

BIC:

1161.

Df Model:

11

Covariance Type:

nonrobust

coef

std err

t

P>|t|

[0.025

0.975]

const

1.617e+08

8.84e+07

1.829

0.086

-2.57e+07

3.49e+08

x1

5.049e+07

8.18e+07

0.617

0.546

-1.23e+08

2.24e+08

x2

-2.167e+08

2.43e+08

-0.893

0.385

-7.31e+08

2.98e+08

x3

-3.492e+07

7.95e+07

-0.439

0.666

-2.03e+08

1.34e+08

x4

1.8e+08

2e+08

0.901

0.381

-2.43e+08

6.03e+08

x5

1.481e+08

1.26e+08

1.178

0.256

-1.18e+08

4.15e+08

x6

-7.138e+07

1.28e+08

-0.559

0.584

-3.42e+08

1.99e+08

x7

-2.52e+07

9.16e+07

-0.275

0.787

-2.19e+08

1.69e+08

x8

6.617e+06

1.25e+08

0.053

0.959

-2.59e+08

2.72e+08

x9

9.968e+07

1.35e+08

0.739

0.471

-1.86e+08

3.86e+08

x10

-3.661e+07

6.49e+07

-0.564

0.581

-1.74e+08

1.01e+08

x11

9.014e+07

2.59e+08

0.348

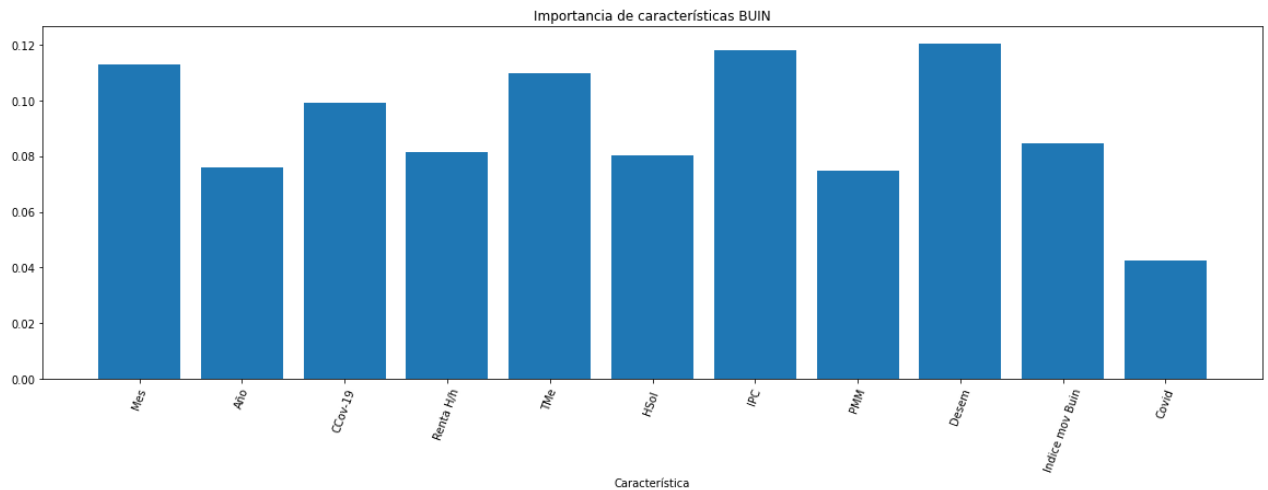
0.732

-4.59e+08

6.39e+08

Listing 6.2: OLS Buin

Figura 6.2: Importancia Variables Buin, del modelo Random Forest



Aquí ocurre una situación un tanto distinta al modelo anterior, ya que de partida el valor de R^2 es bajo para los estándares, y el R^2_{adj} es mucho más bajo aún, lo que da luces de inmediato de que el modelo no es muy bueno al considerar el error que cada variable puede aportar. Esto se confirma con la prueba de hipótesis F, la cual entrega un $F^{statistic} = 0.218$, por lo que no se puede afirmar que el modelo en conjunto es significativo, y si a esto le agregamos los bajos valores en los coeficientes de determinación, se puede concluir que, según el comportamiento de las variables de Buin, no es correcto aplicar una regresión lineal, por lo que todo su análisis se basará en los resultados de Random Forest:

- Como se puede observar en la Figura 6.2, todas las variables tienen asociadas importancias bastante parecidas, lo que puede confirmar nuevamente la relación no lineal que existe dentro de la comuna. Básicamente el consumo no se puede explicar con pocas cuantas variables, y es muy posible que este modelo no haya podido siquiera logrado relacionar su comportamiento.
- Si se ven los resultados de algunos modelos que no fueron seleccionados, en específico la regresión no lineal de SVR: Radial Basis Function (Tabla 5.2), se puede observar que los valores de RMSE de Buin son mucho más grandes a cualquier otra comuna. Esto mismo se repite con SVR: linear kernel (Tabla 5.3), donde el error es bastante más grande, lo que ya deja pocas ventanas para una buena predicción no lineal.
- Considerando todos estos factores, es razonable decir que para Buin no se logró captar de buena manera su consumo eléctrico, por lo que los datos de esta comuna se deberían consultar nuevamente para llegar a conclusiones robustas.

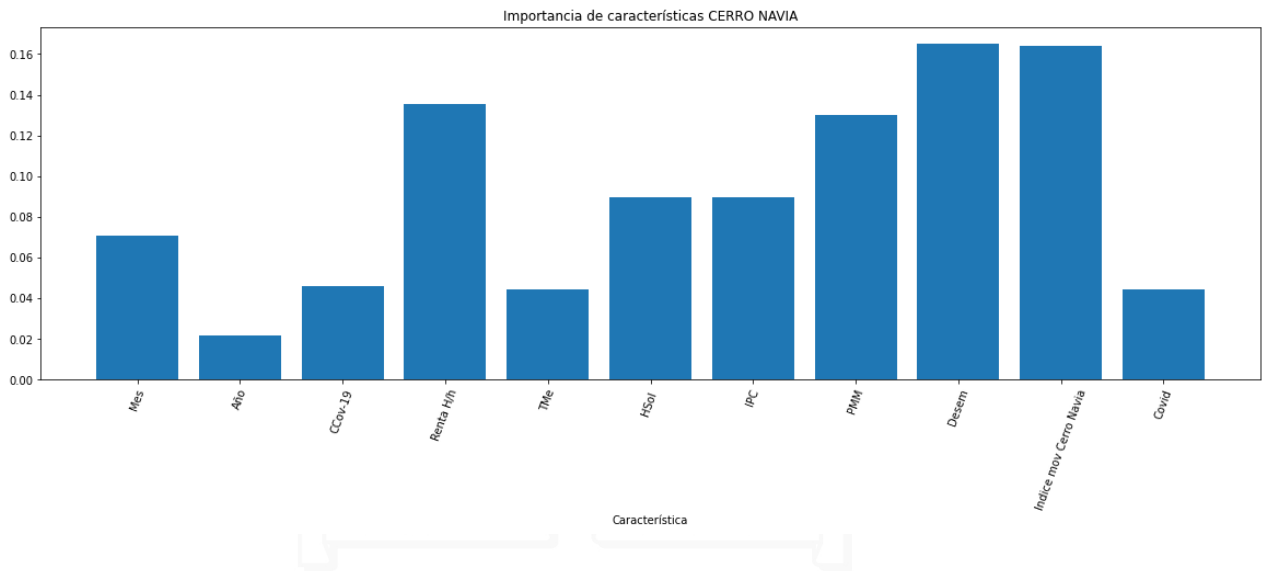
6.3. Cerro Navia

```

1                                     OLS Regression Results
2
3  Dep. Variable:                    CERRO NAVIA    R-squared:                    0.752
4  Model:                            OLS          Adj. R-squared:                0.582
5  Method:                          Least Squares  F-statistic:                  4.421
6  Date:                            Mon, 11 Jul 2022  Prob (F-statistic):          0.00377
7  Time:                            17:26:57       Log-Likelihood:              -508.09
8  No. Observations:                 28            AIC:                        1040.
9  Df Residuals:                     16            BIC:                        1056.
10 Df Model:                          11
11 Covariance Type:                  nonrobust
12
13      coef      std err          t      P>|t|      [0.025      0.975]
14 -----
15 const      1.316e+08    1.58e+07     8.342     0.000     9.81e+07    1.65e+08
16 x1         -3.159e+07    1.27e+07    -2.496     0.024    -5.84e+07    -4.76e+06
17 x2         -1.052e+08    3.76e+07    -2.800     0.013    -1.85e+08    -2.56e+07
18 x3          1.062e+07    1.19e+07     0.891     0.386    -1.47e+07     3.59e+07
19 x4          8.015e+07    3.08e+07     2.601     0.019     1.48e+07    1.45e+08
20 x5         -2.382e+07    1.89e+07    -1.261     0.226    -6.39e+07     1.62e+07
21 x6          1.464e+07    1.91e+07     0.768     0.454    -2.58e+07     5.51e+07
22 x7          9.752e+06    1.42e+07     0.689     0.501    -2.03e+07     3.98e+07
23 x8          4.656e+07    1.94e+07     2.401     0.029     5.45e+06     8.77e+07
24 x9         -4.067e+07    2.09e+07    -1.948     0.069    -8.49e+07     3.59e+06
25 x10         3.649e+07    1.84e+07     1.984     0.065    -2.5e+06     7.55e+07
26 x11        -5.145e+07    4.69e+07    -1.096     0.289    -1.51e+08     4.8e+07
27

```

Listing 6.3: OLS Cerro Navia

Figura 6.3: Importancia Variables Cerro Navia, del modelo Random Forest

Para Cerro Navia, se puede observar que tiene un $R^2 = 0,752$ y $R^2_{adj} = 0,582$, por lo que visto desde un punto de vista meramente de ajuste, no es el mejor que se podría desear ya que el sesgo que agregan las variables y la constante es suficiente como para bajar 0.2 puntos al coeficiente de correlación. Afortunadamente el valor de la prueba F corresponde a $F^{statistic} = 0,00377$, por lo que se puede afirmar que el modelo sí es significativo en conjunto, por lo que se puede analizar las pruebas t sin mayor reparo.

Las variables que más destacan corresponden a x_1 (Mes), x_2 (año), x_4 (Renta H/h) y x_8 (Precio Medio de mercado). Las variables x_9 (Desempleo) y x_{10} (Índice de Movilidad) también presentan significancias relativamente bajas, pero mayores al 5 % que se está buscando, por lo que no es posible afirmar con total certeza que sean relevantes para el modelo, pero hay una buena probabilidad de que sí lo sea. Además, si se ve el resultado del modelo de Random Forest en la Tabla 6.3, estas últimas dos variables son las más importantes para las predicciones, por lo que de todas formas sigue siendo un factor relevante para la comuna. Por esto, del modelo y de sus variables se puede concluir lo siguiente:

- A pesar de que el modelo tiene bastantes variables que son significativas, la mitad no lo son, por lo que de todas formas hay un grado de multicolinealidad entre estas variables con valores $P > |t|$ muy altos.
- Resulta lógico que se dé la situación en que el Precio de la luz y la Renta estén muy relacionadas con el consumo, en general estas dos variables deberían ir de la mano. Junto a esto, Desempleo también podría aportar a esta explicación, ya que afirma que en esta comuna existió un alto grado de desempleo, o desocupación, lo que generó una alta preocupación por el consumo.
- La variable Índice de Movilidad, como se puede ver en la matriz de correlación de la comuna en la Figura A.3, tiene una relación negativa con el consumo, lo que indica que el efecto que las cuarentenas tuvieron en la comuna, se reflejó de manera inversa en el consumo, lo que tiene mucho sentido ya que se les obligó a quedarse en casa, y gastar más energía.
- La variable x_{11} (Covid) no logró afectar drásticamente al consumo, por lo que no existió un quiebre estructural en el consumo, pero como ya se vio en la explicación anterior, el factor de quedarse en casa, efecto indirecto del COVID, sí afectó en buena medida al consumo total de la comuna.
- El hecho que las variables temporales (Mes y Año) sean importantes, deja en claro que para esta comuna existe una fuerte estacionalidad con el consumo eléctrico, la cual se mantuvo antes y después de la llegada del virus al país, debido a lo concluido para la variable dicotómica del COVID.

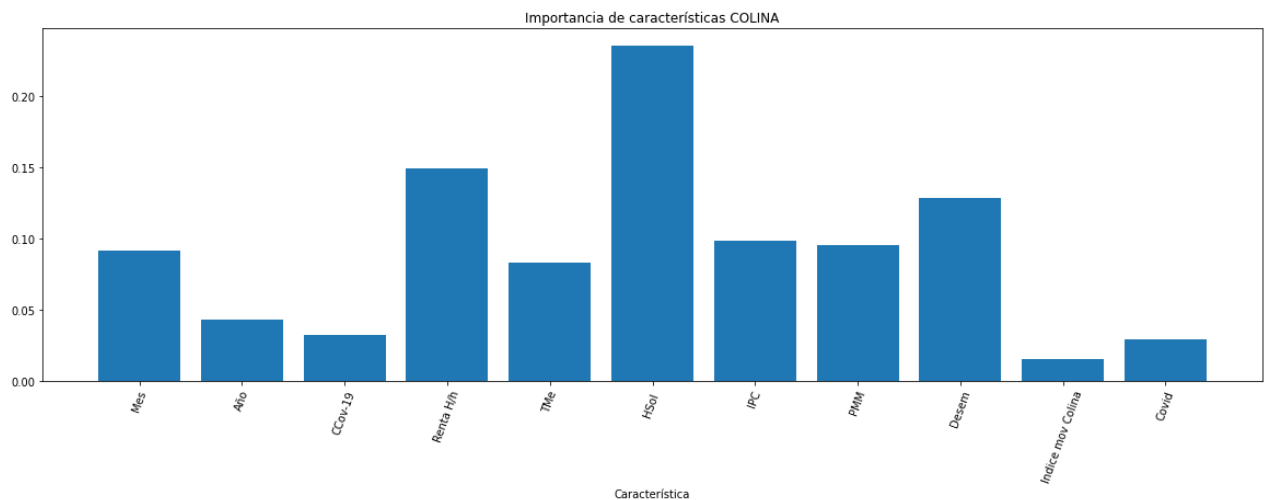
6.4. Colina

OLS Regression Results						
=====						
Dep. Variable:	COLINA		R-squared:	0.873		
Model:	OLS		Adj. R-squared:	0.785		
Method:	Least Squares		F-statistic:	9.956		
Date:	Mon, 11 Jul 2022		Prob (F-statistic):	3.29e-05		
Time:	17:26:57		Log-Likelihood:	-470.13		
No. Observations:	28		AIC:	964.3		
Df Residuals:	16		BIC:	980.3		
Df Model:	11					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]

const	2.56e+07	4.3e+06	5.952	0.000	1.65e+07	3.47e+07
x1	4.726e+06	3.31e+06	1.426	0.173	-2.3e+06	1.18e+07
x2	1.712e+07	1.01e+07	1.694	0.110	-4.3e+06	3.85e+07
x3	7.871e+06	3.52e+06	2.235	0.040	4.06e+05	1.53e+07
x4	-4.676e+06	8.21e+06	-0.569	0.577	-2.21e+07	1.27e+07
x5	-1.903e+06	4.89e+06	-0.389	0.703	-1.23e+07	8.47e+06
x6	-1.02e+07	5.03e+06	-2.027	0.060	-2.09e+07	4.68e+05
x7	7.065e+06	3.65e+06	1.937	0.071	-6.69e+05	1.48e+07
x8	-4.236e+06	5.07e+06	-0.835	0.416	-1.5e+07	6.51e+06
x9	-9.192e+06	5.42e+06	-1.696	0.109	-2.07e+07	2.3e+06
x10	4.537e+06	4.37e+06	1.039	0.314	-4.72e+06	1.38e+07
x11	-2.069e+07	1.29e+07	-1.609	0.127	-4.8e+07	6.58e+06
=====						

Listing 6.4: OLS Colina

Figura 6.4: Importancia Variables Colina, del modelo Random Forest



Para la comuna de Colina, se puede observar que ambos R^2 y R^2_{adj} son altos y la diferencia entre ellos no es tan grande como en modelos anteriores, y si sobre esto se le agrega que el valor de la prueba F es de $F^{statistic} = 3,29E - 5$, se confirma que el modelo en conjunto es significativo.

Lamentablemente sucede lo mismo que en la comuna de Alhué, ya que hay presencia de multicolinealidad debido a que únicamente una variable son significante al 5 % para el modelo, o sea, que ella explica la mayor parte de la varianza, la cual corresponde a Casos Covid 19 (x3). No obstante, hay dos variables que presentan un 6 % y 7 % que también podrían aportar análisis al modelo, que corresponden a Horas Sol (x6)

y el Índice de Precios al Consumidor (x_7), por lo que para corroborar su importancia, se puede observar la Figura 6.9, en donde se puede ver que ellas tienen una importancia por sobre el promedio, en especial Horas Sol que fue clasificada como la variable que más aporta para la predicción en un modelo no lineal, por lo que definitivamente sí es relevante para la comuna. Teniendo esto en consideración, las conclusiones obtenidas para la comuna son las siguientes:

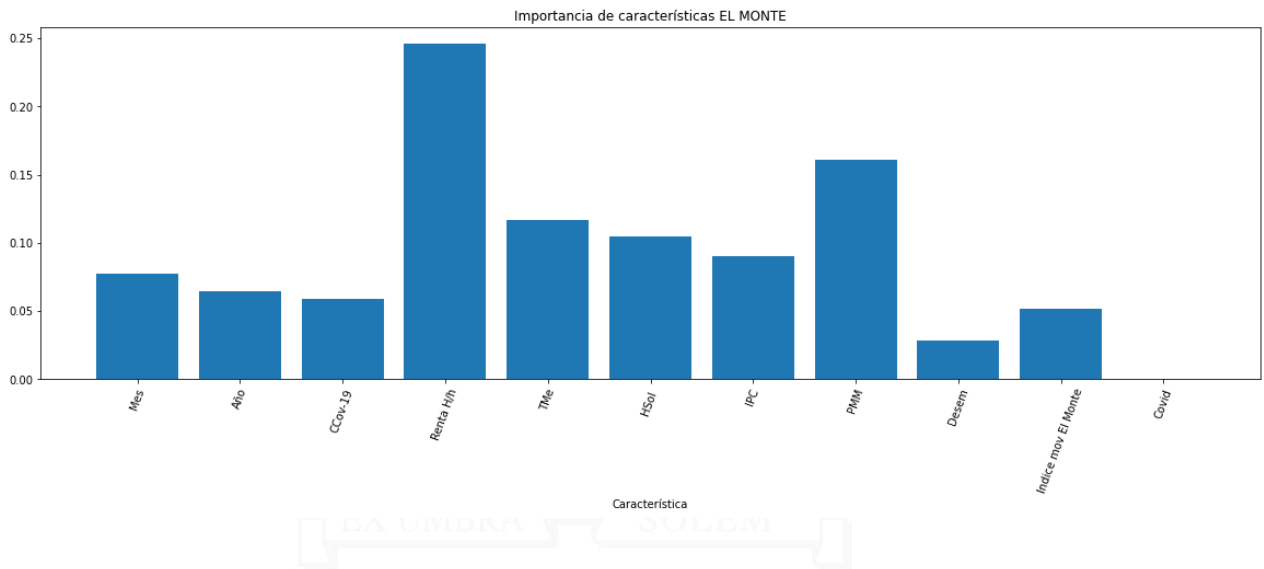
- El problema de la multicolinealidad se justifica por la misma razón que en Alhué, de antemano se pensaba que habría variables que tienen comportamientos parecidos pero claramente dependerá de cada comuna el grado de influencia de este factor.
- La variable de Casos Covid representa la evolución desde marzo del 2020, por lo que al igual que las demás variables relacionadas a la pandemia, cobra total sentido su relación. Con esto, si se observa la matriz de correlación expuesta en la Figura A.4, se puede observar que no existe una proporcionalidad fuerte entre ésta y el consumo, por lo que se puede interpretar que sí existió un cambio, pero no se podría asegurar que fuera estructural ya que no se puede afirmar la significancia de la variable dicotómica para el Covid (x_{11}).
- Finalmente, que las horas de sol, o básicamente el momento del día, haya afectado lo suficiente a Colina, siempre cobrará sentido ya que el gasto energético está altamente relacionado a la iluminación. Además, como se puede observar en su matriz de correlación, Hsol tiene una relación inversa con el consumo, por lo que se mantiene esta lógica de gastar más luz con menos horas de sol, es decir, en la noche.

6.5. El Monte

OLS Regression Results						
=====						
Dep. Variable:	EL MONTE	R-squared:		0.916		
Model:	OLS	Adj. R-squared:		0.859		
Method:	Least Squares	F-statistic:		15.95		
Date:	Mon, 11 Jul 2022	Prob (F-statistic):		1.37e-06		
Time:	17:26:57	Log-Likelihood:		-365.64		
No. Observations:	28	AIC:		755.3		
Df Residuals:	16	BIC:		771.3		
Df Model:	11					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]

const	2.301e+06	8.66e+04	26.561	0.000	2.12e+06	2.48e+06
x1	-1.43e+05	7.82e+04	-1.829	0.086	-3.09e+05	2.27e+04
x2	-2.592e+05	2.32e+05	-1.116	0.281	-7.52e+05	2.33e+05
x3	1.54e+05	7.4e+04	2.082	0.054	-2819.841	3.11e+05
x4	1.885e+05	1.91e+05	0.985	0.339	-2.17e+05	5.94e+05
x5	-3.364e+04	1.16e+05	-0.291	0.775	-2.79e+05	2.12e+05
x6	2.525e+05	1.17e+05	2.150	0.047	3574.598	5.01e+05
x7	6.84e+04	8.74e+04	0.783	0.445	-1.17e+05	2.54e+05
x8	2.648e+05	1.2e+05	2.215	0.042	1.14e+04	5.18e+05
x9	-6.663e+04	1.29e+05	-0.517	0.612	-3.4e+05	2.06e+05
x10	1.123e+05	7.37e+04	1.524	0.147	-4.39e+04	2.69e+05
x11	-5191.7795	2.55e+05	-0.020	0.984	-5.45e+05	5.35e+05
=====						

Listing 6.5: OLS El Monte

Figura 6.5: Importancia Variables El Monte, del modelo Random Forest

La regresión ajustada para El Monte presenta un $R^2 = 0,916$ y un $R^2_{adj} = 0,859$, por lo que está muy bien ajustado considerando que la diferencia entre ambos índices no es de gran magnitud. Además, el resultado de la prueba F indica que el modelo sí es significativo en conjunto debido a un muy bajo valor $F^{statistic} = 1,37E - 6$.

Ahora, mirando las variables, destacan Mes (x_1), Casos Covid 19 (x_3), Hora de Sol (x_6), Precio Medio del Mercado (x_8). Todo esto da indica que puede existir multicolinealidad entre las demás variables, pero no en un grado muy preocupante debido la cantidad de variables que sí son significativas. Los resultados de Random Forest no aportarán en gran medida, ya que las importancias se encuentran en general bajas, pero concuerda con la variable que recibió el mayor valor $P > |t|$, correspondiente a x_{11} . De esto se puede decir lo siguiente:

- Al igual que la comuna anterior, Casos Covid fue un factor importante en el consumo de este período, teniendo una correlación mayor a 0.5 (Figura A.24), por lo que en efecto su consumo aumentó después de la pandemia, pero siempre existió una especial preocupación por el precio del mismo, y el horario en el que más se gastaba electricidad.
- Nuevamente Mes pasa a ser relevante, captando este elemento estacional en el consumo de la comuna. De aquí se puede sacar la conclusión de que la comuna en general tiene un patrón muy marcado por elementos medioambientales ya que Horas de sol también es relevante.
- Como ya se mencionó, Covid no generó un quiebre estructural del consumo eléctrico para la comuna. Esto puede sonar contradictorio con lo visto en el primero punto, pero este quiebre representará al comportamiento de todas las variables, y si ocurrió este quiebre que generó cambios transversales. Esto no ocurrió, y como se puede observar en los resultados de Random Forest, tampoco aportó para la predicción.

6.6. La Pintana

OLS Regression Results			
=====			
Dep. Variable:	LA PINTANA	R-squared:	0.783
Model:	OLS	Adj. R-squared:	0.634
Method:	Least Squares	F-statistic:	5.244
Date:	Mon, 11 Jul 2022	Prob (F-statistic):	0.00153
Time:	17:26:57	Log-Likelihood:	-425.17

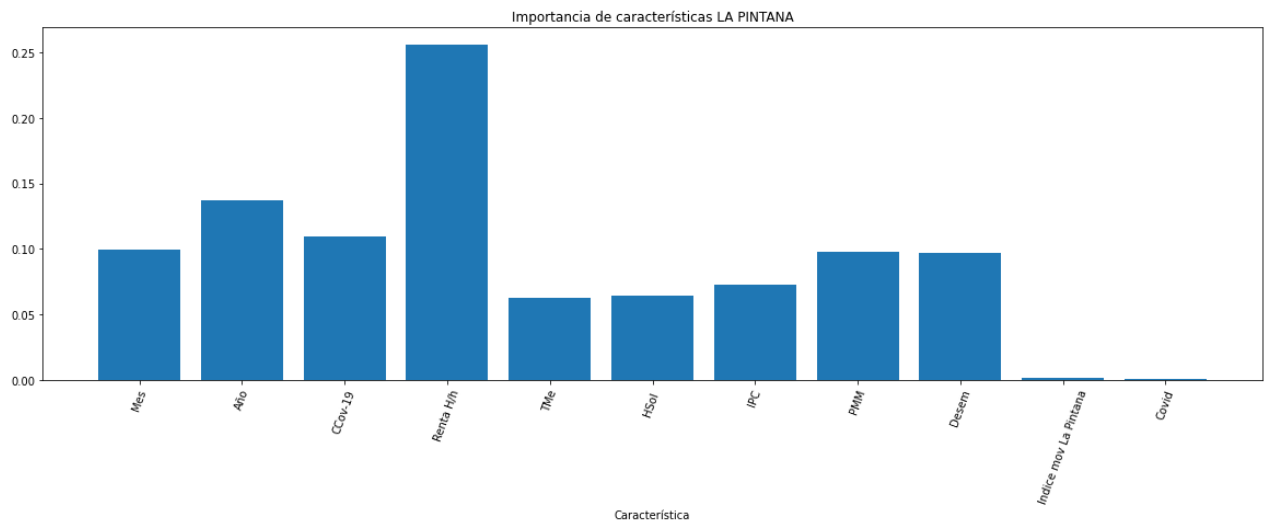
```

8  No. Observations:      28  AIC:      874.3
9  Df Residuals:         16  BIC:      890.3
10 Df Model:             11
11 Covariance Type:      nonrobust
12 =====
13      coef      std err      t      P>|t|      [0.025      0.975]
14 -----
15 const      1.519e+07    1.16e+06    13.071    0.000    1.27e+07    1.77e+07
16 x1          -9.18e+05    6.54e+05    -1.403    0.180    -2.3e+06    4.69e+05
17 x2          6.249e+05    2.03e+06     0.309    0.762    -3.67e+06    4.92e+06
18 x3          1.05e+06    7.15e+05     1.469    0.161    -4.65e+05    2.56e+06
19 x4          2.053e+05    1.61e+06     0.128    0.900    -3.2e+06    3.61e+06
20 x5          -6.54e+05    1.05e+06    -0.624    0.541    -2.87e+06    1.57e+06
21 x6          3.77e+05    1.02e+06     0.369    0.717    -1.79e+06    2.54e+06
22 x7          3.66e+05    7.45e+05     0.491    0.630    -1.21e+06    1.95e+06
23 x8          4.307e+05    1.01e+06     0.426    0.676    -1.71e+06    2.57e+06
24 x9          -1.193e+05    1.59e+06    -0.075    0.941    -3.49e+06    3.26e+06
25 x10         -4.677e+05    8.56e+05    -0.547    0.592    -2.28e+06    1.35e+06
26 x11         9.679e+05    3.54e+06     0.273    0.788    -6.54e+06    8.47e+06
27 =====

```

Listing 6.6: OLS La Pintana

Figura 6.6: Importancia Variables La Pintana, del modelo Random Forest



El modelo para La Pintana entregó resultados bastante aceptables por los coeficientes de determinación, siendo estos $R^2 = 0,783$ y $R^2_{adj} = 0,634$. El ajuste es decente para lo que se necesita lograr pero siempre es necesario asegurar esto con la prueba F, la cual entrega un valor de $F^{statistic} = 0,00153$, por lo que el modelo efectivamente es significativo en conjunto. No obstante, se puede observar que ninguna de las variables se encuentra dentro del umbral del 5 % para el valor de $P > |t|$, teniendo el mejor para la variable mes, que corresponde al 18 %, que es muy alto para afirmarlo.

Ahora, si se observa el resultado del modelo de Random Forest, se considera como la variable mas importante a Renta H/h al momento de realizar predicciones, y todas las demás tienen grados de importancia no nulos. Por esto, se puede concluir que la regresión OLS no logró captar los apportes de la base de datos para entregar explicaciones contundentes, por lo que el análisis se basará en los resultados de Random Forest:

- Como ya se mencionó, Random Forest entrega las importancias de las variables, de donde destacó por sobre las demás Renta H/h. Covid binaria e Índice de movilidad presentaron importancias nulas, y las variables restantes obtuvieron valores rondando el 0.1, por lo que se entiende que el modelo en conjunto es importante para la comuna, pero el sueldo promedio comunal es el factor más importante

a la hora de decidir su consumo, lo que se puede explicar debido a una baja empleabilidad, o bajos sueldos en general.

6.7. Lo Prado

1 OLS Regression Results

2 =====

3 Dep. Variable: LO PRADO R-squared: 0.813

4 Model: OLS Adj. R-squared: 0.684

5 Method: Least Squares F-statistic: 6.321

6 Date: Mon, 11 Jul 2022 Prob (F-statistic): 0.000538

7 Time: 17:26:57 Log-Likelihood: -453.02

8 No. Observations: 28 AIC: 930.0

9 Df Residuals: 16 BIC: 946.0

10 Df Model: 11

11 Covariance Type: nonrobust

12 =====

13

14

15

16

17

18

19

20

21

22

23

24

25

26

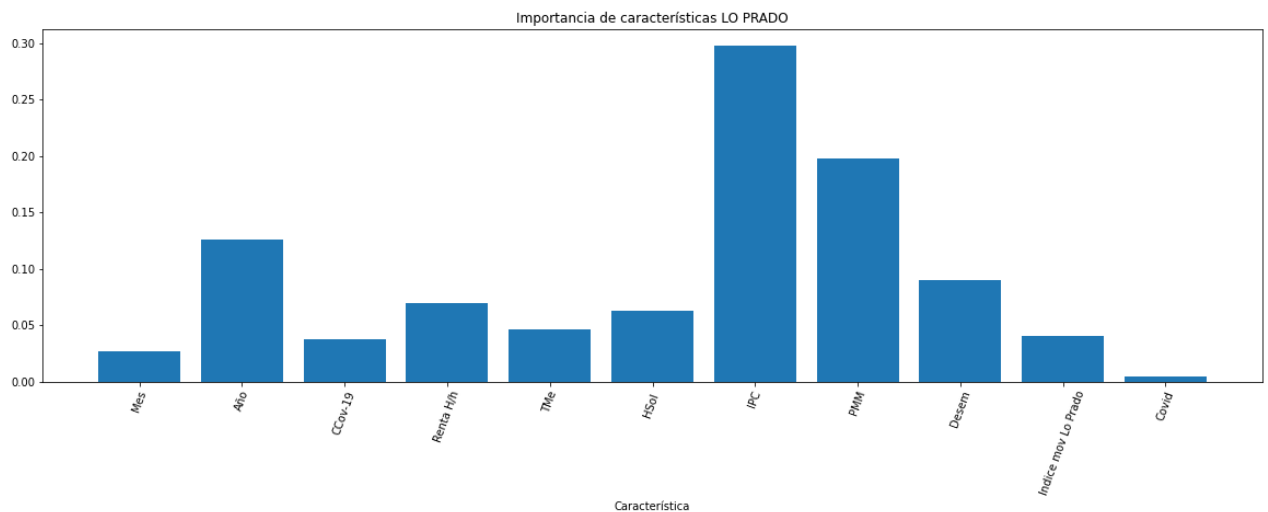
27

	coef	std err	t	P> t	[0.025	0.975]
const	3.251e+06	3.11e+06	1.044	0.312	-3.35e+06	9.85e+06
x1	-9.053e+05	1.81e+06	-0.501	0.623	-4.73e+06	2.92e+06
x2	5.919e+06	6.06e+06	0.976	0.343	-6.93e+06	1.88e+07
x3	5.729e+06	2.34e+06	2.453	0.026	7.77e+05	1.07e+07
x4	-2.592e+06	4.72e+06	-0.549	0.591	-1.26e+07	7.42e+06
x5	-1.668e+06	2.74e+06	-0.609	0.551	-7.47e+06	4.14e+06
x6	-2.67e+05	2.7e+06	-0.099	0.922	-5.99e+06	5.45e+06
x7	9.441e+05	2.06e+06	0.457	0.653	-3.43e+06	5.32e+06
x8	4.502e+06	2.73e+06	1.646	0.119	-1.3e+06	1.03e+07
x9	-1.939e+07	4.92e+06	-3.938	0.001	-2.98e+07	-8.95e+06
x10	-6.948e+06	2.27e+06	-3.066	0.007	-1.18e+07	-2.14e+06
x11	3.22e+07	9.48e+06	3.397	0.004	1.21e+07	5.23e+07

=====

Listing 6.7: OLS Lo Prado

Figura 6.7: Importancia Variables Lo Prado, del modelo Random Forest



Para Lo Prado, se puede observar de inmediato que presenta un muy buen coeficiente de determinación, siendo $R^2 = 0,813$, pero tiene una variación considerable con su versión ajustada, la cual corresponde $R^2_{adj} = 0,684$

por lo que en realidad no todas las variables tienen un efecto muy grande en el rendimiento del modelo. De todas formas si se observa el resultado de la prueba F, $F^{statistic} = 0,000538$, se puede afirmar que el modelo sí es significativo.

Las variables que más destacan corresponden a Casos Covid (x_3), Desempleo (x_9), Índice de Movilidad (x_{10}) y la variable binaria para el Covid 19 (x_{11}). Ahora, el modelo de Random Forest no entregó resultados que puedan aportar mucho a esta significancia ya que las más importantes en este modelo fueron IPC, PMM y Año, bastante distinto a lo que la regresión OLS entregó, pero como se sabe esto puede pasar ya que son modelos completamente distintos, y el gráfico de RF siempre mostrará la capacidad de predicción de las variables para todo el modelo. Junto a esto, las conclusiones para la comuna son las siguientes:

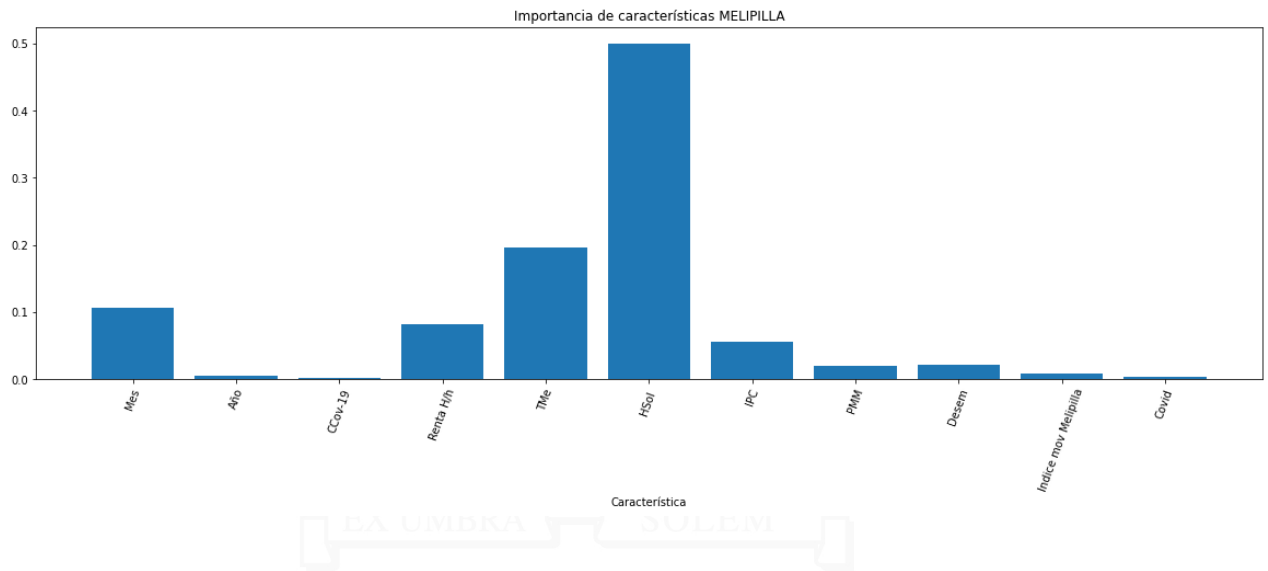
- Para esta comuna, todas las variables que se consideraron significantes para el modelo tienen sinergia, ya que el mismo virus generó desempleo transversal en todas las comunas, por lo que es normal que dentro de algunas haya afectado en mayor o menor medida. Este efecto también se ve reflejado en el índice de movilidad, el cual según la matriz de correlación (Figura A.7) tiene una relación positiva con el consumo, por lo que éste aumentó en la pandemia.
- Todo esto se complementa directamente con la significancia de la variable dicotómica para el Covid 19 (x_{11}). Como se ha dicho hasta ahora, esta variable guarda una especial importancia ya que capta el quiebre estructural de la pandemia, y Lo Prado es la primera comuna en la que efectivamente tuvo un efecto más notorio en el consumo eléctrico, lo que se justifica con la significancia de las otras variables que tienen una estrecha relación con los efectos colaterales de la pandemia.

6.8. Melipilla

OLS Regression Results						
=====						
Dep. Variable:	MELIPILLA		R-squared:	0.848		
Model:	OLS		Adj. R-squared:	0.744		
Method:	Least Squares		F-statistic:	8.142		
Date:	Mon, 11 Jul 2022		Prob (F-statistic):	0.000118		
Time:	17:26:57		Log-Likelihood:	-427.08		
No. Observations:	28		AIC:	878.2		
Df Residuals:	16		BIC:	894.1		
Df Model:	11					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]

const	1.942e+07	1.17e+06	16.572	0.000	1.69e+07	2.19e+07
x1	-1.287e+06	7e+05	-1.838	0.085	-2.77e+06	1.97e+05
x2	-1.984e+06	2.13e+06	-0.932	0.365	-6.5e+06	2.53e+06
x3	4.69e+05	6.79e+05	0.691	0.499	-9.7e+05	1.91e+06
x4	-2.733e+05	1.73e+06	-0.158	0.876	-3.94e+06	3.39e+06
x5	7.947e+05	1.12e+06	0.711	0.487	-1.58e+06	3.16e+06
x6	1.353e+06	1.1e+06	1.226	0.238	-9.87e+05	3.69e+06
x7	1.376e+05	7.84e+05	0.175	0.863	-1.53e+06	1.8e+06
x8	2.305e+06	1.08e+06	2.125	0.050	5027.482	4.6e+06
x9	-5.59e+05	1.31e+06	-0.427	0.675	-3.33e+06	2.21e+06
x10	-4.086e+05	8.75e+05	-0.467	0.647	-2.26e+06	1.45e+06
x11	8.842e+05	3.56e+06	0.248	0.807	-6.66e+06	8.43e+06
=====						

Listing 6.8: OLS Melipilla

Figura 6.8: Importancia Variables Melipilla, del modelo Random Forest

El modelo ajustado para la comuna de Melipilla presenta un $R^2 = 0,848$ y un $R^2_{adj} = 0,744$. Como la variación de ambos es muy pequeña, se puede afirmar que las variables importantes para el primer resultado, lo siguen siendo para su versión ajustada. Además, el modelo también es significativo en conjunto debido al resultado de la prueba F, $F^{statistic} = 0,000118$.

No obstante, existe solo una variable con significancia al 5 %, que corresponde al Precio Medio de Mercado, por lo que definitivamente el modelo presenta multicolinealidad, ya que las demás variables no se encuentran cercanas a este valor de corte, a excepción de Mes, que presenta un 8 %. Ahora, si se observan los resultados de los coeficientes de Random Forest, éste también le entrega la mayor parte de importancia a solo una variable, pero ésta es x_6 , Horas Sol, por lo que se puede concluir lo siguiente:

- Si bien el modelo presenta multicolinealidad, entregó resultados muy buenos para el nivel de significancias de las otras variables. Además, su matriz de correlación no presenta valores muy altos entre variables, por lo que todo parece indicar que se debió recolectar más datos para poder realizar un ajuste más real, debido a que las observaciones por comuna, al mensualizarlas, no son tantas después de dividir la base de datos en entrenamiento y prueba.
- El hecho de que Random Forest también haya dado luces de multicolinealidad, es otra razón más para pensar que esta baja importancia es debido a la falta de datos para entrenar.

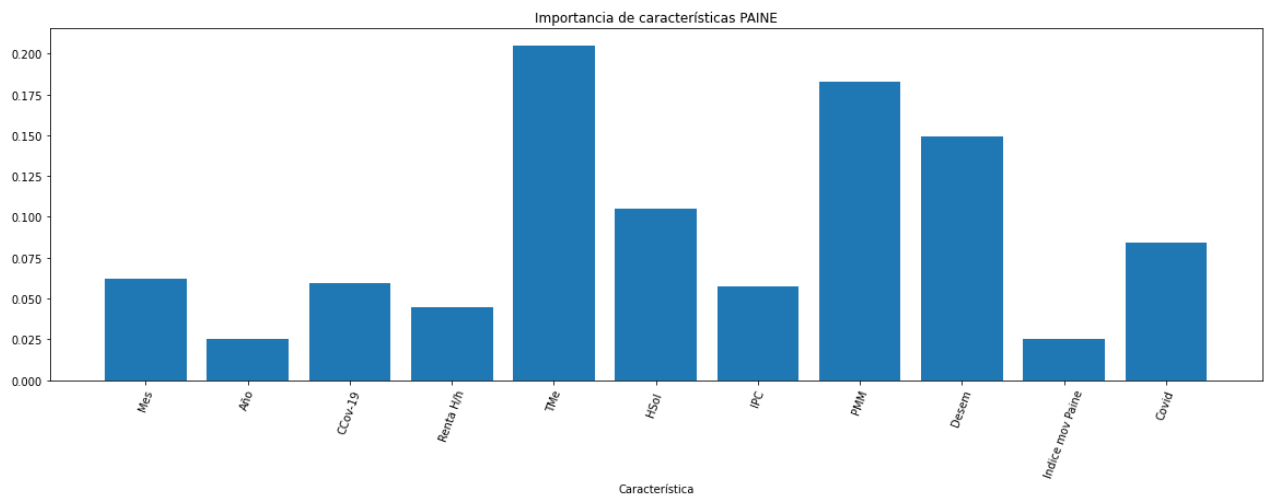
6.9. Paine

OLS Regression Results						
=====						
Dep. Variable:	PAINE		R-squared:	0.369		
Model:	OLS		Adj. R-squared:	-0.065		
Method:	Least Squares		F-statistic:	0.8498		
Date:	Mon, 11 Jul 2022		Prob (F-statistic):	0.600		
Time:	17:26:57		Log-Likelihood:	-461.13		
No. Observations:	28		AIC:	946.3		
Df Residuals:	16		BIC:	962.2		
Df Model:	11					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]

const	1.418e+07	3.98e+06	3.561	0.003	5.74e+06	2.26e+07
x1	-1.35e+06	2.39e+06	-0.564	0.581	-6.42e+06	3.72e+06
x2	-1.617e+06	7.02e+06	-0.230	0.821	-1.65e+07	1.33e+07
x3	-5.947e+05	2.38e+06	-0.250	0.806	-5.64e+06	4.45e+06
x4	-8.905e+05	5.76e+06	-0.155	0.879	-1.31e+07	1.13e+07
x5	1.514e+06	3.92e+06	0.386	0.705	-6.8e+06	9.83e+06
x6	-4.197e+05	3.87e+06	-0.109	0.915	-8.62e+06	7.78e+06
x7	-7.031e+05	2.65e+06	-0.266	0.794	-6.31e+06	4.91e+06
x8	1.135e+06	3.63e+06	0.313	0.759	-6.56e+06	8.83e+06
x9	1.765e+06	4.09e+06	0.432	0.672	-6.9e+06	1.04e+07
x10	-3.835e+06	3.24e+06	-1.184	0.254	-1.07e+07	3.03e+06
x11	7.883e+06	1.21e+07	0.652	0.524	-1.78e+07	3.35e+07
=====						

Listing 6.9: OLS Paine

Figura 6.9: Importancia Variables Paine, del modelo Random Forest



Como puede se puede observar, en este caso ocurrió algo parecido a la comuna de Buin, en el que directamente hay indicadores que por distintas condiciones que no se lograron satisfacer, entregaron resultados que en condiciones normales no deberían ser posibles, como un valor de $R^2_{adj} = -0,065$, y prueba F de $F^{statistic} = 0,8498$, por lo que el modelo no se encuentra en condiciones de ser analizado correctamente, o dicho de otra manera, no se pueden obtener conclusiones robustas a partir de la regresión OLS. No obstante, se tienen los resultados que entrega el modelo no lineal de Random Forest, los cuales indican lo siguiente:

- Las tres variables más importantes para la predicción corresponden a Temperatura Media, Precio Medio de Mercado y Desempleo, por lo que de toda dformas habla de la constante preocupación de la comuna por consumir cuando sea necesario, debido a las temperaturas altas/bajas y el precio de la misma.

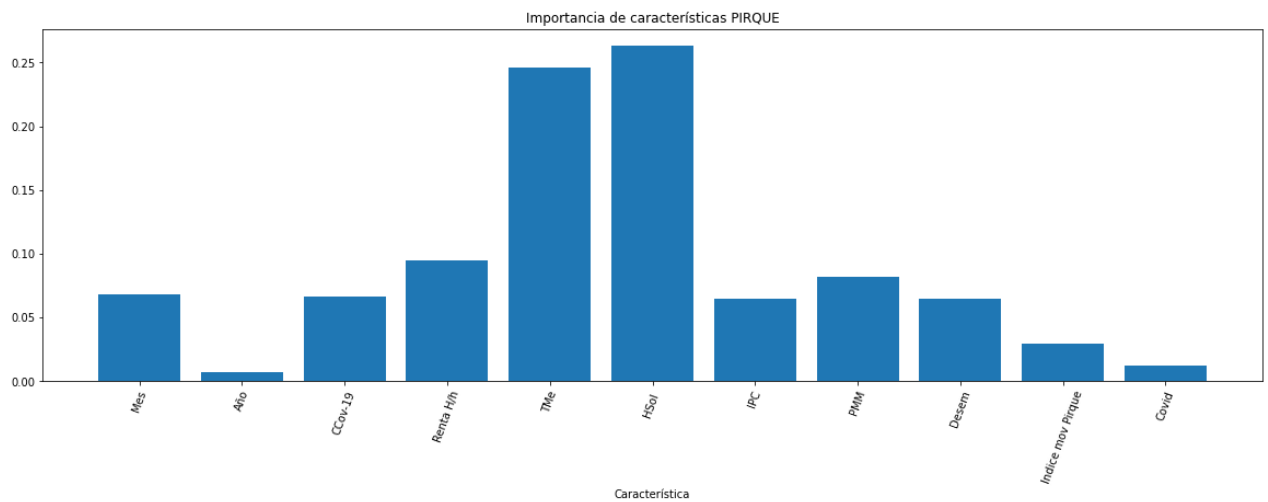
6.10. Pirque

OLS Regression Results						
=====						
Dep. Variable:	PIRQUE	R-squared:	0.856			
Model:	OLS	Adj. R-squared:	0.756			
Method:	Least Squares	F-statistic:	8.621			
Date:	Mon, 11 Jul 2022	Prob (F-statistic):	8.26e-05			
Time:	17:26:57	Log-Likelihood:	-430.53			
No. Observations:	28	AIC:	885.1			
Df Residuals:	16	BIC:	901.0			
Df Model:	11					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]

const	4.337e+06	1.39e+06	3.121	0.007	1.39e+06	7.28e+06
x1	-1.014e+06	8.02e+05	-1.264	0.224	-2.72e+06	6.86e+05
x2	-5.414e+06	2.48e+06	-2.180	0.045	-1.07e+07	-1.5e+05
x3	1.419e+06	7.68e+05	1.847	0.083	-2.1e+05	3.05e+06
x4	3.763e+06	2e+06	1.882	0.078	-4.76e+05	8e+06
x5	1.978e+06	1.2e+06	1.654	0.118	-5.57e+05	4.51e+06
x6	8.144e+05	1.2e+06	0.679	0.507	-1.73e+06	3.36e+06
x7	-1.759e+05	8.9e+05	-0.198	0.846	-2.06e+06	1.71e+06
x8	9.239e+05	1.23e+06	0.752	0.463	-1.68e+06	3.53e+06
x9	-3.523e+05	1.49e+06	-0.236	0.816	-3.51e+06	2.81e+06
x10	-3.403e+05	1.13e+06	-0.302	0.767	-2.73e+06	2.05e+06
x11	5.735e+05	4.23e+06	0.136	0.894	-8.39e+06	9.54e+06
=====						

Listing 6.10: OLS Pirque

Figura 6.10: Importancia Variables Pirque, del modelo Random Forest



El modelo de Pirque presenta un muy buen ajuste, siendo $R^2 = 0,856$, y al corregirlo por la cantidad de variables, este baja a su valor ajustado de $R^2_{adj} = 0,756$, lo que no es mucho por lo que se puede afirmar

que se mantuvo el número de variables relevantes para el ajuste. También se puede afirmar esto desde el punto de vista general del modelo, ya que el valor de la prueba F es suficientemente bajo como para afirmar que el modelo es significativo, entregando un $F^{statistic} = 8,26E - 5$. Ahora, se puede observar que existe solo una variable con significancia al 5 %, la cual corresponde a x_2 , que representa el año. No obstante, también presentan significancias relativamente bajas las variables x_3 (modela los Casos por COVID 19) con sólo un 8 %, y x_4 (representa la Renta H/h) con un 7 %. Ahora, si se miran los resultados de Random Forest, este consideró como las variables más importantes: Temperatura Media y Horas de Sol, lo que cobra sentido dentro para RF ya que son características que están estrechamente relacionadas de por sí. Considerando todos estos factores, se puede interpretar lo siguiente:

- Dado que año es la variable con menor valor $P > |t|$, se puede inferir que para la comuna de Pirque el consumo eléctrico fue aumentando a lo largo de los años, pero no drásticamente como para notar el efecto de la pandemia o un quiebre estructural. Además, si se incluyen Casos COVID y Renta H/h, se puede llegar a la conclusión de que el virus efectivamente afectó al consumo, en especial el año 2020 en el cual comenzaron a registrarse casos por el virus.
- El hecho de que Caso COVID tenga una significancia tan baja en comparación ala variable binaria x_{11} , se puede ver en matriz de correlación de la comuna (Figura A.10), en la cual se puede observar que ellas no presentan tendencias lineales muy marcadas con el consumo, lo que quiere decir que si bien el número de casos puede haber afectado, estos no fueron lo suficientemente importantes, posiblemente por condiciones laborales de distinta índole.

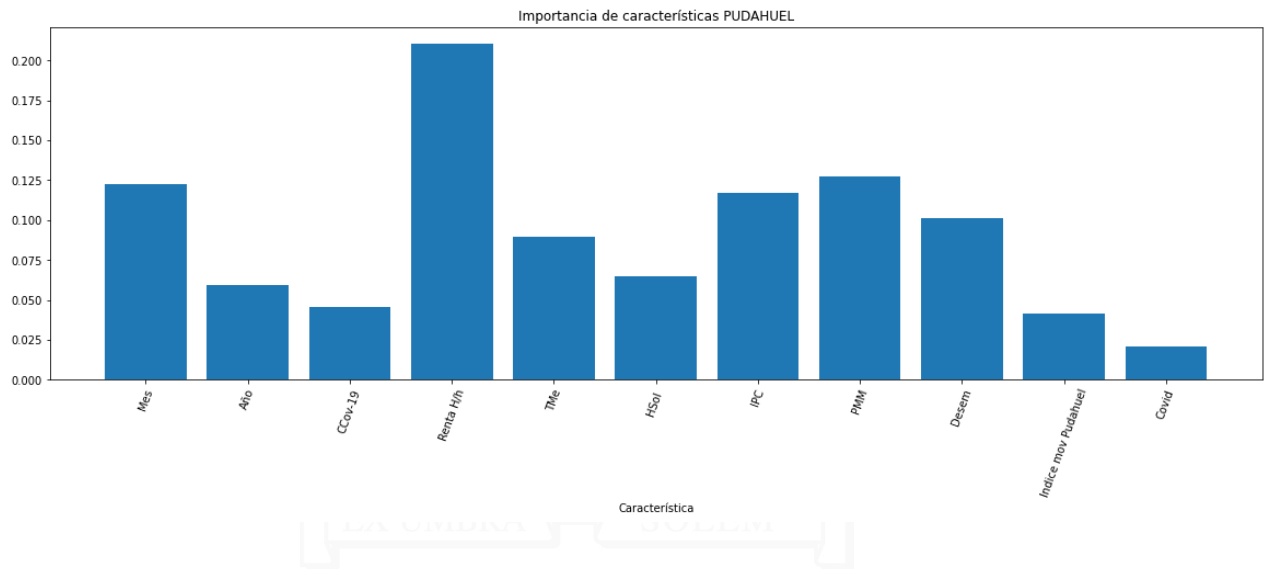
6.11. Pudahuel

```

1      OLS Regression Results
2      =====
3      Dep. Variable:          PUDAHUEL      R-squared:          0.618
4      Model:                  OLS           Adj. R-squared:     0.355
5      Method:                 Least Squares  F-statistic:       2.350
6      Date:                   Mon, 11 Jul 2022  Prob (F-statistic): 0.0587
7      Time:                   17:26:57      Log-Likelihood:    -519.80
8      No. Observations:       28           AIC:              1064.
9      Df Residuals:           16           BIC:              1080.
10     Df Model:                11
11     Covariance Type:        nonrobust
12     =====
13     coef    std err          t      P>|t|      [0.025    0.975]
14     -----
15     const      1.182e+08    3.45e+07      3.423    0.003      4.5e+07    1.91e+08
16     x1          1.117e+07    1.94e+07     0.575    0.573     -3e+07    5.23e+07
17     x2         -2.223e+06    5.71e+07    -0.039    0.969    -1.23e+08    1.19e+08
18     x3          1.045e+07    1.84e+07     0.567    0.578    -2.86e+07    4.95e+07
19     x4          4.995e+05    4.66e+07     0.011    0.992    -9.83e+07    9.93e+07
20     x5          7.189e+07    2.85e+07     2.522    0.023     1.15e+07    1.32e+08
21     x6         -7.671e+07    2.89e+07    -2.656    0.017    -1.38e+08    -1.55e+07
22     x7          2.607e+07    2.16e+07     1.207    0.245    -1.97e+07    7.18e+07
23     x8          7.995e+07    2.94e+07     2.718    0.015     1.76e+07    1.42e+08
24     x9         -6.015e+07    3.25e+07    -1.852    0.083    -1.29e+08     8.7e+06
25     x10         -1.352e+07    3.74e+07    -0.362    0.722    -9.28e+07    6.57e+07
26     x11         -2.356e+06    1.05e+08    -0.022    0.982    -2.25e+08    2.21e+08
27     =====

```

Listing 6.11: OLS Pudahuel

Figura 6.11: Importancia Variables Pudahuel, del modelo Random Forest

En Pudahuel puede observarse que ocurre el efecto de la efectividad para representar el modelo de algunas variables, ya que el valor ajustado varía mucho en relación a su valor normal, pasando desde un $R^2 = 0,618$ a $R^2_{adj} = 0,355$, lo que da luces que el modelo no es significativo. Además, el valor de la prueba F es $F^{statistic} = 0,0587$, justo por encima del 5 %, por lo que no se podría afirmar con certeza que la regresión OLS es significativa en su conjunto, por lo que realizar un análisis de este modelo no será muy acertado. En cambio, se tomarán los resultados de Random Forest presentados en la Figura 6.11, ya que a diferencia de otras comunas, para Pudahuel se presentan importancias relevantes en magnitud, por lo que su análisis está bien fundado.

De aquí, se pueden tomar como los datos más importantes: Mes, Renta H/h, Precio Medio de Mercado, IPC y Desempleo. Al igual que en algunas comunas, las variables económicas resaltan todas juntas por motivos lógicos, y no es de sorpresa que, teniendo esto en consideración, la más importante sea Renta. Por esto, para Pudahuel se puede concluir lo siguiente:

- Como se mencionó, todas las variables económicas resaltan dentro de Random Forest, por lo que la gran parte del consumo se puede explicar con el comportamiento de estas variables. En resumen, la comuna es muy consciente de su consumo debido a que está fuertemente ligado al precio de la misma y al dinero que las familias pueden generar mensualmente.
- Se puede apreciar de la misma figura que ninguna variable relacionada directamente con el Covid 19 tuvo alguna importancia relevante sobre las predicciones por lo que se puede decir que Pudahuel tampoco se vio muy afectada.

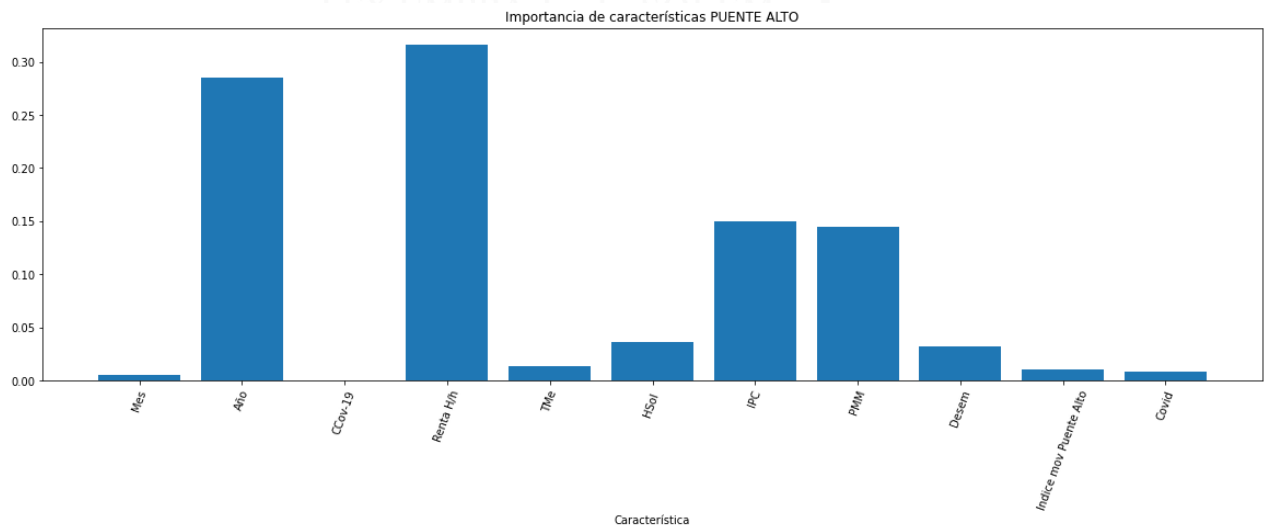
6.12. Puente Alto

OLS Regression Results			
=====			
Dep. Variable:	PUENTE ALTO	R-squared:	0.915
Model:	OLS	Adj. R-squared:	0.857
Method:	Least Squares	F-statistic:	15.69
Date:	Mon, 11 Jul 2022	Prob (F-statistic):	1.53e-06
Time:	17:26:57	Log-Likelihood:	-444.00
No. Observations:	28	AIC:	912.0
Df Residuals:	16	BIC:	928.0
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
const	5.362e+06	1.53e+06	3.496	0.003	2.11e+06	8.61e+06
x1	-1.692e+06	1.3e+06	-1.303	0.211	-4.44e+06	1.06e+06
x2	-6.437e+06	3.81e+06	-1.691	0.110	-1.45e+07	1.63e+06
x3	8.115e+05	1.21e+06	0.672	0.511	-1.75e+06	3.37e+06
x4	4.388e+06	3.11e+06	1.410	0.178	-2.21e+06	1.1e+07
x5	3.115e+06	1.98e+06	1.577	0.134	-1.07e+06	7.3e+06
x6	-3.213e+06	2e+06	-1.609	0.127	-7.45e+06	1.02e+06
x7	5.418e+06	1.44e+06	3.773	0.002	2.37e+06	8.46e+06
x8	1.949e+06	1.97e+06	0.989	0.338	-2.23e+06	6.13e+06
x9	-7.945e+05	2.19e+06	-0.362	0.722	-5.44e+06	3.85e+06
x10	2.891e+05	1.07e+06	0.270	0.791	-1.98e+06	2.56e+06
x11	-1.652e+06	4.55e+06	-0.363	0.721	-1.13e+07	7.99e+06

Listing 6.12: OLS Puente Alto

Figura 6.12: Importancia Variables Puente Alto, del modelo Random Forest



Puente Alto presenta muy buenos resultados por los 3 indicadores que se están utilizando como base. Primero, tiene un muy buen ajuste, entregando un $R^2 = 0,915$, el cual si se contrasta con su versión ajustada, $R^2_{adj} = 0,857$, se puede llegar a la conclusión de que la mayor parte de las variables conservan su capacidad explicativa después de ser analizadas por separado. Además, el valor de la prueba F es bastante bueno, entregando un $F^{statistic} = 1,53E - 6$, por lo que se puede afirmar sin ninguna duda que el modelo es significativo en conjunto.

Ahora, lamentablemente los resultados no son muy comprometedores al analizar la significancia de cada variable, ya que solamente una presenta un valor $P > |t|$ menor al 5 % que se está buscando, por lo que definitivamente se puede afirmar que la regresión OLS presenta multicolinearidad en sus variables, excepto con x7, que corresponde al Índice de Precios al Consumidor. Ahora, si se observa Random Forest, entrega resultados bastante distintos, definiendo de inmediato como importantes las variables Año y Renta H/h. Ahora, IPC también quedó definida como relevante para el modelo, pero el valor de esta importancia es casi la mitad de Renta H/h.

Como ya ha pasado en modelos anteriores, aquí sucede un fenómeno muy parecido a la multicolinearidad, en donde la importancia de muchas variables es despreciable, o muy cercana a 0, por lo que pocas variables explican el comportamiento de todo el modelo. Por esto, de la comuna Puente Alto se puede concluir o siguiente:

- Lamentablemente como existe el problema de multicolinealidad, indica que se deberían realizar ajustes al modelo para que esto no suceda. No obstante, el hecho que el IPC resalte mucho quiere decir que en general este indicador afecta a la comuna transversalmente, muy posiblemente porque los sueldos sean bajos en comparación a las demás comunas, o las municipalidades no administren suficientemente bien sus fondos como para poder entregarle recursos básicos a la mayoría de los habitantes.

6.13. Renca

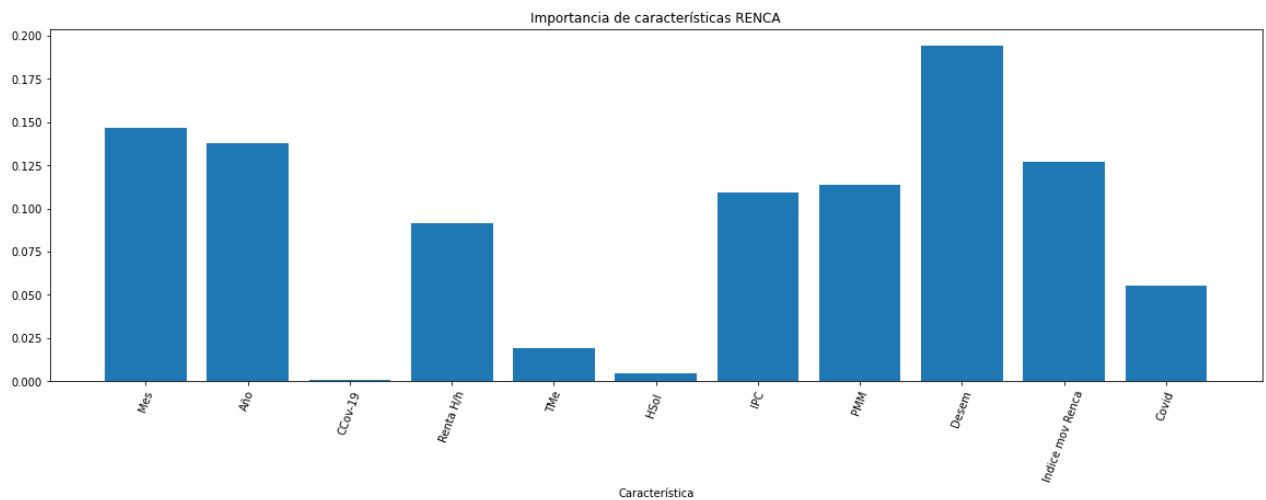
```

1      OLS Regression Results
2
3      Dep. Variable:          RENCA      R-squared:          0.862
4      Model:                OLS      Adj. R-squared:      0.767
5      Method:              Least Squares      F-statistic:        9.075
6      Date:                Mon, 11 Jul 2022      Prob (F-statistic):  5.97e-05
7      Time:                17:26:57      Log-Likelihood:     -515.72
8      No. Observations:      28      AIC:                1055.
9      Df Residuals:          16      BIC:                1071.
10     Df Model:              11
11     Covariance Type:      nonrobust
12
13     =====
14     coef      std err      t      P>|t|      [0.025      0.975]
15     -----
16     const      3.572e+07      2.2e+07      1.620      0.125      -1.1e+07      8.25e+07
17     x1          1.864e+07      1.68e+07      1.111      0.283      -1.69e+07      5.42e+07
18     x2          2.35e+08      5.02e+07      4.677      0.000      1.29e+08      3.42e+08
19     x3          1.187e+07      1.86e+07      0.639      0.532      -2.75e+07      5.13e+07
20     x4         -7.773e+07      4.35e+07     -1.789      0.093      -1.7e+08      1.44e+07
21     x5          3.518e+07      2.54e+07      1.383      0.186      -1.87e+07      8.91e+07
22     x6         -4.32e+07      2.61e+07     -1.657      0.117      -9.85e+07      1.21e+07
23     x7          7.131e+07      1.88e+07      3.789      0.002      3.14e+07      1.11e+08
24     x8         -2.854e+07      2.54e+07     -1.122      0.278      -8.25e+07      2.54e+07
25     x9         -4.574e+07      3.04e+07     -1.503      0.152      -1.1e+08      1.88e+07
26     x10         -1.72e+07      2.31e+07     -0.744      0.468      -6.62e+07      3.18e+07
27     x11         -2.499e+06      6.6e+07     -0.038      0.970      -1.42e+08      1.37e+08
28     =====

```

Listing 6.13: OLS Renca

Figura 6.13: Importancia Variables Renca, del modelo Random Forest



El modelo ajustado para el consumo eléctrico de Renca presenta buenos ajustes en general. Primero, si se observa el coeficiente de determinación, entrega un valor de $R^2 = 0,862$, lo que es muy bueno, pero siempre se debe comparar con su versión ajustada, el cual tiene un valor de $R^2_{adj} = 0,767$, y debido a la poca diferencia entre estos dos indicadores, afirma que el ajuste fue todo un éxito desde el punto de vista explicativo. Ahora, para poder comprobar si el modelo es significativo, se debe ver el valor de la prueba F, el cual corresponde a $F^{statistic} = 5,97E - 5$, por lo que se puede afirmar sin dudas que efectivamente es significativo, por lo que se puede analizar las variables sin ningún reparo. Primero, se observa que las variables que efectivamente son relevantes para el modelo son: x_2 (Año) y x_7 (IPC). Debido a que únicamente dos variables son significativas al 5 %, se puede afirmar que el modelo presenta multicolinealidad, por lo que las demás variables no se analizarán con criterios de significancia.

Ahora, viendo los resultados de la regresión de Random Forest, se puede observar que los resultados son bastantes distintos ya que muchas variables se consideraron como importantes y la escala no es lo suficientemente baja como para no considerarlas, de las que destacan: Desempleo, Mes, Año, Índice de Movilidad e IPC. De esto se puede concluir lo siguiente:

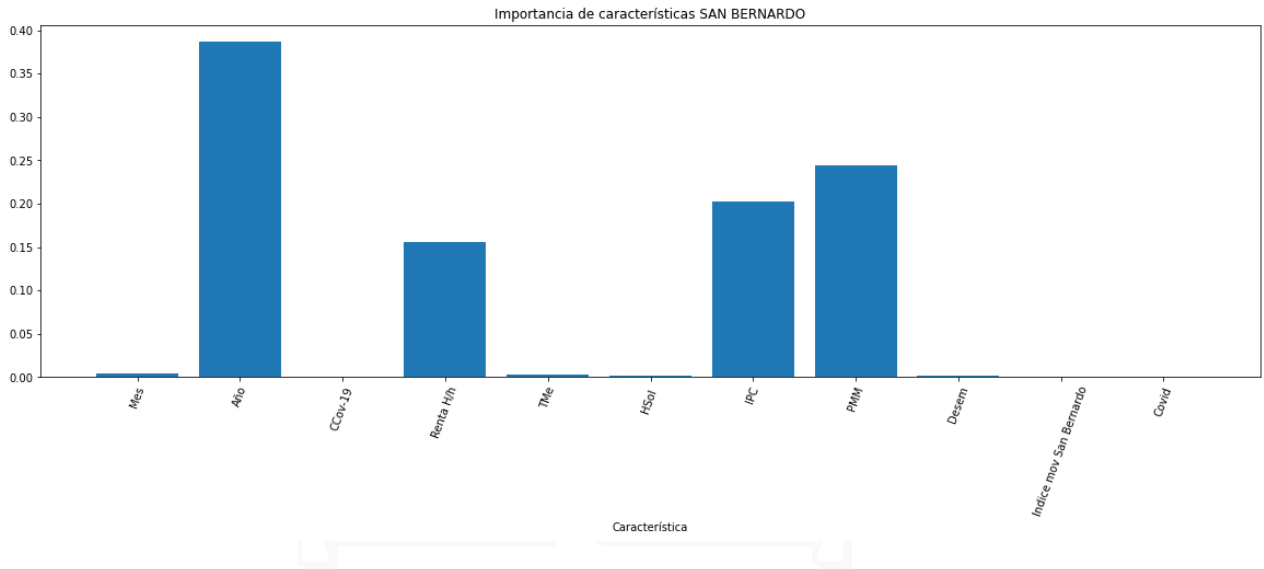
- El hecho de que tanto Año como IPC hayan sido las únicas variables significativas, y que también aparezcan en Random Forest, habla de la preocupación existente por el precio de todos los recursos básicos que año a año siguen en incremento. Como se puede observar en su matriz de correlación (Figura A.13), hay una relación negativa del IPC con el consumo comunal por lo que definitivamente mientras siga aumentando este indicador, el ahorro eléctrico de Renca será mayor a lo largo de los años.
- Si bien el modelo entregó muy buenos indicadores de ajustes, Random Forest logró captar con mucha más claridad los patrones de toda la base de datos, incluyendo muchas más variables. No obstante, su importancia si bien no es tan cercana a 0, siguen siendo relativamente bajas, por lo que si se consideran estos dos factores, se puede concluir que es necesario seguir entrenando los modelos con más datos para poder comprobar efectivamente qué modelo es el más confiable.

6.14. San Bernardo

OLS Regression Results						
=====						
Dep. Variable:	SAN BERNARDO	R-squared:		0.987		
Model:	OLS	Adj. R-squared:		0.978		
Method:	Least Squares	F-statistic:		111.6		
Date:	Mon, 11 Jul 2022	Prob (F-statistic):		5.79e-13		
Time:	17:26:57	Log-Likelihood:		-394.13		
No. Observations:	28	AIC:		812.3		
Df Residuals:	16	BIC:		828.3		
Df Model:	11					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]

const	1.745e+06	3.59e+05	4.855	0.000	9.83e+05	2.51e+06
x1	6.766e+04	2.27e+05	0.299	0.769	-4.13e+05	5.48e+05
x2	2.809e+05	7.46e+05	0.376	0.712	-1.3e+06	1.86e+06
x3	-3.081e+04	2.11e+05	-0.146	0.886	-4.78e+05	4.16e+05
x4	-4.372e+05	6.26e+05	-0.699	0.495	-1.76e+06	8.9e+05
x5	-7.398e+05	3.48e+05	-2.129	0.049	-1.48e+06	-3013.446
x6	6.596e+05	3.45e+05	1.910	0.074	-7.26e+04	1.39e+06
x7	1.844e+06	2.43e+05	7.603	0.000	1.33e+06	2.36e+06
x8	-1.039e+06	3.34e+05	-3.109	0.007	-1.75e+06	-3.31e+05
x9	-7.063e+04	4.06e+05	-0.174	0.864	-9.32e+05	7.9e+05
x10	-1.952e+05	2.68e+05	-0.729	0.477	-7.63e+05	3.72e+05
x11	8.853e+05	1.09e+06	0.811	0.429	-1.43e+06	3.2e+06
=====						

Listing 6.14: OLS San Bernardo

Figura 6.14: Importancia Variables San Bernardo, del modelo Random Forest

El modelo para San Bernardo tuvo excelentes resultados por el lado de los ajustes, entregando valores de $R^2 = 0,987$ y $R^2_{adj} = 0,978$, por lo que se puede afirmar con certeza que efectivamente las variables explican muy bien el consumo eléctrico (explican casi el 100 % de la varianza del modelo), incluso cuando se analizan por separado debido a la muy pequeña variación del coeficiente de determinación con su versión ajustada. Sobre esto, el modelo entregó el menor valor de la prueba F de todos los modelos analizados, siendo un $F^{statistic} = 5,79E - 13$, por lo que teniendo estos 3 indicadores, se puede afirmar que el modelo es significativo en conjunto. No obstante, siempre es necesario tomar en consideración las significancias de todas las variables, y de aquí se observa que solo 3 variables son significantes al 5 %: x_5 (TMe), x_7 (IPC), x_8 (PMM), y además existe otra que se acerca mucho a este valor por lo que también se considerará con un 7 %, que corresponde a x_6 (Horas Sol).

Ahora, Random Forest complementa un poco esta multicolinealidad entre las demás variables que no fueron significantes, ya que es muy robusto en la activación de variables, siendo que en general no lo es, lo que da luces de que hay variables que podrían no considerarse para la formación del modelo. En concreto, se desactivaron 7 variables (i.e su coeficiente fue 0), correspondientes a: Mes, Casos COVID 19, Temperatura Media, Horas de Sol, Desempleo, Índice de movilidad y la variable binaria para COVID. Por esto, se puede concluir lo siguiente para la comuna de San Bernardo:

- A pesar de que los índices de R^2 y R^2_{adj} son ideales, junto al valor de $F^{statistic}$, no quiere decir que el modelo es perfecto para el motivo de este estudio, ya que más allá de tener un buen ajuste, lo que serviría para poder tener modelos predictivos o explicativos, también es muy importante que las variables sean significantes ya que es aquí donde se puede obtener información de la variable por sí sola.
- Considerando estas variables que sí fueron significantes, se pueden relacionar PMM con IPC ya que dentro de IPC está considerado el precio de la luz, por lo que matemáticamente se da naturalmente, y el hecho que ambas sean significativas, habla de la preocupación a nivel comunal del ahorro de energía, y junto a esto, se sigue justificando con la variable de Temperatura Media ya que se consume cuando es necesario.

6.15. San José de Maipo

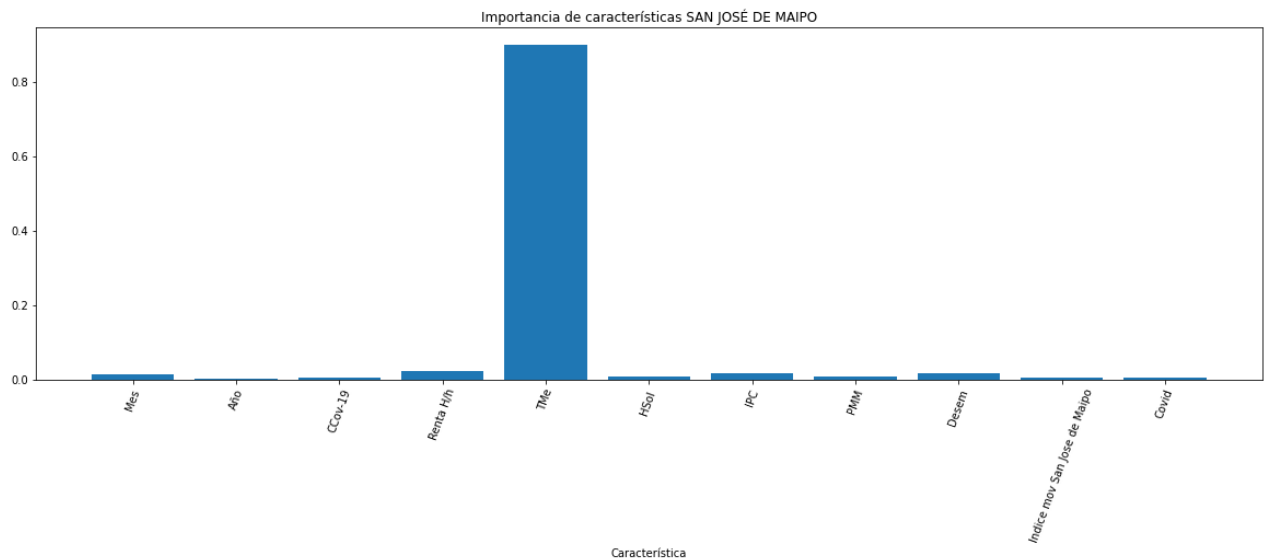
```

=====
2  Dep. Variable:    SAN JOSE DE MAIPO    R-squared:    0.906
3  Model:            OLS    Adj. R-squared:    0.842
4  Method:          Least Squares    F-statistic:    14.08
5  Date:            Mon, 11 Jul 2022    Prob (F-statistic):    3.25e-06
6  Time:            17:26:57    Log-Likelihood:    -496.16
7  No. Observations:    28    AIC:    1016.
8  Df Residuals:    16    BIC:    1032.
9  Df Model:    11
10 Covariance Type:    nonrobust
11
=====
12
13      coef      std err      t      P>|t|      [0.025      0.975]
14 -----
15 const      1.024e+08    1.12e+07     9.158     0.000     7.87e+07    1.26e+08
16 x1          -9.143e+06    8.43e+06    -1.085     0.294    -2.7e+07     8.73e+06
17 x2          -2.251e+07    2.49e+07    -0.904     0.380    -7.53e+07     3.03e+07
18 x3           1.244e+07    8.39e+06     1.483     0.158    -5.35e+06     3.02e+07
19 x4           1.522e+07    2.05e+07     0.744     0.468    -2.81e+07     5.86e+07
20 x5           4.013e+07    1.24e+07     3.232     0.005     1.38e+07     6.64e+07
21 x6          -6.363e+06    1.26e+07    -0.507     0.619    -3.3e+07     2.03e+07
22 x7           5.329e+05    9.25e+06     0.058     0.955    -1.91e+07     2.01e+07
23 x8           1.134e+07    1.26e+07     0.896     0.383    -1.55e+07     3.82e+07
24 x9          -1.487e+07    1.36e+07    -1.093     0.290    -4.37e+07     1.4e+07
25 x10          7.676e+06    1.07e+07     0.717     0.484    -1.5e+07     3.04e+07
26 x11         -1.606e+07    3.35e+07    -0.479     0.638    -8.71e+07     5.5e+07
27
=====

```

Listing 6.15: OLS San José de Maipo

Figura 6.15: Importancia Variables San José de Maipo, del modelo Random Forest



Para la comuna de San José de Maipo ocurre lo mismo que para la comuna anterior. Comenzando por muy buenos valores para el coeficiente de correlación, $R^2 = 0,906$, y para su versión ajustada, $R^2_{adj} = 0,842$. Con ellos se puede observar que al analizar las variables de manera independiente, el ajuste no varía demasiado. Sobre esto, el valor de la prueba F es bastante aceptable, entregando un $F^{statistic} = 3,25E - 6$, por lo que se puede afirmar sin preocupaciones que es un buen modelo al momento de explicar la varianza del consumo. No obstante, como ya se mencionó, ocurre el mismo problema que con muchos otros modelos bien ajustados: multicolinealidad. Esto se comprueba ya que solo una variable de las 11 es significativa, que corresponde a x5, Temperatura Media, por lo que definitivamente el modelo detectó que hay bastantes variables que se

explican entre sí.

Como ya ha sucedido con modelos anteriores, cuando la multicolinealidad es fuerte, esta suele presentarse en otros modelos no lineales debido a que este fenómeno proviene de la base de datos y no del modelo en sí, siendo este uno de esos casos ya que Random Forest entregó un resultado muy parecido, considerando como variable muy importante a TMe (casi 1, el máximo valor que podría tomar), por lo que este comportamiento se comparte y probablemente si se analizara el resultado de SVR: Radial Basis Fuction, el resultado sería muy parecido. Por esto, para San José de Maipo se puede afirmar lo siguiente:

- Si bien se estableció que existe efectivamente multicolinealidad, no quiere decir que el modelo no sea interpretable, sino que estas variables que presentan valores $P > |t|$ muy altos, presentarían conflictos en la interpretación de sus coeficientes, por lo que si bien el modelo funciona muy bien por el lado explicativo y predictivo, no se puede llegar a conclusiones fuertes para cada variable en que no se está seguro que sean significativas.
- Considerando lo anterior, el hecho de que TMe haya aparecido en ambos modelos es una conclusión directa de lo crítico que es esta comuna para el consumo según su temperatura, ya sean muy altas o bajas, es muy probablemente que no existan medidas muy buenas de aislamiento térmico en sus viviendas, por lo que se debe recurrir a aparatos electrónicos para poder remediarlo.

6.16. San Pedro

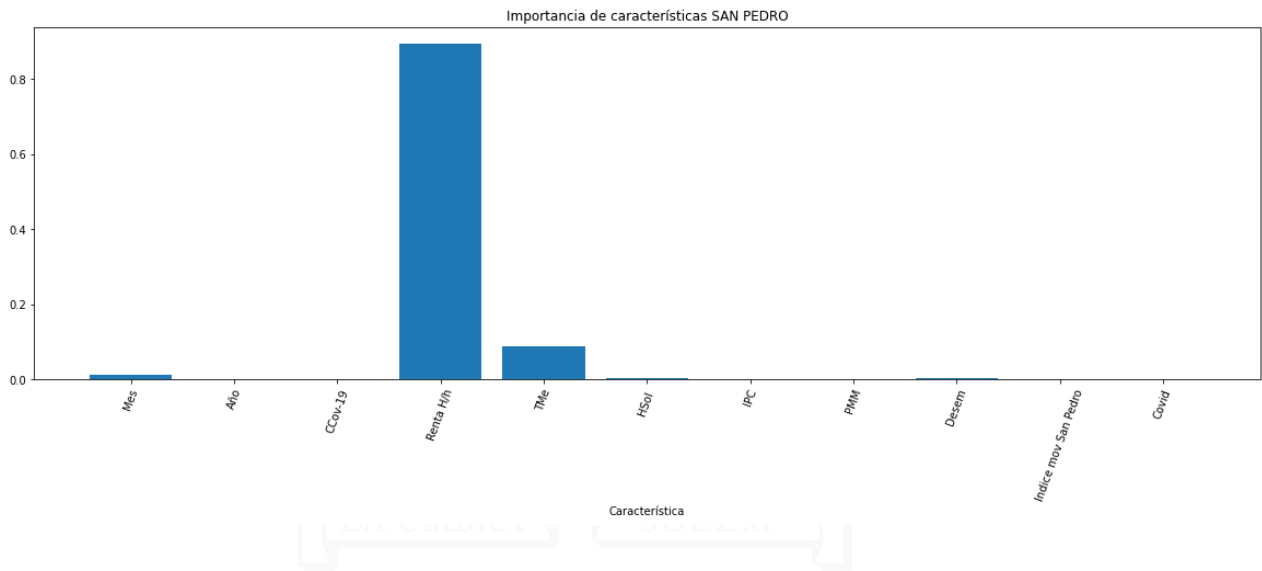
```

1      OLS Regression Results
2      =====
3      Dep. Variable:          SAN PEDRO      R-squared:                0.756
4      Model:                  OLS            Adj. R-squared:          0.589
5      Method:                 Least Squares  F-statistic:             4.518
6      Date:                   Mon, 11 Jul 2022  Prob (F-statistic):      0.00337
7      Time:                   17:26:57       Log-Likelihood:         -466.41
8      No. Observations:       28            AIC:                    956.8
9      Df Residuals:           16            BIC:                    972.8
10     Df Model:                11
11     Covariance Type:        nonrobust
12     =====
13     coef    std err          t      P>|t|      [0.025    0.975]
14     -----
15     const      9.044e+06    2.92e+06      3.094      0.007      2.85e+06    1.52e+07
16     x1         -8.469e+06    2.85e+06     -2.967      0.009     -1.45e+07    -2.42e+06
17     x2         -1.403e+07    8.65e+06     -1.621      0.125     -3.24e+07    4.31e+06
18     x3         -1.269e+06    2.75e+06     -0.462      0.650     -7.09e+06    4.55e+06
19     x4         -1.882e+06    7.01e+06     -0.268      0.792     -1.68e+07    1.3e+07
20     x5         -3.17e+06    4.23e+06     -0.750      0.464     -1.21e+07    5.8e+06
21     x6          7.569e+06    4.31e+06      1.757      0.098     -1.57e+06    1.67e+07
22     x7          2.601e+06    3.19e+06      0.814      0.428     -4.17e+06    9.37e+06
23     x8          1.421e+07    4.43e+06      3.210      0.005      4.83e+06    2.36e+07
24     x9          2.209e+06    4.97e+06      0.444      0.663     -8.33e+06    1.28e+07
25     x10         2.302e+05    3.14e+06      0.073      0.943     -6.44e+06    6.9e+06
26     x11         2.646e+06    8.5e+06       0.311      0.760     -1.54e+07    2.07e+07
27     =====

```

Listing 6.16: OLS San Pedro

Figura 6.16: Importancia Variables San Pedro, del modelo Random Forest



El modelo ajustado para la comuna de San Pedro presenta un aceptable coeficiente de determinación, siendo $R^2 = 0,756$, pero lamentablemente su versión ajustada tiene una variación considerable, tomando un valor de $R^2_{adj} = 0,589$, por lo que no se puede estar tan seguro de que el modelo tuvo un buen desempeño en el ajuste. No obstante, el valor de la prueba F es bastante aceptable, con un valor de $F^{statistic} = 0,00337$, por lo que se puede afirmar que el modelo en conjunto es significativo, por lo que se debe ir a analizar las variables antes de concluir.

Como se dijo, el ajuste no tuvo los valores esperados por lo que también es de esperar que sean pocas las variables que sean significantes para el modelo. dicho esto, sólo destacan 2: x_1 , que corresponde a Mes, y x_8 , que corresponde al Precio Medio de Mercado. De aquí se puede concluir directamente que el modelo efectivamente presenta multicolinealidad, y además que ésta interfiere con el buen ajuste del modelo, cosa que no había sucedido antes, pero el hecho de que en conjunto sea significativo, y ocurra una disminución no menor entre R^2 y R^2_{adj} , da luces de que la inclusión de ciertas variables no aporta al modelo, de hecho, lo entorpece.

Ahora, el modelo Random Forest entregó resultados parecidos por el lado de la multicolinealidad debido a que prácticamente solo activó una variable de las 11 disponibles, siendo esta Renta H/h (x_4 en OLS), por lo que se puede llegar a las siguientes conclusiones:

- Dado que los ajustes no fueron muy buenos, existe una multicolinealidad fuerte, y además el modelo Random Forest entregó la activación de solo una variable, se puede concluir que es necesario seguir entrenando los modelos con una mayor cantidad de datos ya que las interpretaciones que se puede dar son un poco difusas desde el nivel interpretativo
- El hecho de que Random Forest haya activado una variable distinta a la única dignificante al 5 % en OLS, deja en claro que el problema no es el modelo sino que la base de datos. La opción más plausible es que no hay suficientes datos como para encontrar una consistencia entre variables versus consumo eléctrico. Es por esto que es de mucha ayuda tener muchos modelos para comparar porque sucede frecuentemente que simplemente el modelo no se adapta y es necesario encontrar acercamientos no lineales o paramétricos para captar su comportamiento.

6.17. Santiago

1	OLS Regression Results
2	=====

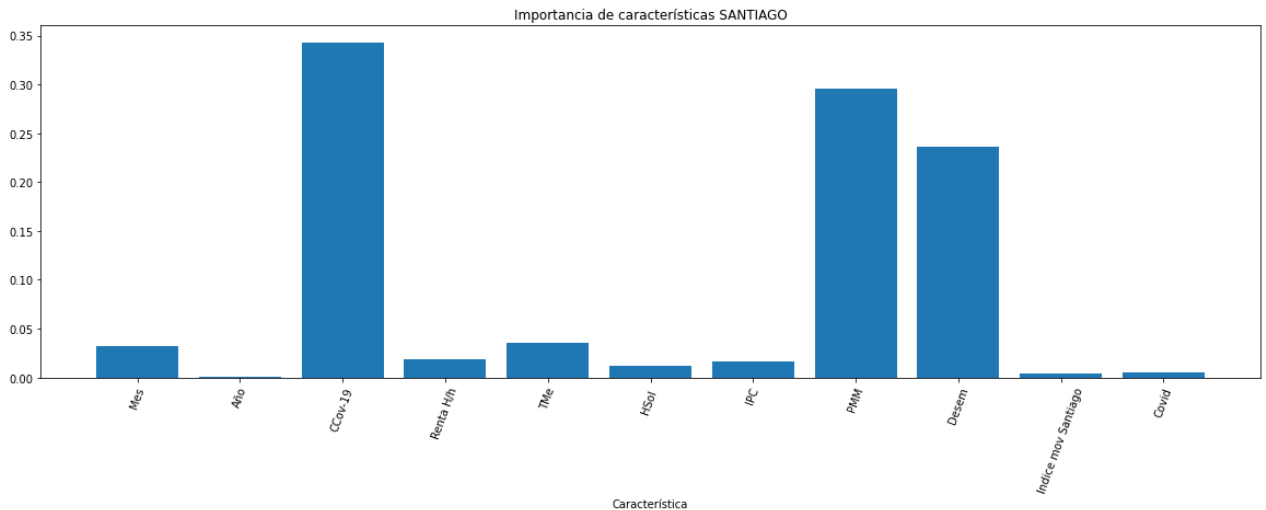
```

3 Dep. Variable:          SANTIAGO    R-squared:          0.881
4 Model:                OLS          Adj. R-squared:      0.799
5 Method:                Least Squares  F-statistic:        10.77
6 Date:                  Mon, 11 Jul 2022  Prob (F-statistic): 1.97e-05
7 Time:                  17:26:57      Log-Likelihood:     -429.02
8 No. Observations:      28          AIC:                882.0
9 Df Residuals:          16          BIC:                898.0
10 Df Model:              11
11 Covariance Type:      nonrobust
12 =====
13      coef      std err      t      P>|t|      [0.025      0.975]
14 -----
15 const      2.211e+07    1.36e+06    16.238    0.000    1.92e+07    2.5e+07
16 x1          -7.811e+04    7.5e+05    -0.104    0.918   -1.67e+06    1.51e+06
17 x2          2.804e+06    2.23e+06    1.259    0.226   -1.92e+06    7.53e+06
18 x3          8.634e+05    7.34e+05    1.177    0.256   -6.92e+05    2.42e+06
19 x4         -1.478e+06    1.82e+06   -0.813    0.428   -5.33e+06    2.37e+06
20 x5          2.686e+04    1.14e+06    0.024    0.981   -2.39e+06    2.44e+06
21 x6          -3.58e+05    1.15e+06   -0.312    0.759   -2.79e+06    2.07e+06
22 x7          5.224e+05    8.64e+05    0.604    0.554   -1.31e+06    2.35e+06
23 x8          -4.119e+05    1.16e+06   -0.357    0.726   -2.86e+06    2.04e+06
24 x9          -2.58e+06    1.58e+06   -1.637    0.121   -5.92e+06    7.61e+05
25 x10         1.63e+06    1.24e+06    1.311    0.208   -1.01e+06    4.27e+06
26 x11        -6.678e+06    4.15e+06   -1.609    0.127   -1.55e+07    2.12e+06
27 =====

```

Listing 6.17: OLS Santiago

Figura 6.17: Importancia Variables Santiago, del modelo Random Forest



Para la comuna de Santiago, los resultados por parte del ajuste fueron bastante buenos. Ambos coeficientes de determinación entregaron valores muy aceptables, con una diferencia que no afecta lo suficiente como para desestimar el poder explicativo de las variables de estudio, siendo estos $R^2 = 0,881$ y $R^2_{adj} = 0,799$. Además, para sustentar estos indicadores, se encuentra la prueba F que puede respaldar desde otro punto de vista, entregando un $F^{statistic} = 1,97E - 5$, lo que quiere decir que el modelo es significativo en conjunto.

Lamentablemente al analizar la significancia de las variables las noticias no son tan buenas como antes ya que ninguna variable es significativa al 5 %, por lo que es necesario apoyar el análisis en los resultados de Random Forest. Aquí, el modelo activó 3 variables: Casos Covid 19, Precio Medio de Mercado y Desempleo. De esto se pueden obtener las siguientes conclusiones:

- Lo primero a destacar es el hecho que los coeficientes de determinación fueron bastante buenos pero ninguna variable fue significativa. Esto puede deberse a muchas razones, pero dado el caso en que esta tabla se realizó en base a una muestra, el 20 % de los datos iniciales, es muy probable que la cantidad de datos sea poca para la cantidad de variabilidad que la misma presenta. Esto tiene mucha concordancia con los resultados de modelos ya analizados.
- Nuevamente, debido a temas de arquitectura de bibliotecas, OLS efectivamente necesita cierta cantidad de datos para ser analizada correctamente, y es por esto que Random Forest es de suma ayuda. Cobra mucho sentido que la comuna con más habitantes de la Región Metropolitana haya sido afectada por la variable de Casos COVID, y este mismo hecho fue el que hizo que Desempleo fuera importante para el consumo ya que muy probablemente por efectos de la pandemia, se despidió a un porcentaje considerable de la población.

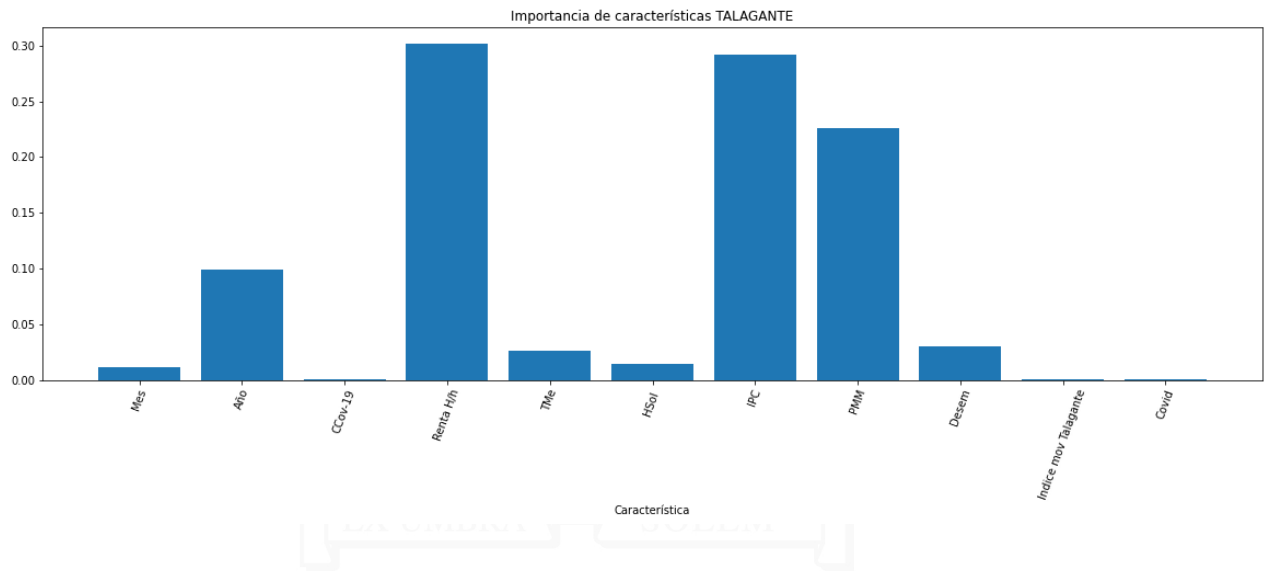
6.18. Talagante

```

1      OLS Regression Results
2      =====
3      Dep. Variable:      TALAGANTE      R-squared:      0.910
4      Model:              OLS      Adj. R-squared:      0.848
5      Method:              Least Squares      F-statistic:      14.72
6      Date:                Mon, 11 Jul 2022      Prob (F-statistic):      2.38e-06
7      Time:                17:26:57      Log-Likelihood:      -438.52
8      No. Observations:      28      AIC:      901.0
9      Df Residuals:          16      BIC:      917.0
10     Df Model:              11
11     Covariance Type:      nonrobust
12     =====
13     coef      std err      t      P>|t|      [0.025      0.975]
14     -----
15     const      1.099e+07      1.53e+06      7.174      0.000      7.74e+06      1.42e+07
16     x1          -8.866e+05      1.05e+06     -0.842      0.412     -3.12e+06      1.35e+06
17     x2          2.155e+06      3.14e+06      0.687      0.502     -4.5e+06      8.81e+06
18     x3          -4.393e+05      1.25e+06     -0.352      0.729     -3.08e+06      2.21e+06
19     x4          -3.203e+06      2.57e+06     -1.246      0.231     -8.65e+06      2.25e+06
20     x5          -2.421e+06      1.56e+06     -1.548      0.141     -5.74e+06      8.95e+05
21     x6          -4.166e+05      1.59e+06     -0.263      0.796     -3.78e+06      2.94e+06
22     x7          7.417e+06      1.18e+06      6.288      0.000      4.92e+06      9.92e+06
23     x8          7.061e+06      1.62e+06      4.369      0.000      3.64e+06      1.05e+07
24     x9          -3.862e+06      1.76e+06     -2.199      0.043     -7.59e+06     -1.39e+05
25     x10         -4.708e+05      1.59e+06     -0.296      0.771     -3.84e+06      2.9e+06
26     x11          5.428e+05      4.61e+06      0.118      0.908     -9.24e+06      1.03e+07
27     =====

```

Listing 6.18: OLS Talagante

Figura 6.18: Importancia Variables Talagante, del modelo Random Forest

La regresión para la comuna de Talagante entregó los siguientes resultados. Para el coeficiente de correlación, $R^2 = 0,910$, lo que quiere decir que las variables están explicando un muy esperanzador 90 % de la varianza del modelo, lo que es muy bueno, y si se analiza el valor de su versión ajustada, $R^2_{adj} = 0,848$, se puede ver que no cambia mucho por lo que el número de variables no está afectando de manera tan negativa al modelo en conjunto. Además, si se analiza el valor de la prueba F, el cual es $F^{statistic} = 2,38E - 6$, se puede afirmar que el modelo es significativo en conjunto por lo que esta dimensión macro ya está cubierta, ahora es necesario analizar variable por variable.

De inmediato se puede identificar que existen 3 variables que son significantes al 5 %, las cuales son x7 (IPC), x8 (PMM), x9 (Desempleo). Esto se repite en los resultados de los coeficientes de Random Forest, en donde destacan como variables importantes a IPC, PMM, Renta H/h y Año, por lo que para esta comuna se puede concluir lo siguiente:

- Se puede ver claramente que en ambos modelos resaltan variables puramente económicas. Como se ha visto en otros análisis, esto suele suceder para las comunas que no son tan céntricas dado que la base n la que estos indicadores se calculan a nivel nacional, les afecta en muchas dimensiones y una de esas es en el consumo, por lo que son conscientes de cuanto se consume a mes y se tiende a ahorrar.
- En general el desempleo siempre será un factor relevante en todas las comunas, y esta no es la excepción, pero si se puede ver que el virus no afectó mucho a la empleabilidad de la comuna debido a que no fue significativa la variable x11 (binaria para el COVID). Es por esto que si se adhiere al análisis anterior, se puede concluir que en general es una comuna con medios-bajos recursos que le afectan mucho estos índices macroeconómicos.

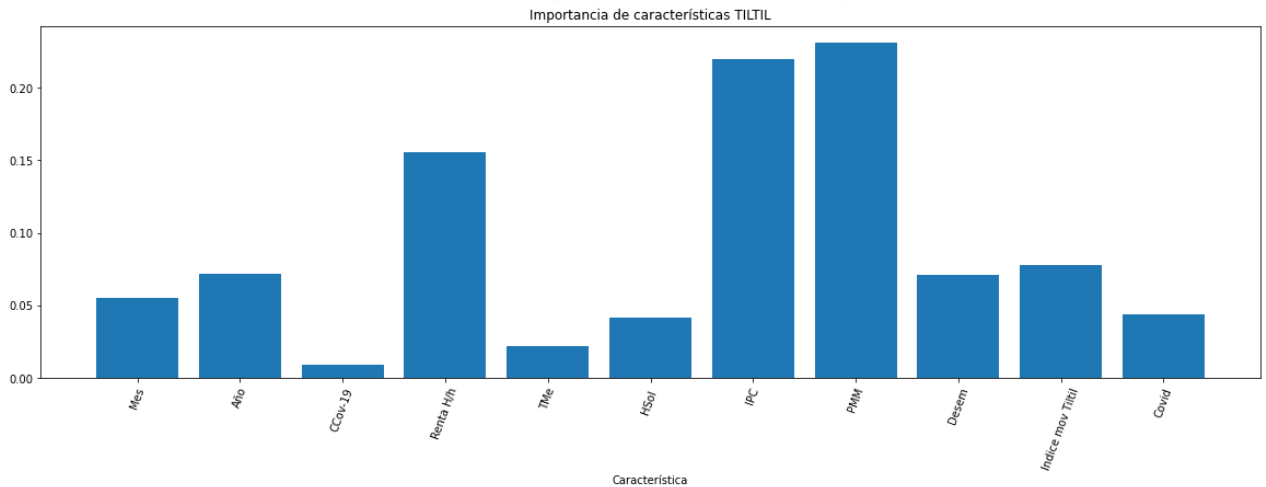
6.19. Til Til

1	OLS Regression Results			
2	=====			
3	Dep. Variable:	TILTIL	R-squared:	0.772
4	Model:	OLS	Adj. R-squared:	0.615
5	Method:	Least Squares	F-statistic:	4.925
6	Date:	Mon, 11 Jul 2022	Prob (F-statistic):	0.00215
7	Time:	17:26:57	Log-Likelihood:	-537.15
8	No. Observations:	28	AIC:	1098.
9	Df Residuals:	16	BIC:	1114.
10	Df Model:	11		

11	Covariance Type: nonrobust						
12	=====						
13		coef	std err	t	P> t	[0.025	0.975]
14		-----					
15	const	1.444e+08	7.11e+07	2.031	0.059	-6.29e+06	2.95e+08
16	x1	-3.57e+07	3.67e+07	-0.973	0.345	-1.14e+08	4.21e+07
17	x2	-1.698e+08	1.06e+08	-1.599	0.129	-3.95e+08	5.53e+07
18	x3	-1.878e+07	3.64e+07	-0.516	0.613	-9.59e+07	5.84e+07
19	x4	6.252e+07	8.81e+07	0.710	0.488	-1.24e+08	2.49e+08
20	x5	-7.684e+07	5.7e+07	-1.347	0.197	-1.98e+08	4.41e+07
21	x6	8.573e+07	5.68e+07	1.510	0.151	-3.46e+07	2.06e+08
22	x7	-1.353e+07	4.05e+07	-0.334	0.743	-9.94e+07	7.23e+07
23	x8	-4.068e+07	5.5e+07	-0.740	0.470	-1.57e+08	7.59e+07
24	x9	5.334e+07	7.17e+07	0.744	0.468	-9.87e+07	2.05e+08
25	x10	1.409e+07	7.27e+07	0.194	0.849	-1.4e+08	1.68e+08
26	x11	3.949e+07	2.17e+08	0.182	0.858	-4.22e+08	5.01e+08
27		=====					

Listing 6.19: OLS Til Til

Figura 6.19: Importancia Variables Til Til, del modelo Random Forest



En la comuna de Til Til se obtuvieron resultados aceptables por parte del ajuste, comenzando por su $R^2 = 0,772$, el cual habla de la buena representación de su comportamiento al incluir las variables de estudio, pero de todas formas existe una variación considerable con su versión ajustada, entregando un valor de $R^2_{adj} = 0,615$, por lo que la inclusión de ciertas variables hizo que el ajuste empeorara por casi 2 décimas. No obstante, la prueba F entregó un $F^{statistic} = 0,00215$ el cual es suficientemente bajo como para asegurar que el modelo efectivamente es significativo en conjunto, por lo que eventualmente se podría analizar sin problemas.

Al ver cada variable, se puede detectar de inmediato un error sistemático de las variables ya que, como ya sucedió anteriormente, ninguna variable es significativa al 5 %, por lo que es necesario basar el análisis en los coeficientes entregados por Random Forest, de los que destacan: IPC, PMM y Renta H/h. Si bien no hay ninguno que sea exactamente 0, las variables con una importancia menor a la de IPC son muy bajas como para poder sacar conclusiones contundentes. De esto se pueden obtener las siguientes conclusiones:

- Dado que OLS no entregó buenos resultados en las significancias de variables, se tuvo que recurrir a RF lo que llama directamente a dos razones: la base de datos es muy pequeña para poder contrarrestar la varianza existente con el consumo, lo que lleva a buenos ajustes pero bajas significancias, o que simplemente un modelo lineal no es el indicado considerando que el ajuste no es suficientemente bueno como para estar convencido de ello.

- Que Random Forest haya captado estas tres variables tiene mucho sentido debido a que tienen naturalezas puramente económicas. En el IPC se encuentra la canasta básica, junto al precio de la luz, por lo que es natural que vayan de la mano, y junto a esto la Renta, que sin ella simplemente no se puede consumir ningún bien de esta canasta.



7 | Conclusiones

A lo largo de este trabajo se realizó un análisis del consumo eléctrico de 19 comunas de la Región Metropolitana. Este se realizó en editor de texto Google Colaboratory, en lenguaje Python, y utilizando distintos modelos y funciones dispuestas en las librerías del listado 4.1. Los datos se obtuvieron de diferentes bases de datos de empresas estatales y privadas que disponen a todo público, y en específico, el consumo eléctrico se obtuvo a través de Coordinador Eléctrico, quien facilitó el balance eléctrico para cada subestación instalada en cada comuna, por esto la información no es del 100 % precisa ya que esta entidad no guarda el consumo de todas las subestaciones de la Región Metropolitana, pero de todas formas entregó una muy buena noción de los órdenes de magnitud del consumo eléctrico entre comunas.

Considerando esto, el análisis se realizó basándose en la comparación de los índices R^2 y $RMSE$, entre 7 modelos distintos, los cuales fueron: Support Vector Machines con kernel “linear” y “radial basis function”, regresión por Mínimos Cuadrados Ordinarios de librerías sklearn y statsmodel.api, regresión Ridge, regresión Lasso, y regresión Random Forest. Además, el modelamiento se realizó sobre la base de datos que contenía 11 variables sin contabilizar el consumo eléctrico, y cada modelo se entrenó sobre 4 escalamientos distintos a la base de datos: Standard, Robust, MinMax y MaxAbs, comparándose un total de 532 modelos.

Cada modelo se iteró en un conjunto de entrenamiento, y se validó en uno de prueba, y después de distintos procesos, se escogieron dos modelos: Mínimos Cuadrados Ordinarios de la librería statsmodel.api, y Random Forest, el cual se utilizó para poder aportar análisis de coeficientes desde una perspectiva no lineal. La principal razón por la cual se escogió esta regresión fue el buen equilibrio entre $RMSE$ y R^2 , ya que un buen coeficiente de correlación entregará una buena explicación de la variable independiente, y el Error Cuadrático Medio medirá que tan bueno es el modelo al realizar predicciones.

En general los modelos entregaron buenos resultados desde el punto de vista matemático, ya que la mayoría entregó valores de ajustes R^2 mayores a 0.8, y R^2_{adj} mayores a 0.75, por lo que la cantidad de variables no afectó en gran medida al ajuste. De todas formas, muchos modelos también presentaron indicios de multicolinealidad, y en este respecto, después de analizar las 19 comunas, se descubrió que la base de datos presentaba multicolinealidad por sí sola, ya que debido a la naturaleza de las variables, muchos modelos captaron que algunas se explicaban entre sí, por lo que se encontraron modelos muy buenos del punto de vista explicativo, pero únicamente 2 o 3 variables eran significativas, de un total de 11. Es más, muchas comunas presentaban un fenómeno parecido al ver los resultados de Random Forest, en donde solo se activaban 2 o 3 variables, lo que quiere decir que las predicciones se podían realizar con éstas y las demás no aportaban mucho en comparación.

Ahora, las variables que más destacaron a lo largo de las comunas, tanto en significancia como en importancia fueron: Temperatura Media, Renta, Número de Casos COVID, Horas Sol, Precio Medio de Mercado e Índice de Movilidad. De antemano se esperaba que Temperatura Media y Horas Sol afectaran a los modelos significativamente, ya que es lógico pensar que hay una tendencia a consumir más luz en la noche (correlación negativa con Horas sol), y también que se vea afectado por la temperatura, lo que se traduce en el uso de ventiladores, aires acondicionados, o calefactores eléctricos. Renta y Precio Medio de Mercado afectaron mayoritariamente a las comunas provinciales, de lo que se entiende que existe una mayor preocupación por mantener un trabajo estable ya que es crítico al momento de consumir el bien básico que es la luz o electricidad.

Para las variables relacionadas con el virus COVID-19, al inicio de este estudio se estableció una variable binaria para poder estudiar si la pandemia generó un cambio estructural en el consumo de algunas comunas, lo

que fue válido únicamente para 2 comunas: Lo Prado y Colina. No obstante, esto no quiere decir que el virus no afectó al consumo, ya que Casos COVID e Índice de Movilidad sí destacaron como significativas, lo que deja en claro que estas dos variables forzaron a que la mayoría de los habitantes de las comunas se quedaran en sus hogares, aumentando el consumo residencial (correlacionado con HSol y TMe) pero disminuyendo el empresarial, e incluso en algunas comunas esto generó una alta significancia en la variable Desempleo (Cerro Navia, La Pintana, Lo Prado, Santiago y Talagante)

De todas formas, existieron modelos en los que las conclusiones no fueron certeras desde el punto de vista de sus significancias, tanto para cada variable como el modelo en conjunto. Estas comunas son Buin, Pudahuel y Paine, en las que se debió recurrir a un análisis con Random Forest, o directamente ambos modelos entregaron resultados y ajustes suficientemente malos como para asegurar la validez de este comportamiento. Esto se analizó y se llegó a la resolución de que sucedió debido a la poca cantidad de datos presentes en el conjunto de prueba de la fase de validación. En general Random Forest no necesita tantos datos para obtener buenos resultados, pero cuando existe mucha varianza y no hay muchas entradas, se dan casos en los que ninguna variable es significativa en la regresión OLS, y Random Forest entregaba importancias muy cercanas a 0, por lo que la única solución es seguir entrenando el modelo con más datos, que lamentablemente no se lograron obtener al momento de realizar el estudio.

Dadas las condiciones en las que se realizaron los modelos, estos se pueden mejorar enormemente si se sigue alimentando la base de datos hasta tener suficientes registros por comuna, idealmente rondando los 50. Además, se pueden aplicar muchas más técnicas de preprocesamiento de datos, como criterios de Random Forest y de índices distintos a la matriz de correlación, que detecten relaciones no-lineales, que ayudarían enormemente a evitar la multicolinealidad, y respecto a esto, se pueden incluir más variables que no se encuentren relacionadas a índices económicos ya que IPC y PMM guardan mucha de esta información, además de ser las variables que más correlación presentaron respecto a las demás. También, con un mejor equipo, se pueden realizar validaciones cruzadas con muchos más parámetros ya que este método de validación le exige mucho a la máquina local y fue una fuerte restricción al momento de realizar los ajustes y predicciones.

Bibliografía

- Agdas, D y Barooah, P (2020). Impact of the COVID-19 Pandemic on the U.S. Electricity Demand and Supply: An Early View From Data. *IEEE Access*, 8, 151523–151534. 1
- Arias Quezada, A (2009). El Mercado Eléctrico en Chile. 3.3.1
- Banco Central de Chile (2020). *La economía chilena frente a la pandemia del COVID-19: fortalezas, desafíos y riesgos*. Technical report. 3.1.2
- Biblioteca del Congreso Nacional de Chile / BCN (2017). *Proyección de población según sexo y tramos de edad, base 2017*. Technical report. 4.2.2
- BNamericas (2021). BNamericas - El impacto del COVID-19 en las principales distribuidoras eléctricas de Perú. 1
- Casagrande, Maria; Favieri, Francesca; Tambelli, Renata; y Forte, Giuseppe (2020). The enemy who sealed the world: effects quarantine due to the covid-19 on sleep quality, anxiety, and psychological distress in the italian population. *Sleep Medicine*, 75, 12–20. 3.1
- Comisión Nacional de Energía (2020). Precio Medio de Mercado - Comisión Nacional de Energía. 4.2.2
- Datosmacro (2022). *Desempleo de Chile*. Technical report. 4.2.2
- Dazarola Leichtle, G (2020). *Fondo Común Municipal (FCM)*. Technical Report 124862. 3.2
- Generadoras de Chile (2020). Generación Eléctrica en Chile. 3.3, 3.2
- Géron, A (2019a). *Support Vector Machines*, (pp. 155–176). O'Reilly Media, 2 edición. 3.5, 3.6, 3.7, 3.8, 3.9, 3.10
- Géron, A (2019b). *The Machine Learning Landscape*, (pp. 3–35). O'Reilly, 2 edición. 3.4, 3.11
- Instituto de Data Science de la Universidad del Desarrollo Bravo, Loreto, and Ferres, Leo. (2020). *The IM (Mobility Index) dataset, electronic dataset, UDD and Ministry of Science, Chile*. Technical report. 4.2.2
- Instituto Nacional de Estadísticas (2021). *Índice nominal de remuneraciones*. Technical report. 4.2.2
- ISCI; SPEC; Suazo, C; Moreno, R; y Otárola, H (2020). *Impactos de las Medidas Asociadas al COVID-19 en el Consumo Eléctrico a Nivel Nacional*. Technical report. 1
- Ley General de Servicios Eléctricos (2007). *Artículo 147*, (pp. 102). Diario Oficial de la República de Chile. 3.3.2
- Ministerio de Salud de Chile (2020). *Informe epidemiológico COVID-19*. Technical report. 4.2.2
- MINSAL (2020a). Ministerio de Salud confirma primer caso de coronavirus en Chile. 3.1.1
- MINSAL (2020b). Ministerio de Salud decreta cuarentena total para la ciudad de Santiago y seis comunas aledañas. 1

- MINSAL (2021). Cifras Oficiales COVID-19. 3.1.1
- Monasterio, F. (2021). El avance de la vacunación del covid-19 en Chile y el mundo. 3.1
- Paramo, A. (2020). Demanda y precio de la energía eléctrica en Argentina: impacto de la pandemia y tendencias. 1
- Raj, R (2020). Supervised, Unsupervised, And Semi-Supervised Learning With Real-Life Usecase. 3.3, 3.4
- Senado de la República de Chile (2017). *Ingreso Municipal Per cápita 2015*. Technical report. 3.2
- Servicio de Impuestos Internos (SII) (2022). *UTM - UTA - IPC 2022*. Technical report. 4.2.2
- Sobre Protección de la Vida Privada (1999). *Ley N°19628*. Diario Oficial de la República de Chile. 4.2.1
- Weather Spark (2022). *El clima y el tiempo promedio en todo el año en Santiago de Chile*. Technical report. 4.2.2
- WHO (2020). WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020. 1

A | Anexos

A.1. Matrices de Correlación

Figura A.1: Matriz de Correlación de Alhué

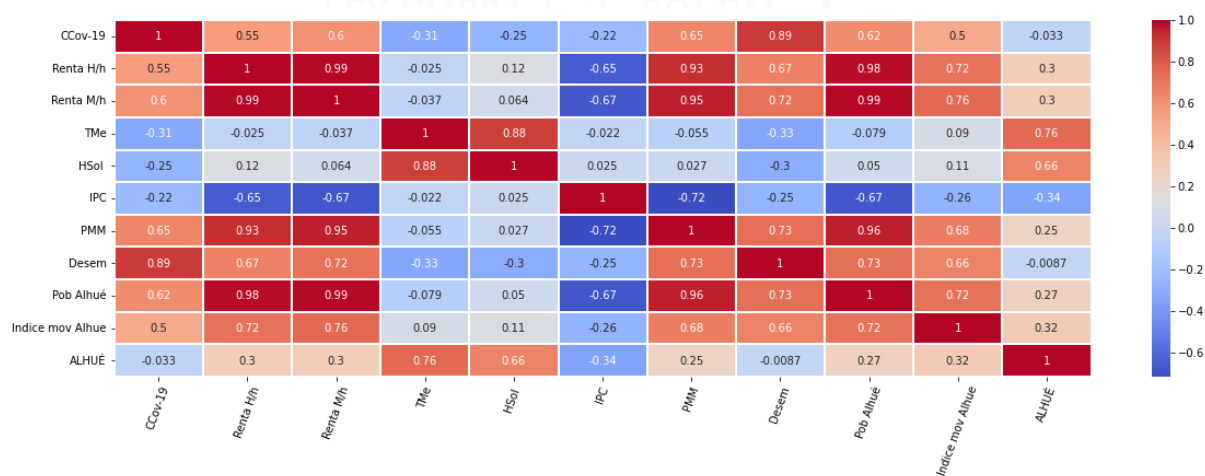


Figura A.2: Matriz de Correlación de Buin

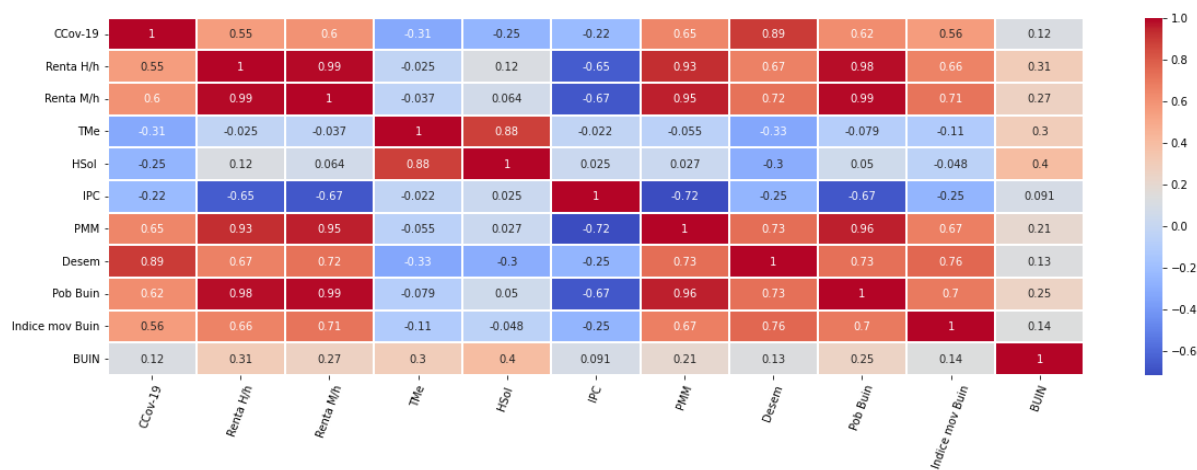


Figura A.3: Matriz de Correlación de Cerro Navia

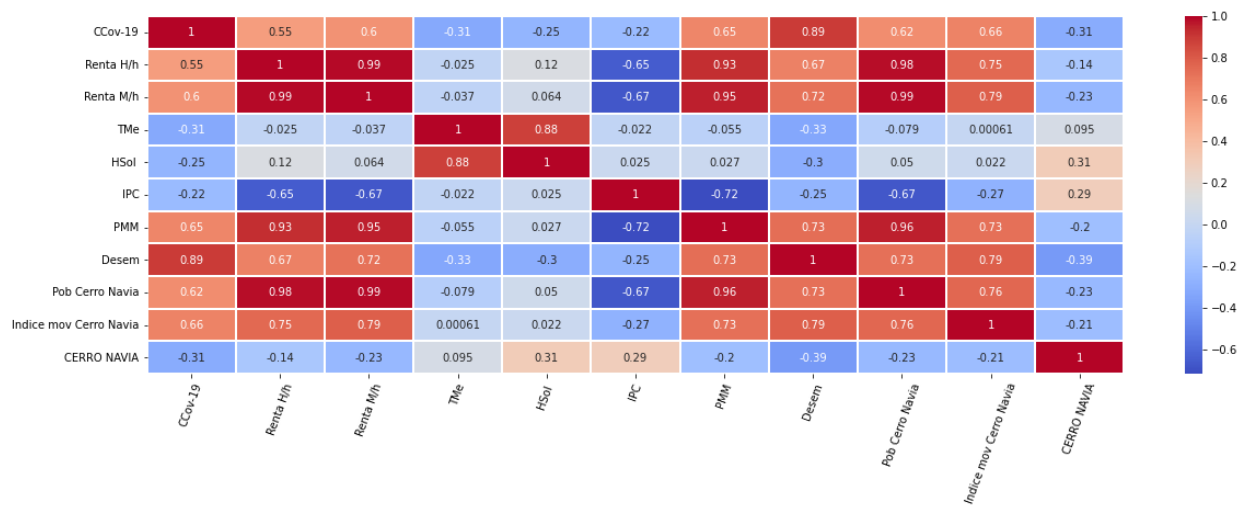


Figura A.4: Matriz de Correlación de Colina

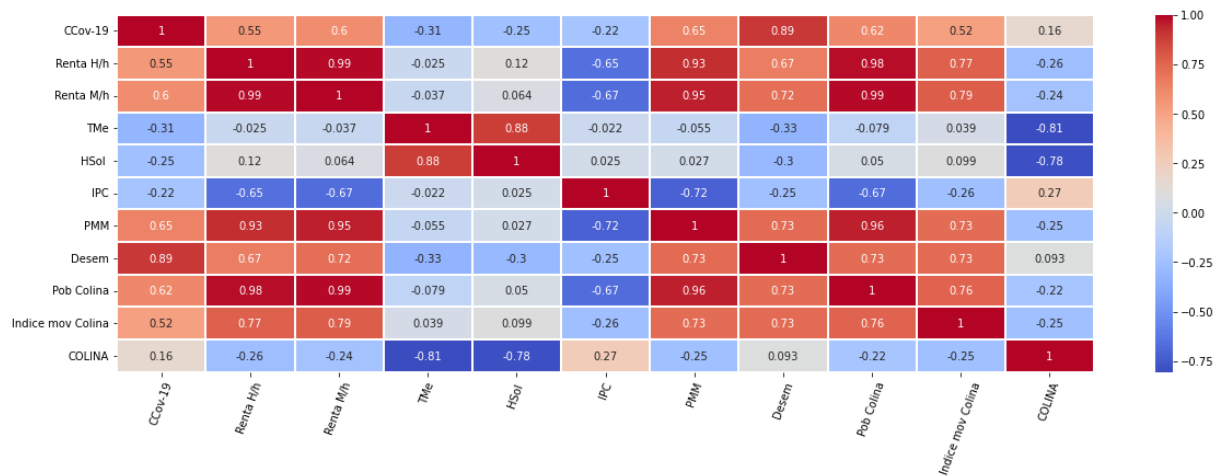


Figura A.5: Matriz de Correlación de El Monte

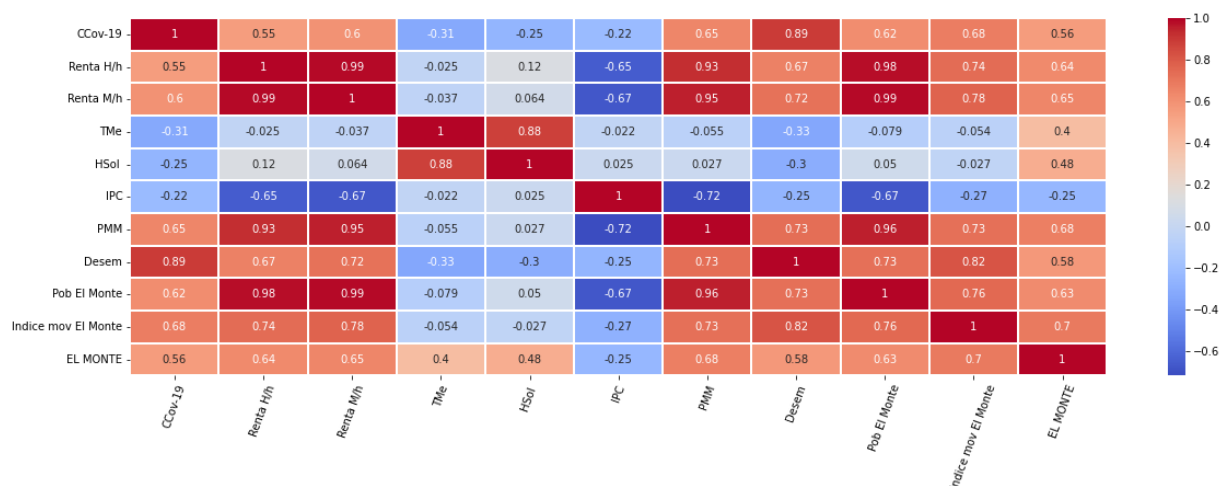


Figura A.6: Matriz de Correlación de La Pintana

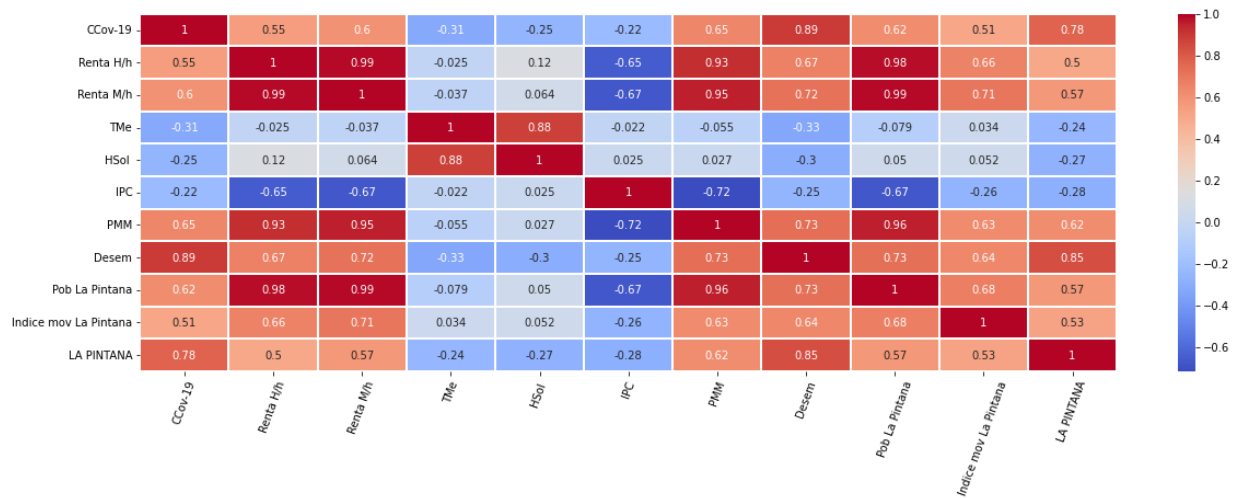


Figura A.7: Matriz de Correlación de Lo Prado

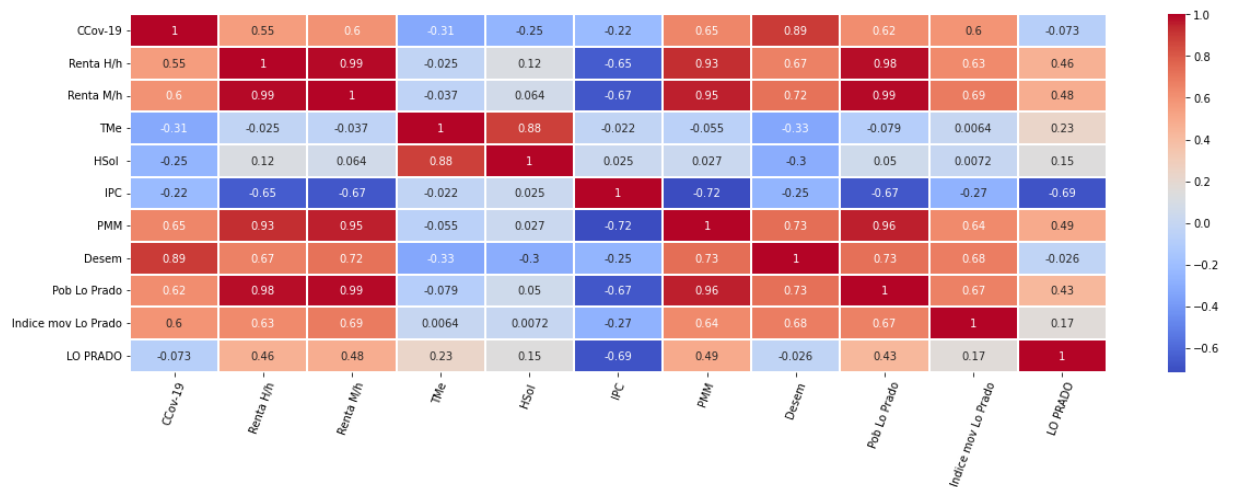


Figura A.8: Matriz de Correlación de Melipilla

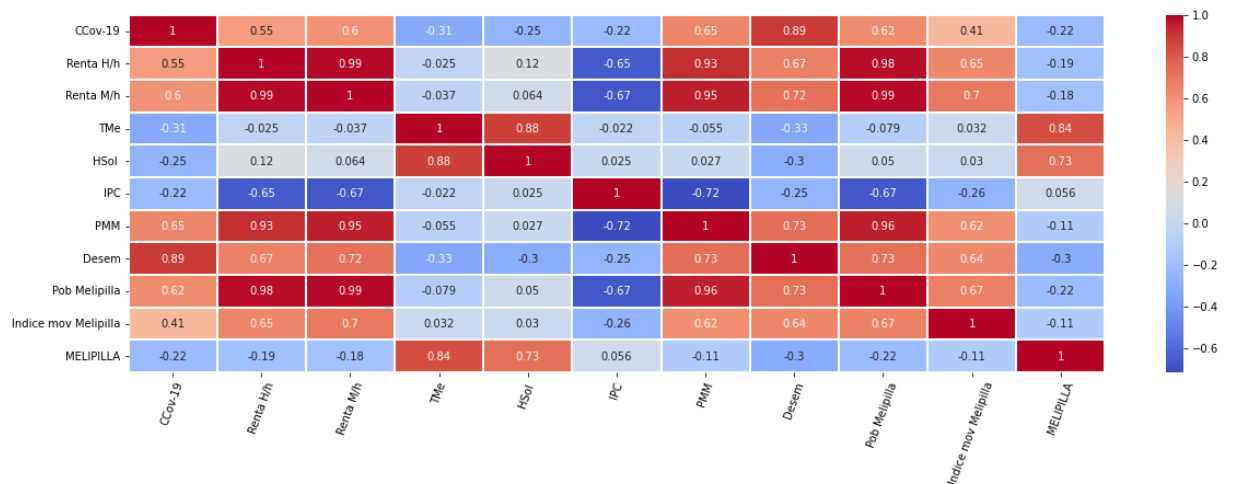


Figura A.9: Matriz de Correlación de Paine

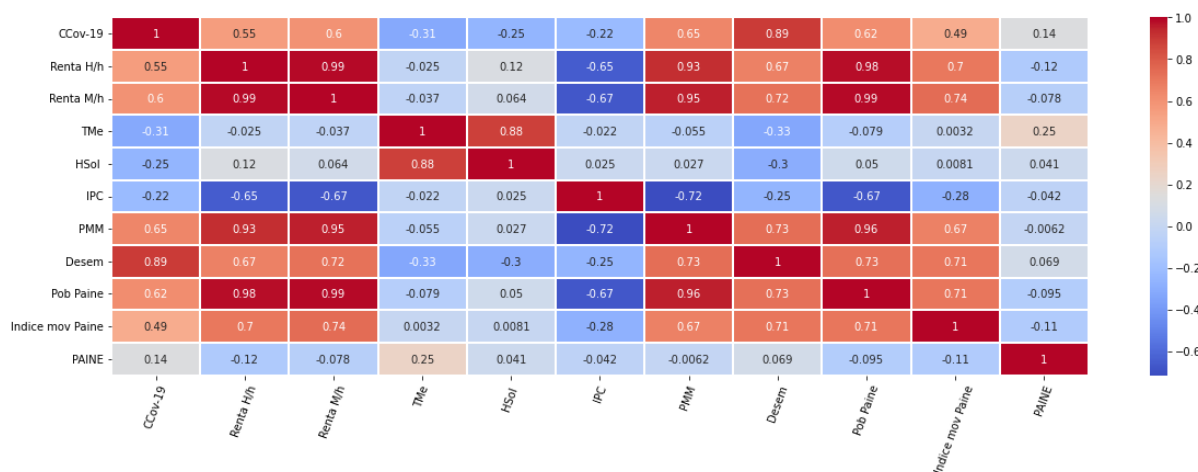


Figura A.10: Matriz de Correlación de Pirque

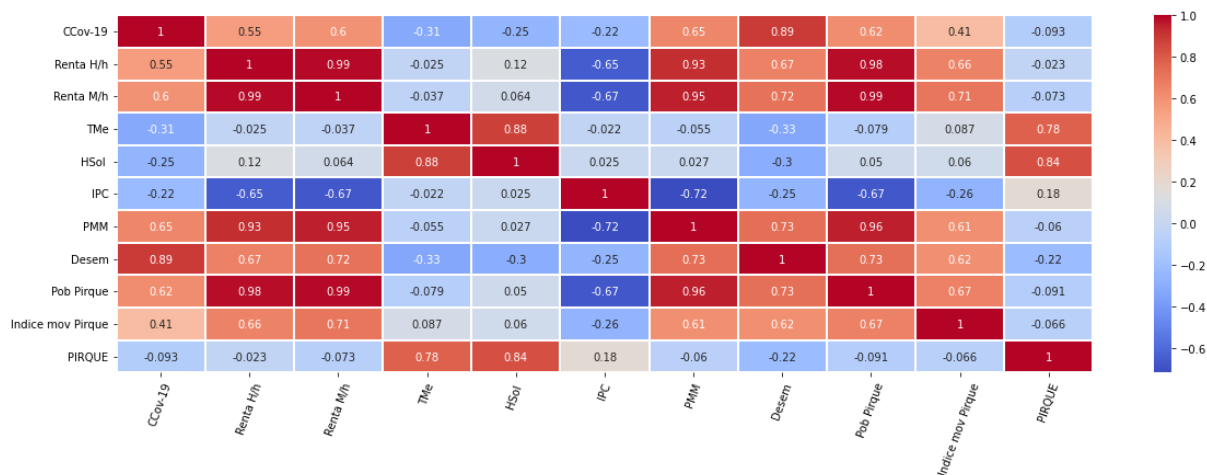


Figura A.11: Matriz de Correlación de Pudahuel

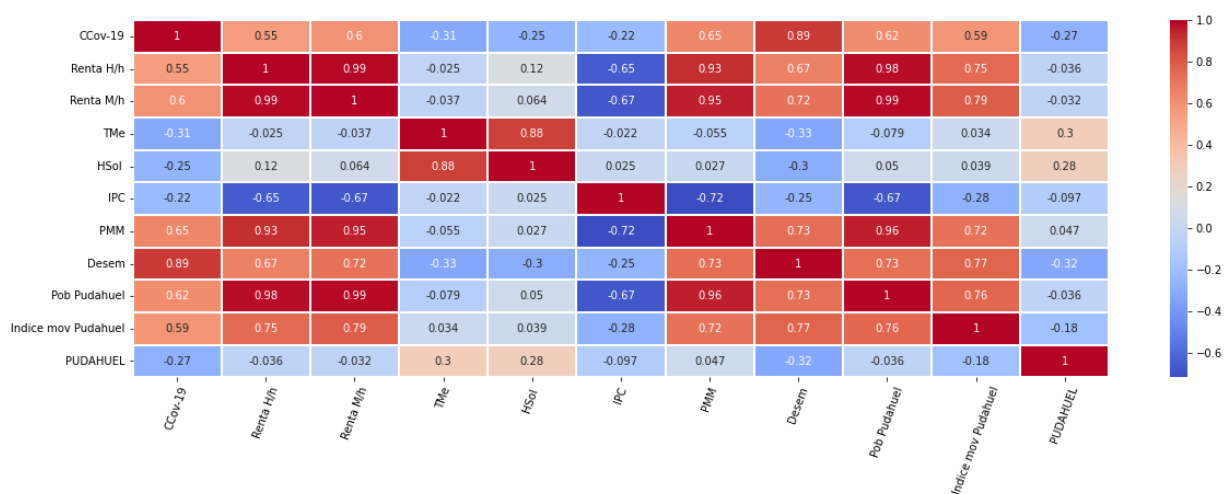


Figura A.12: Matriz de Correlación de Puente Alto

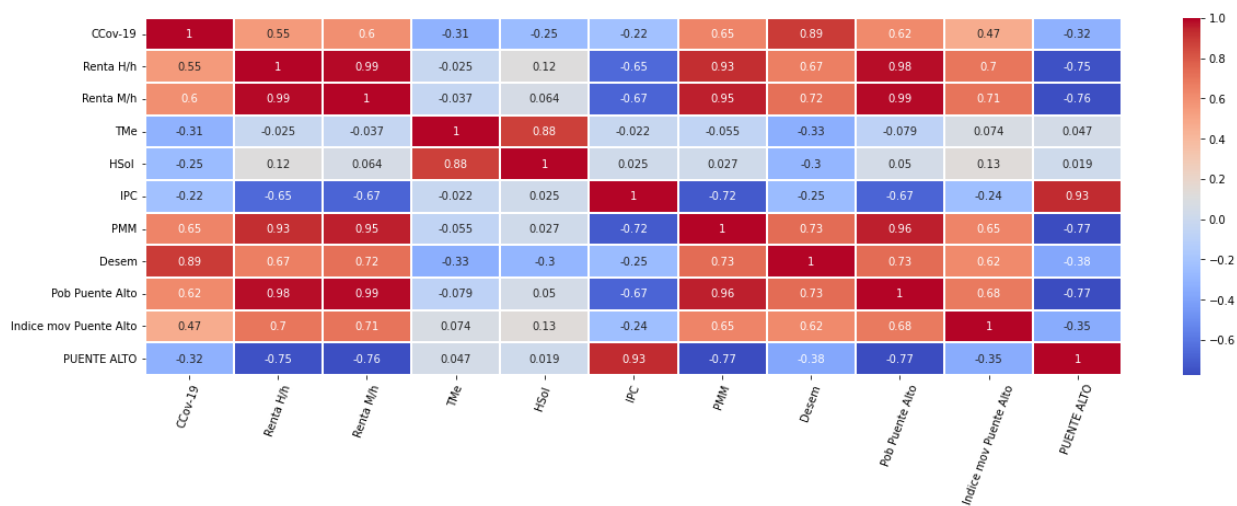


Figura A.13: Matriz de Correlación de Renca

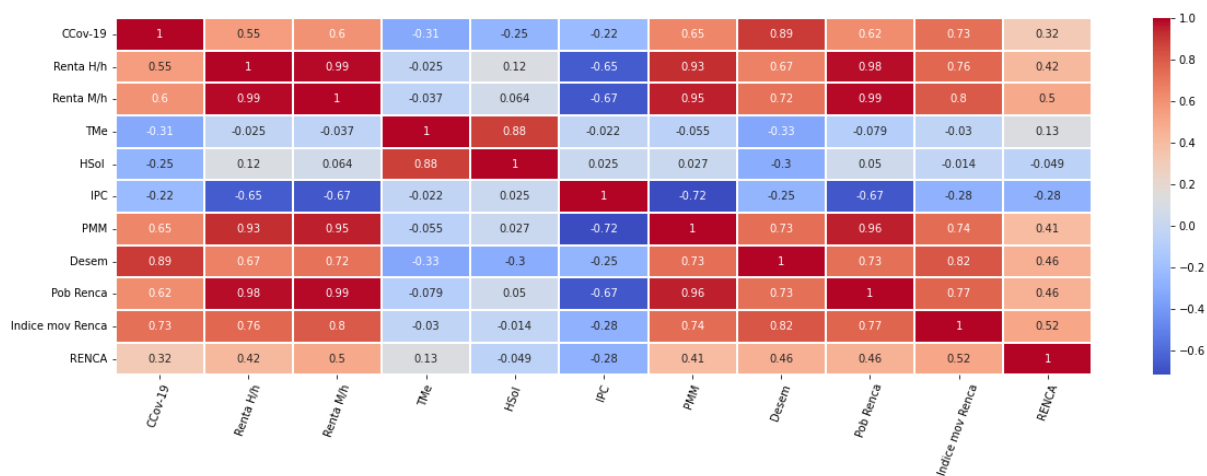


Figura A.14: Matriz de Correlación de San Bernardo

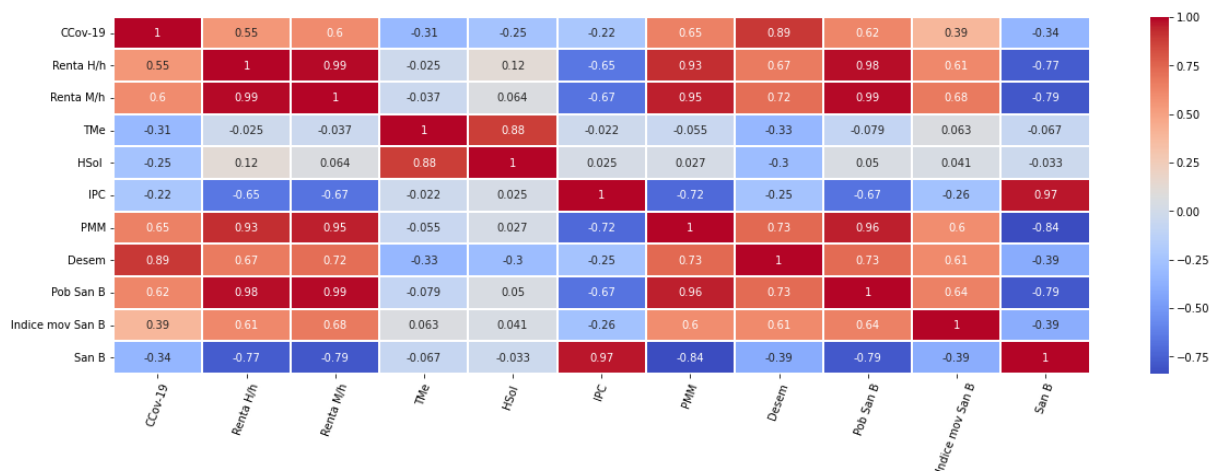


Figura A.15: Matriz de Correlación de San Jose de Maipo

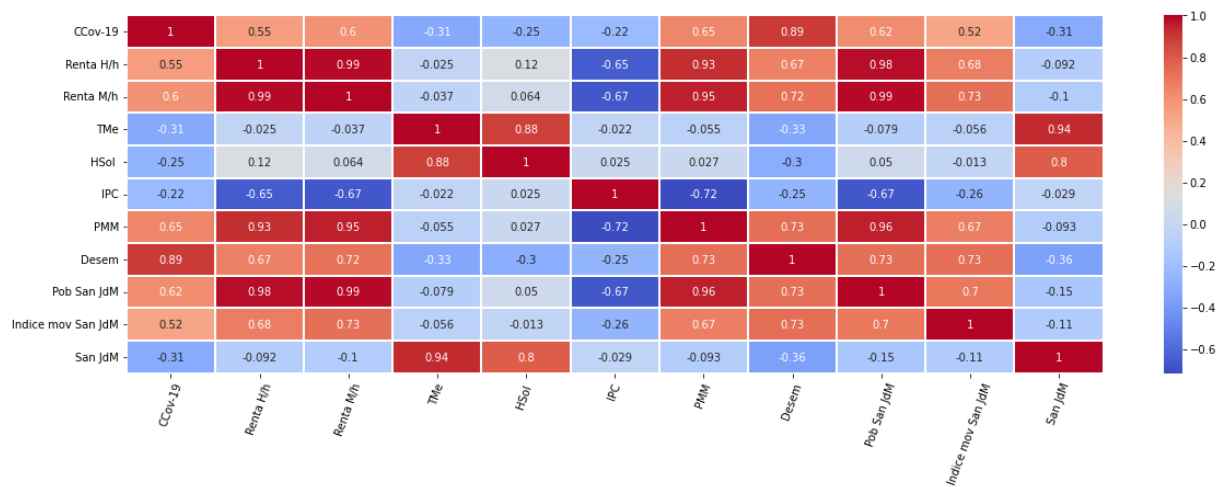


Figura A.16: Matriz de Correlación de San Pedro

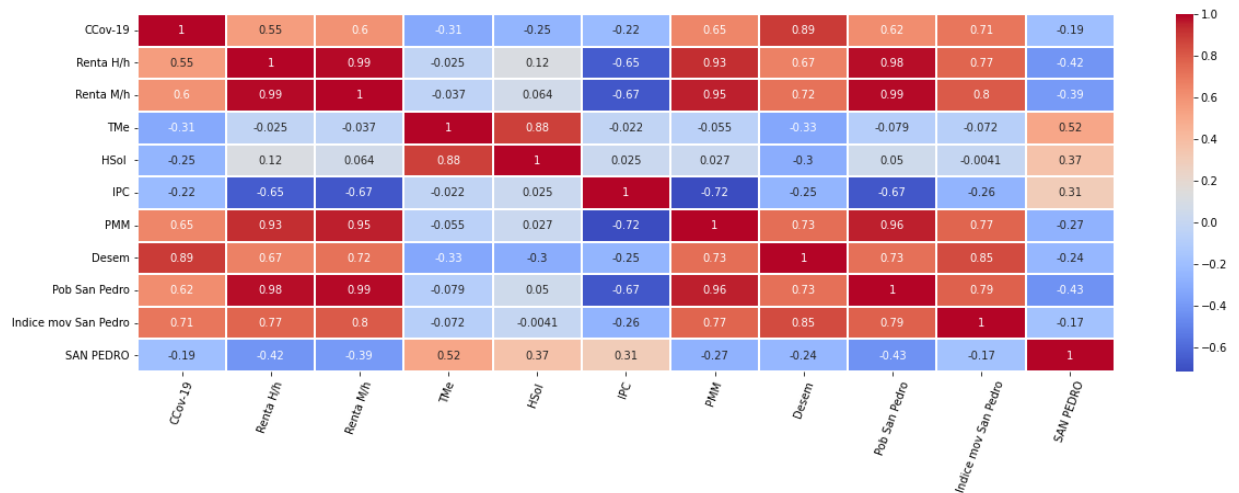


Figura A.17: Matriz de Correlación de Santiago

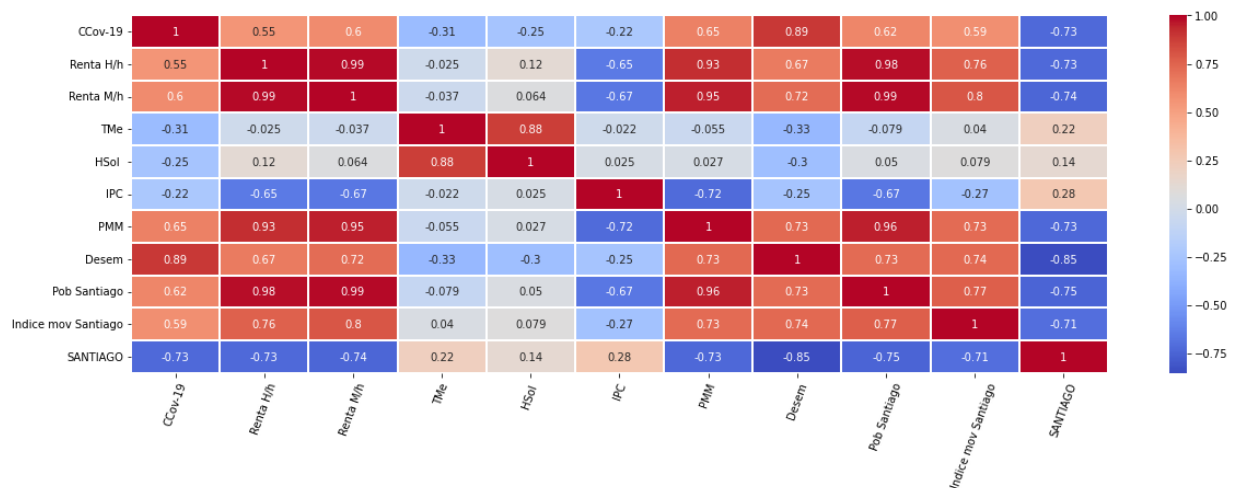
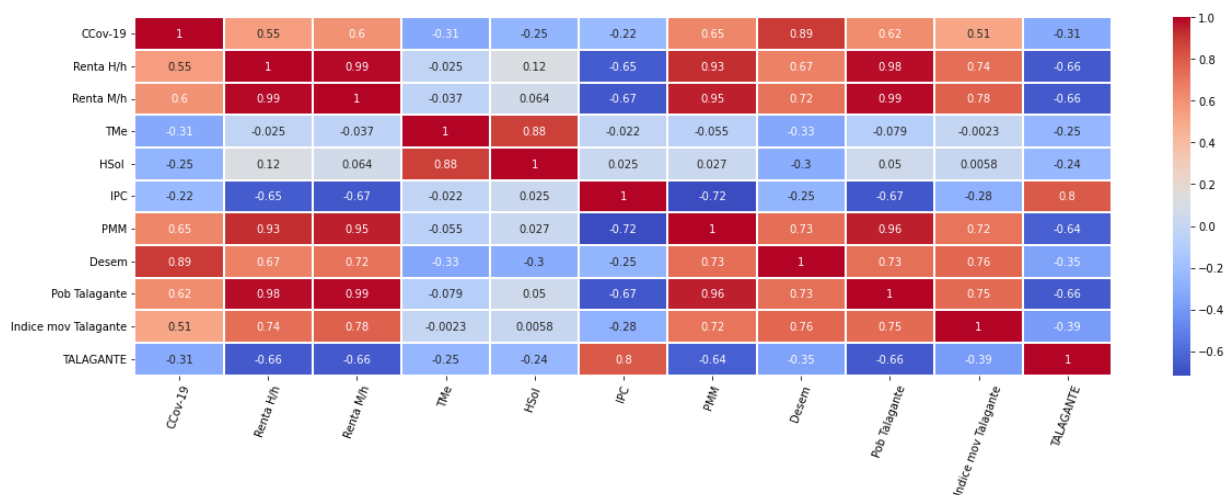
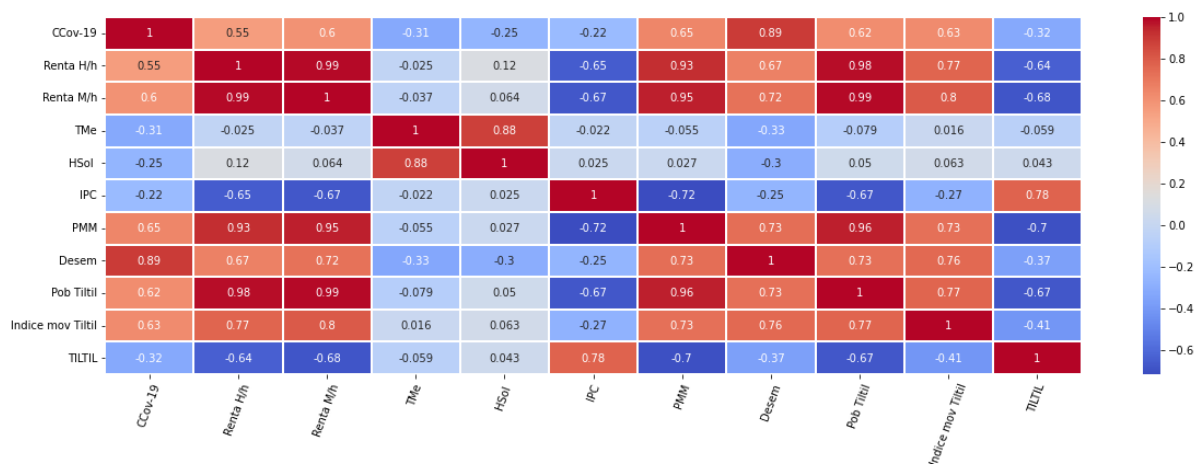


Figura A.18: Matriz de Correlación de Talagante**Figura A.19:** Matriz de Correlación de Til Til



A.2. Gráficos Consumo Eléctrico v/s Renta Mujer y Población

Figura A.20: Consumo v/s Renta M/h y Pob Alhue

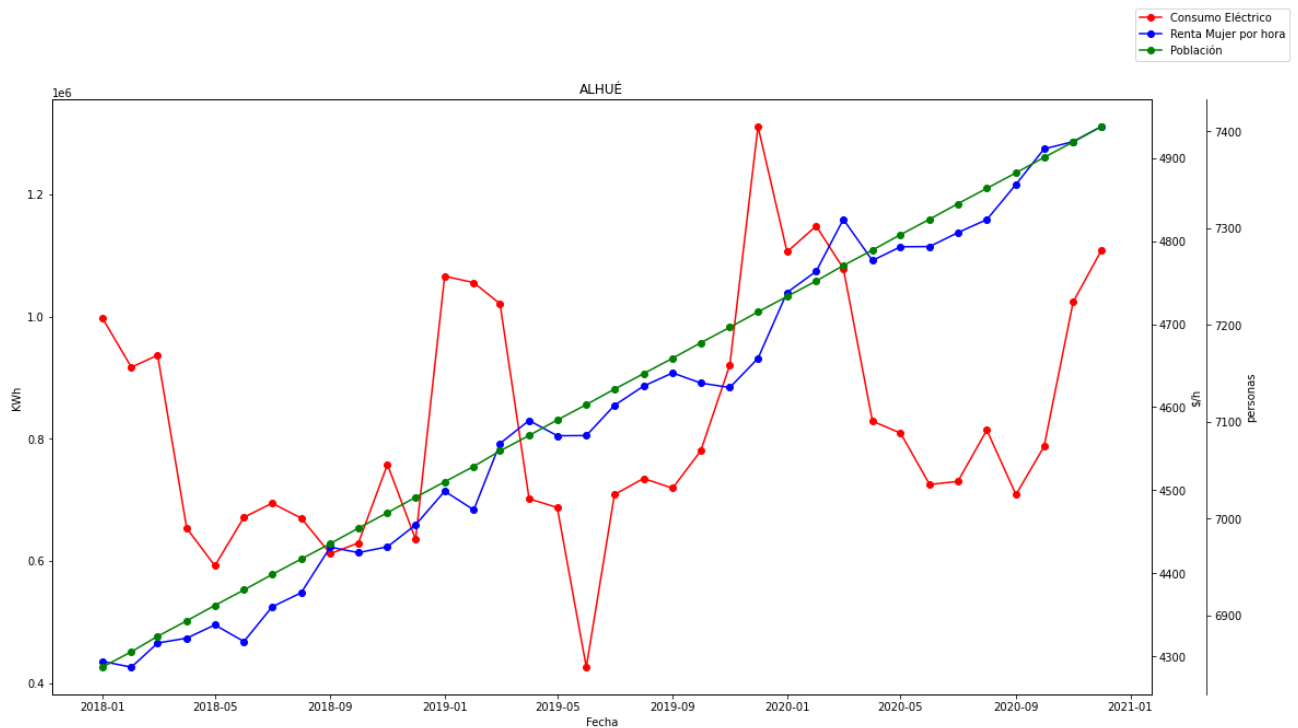


Figura A.21: Consumo v/s Renta M/h y Pob Buin

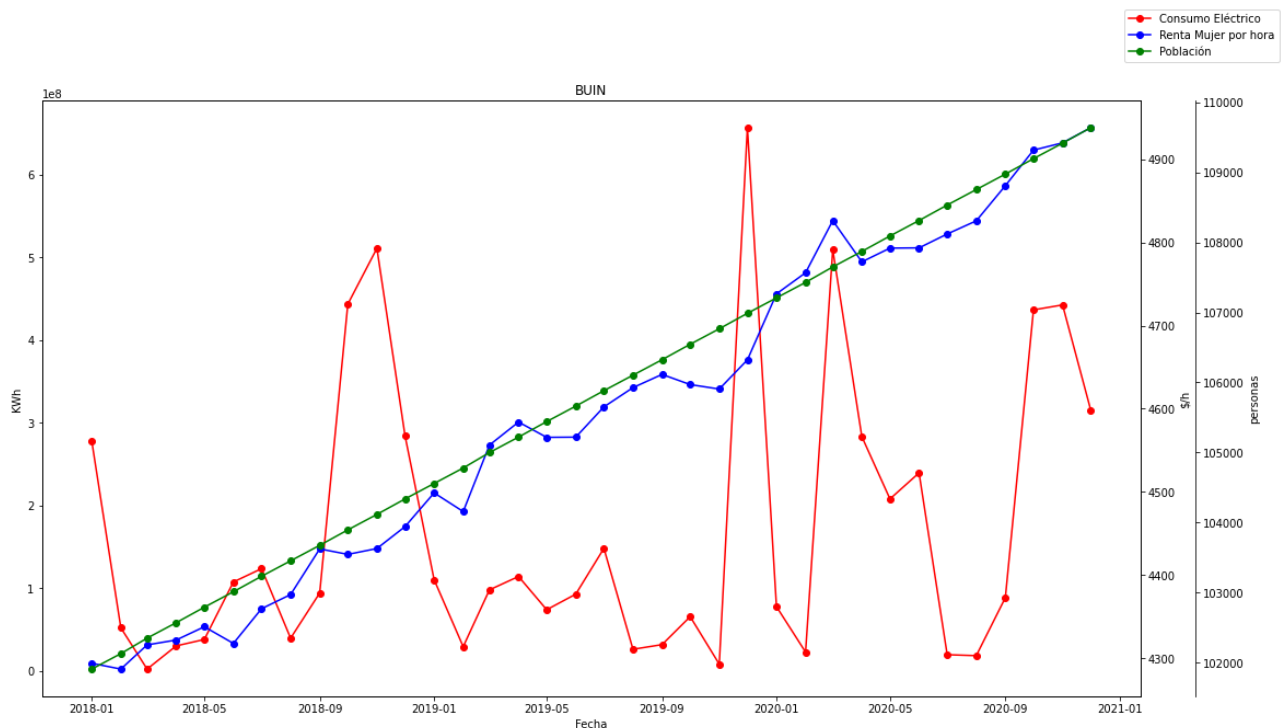


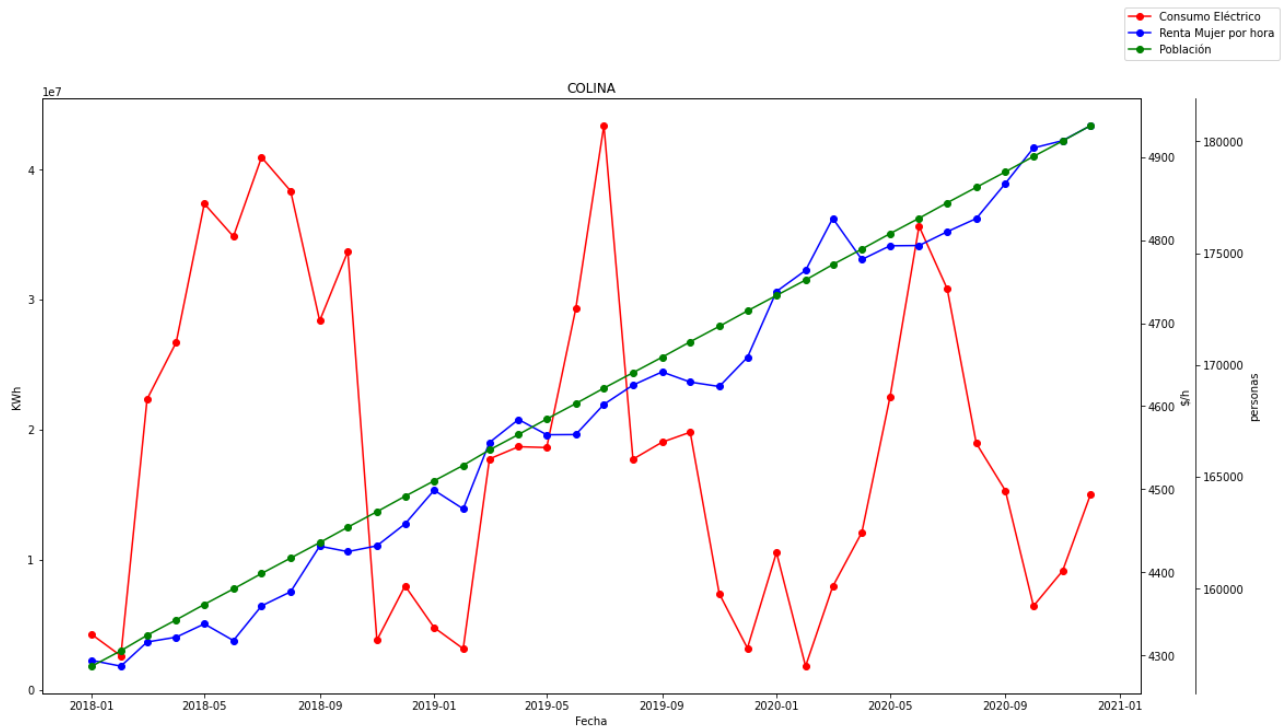
Figura A.22: Consumo v/s Renta M/h y Pob Cerro Navia**Figura A.23:** Consumo v/s Renta M/h y Pob Colina

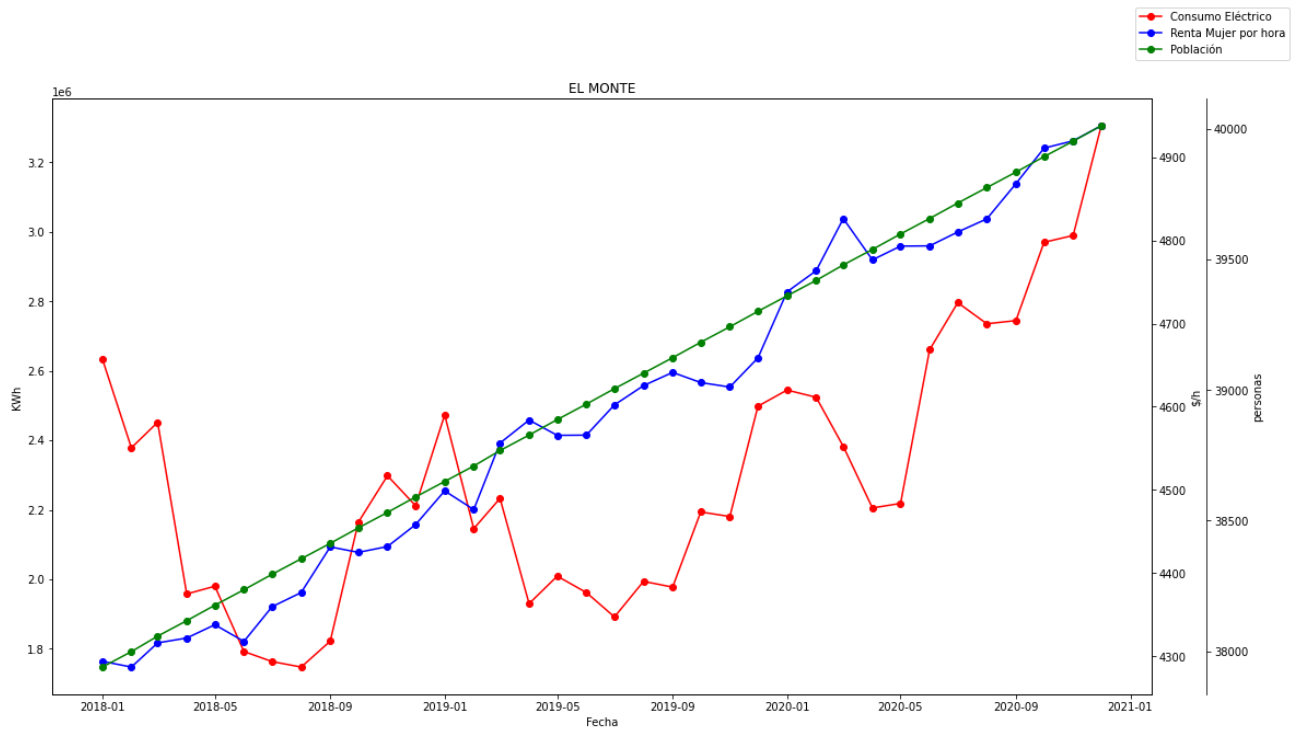
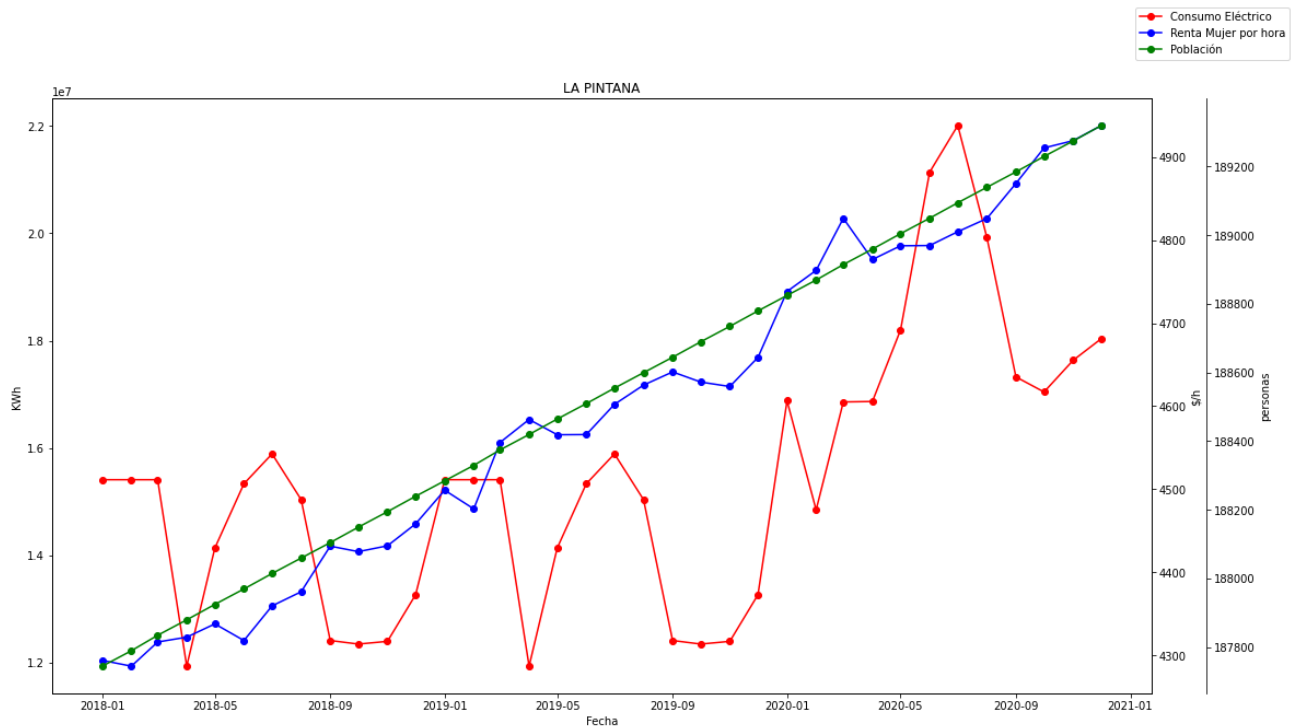
Figura A.24: Consumo v/s Renta M/h y Pob El Monte**Figura A.25:** Consumo v/s Renta M/h y Pob La Pintana

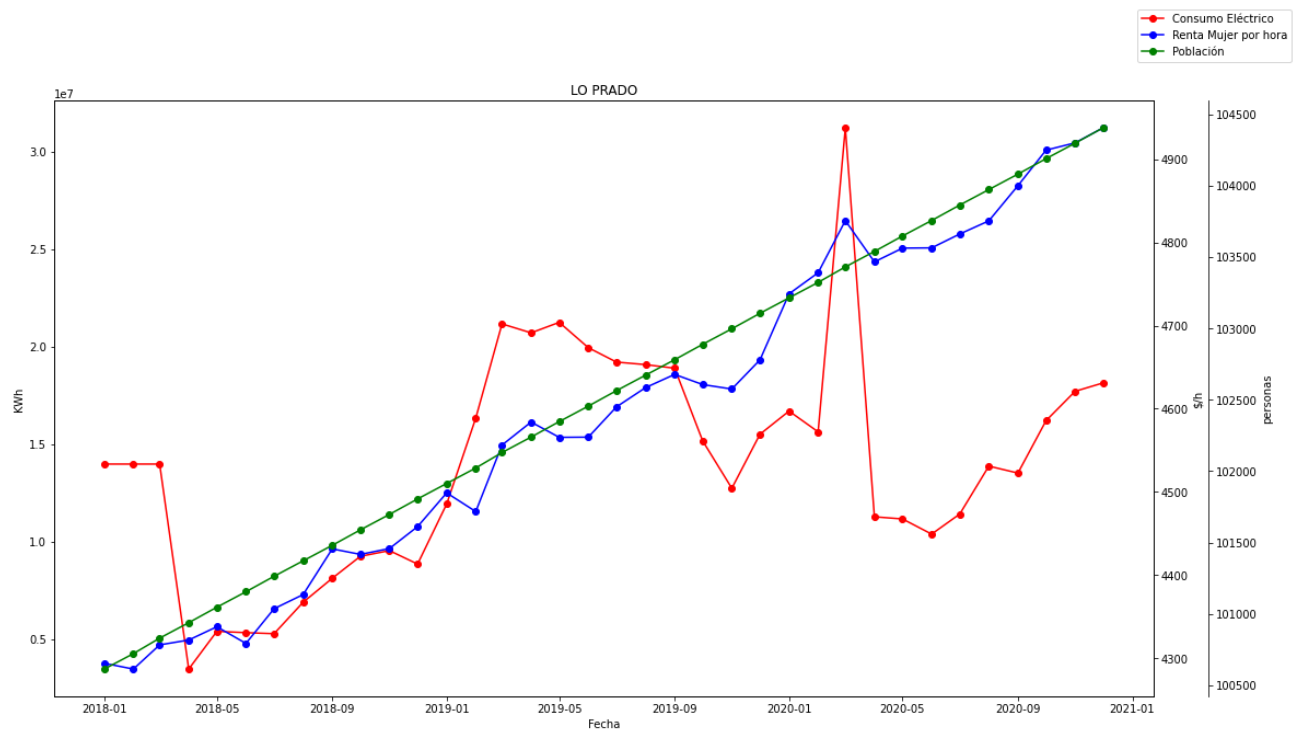
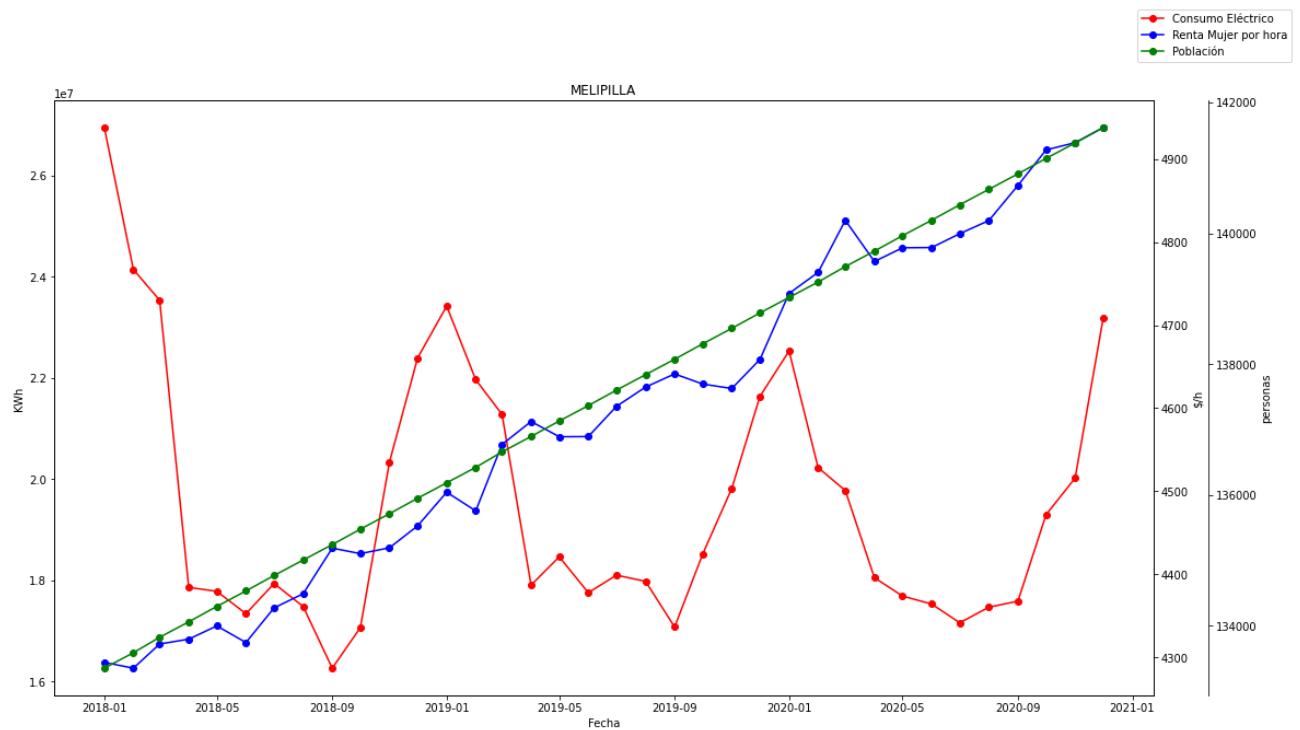
Figura A.26: Consumo v/s Renta M/h y Pob Pirque**Figura A.27:** Consumo v/s Renta M/h y Pob Buin

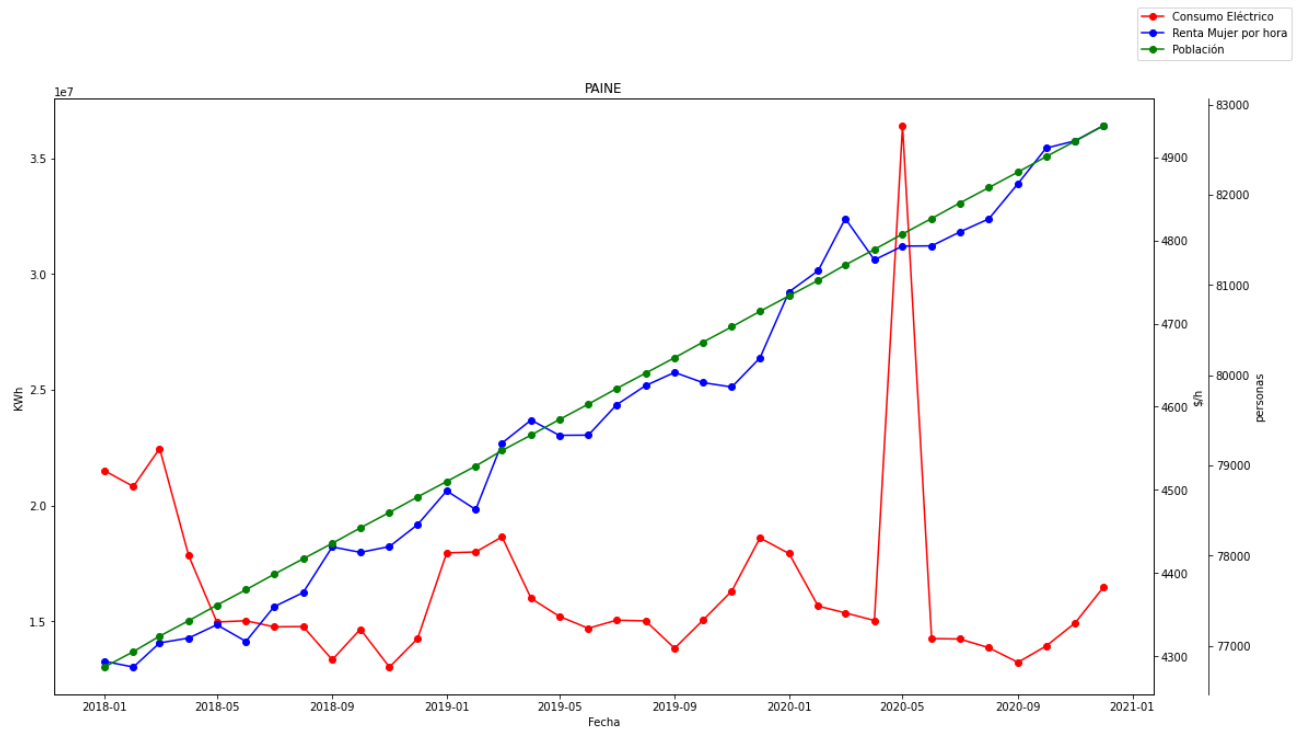
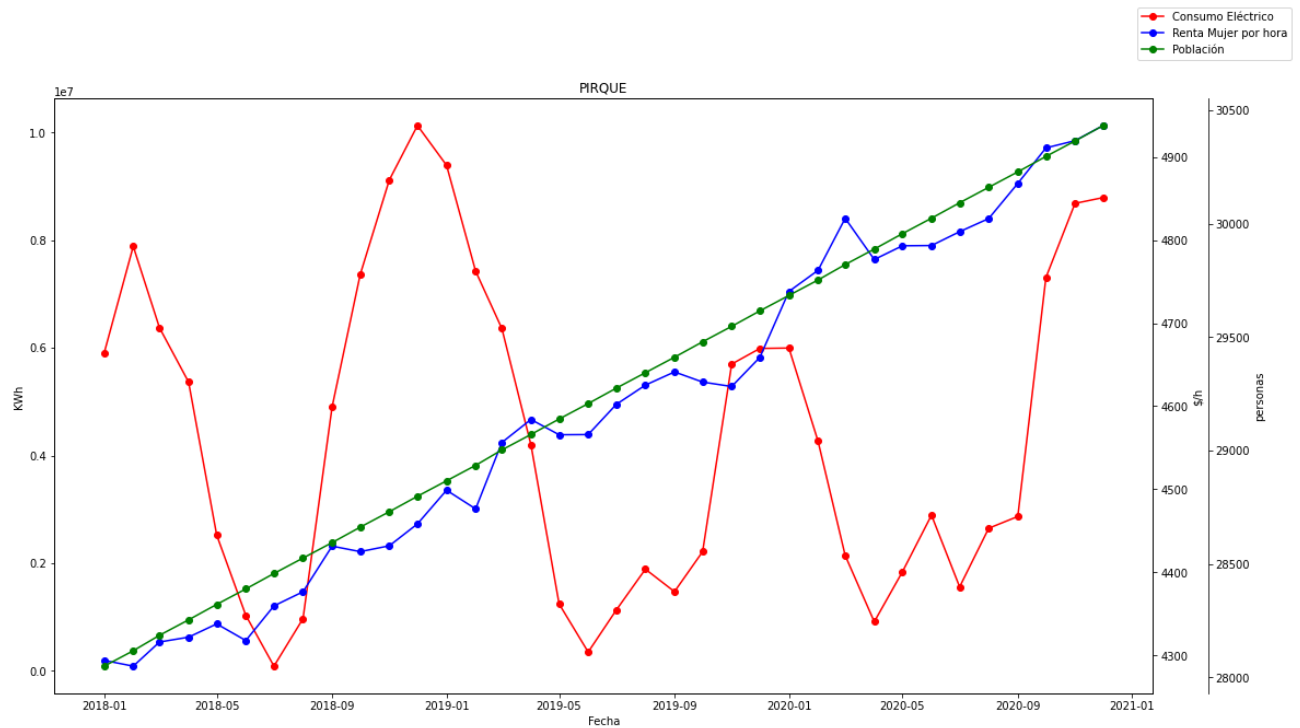
Figura A.28: Consumo v/s Renta M/h y Pob Alhue**Figura A.29:** Consumo v/s Renta M/h y Pob Buin

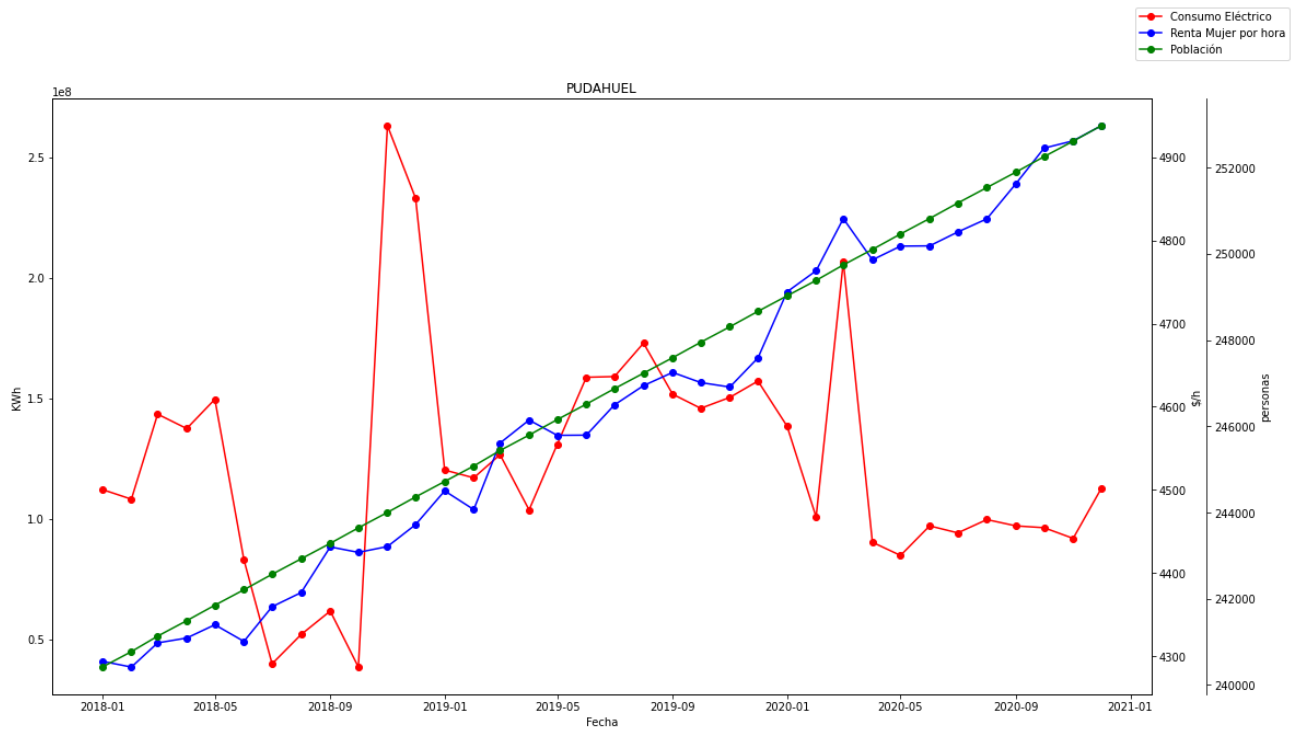
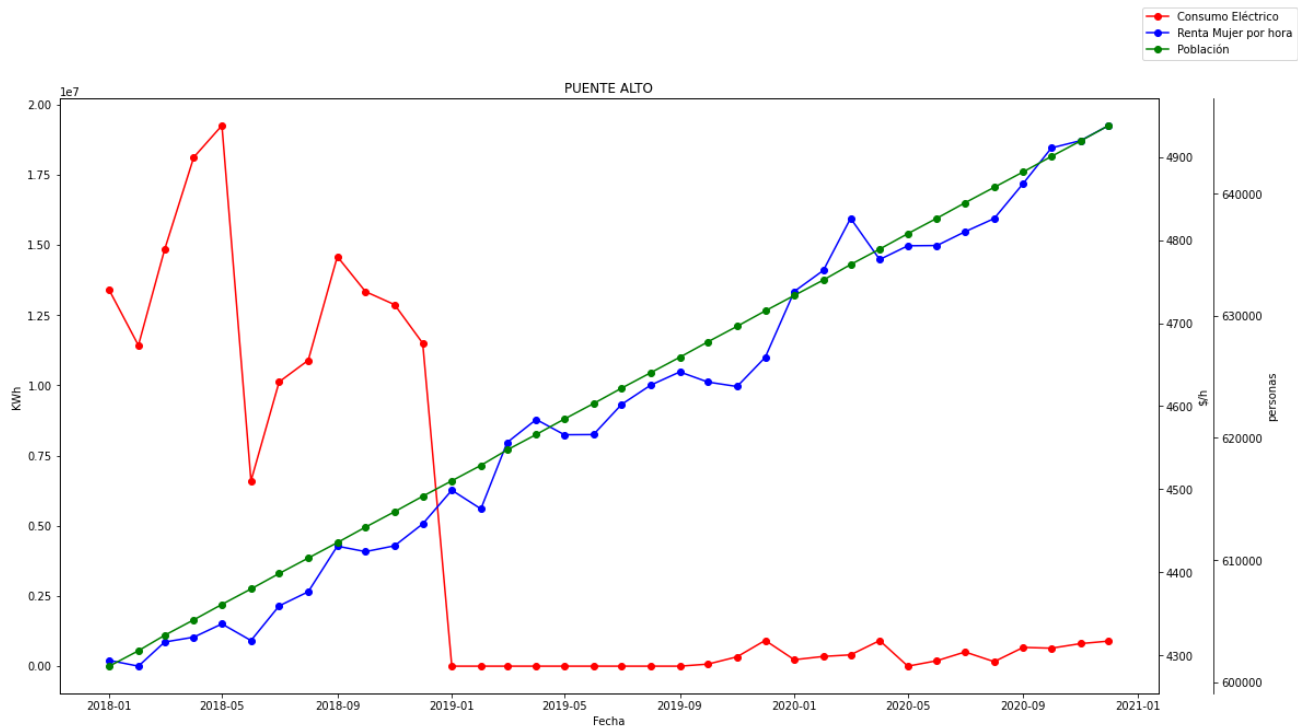
Figura A.30: Consumo v/s Renta M/h y Pob Pudahuel**Figura A.31:** Consumo v/s Renta M/h y Pob Puente Alto

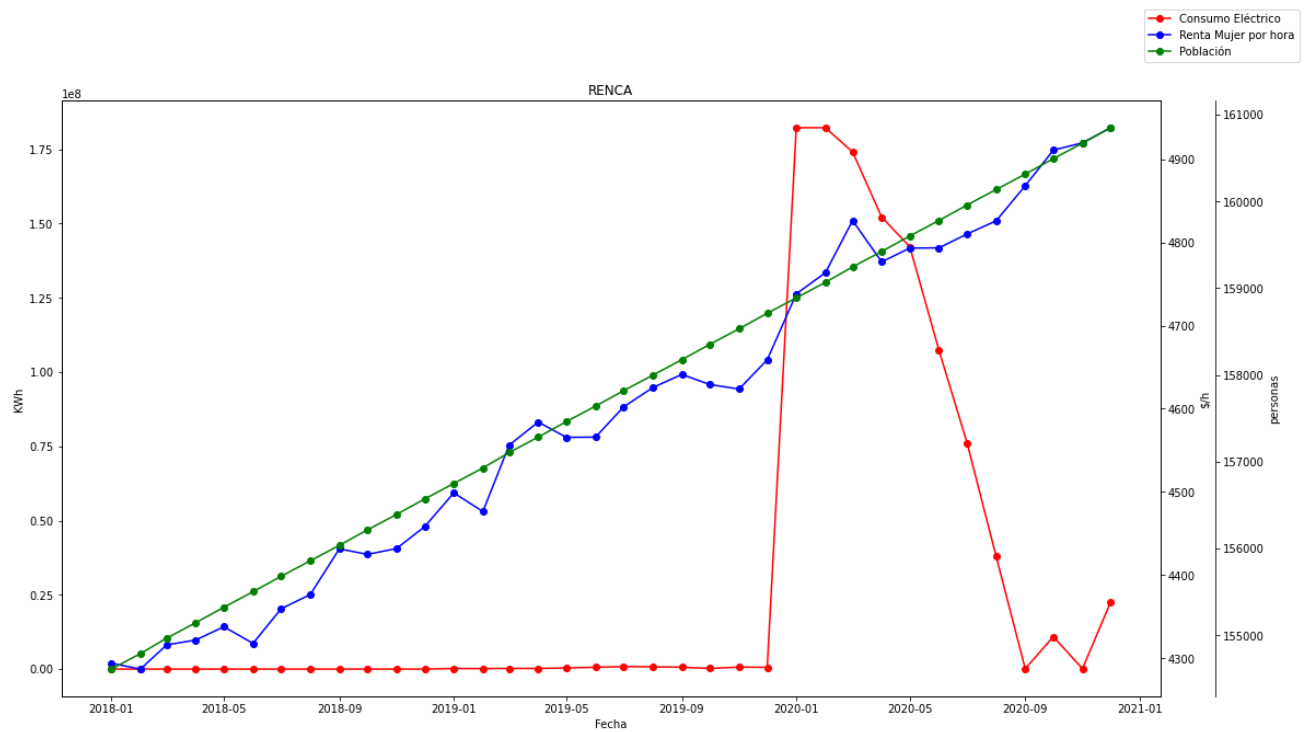
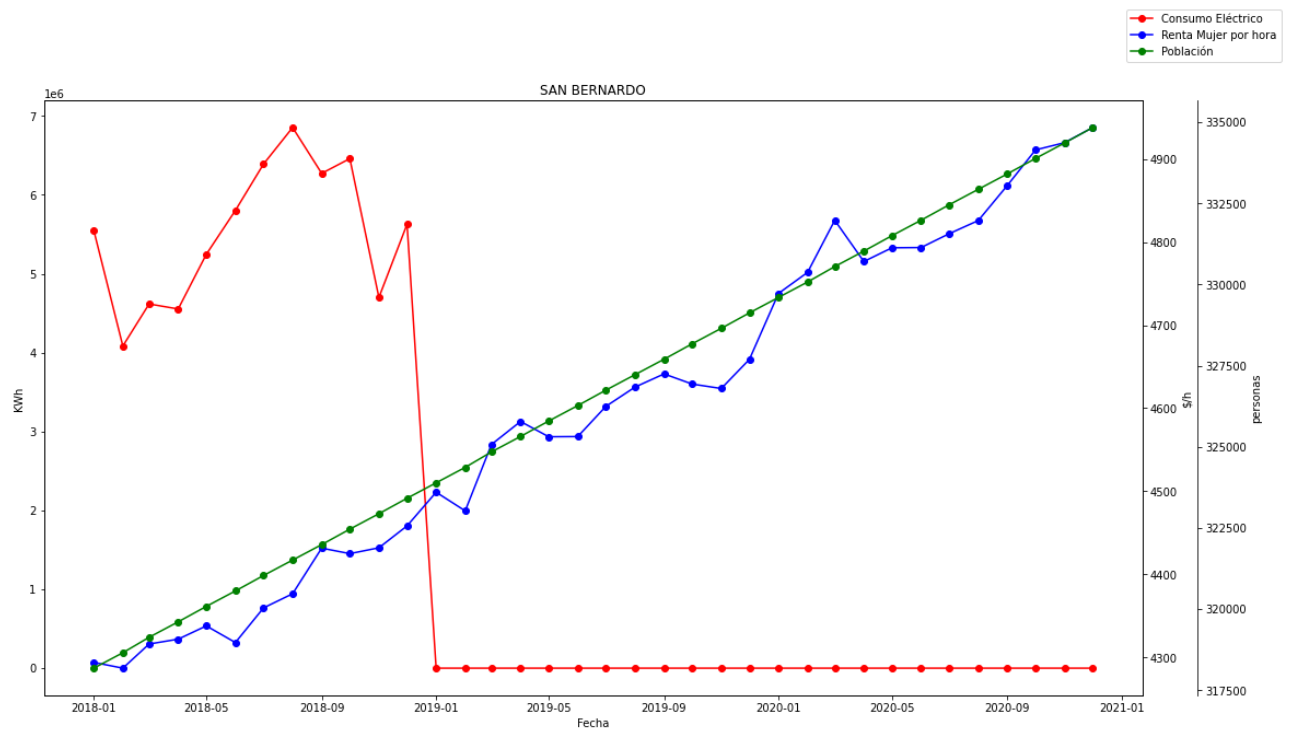
Figura A.32: Consumo v/s Renta M/h y Pob Renca**Figura A.33:** Consumo v/s Renta M/h y Pob San Bernardo

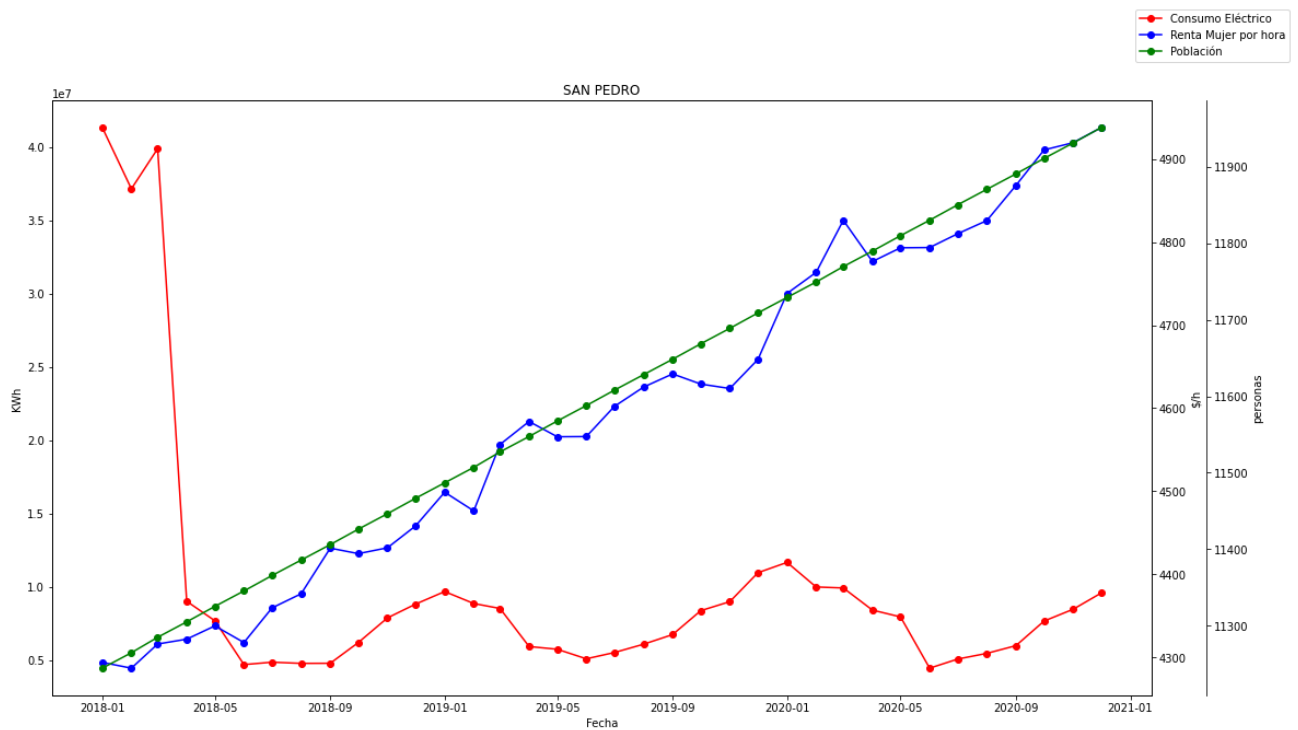
Figura A.34: Consumo v/s Renta M/h y Pob San José de Maipo**Figura A.35:** Consumo v/s Renta M/h y Pob San Pedro

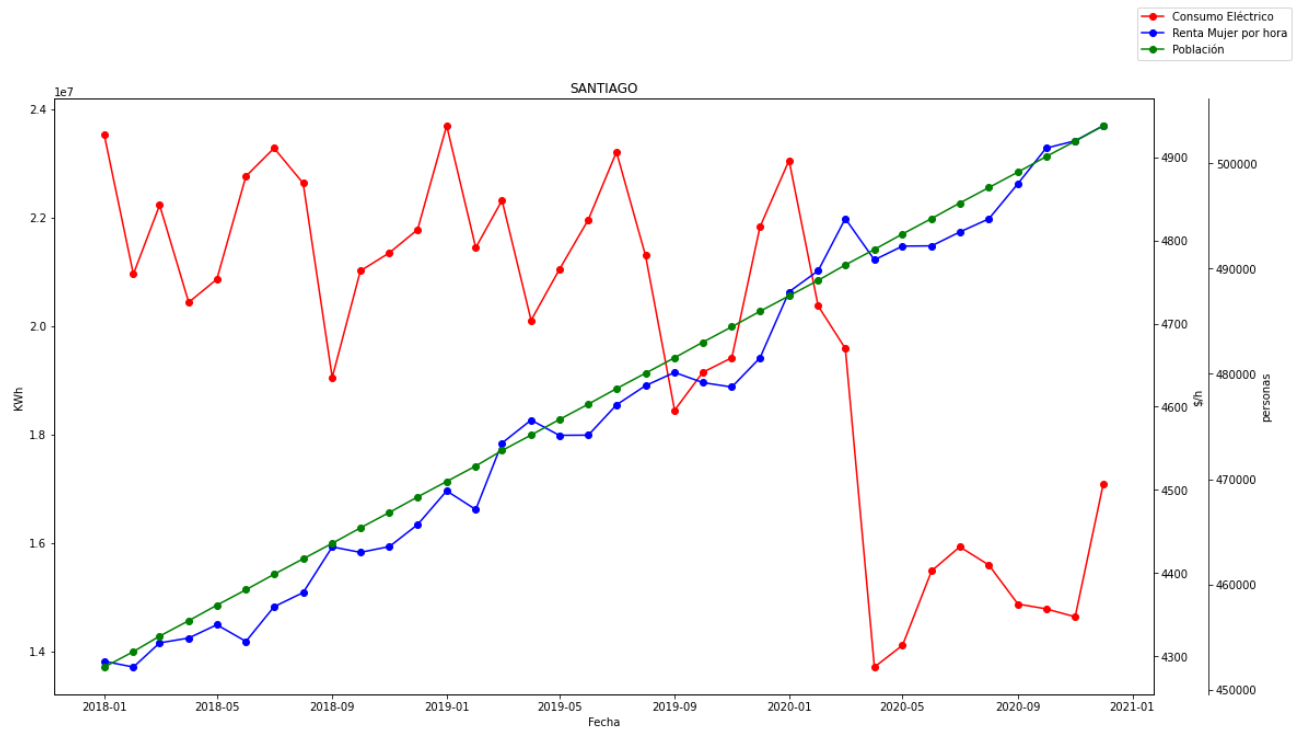
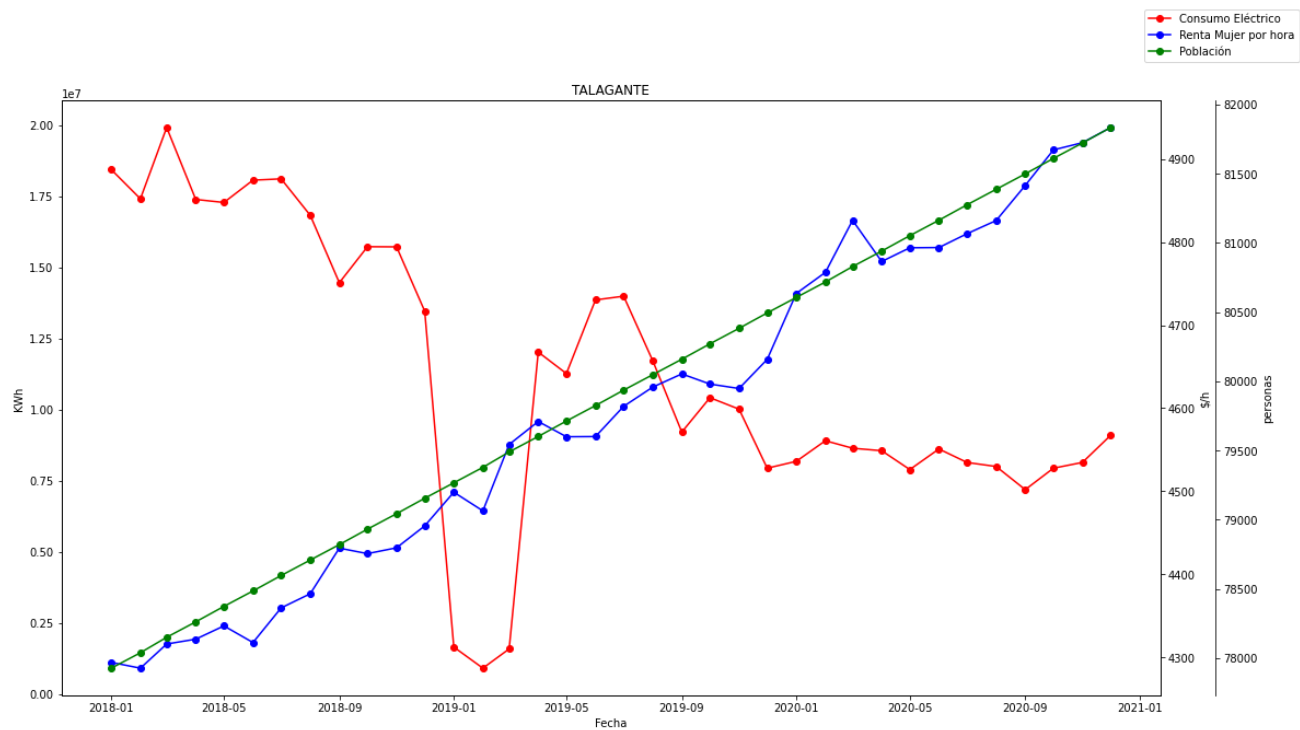
Figura A.36: Consumo v/s Renta M/h y Pob Santiago**Figura A.37:** Consumo v/s Renta M/h y Pob Buin

Figura A.38: Consumo v/s Renta M/h y Pob TilTil

A.3. Diferencia entre escalados

Tabla A.1: Diferencia de escalados MinMax y Robust, para Alhué

$\hat{y}_{pred}^{MinMax} - \hat{y}_{pred}^{Robust}$
-2,33E-09
-1,99E-08
-2,56E-09
-9,78E-09
-6,75E-09
-1,13E-08
-6,52E-09
-8,15E-10
-2,79E-09
-1,51E-09

Tabla A.2: Diferencia de escalados Robust y MaxAbs, para Alhué

$Y_{pred}^{robust} - Y_{pred}^{MaxAbs}$
5,49E-07
1,37E-07
2,40E-07
2,88E-07
4,48E-08
6,92E-07
3,15E-07
3,16E-07
3,38E-07
2,30E-07