UNIVERSIDAD MICHOACANA DE SAN NICOLAS DE HIDALGO

Facultad de Ingeniería Eléctrica

Predicción por hora y región de la demanda eléctrica del Sistema Interconectado Nacional en México utilizando bosques aleatorios y comparación con regresión lineal

T E S I S

que para obtener el título de:

Ingeniero en Computación

presenta

Jehosafat Coria Barragán

Asesor

Dr. Jaime Cerda Jacobo

Estados Unidos Mexicanos

Morelia, Michoacán

Octubre, 2021

Dedicatoria

Dedico este proyecto a los futuros ingenieros mexicanos y a todos los profesionistas que crearon las bases para la creación de esta tesis.

A mi universidad, mi facultad, mis maestros y a todo el cuerpo académico por fungir durante mi desarrollo.

A todas aquellas personas que estuvieron en mi proceso de aprendizaje como ingeniero en computación.

Agradecimientos

Agradezco a mi asesor Dr. Jaime Cerda Jacobo por brindarme el apoyo durante todo el proceso de creación de este proyecto.

A mi madre Noemi Barragán Ortega por brindarme apoyo emocional y estabilidad en el transcurso de la elaboración de la tesis.

A mi hermana Jesse N. Coria Barragán por brindarme consejos y criticas constructivas desde el punto de redacción, ortografía, estilo, sintaxis, entre otros.

A mis sinodales por su apoyo y por enriquecer esta tesis.

A mi familia por apoyarme de manera económica y ayudándome en todo el tramite llevado a cabo.

A mis amigos por darme ánimos en los tiempos mas difíciles y por contar siempre con su apoyo.

Resumen

La Inteligencia Artificial tiene un campo de trabajo vasto y extenso. Se utiliza principalmente en las ramas científicas y de humanidades. Las redes neuronales son uno de los modelos de predicción mas utilizado en la actualidad, esto se debe a su adaptabilidad, escalabilidad y flexibilidad, ya que se pueden adecuar a cualquier problemática, pero han surgido nuevos modelos de predicción entre ellos el modelo de aprendizaje automatizado: Bosques Aleatorios (*Random Forest*), el cual ha tenido un auge bastante grande, ya que si este es armado y entrenado de forma correcta puede tener el mismo potencial que una red neuronal obteniendo un margen de error relativamente pequeño con un tiempo de entrenamiento rápido.

El tema de la tesis actual se eligió para predecir la demanda que se necesitará para dar abasto a todo el Sistema Interconectado Nacional (SIN) y sus respectivas 7 regiones a futuro. Para esto, se toman datos de periodos anteriores brindados por el Centro Nacional de Control de Energía (CENACE), se analizarán dichos datos y se utilizan para el armado y el entrenamiento del modelo de predicción de bosques aleatorios, para que pueda inferir de manera exitosa y tenga una salida con margen de error mínimo.

Los datos utilizados para la ejecución del modelo de predicción de bosques aleatorios están comprendidos en el periodo de tiempo del: 01 de enero de 2020 al 31 de diciembre de 2020, la muestra esta dada por región y hora, se utilizan el 75 % de los datos para el entrenamiento y el resto para la predicción, se analiza y compara el margen de error obtenido utilizando las métricas: error absoluto medio porcentual (MAPE) y distancia media cuadrática mínima (RMSE).

Palabras clave: *Inteligencia Artificial, Bosques Aleatorios, Predicción, Energía Eléctrica, México.*

Abstract

Artificial Intelligence has a wide work field. It is used mainly in the scientific and humanities branches. Neuronal networks are one of the prediction models most used nowadays, because of its adaptable and flexible nature. They are adequate to almost any problematic. Nevertheless, new prediction models have arisen. The automatized learning model called Random Forests is one of them, and it has reached great popularity. This is because, when armed and trained in a correct way, it can have the same potential as a neuronal network, getting an error range that is smaller, in a fast-training time.

The topic of the thesis at hand, has been chosen to predict the demand that will be needed to result sufficient to all de Interconnected National System (Sistema Interconectado Nacional, SIN) and its respective seven regions in the future. For this, data from previous periods given by the National Energy Control Center (Centro Nacional de Control de Energía, CENACE), will be analyzed and used for the manufacturing and training of the random forests model of prediction, so it can successfully conclude and have a way out with a minimal error range.

The data used for the execution of the random forests model of prediction is understood within the time period of: January 1st, 2020 to December 31st, 2020. The sample is given by the region and time. 75 % of the data are used for the training, and the rest of them are used for the prediction. The error range obtained by using the metrics Mean absolute percentage error (MAPE) and Root mean square error (RMSE) is analyzed and compared to this data.

Índice general

Índice de figuras

Índice de tablas

Siglas y Acrónimos

CENACE	Centro Nacional de Control de Energía
SEN	Sistema Eléctrico Nacional
SIN	Sistema Interconectado Nacional
IA	Inteligencia Artificial
MEM	Mercado Eléctrico Mayorista
RNT	Red Nacional de Transmisión
RGD	Redes Generales de Distribución
CV	Coeficiente de Variación
RMSE	Distancia media cuadrática mínima
MAPE	Error absoluto medio porcentual

Capítulo 1

Introducción a la problemática

1.1. Introducción

En la rama de la ingeniería, una ciencia de la computación muy utilizada para la predicción de datos es la inteligencia artificial (IA). Esta rama en la actualidad es utilizada de manera común en cualquier ámbito o ciencia para la predicción de datos, esto gracias a ingenieros e investigadores que fundaron la base y siguen aportando a esta rama aún hoy en día. El principal objetivo de la rama de la IA es obtener inferencias con un error mínimo y con algoritmos simples.

Dentro del campo de la inteligencia artificial existen varios tipos de modelos, pero el más utilizado actualmente son los modelos de predicción, los cuales tienen utilidad dentro de cualquier rama de la ciencia o humanidades. En la construcción y entrenamiento de los modelos de predicción se extrae información útil dentro de un conjunto de datos, para posteriormente producir inferencias; el objetivo es obtener el menor error posible entre los datos estimados y los datos reales, ya que se está hablando de una predicción del futuro y cualquier variable externa o aleatoria puede llegar a afectar el resultado final. Dado que esa variable no se toma en cuenta por el modelo de predicción al momento de hacer el aprendizaje, puede ocasionar que el margen de error aumente. Por esta y otras razones, es que el error de la inferencia nunca resultará en cero; se puede aproximar lo más posible, pero nunca llegará a cero.

Para el entrenamiento del modelo a utilizar en esta tesis se obtuvo información

dentro de la base de datos del Centro Nacional de Control de Energía (CENACE) del Sistema Eléctrico Nacional (SEN). Los datos a utilizar dentro de este proyecto son la estimación de la demanda real que abarca del periodo del 01 de enero del 2020 al 31 de diciembre del 2020, contando con datos de 366 días para el análisis y el aprendizaje del modelo de predicción. Cabe señalar que los datos que brinda CENACE están comprendidos por cada hora en cada zona, si los datos estuvieran comprendidos por unidades de tiempo más pequeña, el modelo de predicción podría obtener un mejor resultado, pero la cantidad de datos sería demasiado grande y el modelo tardaría más en el procesamiento de los datos.

Los datos se analizan y se grafican para entender el comportamiento y tendencia de dichos datos, principalmente para identificar si la variable se comporta de manera lineal o no. En caso de no ser lineal, se debe saber hasta que punto de no linealidad puede llegar la variable, haciendo un análisis de la misma; esto se profundiza más en el capítulo 4, el objetivo es saber si la variable no tiene un comportamiento caótico y puede linealizarse. Se debe observar si existe un patrón o una tendencia, esto se puede notar más si se separan los datos en periodos de tiempo como cuartos de año (3 meses), meses o semanas. Se debe de analizar que factores pueden afectar la demanda eléctrica dentro del país y si esta se comporta como un ciclo estacionario.

Una vez hecho el análisis de los datos se construye un modelo de predicción y se entrena, el cual toma una parte de los datos como aprendizaje para entonces predecir la demanda eléctrica en periodos posteriores. Se utilizan bosques aleatorios como el modelo de predicción de este proyecto. Se comparan datos obtenidos del modelo de predicción con los datos reales para poder obtener un margen de error, el cual se observa y se analiza.

1.2. Antecedentes

Se utiliza el trabajo de aprendizaje automatizado: *"Pronóstico diario de la demanda de electricidad con aprendizaje automático"*(Romero, 2019), como el antecedente inmediatamente próximo de esta tesis la investigación del español: Manuel A. Romero Gracia, dado que dicho trabajo definió las bases para este proyecto. Realiza una predicción de la electricidad consumida de España, el rango de tiempo de su trabajo comprende el periodo del

01 de enero de 2014 al 31 de diciembre del 2019, utilizando solamente un dato por día de toda España.

Se utiliza su trabajo como base a este proyecto pero haciendo ciertas modificaciones para su adaptación a México y al SIN (Sistema Interconectado Nacional). En el trabajo de la presente tesis se observa como se comportan los datos de un solo año (01 de enero al 31 de diciembre del 2020), dado que por la pandemia se espera que sea un año atípico y se desea observar como esto afecta a la demanda en cada una de las regiones de México. También se tendrá mucha más información dado que se tiene una muestra de la demanda eléctrica por cada hora en cada región de México, mientras que en la investigación de Manuel A. Romero Gracia se toma un dato por día para toda España y cabe señalar que la cantidad de datos a utilizar influye al tiempo de entrenamiento de un modelo de predicción.

El modelo de predicción de bosques aleatorios creado por Leo Breiman y Adele Cutler, es un algoritmo de aprendizaje automatizado capaz de predecir valores con un margen de error considerable, es conocido por su poder de predicción bastante acertado y que no es tardado en la etapa de entrenamiento; por esta razón se eligió este modelo de predicción para utilizar en este proyecto.

Los bosques aleatorios es un método de conjunto (agrupa múltiples predictores de árboles de decisión) que fue desarrollado por Leo Breiman en 2001. Breiman afirma que: "el error de generalización de un bosque de clasificadores de árboles depende de la fuerza de los árboles individuales en el bosque y la correlación entre ellos" (Kurtis, 2020).

1.3. Objetivo General

El objetivo general de este proyecto es predecir la demanda real del Sistema Eléctrico Nacional (SIN) con el uso de aprendizaje automático, utilizando como entrenamiento los datos analizados reales proporcionados por el CENACE y codificación con el lenguaje de programación Python.

1.4. Objetivos Particulares

1.- Extraer los datos desde múltiples archivos csv y crear un marco de datos en Python para depurar y clasificar los datos obtenidos en región, fecha y hora.

2.- Analizar y graficar las curvas de demanda del Sistema Interconectado Nacional y sus 7 regiones brindados por CENACE, observar la distribución normal de los mismos y posibles variables que se involucren en el resultado.

3.- Crear y entrenar un modelo de predicción con los datos analizados de la demanda real brindados por la base de datos del CENACE.

4.- Hacer pruebas con el modelo de predicción entrenado y analizar la salida, observar el margen de error obtenido mediante la comparación de los datos proporcionados por el modelo de predicción y la demanda real.

1.5. Metodología

El lenguaje de programación que se utiliza en este proyecto es python, dado que cuenta con un amplio catalogo de librerías de aprendizaje automático, para la creación de modelos de predicción y por ser un lenguaje de programación muy utilizado en la actualidad. Esto se debe a que python tiene librerías que al igual que el lenguaje, son de código abierto, y estas son muy útiles para el desarrollo de cualquier modelo de predicción.

Se utiliza el ambiente de programación Anaconda, dado que tiene un conjunto de herramientas de python para el sistema operativo Windows como pycharm, spyder, jupyterlab, jupyternotebook, entre otras. También vienen instaladas librerías por defecto como: numpy, pandas, sklearn, scipy, matplotlib, entre otras. Dentro de Anaconda la principal herramienta que se utiliza en este proyecto es Jupyter Notebook, esto por que se desea tener un código con un diseño secuencial y explicativo, dado que se puede utilizar como una bitácora y se puede separar el código por bloques para su mayor entendimiento por partes; algo muy útil en proyectos extensos.

Las librerías que se utilizan para el desarrollo de este proyecto son las siguientes:

1.- pandas : útil para la lectura y clasificación de archivos csv [1].

2.- pyplot y seaborn : para la visualización de datos contenidos en listas mediante diferentes tipos de gráficos [2].

3.- numpy : librería matemática/algebraica la cual tiene soporte para la creación de vectores y matrices, también tiene varios métodos para operaciones validas entre vectores y matrices [3].

4.- datatime : librería para la manipulación de variables de tiempo de tipo fecha y hora, también contiene métodos de operaciones aritméticas validas entre estos tipos de variables [4].

5.- scipy : Contiene una gran cantidad de distribuciones de probabilidad, también contiene un vasto diccionario de funciones y variables estadísticas [5].

6.- sklearn : librería de aprendizaje automatizado, con varios algoritmos de clasificación, regresión y análisis de grupos entre los más comunes [6].

7.- pickle : Contiene una gran cantidad de distribuciones de probabilidad, a parte de contener un vasto diccionario de funciones y variables estadísticas [7].

1.6. Justificación

El proyecto que se lleva acabo en esta tesis realiza algo útil para la sociedad mexicana, al predecir la demanda eléctrica que se utilizara en el futuro en México. Obteniendo una inferencia acertada, se podría ahorrar energía en todas las estaciones eléctricas y solo generar la cantidad necesaria. Con este beneficio principal se pueden desglosar varios beneficios indirectamente, como lo es demandar más a las estaciones eléctricas que obtienen la energía

[1] documentación obtenida de: https://pandas.pydata.org/
[2] documentación obtenida de: https://matplotlib.org/stable/tutorials/introductory/pyplot.html
[3] documentación obtenida de: https://numpy.org/
[4] documentación obtenida de: https://docs.python.org/3/library/datetime.html
[5] documentación obtenida de: https://docs.scipy.org/doc/scipy/reference/
[6] documentación obtenida de: https://scikit-learn.org/stable/
[7] documentación obtenida de: https://docs.python.org/3/library/pickle.html

eléctrica de fuentes renovables y reducir la generación de electricidad a las que no. Esto causará la reducción de contaminación en el medio ambiente. Otro beneficio indirecto es que al producir la energía eléctrica necesaria (almacenar un poco en caso de emergencia) no se desperdiciaría tanta energía.

Este proyecto también pretende ser útil como base para proyectos futuros, dado que ya teniendo el marco de datos de la demanda eléctrica de México depurado y seccionado, se le puede aplicar otro modelo de predicción y compararlo con el utilizado en este proyecto, o también usar los datos obtenidos del modelo de predicción utilizando en este proyecto y minimizar el margen de error obtenido con algoritmos dedicados a eso.

1.7. Descripción del Contenido de los Capítulos

En el capítulo 1 se da una breve introducción a este trabajo, donde se muestra la importancia de la IA hoy en día, y como es que es una herramienta que esta lejos a llegar a su fin, dado que cada día se ha mejorado un poco mas, teniendo múltiples áreas de aplicación, en este proyecto esta tendrá la función de la predicción de la demanda eléctrica de México.

En el capítulo 2 se presentan conceptos básicos del funcionamiento del Sistema Eléctrico Nacional (SEN) en México, el diseño y modo de obtención de información de la demanda eléctrica en México, el SIN y sus respectivas regiones, cuáles regiones sí se toman en cuenta, cuales no y por que, también se explica el proceso de obtención de datos de un respectivo periodo de tiempo desde el sitio oficial del CENACE.

En el capítulo 3 se empieza con la obtención de datos desde los archivos csv a un marco de datos en python, una vez obtenidos los datos en crudo estos se deben depuran y ordenar, después los datos se clasifican dependiendo de su región y hora.

En el capítulo 4 se tendrá el marco de datos útil, depurado y clasificado (obtenido en el capítulo anterior), con este se procederá a hacer el análisis y graficado de los datos, observando información importante como: tendencia, distribución, desviación, oblicuidad, media, normalidad, heterocedasticidad, volatilidad, entre otras.

En el capítulo 5 se inicia con el proceso de creación del modelo de predicción, se hace el diseño y posteriormente la construcción del modelo de predicción de bosques aleatorios

utilizando ingeniería de características, una vez armado se proporcionan el 75 % de los datos como entrenamiento de dicho modelo de predicción.

En el capítulo 6 se pone a prueba el modelo de predicción creado y entrenada en el capítulo anterior, se predicen el 25 % de datos faltantes y se comparan con los datos reales de la demanda eléctrica respectiva, de esta manera se obtiene un margen de error, dicho error se calcula con las métricas MAPE y RMSE (véase en el subcapítulo 2.4), también se diseña un modelo en multi-periodo utilizando múltiple regresión lineal para intentar reducir el margen de error obtenido.

En el capítulo 7 se analiza más a fondo la normalidad de los datos y el error obtenido del modelo de predicción, dicho error se compara con el error obtenido con el modelo de regresión lineal y con el modelo de bosques aleatorios con la función de múltiple regresión lineal, se redactan las conclusiones del trabajo con dichos datos obtenidos.

Capítulo 2

Estado Previo

En este apartado se profundizara un poco en los conceptos previos a este trabajo, los cuales son utilizados para el desarrollo del mismo, esto con el objetivo de adquirir un mejor entendimiento del funcionamiento de las métricas utilizadas para el análisis de los datos y del modelo de predicción en si.

2.1. Análisis de datos variables en el tiempo

Para el análisis de datos se opta por utilizar la librería pandas ya que proporciona una forma estructurada de utilizar datos, esto mediante un objeto llamado marco de datos (*DataFrame*). Es un arreglo bidimensional, pero cada columna puede ser de cualquier tipo de dato, aparte de esto, pandas proporciona diferentes funciones para el cálculo de sus propios marcos de datos.

"Cada eje de una estructura de datos de pandas tiene un objeto de índice que almacena información de etiquetado sobre cada marca a lo largo de ese eje" (McKinney y cols., 2011).

	state	color	food	age	height	score
Jane	NY	blue	Steak	30	165	4.6
Niko	TX	green	Lamb	2	70	8.3
Aaron	FL	red	Mango	12	120	9.0
Penelope	AL	white	Apple	4	80	3.3
Dean	AK	gray	Cheese	32	180	1.8
Christina	TX	black	Melon	33	172	9.5
Cornelia	TX	red	Beans	69	150	2.2

Figura 2.1: Marco de datos en python[1].

La figura 2.1 muestra un ejemplo de un marco de datos en python, el cual tiene las columnas: estado, color, comida, edad, altura y puntaje, se tienen el registro de 7 personas, y se puede notar como cada columna puede ser del tipo de variable que se desee, a diferencia de los arreglos de 2 dimensiones usuales los cuales solo pueden tener un único tipo de variable.

"La estructura de datos también contiene ejes etiquetados (filas y columnas). Las operaciones aritméticas se alinean en las etiquetas de fila y columna. Se puede considerar como un contenedor similar a un diccionario para los objetos de la serie. La estructura de datos primaria de los pandas" (Pandas Development Team, 2021).

La librería seaborn nos ofrece la gráfica de tipo caja que sirve para conocer la distribución de los datos de forma visual y atractiva, mostrando por orden los valores: mínimo, primer cuartil, segundo cuartil o mediana, el tercer cuartil y el máximo, también dibuja un cubo en el rango intercuartil (entre el primer y el tercer cuartil) y los datos fuera de rango.

[1] obtenido de: https://medium.com/dunder-data/selecting-subsets-of-data-in-pandas-6fcd0170be9c

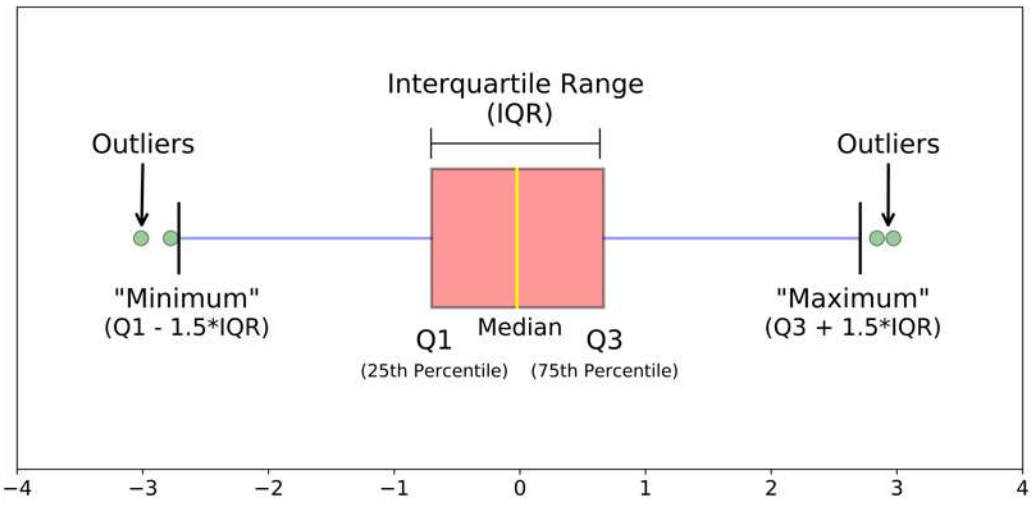

Figura 2.2: Gráfica de tipo caja[1].

En la figura 2.2 se puede ver la representación de un gráfico de tipo caja y todos sus componentes.

Se utilizan medias móviles en python. Las medias móviles es una manera de suavizar los datos de tal manera que para datos muy variantes se obtenga un curva más suave, estos datos nunca deberán tomarse como si fueran los datos reales, solamente para fines experimentales y para ver su comportamiento. Para obtener un dato de una media móvil primero se debe de definir el tamaño de la ventana (cantidad de datos a utilizar para el cálculo). Se debe de definir la orientación de la media móvil, esta es la cantidad de datos anteriores y posteriores que se toman en cuenta para el cálculo (generalmente se toman solamente los datos anteriores). Después se hace un promedio con dichos datos y como resultado es el valor de la media móvil de dicho dato.

[01] obtenido de: https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

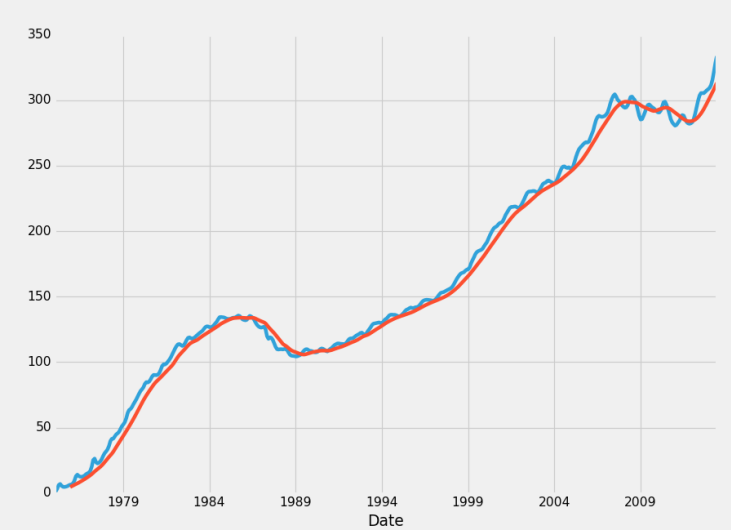

Figura 2.3: Ventana Deslizante por promedios[1].

La figura 2.3 es una gráfica que compara los datos reales (línea azul) y la media móvil (línea roja). Se puede observar como la media móvil empieza después de la curva real y como esta curva tiene un trazado muy suave, como si esta no tuviera ruido.

La prueba de *Shapiro-Wilk* sirve para saber si un conjunto de datos pertenecen a una distribución normal o no, es una de las pruebas más potentes para saber la normalidad de un conjunto de datos y es el que recomienda python para su uso en los marcos de datos.

"El test de Shapiro-Wilk plantea la hipótesis nula que una muestra proviene de una distribución normal. Elegimos un nivel de significanza p, por ejemplo 0,05, y tenemos una hipótesis alternativa que sostiene que la distribución no es normal" (Pedrosa, 2017).

La heterocedasticidad sirve para ver la dispersión de los datos fuera de la media, también para observar si los datos positivos pueden llegarse a anular con los negativos (tomando la media como la referencia). Esta es una característica de las distribuciones normales y deber ser cumplida para considerarse una distribución normal.

"La heterocedasticidad ocurre, en estadística, cuando los errores no son constantes a lo largo de toda la muestra. El término es contrario a homocedasticidad" (Pedrosa, 2017).

[1] obtenido de: https://pythonprogramming.net/rolling-statistics-data-analysis-python-pandas-tutorial/

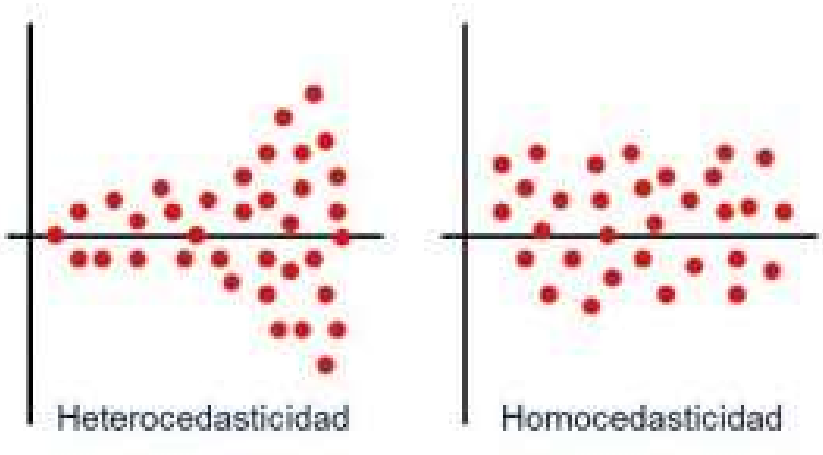

Figura 2.4: Ejemplos de diferencia entre datos con heterocedasticidad y con homocedasticidad[1].

En la figura 2.4 se puede observar ejemplos para cada caso, cuando los datos son heterocedasticos los errores no siguen ningún patrón y es difícil cancelarlos entre si, en cambio cuando los datos son homocedasticos tienen una diferencia con la media más estable.
"En los modelos de regresión lineales se dice que hay heterocedasticidad cuando la varianza de los errores no es igual en todas las observaciones realizadas" (Pedrosa, 2017).

Los cuartiles es una medida estadística utilizada para dividir la muestra de datos en cuatro partes iguales, por lo que cada parte tendría el 25 % de los datos, y el primer cuartil sería el dato más próximo al final de cada parte de datos.

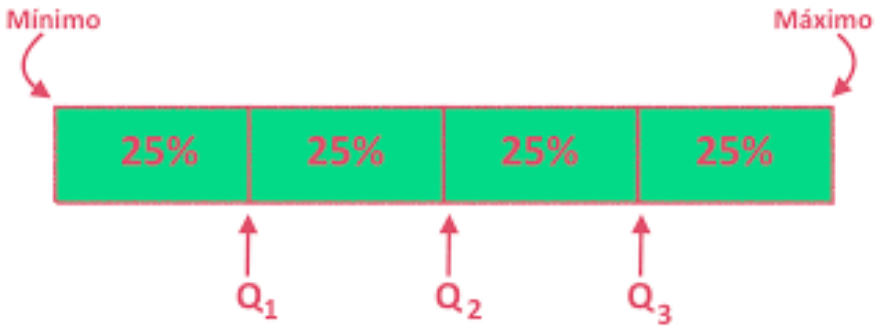

Figura 2.5: Cuartiles[1].

[01] obtenido de: https://economipedia.com/definiciones/homocedasticidad.html

Los percentiles son una medida estadística utilizada para la comparación de datos, consiste en un valor del 0 al 100 que indica el porcentaje de datos que son iguales o menores a dicho valor.

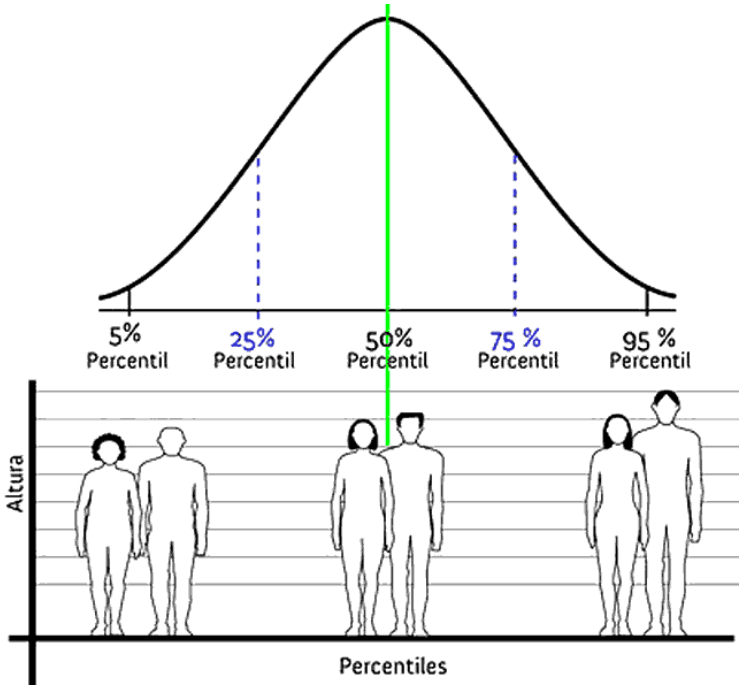

Figura 2.6: Percentiles[2].

En la figura 2.6 se puede ver un ejemplo de los percentiles tomados de la altura de un grupo de personas, estos percentiles se pueden visualizar en la distribución de los datos, si el dato 0 % - 50 % es igual a su contra-parte del 100 % - 50 % respectivamente, significa que los datos tienen una distribución normal.

2.2. Modelo de Predicción

Para el modelo de predicción utilizado en la tesis actual se explican primero algunos conceptos básicos, los cuales son fundamentales para entender el funcionamiento y la construcción del modelo en sí.

[01] obtenido de: https://estadisticalidia.com/tema-2-parte-2-medidas-de-posicion/
[02] obtenido de: https://curiosoando.com/que-son-los-percentiles

2.2.1. Ingeniería de Características

"La ingeniería de características es una tarea central en la preparación de datos para aprendizaje automático. Es la práctica de construir adecuadas características de datos recibidos, que conducen a una mejora predictiva actuación" (Fatemeh Nargesian y cols., 2017); La ingeniería de características se refiere al proceso de extraer información útil o características de los datos ya existentes.

La ingeniería de características es considerado un algoritmo de aprendizaje automático, donde los datos de entrada son procesados y los de salida son generados, entre los datos de entrada se encuentran las características, las cuales deben ser seccionadas. Se le deben dar a los algoritmos rasgos o características para que puedan funcionar. Mediante la ingeniería de características se mejora la productividad y se dispone esos datos de entrada que el algoritmo necesita. La ingeniería de características es importante dado que se requieren resultados precisos. Mientras mejores sean las características, mejores serán los resultados. Aunque se trate de un modelo equivocado, podrá lograrse un buen rendimiento, esto indica la flexibilidad que tiene y se logrará incluso con modelos menos complejos.

Ahora bien, la selección de características es decisiva al momento de crear un modelo adecuado. En primer lugar, existe determinada reducción de la cantidad de características considerados en la creación de un modelo. Los datos incluyen más información de la necesaria en la mayoría de las ocasiones y estos pueden entorpecer el armado del proyecto, por lo que puede haber características con pocos datos, o bien, características duplicadas. En ambos casos, agregarlas de esa manera perjudicaría al modelo. Por eso, la selección de características mejora su calidad y también da al proceso de armado del modelo más eficacia. Sin características sobrantes, el uso de recursos se disminuye y esto impacta al tiempo de armado y predicción del modelo, por esta razón se deben identificar las mejores características las cuales son las que tienen patrones significativos.

Este proceso puede llevarse a cabo, ya sea por un analista, una herramienta de modelado, o un algoritmo; que escogerá atributos de acuerdo con su valor. Es importante identificar ambos, cuando existan datos inútiles y cuando exista una poca cantidad de datos útiles, en el caso de este proyecto se utiliza el coeficiente de Pearson (explicado más

adelante) como el encargado en tomar la decisión de que características tomar y cuales no.

2.2.2. Medidas de dispersión y coeficiente de correlación de Pearson

"Las medidas de dispersión tratan, a través del cálculo de diferentes fórmulas, de arrojar un valor numérico que ofrezca información sobre el grado de variabilidad de una variable" (Francisco, 2019).

Esto quiere decir que mediante estas técnicas se sabe qué tanto actúa una variante con relación a otra, con base a estos datos se puede comparar y decidir.

Algunas de las medidas de dispersión son el rango, la varianza, la desviación estándar y el coeficiente de variación. Ahora, se habla más acerca del coeficiente de correlación de Pearson. Mediante éste se puede cuantificar la intensidad de la relación lineal entre dos variables en un análisis de correlación.

El coeficiente de correlación compara la distancia de cada dato puntual respecto a la medida de la variable. Si trazamos una línea entre los datos, el resultado de la fórmula nos dirá que tanto se ajustan a ella las variables. Sin embargo, se trata sobre las relaciones lineales. Las relaciones en forma curvilínea no podrán ser reveladas satisfactoriamente con el coeficiente de Pearson.

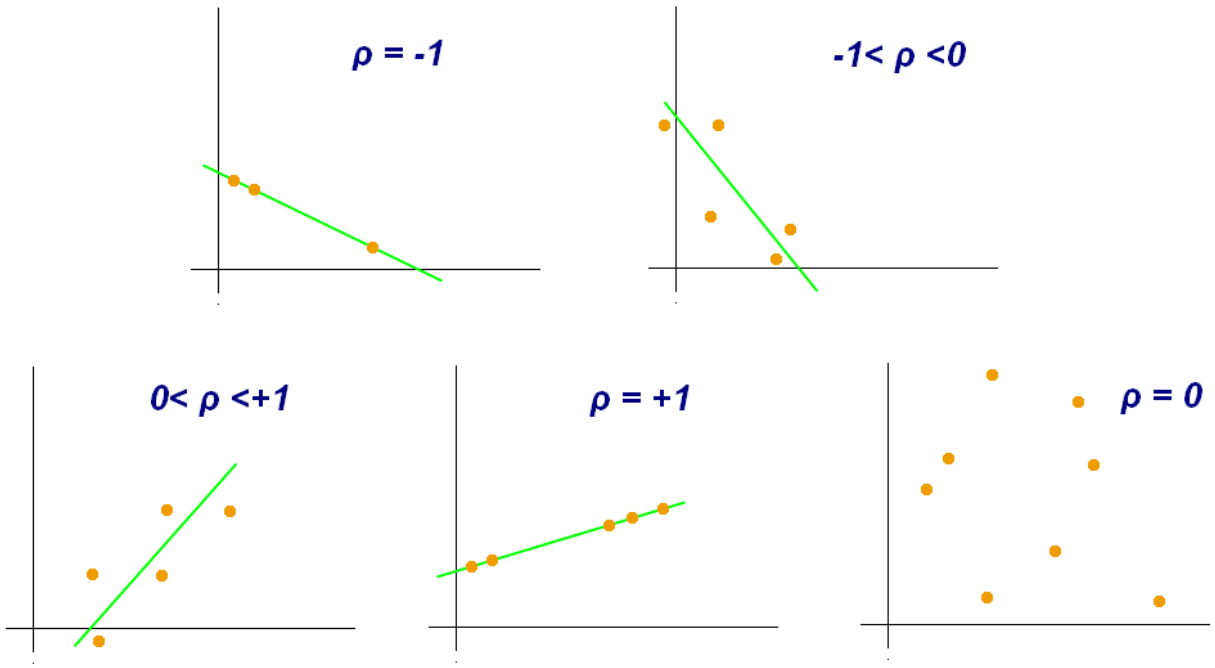

Figura 2.7: Casos del coeficiente de correlación de Pearson[1].

En la figura 2.7 se muestran los casos en los que puede caer el valor del coeficiente de Pearson, y estos casos son los siguientes: si p se aproxima a cero, la relación lineal es poca. Si hay valores de p positivos, la correlación también lo será; y viceversa en el caso negativo. En cambio, si encontramos valores 1 y -1, tendríamos la correlación "perfecta", ya que todos los datos logran asociarse a esta línea recta. El valor p nos ayudará a saber si el coeficiente de correlación es diferente a cero.

2.2.3. Árboles de decisión Y Bosques Aleatorios

Un árbol de decisiones es un esquema que representa las opciones disponibles. Es prácticamente un diagrama de flujo, que representa de forma visual cada una de las decisiones y resultados probables. En el árbol de decisiones se descomponen los datos en subconjuntos cada vez más pequeños, generalmente de valores diferentes. La toma de decisiones se vuelve más sencilla mediante éste, ya que diferentes factores pueden ser tomados en cuenta al

[01] obtenido de: https://es.wikipedia.org/wiki/Coeficiente_de_correlaci%C3%B3n_de_Pearson

mismo tiempo.

Cada uno de sus elementos tienen un significado: los nodos internos significan las opciones posibles. Las ramas serían los resultados de cada prueba y los nodos de hoja son el resultado obtenido. Los caminos que se forman desde la raíz hasta la hoja serían las reglas de clasificación.

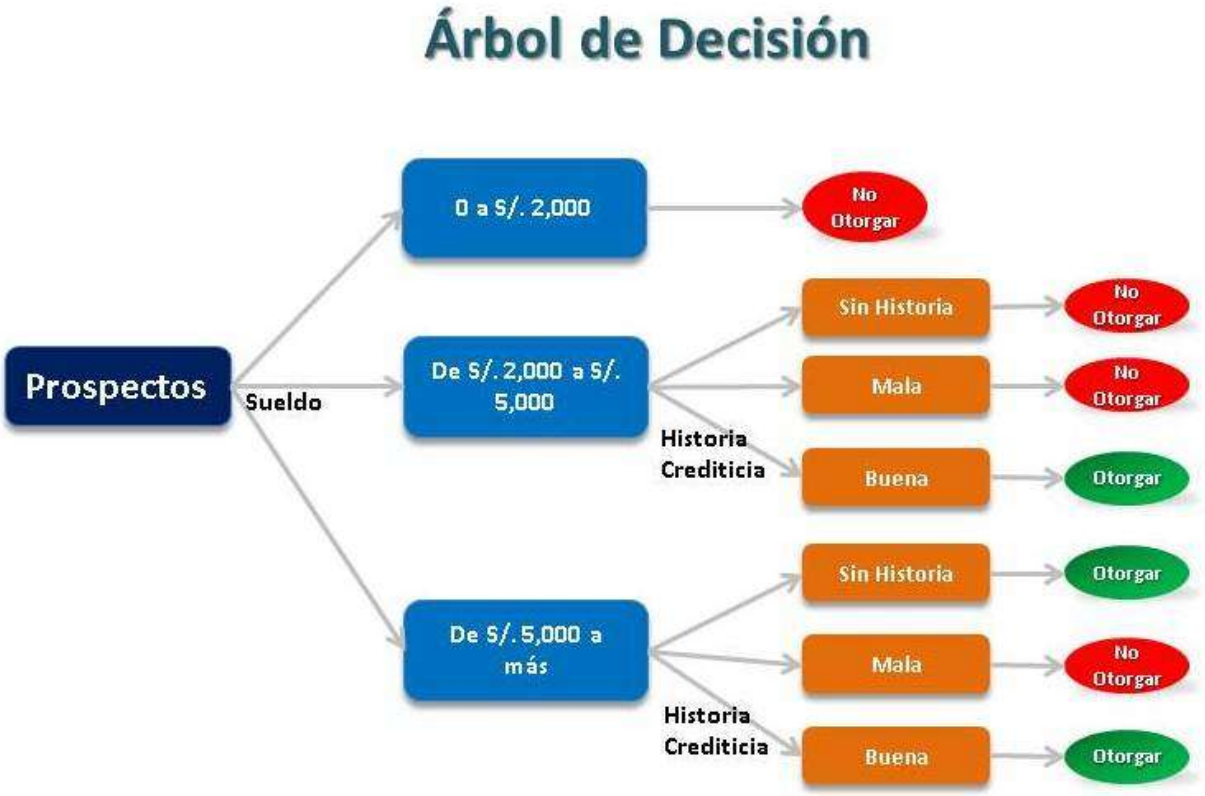

Figura 2.8: Ejemplo sencillo de un árbol de decisión[1].

En la figura 2.8 se muestra un ejemplo sencillo de un árbol de decisión donde se mueve desde el nodo raíz hasta uno de los nodos hojas, pasando por nodos intermedios de decisión en el caso de la figura son decisiones para saber si el banco debería otorgarle a un prospecto una tarjeta de crédito o no.

Generando un conjunto de árboles de decisión es como se crea un bosque aleatorio.

[01] obtenido de: http://herramientas-para-la-toma-de-decisiones.over-blog.com/2018/07/arbol-de-decisiones-1.html

El bosque aleatorio es un algoritmo que busca predecir el valor de una variable, utilizando reglas de decisión simples partiendo de las características de los datos. Es una manera de fortificar al algoritmo que representa un árbol de decisión. Se crean muchos modelos de éste mismo para obtener el resultado de sus predicciones por mayoría de votos. Así se obtiene un resultado mucho más preciso. Los árboles de decisión que conforman al bosque se generan tomando los datos disponibles al azar. El proceso se repite varias veces para obtener los árboles del bosque, todos serán diferentes.

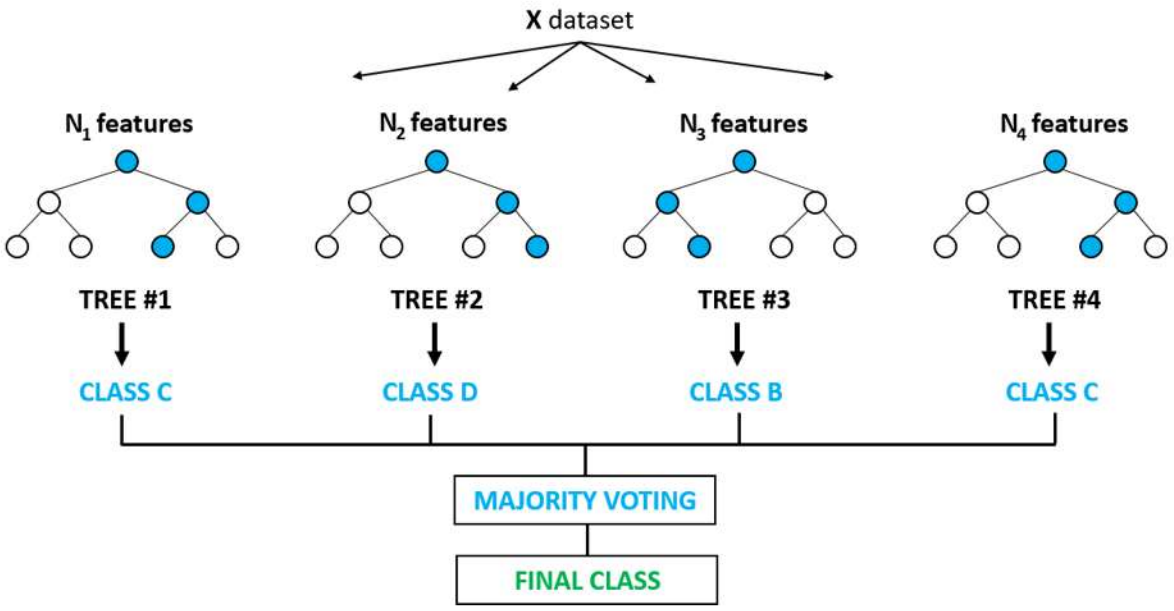

Figura 2.9: Diagrama del diseño de un bosque aleatorio sencillo[1].

En la figura 2.9 se muestra el diagrama de un bosque aleatorio, el cual puede verse que esta compuesto por varios arboles de decisión, generados de manera aleatoria, al final cada árbol obtiene un resultado, los resultados obtenidos se ponen a votación y se decide un resultado final.

[01] obtenido de: https://rpubs.com/Avalos42/randomforest

2.3. Medidas de error utilizados

Para medir la certeza del modelo y para su comparación con la regresión lineal se utilizaron dos variables para medir el error MAPE (error absoluto medio porcentual) y RMSE (distancia media cuadrática mínima). Ambas métricas sirven para la evaluación de la regresión de un modelo y son utilizadas para medir la diferencia entre dos conjuntos de datos diferentes.

$$MAPE = \frac{\frac{\sum_{i=1}^{n} 100|Real_i - Pronóstico_i|}{Real_i}}{n} \tag{2.1}$$

En la ecuación 2.1 se muestra la forma de calcular MAPE, esta trabaja con base en un conjunto de predicciones para medir la magnitud promedio de sus errores. Esta métrica no toma a consideración la dirección de las predicciones. El promedio se toma de una muestra de prueba, tomada de las diferencias absolutas entre la predicción y el valor real. En este caso, todas las diferencias individuales tienen el mismo peso; es una muy buena métrica para medir la exactitud de las predicciones.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \tag{2.2}$$

En la ecuación 2.2 se muestra la forma de calcular RMSE, esta funciona como un precepto de puntuación cuadrática, que también mide la magnitud promediada del error. En este caso se calcula obteniendo la raíz cuadrada del promedio de las diferencias al cuadrado de la predicción y el dato real.

RMSE es una métrica bastante útil. Hay que considerar que, como los errores se elevan al cuadrado antes de promediarlos, esta métrica les concederá un peso alto a los errores grandes. Si se quieren evitar los errores pequeños, RMSE sirve mejor.

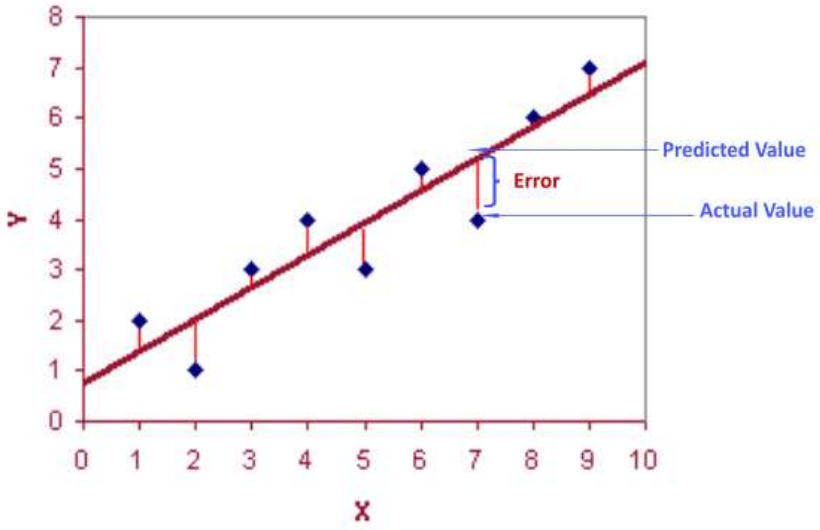

Figura 2.10: Ejemplo del error de la distancia media cuadrática mínima[1].

En la figura 2.10 se puede apreciar el valor de distancia de solamente un margen de error de la predicción de los datos, donde este valor esta alejado del dato real, este cálculo se debe hacer para cada uno de los datos, si un dato es igual a su dato real este tendrá una distancia de 0, por lo que no influirá al margen de error, idealmente un modelo de predicción perfecto debería tener las distancias de todos los datos previstos y de los datos reales en 0, pero esto es muy difícil de conseguir dado que los datos tomados de un sistema real no son lineales.

2.4. Comentarios finales

Se obtuvieron los conceptos previos más importantes de lo que se utilizan en los próximos capítulos, empezando con información acerca del manejo de electricidad en México, continuando con variables y métodos para el análisis de los datos y finalizando con conocimientos del modelo de predicción que se utilizan y las métricas para evaluar el error obtenido, estos son los conocimientos teóricos principales, los cuales pueden considerarse como la base para tener un mejor entendimiento en los próximos capítulos, no se profun-

[01] obtenido de: https://medium.com/@mygreatlearning/rmse-what-does-it-mean-2d446c0b1d0e

dizo tanto dado que al ser un trabajo practico solo es necesario conocer lo indispensable respecto a la teoría.

Capítulo 3

Obtención y creación de un marco de datos

En este capítulo se explica todo el proceso que se realiza para tener un marco de datos funcional, divididos en los siguientes pasos: obtener los datos reales del portal de CENACE, una vez obtenidos se depuran para conseguir los datos funcionales para este proyecto, en seguida se crea un marco de datos en python y se clasifican los datos dependiendo de la región y hora del dato. En pocas palabras se transformaran los datos a una estructura útil para el lenguaje de programación, seccionándolo y eliminando datos, de tal manera que solo se tenga la información útil en un formato eficiente para su uso en el análisis y el modelo de predicción.

3.1. Obtención de datos mediante el portal del CENACE

Los datos que se decidieron utilizar son los datos que proporciona el Centro Nacional de Control de Energía (CENACE), este organismo público es el encargado de ejercer el control operativo del SEN, principalmente la operación del mercado eléctrico de México. CENACE es una dependencia de gobierno que garantiza la imparcialidad en el proceso de obtención de estos datos.

"CENACE es un organismo público descentralizado cuyo objeto es ejercer el control operativo del Sistema Eléctrico Nacional (SEN); la operación del Mercado Eléctrico Mayorista (MEM) y garantizar imparcialidad en el acceso a la Red Nacional de Transmisión (RNT) y a las Redes Generales de Distribución (RGD)" (CENACE, 2021).

Se puede acceder al portal de CENACE por medio de la siguiente url: https://www.cenace.gob.mx//

El sitio donde se obtienen los datos correspondientes a la estimación real de la demanda eléctrica es el siguiente: https://www.cenace.gob.mx/Paginas/SIM/Reportes/EstimacionDemandaReal.aspx/

CENACE hace una estimación promedio de la demanda de energía. Para calcular, la CENACE toma en cuenta muchos factores y la estimación que tiene mayor impacto son todas las solicitudes de energía aceptadas provenientes de la central principal. La estimación de la demanda real la calcula con el consumo neto de energía y el portal proporciona el cálculo de la estimación mediante dos métodos:

Balance: *"La estimación de la Demanda Real del Sistema por Balance se obtiene con base en la generación neta inyectada al sistema en cada hora, menos la energía de exportación. Se incluyen las pérdidas técnicas y no técnicas"* (CENACE, 2021).

Retiros: *"La estimación de la Demanda Real del Sistema por Retiros se obtiene agregando todas las compras de energía que se realizan por las Entidades Responsables de Carga, incluyendo las exportaciones. Se excluyen las pérdidas técnicas y no técnicas de la red que corresponde al Mercado Eléctrico Mayorista"* (CENACE, 2021).

Para el trabajo de esta tesis se opto por la estimación mediante balance, esto con el fin de tener los datos más aproximados a la información real, dado que los cálculos de la demanda por balance están más cerca de la cantidad de energía que realmente se consumió. Cabe señalar que CENACE no proporciona los datos en tiempo real, este cuenta con un retraso de dos semanas, por lo que no es posible obtener los datos del día actual o pocos días anteriores en el sitio, solo se pueden obtener los datos anteriores a 15 días.

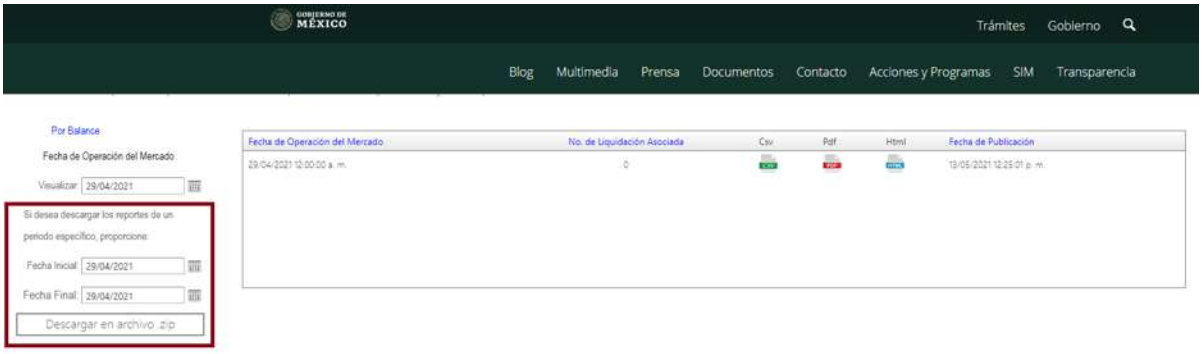

Figura 3.1: Portal de CENACE

En la figura 3.1 se muestra el portal del CENACE. Éste cuenta con una vista similar a un formulario, se puede obtener la información de la energía ingresando el día en especifico en el formulario. Una vez ingresado en la parte derecha, se desplegará un recuadro con los tipos de archivos posibles para la descarga de datos, el archivo contendrá información referente del día y tendrá un registro de cada hora por cada región.

En el caso de este proyecto se necesitan descargar varios días, por lo que sería necesario llenar el formulario de la izquierda mostrado en la figura 3.1. En este se debe de especificar que se descargarán los datos que están comprendidos entre el 01 de enero de 2020 hasta el 31 de Diciembre de 2020. Una vez especificada la fecha, se debe presionar el botón *"Descargar en archivo .zip"*. Esto descargará un archivo comprimido, el cual debería tener n cantidad de archivos csv, donde n es el numero de días proporcionados en el rango de fecha definido en el formulario (un archivo por día). En el caso de esta tesis deberían ser 366 archivos dado que es un año (bisiesto) de datos.

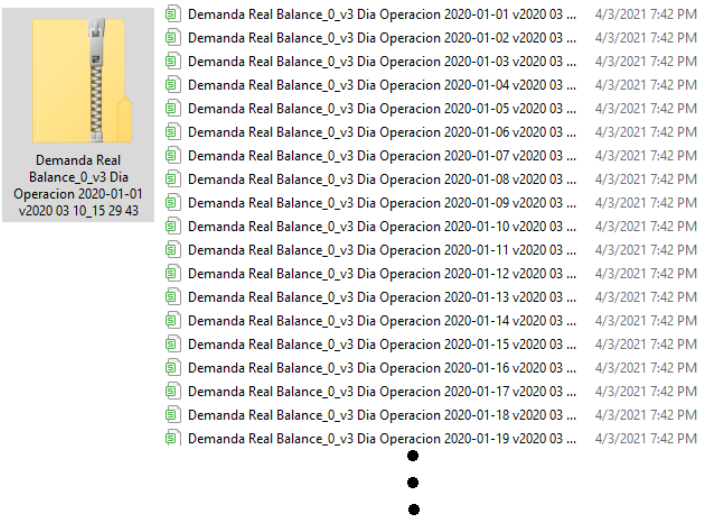

Figura 3.2: Datos descargados de CENACE

Estos archivos se descomprimen en una carpeta, dado que al ser una gran cantidad de archivos es mejor tenerlos ordenados. En el caso de este proyecto se almacenarán los archivos en una carpeta llamada “2020”, pero al observar la cantidad de archivos en la carpeta se puede apreciar que esta no solo contiene 366 archivos como se esperaría, en cambio esta contiene una cantidad mayor y esto se debe a que al ser una estimación, los datos no son perfectos, por lo que puede existir algún error externo al CENACE o que el control central del SEN no tuviera algunos datos exactos, por lo que CENACE publica diferentes versiones de los datos, donde los datos de las últimas versiones pueden ser los más acercados a la realidad.

3.2. Creación de un Marco de Datos a partir de archivos csv

Para la creación de un Marco de Datos primero se deben seleccionar los archivos a utilizar, cada archivo contiene un día de datos, estos vienen separados por hora y región, para leer todos los archivos primero se debe pasar la ubicación del directorio que contiene los archivos, se obtendrán los nombres de todos los archivos contenidos en dicho directorio, y estos nombres se guardan en un arreglo, el arreglo de cadenas de caracteres contiene todas

las rutas con el nombre de los archivos.

Tal como se mencionó en el capítulo 2, una de las problemáticas es que no solo existen 366 archivos csv. Para solucionarla solo se utilizan los archivos con la primera versión de estimación de balance, dado que esta versión si existe para los 366 días. Algunos días pueden tener versiones más recientes de balance, pero no se garantiza que todos los días tengan dicha versión, así que para evitar inconsistencia en versiones se elige tomar todos los archivos con versión de balance inicial (balance 0), de esta manera se evita que algunos archivos tengan mejor o peor margen de error que otros. Cabe señalar que el margen de error en la obtención de los datos por parte del CENACE entre versiones es aproximadamente el mismo, dado que no se cometen muchos errores en la medición de la energía consumida (demanda eléctrica), solo se cometen errores en otros cálculos de medición de datos no importantes para este proyecto, como lo son el intercambio de energía entre estaciones o generación de energía. El único valor importante es la energía consumida (demanda), la fecha y la región; los demás datos se desechan.

Para solamente utilizar los archivos con la primera versión de la estimación de balance, se hace un filtrado. Esto es muy sencillo de hacer ya que que todos los nombres de los archivos de balance de versión 0, ó primera versión, inician con: *Demanda Real Balance_ 0*, y todos los archivos tengan esta sub-cadena como prefijo en su nombre serán guardados en un vector, el cual tiene un tamaño de 366.

3.3. Depuración en un Marco de Datos

La depuración del marco de datos consiste en dejar los datos útiles y desechar los datos innecesarios. Se podrían dejar todos los datos, pero si nunca se utilizan se estarían desperdiciando recursos y visualmente entorpecerían la investigación.

Primero se eliminan todos los datos que obtienen energía de la región de Baja California (BCA), dado que esta región no corresponde al Sistema Interconectado Nacional (SIN) y los únicos que reciben energía de esta región es la misma Baja California y Baja California Sur. Posteriormente, se eliminan todas las columnas innecesarias, dejando solamente las siguientes:

1.- El “área” que correspondería a una de las Regiones del SIN.

2.- La “estimación de Demanda por Balance” lo cual es la demanda eléctrica (energía consumida).

3.- La “fecha” la cual remplazara la columna Hora y obtendrá ese campo con la columna Hora y la fecha en el archivo csv

	A	B	C	D	E	F	G	H	I	J	K
1	Centro Nacional de Control de Energia										
2	Estimacion de la Demanda Real del Sistema -Por Balance										
3	Sistema Electrico Nacional										
4	Reporte Diario										
5	Fecha de Publicacion: 10/mar/2020										
6	Archivo descargado desde el Sistema de Informacion del Mercado (Area Publica) creado el 10/mar/2020 15:29:43 hrs.										
7	Nota 1: Los acentos de este reporte se omiten intencionalmente por sistema.										
8	LIQUIDACION 0 (Dia de Operacion: 01/01/2020)										
9	Sistema	Area	Hora	Generacio	Importacio	Exportacio	Intercambi	Estimacion de Demanda por Balance (MWh)			
10	BCA	BCA	1	1124.1648	121.97846	220.9956	---	1025.1476			
11	BCA	BCA	2	1053.9497	83.21267	131.53384	---	1005.6286			
12	BCA	BCA	3	1047.2082	107.04241	172.11608	---	982.13452			
13	BCA	BCA	4	1042.4578	115.12048	197.85693	---	959.7213			
14	BCA	BCA	5	1026.5618	114.05929	198.96707	---	941.65398			
15	BCA	BCA	6	1039.2282	94.75696	205.94005	---	928.04509			
16	BCA	BCA	7	1024.734	84.6925	206.23433	---	903.19212			
17	BCA	BCA	8	1036.6241	75.03556	244.05527	---	867.60442			
18	BCA	BCA	9	1021.0931	28.88269	175.23318	---	874.74261			
19	BCA	BCA	10	993.9034	66.16441	169.1338	---	890.93401			
20	BCA	BCA	11	987.9028	73.72697	151.84983	---	909.77994			
21	BCA	BCA	12	970.21241	88.26372	136.7054	---	921.77073			
22	BCA	BCA	13	972.2006	97.82353	148.85886	---	921.16527			
23	BCA	BCA	14	968.62488	99.35309	150.40573	---	917.57224			
24	BCA	BCA	15	976.61005	85.33833	149.05476	---	912.89362			
25	BCA	BCA	16	1041.6693	102.03026	219.17842	---	924.52116			
26	BCA	BCA	17	1121.5218	122.96862	281.60854	---	962.88189			

Figura 3.3: Ejemplo de un archivo csv

Como se muestra en la figura 3.3, en el renglón 8 se tiene el valor de la “fecha de operación”. De aquí se saca la fecha, y de la columna “Hora” se saca la hora; de esta manera se puede obtener su día y hora específica.

Los datos empiezan desde el reglón 10 por lo que se saltan las primeras 9 lineas. También si el valor del sistema es ’BCA’ (Baja California), se desecha el dato. Por último, se eliminan las columnas de datos que no son de interés para este proyecto, dejando solo el área (región) y la demanda (Estimación de Demanda por Balance). Finalmente se agrega la fecha completa (día y hora) en una columna adicional.

	Area	Energia	Fecha
0	CEN	5108.53669	2020-01-01 01:00:00
1	CEN	4857.30372	2020-01-01 02:00:00
2	CEN	4625.40121	2020-01-01 03:00:00
3	CEN	4390.01186	2020-01-01 04:00:00
4	CEN	4230.77276	2020-01-01 05:00:00
...	...	...	...
61483	PEN	1461.10692	2020-12-31 20:00:00
61484	PEN	1405.51810	2020-12-31 21:00:00
61485	PEN	1348.38342	2020-12-31 22:00:00
61486	PEN	1284.86587	2020-12-31 23:00:00
61487	PEN	1239.47106	2020-12-31 00:00:00

61488 rows × 3 columns

Figura 3.4: Marco de Datos Totales

En la figura 3.4 se puede observar el marco de datos obtenido despues de la depuración, con un índice y cuyas columnas son área, energía y fecha. Este marco de datos se puede guardar en un archivo csv para su uso posterior.

3.4. Seccionamiento de un marco de datos

Una vez teniendo el marco de datos con todos los registros, se crean otros marcos de datos seccionando todos los datos por región, y este se modifica para su uso posterior en gráficas. Primero se crea un marco de datos filtrando todos los datos que lleven el mismo nombre de región en la columna *Área*. Posteriormente, se elimina la misma columna área (región), esto se hace dado que en el marco de datos ya seccionado tendría el mismo valor de área en todas los registros. Se quita el índice numerado y se pone como índice la columna *Fecha*. Esto por que el objetivo del índice es identificar a cada registro con un valor, y como cada registro tiene una fecha diferente no repetitiva (una hora de diferencia entre cada registro), esta cumple con la función de ser un identificador. Una vez hecho esto, no se necesita la columna *Fecha*, por lo que se elimina.

Fecha	Energia
2020-01-01 01:00:00	5108.53669
2020-01-01 02:00:00	4857.30372
2020-01-01 03:00:00	4625.40121
2020-01-01 04:00:00	4390.01186
2020-01-01 05:00:00	4230.77276
2020-01-01 06:00:00	4165.69996
2020-01-01 07:00:00	4181.54858
2020-01-01 08:00:00	3952.64798
2020-01-01 09:00:00	4000.26235
2020-01-01 10:00:00	4162.62147

Figura 3.5: Marco de Datos Seccionado

En la figura 3.5 se muestra el resultado final de un marco de datos seccionado. Este es muy simple dado que solo tiene el índice y una columna. Este marco tiene 8784 entradas, éstas son las horas que existen en un año bisiesto (366 x 24 = 8784). De esta manera se obtendrían 8 marcos de datos, siete marcos (uno para cada región del SIN) y el ultimo tendrá los valores sumados de los 7 marcos juntos, representando el valor total de la demanda de todo el SIN.

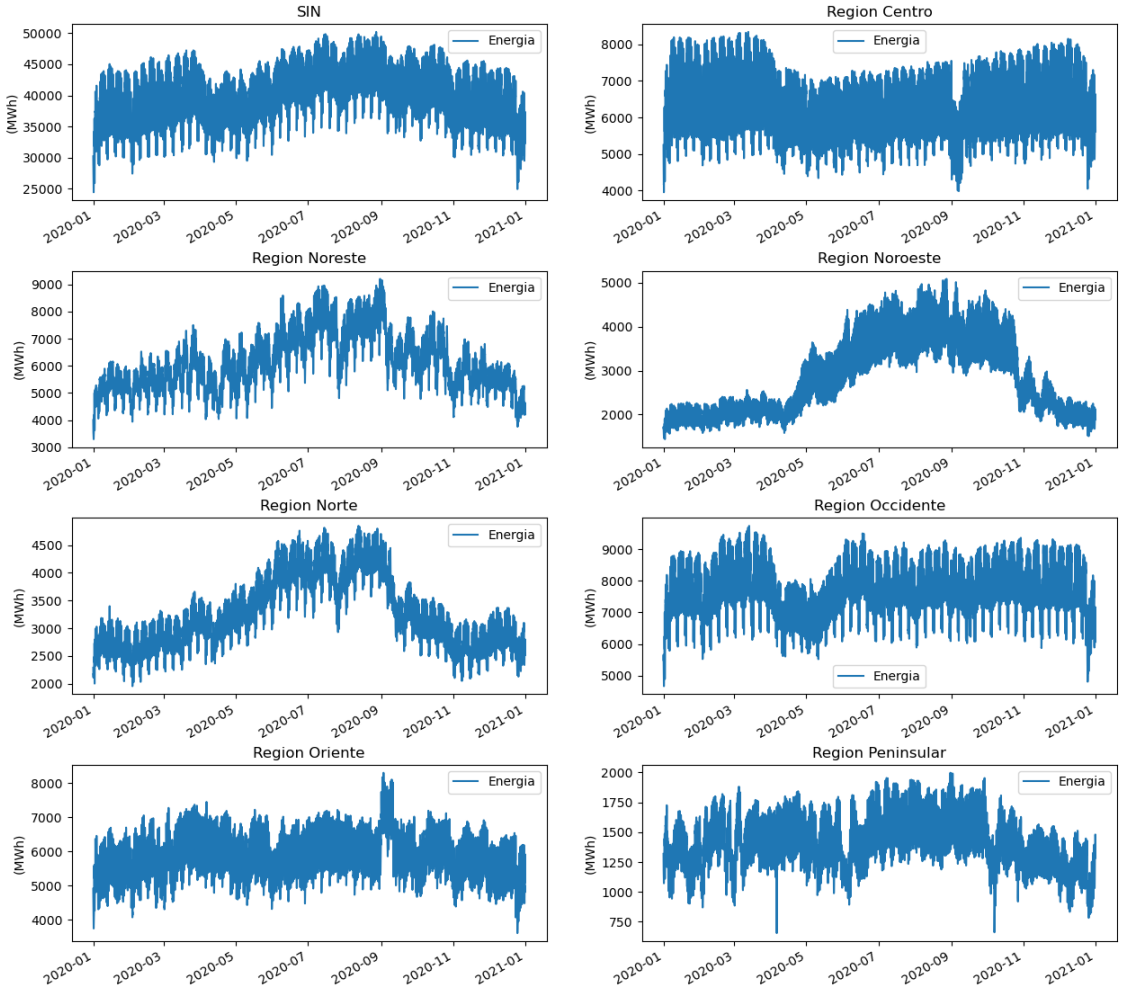

Figura 3.6: Demanda de Energía

En la imagen 3.6 se muestran los datos graficados de la demanda del SIN con sus respectivas 7 regiones. Se puede observar como la demanda varia entre 4000 MWh y 8000 MWh, donde se puede notar que a inicios y fines de año se tiene una demanda mayor a diferencia de mediados de año.

Fecha	Energia	cuarto	mes	semana	dia	ix	ven_prom_1	ven_est_1	ven_prom_7	ven_est_7	ven_prom_30	ven_est_30	ven_prom_90	ven_est_90
2020-01-01 01:00:00	29425.82739	1	1	1	2	0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2020-01-01 02:00:00	28445.97986	1	1	1	2	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2020-01-01 03:00:00	27423.96515	1	1	1	2	2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2020-01-01 04:00:00	26480.24095	1	1	1	2	3	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2020-01-01 05:00:00	25886.69956	1	1	1	2	4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2020-12-31 20:00:00	37346.44173	4	12	53	3	8779	33403.838162	2874.652951	32814.747479	3919.424049	36825.067538	4418.976007	38864.109089	4512.290603
2020-12-31 21:00:00	35764.85033	4	12	53	3	8780	33224.865600	2558.549747	32817.496396	3921.341157	36815.189094	4413.373880	38860.211780	4511.329798
2020-12-31 22:00:00	33800.40446	4	12	53	3	8781	33022.870435	2288.919883	32818.893854	3921.651189	36804.369562	4411.199107	38855.887638	4511.701446
2020-12-31 23:00:00	32320.12888	4	12	53	3	8782	32823.617583	2120.301911	32820.977498	3921.290475	36793.673032	4412.724116	38851.427334	4513.398759
2020-12-31 00:00:00	31000.47975	4	12	53	3	8783	32644.122311	2082.820566	32824.211331	3919.553059	36783.850043	4417.741336	38847.452523	4516.530927

8784 rows × 14 columns

Figura 3.7: Columnas añadidas al marco de Datos

En la imagen 3.7 se muestra como se crean columnas adicionales con fines informativos y algunas para su uso en los próximos capítulos, para seccionar los datos dependiendo de en que categoría cae.

Estas columnas son las siguientes:

1.- cuarto: divide el año en 4 partes (3 meses) y este valor indica a cual de las cuatro partes del año pertenece.

2.- mes: indica a que mes pertenece el dato, tomando un valor entero entre el 1 y el 12, donde el 1 es enero y el 12 es diciembre.

3.- semana: indica la semana del año al que pertenece el dato, tomando un valor entero del 1 al 53, donde el 1 significa la primera semana del año y el 53 a la ultima semana del año.

4.- día: indica a que día de la semana pertenece, tomando un valor entero entre el 0 al 6, donde el 0 es lunes y el 6 es domingo.

5.- ix: una columna que enumera el numero de dato que es, del 1 al 8784.

También se crearon varias columnas adicionales que se calculan mediante medias móviles. Una toma los valores de la media y otra de la desviación estándar para su cálculo. Las ventanas utilizadas están hechas por diferentes cantidades de datos. Por cada día (24 datos), por cada 7 días (7 x 24 datos), por cada 30 días (30 x 24 datos) y por 90 datos (90 x 24 datos). Estos datos son útiles para el siguiente capítulo , en el cual se analizan y grafican estos datos con las columnas recién creadas.

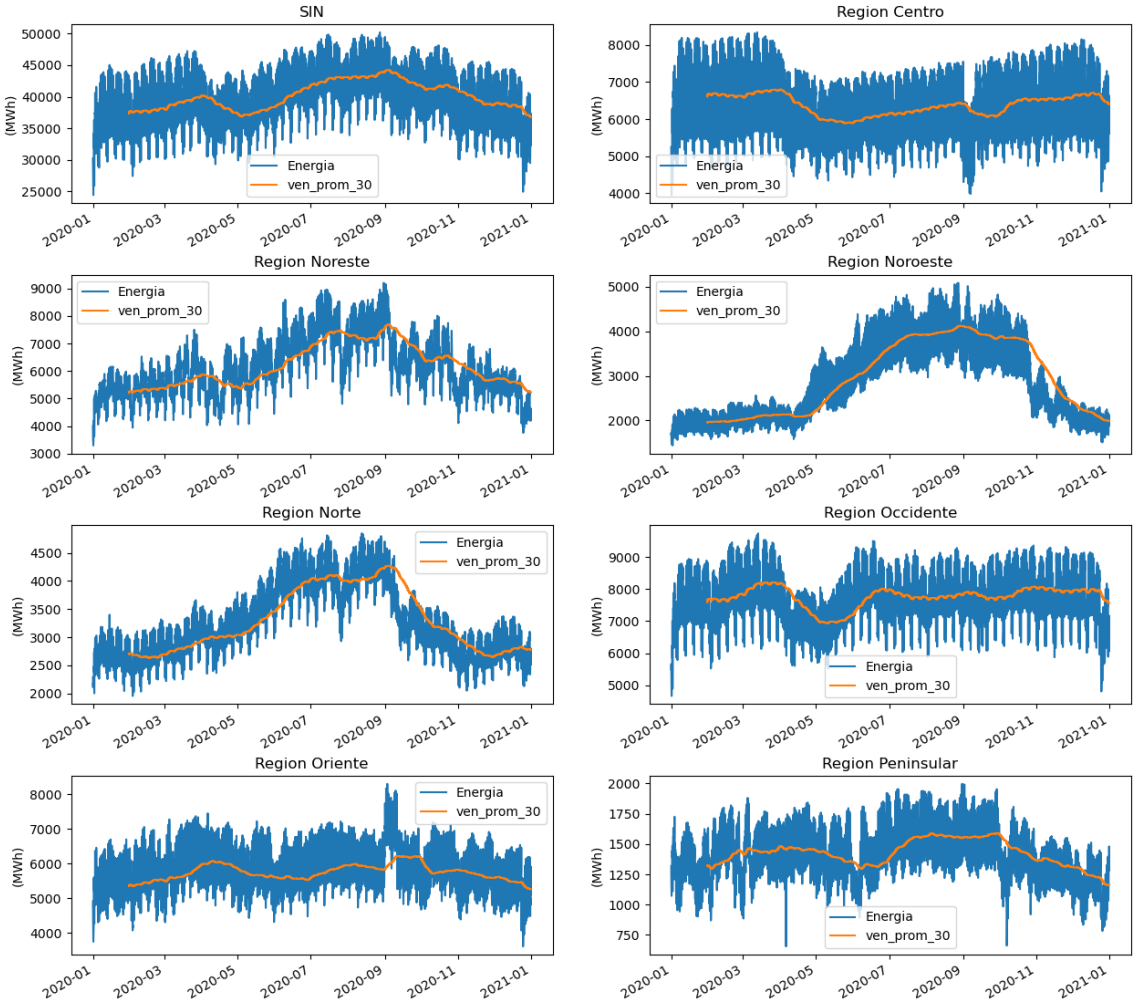

Figura 3.8: Demanda de Energía diaria

En la figura 3.8 se muestra una gráfica de los datos y la ventana deslizante por promedio utilizando ventanas de 30 días de datos. De esta forma se puede ver un poco mejor el comportamiento de los datos. Para calcular la media móvil se toman 30 días (24 x 30 datos), se hace su promedio y el valor obtenido es la media móvil de promedio con 30 días. Este proceso se hace para todos los valores hasta llegar al último. Nótese que los días anteriores al 30 tienen un valor no definido, esto se debe a que no tienen suficiente información para calcular la media móvil, por lo que la curva empieza desde el día 30 para adelante.

3.5. Comentarios finales

En este capítulo se lleva a cabo todo el proceso para obtener marcos datos útiles para el análisis y el modelo de la predicción. Se mostró como obtener los datos reales del portal del CENACE, como se transformo de archivos csv a un marco de datos en python, después se separo y elimino la información no útil para este proyecto, finalmente se le dio un formato y se calcularon algunas variables útiles para el análisis de dichos datos. De esta manera obteniendo marcos de datos con solo la información indispensable, por lo que se puede continuar con los siguientes capítulos de manera adecuada, garantizando su correcto funcionamiento.

Capítulo 4

Análisis de Datos

El objetivo de este capítulo es visualizar el comportamiento de los datos sin necesidad de utilizar métodos estadísticos complejos a excepción de la prueba para normalidad Shapiro-Wilk. Lo principal es identificar si el comportamiento de los datos es normal. En caso de no serlo, se debe de saber hasta que punto estos no son normales, esperando que estos no estén en el extremo de la no normalidad, lo cual es que el comportamiento de los datos sea caótico (de manera aleatoria e impredecible). Para esto se analizara el comportamiento del Sistema Interconectado Nacional en sus respectivas 7 regiones. Para esto, los datos se grafican separándolos en semana, mes y cuarto de año (3 meses). Se utilizan variables estadísticas calculadas como ayuda al entendimiento de las mismas gráficas.

4.1. Distribución Normal

En esta sección se demuestra si los datos pertenecen a una distribución normal o no. Para esto primero se calculan las variables estadísticas más importantes, las cuales son las siguientes:

1.- Media: es el valor que se obtiene al dividir la suma de los datos entre la cantidad de ellos, también llamado promedio.

2.- Desviación estándar: es una medida que nos dice que tan dispersos están los datos de la media.

3.- Varianza: al igual que la desviación estándar es una medida que nos indica que tan dispersos están los datos utilizando el promedio de las diferencias elevadas al cuadrado o simplemente el valor de la desviación estándar al cuadrado.

4.- Sesgo: Indica cual es la inclinación de la gráfica de distribución de los datos, esta puede ser inclinada al centro, a la derecha o a la izquierda.

5.- Curtosis: es una medida que determina que tan concentrados están los datos en la zona central de la gráfica de la distribución normal.

	Promedio	Desviación	Varianza	Sesgo	Curtosis
SIN	39854.18	4608.82	21241205.42	-0.22	2.46
Región Centro	6372.91	863.32	745314.78	-0.15	2.08
Región Noreste	6121.32	1050.04	1102590.19	0.37	2.62
Región Noroeste	2893.72	900.45	810804.67	0.39	1.78
Región Norte	3252.51	644.06	414810.13	0.49	2.31
Región Occidente	7722.67	921.11	848436.11	-0.21	2.24
Región Oriente	5706.99	622.34	387302.42	0.35	3.26
Región Peninsular	1411.14	213.78	45699.13	-0.15	2.65

Tabla 4.1: Variables estadísticas del SIN y sus respectivas 7 regiones.

En la tabla 4.1 se puede observar como las regiones Occidente, Centro y Noroeste tienen el mayor promedio de demanda de las 7 regiones. Cada región tiene diferentes factores que afectan a la cantidad de energía consumida como: cantidad de personas (población) y extensión territorial (km^2), cantidad de industrias, entre otras. Por esta razón, no se deben sacar conclusiones precipitadas o conjeturas; no se deben hacer este tipo de comparaciones entre regiones, solo deberían hacerse comparaciones de datos de la misma región en diferentes periodos de tiempo.

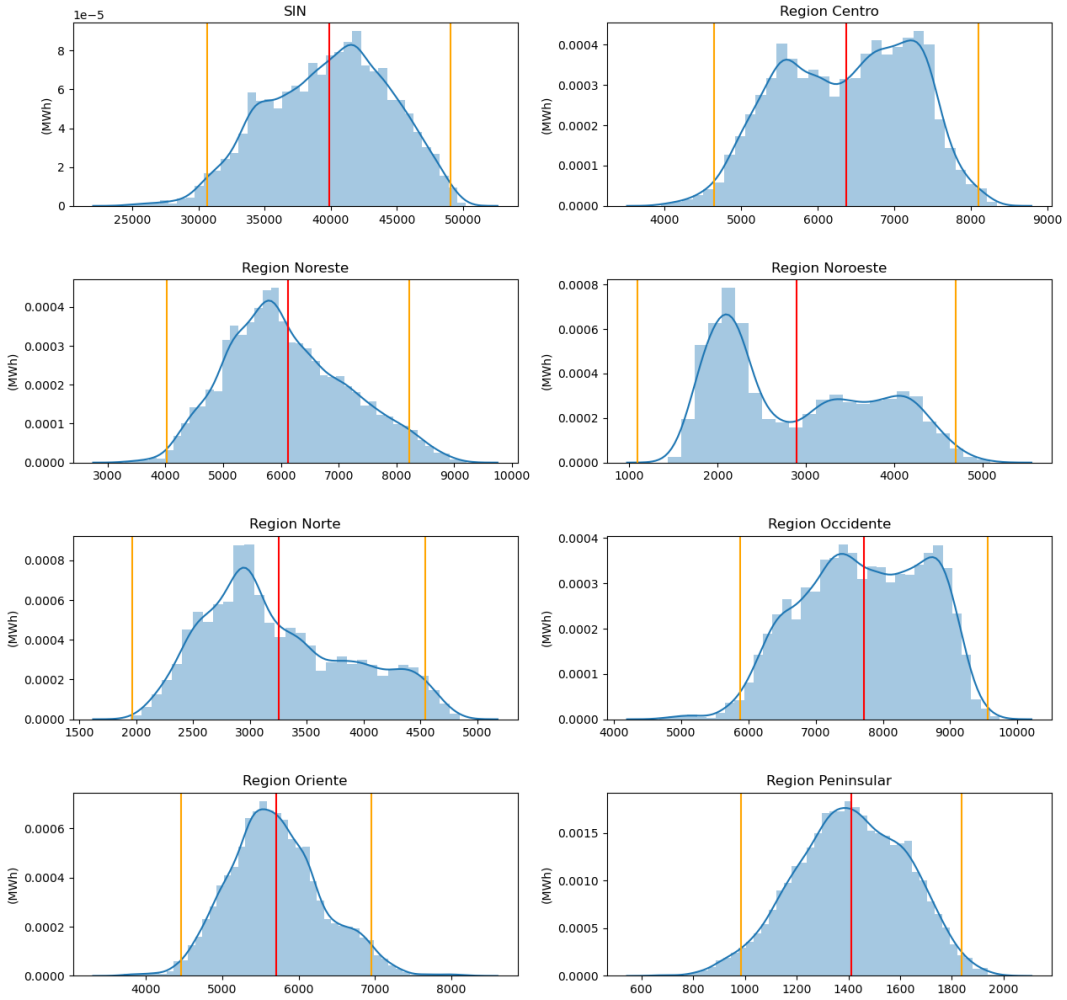

Figura 4.1: Distribución de datos del SIN y sus 7 regiones

En la figura 4.1 se muestra la gráfica de la distribución de los datos, donde la linea roja es la media y las lineas naranjas a los costados muestran 2 veces el valor de la desviación estándar (antes y después de la media). Se puede observar cómo ninguna distribución tiene la forma de una campana de Gauss perfecta, por lo que se difiere que los datos no tienen una distribución simétrica. Esto a simple vista dice que los datos no tienen una distribución normal, pero para comprobarlo será necesario realizar la prueba de Shapiro-Wilk.

Se declara como hipótesis nula (HØ) que: *Los datos se extraen de una distribución normal* y para los 8 casos mostrados en la tabla 4.2 esta hipótesis se rechaza dado que para todos

	Variable p	Variable α
SIN	1.33e-25	0.5
Región Centro	2.06e-35	0.5
Región Noreste	3.01e-30	0.5
Región Noroeste	0.0	0.5
Región Norte	9.81e-45	0.5
Región Occidente	2.66e-34	0.5
Región Oriente	1.44e-22	0.5
Región Peninsular	5.04e-15	0.5

Tabla 4.2: Prueba de normalidad Shapiro Wilk.

los casos el valor de la variable p es menor que el valor de la variable alfa (α). Esto quiere decir que los datos no se obtuvieron de una distribución normal. Una vez sabiendo que la distribución de los datos no es normal, se despierta la curiosidad de saber por qué. Una de las posibles hipótesis es que 2020 es un año bastante "especial" por así llamarlo, dado que aparte de ser un año bisiesto ocurrió por un evento histórico y esta fue la pandemia ocasionada por el virus Covid-19. Esto provoco que a principios de año muchas empresas cerraran y que las personas se quedaran en sus casas, pero realmente ¿esto afecta a la normalidad de la distribución?. Para ver mejor el panorama se obtuvieron los datos del 2019.

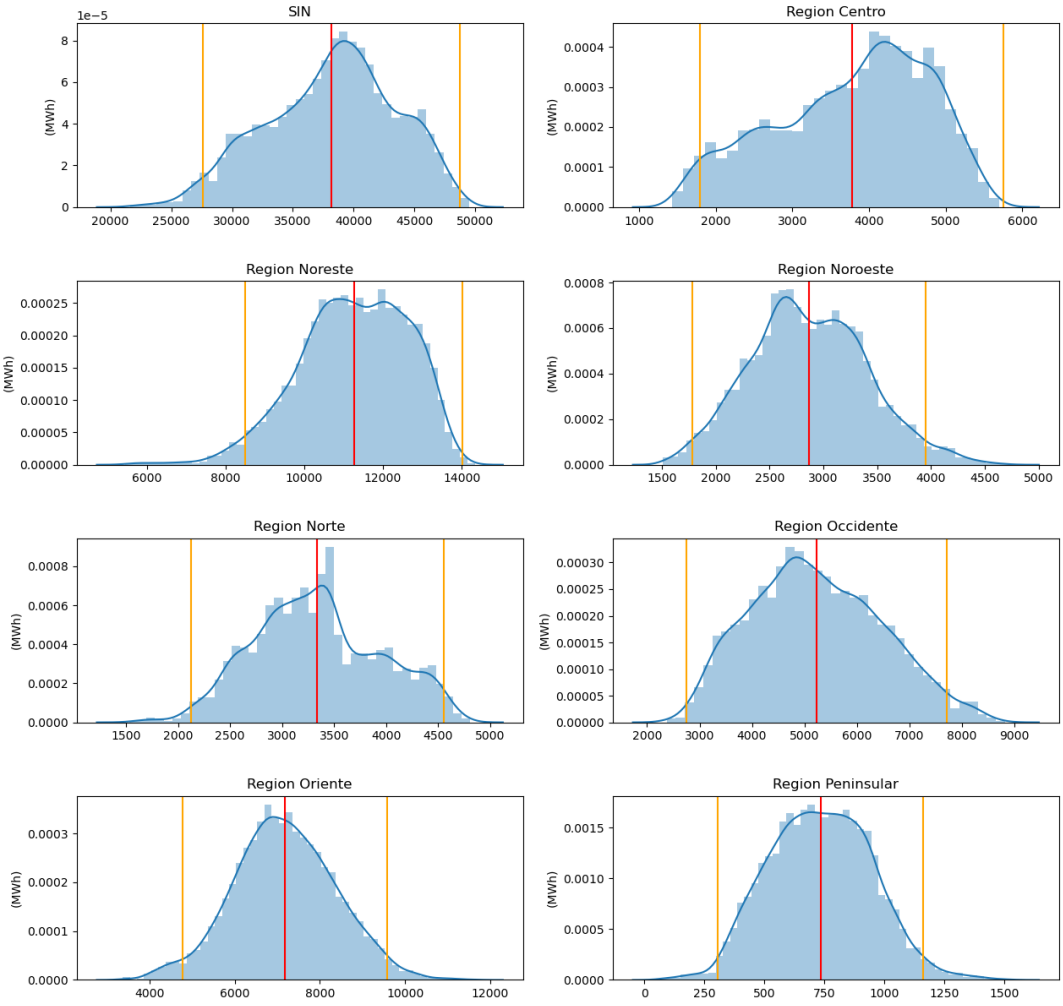

Figura 4.2: Distribución de datos del SIN y sus 7 regiones en el año 2019

Se hizo el mismo procedimiento para los datos obtenidos del 2020, pero esta vez para los datos del 2019 y el resultado de graficar las distribuciones de los datos están en la figura 4.2. Tal vez los datos se vean un poco más "normales" o con una forma más simétrica pero tampoco es una campana perfecta de Gauss, tal vez la región Oriente y Peninsular pueden tener una distribución normal, pero si se nota bien estas tienen una leve inclinación y están un poco asimétricas, por lo que para ninguno de los casos los datos tienen una distribución normal.

La respuesta por la que la distribución no es normal es por que el efecto de "nor-

malidad" de datos, es un concepto meramente teórico y sería el comportamiento ideal de los datos en un sistema, pero es muy difícil que una distribución obtenida de datos reales se comporte de forma "normal". Es cierto que los datos del 2019 varían en comparación a los del 2020, probablemente por la pandemia, pero esto no hace la diferencia de que la distribución sea normal o no, se continuara el proceso de análisis para saber la causa de la no normalidad de los datos.

4.2. Análisis de volatilidad y heterocedasticidad

La volatilidad es muy usada en áreas financieras para ver el comportamiento del mercado, *"es una medida estadística de la dispersión de los rendimientos de un valor"* (Chen, 2021). El análisis de volatilidad nos sirve para ver que tanto varían los datos y si esas variaciones tienen un valor muy grande (alta variabilidad) o muy pequeño (baja variabilidad); lo mejor para el modelo es tener errores poco variables de tal manera que se pueda identificar una tendencia y se puedan corregir de manera sencilla.

"La volatilidad a menudo se mide como la desviación estándar o la variación entre los rendimientos de ese mismo valor o índice de mercado" (Chen, 2021).

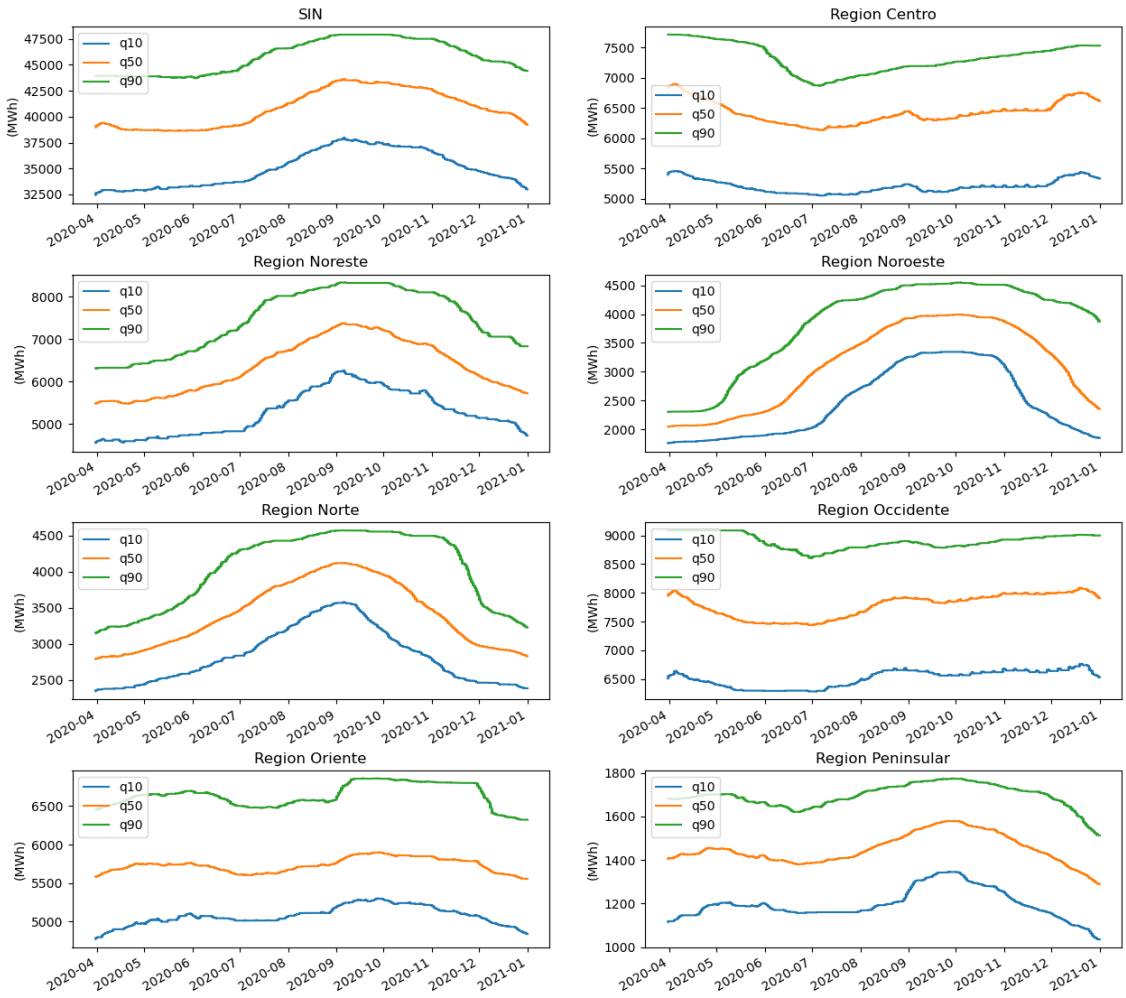

Figura 4.3: Análisis de volatilidad con percentiles de 10 %, 50 % y 90 % de los datos

En la figura 4.3 se muestran los datos separados por percentiles utilizando rangos de datos de 90 días. Se puede ver en las gráficas como el comportamiento del 10 % (q10), del 50 % (q50) y del 90 % es similar, dado que tienen una curva muy parecida entre sí, pero con una amplitud diferente; por lo que se puede decir que la diferencia en amplitud entre el percentil del 90 % y el percentil del 10 % se mantiene casi constante, si las curvas de los percentiles son similares, esto indica que el cambio del error tiene poca variabilidad; lo cual le favorece al modelo de predicción.

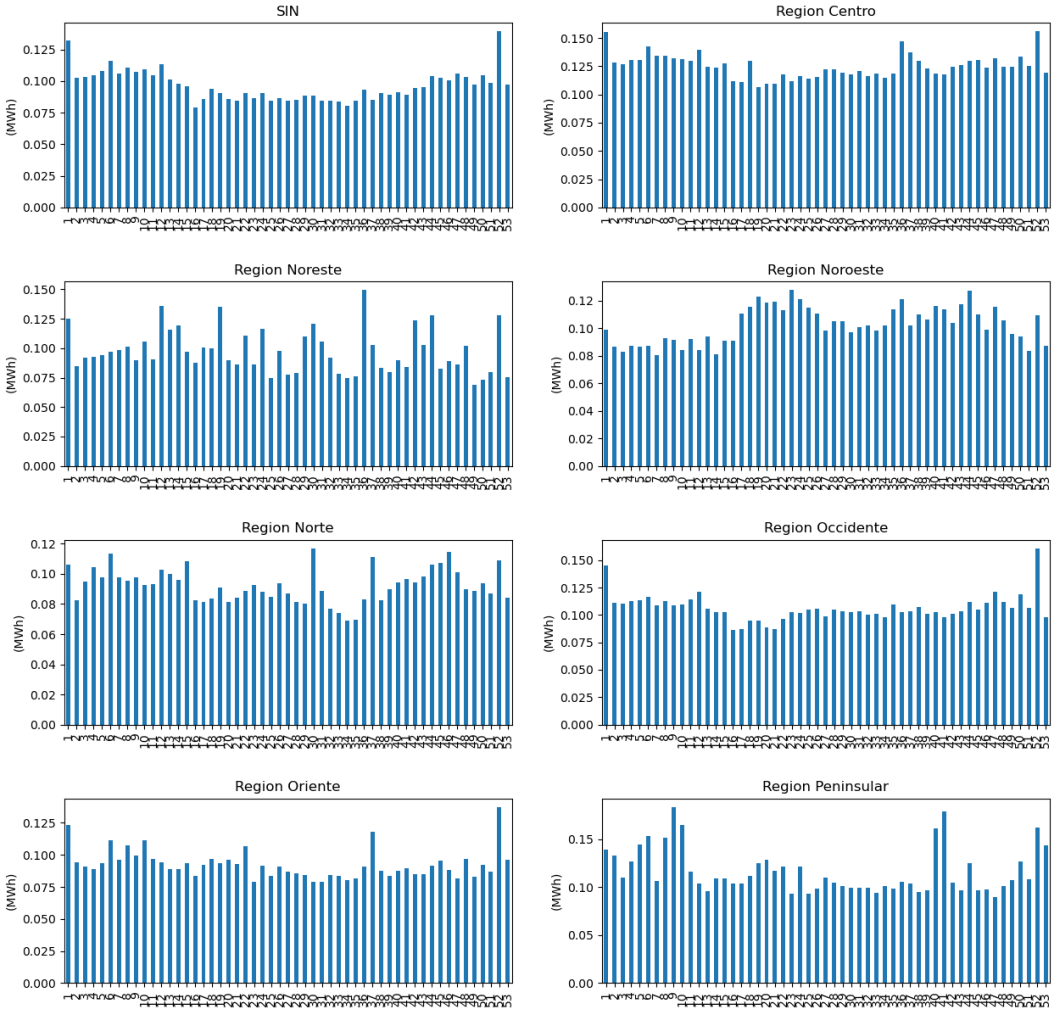

Figura 4.4: Coeficiente de variación por semana

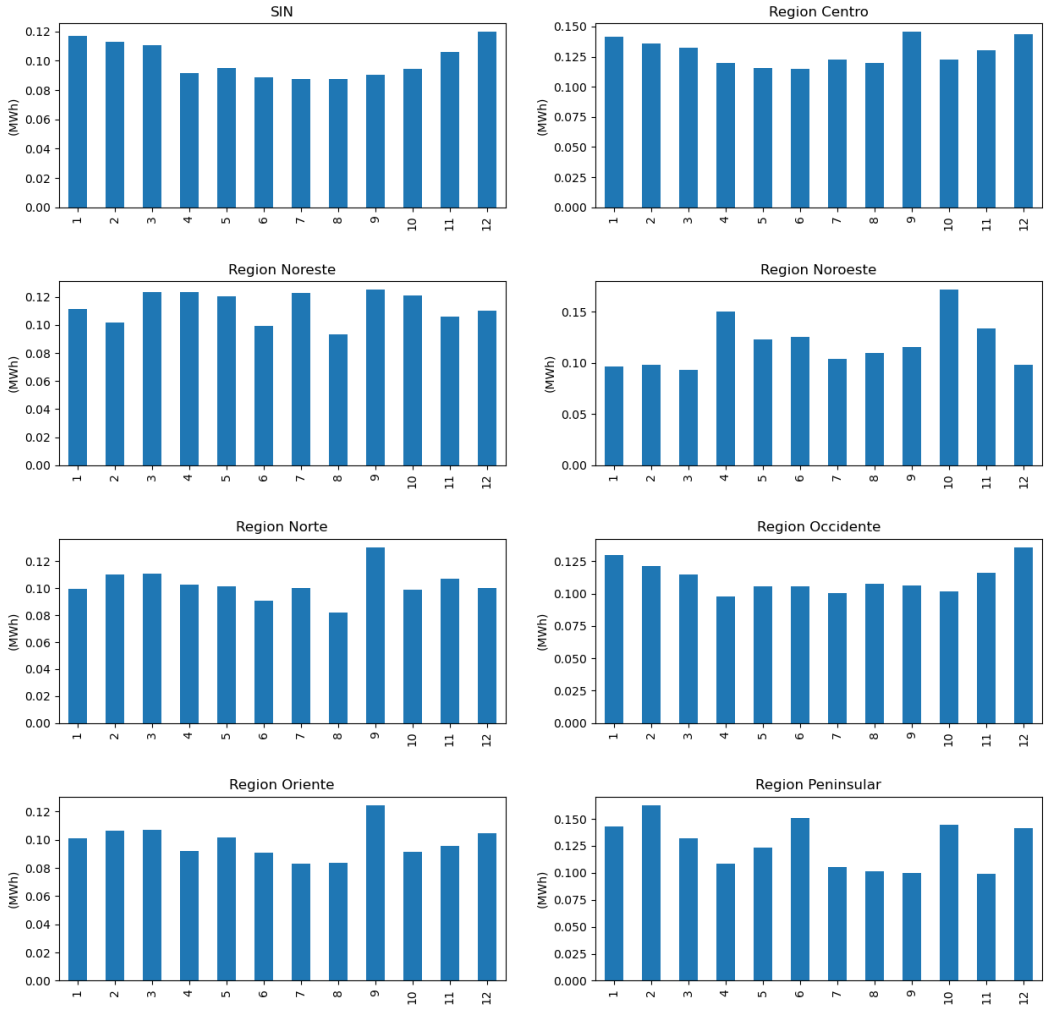

Figura 4.5: Coeficiente de variación por mes

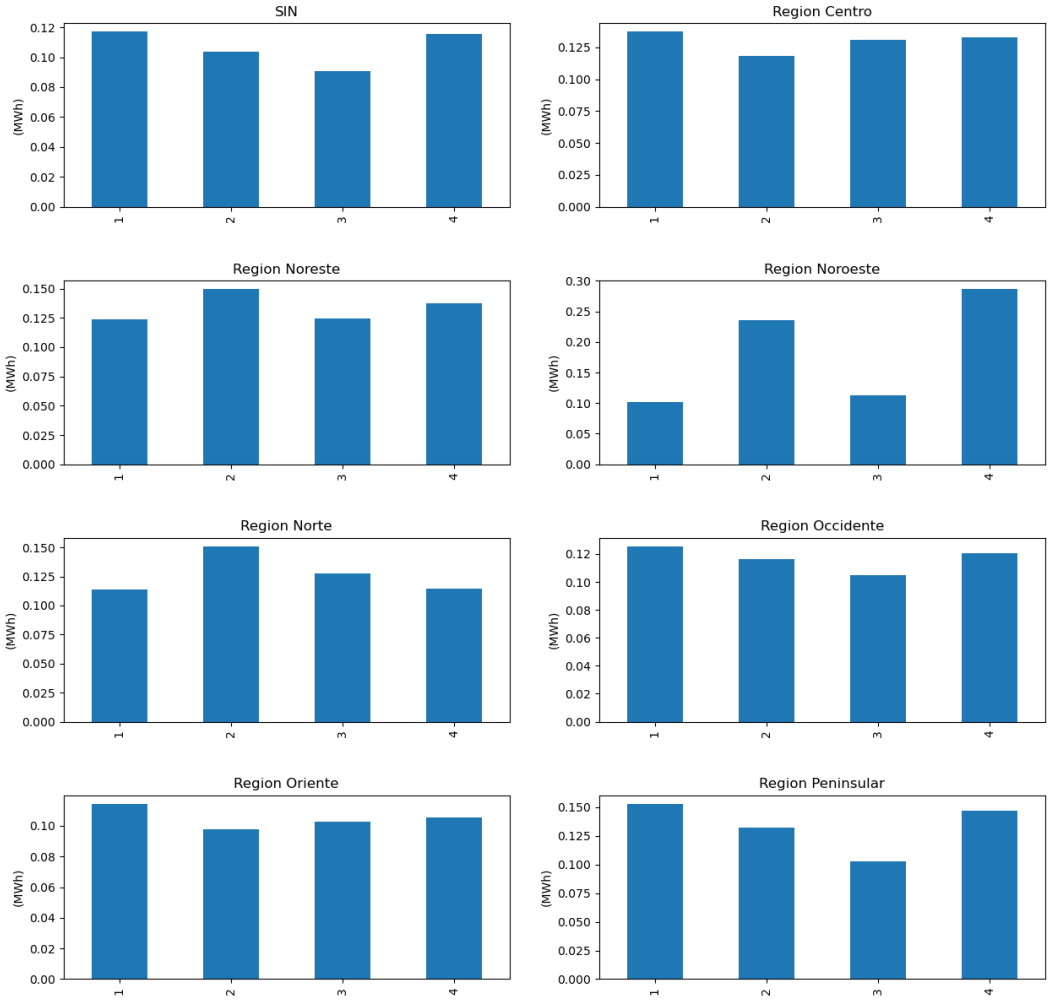

Figura 4.6: Coeficiente de variación por cuarto de año

Las figuras 4.4, 4.5 y 4.6 nos muestran el coeficiente de variación (CV) por periodos de una semana, un mes y un cuarto de año. Se puede observar como para el SIN los datos tienen un coeficiente de variación muy cercano para cada mes. Esto quiere decir que los datos tienen aproximadamente la misma variación en cada mes, lo que indica que la variable siempre esta en constante cambio, ésta no se queda estática en un periodo de tiempo y no disminuye su variabilidad; esto es buen indicio para el modelo dado que entre menor sea la variabilidad de los datos es mas fácil su predicción.

Ahora se procede a hacer el análisis de heterocedasticidad para ver el comporta-

miento de los datos fuera de la media, para ver si estos valores se comportan de manera homogénea o no.

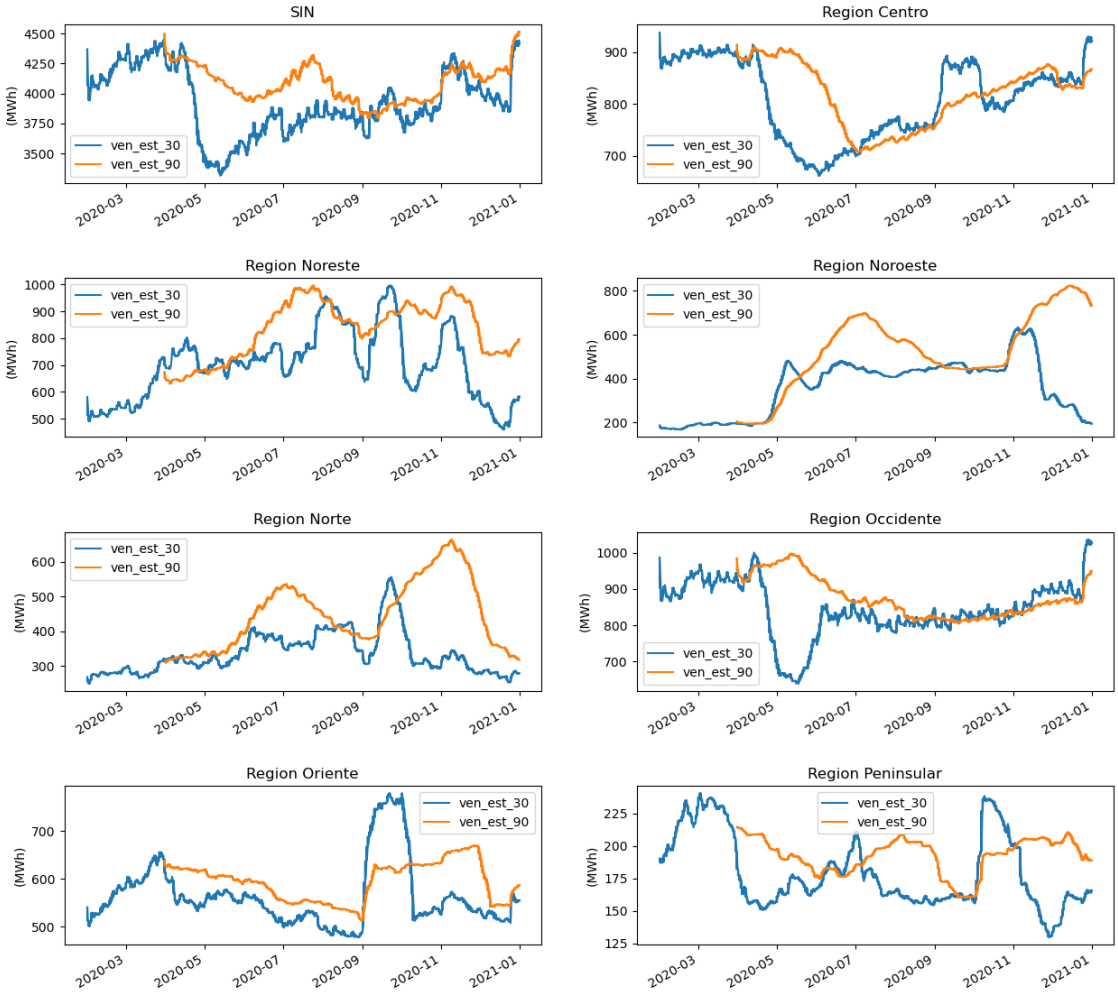

Figura 4.7: Análisis de Heterocedasticidad

La imagen 4.7 nos muestra la media móvil utilizando la desviación estándar en los datos. Tal como se habla en el primer apartado de esta sección, la desviación estándar es la que se utiliza principalmente para analizar la volatilidad de los datos. En caso de tener una desviación estándar constante, se puede concluir que los datos tienen un error también constante, a esto se le conoce como homocedasticidad, pero no es este el caso, dado que el error no es constante en la curva.

Se puede notar que, si el error fuera constante, ambas curvas serían similares, dado que al tomar un promedio de 30 días de datos de la desviación estándar de una variable constante, debería ser el mismo promedio que tomando muestras con 90 días de datos, por lo que se puede decir que los errores son heterocedásticos (no homogéneos), esto ocasiona que los errores entorpezcan el entrenamiento del modelo de predicción.

4.3. Análisis de series de tiempo por estaciones y tendencia

Para descubrir si en los datos existe una tendencia o algún ciclo, se hará el análisis de series de tiempo. Para esto se dividen los datos en cantidades de periodos diferentes.

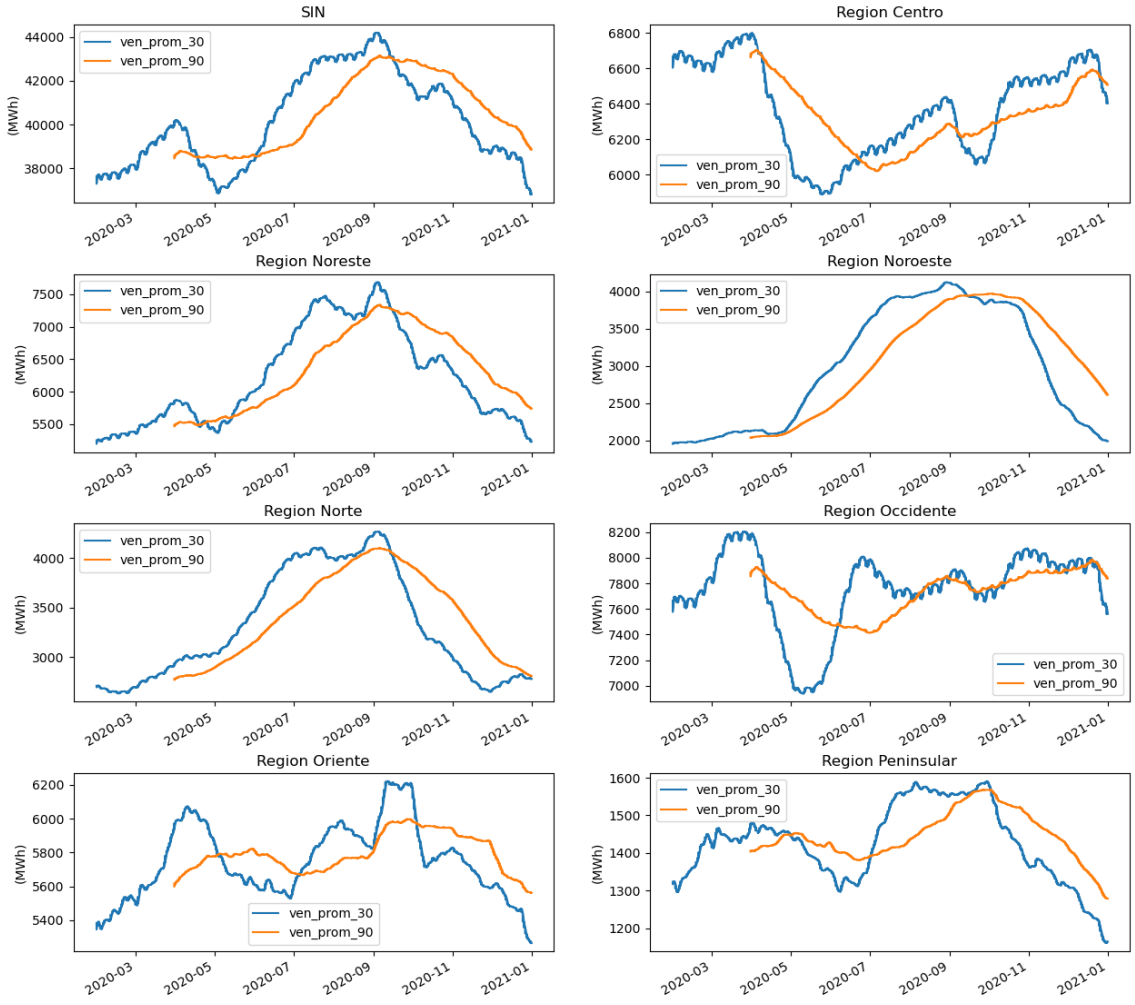

Figura 4.8: Análisis estacional de ventanas deslizantes por promedio

En la figura 4.8 se hizo el cálculo de la media móvil con 30 y 90 días de datos, esto con la finalidad de ver el comportamiento aproximado. Tomando el periodo de un mes y de 3 meses, el resultado visual es que algunos promedios tienen un cambio muy drástico entre periodos y esto se debe a que algunas veces el consumo eléctrico en México es diferente para cada mes o cuarto de año. Se puede apreciar que la curva de 90 días se asemeja un poco a la curva de 30 días y esto se debe a que, entre más datos se tomen para la media móvil, más suave es la curva resultante, pero esto ocasiona que la interpretación se aleje de la realidad.

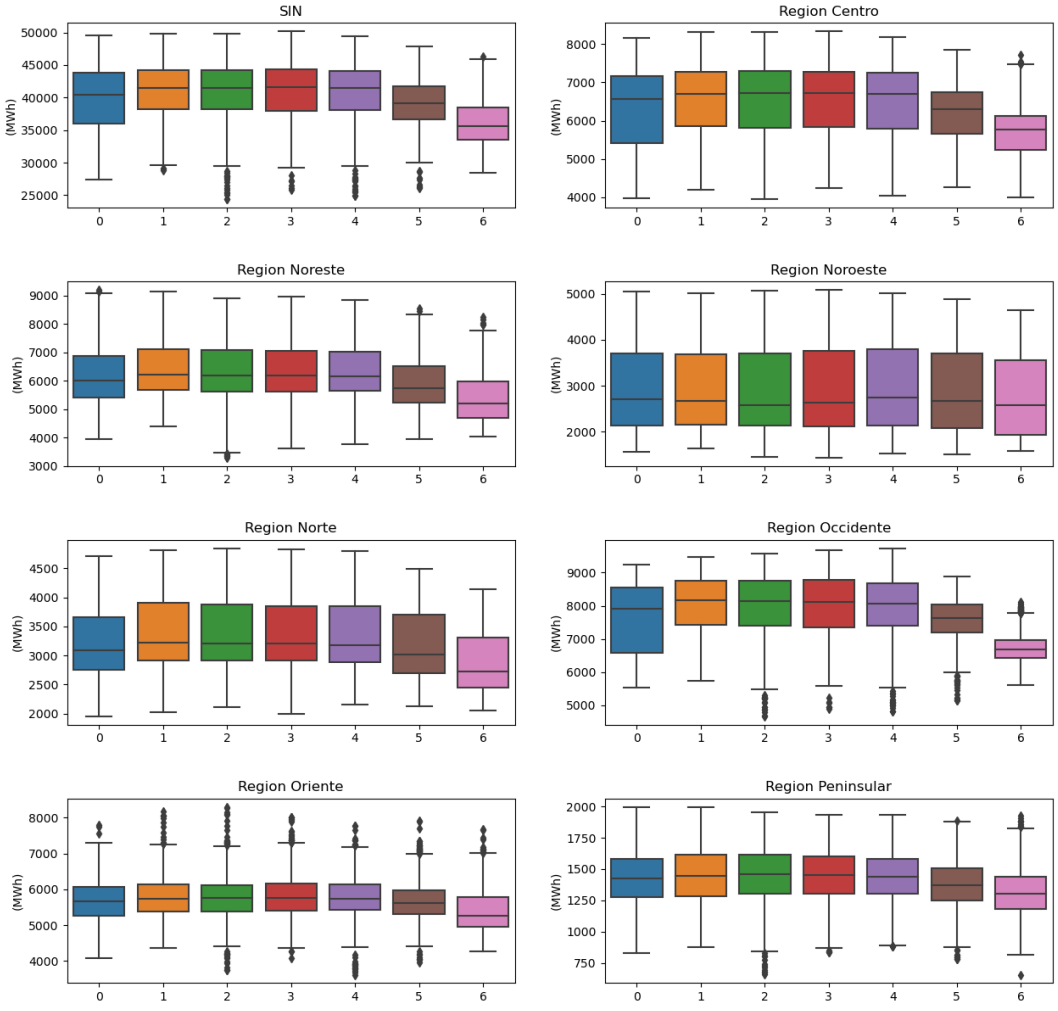

Figura 4.9: Análisis estacional de distribución por día de la semana

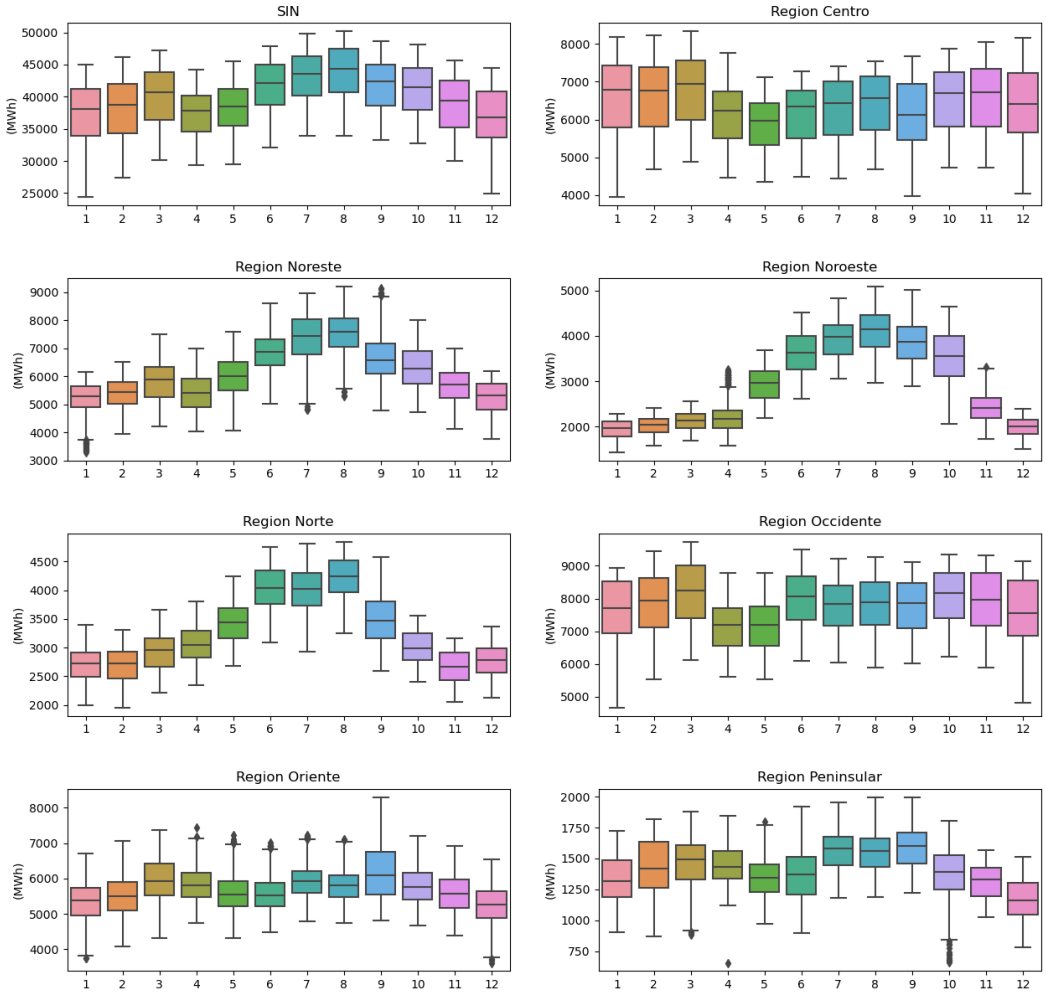

Figura 4.10: Análisis estacional de distribución mensual

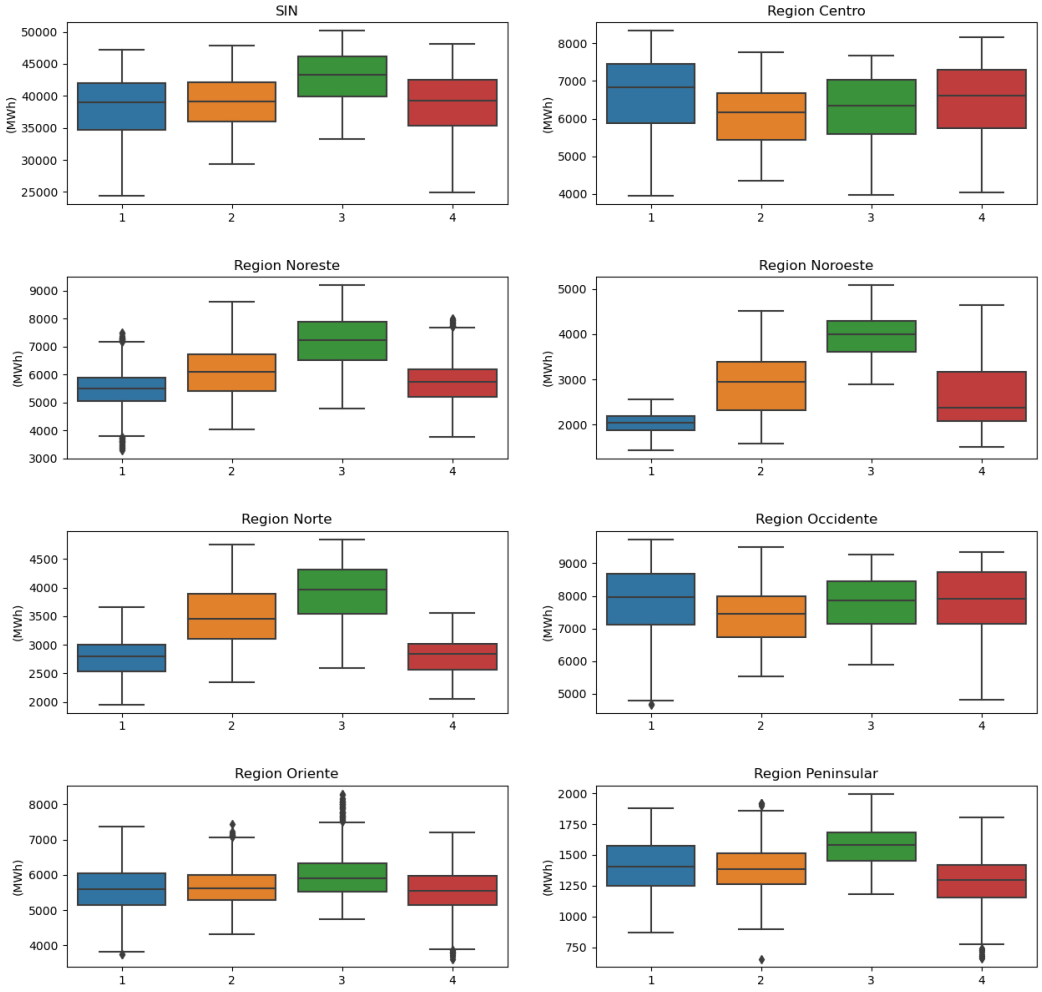

Figura 4.11: Análisis estacional de distribución por cuartos de año

En las imágenes de la 4.9, 4.10 y 4.11 se grafican por día de la semana, mes, cuartos de año respectivamente. En estos se puede notar un pequeño patrón estacional en las gráficas correspondientes a los días de la semana y el mes, donde se ve como los datos se comportan casi de forma cíclica. Se tomaron todos los días de cada semana y se hizo un promedio.

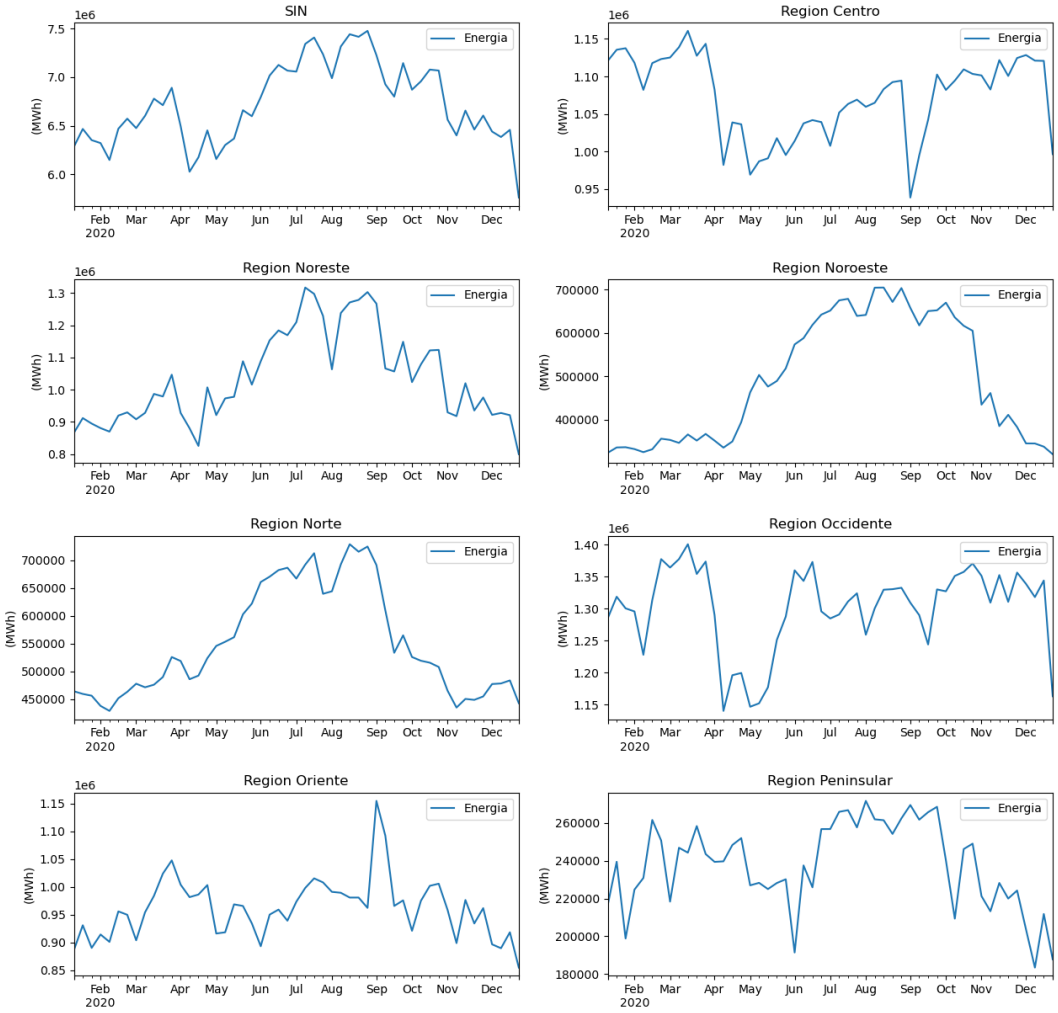

Figura 4.12: Demanda promedio por semanal

En la imagen 4.12 se graficaron los promedios de cada una de las semanas del año, donde se toman los 7 días de la semana y se calcula el promedio de la demanda semanal. Este proceso se hace para 51 semanas del año. Se decide ignorar la primer y ultima semana del año, por que el año no empieza en lunes y tampoco termina en domingo, por lo que no tendrían los 7 días. Esto quiere decir que si se toman en cuenta estas semanas, su valor estaría fuera del rango aproximado, en comparación con las semanas que sí están completas.

También se puede observar en la figura 4.12 cómo el consumo para cada semana

es diferente y esto depende de muchas variables externas (no tomadas en cuenta en este proyecto) como los días festivos, las vacaciones, entre otras.

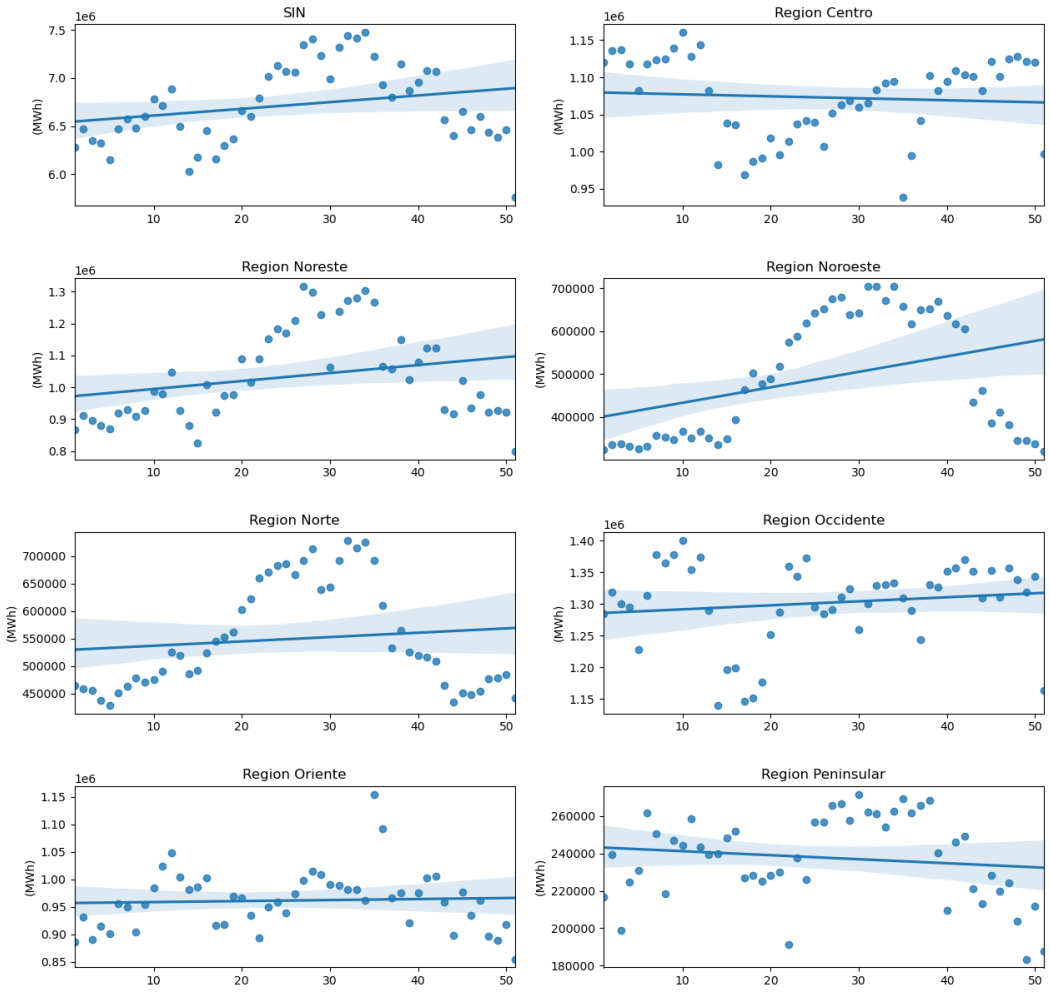

Figura 4.13: Análisis de tendencias con Regresión

En la figura 4.13 se realizo la regresión lineal utilizando los promedios semanales anteriormente calculados. No es objetivo de esta tesis presentar la ecuación obtenida de la regresión lineal (aunque generalmente al utilizar regresión lineal se consigue la ecuación de una recta). Como se muestra en la figura, la mayoría de los datos no están cerca de la recta, por lo que al elegir este método ocasionaría un margen de error muy grande. Esta regresión solo se hace con fines informativos y para su comparación con el modelo de

predicción en los próximos capítulos.

4.4. Comentarios finales

Después de analizar el comportamiento de los datos, se obtuvo el resultado de que ninguna de las distribuciones tienen un comportamiento normal o ideal. Algunas se acercaron a la forma de una campana de Gauss pero no a la perfección. Una vez obtenido este resultado, se hicieron una serie de análisis donde se descubrió que los datos no tienen variabilidad alta, sino que ésta se comporta con cambios leves, no drásticos. Sus errores no siguen un comportamiento homogéneo, pero se puede notar un comportamiento estacional, lo que quiere decir que los datos se comportan de forma cíclica.

Con este análisis quedo claro que los datos no se comportan de forma normal o ideal pero tampoco están en el extremo de ser una distribución caótica dado que se puede definir una tendencia en los mismos, por lo que el resultado en la predicción del modelo debería ser precisa o acertada, pero no perfecta.

Capítulo 5

Construcción del modelo de predicción

En este capítulo se explica el proceso del construcción del modelo de predicción. Para esto se utiliza la ingeniería de características, la cual consiste en separar los datos en variables de comportamiento las cuales son llamadas ”características”, en este proceso se pueden obtener bastantes características y dentro de estas se deben escoger solamente las que estén más relacionadas con la variable principal y fuertemente correlacionadas entre dichas características (utilizando el coeficiente de correlacion de Pearson). De esta manera al utilizar las mejores características, se asegura que el desempeño del modelo sera lo mejor posible.

5.1. Creación y selección de características

El objetivo ideal del modelo es igualar la distribución de datos original, por lo que se separan los datos en las variables que hacen que estos varíen. En esta sección se habla acerca de la creación de estas variables llamadas características.

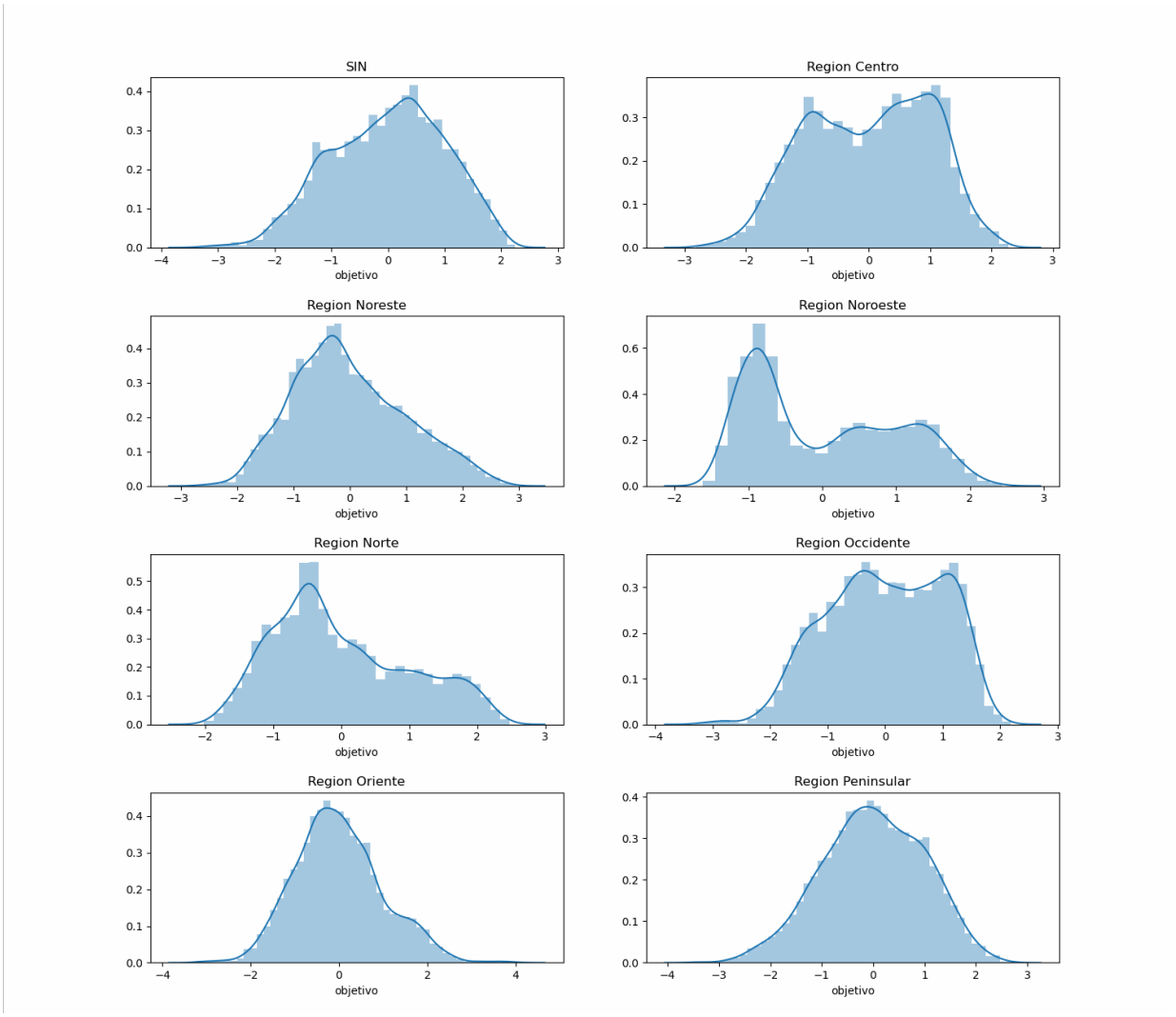

Figura 5.1: Distribución Objetivo

La figura 5.1 muestra la distribución de los datos, la cual se pretende imitar, pero dado a que la distribución no es normal, se dificulta un poco la construcción del modelo y por consiguiente la predicción.

Una vez sabiendo la distribución objetivo se procede a la creación de las características. Para esto se consiguen diferentes tipos de características utilizando 30 periodos de tiempo. Esto quiere decir que los datos se dividirán en 30 partes iguales. Los tipos de características que se tomaron en cuenta son las siguientes:

1.- Objetivo: Utilizando la distribución objetivo pero retrocediendo los datos en 1 pe-

riodo.

2.- Regresión : Utilizando la distribución objetivo pero adelantado los datos en 1 periodo.

3.- Media móvil con el promedio: Calculando la media móvil de cada 7, 14 y 30 días de datos utilizando el valor del promedio.

4.- Media móvil con la desviación estándar : Calculando la media móvil de cada 7, 14 y 30 días de datos utilizando el valor de la desviación estándar.

5.- Media móvil con el valor mínimo: Calculando la media móvil de cada 7, 14 y 30 días de datos utilizando los valores mínimos.

6.- Media móvil con el valor máximo: Calculando la media móvil de cada 7, 14 y 30 días de datos utilizando los valores máximos.

7.- Mes: Utilizando el mes en el que esta, donde el 2 es febrero y el 12 es diciembre.

8.- Día: Utilizando el día de la semana en el que este, donde el 1 es Martes y el 6 es Domingo.

Una vez obtenidas todas las características el siguiente paso es hacer el proceso de selección. En el caso de este modelo se seleccionan las 10 características que impactan más a la distribución objetivo.

objetivo	caract_ar1	caract_mediaprom7
objetivo_t1	caract_ar2	caract_mediades7
objetivo_t2	caract_ar3	caract_mediamin7
objetivo_t3	caract_ar4	caract_mediamax7
objetivo_t4	caract_ar5	caract_mediaprom14
objetivo_t5	caract_ar6	caract_mediades14
objetivo_t6	caract_ar7	caract_mediamin14
objetivo_t7	caract_ar8	caract_mediamax14
objetivo_t8	caract_ar9	caract_mediaprom30
objetivo_t9	caract_ar10	caract_mediades30
objetivo_t10	caract_ar11	caract_mediamin30
objetivo_t11	caract_ar12	caract_mediamax30
objetivo_t12	caract_ar13	mes1
objetivo_t13	caract_ar14	mes2
objetivo_t14	caract_ar15	mes3
objetivo_t15	caract_ar16	mes4
objetivo_t16	caract_ar17	mes5
objetivo_t17	caract_ar18	mes6
objetivo_t18	caract_ar19	mes7
objetivo_t19	caract_ar20	mes8
objetivo_t20	caract_ar21	mes9
objetivo_t21	caract_ar22	mes10
objetivo_t22	caract_ar23	mes11
objetivo_t23	caract_ar24	mes12
objetivo_t24	caract_ar25	dia1
objetivo_t25	caract_ar26	dia2
objetivo_t26	caract_ar27	dia3
objetivo_t27	caract_ar28	dia4
objetivo_t28	caract_ar29	dia5
objetivo_t29	caract_ar30	dia6
objetivo_t30		

Tabla 5.1: Características calculadas.

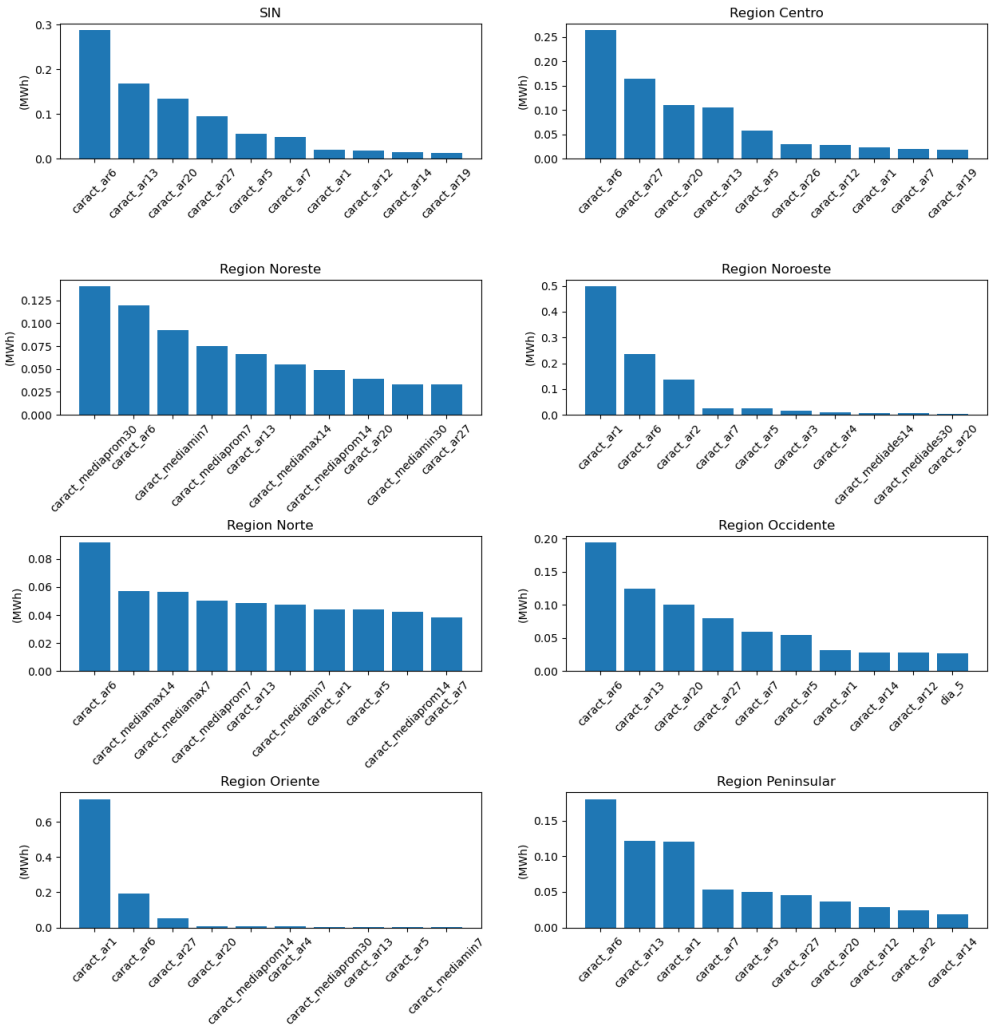

Figura 5.2: Análisis de la importancia de las características

En la figura 5.2 se pueden observar las 10 características que más impactan al modelo. Ahora, el siguiente paso es observar la relación que existe entre las mismas características en sí. Para lograr esto se hará una multiplicación matricial de $A * A^t$, después se normalizara la matriz resultante y en el resultado final se verán cuales características tienen más relación entre ellas y cuáles no.

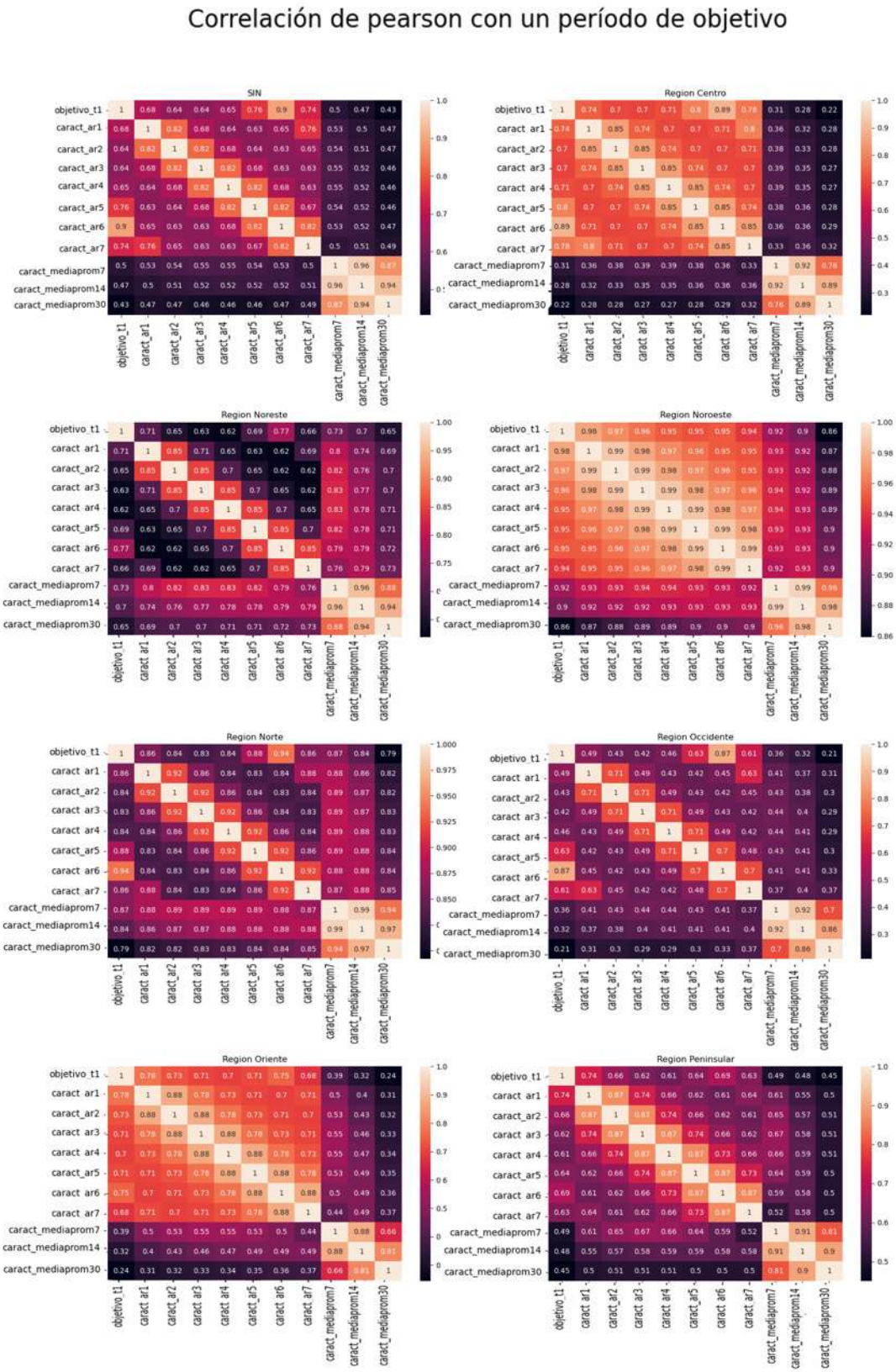

Figura 5.3: Correlación de Pearson con un periodo de objetivo

En la figura 5.3 se muestra el resultado de la multiplicación matricial y el valor del coeficiente de correlación de Pearson para cada uno de ellos, donde un color más claro significa que las características que lo conforman están fuertemente relacionadas, mientras que las que tienen un color obscuro significan que están débilmente relacionadas.

Una vez haciendo este proceso se volverán a dividir las características, reduciendo en las 5 características más importantes y más relacionadas. Para hacer esto se deben ver cuáles características tienen una mayor cantidad de coeficientes de correlación altos.

5.2. Separación de datos

El siguiente paso para armar el modelo es realizar la separación de datos. Es recomendable utilizar más del 60 % de datos para el entrenamiento del modelo, y los datos restantes para comparar el resultado obtenido de la predicción. En este proyecto se utilizan 3/4 (75 %) de datos para el entrenamiento y 1/4 (25 %) de datos para la prueba de la predicción.

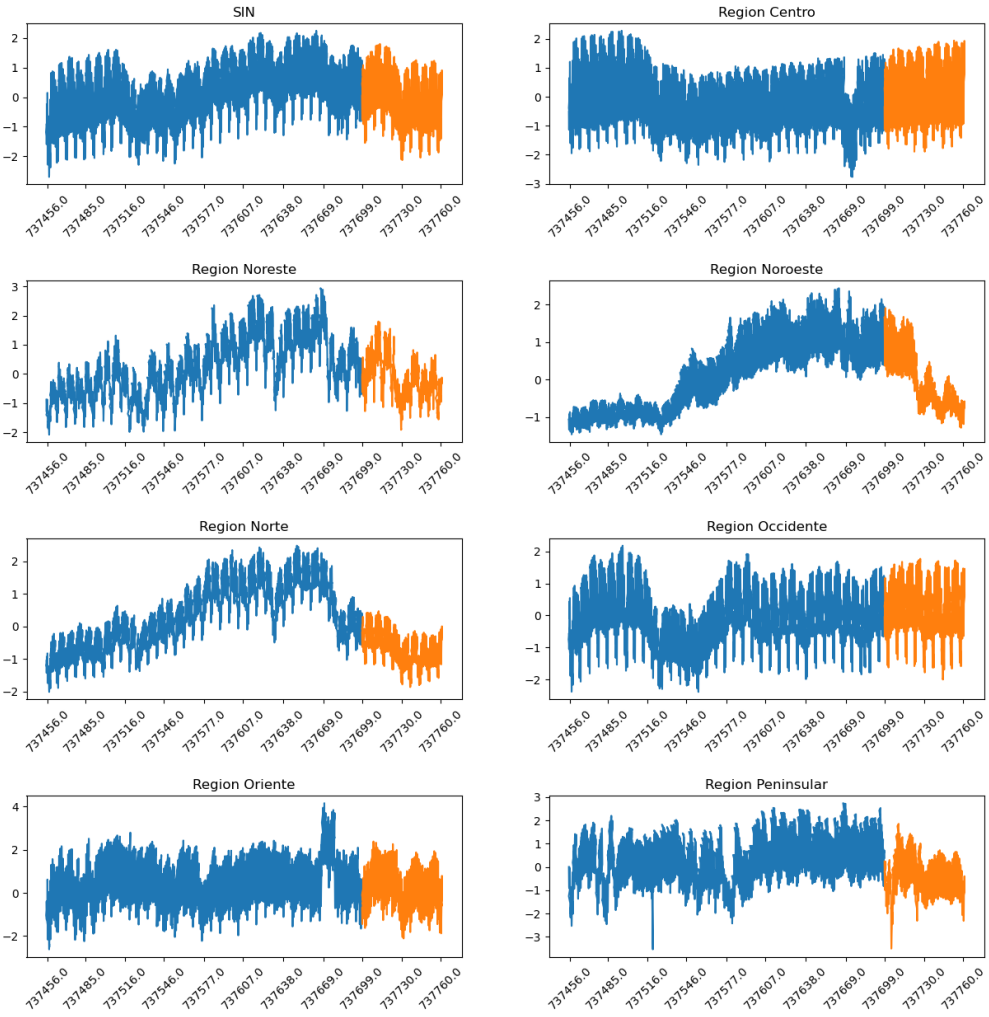

Figura 5.4: Separación de los datos de entrenamiento y prueba

En la figura 5.4 se pueden apreciar los datos que se utilizan para el entrenamiento (azul) y para la prueba de la predicción (naranja). Cabe señalar que se recomienda que los datos utilizados para la prueba tengan un comportamiento similar, por que no tendrá sentido entrenar con datos muy alejados a los utilizados para la predicción.

5.3. Comentarios finales

En este capítulo se explica el proceso de la creación del modelo de predicción de bosques aleatorios desde cero. Se dividió el comportamiento del modelo en diferentes variables (características), pero de todas estas se hizo un proceso cauteloso de elección de características para elegir las que están mas relacionadas a la variable principal y fuertemente correlacionadas a las demás características, de esta manera se asegura que en el resultado de la predicción tiene las mejores características de todas las creadas y que no se toman características que podrían indicar el comportamiento del error o de una variable externa en el sistema. Este desempeño se vera reflejado en el resultado del modelo de predicción, lo cual se puede observar en los capítulos próximos.

Capítulo 6

Predicción del modelo

Este capítulo puede considerarse el más importante de todos, ya que a fin de cuentas lo que más importa es el resultado del modelo para ver si este es satisfactorio o no. Para tener un punto de comparación, también se obtienen los datos de predicción por medio del modelo de regresión lineal, el cual puede considerarse como uno de los modelos de "predicción" mas sencillos de utilizar, el cual es obtener la recta que mejor se acople a los datos. Una vez obtenidos los datos, se consiguen los errores, los cuales se analizan y comparan, finalmente se tratan de mejorar dichos errores utilizando el mismo modelo de bosques aleatorios pero ahora con la función de regresión lineal múltiple.

6.1. Resultado obtenido con regresión lineal y bosques aleatorios

Para tener un punto de referencia de si el error obtenido es bueno o malo, se utilizo el método de regresión lineal y se utiliza para comparar su resultado con los datos de entrenamiento y los de prueba.

	Entrenamiento	Prueba
SIN	0.2294	0.6569
Región Centro	0.2568	0.3415
Región Noreste	0.3699	0.5221
Región Noroeste	0.1485	0.7472
Región Norte	0.1855	0.2569
Región Occidente	0.2711	0.9248
Región Oriente	0.3947	1.4019
Región Peninsular	0.4288	0.6122

Tabla 6.1: RMSE obtenido utilizando regresión lineal.

El resultado del error RMSE obtenido utilizando regresión lineal se puede observar en la tabla 6.1. Éste se obtuvo para compararlo con los valores obtenidos de la predicción realizada para bosques aleatorios. Teniendo este error como punto de partida se puede proceder a iniciar el entrenamiento del modelo. Esto puede tardar algo de tiempo dependiendo de que tan rápido es el procesador de la computadora, pero a diferencia del modelo de redes neuronales, los bosques pueden llegar a ser igual de potentes que algunos modelos de redes neuronales.

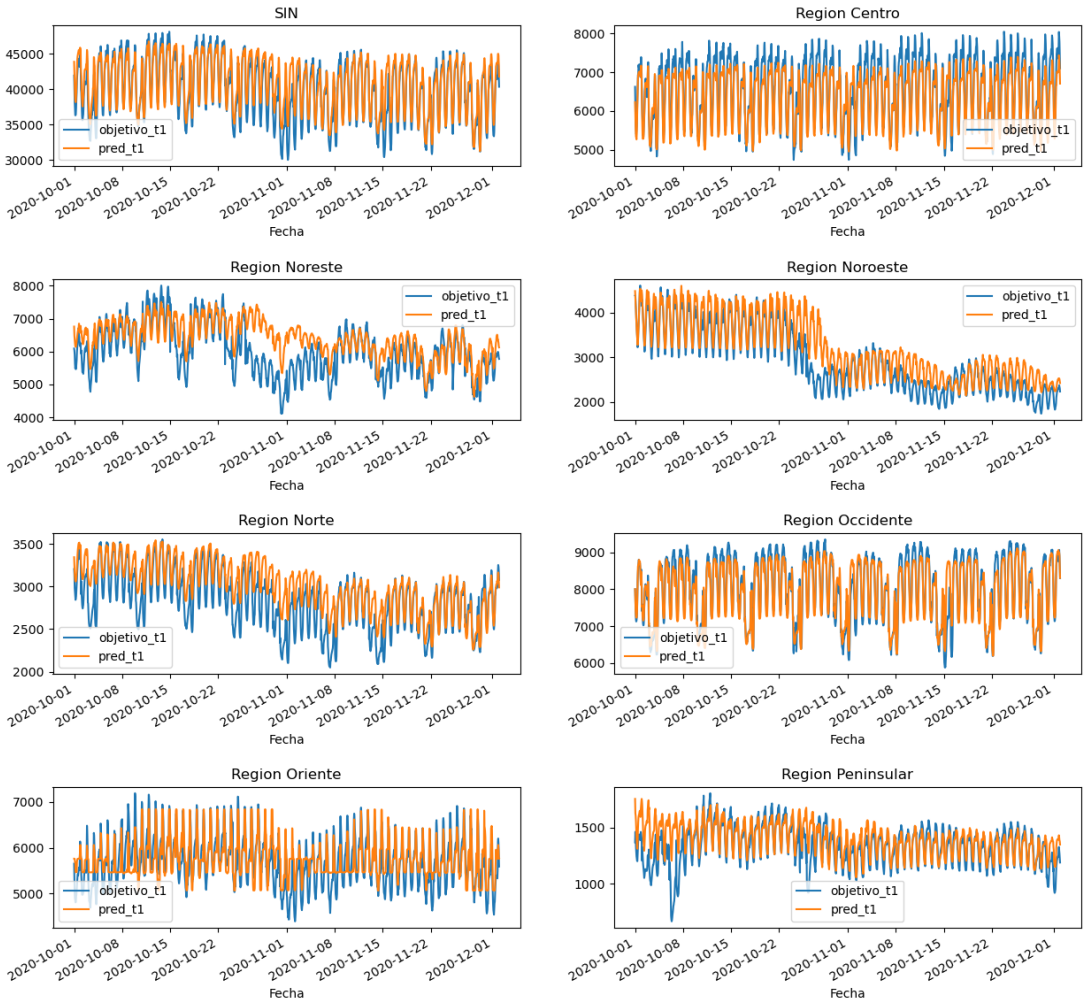

Figura 6.1: Pronostico de 90 días adelante

En la figura 6.1 se puede apreciar de manera visual la comparación de los datos pronosticados y los datos reales. A diferencia de la regresión lineal, la mayoría de los datos obtenidos tienen el mismo valor que los datos reales o uno muy aproximado.

En la tabla 6.3 se aprecia el resultado obtenido del cálculo del error RMSE, pero obtenido con el modelo de Árboles Aleatorios, se obtiene el valor del error RMSE para los datos de entrenamiento y para los datos de prueba.

	MAPE
SIN	2.1323 %
Región Centro	2.9751 %
Región Noreste	7.4694 %
Región Noroeste	10.2459 %
Región Norte	6.5730 %
Región Occidente	0.4856 %
Región Oriente	1.3688 %
Región Peninsular	5.9621 %

Tabla 6.2: MAPE Bosques aleatorios.

	Hiperparámetros de mejor plegado				Entrenamiento de mejor pliegue RMSE	Valor de mejor pliegue RMSE
	Estado aleatorio	No. de Estimadores	Decisiones max.	Profundidad max.		
SIN	123	500	8	3	0.1022	0.2444
Región Centro	123	500	32	3	0.1154	0.3029
Región Noreste	123	500	16	10	0.0040	0.1408
Región Noroeste	123	500	8	10	0.0077	0.1922
Región Norte	123	500	59	3	0.1297	0.0996
Región Occidente	123	500	16	30	0.0028	0.2756
Región Oriente	123	500	8	3	0.0929	0.3996
Región Peninsular	123	500	59	30	0.0021	0.7101

Tabla 6.3: RMSE de bosques aleatorios

6.2. Análisis de los resultados obtenidos

Una vez obtenidos los datos por el modelo de bosques aleatorios, estos se deben analizar para poder observar los puntos en el que este falló y si realmente el resultado obtenido es satisfactorio.

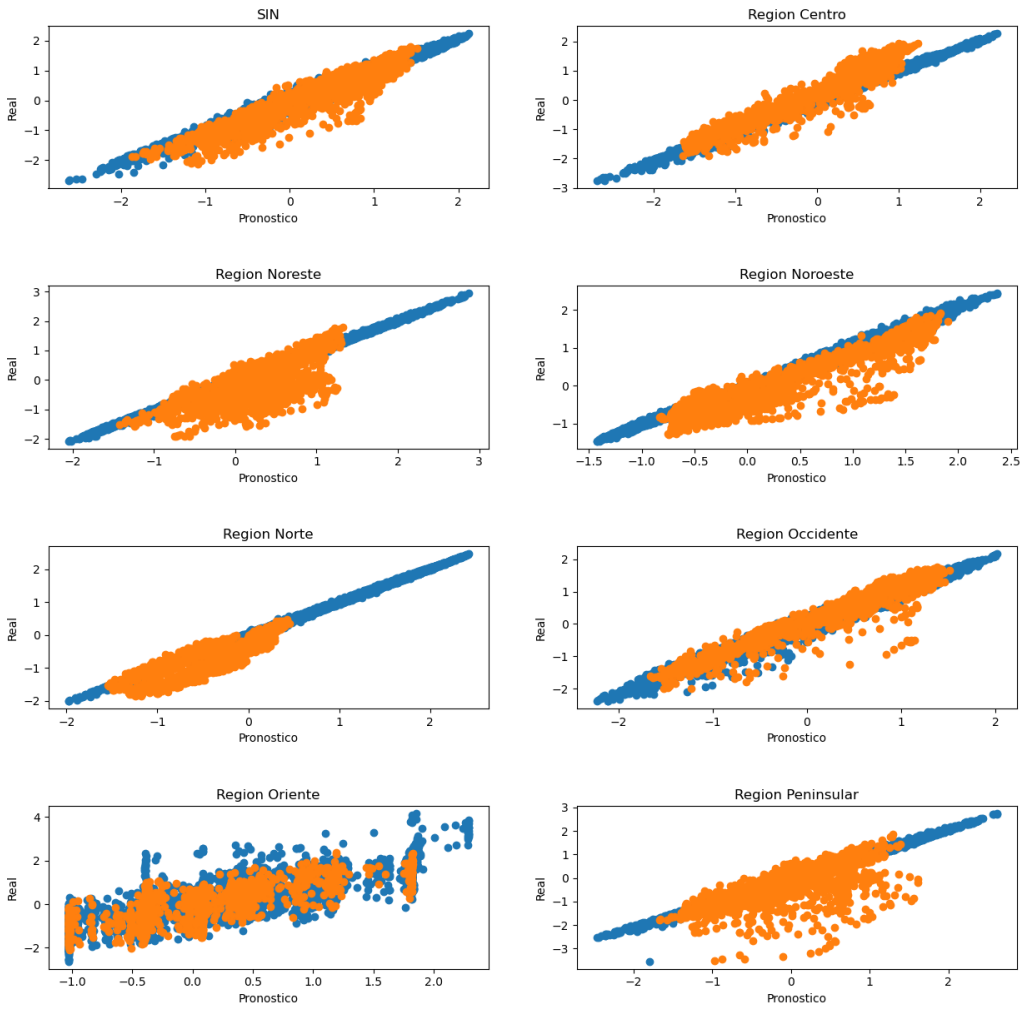

Figura 6.2: Periodo siguiente de los datos reales contra los obtenidos en el pronostico

En la figura 6.2 se pueden apreciar las gráficas de la dispersión de los datos. Para esto se grafican los datos del pronostico en el eje horizontal y los datos reales en el eje vertical, el resultado son los puntos de color azul; en cambio los puntos naranjas son los datos

obtenidos de la predicción. Se puede notar que no todos los puntos corresponden, pero sí la mayoría y que los datos están más concentrados en el centro de la gráfica.

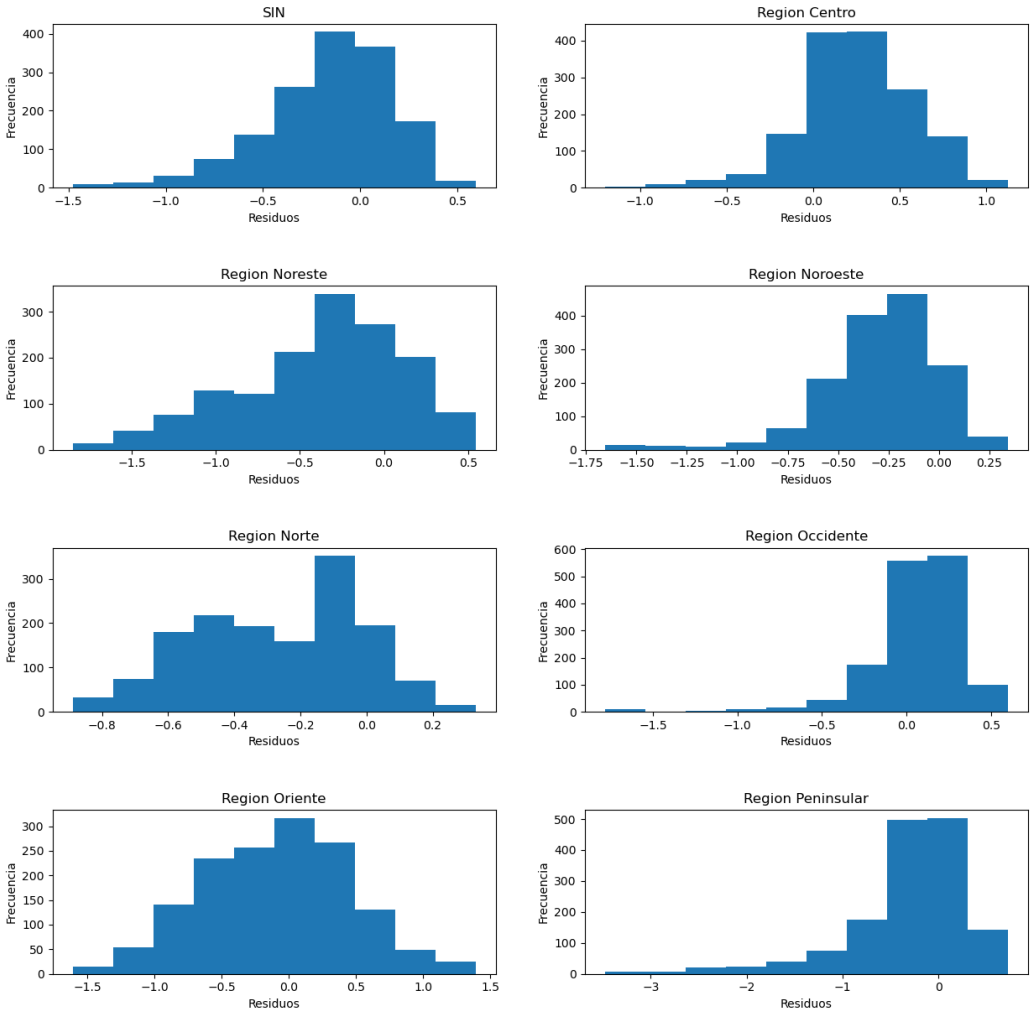

Figura 6.3: Prueba de 1 paso adelante de distribución residual

En la figura 6.3 se hace el análisis desde otro punto de vista, donde se gráfica la cantidad de datos que no encajaron totalmente (eje vertical) y que tan separado estuvo de su verdadero valor en MWh (eje horizontal)

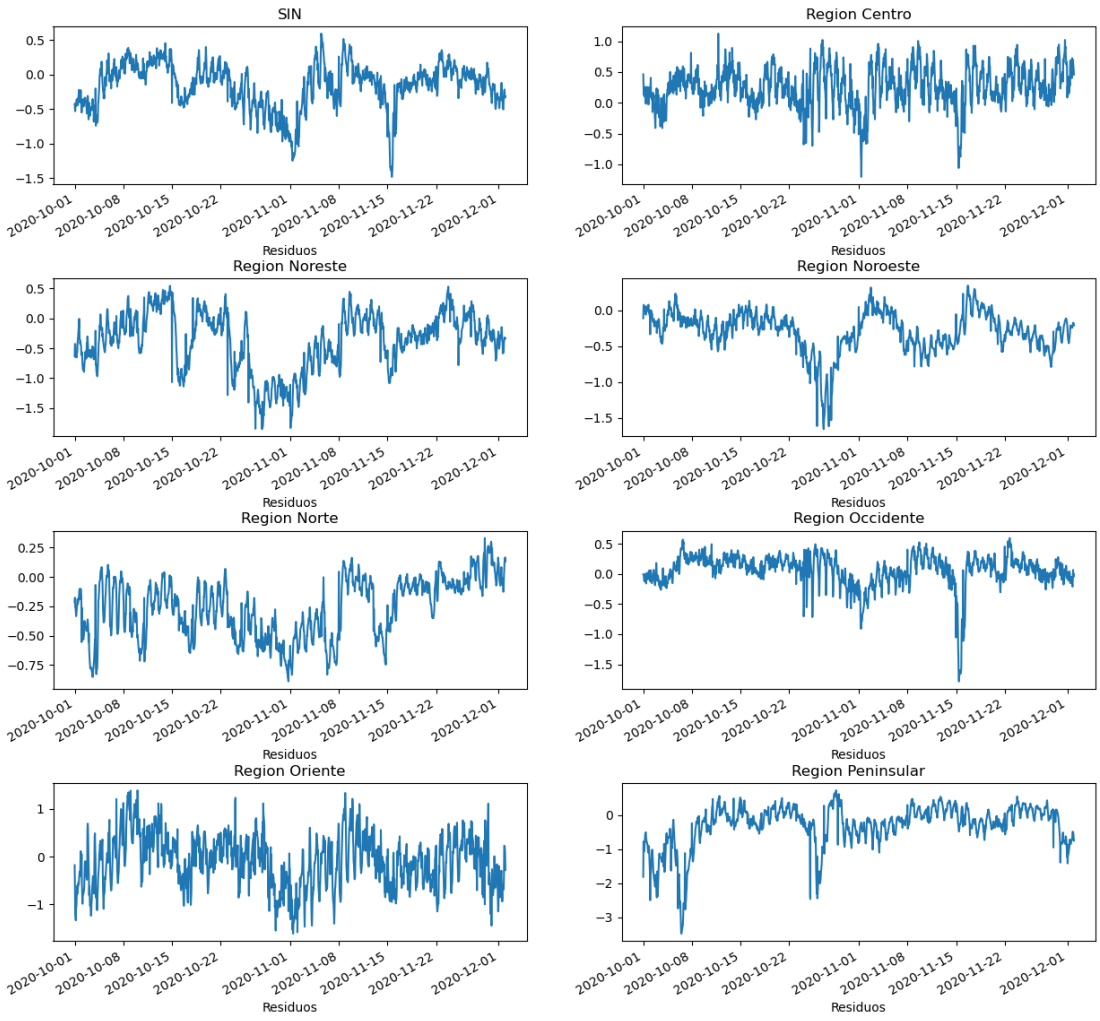

Figura 6.4: Prueba de 1 paso adelante de las series de tiempo

En la figura 6.4 se puede observar los valores "residuales" graficados, estos son los valores que no correspondieron con los datos reales. Esto se hace con el objetivo de ver la variabilidad de los errores en el tiempo; entre más baja sea significa que es más fácil es corregirlos. Se pueden reducir la cantidad de errores que se tienen, pero esta cantidad nunca llega a cero.

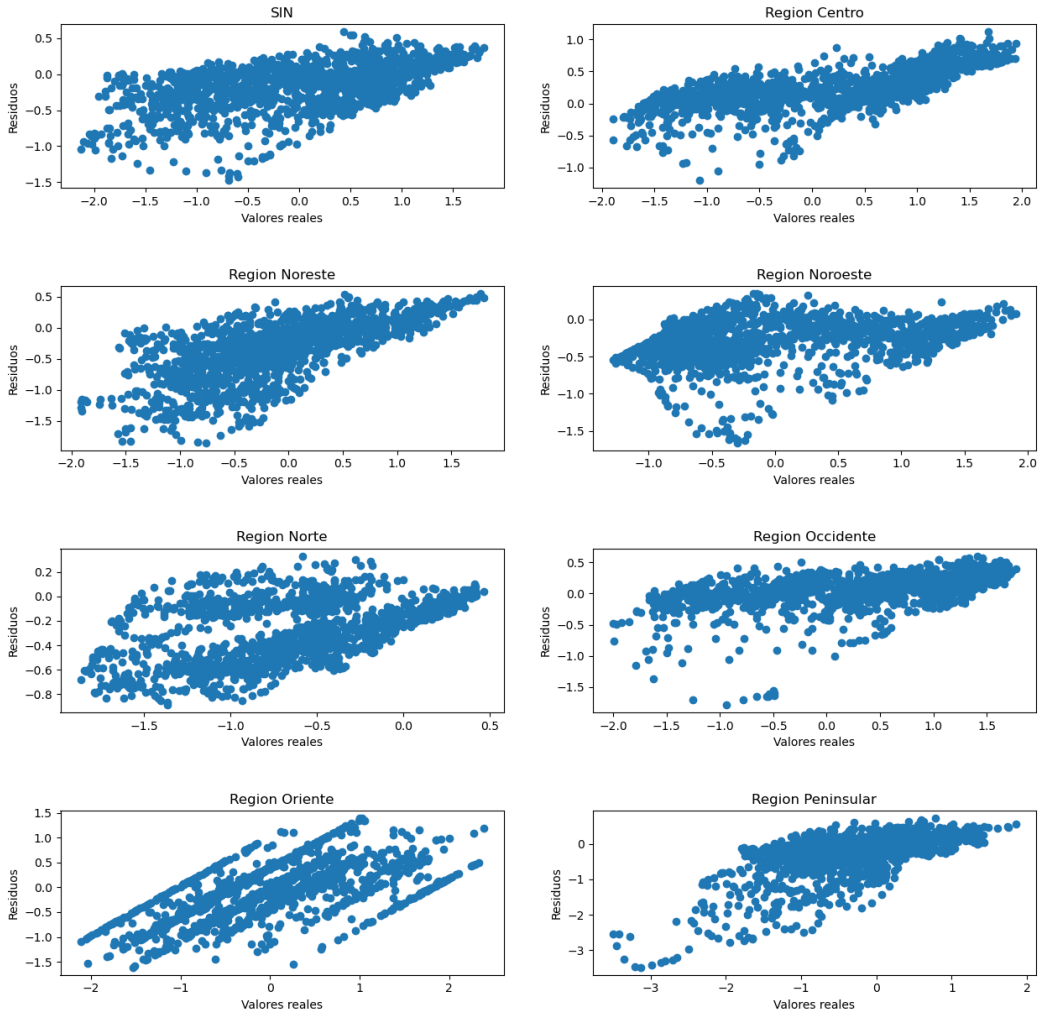

Figura 6.5: Prueba de 1 periodo de los adelante valores reales contra los valores Residuales

En la figura 6.5 se grafican los datos de los valores reales (eje horizontal) y los valores residuales (eje vertical), para observar qué tanto varían los datos el uno del otro, con el objetivo de saber si los errores están concentrados o no. Entre más concentrados están los datos erróneos es más fácil mejorar la predicción, pero no se debe de forzar la predicción, ya que con esto se puede conseguir que el modelo ocasione un sobre ajuste que solo funciona específicamente para ese rango de datos de prueba.

6.2.1. Construcción del modelo de muti-periodo próximo

"La regresión múltiple es una extensión de la regresión lineal simple. Se utiliza cuando queremos predecir el valor de una variable en función del valor de otras dos o más variables. La variable que queremos predecir se llama variable dependiente (o, a veces, variable de resultado, objetivo o criterio). Las variables que utilizamos para predecir el valor de la variable dependiente se denominan variables independientes (o, a veces, variables predictoras, explicativas o regresoras)"(Statistics, 2018).

Se utiliza la funcionalidad de regresión múltiple ya diseñada en python con la librería sklearn, la cual permite utilizarse en cualquier modelo de predicción para intentar mejorar su resultado. Esto hará que el entrenamiento del modelo sea más tardado. También cabe señalar que se hará la predicción en multi-periodo, esto quiere decir que se obtendrá una predicción para los próximos 1, 7, 14 y 30 días. Una vez utilizada esta función con el modelo de bosques aleatorios, se recalcularon los datos obtenidos y se calcularon los errores MAPE.

	MAPE			
Días	1	7	14	30
SIN	3.05	3.94	5.01	8.58
Región Centro	3.76	4.85	4.83	5.89
Región Noreste	9.48	11.35	15.89	24.71
Región Noroeste	14.47	28.52	41.30	64.54
Región Norte	6.99	10.68	13.76	20.5
Región Occidente	2.80	3.23	3.57	5.97
Región Oriente	5.03	6.25	6.46	8.76
Región Peninsular	8.89	8.38	11.19	15.36

Tabla 6.4: MAPE en el multi-periodo

En la tabla 6.4 se puede observar un vector de 4 valores correspondiente al error MAPE obtenidos para los 1, 7, 14, 30 días próximos, puede verse que conforme se va alejando de los datos de entrenamiento este error comienza a aumentarse.

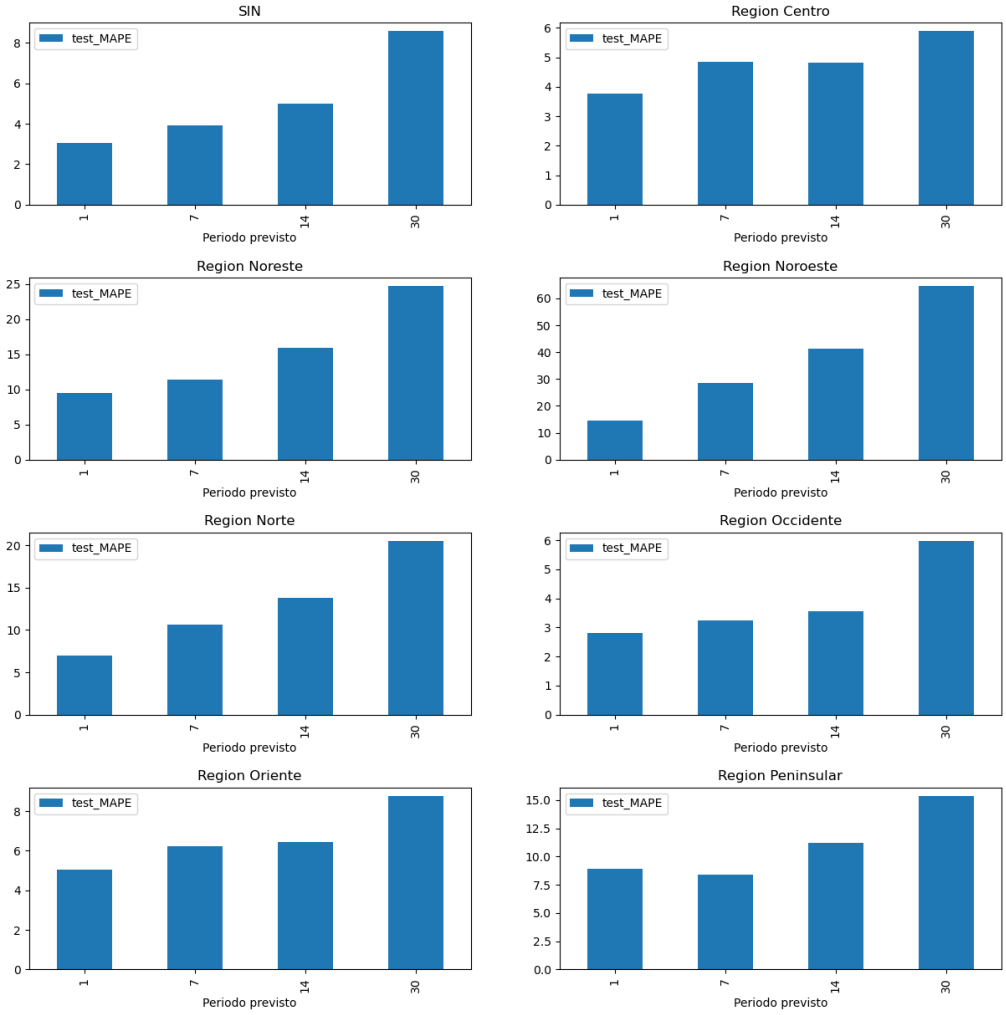

Figura 6.6: MAPE en la prueba

En la imagen 6.6 se pueden observar los mismos datos que la tabla 6.4, para ver como el error va evolucionando dependiendo de la cantidad de días que se tomen para hacer la predicción.

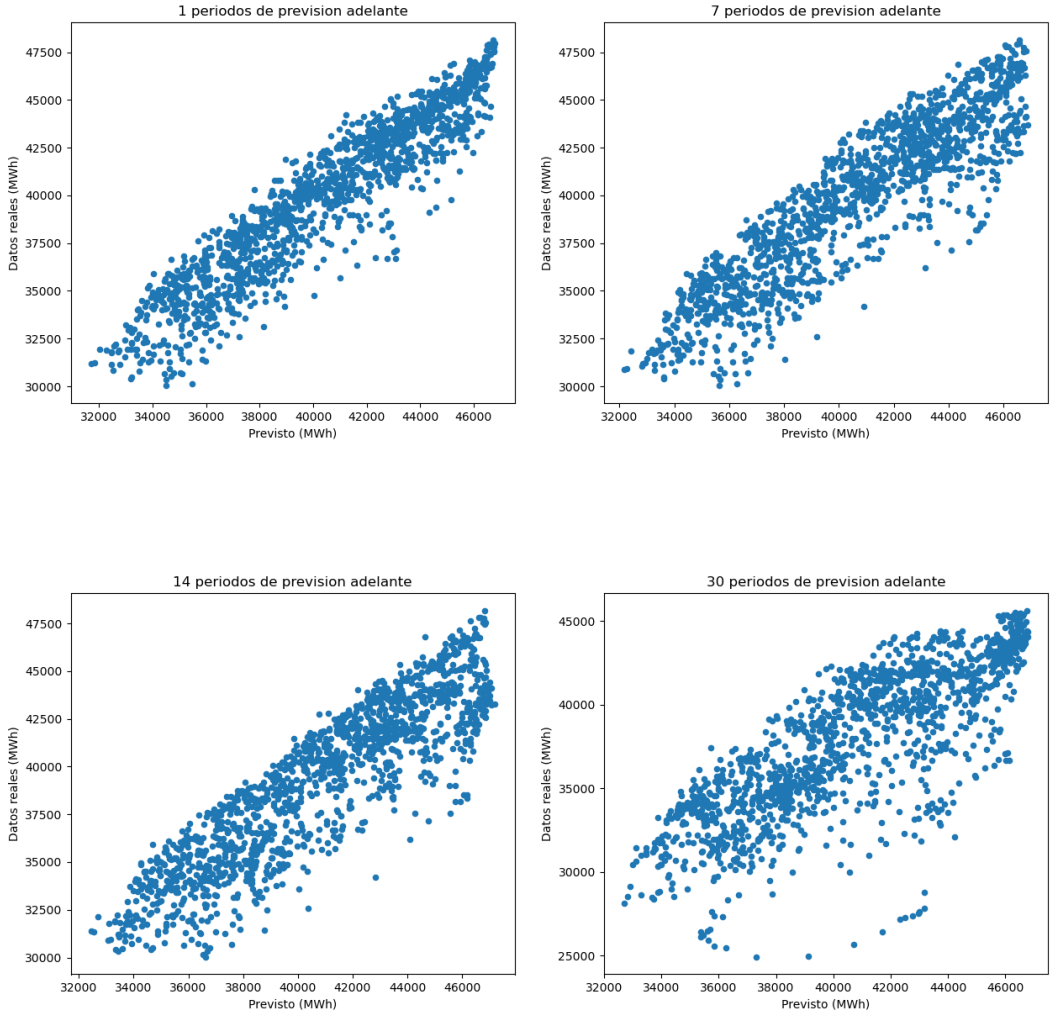

Figura 6.7: Periodos de previsión adelante SIN

En la imagen 6.7 se puede observar el resultado de graficar los datos previstos contra los datos de la demanda real de los periodos de prevision (para 1, 7, 14 , 30 dias).

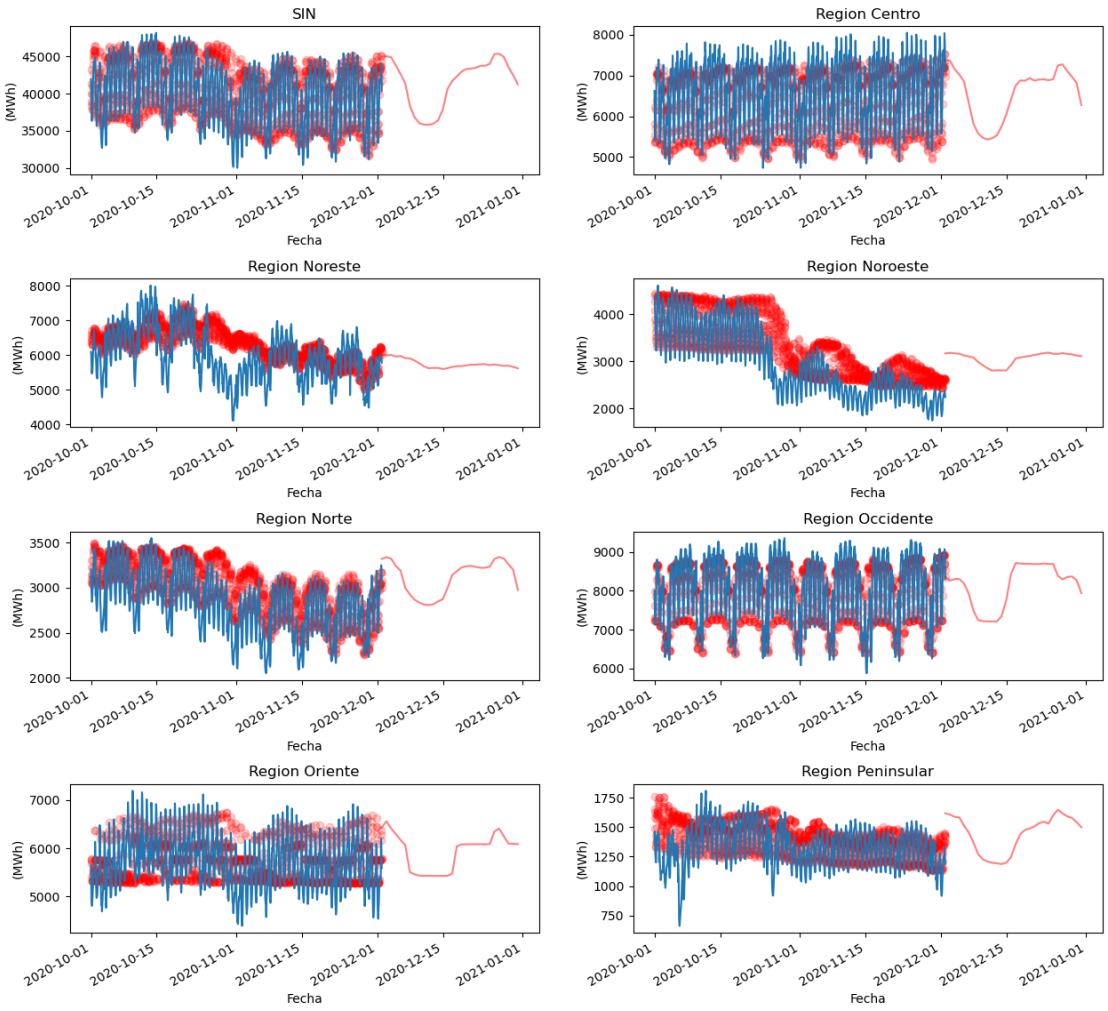

Figura 6.8: Demanda prevista utilizando regresión múltiple por hora en SIN y sus 7 áreas

En la imagen 6.8, se puede visualizar finalmente el resultado de aplicar el modelo de regresión junto con la funcionalidad de regresión lineal múltiple. La linea roja representa el resultado de la predicción mientras que la linea azul representan los datos de prueba obtenidos. Puede ser que en algunos puntos mejore la predicción y en otros empeore. Es aquí donde se debe comparar el error obtenido con regresión múltiple y sin ella para ver si realmente mejora el error o se tiene un caso típico en aprendizaje automático de sobre ajuste. Esta comparación se analizara más a fondo en la proxima sección.

6.3. Resultado del Análisis y Predicción de los datos

	Regresión Lineal	Bosques Aleatorios
SIN	0.6568	0.1021
Región Centro	0.3415	0.1154
Región Noreste	0.5220	0.0040
Región Noroeste	0.7471	0.0076
Región Norte	0.2569	0.1229
Región Occidente	0.9247	0.0028
Región Oriente	1.4019	0.0929
Región Peninsular	0.6124	0.0020

Tabla 6.5: Comparación del error RMSE obtenido en los modelos de bosques aleatorios y regresión lineal.

Se puede apreciar en la tabla 6.5 el valor del error RMSE obtenido en ambos modelos. Nótese que el valor obtenido en el modelo de predicción por bosques aleatorios es mucho menor para el SIN y sus 7 regiones en comparación con el modelo de regresión lineal, por lo que se puede decir que el modelo de predicción de bosques aleatorios logra cumplir su objetivo, el cual es predecir la demanda eléctrica con un margen de error bastante bajo. Este error podría bajar aún más con ayuda de algoritmos mas complejos, que ayudan a ajustar el modelo de predicción aún más a la curva de datos reales. Sin embargo, se debe de tener cuidado de no tener un caso de sobre ajuste a la curva. Esto pasa cuando los datos son tan fieles a la curva de entrenamiento, que cuando se prueban con datos diferentes se suele obtener un margen de error alto, por lo que no es conveniente utilizar un modelo con sobre ajuste.

	Bosques Aleatorios con multi Regresión Lineal				Bosques Aleatorios
Días	1	7	14	30	90
SIN	3.05 %	3.94 %	5.01 %	8.58 %	2.13 %
Región Centro	3.76 %	4.85 %	4.83 %	5.89 %	2.87 %
Región Noreste	9.48 %	11.35 %	15.89 %	24.71 %	7.46 %
Región Noroeste	14.47 %	28.52 %	41.3 %	64.54 %	10.24 %
Región Norte	6.99 %	10.68	13.76 %	20.5 %	6.57 %
Región Occidente	2.80 %	3.23 %	3.57 %	5.97 %	0.48 %
Región Oriente	5.03 %	6.25 %	6.46 %	8.76 %	1.36 %
Región Peninsular	8.89 %	8.38 %	11.19 %	15.36 %	5.96 %

Tabla 6.6: Comparación del error MAPE obtenido en los modelos de bosques aleatorios con y sin la funcionalidad de múltiples regresiones lineales

En la tabla 6.6 se puede ver como el modelo de predicción de bosques aleatorios con la función de multi-periodo de regresión lineal no logra cumplir su objetivo, dado que el error obtenido es mucho mayor al obtenido sin dicha función, por lo que como se mencionó, esta función solo es útil cuando una distribución de datos se comporta de forma lineal o se puede linealizar. Éste es un requisito que los datos no cuentan, por lo que es mejor dejar el modelo de predicción tal como esta, o intentar otro método para minimizar el error obtenido.

"La regresión lineal múltiple trata de ajustar modelos lineales o linealizables entre una variable dependiente y más de una variable independiente. En este tipo de modelos es importante testear la heterocedasticidad, la multicolinealidad y la especificación"(Roberto, 2016) .

6.4. Comentarios finales

Al obtener el resultado de predicción del modelo de bosques aleatorios, se calculan las métricas de error RMSE y MAPE y se comparan con los resultados del modelo de regresión

lineal, al hacer esto se puede notar una muy grande diferencia entre ambos para todas las regiones del SIN a favor del modelo de bosques aleatorios, después se hizo un análisis de manera gráfica de el comportamiento de los datos obtenidos contra los datos reales, comparando sus distribución se obtiene que en algunos puntos el error es grande pero en la mayoría de los casos este error es pequeño.

Para observar el comportamiento del error primero se gráfican los datos cuyo error esta muy alejado al valor real para saber si este error cuenta con una tendencia, en caso de ser así este podría mejorarse notablemente, una vez hecho esto se volvió a obtener el resultado mediante bosques aleatorios pero ahora con la función de regresión lineal múltiple, el resultado esperado fue contrario al esperado, dado que al obtener las métricas de error se noto que el error empeoro, esto se debe a que esta funcionalidad esta pensada para distribuciones lineales, linealizables o aquellas compuestas de múltiples lineas rectas, lo cual ninguno de los casos es la distribución utilizada en esta tesis.

Capítulo 7

Conclusiones

En este apartado se explica en resumen los resultados obtenidos de cada etapa del proyecto, mencionando los puntos importantes, haciendo comparaciones entre modelos de predicción. Al final se da un conclusión acerca del resultado obtenido con la predicción de los bosques aleatorios.

Empezando por el capítulo 3 "Análisis de Datos" se descubrió que el SIN y sus 7 regiones no poseían una distribución normal (con una forma de campana de Gauss perfecta). Podían acercarse a dicha forma, pero nunca igualarse. Esto se demostró con la prueba de normalidad de *shapiro-wilk*, donde ninguna distribución pudo aprobar dicha prueba, y la distribución normal es lo que impacta a qué tanto podrán acercarse los modelos de predicción a los valores reales, entre mas "normal" sea la distribución es mas sencillo pronosticar dicho comportamiento.

El año tomado para la predicción fue uno "especial", dado que aparte de ser un año bisiesto en este año sucedió un acontecimiento atípico, el cual es la pandemia ocasionada por el covid-19, por lo que se compararon las distribuciones obtenidas del 2020 y del 2019. En esta comparación se puede notar que en esta distribución si existe una variación, pero no a tal grado de ocasionar que se volviera "menos normal", dado que ninguna de las distribuciones del SIN en el 2019 tiene una distribución normal perfecta al igual que las distribuciones del 2020. Una explicación bastante lógica a la razón por la que los datos no son normales, es que al ser datos tomados de un sistema real es muy difícil que este se comporte de forma idónea.

Una vez observado que ninguna distribución era normal, se debía observar la linealidad de dichos datos, dado que si los datos tienen un comportamiento caótico será casi imposible predecirlos. Al hacer un análisis de volatilidad para observar la variabilidad de datos, se pudo observar que al tomar diferentes percentiles se obtuvo una curvatura similar para todo ellos, pero la amplitud era diferente. En el análisis de heterocedasticidad se demostró cómo el comportamiento de los valores de los errores no es lineal (homocedasticidad). También se analizó el coeficiente de varianza de los datos separándolos en: semanas, meses y cuartos de año, el resultado de dicho análisis fue que el coeficiente de varianza no era el mismo para cada separación pero sus valores eran muy cercanos el uno del otro. Finalizando, se realizó un análisis estacional donde se pudo observar como los datos se comportaban de forma cíclica, los resultados de este análisis nos dicen que es una distribución de datos no normal pero a pesar de esto, los datos tienen una tendencia lo cual los convierte en una distribución predecible.

Una vez obtenido el análisis de los datos se procedió a hacer la creación y el entrenamiento del modelo de predicción de bosques aleatorios. En el resultado gráfico se puede observar como la mayoría de datos corresponden a su contraparte real, pero para poder saber el poder de predicción de dicho modelo se comparó con el modelo de regresión lineal y para observar el error de manera cuántica se utilizó la distancia media cuadrática mínima (RMSE), el resultado de dicha comparación se puede notar en la tabla 6.5, en donde se puede ver una gran diferencia a favor del modelo de predicción donde en todas las distribuciones de las regiones del SIN tiene una gran ventaja y un valor de error bastante pequeño.

Se realizó un modelo de predicción un poco mas complejo utilizando como base el modelo de predicción de bosques aleatorios y añadiendo la función de regresión lineal múltiple para tratar de mejorar el margen de error obtenido, el resultado fue peor al esperado como se muestra en la tabla 6.6, donde el modelo por si solo tiene un valor de error mas pequeño, esto se debe a que esta funcionalidad es útil cuando la tendencia de los datos se puede separar en múltiples rectas, lo cual no aplica para los datos analizados.

Finalizando con el proyecto cabe señalar que el resultado obtenido fue bastante satisfactorio a pesar de que la distribución de los datos no fue normal y no haya pasado de

manera positiva todos los análisis hechos. Esto nos puede decir que el modelo de predicción de bosques aleatorios es bastante potente tomando en cuenta estos aspectos.

7.1. Trabajo a futuro

En (Romero, 2019) se utilizan 4 años de datos (1 dato por día), en mi caso se hicieron las correspondientes modificaciones para que esto fuera útil para solamente un año de datos, pero usando un dato por hora para el SIN y cada una de sus 7 regiones. De igual manera se puede partir de este proyecto para predecir datos de otro tipo o de una diferente parte del planeta. También se podrían hacer modificaciones a este trabajo para armar diferente el modelo de predicción, o utilizar otro para compararlo con el actual.También se puede utilizar un algoritmo para mejorar el error obtenido sin ocasionar un sobre ajuste, o incluso tomar un diferente año para predecir, estas son solo algunas de las posibilidades de expansión de las ramas de este proyecto.

Referencias

CENACE. (2021). *Estimación de la demanda real.* (https://www.cenace.gob.mx/Paginas/SIM/Reportes/EstimacionDemandaReal.aspx and https://www.gob.mx/cenace/que-hacemos [Online; accessed 13-June-2021])

Chen, J. (2021). *Volatility.* https://www.investopedia.com/terms/v/volatility.asp. ([Online; accessed 13-June-2021])

Fatemeh Nargesian, U. K., Horst Samulowitz, y cols. (2017). Learning feature engineering for classification. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 2529.

Francisco, L. J. (2019). *Medidas de dispersión.* https://economipedia.com/definiciones/medidas-de-dispersion.html. ([Online; accessed 12-July-2021])

Kurtis, P. (2020). *Random forest overview.* https://towardsdatascience.com/random-forest-overview-746e7983316. ([Online; accessed 12-July-2021])

McKinney, W., y cols. (2011). pandas: a foundational python library for data analysis and statistics. *Python for High Performance and Scientific Computing*, *14*(9), 1–9.

Pandas Development Team. (2021). *Dataframe.* https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.html. ([Online; accessed 13-June-2021])

Pedrosa, P. S. J. (2017). *Heterocedasticidad.* https://economipedia.com/definiciones/heterocedasticidad.html. ([Online; accessed 13-June-2021])

Roberto, M. G. (2016). Modelos de regresión lineal múltiple. *Documentos de Trabajo en Economía Aplicada*, 1.

Romero, M. A. (2019). *Machine learning applied to forecasting: Daily electricity production in spain 2014-2018.* https://www.kaggle.com/manualrg/daily-electricity

-demand-forecast-machine-learning. ([Online; accessed 26-Mayo-2021])

Statistics, L. (2018). *Multiple regression analysis using spss statistics.* https://statistics.laerd.com/spss-tutorials/multiple-regression-using-spss-statistics.php. ([Online; accessed 12-July-2021])