UNIVERSIDAD DE LA REPÚBLICA URUGUAY

FACULTAD DE INGENIERÍA

Predicción de consumo eléctrico en Montevideo

Informe de Proyecto de Grado presentado por

Ignacio Otero, Agustín Ruíz Díaz, Nicolás San Martín

en cumplimiento parcial de los requerimientos para la graduación de la carrera de Ingeniería en Computación de Facultad de Ingeniería de la Universidad de la República

Supervisor

Guillermo Moncecchi

Montevideo, 1 de septiembre de 2023

Agradecimientos

Queremos agradecer a todas las personas que nos han ayudado en la realización del proyecto. Agradecemos a los integrantes del MIEM, Federico Matonte, Camila Cosentino, Alejandra Reyes, quienes nos permitieron hacer uso de sus datos y que siempre estuvieron dispuestos a dar una mano con el proyecto. Agradecemos a Guillermo Moncecchi, nuestro tutor, por guiarnos en las distintas etapas de la tesis y por siempre darnos para adelante. Agradecemos a Sebastián García Parra, que se interesó en el aspecto de causalidad de nuestro proyecto y con quien pudimos conversar sobre el tema. Agradecemos a Madeleine Renom, que nos orientó sobre el uso y obtención de datos meteorológicos de INUMET y por último, queremos agradecer a nuestras familias y amigos, por todo el apoyo e interés que nos brindaron para realizar nuestra tarea.

Resumen

Este proyecto se enfoca en el análisis y la predicción del consumo eléctrico en Uruguay, específicamente en el departamento de Montevideo. Se busca recopilar e integrar distintas fuentes de información de la región que puedan ser de utilidad para esta tarea. Principalmente interesan datos socioeconómicos, meteorológicos y de consumo eléctrico, los cuales se obtienen de la Encuesta Continua de Hogares (ECH), el Instituto Uruguayo de Meteorología (INUMET) y la Administración Nacional de Usinas y Trasmisiones Eléctricas (UTE) respectivamente. Con esta información se busca predecir mensualmente el consumo eléctrico diario promedio por hogar en distintas regiones de Montevideo, denominadas como segmentos. Para realizar las predicciones se pueden utilizar diversos modelos de aprendizaje, en nuestro proyecto se decide utilizar Regresión Lineal, *K-nearest neighbors* (KNN) y *Random Forest*, siendo este último el modelo que presenta mejores resultados con un MAPE de 5.6 % y un R^2 de 0.85. También resulta interesante analizar el impacto que tienen las variables en la predicción, por lo que se realiza un estudio causal en el que se busca determinar si realmente causan aumentos o disminuciones en el consumo eléctrico. Los resultados de este estudio sugieren, a partir del modelo de la realidad propuesto, que los ingresos de los hogares y vivir en una casa son factores que causan diferencias en el consumo eléctrico

Palabras clave: Predicción de consumo eléctrico, Aprendizaje Automático, Causalidad

Índice general

Capítulo 1

Introducción

En el mundo moderno es imposible imaginar la vida sin energía eléctrica, ya que está relacionada con todo tipo de actividad humana; desde la industria, el transporte, la ciencia y la agricultura, hasta la vida cotidiana de la gran mayoría de la población. La generación de electricidad es una de las bases para el desarrollo de la sociedad.

Las predicciones de consumo eléctrico son de gran importancia, ya que esto permite a las organizaciones que proveen electricidad generar la cantidad necesaria. Las predicciones a corto plazo intentan determinar qué tanta electricidad se deberá generar con la infraestructura actual para su consumo cercano, y las predicciones a largo plazo ayudan a las organizaciones a evaluar si deben mejorar su infraestructura, qué tanto y en qué momento hacerlo.

Tanto la demanda como la producción eléctrica han aumentado drásticamente desde su origen hasta el día de hoy en todo el mundo. Un ejemplo de esto es el consumo eléctrico per cápita en Uruguay, que se ha sextuplicado en el lapso de 1965 a la actualidad (Ministerio de Industria, 2020a). Por este motivo resulta de gran importancia poder hacer predicciones precisas en esta área tan cambiante. De esta forma se podrían detectar tendencias de consumo en determinadas regiones y poder generar un plan de acción en cuanto a la cantidad de energía será necesario producir, en qué sectores se deberá invertir, etc.

Dentro de los componentes del consumo eléctrico el sector residencial suele ser el más grande en muchos países, como en Estados Unidos, Gran Bretaña (Ma y Cheng, 2016) o Uruguay (Ministerio de Industria, 2020b).

Existen muchos factores que impactan el consumo eléctrico. Desde factores ambientales, hasta sociales. Todos estos tienen diferente impacto relativo dependiendo del período sobre el cual se quiera predecir el consumo eléctrico.

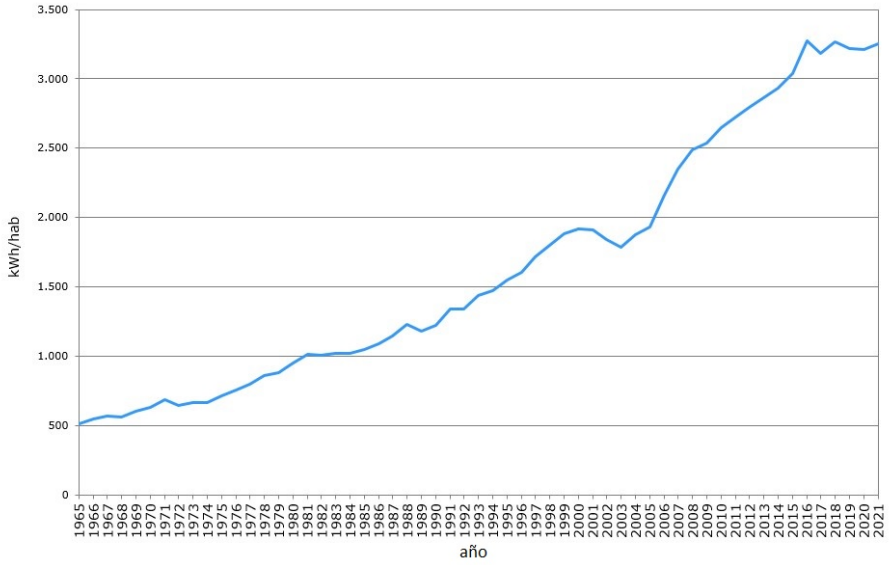

Figura 1.1: Demanda de consumo eléctrico anual en Uruguay (Ministerio de Industria, 2020a)

1.1. Objetivos y alcance

El primer objetivo de este trabajo es determinar la información con la que se cuenta para llevar a cabo la predicción. Es necesario realizar un correcto análisis y una selección de los datos, ya que de esto dependerá la precisión de los resultados que se obtengan. Esto es responder a la pregunta: "¿Qué datos que pueden tener relación con el consumo eléctrico están disponibles en Uruguay?"

Dados los datos disponibles, el segundo objetivo consiste en determinar qué predecir exactamente. Esto incluye el lapso y el área geográfica que se definen para realizar las predicciones de consumo eléctrico. Se responderá a la pregunta: "¿Qué tiene sentido predecir, sobre el consumo eléctrico, en Uruguay?"

Una vez que se hayan analizado los datos con los que se cuentan y determinado el enfoque, se deberá definir cómo realizar la predicción. Para esto se hará uso de modelos de aprendizaje automático, decisión que surge de la literatura y el estado del arte estudiado. Será necesario determinar cuáles métodos utilizar y las métricas de rendimiento con las cuales se evaluarán.

Además de realizar las predicciones, es de interés poder estudiar la importancia que tienen las diferentes variables utilizadas, si el modelo de aprendizaje automático permite determinarlas.

A su vez, relacionado al objetivo anterior, será de interés determinar el efecto causal entre las distintas variables estudiadas y el consumo eléctrico. Es por esto que se realizará un pequeño estudio de causalidad, un proceso en el cual se intenta determinar si una variable realmente causa el consumo eléctrico, o simplemente aparenta estar relacionada. El objetivo de este experimento es solamente aproximarse al problema de la causalidad a partir de los datos disponibles, ya que este es un campo emergente y novedoso.

1.2. Estructura

El informe está estructurado de forma tal que el capítulo 2 se centra la revisión de antecedentes, sintetizando literatura vinculada a la predicción de consumo eléctrico y las métricas comúnmente utilizadas para evaluarla.

El capítulo 3 hace un recorrido por el proceso mediante el cual se obtuvieron los distintos conjuntos de datos utilizados en la experimentación, en qué consisten (dando ejemplos de ellos). A su vez se presenta una sección sobre la ingeniería de atributos que se realiza sobre estos conjuntos. Se realiza un control de calidad de los datos, para comprobar que los datos sean adecuados antes de utilizarlos para la predicción. En las últimas dos secciones de este capítulo se integran los datos, y se presentan visualizaciones para tener una mejor comprensión sobre ellos.

En el capítulo 4 se plantea explícitamente la discusión sobre qué predecir dentro del consumo eléctrico, definiendo plazo y lugar geográfico. Luego se definen los modelos que se van a utilizar para la predicción, y líneas base para comparar su rendimiento.

El capítulo 5 contiene el análisis experimental, conteniendo la ejecución del aprendizaje y predicción de los modelos, con sus respectivos resultados. También se presentan conclusiones de los resultados obtenidos.

El capítulo 6 es un capítulo autocontenido sobre causalidad. Aquí se presenta un marco teórico para comprender de qué se trata este tema, a su vez que una introducción a la biblioteca *Dowhy* que será utilizada para el estudio causal. Se plantea un ejemplo práctico simple, donde se aplican los conocimientos presentados en el caso de la predicción de energía eléctrica.

Finalmente, el capítulo 7 está destinado a las conclusiones del proyecto y presenta un resumen de lo trabajado.

Capítulo 2

Revisión de antecedentes

Ya presentado el enfoque de este proyecto, en este capítulo se presenta una investigación documental que permite el estudio del conocimiento acumulado en el área de la predicción del consumo eléctrico, para determinar tendencias de investigación y para tener un punto de partida para tomar decisiones sobre las distintas interrogantes que se plantearon en los objetivos. Se documentará lo relevado sobre los siguientes puntos:

- Problemas que se resuelven: se discutirán los diferentes objetivos que intentan conseguir los investigadores. Más precisamente, esta discusión resulta en la variable que intentan predecir con sus cálculos y si se predice a corto, mediano o largo plazo.
- Datos que se disponen: se investigará qué datos fueron utilizados en predicciones de consumo eléctrico y cuáles de ellas se han considerado más relevantes.
- Métodos que se utilizan: se mencionarán los métodos de aprendizaje aplicados, buscando analizar los resultados obtenidos con cada uno, así como sus ventajas y desventajas.

El relevamiento parte del trabajo de Mosavi y Bahmani (2019), recomendado por miembros del MIEM, que recopila información sobre una gran cantidad de trabajos relacionados a la predicción de consumo energético, como en varios de los allí citados.

2.1. Medidas de rendimiento

Antes de presentar el análisis de los trabajos relacionados en esta sección se definirán y presentarán distintas medidas de rendimiento, comúnmente utilizadas. Estas medidas calculan el error de las predicciones, que determina qué tan cercanas fueron a los valores reales.

Se define la siguiente nomenclatura:

- $EEC(t)$: predicción del consumo eléctrico que se obtuvo para un tiempo t
- $EEC(t)_{observado}$: valor real de consumo eléctrico para un tiempo t
- σ_X^2: varianza de la variable X
- σ_{XY}: covarianza de las variables X e Y

$\mathbf{R}^2$

El coeficiente de determinación (R^2) es una medida estadística que determina la calidad del modelo para replicar los resultados y la proporción de variación de los resultados que puede explicarse por el modelo, o sea R^2 = variación explicada/variación total. Un mayor valor de R^2 es positivo, y generalmente indica que el modelo se ajusta bien a los datos. El valor máximo de R^2 es 1 y el mínimo es 0.

Su fórmula para los casos de regresión lineal es:

$$R^2 = \frac{\sigma_{XY}^2}{\sigma_X^2 \sigma_Y^2}$$

APE

El Error Porcentual Absoluto (APE o *Absolute Percentage Error)* es un indicador del desempeño que mide el tamaño del error (absoluto) en términos porcentuales. El hecho de ser una medida de error porcentual es la razón de porque se suele utilizar tanto cuando se elaboran predicciones, debido a su fácil interpretación. Su fórmula es:

$$APE = \frac{|EEC(t)_{observado} - EEC(t)|}{EEC(t)_{observado}} \times 100$$

MAPE

El Error Porcentual Absoluto Medio (MAPE o *Mean Absolute Percentage Error*) es una medida que consta en tomar el promedio de los APE para los distintos intervalos de tiempo en los que se realizan las predicciones. Por ejemplo, si se hicieron predicciones para todo un año, mes a mes, cada intervalo de tiempo sería un mes. Su fórmula es la siguiente:

$$MAPE = \frac{1}{N}\sum_{t=1}^{N}\frac{|EEC(t)_{observado} - EEC(t)|}{EEC(t)_{observado}} \times 100$$

WAPE

El Error Porcentual Absoluto Ponderado (WAPE o *Weighted Average Percentage Error*) es una métrica de la suma del error absoluto normalizado por la demanda total. Esta medida penaliza de forma equitativa predicciones por debajo o por encima de lo esperado, favoreciendo errores en predicciones de bajo peso relativo. Se utiliza la media del valor esperado de la predicción para calcular el error absoluto. Esta medida se recomienda utilizar en los casos en que no hay diferencia en costos en caso de predecir por debajo o por encima del valor real. Su fórmula es la siguiente:

$$WAPE = \frac{\sum_{t=1}^{N}|EEC(t)_{observado} - EEC(t)|}{\sum_{t=1}^{N}EEC(t)_{observado}} \times 100$$

MSE

El Error Cuadrático Medio (MSE o *Mean Squared Error*) es una medida de dispersión del error de predicción. Esta medida maximiza el error al elevar al cuadrado, castigando los periodos en los que la diferencia fue más alta que en otros. Su fórmula es:

$$MSE = \frac{1}{N}\sum_{t=1}^{N}(EEC(t)_{observado} - EEC(t))^2$$

RMSE

La Raíz del Error Cuadrático Medio (RMSE o *Rooted Mean Squared Error*) es una medida muy similar a la anterior ya que $RMSE = \sqrt{MSE}$. La raíz se agrega para que la escala de los errores sea igual a la escala de los objetivos. MSE y RMSE son similares en el sentido de que si tenemos dos predicciones A y B se cumple que si $MSE(A) > MSE(B) \longleftrightarrow RMSE(A) > RMSE(B)$ y también funciona en la dirección opuesta.

Es por esta razón que no se suelen utilizar ambas medidas a la vez, sino se suele trabajar con una de las dos y en general se elige MSE, ya que es más simple.

Su fórmula es:

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(EEC(t)_{observado} - EEC(t))^2} = \sqrt{MSE}$$

MAE

El Error Absoluto Medio (MSE o *Mean Absolute Error*) es el promedio de la diferencia absoluta entre el valor observado y los valores de predicción. Todas las diferencias individuales se ponderan por igual en el promedio. Su fórmula es:

$$MAE = \frac{1}{N}\sum_{t=1}^{N}|EEC(t)_{observado} - EEC(t)|$$

2.2. Predicción de consumo eléctrico

Cuando se habla de predicciones en el consumo eléctrico, resulta importante determinar el lapso para el cual se realizan estas predicciones, debido a que los objetivos y variables que se tienen en cuenta serán diferentes dependiendo del plazo. Shao y cols. (2017) definen las siguientes tres categorías:

- Predicción de carga de largo plazo (*Long-term load forecasting*, LTLF): hace referencia a las predicciones de largo plazo. Se utilizan principalmente para predecir el crecimiento en demanda y las expansiones que serán necesarias en los próximos años. Se define que una predicción es a largo plazo si predice en intervalos de un año o más grandes.
- Predicción de carga de medio plazo (*Medium-term load forecasting*, MTLF): predicciones de mediano plazo, principalmente utilizadas para poder plantear cronogramas de abastecimientos y mantenimientos. Se define que es a mediano plazo si se trata de intervalos entre semanas y meses.
- Predicción de carga de corto plazo (*Short-term load forecasting*, STLF): predicciones a corto plazo, utilizadas para tomar acciones sobre cambios inesperados en el consumo, como por ejemplo hacer ajustes en la producción y distribución de la energía. Se define una predicción a corto plazo si los intervalos son de un día o menores.

Los datos utilizados varían mucho dependiendo del plazo de la predicción. Hay estudios que realizan predicciones únicamente con datos meteorológicos y de consumo eléctrico histórico (Lara-Benítez y cols., 2020), y por otro lado hay estudios que hacen uso de 171 variables diferentes (Ma y Cheng, 2016). También existen trabajos cuyo principal propósito es discutir la relevancia de los datos que afectan el consumo eléctrico (Ma y Cheng, 2016; Porse y cols., 2016).

El lapso en el que se realizan las predicciones resulta ser una pieza de información no despreciable para determinar qué métodos y qué datos se utilizan. En las próximas secciones se presenta como éstos varían según las categorías que fueron definidas previamente.

2.2.1. STLF

Sobre los intervalos en los que se predice el consumo eléctrico, dentro de los trabajos categorizados como STLF realizados por otros investigadores, es común predecir en intervalos de horas (Lei y cols., 2021; Wang y cols., 2018; Carpinteiro y cols., 2000) o minutos (Lara-Benítez y cols., 2020; Popoola, 2016).

Dada la magnitud de los datos que se utilizan, generalmente se estima usando pocas variables para la suma de una gran población (Lara-Benítez y cols., 2020; Carpinteiro y cols., 2000), o se estima con muchas variables con objetivo de calcular la electricidad para cada hogar específico (Wang y cols., 2018; Popoola, 2016).

En este tipo de predicciones, usualmente los datos se extraen individualmente de hogares, y se toma un número reducido de estos, desde tan poco como 2 edificios (Wang y cols., 2018) como 102 (Popoola, 2016). Esto es común ya que se suelen medir muchos atributos de los hogares, incluso llegando a colocar sensores en cada uno para mayor precisión de las medidas, por lo que el costo de hacerlo a escalas más grandes sería enorme. Sin embargo, existen estudios como el que se realizó sobre la Red Eléctrica Española (Lara-Benítez y cols., 2020), cuya predicción tiene un enfoque a nivel país. Esto fue posible porque el estudio únicamente tomó como dato el consumo eléctrico y el tiempo.

Es común incorporar datos meteorológicos en estos estudios, ya que el consumo puede variar mucho de un día al siguiente por este tipo de fenómenos. Según un estudio del estado del arte (Shao y cols., 2017) los factores que más impactan en el consumo a corto plazo son los factores meteorológicos y el precio de la electricidad, y según el estudio de Nassif y cols. (2022) cuando se trata de predicciones a corto plazo, los factores más importantes son: variables de tiempo (temperatura, humedad, etc..), crecimiento de la población y el tipo de día (día de semana o fin de semana).

En cuanto a los métodos y técnicas que más se utilizan para las predicciones a corto plazo, se entrará más en detalle en el estudio realizado por Nassif y cols.

(2022). Este estudio analiza 240 trabajos, recompilando las técnicas que más se utilizan y con las que mejores resultados se obtienen.

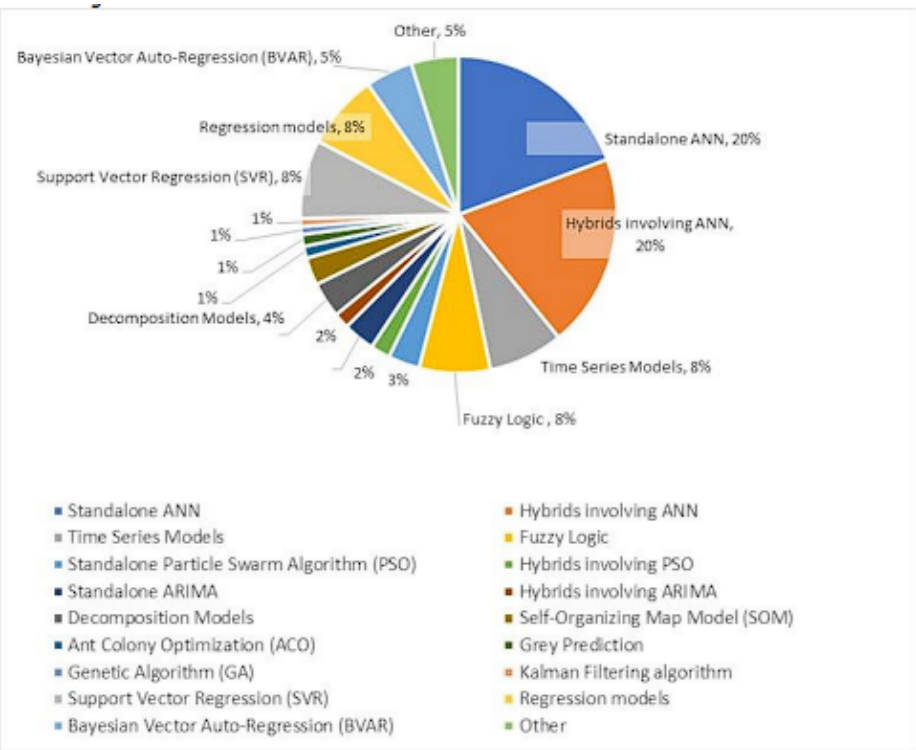

Figura 2.1: Métodos de aprendizaje automático más utilizados según estudio de Nassif y cols. (2022)

En los trabajos analizados se utilizan desde técnicas no basadas en AA, hasta Redes neuronales artificiales (ANN, por sus siglas en inglés), siendo la técnica más utilizada junto a otros métodos híbridos usando también ANN. En cuanto a datos, un 61 % de los trabajos utilizaron conjuntos de datos privados para sus modelos, mientras que un 36 % conjuntos de datos públicos. Dentro de estos datos, los factores más comunes que se utilizaron para las predicciones STLF son variables climáticas (humedad, temperatura, etc.), crecimiento poblacional, y el tipo de día (de semana, fin de semana). Y por último las medidas de rendimiento que más se utilizaron fueron Error absoluto medio porcentual (MAPE), Error cuadrático medio (MSE), Raíz del error cuadrático medio (RMSE) y Error absoluto medio (MAE).

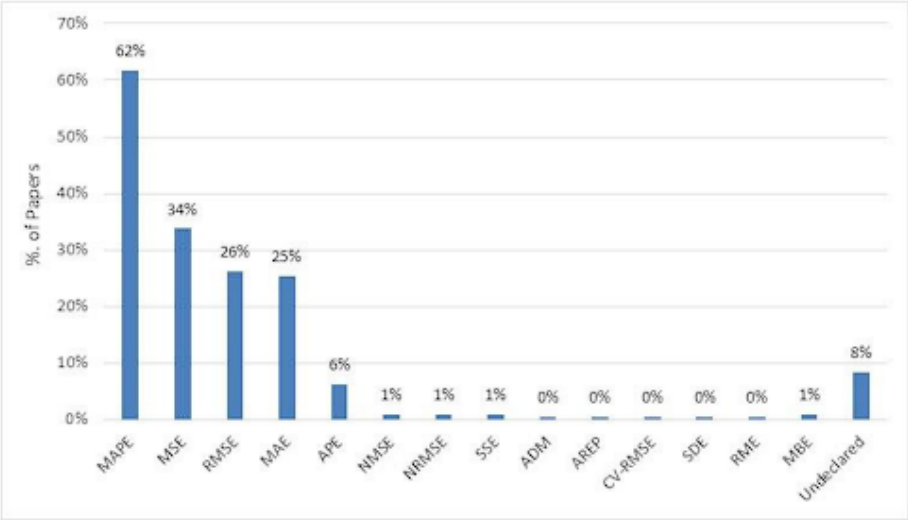

Figura 2.2: Porcentaje de uso de criterios de evaluación de rendimiento en estudio Nassif y cols. (2022)

El estudio de Lara-Benítez y cols. (2020) coincide en que existe una cantidad significativa de trabajos en la literatura reciente que aplican redes neuronales a resolver problemas de predicción de consumo eléctrico. En este caso se utilizan las redes LSTM (long short-term memory), identificadas como las más populares, y Redes Neuronales Convolucionales (CNN, por sus siglas en inglés). Como datos para la predicción se utilizó el conjunto de datos más simple que cualquier otro estudio evaluado, simplemente el tiempo de la medida y la demanda eléctrica en el período. Estos datos los obtuvieron en intervalos de 10 minutos, en un rango desde 2014 a 2019, brindados por la Red Eléctrica Española. Esto resulta en un total de 306.721 entradas en su tabla de datos. Utilizando el WAPE para medir el rendimiento de estas redes neuronales obtuvieron un 0,93 % para la mejor variante de CNN y 1,51 % en el caso de LSTM, sobre las predicciones con ventana de tiempo de 4 horas en períodos de 2018 a 2019.

Por otra parte, se presentan estudios más reducidos, como el de Wang y cols. (2018) en el cual dedican su investigación únicamente al consumo de 2 edificios específicos. Esto les permitió observar una respetable cantidad de datos: datos meteorológicos de una estación a 500 metros, datos de la ocupación de los edificios en intervalos de 1 hora entre ellos. Se obtuvo un total de 8760 entradas de datos para cada edificio. El método de evaluación de rendimiento del algoritmo fue uno personalizado, el cual tiene en cuenta: R^2, RMSE y MAPE, y se utilizó para comparar el algoritmo contra Árbol de regresión y SVR. Se consiguió un MAPE de 7.75 %, R^2 de 0.73, y RMSE de 5.12 para Random Forest, siendo el mejor de los evaluados con todas las métricas. En este trabajo no presentan el valor de las predicciones por lo que el valor de RMSE queda un poco fuera de

contexto.

Un factor interesante del trabajo es que al haber utilizado Random Forest para la predicción obtuvieron la importancia relativa de cada atributo para la predicción. Lamentablemente no hacen pública la información específica de la importancia de cada atributo, pero esto ilustra la utilidad de este método de predicción para conocer información sobre las causas de consumo.

Otro caso de uso de Random Forest se da en el trabajo realizado por Porse y cols. (2016). Este estudio se ubica en un pueblo rural llamado Bruce County, en la región de Ontario, Canadá, trabajando a nivel de distrito, donde se cuenta con datos por hora del consumo eléctrico del pueblo, así como también la temperatura, cuya influencia en el comportamiento de los habitantes es destacada por los autores, y la velocidad del viento en el intervalo entre 2010 y 2019 lo cual se traduce en un conjunto de datos considerablemente grande. Los investigadores entrenan Random forest, Support Vector Machines, Long short term memory y Nonlinear autoregressive exogenous para predecir el consumo eléctrico por hora.

2.2.2. MTLF

Para las predicciones a mediano plazo, hay mucha diversidad en las características de los datos e implementaciones, ya que comparten algunas características con predicciones a corto plazo (como por ejemplo el aspecto cíclico sobre el consumo) y algunas con predicciones a largo plazo (mayor variabilidad de los resultados o estar generalmente enfocados a grandes poblaciones)

La mayoría de los estudios realizan sus predicciones en intervalos de un mes. Esto se debe principalmente a que las fuentes de información que utilizan registran sus datos en este intervalo de tiempo.

Partiendo de los trabajos que más abarcan geográficamente, se encuentran algunos como los de Oreshkin y cols. (2021) y Chen y cols. (2017), que buscan predecir el consumo mensual de 35 países europeos. Ambos trabajan con el mismo conjunto de datos, utilizando el consumo eléctrico mensual partiendo del año 1991 hasta el 2014.

En el trabajo de Dudek y Pełka (2021) la intención es mostrar al lector la predicción de series temporales, en este caso de consumo eléctrico, a través de métodos PSFM (pattern similarity-based forecasting model) comparándolo contra otros tipos de modelos de aprendizaje automático más populares, marcando las diferencias en cuanto a implementación. Finalmente, el trabajo muestra los resultados de las predicciones de cada mes para el año 2014, mostrando el uso de cada método, e incluso sumando híbridos de estos. Los resultados obtenidos son muy alentadores para este tipo de métodos, ya que comparados, por ejemplo,

con LSTM, el error obtenido en las predicciones es mucho menor.

Por otro lado Oreshkin y cols. (2021) se enfocan en la utilización de N-Beats, que se trata de un tipo particular de red neuronal recurrente (RNN, por sus siglas en inglés) y que ha demostrado buen rendimiento en conjuntos de datos de grandes tamaños. En este caso no se cuenta con un conjunto de datos grande. Para implementar N-Beats los autores utilizan el MAPE para entrenarla y se comparan los resultados junto con otros 10 métodos entrenados con el mismo conjunto de datos, prediciendo los 12 meses del 2014. El modelo N-BEATS muestra una mejora sustancial en el MAPE con respecto a otros métodos con un error de 3,8 % mientras que el resto de los métodos comparados presentan un MAPE de hasta 6.18 %.

Pasando a espacios geográficos más reducidos se analizó el estudio realizado por Chen y cols. (2017), que se ubica en la región de Hanoi, Vietnam, región que sufre cambios meteorológicos sustanciales a lo largo de las diferentes estaciones del año. Este estudio busca aplicar dos algoritmos heurísticos llamados *Gravitational Search Algorithm*(GSA) y *Cuckoo Optimization Algorithm* (COA) en el entrenamiento de redes neuronales con el fin de obtener predicciones más precisas que los métodos más tradicionales. Para este estudio se utilizaron datos mensuales que van desde el 2003 hasta el 2013. Donde se encuentran: consumo eléctrico, temperatura promedio, velocidad del viento promedio, tiempo de lluvias, humedad relativa promedio y tiempo de sol en el día. Los resultados del trabajo son favorables para el algoritmo COA, siendo el que mejor se desempeñó.

Finalmente, el último estudio de este tipo que se presenta es el de Gul y cols. (2021), que se ubica en el distrito de Guro, en la ciudad de Seúl, capital de Corea del Sur. Se predice a partir de los datos de consumo eléctrico mensual de distintas áreas de consumo como residencial, industrial, comercial, etc. Como en los casos anteriores se entrenaron varios modelos, ARIMA y varias versiones de redes neuronales LSTM y CNN. Los resultados obtenidos muestran que ARIMA fue el método que mejores resultados arrojó, aunque a su vez es el que mayor preprocesamiento de datos requiere.

2.2.3. LTLF

Si bien la categoría abarca predicciones en intervalos de cualquier tamaño superior a un año, todos los trabajos relacionados relevados utilizan intervalos de exactamente un año.

Investigadores han logrado predicciones de consumo energético para un país entero hasta el año 2040 (Rivera-González y cols., 2019) y también para regiones de países similares hasta el año 2030 (Kaboli y cols., 2017)

Existen investigaciones sobre el consumo en ciudades (Ma y Cheng, 2016),

en países (Mohamed y Bodger, 2005; Kaboli y cols., 2017; Rivera-González y cols., 2019) y por edificios (Yu y cols., 2010). A su vez, hay investigaciones que se hacen con el objetivo de encontrar relaciones entre las variables y el consumo eléctrico (Ma y Cheng, 2016; Yu y cols., 2010).

Como ya fue mencionado anteriormente, los principales objetivos que se buscan, realizando este tipo de predicciones, es poder generar un plan de acción futuro para los posibles cambios en la demanda energética de una región, ya sea un edificio o un país entero. La precisión impacta directamente en el mercado de energía, costos de mantenimiento, futuras inversiones y planificaciones.

Según lo que establece Khuntia y cols. (2018), lograr predicciones precisas a largo plazo puede llegar a ser bastante más complejo que para otros plazos, por la naturaleza del crecimiento en la demanda y de varios parámetros que influyen, siendo muchos de estos impredecibles e incontrolables. Un ejemplo son las variables climáticas. En predicciones de corto plazo se conoce con mucha exactitud el pronóstico para el día siguiente, pero intentar determinar estas variables con años de antelación es una tarea muy complicada.

En cuanto a los datos que se suelen utilizar en los estudios más a largo plazo, las variables que se toman son mucho más estables y no se suelen tomar datos específicos de los hogares, no porque no sean de utilidad, sino porque en general, estos estudios a largo plazo involucran regiones muy grandes, y obtener datos específicos de cada hogar en estas regiones puede ser una tarea muy complicada. Según Shao y cols. (2017) los factores que suelen ser más importantes para este tipo de predicción son el desarrollo económico y la política macroeconómica. Khuntia y cols. (2018) también menciona que, en predicciones para plazos tan largos, otros tipos de parámetros que pueden ser de utilidad son la ocurrencia de eventos especiales en la región de estudio, como también requerimientos regulatorios (por ejemplo, leyes que regulen el consumo).

Los datos que se utilizan no suelen presentar naturaleza cíclica, sino que presentan una fuerte tendencia al crecimiento de un año al siguiente (Shao y cols., 2017; Mohamed y Bodger, 2005). Es por esto que la Regresión aquí es más común que en las predicciones STLF y MTLF. Según Shao y cols. (2017), los métodos más comunes son Redes neuronales, SVM y Regresión.

Se realizó un estudio (Kaboli y cols., 2017) en la región de ASEAN-5 (refiere a Indonesia, Malasia, Filipinas, Singapur, y Tailandia), en la cual se realizan predicciones de consumo eléctrico en los países de esta región hasta el año 2030 inclusive. Aquí comparan modelos basados en Regresión y modelos basados en Redes Neuronales, entre otros. Este trabajo muestra los efectos de dos tipos de datos diferentes en consumo eléctrico para los países, indicadores socioeconómicos de los países que integran la región, como por ejemplo GDP, población, importación y exportación de bienes y servicios, entre otros. Por otro lado, también se toman los datos históricos del consumo eléctrico para los 3 años

anteriores (esto es para poder detectar tendencias de crecimiento en los últimos años). En este trabajo se hace énfasis en *Gene Expression Programming* (GEP) ya que es con el que mejores resultados se obtuvieron. Los modelos fueron validados con datos de los años 1971 a 2011, y utilizaron MAPE, RMSE, U-static, R2 y RCAF como indicadores de rendimiento. Dentro de la gran cantidad de métodos que probaron, los valores de MAPE rondan entre 2,8 % y 10,0 %. Aquí se predice que consumo eléctrico de los países de la región irá incrementando un 3.01 % por año, desde 2011 hasta el 2030 inclusive.

Un trabajo similar llevado a cabo por Rivera-González y cols. (2019), realiza predicciones del consumo y de la generación eléctrica en todo el país de Ecuador, desde el año 2018 hasta el 2040 inclusive, utilizando la metodología LEAP (*Long-range Energy Alternatives Planning*). Los datos que se tomaron para llevar a cabo este estudio fueron datos demográficos y macroeconómicos del país (GDP y crecimiento poblacional), así también como datos históricos y estadísticos sobre el sistema de generación de energía de Ecuador. Junto con estos datos se plantean distintos escenarios futuros que afectan el crecimiento del consumo eléctrico a futuro en el país, como la aplicación de propuestas del ámbito gubernamental sobre políticas energéticas.

El trabajo de Ma y Cheng (2016) es el que, dentro de la revisión hecha, incorpora un mayor número de variables en su predicción. Utilizan un total de 171 variables, que se presentan en la figura 2.3, con nivel de detalle por edificio de 3640 edificios distintos. Estos datos se obtuvieron de la ciudad de Nueva York, en donde se pueden obtener de forma pública y a su vez incorporaron datos del censo nacional y de una base de datos sobre vegetación llamada PLUTO. Este estudio tiene como objetivo establecer los factores más influyentes en el uso de energía de edificios residenciales, y para hacerlo utilizan como método de predicción *Random Forest* (RF), que luego permite generar una clasificación de las variables según su impacto en el algoritmo, con una estimación *out-of-bag*. A su vez se evaluó este algoritmo a la par que MLR, Lasso y SVM con las métricas de MSE y RMSE. RF logró los mejores resultados con 0.773 de MSE y 0.879 de RMSE. Para poner en contexto estas métricas, los valores que predicen están en el rango de 0 a 400 kwh/m^2, excepto alguna anomalía.

Un trabajo mucho anterior a los demás citados (Yu y cols., 2010) consistió en desarrollar un predictor de consumo eléctrico para 80 edificios en Japón. Esto lo hicieron con árboles de decisión. La principal desventaja de su predictor es que el mismo sólo predice si la energía consumida será “alta” o “baja”, considerando un límite arbitrario para decidir entre estos dos valores. Sin embargo, es una demostración de un algoritmo capaz de generar una estructura de árbol que representa cómo los factores influencian la predicción del algoritmo, y de ahí su utilidad. Los datos utilizados para este modelo fueron: Temperatura anual promedio, Apartamento (Sí/No), Madera (Sí/No), tamaño del hogar, coeficiente de pérdida de calor, cantidad de integrantes y métodos de calefacción y de cocina.

Top 20 features that affect the estimation of the average regional EUI of residential buildings.

Rank	Feature	1st level aspect	2nd level category	RF importance	Correlation coefficient
1	Percentage of people with bachelor degree or higher among the population over 25	Education	Education	30.71	−0.29
2	Percentage of households heated by fuel oil	Building	Heating source	28.84	+0.28
3	Median household income	Economy	Family/household income	26.83	−0.31
4	Percentage of people with high school degree or lower among the population over 25	Education	Education	23.84	+0.28
5	Number of residential complaints per capita	Building	Residential complaints	19.40	+0.23
6	Percentage of households heated by natural gas	Building	Heating source	18.51	−0.25
7	Median house value	Economy	House value	17.99	−0.27
8	Density of population over 25 that has high school degree or lower	Education	Education	17.94	+0.32
9	Median gross rent	Economy	Gross rent	17.68	−0.28
10	Average assessed land value	Economy	Land value	16.76	−0.14
11	Percentage of families whose income is below the poverty level	Economy	Poverty	16.16	+0.29
12	Average household size	Population and household	Household size	14.78	+0.13
13	Average percentage of the monthly costs over the household income	Economy	Owner costs	14.50	+0.15
14	Density of public schools	Surrounding	School	14.29	+0.25
15	Median family income	Economy	Family/household income	13.62	−0.29
16	Unemployment rate	Economy	Unemployment	13.12	+0.21
17	Percentage of households with more than 5 persons	Population and household	Household size	13.08	+0.12
18	Density of private school	Surrounding	School	12.12	−0.13
19	Density of people whose income is below poverty level	Economy	Poverty	12.02	+0.28
20	Percentage of households with 1 person	Population and household	Household size	11.40	−0.10

Figura 2.3: Fuente: Lista de 20 variables que más afectaron al consumo eléctrico según modelo de Ma y Cheng (2016)

2.3. Conclusiones

Gracias al estudio de los trabajos relevados se pudo conocer más sobre el panorama actual en cuanto a predicciones de consumo eléctrico.

Las características más distintivas en estos estudios surgen por el intervalo en el que se busca realizar las predicciones. Este intervalo tiene un gran impacto en los datos que se pueden utilizar o qué se suele tener disponible para la predicción, como también en su importancia.

Los datos que son más relevantes para las predicciones a corto plazo también parecen ser útiles para realizar predicciones a plazos mayores, ya que se puede pensar como la combinación de muchas predicciones de plazos más cortos. Sin embargo, en la práctica, los datos que se utilizan suelen variar bastante dependiendo del intervalo de predicción. El uso de variables volátiles parece ser más adecuado para predicciones a corto plazo, y por otro lado, las variables que son más estables y consistentes a lo largo del tiempo se muestran más apropiadas en plazos más largos.

Un claro ejemplo es el uso de variables meteorológicas. Los resultados obtenidos en los distintos estudio sugieren que son factores determinantes en predicciones de corto y mediano plazo, pero no tanto en plazos más largos dada su naturaleza cíclica, ya que a lo largo del año se recorren todas las estaciones.

Por otro lado, diversas investigaciones han mostrado que los factores económicos son esenciales para obtener resultados precisos, sin importar el plazo de las predicciones realizadas.

Los trabajos se centran en zonas de las que se pueden extraer una gran cantidad de datos, ya sea porque estos estaban recopilados con anterioridad o en otros casos (generalmente en STLF) se consiguen datos exclusivamente para la investigación. Contar con un gran volumen de datos parece ser importante para la precisión de los resultados, ya que los modelos logran mejores predicciones al estar entrenados con más información.

Los modelos que se utilizan son muy dependientes del tipo de problema que se quiere resolver. Los problemas a corto y medio plazo tienen características cíclicas que se deben capturar en el algoritmo de predicción. También se puede notar que en la mayoría de los trabajos se utilizan varias metodologías para la misma tarea, comparando los resultados que se obtienen entre los distintos modelos. Otro aspecto para resaltar en cuanto a los modelos es que, si es de interés para el estudio conocer el impacto de las variables en la predicción, se suelen utilizar métodos como *Random Forest* o Árboles de decisión.

El estudio también ha permitido conocer cómo los investigadores evalúan los resultados generados por sus modelos. Se suelen mantener las mismas medidas

de rendimiento entre los distintos trabajos y por lo general, los investigadores se encuentran satisfechos con la precisión de las predicciones, lo que resulta alentador para trabajos futuros.

En este punto consideramos que se ha relevado suficiente conocimiento en el área de la predicción del consumo eléctrico como para comenzar a aplicarlo en el proyecto, comenzando con la obtención de datos en Uruguay, que es de lo que trata el siguiente capítulo.

Capítulo 3

Datos para la predicción del consumo eléctrico

Hasta el momento se ha presentado el proyecto junto con los objetivos que lo motivan. Se han analizado varios estudios realizados en el área de la predicción del consumo eléctrico con el fin de adquirir conocimientos para tomar decisiones en el presente proyecto. Este análisis ha permitido entender cuáles son los problemas más comunes, así como los datos y metodologías más utilizados para resolverlos.

El primer y más importante paso de nuestro trabajo es la obtención de datos, que es de lo que trata este capítulo. Se debe determinar qué tipos de datos pueden estar relacionados con el consumo eléctrico, así como las distintas fuentes que se tienen a disposición en Uruguay. La disponibilidad de los datos puede llegar a ser un factor limitante para el plazo en el que se realizan las predicciones. Por ejemplo, si solo se cuenta con información anual sobre el comportamiento de una variable, puede ser difícil realizar predicciones precisas a nivel diario o semanal.

Existe una amplia variedad de datos que pueden ser de utilidad, pero por lo general destacan los datos meteorológicos y socioeconómicos, sumado a datos históricos del consumo eléctrico. Se busca información relevante que esté disponible en Uruguay para estas tres categorías teniendo en cuenta que las fuentes que se utilicen cuenten con muchos registros, ya que cuanta más información se les brinde a los modelos de aprendizaje, más precisas son las predicciones.

En cuanto a datos meteorológicos de Uruguay, la fuente más relevante, confiable y extensa proviene del Instituto Uruguayo de Meteorología [1], que tiene como finalidad prestar los servicios públicos meteorológicos y climatológicos, consistentes en observar, registrar y predecir el tiempo y el clima en el territorio

[1] Sitio web INUMET: https://www.inumet.gub.uy

nacional.

Por otro lado, una gran fuente de datos socioeconómicos proviene de la Encuesta Continua de Hogares [2]. Esta encuesta es realizada anualmente por el Instituto Nacional de Estadística [3] y consta en tomar un muestreo de hogares de distintas regiones del país y realizarles a las personas que allí viven una serie de preguntas, como por ejemplo la cantidad de personas que viven en el hogar, los ingresos que perciben, su nivel de educación, etc.

Por último, en cuanto a datos de consumo eléctrico en Uruguay solo existe una fuente que proviene de la Administración Nacional de Usinas y Trasmisiones Eléctricas [4], el único proveedor eléctrico del país. UTE no solo cuenta con datos históricos de consumo eléctrico de los hogares, sino que además cuenta con otros atributos que son de utilidad, como por ejemplo las tensiones, potencias contratadas, tarifas, etc.

Habiendo fuentes de información disponibles para cada una de las categorías que se buscaba cubrir, se puede proceder con su análisis. Es importante tener en cuenta que algunos de estos conjuntos de datos son de acceso público y otros de acceso privado. Específicamente los datos de UTE tuvieron que ser solicitados y deben pasar por un proceso de sanitización, que consta en analizar y tratar los datos de forma de no exponer información personal, como por ejemplo la dirección de los hogares.

El resto del capítulo se divide en cuatro secciones: una por cada una de las fuentes de datos con las que se trabaja, partiendo con los datos tales y como fueron entregados, detectando problemas con sus correspondientes soluciones y una sección final, en la que se detalla el proceso de integración de todas las fuentes en una única.

Se evalúa la calidad de los datos con las siguientes medidas:

- Corrección: que tan precisos, válidos y libres de errores están los datos.
- Completitud: indica si se contiene toda la información de interés, es decir, si se representan todos los objetos de la realidad y si se tienen todos los datos que describen a los objetos.
- Frescura: qué tan actualizados están los datos con respecto a la realidad que se busca modelar y el período en el que son válidos. Con respecto a los datos para este trabajo se busca que el rango de años que abarcan sea lo más extenso y reciente posible. Si los datos que se utilizan son muy

[2] Portal ECH https://www.gub.uy/instituto-nacional-estadistica/politicas-y-gestion/microdatos-metadatos-cuestionarios-manuales-ech-edicion

[3] Portal del INE: https://www.gub.uy/instituto-nacional-estadistica/

[4] Sitio web UTE: https://www.ute.com.uy/

antiguos no serán de mucha utilidad para realizar predicciones de consumo eléctrico en la actualidad. En este caso nos limita la ECH, por razones que se detallan en la siguiente sección.

- Consistencia: captura la satisfacción de reglas semánticas sobre los datos, por ejemplo, si se cumplen las reglas de dominio y si no hay contradicciones.
- Unicidad: indica el nivel de duplicación entre los datos, que ocurre cuando una misma entidad está representada dos o más veces.

3.1. Encuesta Continua de Hogares (ECH)

Se comienza el análisis con los de datos de la ECH. El Instituto Nacional de Estadística (INE) es el encargado de llevar a cabo las encuestas anuales. Se toma un muestreo al azar de hogares en distintas regiones del país para hacerles un cuestionario, registrando las respuestas. Esta información luego es procesada y queda disponible públicamente.

Los resultados de las encuestas tienen aproximadamente 160 variables, pero dependiendo del año puede haber diferencias. A los modelos de aprendizaje se les deberá alimentar con los mismos parámetros de entrada en cada instancia, por lo que para poder entrenarlos con información de distintos años es necesario que se mantengan las mismas variables. Por esta razón es que se toman las encuestas entre 2013 y 2019 inclusive. Para las encuestas de años posteriores y anteriores existen grandes diferencias en las variables que se utilizan. Esta restricción también aplica para las otras fuentes de datos ya que como se mencionó previamente, en determinado punto se unifican y no se pueden mezclar periodos distintos.

No todas las variables de la encuesta son útiles para el estudio. Interesan aquellas que puedan tener relación con el consumo eléctrico del hogar, como por ejemplo variables relacionadas a los materiales con los que se construyó el hogar, ingresos, electrodomésticos, cantidad de personas que habitan el hogar, etc. Se toman 38 variables de las 160 totales. El listado de variables seleccionadas se encuentra disponible en el anexo “Variables utilizadas de la ECH”, mientras que la lista completa está definida en el diccionario de variables del INE.

En las figuras 3.1 y 3.2 se puede observar el comportamiento de algunas de las variables que fueron seleccionadas en relación con al consumo eléctrico diario por hogar en Montevideo.

Como los resultados de las encuestas son públicos, no pueden incluir información personal como la dirección exacta del hogar o los nombres de las personas que allí viven. En la mayor parte del país los resultados de la encuesta indican únicamente su departamento, a excepción de Montevideo. Al ser el departamento más poblado se busca ser más precisos, por lo que se hace uso de una división

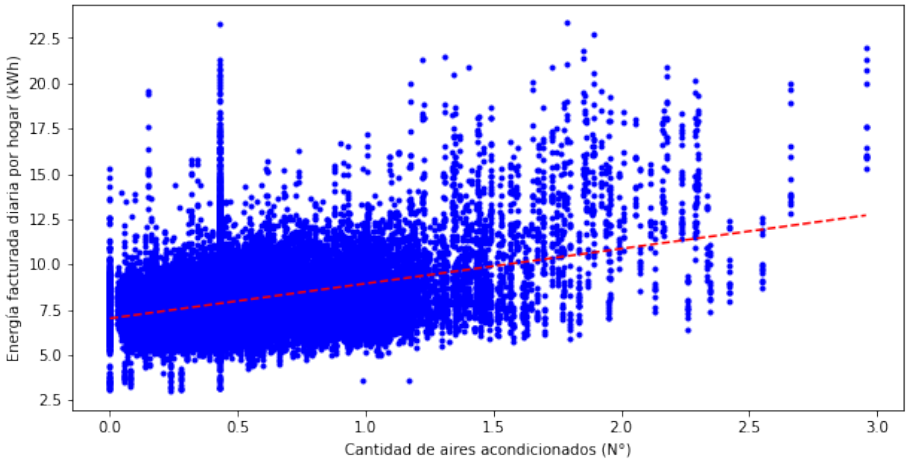

Figura 3.1: Energía facturada diaria por hogar, en relación al promedio de la cantidad de aires acondicionados en un segmento. Esta figura incluye datos de todos los meses del año, donde los valores de la ECH se repiten, causando que en el mismo valor de la variable tenga distintos consumos para un mismo segmento.

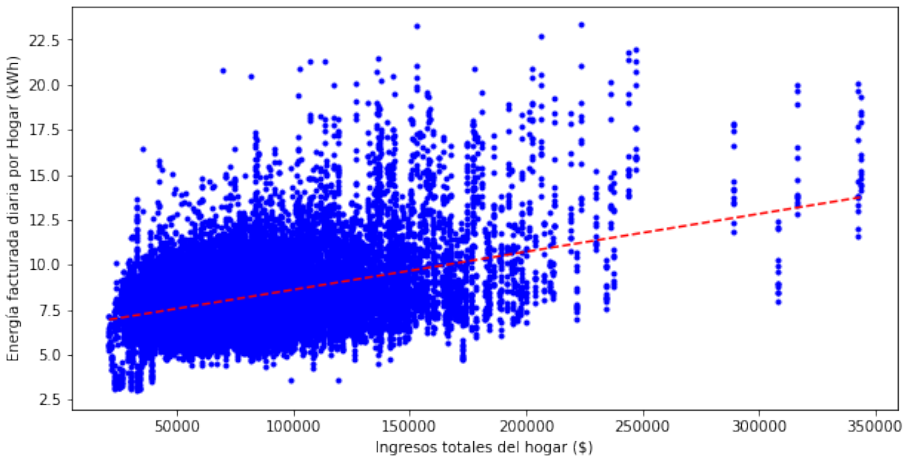

Figura 3.2: Energía facturada diaria por hogar, en relación al promedio de ingresos totales del hogar en un segmento. Esta figura incluye datos de todos los meses del año, donde los valores de la ECH se repiten, causando que en el mismo valor de la variable tenga distintos consumos para un mismo segmento.

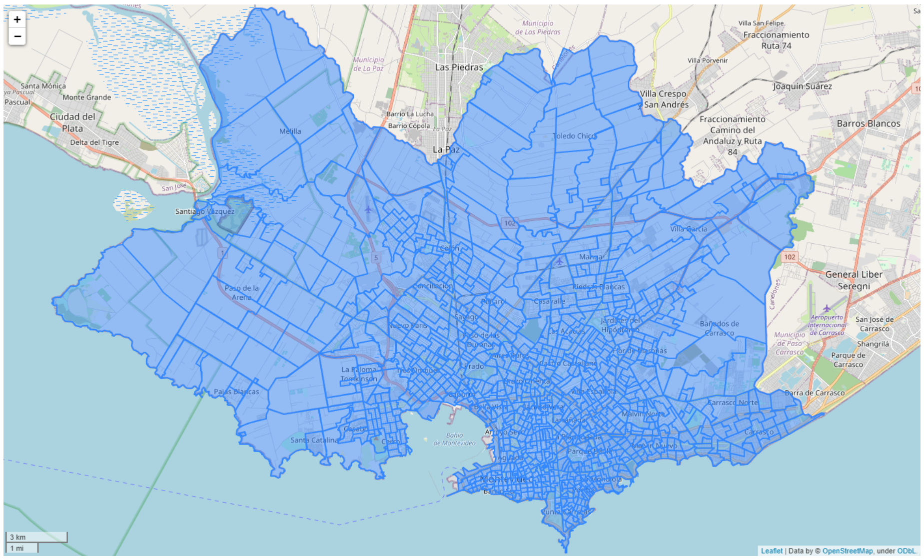

Figura 3.3: Montevideo dividido en segmentos, según el INE

en regiones más pequeñas del departamento denominadas *secciones* y *segmentos* para ubicar cada encuesta, como se puede observar en la figura 3.3.

Una sección es de un tamaño considerablemente menor que el departamento y a su vez, cada segmento divide a una sección en regiones aún más chicas (de al menos unas cuadras). Montevideo se fracciona en 1063 segmentos y la cantidad de hogares dentro de cada uno es muy variable. Existen segmentos con 10 hogares o menos y otros con más de 1600. Cada segmento se identifica a partir de una sección, su unión cubre la totalidad de Montevideo y son disjuntos entre sí. Esta división en regiones más pequeñas del departamento es el principal motivo por el cual se decide trabajar únicamente en Montevideo, ya que permite ser más precisos con las predicciones. Se entrará más en detalle sobre esta decisión en el capítulo 4.

La idea es entonces, tomar todas las encuestas que se hicieron en un segmento y promediar sus resultados. De esta forma se tendría un único registro por segmento para cada año y debido a que se cuenta con 7 años de información se obtendrían un total de $1063 \times 7 = 7441$ registros. Pero hay un inconveniente; no todos los segmentos fueron encuestados todos los años, es decir que existen segmentos que por lo menos para un año (entre 2013 y 2019) no se hizo ninguna encuesta. No se cree que se trate de un problema de completitud, ya que probablemente no se busque cubrir todos los segmentos de Montevideo con la muestra de hogares a encuestar. Esto es un problema, ya que para promediar los resultados por segmento debe existir por lo menos una encuesta. Es por esta

razón que se decide trabajar únicamente con el subconjunto de segmentos que hayan sido encuestados todos los años, que se pueden visualizar en la figura 3.4.

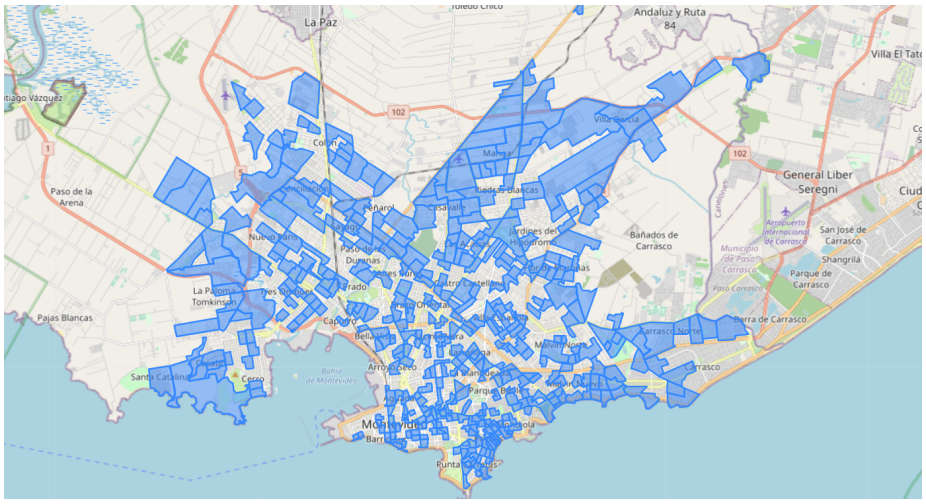

Figura 3.4: Segmentos encuestados todos los años en Montevideo (desde 2013 hasta 2019)

Existen 431 segmentos de los 1063 que cumplen esta condición, con los que se trabaja de ahora en adelante. La cantidad total de registros de esta fuente de datos es de $431 \times 7 = 3017$.

3.2. Administración Nacional de Usinas y Trasmisiones Eléctricas (UTE)

Se continúa el análisis con los datos de UTE, la empresa estatal encargada de la generación, transmisión y distribución de energía eléctrica en el país. UTE suministra a la gran mayoría de los hogares, manteniendo registros mensuales de su consumo eléctrico, planes tarifarios, tensiones, etc. El Ministerio de Industria, Energía y Minería (MIEM) compartió estos datos para el departamento de Montevideo en el rango de años entre 2013 y 2019 inclusive para la realización del proyecto.

Estos datos tienen la particularidad de ser de acceso privado al contener información personal de los hogares de Uruguay (dirección, coordenadas geográficas, teléfonos, tarifas contratadas, etc.). Esto implica un esfuerzo adicional para transformar los datos y no exponer esta información. Por otro lado, al igual

que la ECH, esta fuente cuenta con muchas variables y no todas son de interés. Con la ayuda del personal del MIEM se hizo una selección de las variables más importantes, que se lista a continuación:

- *id_predio*: Identificador de un predio (en nuestro caso, hogares). Un predio puede tener más de un acuerdo de servicio.
- *año y mes*: Indican la fecha en la que se toma la medida del consumo.
- *latitud y longitud*: Coordenadas geográficas que ubican al predio.
- *tensión*: Voltaje contratado en el acuerdo de servicio, este puede ser de baja, media o alta tensión (BT, MT y AT respectivamente).
- *tarifa*: Tipo de tarifa contratada en el acuerdo de servicio. Estas pueden ser: Tarifa Residencial Simple, Doble Horario, Triple Horario, Tarifa de Consumo Básico Residencial, entre otras.
- *energía_facturada*: Nuestra variable objetivo. Medida de consumo eléctrico del predio en determinado mes.

El costo de la electricidad es un factor muy importante con respecto al consumo eléctrico. La UTE ofrece distintas tarifas que los hogares pueden contratar dependiendo de sus necesidades. Estas pueden tener precios especiales en ciertos horarios, o requerir un consumo limitado. Las tres tarifas más comunes en Montevideo entre los años 2013 y 2019 son:

- Tarifa residencial simple (TRS): Se mantiene el mismo costo de la electricidad a lo largo del día. Sin restricciones. Si se realizan consumos elevados, el precio de los kWh aumenta en diferentes niveles.
- Tarifa residencial doble horario (TRD): El costo de la electricidad es más caro que TRS en horas pico, pero más barato (cerca de la mitad) el resto del día.
- Tarifa de consumo básico residencial (TCB): Costo más bajo que TRS si se consume dentro de un límite de kWh. Si se pasa de este límite dos veces, se cambia automáticamente a TRS.

A continuación, se muestran algunos registros para el año 2019, luego de haber filtrado las variables de interés:

año	**mes**	**tensión**	**tarifa**	**energía_facturada (kWh)**
2019	1	BT 230V	TRS	314
2019	4	BT 230V	TRD	370
2019	7	BT 230V	TRS	633
2019	10	BT 230V	TCB	23

Tabla 3.1: Registros de UTE, 2019

Notar que no se incluyen las variables *id_predio*, *latitud* y *longitud* en la tabla 3.1 por cuestiones de privacidad, pero sí se utilizan en el trabajo.

Analizando en profundidad los datos se detecta un inconveniente en cuanto a su corrección, y es que a pesar de que los registros están clasificados como de hogares, algunos de los valores de consumo eléctrico son extremadamente altos al punto que no es posible que se correspondan al consumo de un hogar. Por ejemplo, existen los siguientes registros:

año	**mes**	**tensión**	**tarifa**	**energía_facturada (kWh)**
2019	2	BT 400 V	TRD	1.843,8
2019	4	MT 6,4 KV	MC2	17.810,1
2019	6	MT 22 KV	GC2	298.164,2
2019	9	MT 31,5 KV	GC3	1.158.114,0

Tabla 3.2: Registros de UTE con anomalías, 2019

Muchos de estos registros cuentan con tensiones medias (MT) o superiores cuando estas no pueden ser contratadas por hogares comunes y corrientes. Este tipo de registros se deben eliminar. Antes realizar el filtro, para el año 2019 existen un total de 5.654.094 registros, el promedio de *energia_facturada* por hogar es de 266,9 kWh y la desviación estándar es de 1.148,6 kWh.

Como primer paso, se elimina todo registro con un valor de tensión distinto a BT 230V ya que, al tratarse de hogares, esta es la tensión a la que acceden. A su vez, se elimina todo registro con un consumo eléctrico menor a 5 kWh que es lo suficientemente bajo como para asumir que el hogar no estuvo habitado ese mes. La proporción de registros con una tensión distinta a BT 230V para el año 2019 es del 1,32 % y la de registros con un consumo menor de 5 kWh es de 1,30 %. Una vez que se remueven estos registros atípicos se llega a una nueva media de consumo eléctrico de 261,3 kWh con una desviación estándar de 342,2 kWh. La desviación disminuye considerablemente, y la media no varía en gran cantidad.

Este filtro es válido, pero no cubre todos los casos ya que existen registros que a pesar de tener baja tensión cuentan con un consumo eléctrico muy alto. Se presentan algunos ejemplos en la tabla 3.3.

año	**mes**	**tensión**	**tarifa**	**energía_facturada (kWh)**
2019	2	BT 230 V	THE	100.007,0
2019	4	BT 230 V	TRS	13.417,0
2019	8	BT 230 V	MC1	28.082,0
2019	9	BT 230 V	TRS	59.210,0
2019	10	BT 230 V	TGS	67.822,0

Tabla 3.3: Registros que no corresponden a consumos de hogares, 2019

Filtrar por tarifa no elimina todas las anomalías por lo que se hace uso de la medida *z-score* para determinar qué registros quitar. Esta medida indica la cantidad de desviaciones estándar de diferencia entre el valor de un atributo y su media (en este caso, el consumo eléctrico). Se toma un *z-score* de 4 como margen, por lo que se eliminan los registros con un consumo mayor o igual a 1630,2 kWh (calculando $261,3 + 342,2 * 4$). La cantidad de registros en estas condiciones es el 0,44 % de los restantes.

Esto resulta en un total de 5.472.435 registros para el año 2019, con un consumo eléctrico promedio por hogar de 249,6 kWh y una desviación estándar de 194,4 kWh. Este mismo proceso se aplica para el resto de los años. En la figura 3.5 se puede visualizar la Energía Facturada Total en kWh para Montevideo en cada mes del 2019.

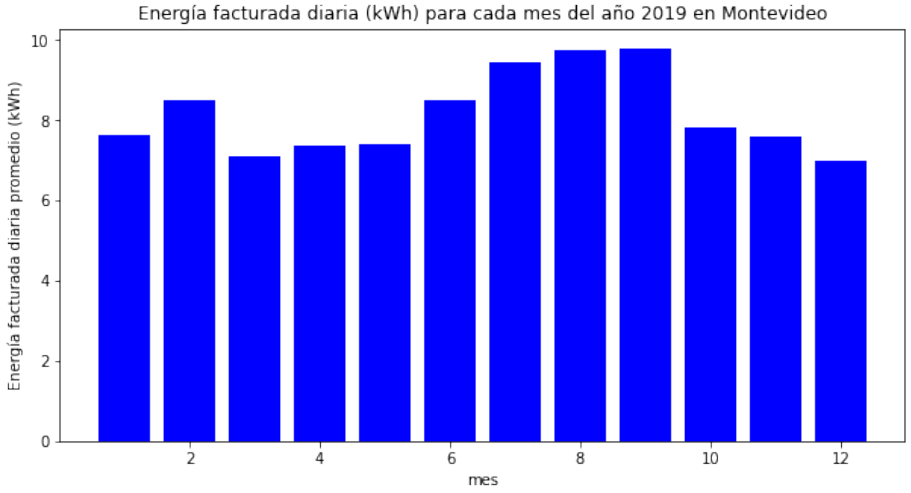

Figura 3.5: Energía facturada total diaria por hogar, para cada mes del año 2019

Se puede observar que en los meses más fríos del año el consumo es mayor, probablemente debido a la calefacción eléctrica que se utilice en los hogares. En los meses calurosos el consumo no es tan alto como en los meses fríos, lo que resulta curioso debido a que la mayoría de las opciones para enfriar implican el uso de energía eléctrica (aires acondicionados, ventiladores), mientras que para la calefacción existen otras opciones muy comunes como el gas o la leña.

Una posible explicación para esto es que los meses más calurosos (diciembre, enero, febrero y marzo) coinciden con los meses de turismo más activo. Muchas familias se toman vacaciones por fuera de Montevideo, por lo que sería razonable que el consumo no fuera tan alto en estas fechas. Para comprobar esta idea se cuentan la cantidad de hogares con energía facturada menor a 50 kWh, un valor

de consumo lo suficientemente bajo (teniendo en cuenta que el promedio por hogar es de 249,6 kWh) como para que en la mayoría de los casos se deba a la ausencia temporal de personas en sus hogares, aunque sea por unos días del mes.

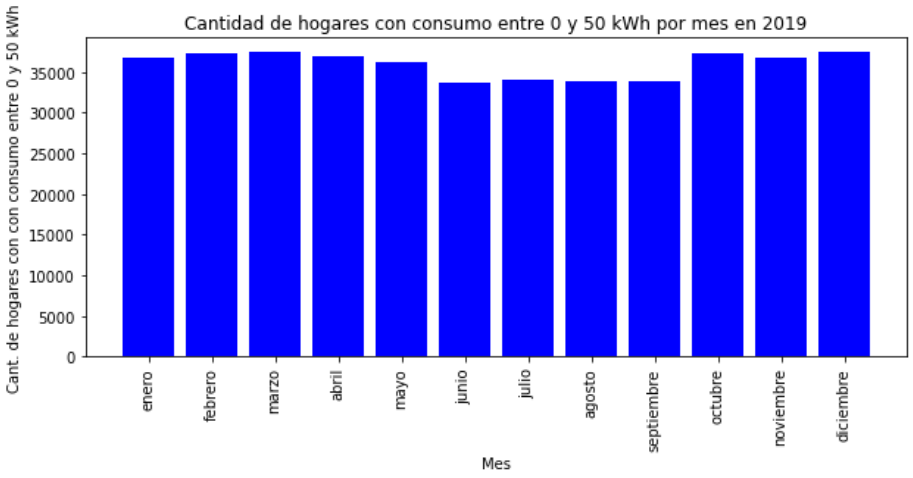

Figura 3.6: Cantidad de hogares con un consumo menor a 50 kWh para cada mes, en 2019

Como se puede observar en la gráfica 3.6, en los meses calurosos hay más hogares con muy poco consumo, lo que respalda nuestra idea. Existen otros factores que también pueden explicar este fenómeno, como los costos de calefacción o enfriamiento, la eficiencia energética de los equipos, entre otros.

3.3. Instituto Uruguayo de Meteorología (INUMET)

Por último, se presenta la fuente de datos de INUMET. Las variables meteorológicas están muy relacionadas al consumo eléctrico, como ya se puede intuir por la figura 3.5. Es por esto que esta fuente de datos es de gran importancia para el trabajo. Nuevamente se toman los datos entre los años 2013 y 2019 inclusive.

En este caso se debieron solicitar específicamente las variables meteorológicas deseadas para obtener los datos. Para decidir qué datos pedir se contó con la ayuda de Madeleine Renom (expresidente de INUMET y grado 4 en meteorología), quien recomendó utilizar únicamente las temperaturas máximas y mínimas registradas día a día en tres estaciones meteorológicas de Montevideo (Prado, Melilla y Carrasco). Se consultó sobre la utilidad de variables como la

humedad y el viento, ya que son factores que se tienen en cuenta para determinar la sensación térmica, sin embargo ella recomendó que no se utilicen. Por esta razón y sumado a que no son variables que suelen aparecer en los trabajos relacionados es que se decide no solicitarlas.

Las diferentes estaciones meteorológicas reportan ciertas diferencias en sus medidas, a pesar de que el departamento de Montevideo sea pequeño. Esto es debido a que las zonas que se encuentran más cerca del mar suelen tener temperaturas distintas, lo suficiente como para que sea de utilidad considerar diferentes estaciones.

Los datos obtenidos de INUMET presentan cuatro variables, que son:

- *fecha*: Fecha en formato "AAAA/MM/DD" del día en el que se realiza la medida.
- *estacion*: Estación meteorológica en la que se realiza la medida (solo puede tomar tres valores: "Prado", "Melilla" o "Carrasco").
- *temperatura_maxima*: Valor de temperatura más alto que se registró en el día.
- *temperatura_minima*: Valor de temperatura más bajo que se registró en el día

Al observar los datos se detecta un problema de completitud. Algunos de los registros tienen valores nulos en las temperaturas (máxima, mínima o ambas). Es de interés contar con todos los registros entre los años 2013 a 2019 sin campos nulos, por lo que se decide rellenarlos con los valores promedio de las temperaturas. A modo de ejemplo, en la gráfica 3.7 se puede observar el promedio de las temperaturas máximas (rojo) y mínimas (azul) todos los años, para cada día del año, en la estación de Prado. Estos son los valores que se utilizan para rellenar los campos faltantes de esta estación, y se procede de manera análoga para las demás estaciones.

La figura 3.8 muestra las temperaturas de cada día entre los años 2013 y 2019. Se muestra tanto la temperatura máxima (rojo) como la mínima (azul):

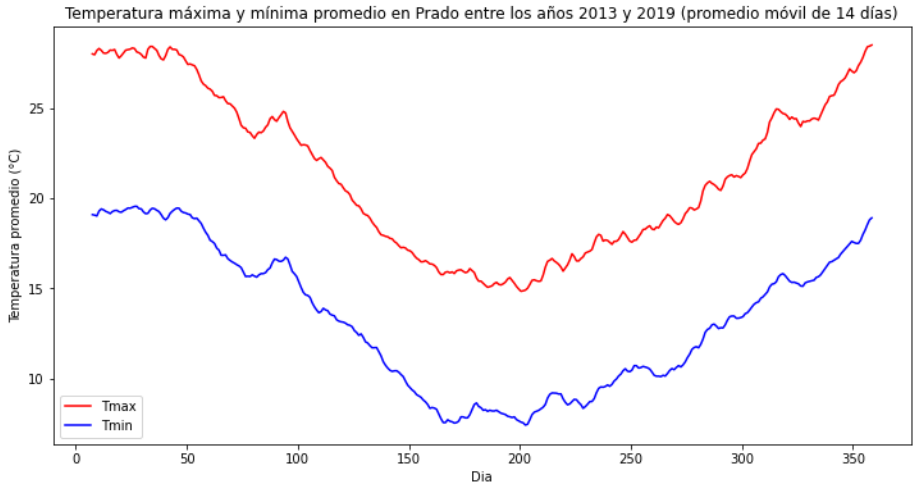

Figura 3.7: Temperaturas máximas y mínimas promedio para cada día del año a lo largo de 2013-2019, en la estación del Prado

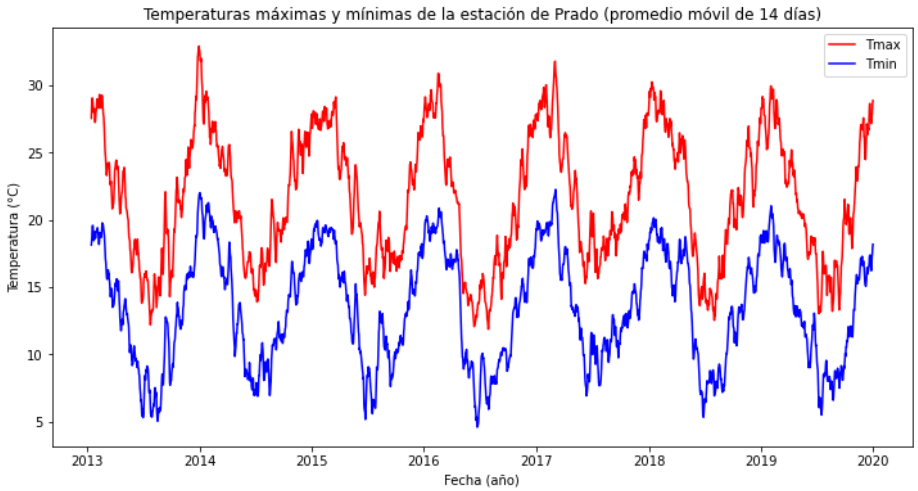

Figura 3.8: Temperaturas máximas y mínimas en la estación del Prado a lo largo de 2013-2019

También se extraen algunos datos estadísticos de todas las medidas que se tienen de la estación de Prado:

Medida	**Valor**
Temperatura máxima promedio	21,6°C
Temperatura mínima promedio	13,5°C
Temperatura media	17,6°C
Máxima temperatura registrada	38,1°C
Mínima temperatura registrada	0,1°C
Desviación estándar de la temperatura máxima	6,0°C
Desviación estándar de la temperatura mínima	5,1°C

Tabla 3.4: Datos estadísticos de las medidas meteorológicas de la estación de Prado, entre los años 2013 y 2019

Anticipándose al proceso de integración (que se ve más en detalle en la siguiente sección del capítulo) se puede notar un inconveniente entre las distintas fuentes, y es que cada una registra sus datos en periodos temporales distintos (anual, mensual y diario). Esta diferencia temporal complica la comparación y el análisis de los datos de manera coherente, por lo que debemos sincronizar estos marcos temporales entre las distintas fuentes para poder integrarlas. Se entra en detalle del motivo en el siguiente capítulo, pero se realizarán predicciones mensuales. Por esto es necesario resumir los valores diarios de manera mensual.

Transformar estos datos a un periodo mensual implica combinar todos los registros un mismo mes obteniendo un único registro por estación meteorológica (es decir, tres en registros por cada mes) con un total de $3 * 12 * 7 = 252$ registros. Se combinan los registros promediando las medidas de temperatura, tanto para la máxima como la mínima registrando también sus máximos, mínimos y desviación estándar.

En la figura 3.9 se puede observar la relación entre el promedio de temperatura máxima con el consumo eléctrico diario por hogar en Montevideo, entre los años 2013 y 2019.

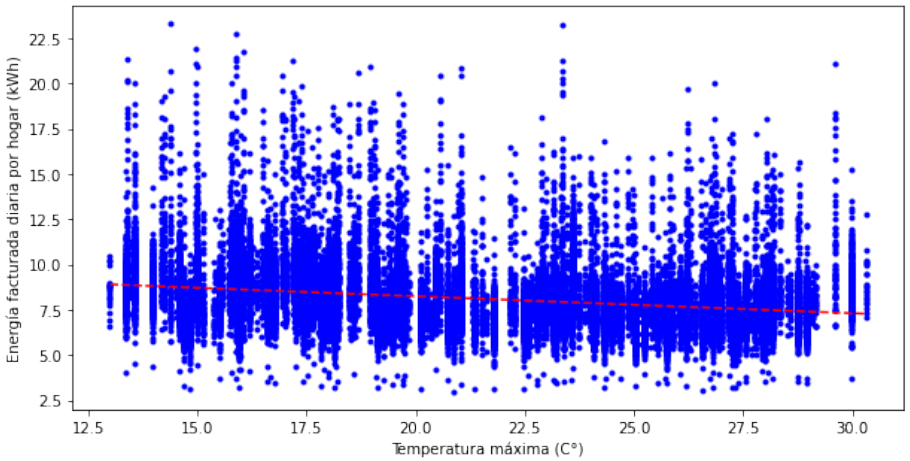

Figura 3.9: Energía facturada diaria por hogar, en relación a la temperatura máxima. Esta figura incluye datos de los diferentes segmentos que tienen distinta energía facturada para un mismo valor de temperatura.

Puede parecer extraño que a mayor temperatura máxima menor consumo eléctrico, sin embargo esto sucede al intentar aproximar el consumo de manera lineal. Ya se ha analizado en la sección de UTE que el consumo eléctrico es mayor en meses fríos, lo que se ve reflejado en esta figura.

Para evitar el problema mencionado anteriormente se introducen medidas de *Degree Days*. Estas son métricas que intentan representar la necesidad de calefacción o refrigeración, y pretenden tener una relación lineal con el consumo eléctrico. Se calculan en base a una temperatura registrada como cómoda, que se toma 15.5°C, por convención de la Unión Europea. El cálculo de HDD *(Heating Degree Days)* representa qué tan fríos fueron los días a lo largo de un período de tiempo y queda definido por la primera condición que se cumpla en la tabla 3.5[5]:

Condición	**Fórmula**
Tmin > Tbase	0
(Tmax + Tmin) / 2 > Tbase	(Tbase - Tmin) / 4
Tmax >= Tbase	(Tbase - Tmin) / 2 - (Tmax - Tbase) / 4
Tmax < Tbase	Tbase - (Tmax + Tmin) / 2

Tabla 3.5: Cálculo de HDD

En donde *Tmin* y *Tmax* corresponden a las temperaturas mínimas y máximas del día, y *Tbase* corresponde al valor de temperatura de comodidad, 15.5°C.

[5]Fórmula para calcularlo recuperada de http://www.vesma.com/ddd/ddcalcs.htm

El cálculo de *Degree Days* para un período (en este caso, un mes) es la suma acumulada de los resultados diarios.

De manera similar, el cálculo de la métrica CDD *(Cooling Degree Days)* queda definida por la siguiente tabla:

Condición	**Fórmula**
Tmax < Tbase	0
(Tmax + Tmin) / 2 < Tbase	(Tmax - Tbase) / 4
Tmin <= Tbase	(Tmax - Tbase) / 2 - (Tbase - Tmin) / 4
Tmin > Tbase	(Tmax + Tmin) / 2 - Tbase

Tabla 3.6: Cálculo de CDD

En conclusión, la lista de variables meteorológicas utilizada es la siguiente:

- *Año*: Año en el que se realizan las medidas.
- *Mes*: Mes en el que se realizan las medidas.
- *Estación*: Estación meteorológica que realiza las medidas (solo puede tomar tres valores: "Prado", "Melilla" o "Carrasco").
- *TMaxPromedio*: Temperatura máxima promedio en el mes.
- *TMinPromedio*: Temperatura mínima promedio en el mes.
- *TMaxRegistrada*: Temperatura máxima registrada en el mes.
- *TMinRegistrada*: Temperatura mínima registrada en el mes.
- *TMaxDesviacion*: Desviación estándar de la temperatura máxima.
- *TMinDesviacion*: Desviación estándar de la temperatura mínima.
- *HDD* (Heating Degree Days): Rigor del calor en el mes.
- *CDD* (Cooling Degree Days): Rigor del frío en el mes.

A continuación, se muestran algunos registros con las nuevas variables para la estación de Prado:

Año	**Mes**	**Estación**	**TMaxProm. (°C)**	**TMinProm. (°C)**	**TMaxReg. (°C)**	**TMinReg. (°C)**
2013	Enero	Prado	27,79	17,37	33,8	10,2
2013	Febrero	Prado	26,21	17,52	34,6	9,9
2014	Julio	Prado	15,78	6,63	25,2	1,0
2014	Agosto	Prado	18,69	7,91	30,0	1,6
2015	Noviembre	Prado	22,87	13,62	29,4	8,0
2015	Diciembre	Prado	26,76	16,72	35,0	9,5

Tabla 3.7: Registros de INUMET, con las nuevas variables

Año	Mes	Estación	TMaxDesv. (°C)	TMinDesv. (°C)	HDD	CDD
2013	Enero	Prado	3,39	2,97	4,18	223,73
2013	Febrero	Prado	3,83	3,22	4,18	182,33
2014	Julio	Prado	2,98	3,29	117,90	15,10
2014	Agosto	Prado	4,92	3,93	112,25	44,05
2015	Noviembre	Prado	2,96	3,05	20,13	102,58
2015	Diciembre	Prado	4,21	3,10	5,95	199,35

Tabla 3.8: Registros de INUMET, con las nuevas variables (continuación)

3.4. Integración de Datos

Se han presentado y analizado las tres fuentes de datos que se utilizan en el proyecto de manera individual, solucionando los problemas que se detectaron durante el procedimiento. En esta sección se detalla el proceso de integración que consta en juntar estas distintas fuentes de datos en un único conjunto de datos, que será el insumo de los modelos de aprendizaje que se definen en el capítulo siguiente.

El proceso de integración cuenta con dos problemas principales a resolver; se debe lograr una uniformidad temporal y espacial entre las distintas fuentes que se van a integrar. La solución en este caso es llevar todas las fuentes de datos a periodicidad mensual y agrupar los registros en las secciones y segmentos definidas por el INE. La discusión de los motivos que llevan a estas decisiones se da en el siguiente capítulo.

La uniformidad temporal se logra fácilmente. Los registros de consumo eléctrico de UTE se registran mensualmente por lo que no requieren cambios. Los datos de INUMET fueron procesados en la sección 3.3 para obtener registros mensuales. Las encuestas de la ECH son anuales, pero se asume que para todo mes dentro de un mismo año se mantienen los mismos valores en los atributos.

Se procede con el análisis de la uniformidad espacial, que es un poco más complejo. Los resultados de las encuestas de la ECH ya están categorizados por sección y segmento. Los registros de UTE están asociados a un predio particular, ubicado en Montevideo por sus coordenadas geográficas (*latitud* y *longitud*). Por último, los datos de INUMET están asociados a una estación meteorológica, la que también podemos ubicar en Montevideo por sus coordenadas geográficas.

Comenzando con INUMET, dado a que se cuentan con tres estaciones meteorológicas distintas se debe establecer una relación entre los distintos segmentos y las estaciones, de las que tomarán sus medidas. Surge una idea muy intuitiva, se le asigna a cada segmento la estación meteorológica más cercana en cuanto a distancia. Se puede observar el resultado de la asignación en la figura 3.10. Las regiones pintadas de amarillo son las más cercanas a la estación de Melilla, las verdes a la estación de Prado y las violetas son las más cercanas a la estación de Carrasco.

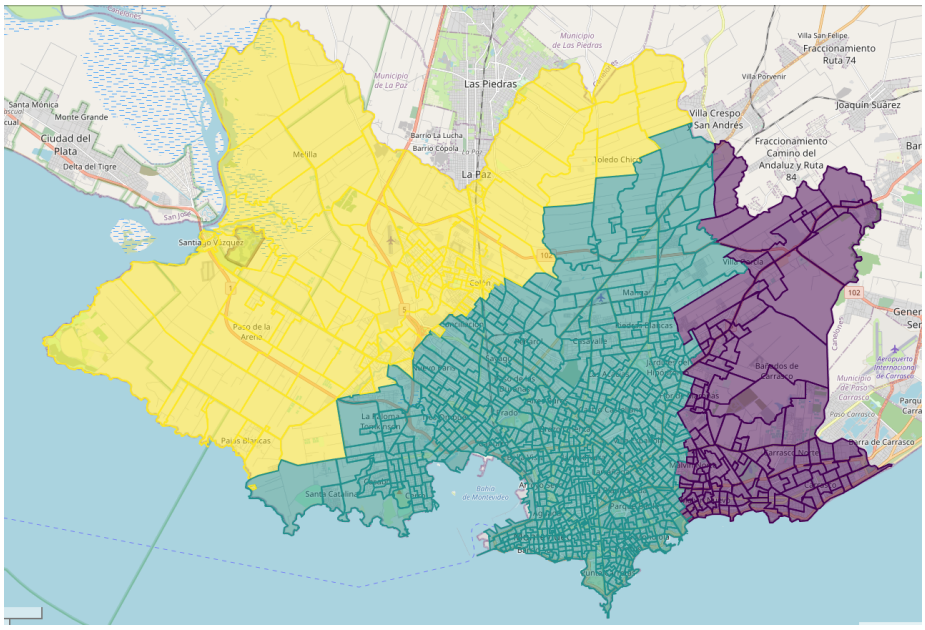

Figura 3.10: Asignación entre estaciones meteorológicas y segmentos

Pasando a analizar los datos de UTE, se debe realizar una correspondencia entre las coordenadas geográficas de los hogares y los segmentos definidos por el INE. Esto a su vez soluciona el problema de tratar con datos personales, ya que al transformar los atributos *latitud* y *longitud* en su correspondiente *sección* y *segmento* ya no es posible ubicar un hogar de forma precisa.

En la tabla 3.9 se muestran algunos registros de UTE luego de la transformación espacial

año	**mes**	**sección**	**segmento**	**tensión**	**tarifa**	**energía_facturada (kWh)**
2019	1	9	44	BT 230 V	TRS	314
2019	4	17	430	BT 230 V	TRD	370
2019	7	99	1	BT 230 V	TRS	633
2019	10	22	134	BT 230 V	TCB	23

Tabla 3.9: Registros de UTE luego de la correspondencia entre hogares y regiones

Por último, se deben tomar todos los registros que coincidan en año, mes, sección y segmento y promediar sus valores, similar a como se hizo con los datos de la ECH en la sección 3.1. El problema es que no todos los atributos de esta fuente son numéricos. El atributo *tensión* es categórico, ya que se filtran los registros que no cumplen *tensión* = BT 230 V, este atributo dejará de figurar ya que no aporta información. El único atributo categórico restante es *tarifa*.

Se decide dividir la variable en 3 nuevas variables que muestren el porcentaje de hogares que utilizan cada tarifa, para las tarifas más utilizadas: Tarifa Residencial Simple (TRS), Tarifa Residencial Doble horario (TRD) y Tarifa de Consumo Básico residencial (TCB).

Otro detalle es que no se promedia la *energía_facturada* de todos los predios de la región, sino que se suman los consumos obteniendo el total de energía facturada por toda la región en el mes. Se denomina este atributo como *energía_facturada_total*. Para que no se pierda la noción de cuántos predios hay dentro de la región, se agrega el atributo *cantidad_predios* que también nos permite obtener el consumo promedio por hogar con una simple división.

Teniendo todo esto en cuenta, el listado final de variables de UTE es el siguiente:

- *año*: Año en el que se realizan las medidas del consumo.
- *mes*: Mes en el que se realizan las medidas del consumo.
- *sección*: Identificador de una sección de Montevideo.
- *segmento*: Identificador de un segmento dentro de una sección de Montevideo.
- *cantidad_predios*: Indica la cantidad de predios presentes en la región.
- *tarifa_ %TRS*: Porcentaje de uso de la Tarifa Residencial Simple en la región.
- *tarifa_ %TRD*: Porcentaje de uso de la Tarifa Residencial Doble horario en la región.
- *tarifa_ %TCB*: Porcentaje de uso de la Tarifa de Consumo Básico residencial en la región.
- *energía_facturada_total*: Suma del consumo eléctrico mensual de todos los predios presentes en la región.

Luego de las transformaciones, se cuenta con un total de 88.985 registros en el conjunto de datos. Se pueden visualizar algunos en la siguiente tabla:

año	mes	sección	segmento	cantidad_predios	tarifa_%TRS	tarifa_%TCB	tarifa_%TRD	energía_facturada_total
2019	9	9	25	511	0,744639	0,132554	0,089669	183.242,964
2019	7	20	42	382	0,725131	0,178010	0,065445	123.892,249
2019	6	17	21	411	0,656885	0,284424	0,042889	100.211,344
2019	2	99	70	246	0,692308	0,210526	0,020243	60.055,770
2019	6	13	51	486	0,646091	0,318930	0,010288	85.162,661
2019	4	16	42	483	0,780992	0,148760	0,047521	113.910,330

Tabla 3.10: Registros finales de UTE, luego de las transformaciones

A este punto ya se tienen todas las fuentes de datos adaptadas para la integración, por lo que se procede a unificarlas como último paso de este capítulo. Cada uno de estos conjuntos de datos tiene como atributos identificadores a la tupla *sección*, *segmento*, *año* y *mes*. La condición de unión estará determinada por dichos atributos.

La cantidad de registros totales luego de la unión es de 36.204 (corresponde a datos de los 431 segmentos, sobre los 12 meses de 7 años), la misma cantidad que la fuente transformada de la ECH. Esto es debido a que este conjunto de datos es el más restringido al no contar con encuestas para todas las regiones de Montevideo. El resultado de la unión solo tomará las regiones que figuren en los tres conjuntos de datos.

Una vez integrados los datos, se presentan las figuras 3.11, 3.12 y 3.13 con mapas que ilustran la relación entre atributos.

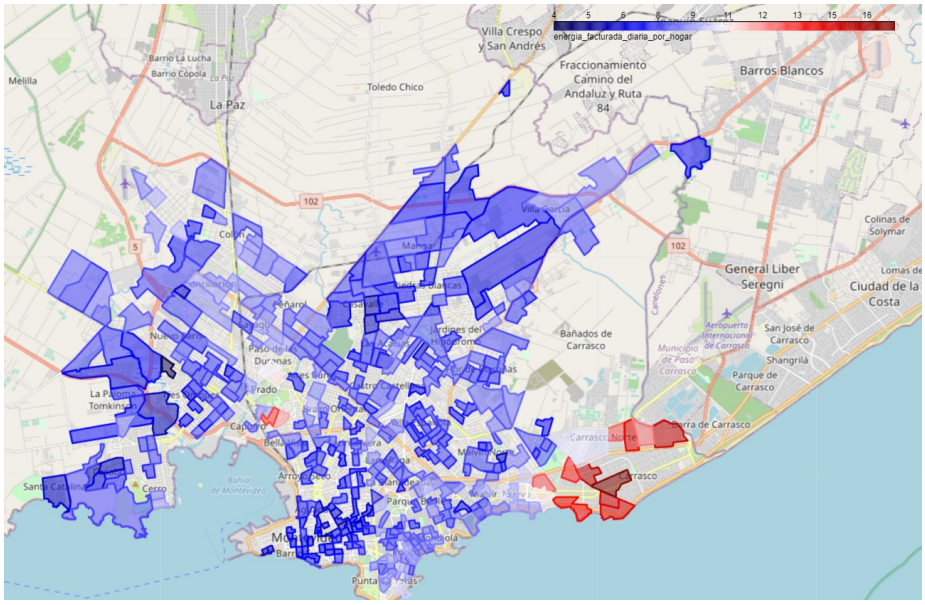

Figura 3.11: Energía demandada por segmento

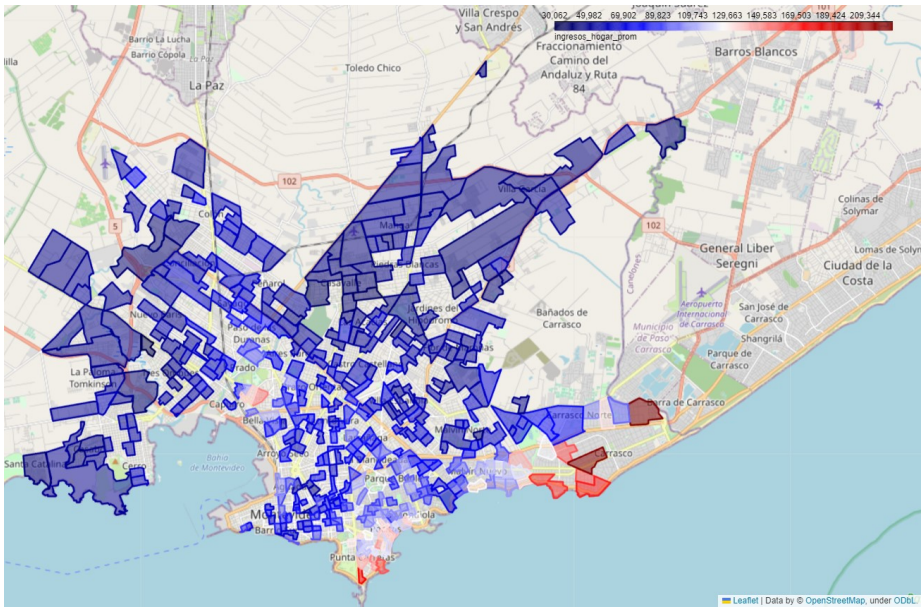

Figura 3.12: Ingresos promedio del hogar por segmento

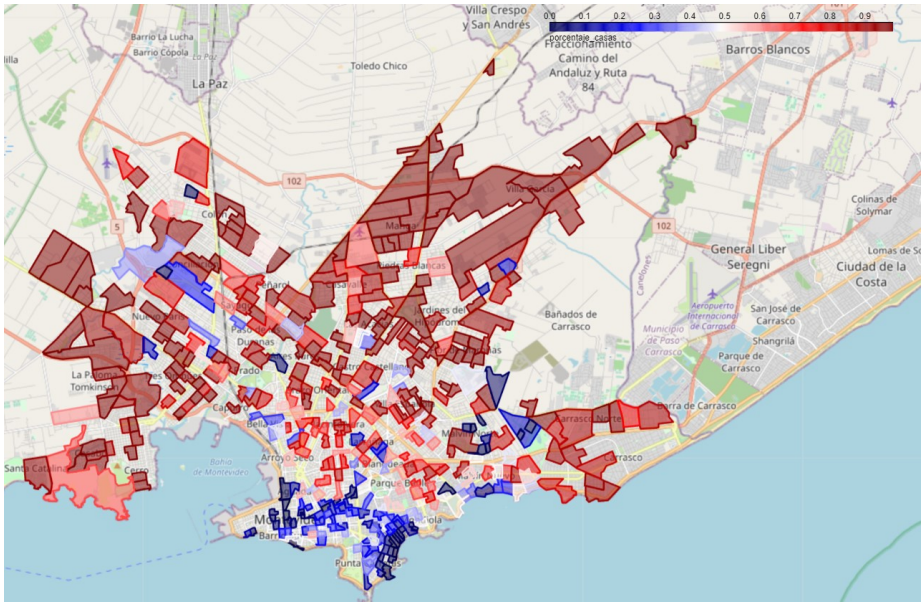

Figura 3.13: Porcentaje de casas por segmento

Resumiendo el capítulo, se han encontrado fuentes de datos en Uruguay para las categorías más relevantes en cuanto a consumo eléctrico que serán utilizadas para nuestro estudio. Los distintos conjuntos de datos fueron tratados e integrados en una única fuente de información, la que será insumo de los modelos de aprendizaje que se definen en el capítulo siguiente.

Capítulo 4

Predicción de consumo eléctrico

Habiendo determinado los datos disponibles, en este capítulo se plantea cómo estos serán utilizados para la predicción de consumo eléctrico. Se adelantó anteriormente que la predicción se realizará en un lapso mensual y sobre regiones dentro de Montevideo, denominadas como segmentos. A su vez se mencionó que se aplicará aprendizaje automático para realizar las predicciones. A lo largo del capítulo se presentan los distintos métodos que se utilizarán y las métricas que evaluarán su rendimiento.

Las predicciones se realizan mensualmente debido a la periodicidad de los registros en las fuentes de datos originales. Como ya se ha visto en el capítulo anterior, cada una trabaja con lapsos temporales distintos. Los datos de INUMET tienen el periodo más frecuente, registrando diariamente las medidas meteorológicas. Por otro lado, los registros de consumo eléctrico de UTE son mensuales y las encuestas de la ECH se realizan anualmente.

Realizar predicciones diarias no sería la opción más adecuada ya que a partir del consumo mensual de los hogares se debería aproximar el consumo diario. la idea más simple en este caso sería dividir el consumo mensual por la cantidad de días del mes, pero esto induce mucho error ya que el consumo no es similar para todos los días de la semana. Otra posible solución sería realizar un estudio de hábitos de consumo eléctrico para intentar aproximar el consumo diario, pero podría ser muy complejo e impreciso, sumado a que se requiere más información de la que se cuenta. A su vez, según lo relevado en el estado del arte, los datos socioeconómicos que proporciona la ECH tienen mayor relevancia a mediano y largo plazo mientras que los factores meteorológicos que brinda INUMET son más adecuados para las predicciones a mediano o corto plazo. Por estos motivos es que se opta por un lapso mensual.

Sobre las regiones en las que se realizan las predicciones, se elige trabajar sobre Montevideo ya que los datos de la ECH cuentan con información extra sobre la ubicación de los hogares para este departamento en particular. En Montevideo se pueden distinguir distintas secciones y segmentos, que son regiones de un tamaño mucho menor al del departamento, pero para el resto del país no. Esta subdivisión permite obtener datos más precisos y diversos en las distintas zonas de Montevideo, lo que se traduce en mejores predicciones para estas regiones.

Antes de presentar los modelos de aprendizaje, se trabaja sobre el conjunto de datos que será utilizado por estos modelos para realizar las predicciones.

4.1. Transformaciones sobre los datos

Como resultado del capítulo 3, obtuvimos una fuente de datos que unifica la información proveniente de la ECH, UTE e INUMET. Esta fuente cuenta con un total de 36.204 instancias y 53 variables: 11 de INUMET, 9 de UTE (que se listaron en el capítulo anterior) y 38 de la ECH (que se pueden observar en el Anexo *Variables utilizadas de la ECH*). Debido a que entre estas fuentes se tienen algunas variables en común, al integrarlas se llega al total de 53 variables. A continuación, se muestran algunas de estas instancias (se incluyen solo 13 de las 53 variables por cuestiones de tamaño):

sección	segmento	año	mes	cantidad_predios	energía_facturada_total	nivel_economico_hogar_prom
13	139	2013	3	618	146.637,00	0,25
9	1	2016	5	100	22.842,00	0,00
20	35	2015	1	510	128.854,00	0,50
24	217	2016	4	551	120.026,00	0,75
12	236	2019	3	705	144.931,00	0,50
15	306	2019	10	1113	213.069,58	0,50

Tabla 4.1: Muestreo de instancias con algunas de las variables más interesantes de la fuente de datos que será utilizada para entrenar los modelos de aprendizaje

%calefaccion_electrica	cant_dias_mes	TmaxPromedio	TminPromedio	HDD	CDD
0,33	31	23,27	16,01	8,80	137,15
0,36	31	14,39	9,50	114,00	3,90
0,00	31	26,94	18,74	2,05	229,55
0,25	30	19,69	15,24	26,28	85,28
0,41	31	24,09	16,77	4,50	157,20
0,40	31	19,74	12,39	47,50	65,00

Tabla 4.2: Muestreo de instancias con algunas de las variables más interesantes de la fuente de datos que será utilizada para entrenar los modelos de aprendizaje (continuación)

Esta fuente contiene toda la información necesaria para los modelos, pero requiere de algunas modificaciones antes de que pueda ser utilizada.

Primero que nada, se decide quitar las variables de sección y segmento. Estas fueron de utilidad para integrar las distintas fuentes de datos por región, pero por sí solas no causan un mayor o menor consumo eléctrico. Son las variables socioeconómicas y meteorológicas las que pueden tener un impacto en el consumo y para no opacar su importancia es que se decide removerlas. Otro problema que generan las variables de sección y segmento es el sobreajuste de los modelos a las regiones. El sobreajuste es un fenómeno que ocurre cuando los modelos se adaptan demasiado a los datos en la etapa de entrenamiento y no son capaces de generalizar bien sobre datos aún no explorados. Se quiere evitar que las predicciones estén sesgadas por su región.

La variable correspondiente al mes se mantiene en el conjunto de datos, pero debe ser modificada. El problema de representarla con números en el rango $[1, 12]$ es que se introduce una diferencia de magnitud entre los meses. Por ejemplo, el mes de diciembre (12) tiene un valor numérico mayor que el mes de enero (1). Esto puede llevar a que el algoritmo de aprendizaje interprete que diciembre es más importante o tiene una influencia mayor en los resultados. Otro problema es que se puede interpretar que diciembre está muy lejos de enero en términos de magnitud, cuando en realidad están muy cerca en el sentido cíclico. Para solucionar estos problemas se decidió utilizar *One Hot Encoding*, que consta en representar cada mes como su propia variable con valores binarios (0 o 1), lo que preserva la naturaleza categórica de los meses y evita interpretaciones erróneas por parte de los algoritmos. Tiene como desventaja que no representa el ciclo temporal de los meses. Esto podría lograrse con otros enfoques, como por ejemplo la representación cíclica con seno y coseno.

Por otro lado, la variable que corresponde al año del registro se decide eliminar del conjunto datos. Esta podría llegar a ser de utilidad para los modelos ya que el consumo eléctrico tiene una tendencia a aumentar con el tiempo, como se pudo observar en la figura 1.1. El problema es que no se cuenta con un espectro de datos lo suficientemente amplio como para que se pueda notar esta tendencia, por lo que podría inducir al sobreajuste.

Otro problema que tiene el conjunto de datos es que sus variables toman valores en rangos de magnitudes muy distintas. Por ejemplo, la variable *ingresos_hogar_promedio* llega a tomar valores de 343.586,42 mientras que otras variables, como por ejemplo las que representan porcentajes, toman valores en el rango $[0, 1]$. Esta diferencia en la escala puede afectar el rendimiento de algunos algoritmos de aprendizaje, ya que pueden otorgarles más importancia a los atributos con rangos más grandes. La normalización transforma los atributos para que tengan una escala similar, lo que ayuda a evitar estos problemas. Para este trabajo se decide utilizar el algoritmo *min-max*, que normaliza los valores numéricos y los transforma dentro del rango $[0, 1]$.

Como siguiente paso, se debe determinar cuál será la variable objetivo. Esta variable es la que se pretende predecir o estimar utilizando los modelos de apren-

dizaje. Como ya se ha mencionado, el objetivo en nuestro caso es la predicción del consumo eléctrico.

Se cuenta con la variable *energía_facturada_total*, que refleja el consumo total de energía de todos los predios dentro de un segmento para un mes. Tomar esta variable como objetivo tiene el inconveniente de que los modelos se ajustan demasiado a la cantidad de hogares que hay en el segmento (variable *cantidad_predios*), opacando la importancia del resto de variables. Otro inconveniente es que algunos meses tienen más días que otros, lo que también se ve reflejado directamente en el consumo total del mes. Para abstraer a los modelos de esto es que se decide crear una nueva variable denominada *energía_facturada_diaria_por_hogar*, que se obtiene al dividir la *energía_facturada_total* de una región por la cantidad de días del mes y por la *cantidad_predios* en la región. Esta nueva variable será entonces la variable objetivo.

4.2. Instancias de aprendizaje

Luego de las transformaciones se sigue contando con un total de 36.204 instancias ya que no fue modificada la cantidad, y 60 variables (incluyendo a la variable objetivo) para realizar el entrenamiento. Se presentan instancias con algunas de las variables del conjunto de datos:

mes_1	**mes_2**	**mes_3**	**cantidad_predios**	**energia_facturada_diara_por_hogar**	**tarifa_%TRS**	**%casas**
1	0	0	0.4666	7.9760	0.7818	0.1126
1	0	0	0.1881	7.8928	0.8509	1.0000
0	0	0	0.2914	6.4057	0.8610	0.7921
0	0	0	0.1738	8.7544	0.7884	0.3729
1	0	0	0.1475	7.4458	0.7009	0.8246
0	0	0	0.2379	8.3974	0.7582	0.8020

Tabla 4.3: Muestreo de instancias de la fuente de datos que será utilizada para entrenar los modelos de aprendizaje luego de las transformaciones con algunas de las variables más interesantes

nivel_economico_hogar_prom	**%calefaccion_electrica**	**TmaxPromedio**	**TminPromedio**	**HDD**	**CDD**
1,00	0,45	0.9095	0.8947	0.0056	0.9175
0,00	0,33	0.8674	0.8718	0.0067	0.8590
0,75	0,33	0.1064	0.1225	0.8716	0.0386
0,50	0,17	0.0214	0.0721	0.9766	0.0001
0,00	0,43	0.9210	0.8955	0.0019	0.9269
0,50	0,30	0.2447	0.2589	0.5708	0.0988

Tabla 4.4: Muestreo de instancias de la fuente de datos que será utilizada para entrenar los modelos de aprendizaje luego de las transformaciones con algunas de las variables más interesantes (continuación)

De las 36.204 instancias que se tienen se forman dos conjuntos de datos separados, uno de entrenamiento con el 80 % de la información, y otro con el

20 % restante para *test*. Esta división se realiza para evaluar el rendimiento de un modelo de manera objetiva y realista. Los modelos aprenden del conjunto de entrenamiento y una vez que termina este proceso se evalúan utilizando el conjunto de prueba para medir su rendimiento en datos no vistos previamente. La repartición de instancias entre estos conjuntos se realiza al azar para garantizar una representación diversa de los datos en ambos conjuntos. No se utiliza ninguna técnica de estratificación ya que no se considera necesaria. Tampoco se hace uso de un conjunto de validación, ya que se realiza validación cruzada de 5 iteraciones (o *folds*) para elegir los mejores hiperparámetros para los modelos.

Habiendo presentado los datos, se continúa el capítulo especificando los distintos modelos que se utilizan.

4.3. Línea Base

Como primer paso se definen líneas base. Una línea base es un punto de referencia inicial para comparar el desempeño de otros modelos. Por lo general son sencillos y se espera que los modelos más complejos sean capaces de superarlos y proporcionar mejoras significativas en términos de precisión.

Para este estudio se definen dos líneas base:

La **Línea Base 1** consiste en que, dado un mes a predecir, se tome el promedio del consumo de ese mismo mes para todos los años anteriores que se tengan en los datos del conjunto de entrenamiento. Este valor será el resultado de la predicción.

La **Línea Base 2** es similar a la Línea Base 1. Dado un mes, un segmento y una sección, toma el promedio del consumo del mes a predecir en todos los años anteriores, pero sólo para el segmento y sección correspondientes. Esta línea base toma atributos que no tendrán los otros modelos (sección y segmento), por lo que tiene cierta ventaja en ese aspecto

La diferencia entre ambas es que en una se obtiene el consumo del mes para todos los segmentos juntos, mientras que en la otra para uno en particular.

4.4. Métodos

Para escoger los modelos de aprendizaje automático a utilizar, se tuvieron en cuenta los siguientes criterios:

- Deben haber sido aplicados en la literatura y en el estado del arte estudiado. Esto tiene como objetivo seleccionar métodos ya probados.

- Deben ser modelos de diferente tipo para poder experimentar varios enfoques. Cada tipo de modelo tiene sus fortalezas y debilidades, por lo que probar con varios ayuda a determinar cuál es el más adecuado para la tarea.

- Deben ser modelos que se espere que den buenos resultados con la cantidad de datos disponibles para este enfoque. Por este motivo se descarta el uso de redes neuronales para este trabajo.

- En lo posible, deben ser métodos que permitan jerarquizar las variables por su impacto en la predicción.

Teniendo en cuenta estos criterios, se escogen los modelos: KNN como método *lazy* de vecino cercano, Regresión lineal como modelo clásico y *Random Forest* como método más moderno de árboles de decisión.

Todos estos modelos son formas de aprendizaje supervisado, los algoritmos entrenan haciendo uso de un conjunto de datos etiquetados, donde cada instancia de aprendizaje tiene asociada una etiqueta que indica su categoría o valor esperado (en nuestro caso, la variable objetivo energía_facturada_diaria_por_hogar). Estas etiquetas se utilizan tanto para entrenar como para evaluar.

Regresión Lineal

Este método busca establecer una relación lineal entre las variables para predecir valores numéricos, o sea que partiendo de un vector $x^T = (x_1, x_2, \ldots, x_n)$ con n variables se busca construir una función (hipótesis) $h_\theta(x) : R^n \to R$ que prediga la salida $y \in R$, continua, a través del siguiente modelo:

$$h_\theta(x) = \theta_0 + \sum_{j=1}^{n} x_j \theta_j$$

La regresión lineal permite obtener la importancia relativa de las variables para la predicción a través de los coeficientes θ. Esto sólo es aplicable si todas las variables están normalizadas, ya que de lo contrario esto afecta a la magnitud de los coeficientes.

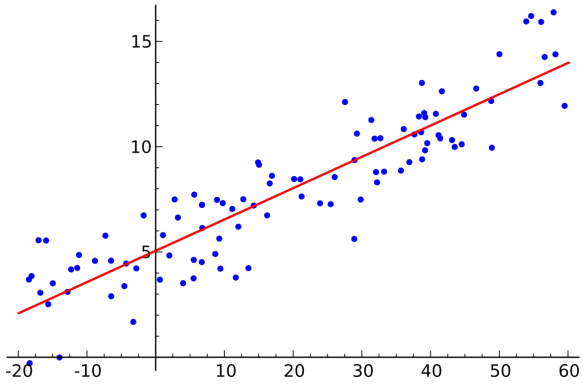

Figura 4.1: Representación gráfica del modelo de Regresión Lineal

Un detalle es que este modelo asume que la relación entre variables puede aproximarse con una recta, por lo que podría llegar a no capturar relaciones más complejas o no lineales presentes en los datos.

Regresión con KNN

El algoritmo *K-Nearest Neighbors* (KNN) se utiliza en problemas de clasificación y de regresión. La diferencia entre estos radica en el tipo de problema que se está abordando y en lo que se quiere predecir. En los problemas de clasificación se quiere determinar una etiqueta o categoría para las instancias a evaluar, mientras que en los problemas de regresión lo que se quiere predecir son valores numéricos continuos, lo que aplica a nuestro caso.

El algoritmo se basa en la idea de que dos instancias que están cerca en el espacio son similares entre sí. En otras palabras, el valor a predecir de una instancia nueva se obtiene de los valores de las k instancias más cercanas ya evaluadas. Se debe definir como calcular la cercanía entre instancias.

Entonces, dada una nueva instancia a predecir representada por un vector $x^T = (x_1, x_2, \ldots, x_n)$, donde cada x_i refiere a las distintas variables de la instancia, se buscan los k vectores más cercanos dentro del conjunto de entrenamiento. La función para calcular la distancia puede variar, pero es común utilizar la distancia euclídea y este es el caso en este trabajo. Para determinar el valor de la variable objetivo a partir de los k vecinos más cercanos, se utiliza interpolación a partir de las variables pertenecientes a los k vectores.

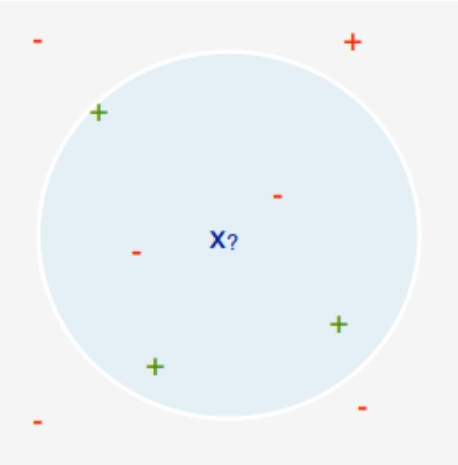

Figura 4.2: Representación gráfica del modelo KNN[1]

Para este modelo se cuenta con el hiperparámetro k de cantidad de vecinos. Se utiliza validación cruzada para probar cuál de los valores de $k \in \{3, 5, 7, 11\}$ es el más adecuado en este caso.

Random Forest

Random Forest también puede ser utilizado para problemas de clasificación y regresión. En nuestro caso la variable objetivo es numérica por lo que se busca solucionar un problema de regresión.

Este algoritmo utiliza varios árboles de decisión, que son estructuras de datos que sirven para tomar decisiones o realizar predicciones basándose en ciertas reglas y en las características de los datos de entrada. Los árboles están compuestos por nodos y ramas, en donde un nodo representa una variable del conjunto de datos y las ramas las posibles decisiones que se toman dependiendo de las reglas que se definan sobre la variable.

Los árboles de decisión están entrenados en diferentes subconjuntos de datos, cada uno con algunas variables elegidas al azar. Cada árbol emite su propia predicción y se combinan los resultados para obtener el resultado final. Como en nuestro caso se predice un valor numérico, se calcula la media de los valores que cada árbol predice para una entrada.

Además, utilizando *Random Forest* se puede obtener la importancia de las variables para la predicción. Esto se determina en base a la media y la desvia-

[1]Representación gráfica del modelo KNN extraída del material del curso "Aprendizaje Automático" de la Facultad de Ingeniería, UdelaR

ción estándar de la impureza en cada árbol.

Este modelo cuenta con hiperparámetros que refieren a la $profundidad$ máxima del árbol y a la cantidad de $árboles$ dentro del modelo. Se probarán los valores: $profundidad \in \{5, 8, 15, 20, 40, 100, 200, 300\}$, $árboles \in \{50, 100, 200, 300\}$ mediante validación cruzada.

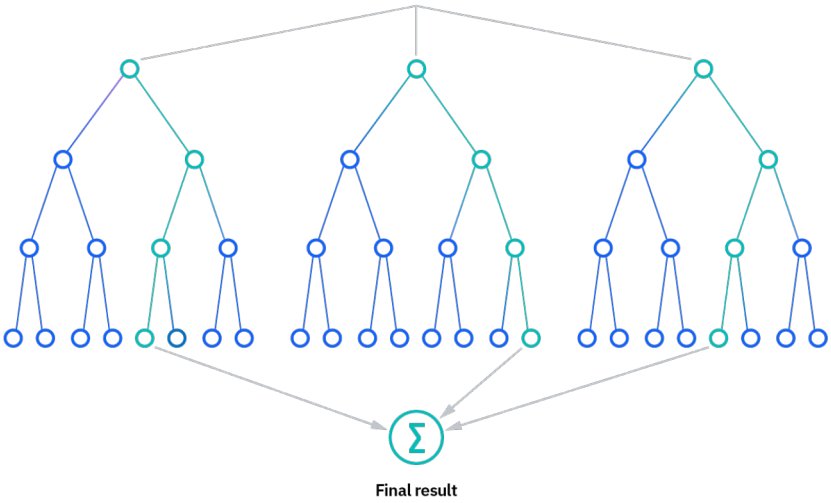

Figura 4.3: Representación gráfica del modelo *Random Forest*. Imagen recuperada del sitio web de IBM[3]

4.5. Métricas

Para evaluar el rendimiento de los distintos modelos se utilizarán las medidas $MAPE$, que consiste en calcular la media de los valores de error porcentual absoluto de todos los valores predichos, y R^2, que es el porcentaje de la variación en la variable de respuesta que es explicado por un modelo lineal. Se escogen estas métricas ya que son ampliamente usadas en otros trabajos relacionados.

Habiendo definido las métricas, los modelos, y líneas base para compararlos, se presenta el análisis experimental en el siguiente capítulo.

[3]Sitio web de IBM sobre *Random Forest*: https://www.ibm.com/topics/random-forest

Capítulo 5

Resultados

Una vez definidos los modelos a utilizar y las maneras de evaluarlos, se procede con la etapa de entrenamiento seguida de la obtención de los resultados sobre el conjunto de *test*.

Las predicciones realizadas utilizando los datos socioeconómicos y meteorológicos obtuvieron mejores resultados que las líneas base definidas, según las métricas seleccionadas previamente. Se puede destacar que la ejecución del aprendizaje de estos métodos no fue costosa en términos de tiempo [1]. Esta evaluación presenta al método *Random Forest* como el mejor, superando a todos los demás en ambas métricas. Utilizando este método se obtienen predicciones con un error absoluto promedio de 5.6 %. Los resultados totales se pueden ver en la tabla 5.1. Es importante recordar que un MAPE bajo es una mejor evaluación, debido a que establece que el error promedio de las predicciones es más pequeño. Con respecto a R^2 esto es al revés, valores mayores, y por lo tanto más cercanos al máximo de 1, indican que el modelo predice con mayor precisión.

Modelo	MAPE	R^2
Random Forest	**0.056**	**0.85**
Línea base 2	0.077	0.71
KNN	0.083	0.69
Regresión Lineal	0.091	0.64
Línea base 1	0.130	0.14

Tabla 5.1: Rendimiento de los distintos modelos y líneas base

Como se planteó anteriormente, es de interés analizar la importancia relativa de las diferentes variables utilizadas para la predicción. Los modelos de *Random*

[1] Tiempo promedio (sobre 3 pruebas) de ejecución de entrenamiento y testeo de los modelos con una CPU AMD RYZEN 7 3700X: Regresión Lineal 1 segundo, KNN 6 segundos, *Random Forest* 1670 segundos

Forest y Regresión Lineal proveen maneras de obtener información sobre esto según el peso que tiene cada variable en la predicción. Dado que *Random Forest* obtuvo los mejores resultados para las métricas seleccionadas, se utiliza este modelo para representar la importancia de las variables, que se puede observar en la figura 2.

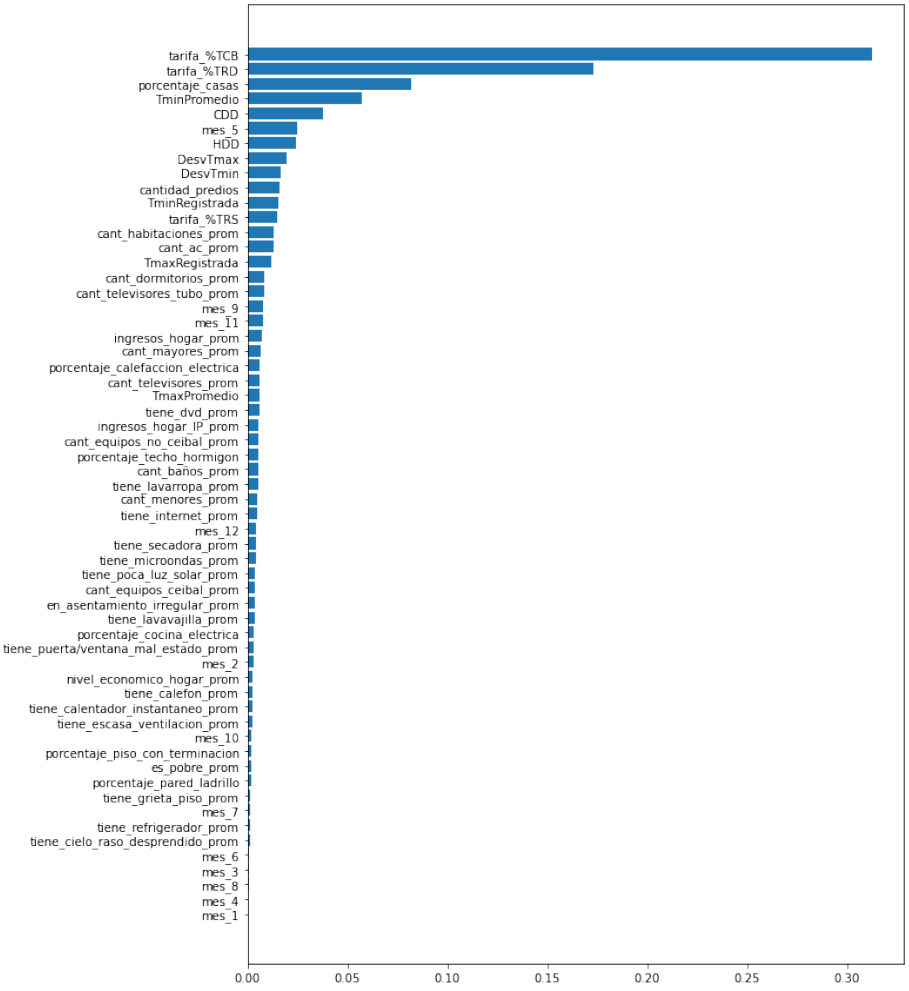

Figura 5.1: Importancia de las variables en el modelo de Random Forest

Los resultados obtenidos coinciden con lo esperado, según el estudio de trabajos relacionados. Se observa una gran importancia de las variables relacionadas con la temperatura y de algunas variables socioeconómicas, por ejemplo cantidad de habitaciones o ingresos del hogar entre otros.

Los resultados sugieren que las diferentes tarifas eléctricas de los hogares están fuertemente relacionadas con el consumo eléctrico que estos presentan. Un claro ejemplo de esto es el porcentaje de hogares que utilizan la "Tarifa de consumo básico", siendo la variable más importante en el modelo de Random Forest. Esto es esperable, ya que los hogares que la utilizan no pueden consumir más de 230 kWh/mes sin penalizaciones económicas.

A su vez, el porcentaje de hogares con la "Tarifa residencial doble" también es considerada extremadamente importante en ambos modelos. Una posible explicación es que los hogares consuman más electricidad, sabiendo que en ciertos horarios es más barata.

Las variables relacionadas a la temperatura también son consideradas muy importantes en el modelo, teniendo más relevancia que la mayoría de los factores socioeconómicos.

Dentro de las variables socioeconómicas, se puede destacar la importancia del porcentaje de casas en el segmento, siendo más importante que la temperatura para determinar el consumo con el modelo. Una gran parte de las variables que se relevan en la ECH están relacionados con el nivel socioeconómico del hogar. Por lo tanto, al estar fuertemente ligadas la importancia se puede ver repartida en cada una.

Inesperadamente, un factor importante para determinar el consumo es si el mes es mayo. Dado que no se conoce una razón ligada a la realidad que explique un aumento o decremento de consumo en este mes, se cree que el modelo lo utiliza como una variable similar a la temperatura.

Se puede observar que todos los conjuntos de datos aportan información de utilidad para la predicción, ya que, entre las cuatro variables más importantes para el modelo aparece al menos un dato de cada conjunto (dos de UTE, uno de ECH y uno de INUMET).

En el siguiente capítulo se lleva a cabo una aproximación al estudio causal con este mismo conjunto de datos.

Capítulo 6

Causalidad

Con el siguiente estudio de causalidad se pretende complementar el análisis anterior, ya que la importancia de las variables para la predicción no determinan la existencia o inexistencia de relaciones causales con respecto al consumo eléctrico. Es de interés determinar si pueden existir estas relaciones, a través de un estudio causal simple.

Es común tener una intuición de lo que son las relaciones causa-efecto (cambios en una variable que causan que otra variable también presente cambios), pero no es trivial demostrar que estas existan realmente. La disciplina de la causalidad tiene como objetivo presentar la relación causa-efecto desde un punto de vista estadístico.

La inferencia causal (o causalidad) es el proceso en el cual las causas son determinadas a partir de datos y una hipótesis de relaciones de causa-efecto entre ellos. A través de procesos estadísticos es que se valida la hipótesis de causalidad sobre los datos.

Dado que este estudio se basa en la estadística, cuando se intenta probar una relación causal, lo máximo que se puede afirmar es que la relación no pudo ser probada como falsa (refutar la hipótesis nula), utilizando un umbral probabilístico estándar de 5 %.

Es muy importante distinguir entre los conceptos de causalidad y correlación entre variables. Dos variables que estén relacionadas no implican causalidad. Un ejemplo de esto es la relación entre dormir con zapatos puestos y el dolor de cabeza al levantarse. En un análisis estadístico, estas dos variables pueden estar relacionadas, ya que se podría determinar que no son independientes, pero no implica que una cause la otra. Si a su vez se incorpora el dato de consumo alcohólico la noche anterior, se puede ver cómo esta variable podría afectar ambas, y ser una causa común para las otras dos. Nótese que esto se puede determinar sólo con conocimiento de la realidad, porque únicamente con un estudio

estadístico no se puede llegar a esta conclusión.

También puede ocurrir que las variables parezcan dependientes, pero no tengan ninguna relación entre ellas en la realidad. Un posible ejemplo de esto es la figura 6.1, que presenta una correlación muy fuerte entre los divorcios en Maine (estado de EE. UU.) con el consumo per cápita estadounidense de queso. Quizás haya una causa común para ambas variables, o es simple coincidencia, pero no es posible determinarlo exclusivamente con los datos.

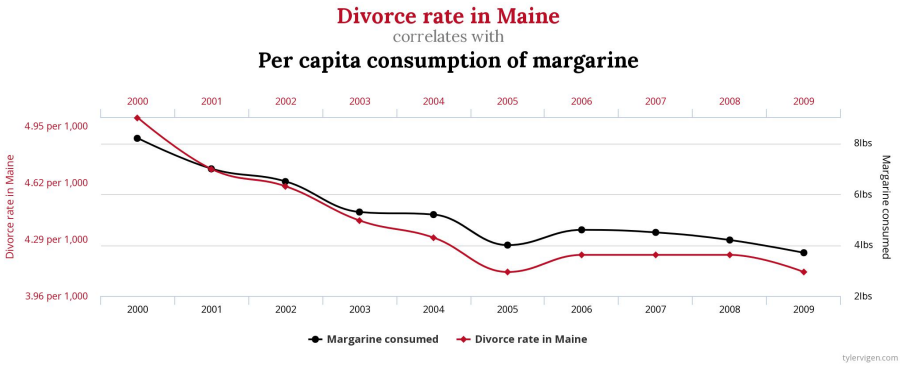

Figura 6.1: Relación entre Divorcios en Maine (estado de EE. UU.) con el consumo per cápita estadounidense de margarina. Recuperado de https://www.tylervigen.com/spurious-correlations

Por este motivo es necesario incluir una suposición de la realidad antes de realizar un estudio causal. Sin esto, no habría manera de diferenciar simples correlaciones de reales efectos causales.

Este capítulo presentará un análisis superficial de la inferencia causal, comenzando con una introducción teórica de la causalidad, para continuar con su aplicación, en nuestro caso de estudio. Este estudio se basa principalmente en los trabajos de Neal (s.f.), Pearl (2009) y la biblioteca de programación DoWhy[1].

6.1. Conceptos

A continuación, se listan una serie de conceptos básicos, de manera resumida, que forman parte de la teoría de la inferencia causal, que permitirán entender las características de este tipo de análisis.

[1] https://github.com/py-why/dowhy

Correlación

El coeficiente de correlación de Pearson, pensado para variables cuantitativas, es un índice que mide el grado de covarianza entre distintas variables relacionadas linealmente. Intuitivamente expresa cuanto varía una variable observando la variación de otra variable.

Causa y efecto

Sean dos variables T, Y, se dice que Y es una causa de T si al variar el valor de T también lo hace el valor de Y

Se entiende que T causa Y si un cambio de T modifica a Y, siempre y cuando todas las demás variables no se modifiquen. Esto último es algo muy importante y es la diferencia entre causa y correlación. Es complejo el análisis causal debido a que, para estar totalmente seguros de una implicancia, se deberían conocer todas las variables que afectan al resultado, lo que es difícil de constatar en la gran mayoría de los casos.

El operador *do*

El operador *do* denota una intervención en las entidades dentro de un estudio. A diferencia de condicionar, que se denota $P(Y = y|X = x)$, donde se quiere estudiar cual es la probabilidad de que la variable Y tome el valor y dado $X = x$, es decir, se trabaja con el subconjunto de las entidades en donde la variable X es x, intervenir, que se denota $P(Y = y|do(X = x))$, quiere estudiar la misma probabilidad pero siendo x el valor de la variable X, trabajando en este caso con todas las entidades y forzándolas a tomar $X = x$.

A modo de ejemplo, tomamos un estudio clínico donde se quiere estudiar el efecto de un nuevo medicamento M en el tratamiento de una enfermedad E. La variable objetivo en este caso se llamará C, y tomará el valor 0 si el paciente no se curó y 1 si se curó, la variable de tratamiento T tomará el valor 0 si el paciente se trató con bajo el método convencional y 1 si fue tratado con M. Cuando se quiere calcular $P(C = 1|T = 1)$ se busca obtener la probabilidad de que un paciente se curase dentro del subconjunto de los pacientes que recibieron el nuevo tratamiento. Por otra parte $P(C = 1|do(T = 1))$ busca obtener la probabilidad de que un paciente se curase sometiendo a todos los pacientes al nuevo tratamiento.

Efecto causal

El efecto causal de X en Y es la magnitud del cambio de Y dado un cambio en T. Se presenta la definición de efecto causal para un caso con T tomando valores 0 o 1.

Sea Y_i la variable resultado, y T_i la variable tratamiento binaria, para la i-ésima entidad. El efecto causal esta dado por la siguiente expresión, comúnmente conocida como *Individual Treatment Effect* (ITE)

$$Y_i(do(T_i = 1)) - Y_i(do(T_i = 0))$$

Promediando el ITE para toda la población determina el *Average Treatment Effect* (ATE).

La complejidad del cálculo del efecto es que en la realidad si $T_i = 0$, entonces no se conoce el resultado de haber ocurrido $T_i = 1$, llamado el contrafáctico (*counterfactual* en inglés), por lo que siempre es necesario estimarlo. Esta imposibilidad de medir el resultado bajo un tratamiento distinto al observado es comúnmente llamado el "problema fundamental de la inferencia causal".

Grafo causal

Los grafos causales utilizan grafos acíclicos dirigidos (DAG, por sus siglas en inglés) se pueden utilizar como un modelo gráfico para representar relaciones causales entre variables. Son esenciales para el estudio causal, ya que no se pueden realizar sin tener un modelo de la realidad con el que partir.

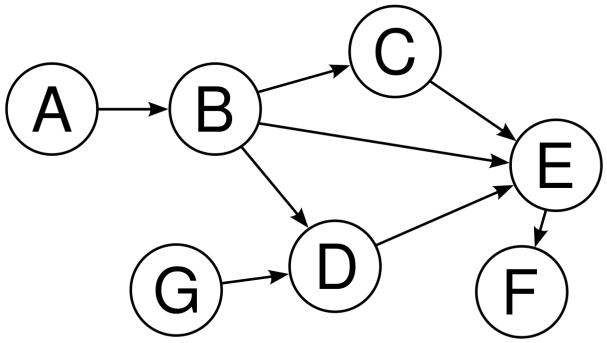

Figura 6.2: Ejemplo de un DAG

Cada nodo en el grafo representa una variable, y dados dos nodos A y B, una flecha de A hacia B indica que existe una relación causal entre A y B, en este caso, cambios en la variable A generan cambios en la variable B.

Confounder o causa común

Un *confounder* es una variable que tiene relaciones causales con dos o más variables con las cuales se quiere probar si existe una relación causal. La palabra "Confounder" en inglés, ilustra la confusión que puede llegar a generar esta

variable, ya que genera correlación entre ellas.

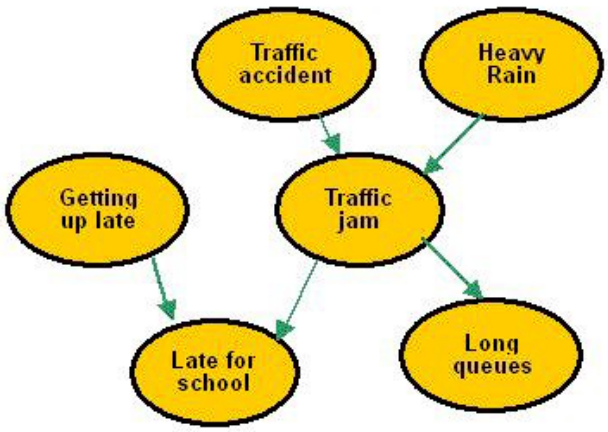

Figura 6.3: Ejemplo para ilustrar la causa común, recuperado de https://towardsdatascience.com/implementing-causal-inference-a-key-step-towards-agi-de2cde8ea599

En la figura 6.3, "traffic jam" (atasco de tráfico) es una causa común para "late for school" (llegar tarde a la escuela) y "long queues" (filas largas). Si se agregara una arista apuntando desde "late for school" a "long queues" y se quisiera calcular el efecto causal de esa arista, se debería de alguna manera poder diferenciar el efecto de la variable "traffic jam" sobre ambas variables de manera de poder calcular meramente el efecto de "late for school" sobre "long queues". Si el modelado del DAG de la realidad es correcto, se puede afirmar que "traffic jam" es un confounder para las otras dos variables, y que entre estas dos últimas no hay relación causal.

En un estudio causal siempre se deben de buscar las causas comunes, ya que estas introducen correlaciones entre variables que pueden llegar a confundir. Incluso existe la posibilidad de que alguna variable que no haya sido considerada en el estudio pueda ser un *confounder* de variables que forman parte. Es entonces que el problema radica en controlar el efecto de estos *confounders* de manera que no produzca un sesgo en el cálculo del efecto causal. Este problema es muy desafiante, ya que como siempre se trabaja con modelos de la realidad, siempre se debe intentar tomar en cuenta esto como posibilidad. Más adelante, se verán métodos que permiten evaluar el impacto de esto en los estudios (utilizando un estudio de refutación).

6.2. Biblioteca DoWhy

La biblioteca DoWhy, es una biblioteca mantenida por Microsoft para facilitar el análisis Causal. Está implementada en Python y cuenta con documentación de uso e introducciones al pensamiento causal.

Para el uso de la biblioteca se incita a acercarse al análisis con una guía de 4 etapas, los cuales se describen a continuación:

Modelado
Esta etapa se trata principalmente de construir un grafo causal que modele las suposiciones sobre el dominio. De esta manera quedan definidas las relaciones causales entre las distintas variables que serán posteriormente contrastadas con el análisis de los datos tanto para ponderar el efecto causal de una variable de tratamiento sobre la variable objetivo, así como para intentar contradecir estas suposiciones.

Identificación
En esta etapa a partir del grafo anterior se busca definir un estimando que permita calcular el efecto causal de la variable de tratamiento sobre la variable objetivo. En esta etapa se busca aplicar algún criterio, como es el *backdoor*, que funciona bajo las suposiciones de la inferencia causal y que permite expresar el efecto causal de tratamiento sobre la variable objetivo como una función matemática. Aquí no se profundizará en cómo funciona el ajuste *backdoor*, y se tomará como método de la biblioteca.

Estimación
En esta etapa se computa el estimando a través de algún método de aproximación. Esta etapa resulta en una estimación del efecto causal que permite calcular el valor de la variable objetivo que se obtendría al variar el valor de la variable de tratamiento.

Para la etapa de estimación la biblioteca dispone distintos métodos para generarla. Entre las posibilidades se encuentran modelos de aprendizaje automático o simples cálculos estadísticos.

Refutación
En esta etapa se ejecutan varias pruebas para intentar refutar la estimación anterior y calcular qué tan confiable es la estimación obtenida.

Las pruebas provistas por la biblioteca son:

- **Random common cause** Se agrega una variable aleatoria independiente como una causa común al tratamiento y la variable objetivo. Si la suposición es correcta, la estimación no debería cambiar.

- **Data subset refuter** Reemplaza el conjunto de datos con un subconjunto de este, seleccionado al azar. Donde si la suposición es correcta la estimación no debería cambiar de forma drástica.

- **Placebo treatment refuter** Se reemplazan los valores de la variable de tratamiento con valores elegidos de forma aleatoria. En este caso si la suposición es correcta la nueva estimación, que se hace a partir de esta nueva variable, debería acercarse a cero.

6.3. Estudio causal

A continuación, se presenta la utilización de la biblioteca DoWhy para nuestro caso de estudio: la causalidad en el consumo eléctrico hogareño.

Este estudio busca generar una estimación del efecto causal de múltiples variables de índole social y meteorológico sobre el consumo eléctrico en los hogares. Por lo tanto, es necesario estimar el efecto causal individualmente, generando una ejecución de identificación, estimación y refutación por cada variable que se quiera estudiar.

En primer lugar, se genera un grafo partiendo de las suposiciones de relaciones causales entre las distintas variables. Se toma un grafo simplificado, para que sea fácil de entender. Dado que se tienen una gran cantidad de variables dentro del conjunto de datos, colocarlas todas en este ejemplo de acercamiento a la causalidad complejizaría innecesariamente el escenario. Para seleccionar las variables se utilizó como entrada la importancia de las variables a la hora de la predicción.

El resto de las variables que no aparecen en el grafo, se consideran posibles causas comunes o *confounders*, es decir, aparecerán como nodos apuntando tanto al tratamiento como al objetivo. En la imagen 6.4 se puede observar el grafo utilizado para la experimentación. La variable objetivo siempre será la energía facturada diaria por hogar, sin embargo, se realizan pruebas con distintas variables de tratamiento.

Se debe considerar que la importancia de las variables en el aprendizaje, y el efecto causal que tenga cada una, no son dos factores estrictamente relacionados. Esto es debido a que ambos estudios tienen objetivos distintos. El estudio causal posee como insumo el grafo, y busca calcular el efecto real del tratamiento, mientras que la relevancia dentro de los algoritmos de aprendizaje automático utilizados simplemente responden a las variables que más influyen en el resultado, pudiendo responder a simples correlaciones. Muchas de las variables del conjunto de datos están relacionadas, ya sea porque algunos factores socioeconómicos como los ingresos afectan a la mayoría de los otros factores, o porque algunas variables dentro del conjunto son calculadas a partir del mismo dato original.

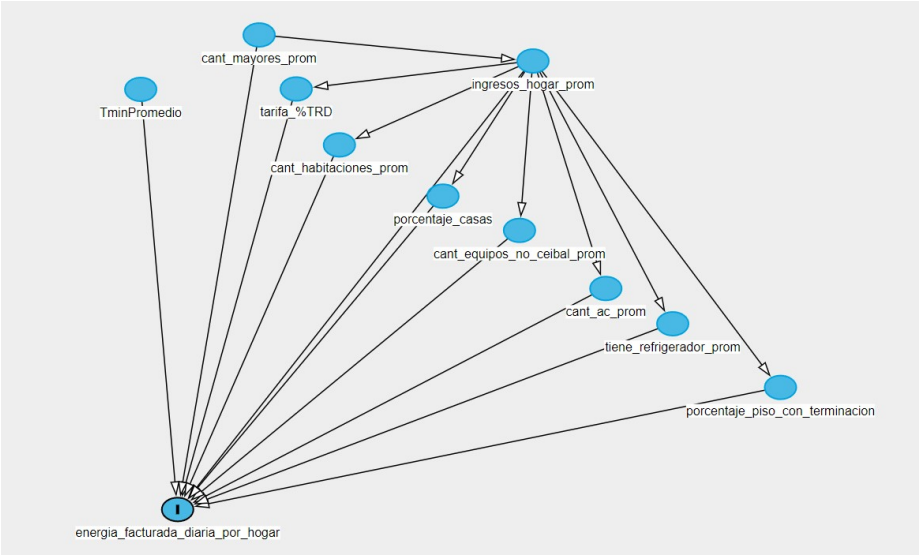

Figura 6.4: Grafo causal correspondiente a ambos casos

En la etapa de identificación, como se mencionó anteriormente, los estimandos están limitados por las características del grafo causal. Específicamente el estimando que la biblioteca permite aplicar para el ejemplo es el correspondiente al de aplicar el *backdoor criterion*.

La naturaleza continua de las variables a estudiar, tanto de tratamiento como también la variable objetivo, marca fuertemente los métodos de estimación que la biblioteca permite utilizar. En particular la utilización de una variable de tratamiento continua impide la estimación mediante la familia de métodos de *propensity score*. A su vez, en la realidad modelada, no aplican métodos de variable instrumental (*instrumental variables* en inglés). Esto significa que se deben utilizar métodos de regresión, como por ejemplo la Regresión Lineal que se encuentra presente en la biblioteca. Otros modelos de regresión pueden ser importados de bibliotecas de terceros.

6.4. Resultados

Se presenta el caso de análisis del efecto causal para la variable ingreso_hogar_prom, que se trata del ingreso del hogar promedio, sobre la variable objetivo que es energia_facturada_diaria_por_hogar.

La identificación del estimando para el efecto causal se calcula a partir del grafo causal utilizado. DoWhy se encarga internamente de definir cómo se hará el cálculo en el siguiente paso de estimación, y se menciona que dado el grafo la variable de cant_mayores_prom es un *confounder*.

En la etapa de estimación, debido a que el tratamiento es una variable continua se aproxima el estimando utilizando regresión lineal. La biblioteca calcula la media de efecto causal para el intervalo definido por los valores ingreso_hogar_prom. El efecto causal medio calculado corresponde a $1{,}868 * 10^{-5}$.

Este valor parece muy pequeño en primera instancia, pero representa la diferencia en la variable objetivo al aumentar en 1 la variable de tratamiento. Esto quiere decir que si para un segmento se mantienen fijos todos los valores de las variables y se aumenta el ingreso del hogar en $10.000, la energía consumida diaria será incrementada en 0,18868 kWh. Teniendo en cuenta que el promedio de energía facturada por mes en los hogares de Montevideo para el año 2019 es de 249,6 kWh, se tiene aproximadamente un promedio de consumo diario de 8,32 kWh. Entonces tenemos que un aumento del ingreso en $10.000 genera un incremento del 2,2678 % en el consumo eléctrico.

Finalmente se llevan a cabo las pruebas de refutación. En ninguna de las tres pruebas es posible afirmar que no existe una relación causal entre los ingresos del hogar y la energía consumida diaria. Por lo tanto, se puede pensar que existe un efecto causal entre ambas, dado el grafo de la realidad planteado y el conjunto de datos utilizado.

Se realizó un estudio análogo con la variable correspondiente al porcentaje de casas, y se obtuvo un efecto causal de 1,139 kWh, que corresponde a que un segmento tuviese 1 % más de casas entonces consumiría 1,139 kWh más energía diaria por hogar. Al igual que fue mencionado previamente, esto es considerando que el resto de las variables se mantienen exactamente iguales para ambos segmentos.

Nuevamente, como en el caso anterior, las pruebas de refutación no permiten descartar la relación causal entre el porcentaje de casa y el consumo eléctrico diario por hogar.

Es importante recordar que estos valores reflejan el efecto causal promedio, lo que quiere decir que muchas de estas variables podrían tener un efecto causal mayor o menor dependiendo del valor que éstas tomen.

Habiendo presentado este ejemplo práctico de causalidad, se presentan las conclusiones sobre todo el proyecto en el siguiente capítulo.

Capítulo 7

Conclusiones

Este estudio buscaba generar un análisis del consumo de energía eléctrica en Uruguay a través del uso de aprendizaje automático para su predicción, para lo que se indagó dentro de distintos organismos sobre los datos disponibles, desde el consumo propiamente dicho, así como también a otras variables que incidieran sobre él.

Se logró relevar un estado del arte, estudiando trabajos relacionados actuales. Gracias a este relevamiento pudimos ver las tendencias de investigación en esta área con lo cual tener una visión hacia donde poder dirigir nuestro estudio. Existe mucha literatura reciente al respecto y variada en cuanto a diferentes plazos de predicción y métodos utilizados.

Se investigó sobre la disponibilidad de información en Uruguay que pudiera ser de utilidad para nuestra tarea, basándonos en lo aprendido de la literatura. Se obtuvo información de tres fuentes distintas, de forma que se pudo cubrir con las categorías más relevantes de información para la predicción del consumo eléctrico. Se realizaron controles de calidad y transformaciones sobre estos datos para que pudieran integrarse en una única fuente de información para los modelos de aprendizaje.

A partir de los datos que se obtuvieron, se tomó la decisión de trabajar en predicciones mensuales y únicamente para el departamento de Montevideo, en regiones denominadas como segmentos que son definidos por el INE.

Se definieron dos líneas base para poder evaluar los resultados de la predicción de consumo contra métodos más simples. Los modelos de aprendizaje que se decidieron utilizar son: Regresión Lineal, *Random Forest* y KNN. Esta selección busca cubrir métodos de distinto tipo que hayan sido mencionados en la literatura, con la finalidad de compararlos entre si y ver cuál es más adecuado para esta tarea.

Se pudo observar que los mejores resultados para las predicciones se obtuvieron con *Random Forest*, con un MAPE de 5,6 %. Esta precisión es acorde a otros trabajos relacionados, estando a la par del error conseguido por Wang y cols. (2018) de 7,75 %, el rango de errores de 2,8 % a 10,0 % de Kaboli y cols. (2017) o el rango de errores de 3,8 % a 6,18 % de Oreshkin y cols. (2021). Cabe destacar que estos trabajos no resuelven exactamente el mismo problema, ya que no trabajan con el mismo conjunto de datos.

A su vez, se logra un MAPE 60 % mejor que predecir el consumo mensual como el promedio de los consumos previamente registrados. Esto evidencia la utilidad de incorporar los diferentes conjuntos de datos que se utilizaron.

El tiempo de ejecución del aprendizaje no es tan largo comparado con aprendizajes que utilizan otros modelos, por lo que es fácilmente reproducible y extensible en el futuro.

Se realizó una investigación sobre causalidad, presentando una breve introducción al tema que permite luego entender el ejemplo práctico basado en el consumo eléctrico. Los resultados sugieren, a partir del modelo de la realidad propuesto, que existe un efecto causal entre los ingresos de los hogares y el consumo eléctrico diario. Este efecto causal implica que, a más ingresos del hogar, más consumo se produce. De forma similar, la proporción de casas (en oposición a edificios) en un segmento parece ser un factor que causa mayor consumo eléctrico. Es decir, se obtuvo que las casas poseen un consumo mayor que los edificios.

Los estudios de la inferencia causal y principalmente en el análisis con variables objetivo y tratamiento continuas, son un campo con mucho margen de crecimiento. Como trabajo futuro se puede profundizar en este análisis y esperamos que esta prueba de concepto motive a la investigación en esta área ya que resulta de gran utilidad, no solo a empresas sino también a las distintas entidades gubernamentales para la optimización de los recursos.

Referencias

Carpinteiro, O., da Silva, A., y Feichas, C. (2000). A hierarchical neural model in short-term load forecasting. En *Proceedings of the ieee-inns-enns international joint conference on neural networks. ijcnn 2000. neural computing: New challenges and perspectives for the new millennium* (Vol. 6, p. 241-246 vol.6). doi: 10.1109/IJCNN.2000.859403

Chen, J.-F., Lo, S.-K., y Do, Q. H. (2017). Forecasting monthly electricity demands: An application of neural networks trained by heuristic algorithms. *Information*, *8*(1).

Dudek, G., y Pełka, P. (2021). Pattern similarity-based machine learning methods for mid-term load forecasting: A comparative study. *Applied Soft Computing*, *104*, 107223.

Gul, M., Urfa, G., Paul, A., Moon, J., Rho, S., y Hwang, E. (2021, octubre). Mid-term electricity load prediction using cnn and bi-lstm. *The Journal of Supercomputing*, *77*(10), 10942–10958.

Kaboli, S. H. A., Fallahpour, A., Selvaraj, J., y Rahim, N. (2017). Long-term electrical energy consumption formulating and forecasting via optimized gene expression programming. *Energy*, *126*, 144-164.

Khuntia, S. R., Rueda, J., y Meijden, M. (2018, 11). Long-term electricity load forecasting considering volatility using multiplicative error model. *Energies*, *11*, 3308.

Lara-Benítez, P., Carranza-García, M., Luna-Romera, J. M., y Riquelme, J. C. (2020). Temporal convolutional networks applied to energy-related time series forecasting. *Applied Sciences*, *10*(7).

Lei, L., Chen, W., Wu, B., Chen, C., y Liu, W. (2021). A building energy consumption prediction model based on rough set theory and deep learning algorithms. *Energy and Buildings*, *240*.

Ma, J., y Cheng, J. C. (2016). Identifying the influential features on the regional energy use intensity of residential buildings based on Random Forests. *Applied Energy*, *183*(C). doi: https://doi.org/10.1016/j.apenergy.2016.08.096

Ministerio de Industria, E. y. M. (2020a). *Consumo electricidad per cápita.* https://ben.miem.gub.uy/indicadores3.php. (Recuperado el 22 de mayo de 2022)

Ministerio de Industria, E. y. M. (2020b). *Informe energía eléctrica series estadísticas.* https://observatorio.miem.gub.uy/obs/sites/default/

files/documentos/informe_electrica_20201.pdf. (Recuperado el 22 de mayo de 2022)

Mohamed, Z., y Bodger, P. (2005). Forecasting electricity consumption in new zealand using economic and demographic variables. *Energy*, *30*(10), 1833-1843.

Mosavi, A., y Bahmani, A. (2019, 03). Energy consumption prediction using machine learning; a review.

Nassif, A. B., Soudan, B., Azzeh, M., Attilli, I., y AlMulla, O. (2022). *Artificial intelligence and statistical techniques in short-term load forecasting: A review.* arXiv.

Neal, B. N. (s.f.). *Introduction to causal inference from a machine learning perspective.*

Oreshkin, B. N., Dudek, G., Pełka, P., y Turkina, E. (2021). N-beats neural network for mid-term electricity load forecasting. *Applied Energy*, *293*, 116918.

Pearl, J. (2009). *Causality* (2.ª ed.). Cambridge, UK: Cambridge University Press.

Popoola, O. (2016). Modeling of residential lighting load profile using adaptive neuro fuzzy inference system (anfis). *International Journal of Green Energy*, *13*(14).

Porse, E., Derenski, J., Gustafson, H., Elizabeth, Z., y Pincetl, S. (2016). Structural, geographic, and social factors in urban building energy use: Analysis of aggregated account-level consumption data in a megacity. *Energy Policy*, *96*.

Rivera-González, L., Bolonio, D., Mazadiego, L. F., y Valencia-Chapi, R. (2019). Long-term electricity supply and demand forecast (2018–2040): A leap model application towards a sustainable power generation system in ecuador. *Sustainability*, *11*(19).

Shao, Z., Chao, F., Yang, S.-L., y Zhou, K.-L. (2017). A review of the decomposition methodology for extracting and identifying the fluctuation characteristics in electricity demand forecasting. *Renewable and Sustainable Energy Reviews*, *75*.

Wang, Z., Wang, Y., Zeng, R., Srinivasan, R. S., y Ahrentzen, S. (2018). Random forest based hourly building energy prediction. *Energy and Buildings*, *171*, 11-25.

Yu, Z., Haghighat, F., Fung, B. C., y Yoshino, H. (2010). A decision tree method for building energy demand modeling. *Energy and Buildings*, *42*(10).

Variables utilizadas de la ECH

La lista completa de las 38 variables utilizadas de la encuesta son las siguientes:

secc	Sección censal. Dato categórico.
segm	Segmento censal (junto con secc, identifican el segmento en Montevideo). Dato categórico.
pesoaño	Ponderador de cada entrada para cálculos anuales, indica la cantidad de personas que este dato representa estadísticamente. Dato numérico. Utilizado para calcular la variable "peso_año".
estred13	Estratos sociales que clasifican a los hogares en Montevideo con 5 niveles económicos desde Bajo a Alto. Dato categórico ordenado. Utilizado para calcular la variable "nivel_economico_hogar_prom".
c1	Tipo de vivienda, dato categórico que clasifica la vivienda si es una casa, un apartamento de distintos tipos o un local no construído para una vivienda. Utilizado para calcular la variable "porcentaje_casas".
c2	Material predominante en paredes externas. Dato categórico. Utilizado para calcular la variable "porcentaje_pared_ladrillo".
c3	Material predominante en techo. Dato categórico. Utilizado para calcular la variable "porcentaje_techo_hormigon".
c4	Material predominante en pisos. Dato categórico. Utilizado para calcular la variable "porcentaje_piso_con_terminacion".
c5_4	Puertas o ventanas en mal estado. Dato categórico, sí o no. Utilizado para calcular la variable "tiene_puerta/ventana_mal_estado_prom".
c5_5	Grietas en pisos. Dato categórico, sí o no. Utilizado para calcular la variable "tiene_grieta_piso_prom".
c5_7	Cielos rasos desprendidos. Dato categórico, sí o no. Utilizado para calcular la variable "tiene_cielo_raso_desprendido_prom".
c5_8	Poca luz solar. Dato categórico, sí o no. Utilizado para calcular la variable "tiene_poca_luz_solar_prom".
c5_9	Escasa ventilación. Dato categórico, sí o no. Utilizado para calcular la variable "tiene_escasa_ventilacion_prom".
d8_4	Vivienda ubicada en asentamiento irregular. Dato categórico. Utilizado para calcular la variable "en_asentamiento_irregular_prom ".
d9	Cantidad de habitaciones residenciales. Dato numérico. Utilizado para calcular la variable "cant_habitaciones_prom".
d10	Cantidad de habitaciones para dormir. Dato numérico. Utilizado para calcular la variable "cant_dormitorios_prom".
d14	Cantidad de baños. Dato numérico. Utilizado para calcular la variable "cant_baños_prom".
d20	Fuente de energía para cocinar, considerando energía eléctrica o varias alternativas. Dato categórico. Utilizado para calcular la variable "porcentaje_cocina_electrica".
d21_1	Posesión de calefón. Dato categórico, si o no. Utilizado para calcular la variable "tiene_calefon_prom".
d21_2	Calentador instantáneo de agua. Dato categórico, si o no. Utilizado para calcular la variable "tiene_calentador_instantaneo_prom".
d21_3	Refrigerador (con o sin freezer). Dato categórico, si o no. Utilizado para calcular la variable "tiene_refrigerador_prom".
d21_4_1	Cantidad de Televisores de tubo. Dato numérico. Utilizado para calcular la variable "cant_televisores_tubo_prom".
d21_5_1	Cantidad de Televisores de pantalla plana, tecnologías LCD, Plasma, o similar. Dato numérico. Utilizado para calcular la variable "cant_televisores_prom".
d21_9	Reproductor de DVD. Dato categórico, si o no. Utilizado para calcular la variable "tiene_dvd_prom".
d21_10	Lavarropa. Dato categórico, si o no. Utilizado para calcular la variable "tiene_lavarropa_prom".
d21_11	Secadora de ropa. Dato categórico, si o no. Utilizado para calcular la variable "tiene_secadora_prom".
d21_12	Lavavajilla. Dato categórico, si o no. Utilizado para calcular la variable "tiene_lavavajilla_prom".
d21_13	Horno microondas. Dato categórico, si o no. Utilizado para calcular la variable "tiene_microondas_prom".
d21_15_2	Cantidad de microcomputadores del Plan Ceibal. Dato numérico. Utilizado para calcular la variable "cant_equipos_ceibal_prom".
d21_15_4	Cantidad de microcomputadores que no son del Plan Ceibal. Dato numérico. Utilizado para calcular la variable "cant_equipos_no_ceibal_prom".
d21_16	Acceso a internet. Dato categórico, si o no. Utilizado para calcular la variable "tiene_internet_prom".
d23	Cantidad de personas de 14 o más años. Dato numérico. Utilizado para calcular la variable "cant_mayores_prom".
d24	Cantidad de personas menores de 14 años. Dato numérico. Utilizado para calcular la variable "cant_menores_prom".
YHOG	Ingresos del hogar no imputables a personas. Dato numérico. Utilizado para calcular la variable "ingresos_hogar_IP_prom".
ht11	Ingresos del hogar con valor locativo, sin servcio doméstico. Dato numérico. Utilizado para calcular la variable "ingresos_hogar_prom".
pobre06	Indicador de pobreza según metodología de 2006. Dato categórico, sí o no. Utilizado para calcular la variable "es_pobre_prom".
d260	Fuente energía para calefaccionar. Dato categórico. - desde ECH 2014 Utilizado para calcular la variable "porcentaje_calefaccion_electrica".
d21_14_1	Cantidad de aires acondicionados. Dato numérico - desde ECH 2014 Utilizado para calcular la variable "cant_ac_prom".

Tabla 1: Lista completa de las 38 variables utilizadas de la ECH en este trabajo

Bibliotecas y herramientas utilizadas

La solución y las figuras se realizaron utilizando el lenguaje de programación Python.

Para la elaboración de figuras con el mapa de montevideo se utilizó la biblioteca "geopandas" y para el resto se utilizó "matplotlib".

Los modelos se implementaron utilizando la bilbioteca "scikit-learn". Se eligió esta biblioteca porque es ampliamente utilizada, ya teníamos experiencia previa utilizándola, y permite implementar todos los modelos que deseamos probar.

El manejo e ingeniería de datos se realizaron utilizando "pandas". Esta biblioteca provee muchas utilidades al momento de trabajar con datos y es muy popular para estas tareas.

Como se menciona en el capítulo 6, los cálculos de causalidad se realizaron utilizando la biblioteca "DoWhy".

Cronograma del proyecto

En las figuras 1 y 2 se presenta el cronograma de las actividades realizadas durante el proyecto.

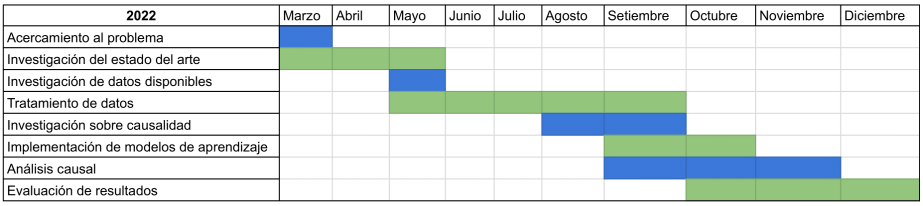

Figura 1: Cronograma 2022

Figura 2: Cronograma 2023