Revista de investigación de sistemas e informática
17(2): 33-44 (2024)
DOI: https://doi.org/10.15381/risi.v17i2.29154

ISSN: 1815-0268 (Impreso) / ISSN: 1816-3823 (Electrónico)
Facultad de Ingeniería de Sistemas e Informática - UNMSM

Uso de herramientas de balanceo de clases para la mejora del desempeño de modelos de clasificación en la estimación de la hora de la demanda eléctrica máxima

Use of class balancing tools to improve the performance of classification models in estimating the time of the maximum electrical demand

César Aristóteles Yajure Ramírez
https://orcid.org/0000-0002-3813-7606
cyajure@gmail.com
Universidad Central de Venezuela. Caracas, Venezuela

Recibido: 30/10/2024 - Aceptado: 22/11/2024 - Publicado: 31/12/2024

RESUMEN

La estimación de la hora de ocurrencia de la demanda eléctrica máxima es útil para establecer la magnitud de la generación eléctrica requerida para satisfacer dicha demanda, para establecer tarifas diferenciadas con el fin de "aplanar" la curva de carga, entre otras razones. Esta hora de ocurrencia de la demanda podría pertenecer al horario diurno o al horario nocturno, con un claro desequilibrio hacia este último. El objetivo de esta investigación es mejorar el desempeño de los modelos de pronóstico de la hora de ocurrencia de demanda eléctrica máxima al aplicar herramientas para el balance de clases. Se trabajó con los datos históricos de un país suramericano del período 2021-2024, y se utilizaron los algoritmos de máquina de soporte vectorial y regresión logística para generar los modelos de clasificación. Los métodos de balanceo de clases considerados fueron SMOTE, SMOTE-NC, así como el argumento de ajuste de los pesos de las clases de los propios algoritmos de aprendizaje automático. Para cada algoritmo de generaron cuatro modelos: uno con las clases desbalanceadas, otros con las clases balanceadas con el argumento de ajuste de los pesos, otro utilizando el método SMOTE-NC, y el cuarto utilizando el método SMOTE. Como resultado se obtuvo que los modelos en los que estuvo presente el método SMOTE-NC tuvieron las mayores mejoras de sus métricas de desempeño, las cuáles fueron: Exactitud, Precisión, F1, y Recordatorio.

Palabras clave: Clases desbalanceadas, máquina de soporte vectorial, regresión logística, sobre-muestreo, sub-muestreo.

© Los autores. Este artículo es publicado por la Revista de investigación de sistemas e informática de la Facultad de Ingeniería de Sistemas e Informática de la Universidad Nacional Mayor de San Marcos. Este es un artículo de acceso abierto, distribuido bajo los términos de la licencia Creative Commons Atribución 4.0 Internacional (CC BY 4.0) [https://creativecommons.org/licenses/by/4.0/deed.es] que permite el uso, distribución y reproducción en cualquier medio, siempre que la obra original sea debidamente citada de su fuente original. Para mas información, por favor póngase en contacto con revistarisi.fisi@unmsm.edu.pe

ABSTRACT

Estimating the time of occurrence of the maximum electrical demand is useful to establish the magnitude of the electrical generation required to satisfy it, to establish differentiated rates to "flatten" the load curve, among other reasons. This time of demand occurrence could belong to daytime or nighttime, with a clear imbalance towards the latter. The objective of this research is to improve the performance of forecast models for the time of occurrence of maximum electrical demand by applying tools for class balance. We worked with historical data from a South American country from the period 2021-2024, and support vector machine and logistic regression algorithms are used to generate the classification models. The class balancing methods considered were SMOTE, SMOTE-NC, as well as the class weight adjustment argument of the machine learning algorithms themselves. For each algorithm, four models were generated: one with unbalanced classes, others with balanced classes with the weight adjustment argument, another using the SMOTE-NC method, and the fourth using the SMOTE method. As a result, it was obtained that the models in which the SMOTE-NC method was present had the greatest improvements in their performance metrics, which were: Accuracy, Precision, F1, and Recall.

Keywords: Unbalanced classes, support vector machine, logistic regression, oversampling, undersampling.

I. INTRODUCCIÓN

Las metodologías de pronóstico de la demanda eléctrica usualmente se concentran en su magnitud, pero adicionalmente, la determinación de la hora en la que ocurren tanto su valor máximo como su valor mínimo, son muy importantes por varias razones. Por ejemplo, se requiere para efectos de determinar la generación eléctrica que debe estar disponible para satisfacer esta demanda máxima y así suavizar su pico (Xie, 2022). De igual forma, la empresa encargada del sistema eléctrico podría buscar disminuir esa demanda máxima o trasladarla temporalmente para "aplanar" la curva de carga, lo que nos lleva a la segunda razón, la cual es el establecimiento de tarifas diferenciadas para esas horas de alta demanda (Candia, 2024).

Asimismo, se ha hecho evidente la existencia de una dependencia temporal entre las practicas sociales de la población y su consumo de energía eléctrica, con algunas de estas prácticas específicas con mayor dependencia temporal que otras (Torriti, 2017). Entonces, al conocerse la hora a la que ocurre la demanda máxima, el ente regulador podría implementar políticas de uso racional y eficiente de la energía eléctrica de acuerdo con los tipos de usuarios del sistema eléctrico. Por otra parte, conocer la hora usual en la que ocurre la demanda mínima de potencia es útil para efectos de planificar la parada obligatoria de unidades de generación para efectos de mantenimiento, o incluso para planificar el mantenimiento de elementos del sistema de transmisión y/o distribución eléctrica.

Adicionalmente, la hora de ocurrencia de la demanda máxima varía dependiendo de las actividades prevalentes de la población en las distintas horas del día. Es así como podría haber un pico de la demanda en horas de la tarde, adjudicable a las actividades laborales, mientras que, podría haber un pico de demanda en horas de la noche adjudicable a las actividades de la mayoría de la población en sus respectivos hogares. Este último pico de demanda es el que mayoritariamente ocurre, haciendo que los datos históricos estén normalmente desequilibrados hacia el pico de demanda nocturno.

Debido a lo anterior, el objetivo de esta investigación es mejorar el desempeño de los modelos de clasificación para la estimación de la hora de ocurrencia de la demanda máxima haciendo uso de herramientas de balanceo de clases. Los algoritmos de aprendizaje automático supervisado considerados fueron: máquina de soporte vectorial, y regresión logística. Para el balanceo de las clases se utilizaron las técnicas de sobre-muestreo SMOTE (*Synthetic Minority Over-sampling Technique*), y SMOTE-NC (*Synthetic Minority Over-sampling Technique for Nominal and Continuous*).

El tema asociado a modelos de clasificación para la estimación de la hora de ocurrencia de la demanda máxima ha sido ampliamente investigado previamente, pero usualmente sin considerar los métodos de balanceo de clases. Asimismo, el tema de uso de las herramientas de balanceo de clases para mejorar los modelos de clasificación también ha sido desarrollado previamente, pero no aplicado a la estimación de la hora de ocurrencia de la demanda eléctrica máxima. Por ejemplo, Fu et al. (2022) desarrollan distintos modelos, uno para predecir si el día siguiente será el día de máxima demanda del mes, y otro para predecir la hora de demanda pico. Hacen uso de algoritmos de aprendizaje automático para el desarrollo de los modelos, y consideran las temperaturas máxima y mínima, entre otras, como variables explicativas de los modelos. La metodología fue aplicada al sistema de energía de Duke en Carolina del Norte Estados Unidos. De los 72 meses de datos considerados, en 69 de ellos

los modelos acertaron el día de demanda máxima, y en el 90% de los días pico, la hora real de demanda máxima estuvo entre las 2 horas con mayor probabilidad. Hacen mención que algunos modelos no generan buenos resultados debido a su inhabilidad para manejar datos desbalanceados. Asimismo, Liu & Brown (2019) construyen modelos de clasificación para predecir la hora de la demanda máxima diaria con 24 horas de antelación. Utilizan varios algoritmos de aprendizaje automático de clasificación: Näive Bayes, máquinas de soporte Vectorial (SVM), bosques aleatorios, Incremento adaptativo (AdaBoost), red neuronal de convolución (CNN), red neuronal LSTM, y red neuronal artificial del tipo codificador automático. Los datos utilizados corresponden a la demanda máxima horaria de la ciudad de Ontario en Canadá del período desde mayo 2003 hasta abril del 2008, diferenciando el período de invierno del período de verano. Obtienen que la red neuronal artificial es la que tiene mejor desempeño tanto para el período de invierno como para el período de verano.

Además, Voronin et al. (2023) presentan un enfoque para el desarrollo de un modelo de pronóstico de las horas pico, seleccionando el modelo óptimo a partir de la aplicación de una serie de algoritmos de aprendizaje automático. Para evaluar el modelo, se trabaja con los datos de 57 regiones de Rusia correspondientes al período que va desde enero del 2016 hasta enero del 2020. Los algoritmos considerados fueron: clasificador de bosque aleatorio (RFC), clasificador de árbol de decisión (DTC), clasificador de vecinos más cercanos (K-NN), clasificador de extra-árboles (ETC). La evaluación del desempeño del modelo se realiza sobre la base de la exactitud de la hora pico real con respecto a la hora pico pronosticada, tomando en cuenta intervalos de una, dos, y tres horas de las horas pico más probables. La exactitud más alta se obtuvo utilizando el clasificador de extra-árboles. En su investigación, Wongvorachan, He, & Bulut (2023) desarrollan una comparación de las técnicas de balanceo de clases, tanto de sub-muestreo como de sobre-muestreo, para tratar el desbalanceo de clases en modelos de clasificación en aplicaciones educativas. Utilizan el algoritmo de bosques aleatorios para generar los modelos de clasificación, y evaluar las técnicas de balanceo consideradas. Los resultados obtenidos muestran que para datos moderadamente desbalanceados las técnicas de sobre-muestreo aleatorio son las más convenientes, mientras que para datos extremadamente desbalanceados las técnicas híbridas (SMOTE-NC combinada con sub-muestreo aleatorio) arrojan un mejor desempeño. Adicionalmente, en (Bovornkeeratiroj et al., 2022) presentan una herramienta de código abierto para el pronóstico de energía eléctrica a través del cual se implementa una variedad de métodos de pronóstico de la demanda máxima, así como su tiempo de ocurrencia. Uno de los casos de estudio presentados consistió en pronosticar las horas del día o días en un mes o año, en las cuáles ocurre la demanda máxima. Trabajaron con los datos de consumo de electricidad, con resolución de 5 minutos, de la región de Nueva Inglaterra en los Estados Unidos del año 2020. Resultó que el modelo obtenido con una red neuronal artificial del tipo LSTM tuvo mejor desempeño, con valores para las métricas precisión, *recall*, y exactitud, de 0,84, 0,84, y 0,83, respectivamente. Mencionan que utilizan herramientas para el remuestreo de conjunto de datos desbalanceados. De igual forma, Dube & Verster (2023) mejoran el desempeño de los modelos de clasificación para el pronóstico de riesgo de créditos financieros, utilizando técnicas de balanceo de datos, tanto de sub-muestreo como de sobre-muestreo. Entre sus resultados obtienen que los modelos derivados de los algoritmos de bosques aleatorios y árboles de decisión se muestran robustos frente a clases desbalanceadas mejorando sus desempeños al aplicar las técnicas de balanceo de clases, y los modelos provenientes de los algoritmos de regresión logística y *Naive Bayes* no se presentan tan robustos frente a las clases desbalanceadas. Finalmente, Donaldson et al. (2023) presentan una metodología para pronosticar la magnitud de la demanda máxima, así como su hora de ocurrencia, utilizando dos algoritmos de aprendizaje automático: Regresión Lineal Múltiple y Máquina de Incremento de Gradiente. Entre las variables explicativas consideran la hora, el día de la semana, el mes, feriados, y temperatura. Los resultados indican que la regresión tiene mejor desempeño durante las temporadas con baja variabilidad horaria, mientras que los métodos de conjunto muestran una mayor exactitud en general.

El resto del artículo se distribuye de la siguiente manera. En la sección 2 se muestra la metodología empleada en el estudio. En la sección 3 se presenta la discusión y análisis de resultados. Seguidamente, en la sección 4 se plantean las conclusiones que se derivan del trabajo realizado. Finalmente, se listan las referencias bibliográficas utilizadas durante el desarrollo de la investigación.

II. MATERIALES Y MÉTODOS

Para desarrollar la investigación, se siguieron las etapas que conforman un proyecto típico de ciencia de datos. Estas etapas son: establecimiento del o de los objetivos de proyecto, adquisición de los

datos a utilizar, procesamiento de los datos, análisis exploratorio de los datos, modelación de los datos, y la toma de decisiones (Cielen et al., 2016). En la Figura 1 se presenta un esquema de las etapas seguidas en la metodología, en la que se incluye explícitamente una etapa de balanceo de datos, y una etapa de evaluación de los modelos.

Si nos circunscribimos al enfoque de ciencia de datos, el objetivo sería determinar la hora de ocurrencia de la demanda máxima de una zona geográfica. Los datos se consiguen de fuentes externas y/o internas, en este caso corresponden a las mediciones de la demanda eléctrica máxima y mínima, y sus horas de ocurrencia, para la zona bajo estudio. Luego de adquirir los datos, a éstos usualmente se les debe limpiar y procesar, aplicando las técnicas descritas en (Mukhiya & Ahmed, 2020). Para el análisis exploratorio se utilizan técnicas de la estadística descriptiva gráficas y no gráficas, y de este análisis ya se podría obtener información significativa de los datos. Luego se hace la modelación utilizando algoritmos de aprendizaje automático. Específicamente se utilizan los algoritmos de clasificación para pronosticar la hora a la que ocurrirá la demanda máxima en los próximos días. Con los resultados obtenidos en las dos etapas previas, se procede a la de toma de decisiones. Las etapas mencionadas hasta ahora, correspondientes a los bloques azules de la Figura 1, son las típicas de un proyecto de ciencia de datos.

En esta investigación se incorpora una etapa de balanceo, ya que los datos utilizados son desbalanceados y esa característica afecta el resultado arrojado por los modelos. Asimismo, se incorpora la etapa de evaluación de los modelos, en los que se aplican las métricas de desempeño, y se evalúa el efecto de las técnicas de balanceo. Todas las etapas recién presentadas son desarrolladas utilizando el lenguaje de programación Python y sus respectivas librerías.

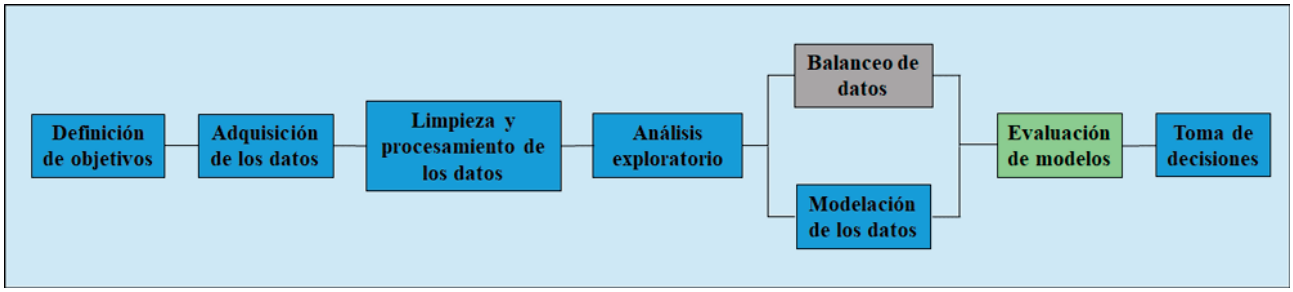
TÉCNICAS DE BALANCEO DE CLASES

Se dice que un conjunto de datos está desequilibrado si la clase en la que estamos interesados cae en la clase minoritaria y aparece escasamente en comparación con la clase mayoritaria, la clase minoritaria también se conoce como clase positiva, mientras que la clase mayoritaria también se conoce como clase negativa (Ebenezer et al., 2021). Por otra parte, el desequilibrio de clases se describe como una gran discrepancia entre dos clases de la misma variable objetivo, donde una clase está representada por muchas instancias, mientras que la otra solo está representada por un pequeño número de casos (Wongvorachan et al., 2023).

Se pueden considerar dos estrategias para contrarrestar el desequilibrio de las clases: el enfoque del muestreo de los datos, y el enfoque de aprendizaje sensible a los costos. En esta investigación se trabajó con el primer enfoque, el cual se podría abordar utilizando un método de sub-muestreo o un método de sobre-muestreo. La metodología de muestreo que selectivamente despoja la clase mayoritaria, mientras asegura que el conjunto de datos retenga la información significativa asociada a esta clase mayoritaria, se conoce como enfoque de sub-muestreo. Mientras que el enfoque de muestreo en la que las instancias de la clase minoritaria son frecuentemente replicadas hasta que se alcanza una distribución de clases balanceadas, se conoce como enfoque de sobre-muestreo. En este trabajo se utilizan los métodos de sobre-muestreo, entre los que destacan el método SMOTE, el método SMOTE-NC, entre otros. Poddar et al. (2024) plantean que el método SMOTE trabaja bien cuando el conjunto de datos es de baja dimensión, y que además todas las variables deben ser numéricas, por lo que las de tipo categóricas deben ser procesadas. Asimismo, indican que el método SMOTE-NC tiene un mejor desempeño cuando hay

Figura 1

Metodología utilizada



Elaboración propia.

variables tanto categóricas como numéricas, y no solamente numéricas.

Además, los algoritmos de clasificación presentes en las librerías de Python cuentan con un argumento que permite hacer el balanceo de las clases, ajustando los "pesos" de cada una de las clases de manera inversamente proporcional a la frecuencia de dichas clases.

Métricas para la evaluación de desempeño de modelos de clasificación

Se consideran las siguientes métricas para evaluar el desempeño de los modelos de clasificación: exactitud, precisión, recordatorio (*Recall*), y F1. De acuerdo con lo mencionado por Lee (2019, p. 170), la exactitud "es definida como la suma de todas las predicciones correctas dividida entre la suma de todas las predicciones", se obtiene utilizando (1). La precisión "está relacionada con el número de predicciones positivas correcta", y se genera a través de (2). El recordatorio "está relacionada con el número de eventos positivos predichos correctamente", se consigue haciendo uso de (3). La métrica F1 "es conocido como la media armónica entre la precisión y el recall", se computa con (4).

$$Exactitud = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precisión = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1 = \frac{2(precisión \cdot recall)}{precisión + recall} \quad (4)$$

Donde:

TP: Verdaderos positivos

TN: Verdaderos negativos

FP: Falsos positivos

FN Falsos negativos

Limpieza y procesamiento de los datos

Los datos utilizados consisten en la serie temporal histórica de la demanda eléctrica máxima horaria del período 2021-2024 de un país de Suramérica. Esta serie se procesa con el fin de generar una serie temporal con resolución diaria, que contiene la demanda máxima del día, la hora a la que ocurre esta demanda, la demanda mínima día, y la hora a la que ocurre esta demanda mínima. Además, la serie incluye el año, el mes, la semana, y el día de la semana. A esta serie temporal diaria se incorpora información de los días feriados y laborables, así como una columna para la temperatura ambiente máxima, y otra para la temperatura ambiente promedio.

Tomando en cuenta la columna de los meses del año, se crea una columna que indica si el día correspondiente pertenece al período lluvioso histórico o al período de sequía histórica, puesto que durante el período de sequía la temperatura ambiente crece, y por consiguiente la demanda eléctrica crece, afectando de alguna manera la hora de ocurrencia de esta demanda.

Finalmente, a partir de la columna de la hora de demanda máxima, se crea una columna que indica si para el día respectivo, la hora de demanda máxima pertenece a las horas de la tarde (alrededor de las 2 pm), o pertenece a las horas de la noche (alrededor de las 8 pm).

III. ANÁLISIS Y DISCUSIÓN DE RESULTADOS

En esta sección se analizan los resultados obtenidos en las etapas de análisis exploratorio, modelación de los datos, y evaluación de los modelos.

Análisis exploratorio de los datos

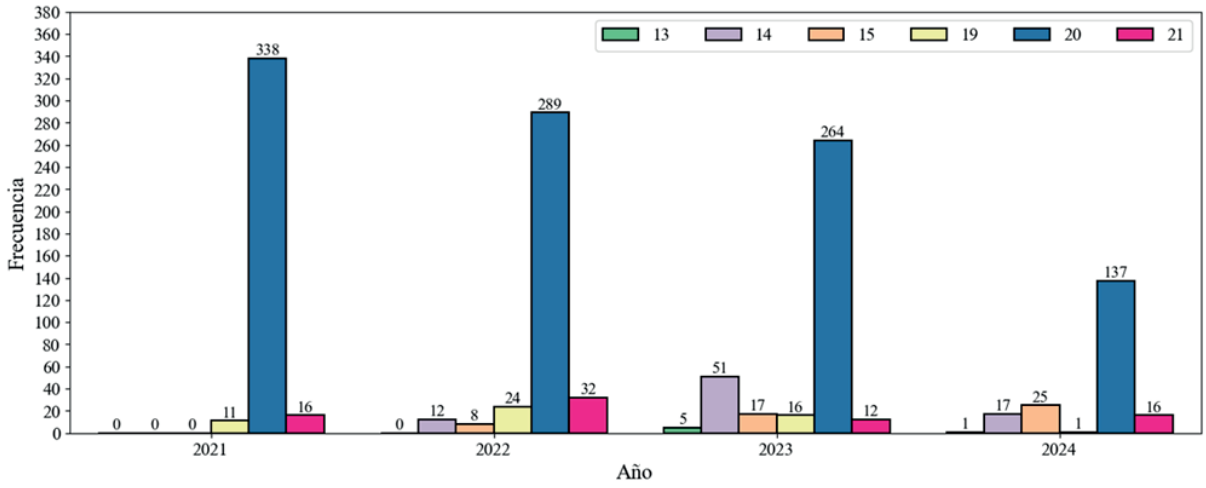
Inicialmente, se determina la distribución de las horas de demanda máxima a lo largo del período de estudio 2021-2024, la cual se presenta en la Figura 2. Se puede notar que las horas de ocurrencia de la demanda máxima incluyen las horas: 13, 14, 15, 19, 20, y 21.

De la Figura 2 se puede ver que, durante el 2021, las horas de demanda máxima fueron todas en el horario nocturno (horas 19, 20, y 21), lo cual concuerda con el hecho que en ese año hubo mayor presencia de teletrabajo debido a la pandemia. Asimismo, se puede observar que para todo el período la hora de demanda máxima fue mayoritariamente las 8 pm, y que a partir del año 2022 las horas de la tarde empezaron a tener cierta presencia en los datos de hora de demanda máxima.

Por otra parte, en la Figura 3 se presenta la distribución de las horas de demanda máxima para cada uno de los meses del año. Se observa que durante los meses del último tercio del año es cuando tiene mayor incidencia las 7 pm como hora de demanda máxima, mientras que hasta el mes de agosto su presencia es casi nula. Adicionalmente, se nota que

Figura 2

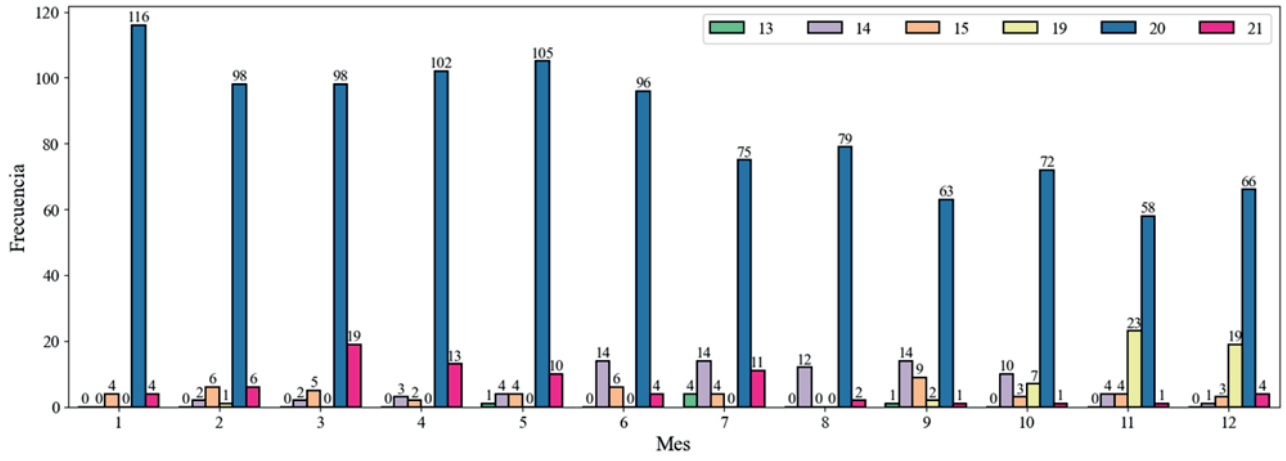
Metodología utilizada



Elaboración propia.

Figura 3

Distribución de horas de demanda máxima por año.



Elaboración propia.

las horas de demanda máxima diurnas se intensifican a partir del mes de junio, y hasta el mes de octubre.

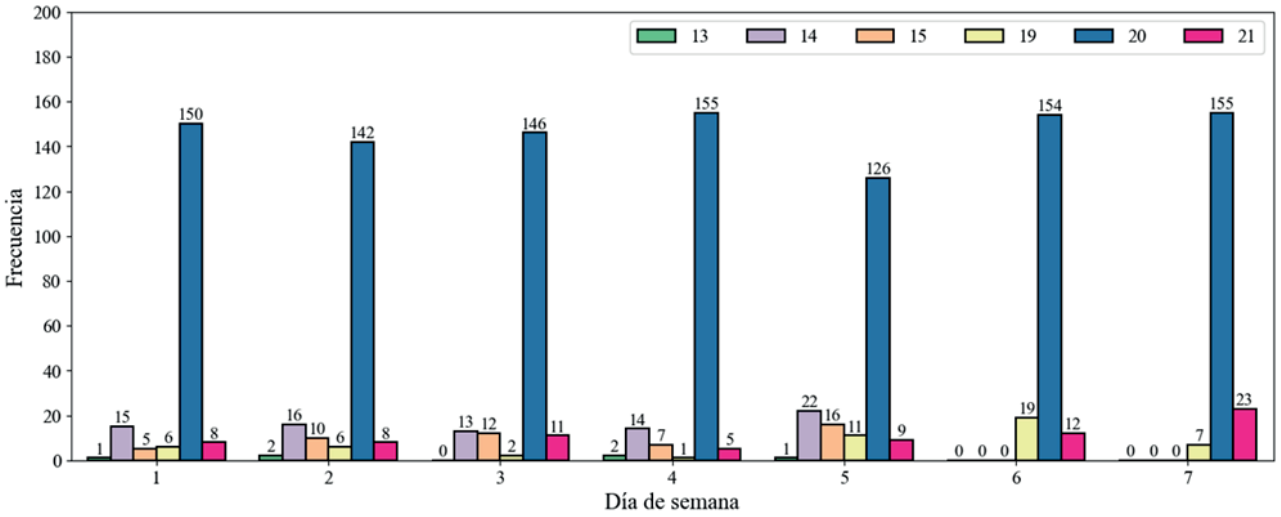
De igual manera, en la Figura 4 se presenta la distribución de las horas de demanda máxima por día de semana. En este caso, el lunes es representado por el número “1”, y el domingo se representa con el número “7”.

Se puede notar que, durante los sábados y domingos, las horas de demanda máxima corresponden sólo a horas de la noche. Por otra parte, se observa que los viernes (día 5) son los días en que hay más horas de demanda máxima durante la tarde, y además son los días en que menos horas de demanda máxima coinciden con las 8 pm.

Adicionalmente, en la Figura 5 se presenta la distribución de horas de demanda máxima de acuerdo con el tipo de día, con el fin de comparar el comportamiento de los días laborables con los feriados y los fines de semana. Se puede ver que durante los días laborables predomina las 8 pm, y un 15,4% de las horas de demanda máxima ocurren durante las tardes. El comportamiento de los días feriados es similar al de los sábados y domingos, con el 100% de las horas de demanda máxima durante las noches

Figura 4

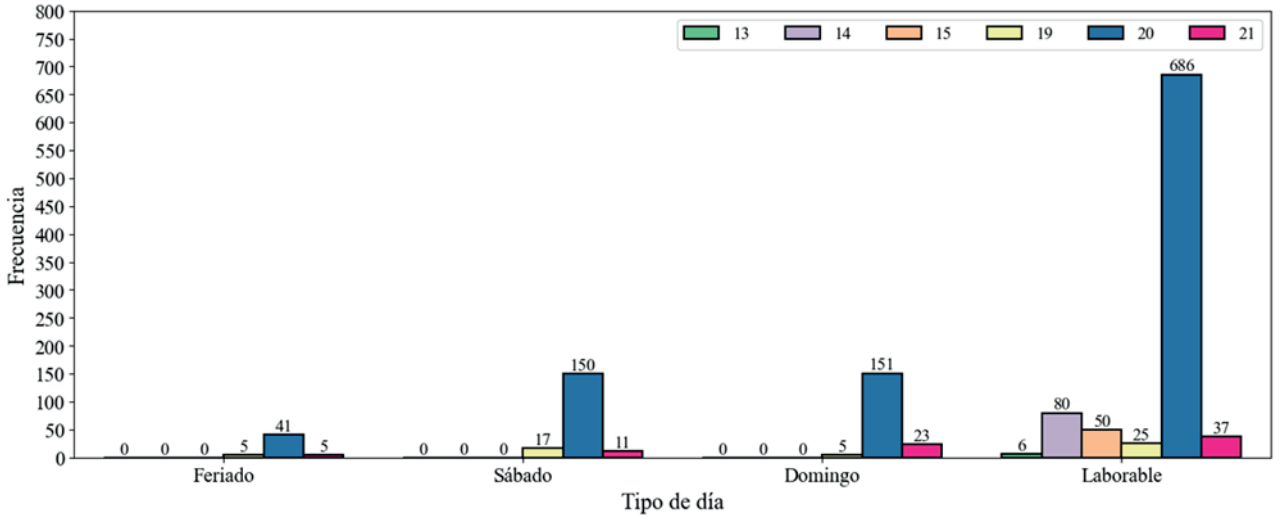
Distribución de horas de demanda máxima por día de semana



Elaboración propia.

Figura 5

Distribución de horas de demanda máxima por tipo de día



Elaboración propia.

y con una proporción de alrededor del 80% de las horas coincidiendo con las 8 pm.

Modelación de los datos

Seguidamente, se presenta la modelación de los datos para la estimación de la hora de ocurrencia de la demanda eléctrica máxima, para lo que se utilizan algoritmos de aprendizaje automático supervisado de clasificación: máquina de soporte vectorial, y regresión logística. En (Yajure Ramírez, 2023) se utilizan los mismos algoritmos de clasificación, además de K vecinos más cercanos y bosques aleatorios, para estimar la relación de desempeño de una planta solar fotovoltaica, lo cual también corresponde a un problema de clasificación binaria. Los modelos obtenidos son evaluados utilizando las métricas: Exactitud, Precisión, F1, y Recordatorio, las cuales son también utilizadas en (Liu & Brown, 2019) para evaluar los modelos de clasificación.

En esta investigación, se hace uso de los datos diarios de demanda eléctrica máxima del período 2023-2024, hasta el 14/07/2024, para generar y

entrenar los modelos, y los diecisiete días restantes hasta el 31/07/2024 se utilizan para calcular nuevamente las métricas de evaluación de desempeño de dichos modelos. Asimismo, se utilizan técnicas de balanceo de datos en combinación con cada uno de los modelos, puesto que tal como se observó en el análisis exploratorio de los datos, éstos están desbalanceados hacia las 8 pm, como hora de demanda máxima mayoritaria. Específicamente, se utilizan el método SMOTE, el método SMOTENC, el cual es conveniente cuando se tienen conjunto de datos tanto con variables numéricas como con variables categóricas, y el argumento de clases balanceadas con el cuentan los propios algoritmos.

Para cada uno de los modelos, el conjunto de datos se dividió en dos partes: entrenamiento y prueba, con una proporción de 80% para el entrenamiento y 20% para la prueba. Se consideran dos posibles horas: las 2 pm y las 8 pm, ambas con una variación de más o menos una hora.

Algoritmo de máquina de soporte vectorial (SVC)

Las máquinas de soporte vectorial se pueden utilizar tanto para problemas de clasificación como para problemas de regresión, utilizando el mismo principio de funcionamiento. Este algoritmo utiliza el concepto de *kernel* para convertir los datos dados en una dimensión superior, para así conseguir los llamados hiperplanos. Los puntos ubicados a cada lado del hiperplano y que estén más próximos a él se conocen como vectores de soporte. Hay cuatro tipos principales de *kernel*, a saber, lineal, polinómico, sigmoidea y función de base radial (Muthukrishnan & Jamila, 2020).

Para nuestro caso de estudio se utiliza el clasificador de soporte vectorial, y se consideran sus parámetros por defecto, lo que incluye un *kernel* de función de base radial. Se generan cuatro modelos utilizando este algoritmo: el primero manteniendo las clases desbalanceadas, luego balanceando las clases haciendo uso del hiperparámetro que el algoritmo posee para tal fin, un tercer modelo balanceando las clases previamente con el método SMOTE-NC, y un cuarto modelo con el balanceo previo de las clases con el método SMOTE. Con cada modelo se obtienen pronósticos a partir de los datos de prueba, y se evalúan los resultados obtenidos a través de las métricas correspondientes. Los valores de las métricas se presentan en la Tabla 1.

De la tabla 1 se puede ver que las métricas del modelo con clases desbalanceadas son mejores que las correspondientes al modelo cuando se balancean las clases utilizando el argumento de ajuste de clases del algoritmo, lo cual indica que esa técnica no mejora el desempeño del modelo con los datos de prueba. Por otra parte, las métricas de los modelos con las clases balanceadas a través de los métodos SMOTE-NC y SMOTE si mejoran al compararse con las métricas del modelo con clases desbalanceadas.

Posteriormente, estos modelos se utilizan para estimar la hora de demanda máxima para los siguientes diecisiete días (datos nuevos), desde el 15 al 31 de julio, considerando dos opciones: 2 pm y 8 pm, ambas clases con una tolerancia de 1 hora, es decir, un rango de 1 pm a 3 pm (13_14_15), y el otro rango desde las 7 pm a las 9 pm (19_20_21). Los resultados se comparan con los valores reales, y los nuevos valores de las métricas de evaluación se presentan en la Tabla 2, mientras que las horas estimadas se presentan en la Tabla 3.

De la Tabla 2 se puede observar que los valores de las métricas del modelo con clases desbalanceadas son los menores, es decir, las métricas de desempeño mejoran en cada uno de los tres modelos cuyas clases fueron balanceadas. De esos tres modelos, destaca aquel en el que las clases fueron previamente balanceadas utilizando el método SMOTE-NC.

En la Tabla 3 se ve que, utilizando el modelo con clases desbalanceadas, las horas estimadas para

Tabla 1

Métricas de desempeño de modelos SVC – Datos de prueba

Métrica	Desbalanceadas	Balanceadas	SMOTE-NC	SMOTE
Exactitud	0,752	0,717	0,776	0,782
Precisión	0,686	0,649	0,807	0,793
F1	0,695	0,644	0,769	0,779
Recordatorio	0,777	0,764	0,772	0,779

Tabla 2

Métricas de desempeño de modelos SVC – Datos nuevos

Métrica	Desbalanceadas	Balanceadas	SMOTE-NC	SMOTE
Exactitud	0,529	0,529	0,882	0,529
Precisión	0,265	0,517	0,882	0,517
F1	0,346	0,433	0,882	0,433
Recordatorio	0,500	0,507	0,882	0,507

Tabla 3

Horas estimadas modelos SVC

Fecha	Hora real	Rango real	Desbalanceadas	Balanceadas	SMOTENC	SMOTE
15-Jul-24	14	13_14_15	19_20_21	19_20_21	13_14_15	19_20_21
16-Jul-24	20	19_20_21	19_20_21	19_20_21	19_20_21	19_20_21
17-Jul-24	15	13_14_15	19_20_21	19_20_21	13_14_15	19_20_21
18-Jul-24	15	13_14_15	19_20_21	19_20_21	19_20_21	19_20_21
19-Jul-24	20	19_20_21	19_20_21	13_14_15	13_14_15	13_14_15
20-Jul-24	20	19_20_21	19_20_21	19_20_21	19_20_21	19_20_21
21-Jul-24	20	19_20_21	19_20_21	19_20_21	19_20_21	19_20_21
22-Jul-24	14	13_14_15	19_20_21	19_20_21	13_14_15	19_20_21
23-Jul-24	15	13_14_15	19_20_21	19_20_21	13_14_15	19_20_21
24-Jul-24	20	19_20_21	19_20_21	19_20_21	19_20_21	19_20_21
25-Jul-24	14	13_14_15	19_20_21	19_20_21	13_14_15	19_20_21
26-Jul-24	14	13_14_15	19_20_21	13_14_15	13_14_15	13_14_15
27-Jul-24	20	19_20_21	19_20_21	19_20_21	19_20_21	19_20_21
28-Jul-24	20	19_20_21	19_20_21	19_20_21	19_20_21	19_20_21
29-Jul-24	13	13_14_15	19_20_21	19_20_21	13_14_15	19_20_21
30-Jul-24	20	19_20_21	19_20_21	19_20_21	19_20_21	19_20_21
31-Jul-24	20	19_20_21	19_20_21	19_20_21	19_20_21	19_20_21

los últimos diecisiete días de julio es siempre las 8 pm, lo cual era de esperarse puesto que las clases están desbalanceadas hacia esa hora. También se puede observar que los resultados asociados al modelo con clases balanceadas a través del uso de SMOTE-NC, acertó la hora en quince de los diecisiete días considerados, mientras que con los otros tres modelos sólo se acertaron nueve de los diecisiete días.

Al considerar las métricas del modelo que tuvo más aciertos y compararlas con las métricas del modelo con clases desbalanceadas, para los datos nuevos, se tiene que la Exactitud mejoró alrededor de 67%, la Precisión mejoró en más de 200%, la métrica F1 mejoró en más del 150%, y el Recordatorio mejoró alrededor de 76%.

Algoritmo de regresión logística (RL)

A través de la aplicación de este algoritmo, se pueden generar modelos para clasificación binaria. Según (Kirasich et al., 2018, p. 8), esta técnica "Es uno de los modelos estadísticos lineales más utilizados para análisis discriminante". Posterior a la realización de una regresión lineal, el algoritmo convierte la salida de esta regresión a través de una función logística (de allí su nombre), que comúnmente es la función sigmoide. Esta última función asigna una probabilidad condicional para cada una de las clases. Al aplicar este algoritmo se tomaron todos los parámetros por defecto, lo que incluyó el método de optimización *lbfgs*. Tal cual se procedió con el algoritmo anterior, se generan cuatro modelos: el primero manteniendo las clases desbalanceadas, luego balanceando las clases haciendo uso del argumento de ajuste de clases que el algoritmo posee para tal fin, un tercer modelo balanceando las clases previamente con el método SMOTE-NC, y un cuarto modelo con el balanceo previo de las clases con el método SMOTE. Con cada modelo se obtienen pronósticos a partir de los datos de prueba, y se evalúan los resultados

obtenidos a través de las métricas correspondientes. Los valores de las métricas se presentan en la Tabla 4.

Revisando la Tabla 4, se puede decir que las cuatro métricas de desempeño mejoran cuando las clases son balanceadas utilizando el método SMOTE. Sí se utiliza el método SMOTE-NC mejoran las métricas, a excepción de la exactitud. Al balancear las clases utilizando el argumento de ajuste de clases del algoritmo, mejoran sólo las métricas F1 y Recordatorio.

Posteriormente, estos modelos se utilizan para estimar la hora de demanda máxima para los siguientes diecisiete días (datos nuevos), desde el 15 al 31 de julio, considerando dos opciones: 2 pm y 8 pm, ambas clases con una tolerancia de 1 hora, es decir, un rango de 1 pm a 3 pm, y el otro rango desde las 7 pm a las 9 pm. Los resultados se comparan con los valores reales, y los nuevos valores de las métricas de evaluación se presentan en la Tabla 5, mientras que las horas estimadas se presentan en la Tabla 6.

Tabla 4

Métricas de desempeño de modelos RL – Datos de prueba

Métrica	Desbalanceadas	Balanceadas	SMOTE-NC	SMOTE
Exactitud	0,788	0,752	0,737	0,799
Precisión	0,731	0,717	0,735	0,799
F1	0,563	0,719	0,726	0,792
Recordatorio	0,565	0,799	0,723	0,789

Tabla 5

Métricas de desempeño de modelos RL – Datos nuevos

Métrica	Desbalanceadas	Balanceadas	SMOTE-NC	SMOTE
Exactitud	0,529	0,588	0,823	0,647
Precisión	0,265	0,781	0,826	0,800
F1	0,346	0,471	0,824	0,575
Recordatorio	0,500	0,563	0,826	0,625

Tabla 6

Horas estimadas modelos RL

Fecha	Hora real	Rango real	Desbalanceadas	Balanceadas	SMOTE-NC	SMOTE
15-Jul-24	14	13_14_15	19_20_21	19_20_21	13_14_15	19_20_21
16-Jul-24	20	19_20_21	19_20_21	19_20_21	13_14_15	19_20_21
17-Jul-24	15	13_14_15	19_20_21	19_20_21	13_14_15	19_20_21
18-Jul-24	15	13_14_15	19_20_21	19_20_21	19_20_21	19_20_21
19-Jul-24	20	19_20_21	19_20_21	19_20_21	13_14_15	19_20_21
20-Jul-24	20	19_20_21	19_20_21	19_20_21	19_20_21	19_20_21
21-Jul-24	20	19_20_21	19_20_21	19_20_21	19_20_21	19_20_21
22-Jul-24	14	13_14_15	19_20_21	19_20_21	13_14_15	19_20_21
23-Jul-24	15	13_14_15	19_20_21	19_20_21	13_14_15	19_20_21
24-Jul-24	20	19_20_21	19_20_21	19_20_21	19_20_21	19_20_21
25-Jul-24	14	13_14_15	19_20_21	19_20_21	13_14_15	19_20_21
26-Jul-24	14	13_14_15	19_20_21	13_14_15	13_14_15	13_14_15
27-Jul-24	20	19_20_21	19_20_21	19_20_21	19_20_21	19_20_21
28-Jul-24	20	19_20_21	19_20_21	19_20_21	19_20_21	19_20_21
29-Jul-24	13	13_14_15	19_20_21	19_20_21	13_14_15	13_14_15
30-Jul-24	20	19_20_21	19_20_21	19_20_21	19_20_21	19_20_21
31-Jul-24	20	19_20_21	19_20_21	19_20_21	19_20_21	19_20_21

De la Tabla 5 se puede ver que todas las métricas de los modelos con clases balanceadas mejoran al compararse con las métricas derivadas del modelo con clases desbalanceadas. De los tres modelos con métricas mejoradas, aquel en el que se utilizó el método SMOTE-NC es el que tiene los mejores valores de las métricas.

Al observar la Tabla 6, se puede decir que el modelo con clases balanceadas utilizando el método SMOTE-NC acertó la hora de demanda máxima en catorce de los diecisiete días evaluados, seguido con el modelo con las clases balanceadas haciendo uso del método SMOTE con once días. Para el caso del modelo con clases balanceadas utilizando el argumento de ajuste de clases los aciertos fueron en diez de los diecisiete días, y para el modelo con clases desbalanceadas fueron en sólo nueve días.

Sí se comparan las métricas del modelo que tuvo más aciertos con las métricas del modelo con clases desbalanceadas, para los datos nuevos, se tiene que la Exactitud mejoró alrededor de 56%, la Precisión mejoró en más de 200%, la métrica F1 mejoró en más del 100%, y el Recordatorio mejoró alrededor de 65%.

IV. CONCLUSIONES

Se presenta una metodología que permite mejorar el desempeño de modelos de clasificación utilizados para pronosticar la hora de ocurrencia de la demanda eléctrica máxima, cuando hay un desequilibrio en las clases de los datos. Para ello se utilizaron métodos de balanceo de clases del tipo de sobre-muestreo, tales como: SMOTE y SMOTE-NC, así como el ajuste de clases a través del argumento propio de los algoritmos implementados utilizando el lenguaje de programación Python.

Para cada uno de los algoritmos utilizados: máquina de soporte vectorial y regresión logística, se generaron cuatro modelos, uno con clases desbalanceadas y tres con clases balanceadas, resultando en mejoras de las métricas de desempeño, tanto con los datos de prueba como con nuevas instancias, principalmente cuando se utilizaron los métodos SMOTE y SMOTE-NC.

Los dos modelos de clasificación generados con clases balanceadas utilizando el método SMOTE-NC, tuvieron los mayores aciertos de las horas de ocurrencia de la demanda máxima al considerar los datos nuevos, con quince aciertos para el modelo del algoritmo de máquina de soporte vectorial, y catorce aciertos para el modelo del algoritmo de regresión logística.

De acuerdo con los datos históricos, la hora de ocurrencia es en la tarde alrededor de las 2 pm, o en la noche alrededor de las 8 pm. Hay un desequilibrio en estas horas de ocurrencia, con una clara inclinación hacia las 8 pm, con una proporción de 75/25 con respecto a todas las otras horas para el período 2021-2024, y una proporción de 88/12 en las horas de ocurrencia de la noche con respecto a las horas del día, para el mismo período de estudio.

REFERENCIAS

[1] Babajide Ebenezer1, A., Boyinbode, O., & Oladunjoye, M. (2021). A Comprehensive Analysis of Handling Imbalanced Dataset. *International Journal of Advanced Trends in Computer Science and Engineering*, 454-463. https://doi.org/10.30534/ijatcse/2021/031022021.

[2] Bovornkeeratiroj, P., Wamburu, J., Irwin, D., & Shenoy, P. (2022). PeakTK: An Open Source Toolkit for Peak Forecasting in Energy Systems. *Computing and Sustainable Societies.* Seattle, WA, USA: ACM SIGCAS & SIGCHI. https://doi.org/10.1145/3530190.3534791.

[3] Candia, C. (10 de Julio de 2024). *ENERLINK*. Recuperado el 10 de Julio de 2024, de https://blog.enerlink.com/las-claves-del-cobro-de-demanda-en-horas-punta

[4] Cielen, D., Meysman, A., & Ali, M. (2016). *Introducing Data Science.* Shelter Island, NY: Manning Publications Co.

[5] Donaldson, D. L., Browell, J., & Gilbert, C. (2023). Predicting the magnitude and timing of peak electricity demand: A competition case study. *IET Smart Grid*, 1-12. https://doi.org/10.1049/stg2.12152.

[6] Dube, L., & Verster, T. (2023). Enhancing classification performance in imbalanced datasets: A comparative analysis of machine learning models. *Data Science in Finance and Economics*, 354-379. DOI:10.3934/DSFE.2023021.

[7] Fu, T., Zhou, H., Ma, X., Hou, Z., & Wu, D. (2022). Predicting peak day and peak hour of electricity demand with ensemble machine learning. *Frontiers in Energy Research*, 1-11. https://doi.org/10.3389/fenrg.2022.944804.

[8] Kirasich, K., Smith, T., & Sadler, B. (2018). Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets. *SMU Data Science Review*, https://scholar.smu.edu/datasciencereview/vol1/iss3/9.

[9] Lee, W. M. (2019). *Python Machine Learning.* Indianapolis: John Wiley & Sons, Inc.

[10] Liu, J., & Brown, L. (2019). Prediction of Hour of Coincident Daily Peak Load. *IEEE Power & Energy Society Innovative Smart Grid Technologies Conference.* Washington DC, USA : IEEE Explore. https://doi.org/10.1109/ISGT.2019.8791587.

[11] Mukhiya, S., & Ahmed, U. (2020). *Hands-On Exploratory Data Analysis with Python.* Birmingham, UK: Packt Publishing Ltd.

[12] Muthukrishnan, R., & Jamila. S, M. (2020). Predictive Modeling Using Support Vector Regression. *International Journal of Scientific & Technology Research*, 4863-4865.

[13] Poddar, G., Patill, R., & Kumar, S. (2024). Approaches to handle Data Imbalance Problem in Predictive Machine Learning Models: A Comprehensive Review. *INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING, 12*(21S), 841-856.

[14] Torriti, J. (2017). Understanding the timing of energy demand through time use data: Time of the day dependence of social practices. *Energy Research and Social Science*, 37-47. https://doi.org/10.1016/j.erss.2016.12.004.

[15] Voronin, V., Nepsha, F., & Krasilnikov, M. (2023). Short term forecasting peak load hours of regional power systems using machine learning methods. *CIGRE Science & Engineering (CSE)*, 1-18.

[16] Wongvorachan, T., He, S., & Bulut, O. (2023). A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining. *Information MDPI*, 4-15. https://doi.org/10.3390/info14010054.

[17] Xie, Y. (01 de 01 de 2022). *Peak Load Hour Prediction.* Obtenido de Exergy Energy: https://exergyenergy.com/wp-content/uploads/2022/01/Peak-load-Prediction-White-Paper.pdf

[18] Yajure Ramírez, C. A. (2023). Selección del modelo óptimo de predicción de la relación de desempeño de una planta solar fotovoltaica. Un enfoque multicriterio basado en algoritmos de aprendizaje automático. *Ciencia, Ingenierías y Aplicaciones INTEC*, 7-29. DOI: https://doi.org/10.22206/cyap.2023.v6i2.2935

Fuentes de financiamiento:

Propia.

Conflictos de interés:

El autor declara no tener conflictos de interés.

Contribuciones de autoría:

El único autor de esta investigación desarrolló todas sus etapas. Es decir, extrajo el conjunto de datos utilizados, aplicó las técnicas de preparación a dichos datos, realizó el análisis exploratorio, generó los modelos de aprendizaje automático, y redactó las conclusiones. Asimismo, participó en la redacción y revisión final del artículo.