UNIVERSIDAD AUTÓNOMA DEL ESTADO DE

MÉXICO

TESIS

Predicción de la Demanda Eléctrica Utilizando Técnicas de Aprendizaje Automático

---

Tesis que presenta

Miguel Ventura Cruz

Para obtener el Grado de

Ingeniero en Computación

Asesor de Tesis:

Dr. Jair CERVANTES

Revisores:

Dr. Joel Ayala de la Vega

Dr. Josué Espejel Cabrera

Texcoco, Estado de México. Diciembre del 2025

Resumen

El crecimiento del sistema y la incorporación de nuevas fuentes de generación exigen una gestión con mayor eficiencia de la generación, el consumo y los precios de la energía. La predicción de demanda eléctrica es afectada por el uso de fuentes renovables (como energía solar y eólica) y la variabilidad en la demanda eléctrica complica la predicción precisa de estos factores. Sin embargo, factores como las variaciones climáticas los patrones de consumo y el crecimiento de la demanda hacen que tarea sea un desafío. Esto puede llevar a ineficiencias en la operación de la red y precios volátiles. Esta investigación propone el desarrollo de modelos basados en aprendizaje automático (machine learning) y aprendizaje profundo (deep learning) para mejorar la precisión en la predicción de precios y demanda eléctrica en el mercado. Se analizarán y compararán técnicas como Random Forest, XGBoost y redes neuronales recurrentes (RNN, LSTM), tomando en cuenta variables históricas de consumo y factores meteorológicos clave. El objetivo es la optimización de estos modelos predictivos para mejorar la gestión del sistema eléctrico, contribuir a una gestión más eficiente y sostenible de la energía.

Índice general

Índice de figuras

Lista de Tablas

Capítulo 1

Introducción

1.1. Introducción

Con el crecimiento urbano e industrial, la demanda energética ha aumentado exponencialmente. La predicción de la demanda eléctrica representa un elemento clave en la planificación y operación de los sistemas energéticos modernos. Una estimación precisa de esta variable permite optimizar la generación, distribución y almacenamiento de energía, así como prevenir posibles fallas en la red, minimizar costos y favorecer la integración eficiente de fuentes renovables.

La predicción de estas tendencias es de suma importancia y a la vez estrategica para los operadores del sistema electrico y para los gobiernos.

Para abordar este problema, el uso de técnicas avanzadas de aprendizaje automático (machine learning) y aprendizaje profundo (deep learning) ha surgido como una alternativa. Esto permite analizar grandes cantidades de datos en tiempo real, identificar patrones y generar predicciones más precisas. La incorporación de variables meteorológicas y de mercado en modelos basados en algoritmos como Random Forest, XGBoost y redes neuronales (RNN, LSTM) pueden contribuir a mejorar la predicción de precios y demanda eléctrica, optimizando la estabilidad y la eficiencia del mercado energético.

Tradicionalmente, la predicción de la demanda se ha abordado mediante modelos estadísticos clásicos como ARIMA o regresiones lineales. Sin embargo, estos enfoques

presentan limitaciones cuando se enfrentan a series temporales con dinámicas complejas, no lineales, estacionales o afectadas por múltiples variables externas. En contraste, los métodos de aprendizaje automático han demostrado una mayor capacidad para modelar relaciones complejas, adaptarse a patrones de comportamiento cambiantes y ofrecer mejores resultados en tareas de regresión, lo que los convierte en una alternativa prometedora para esta problemática.

En este contexto, la presente investigación propone aplicar técnicas de aprendizaje automático para predecir la demanda eléctrica mensual, utilizando como base el conjunto de datos proporcionado por la Agencia Internacional de Energía (IEA), el cual está disponible públicamente a través de la plataforma Kaggle bajo el nombre "Monthly Electricity Production in GWh [2010–2022]". Este conjunto de datos ofrece registros mensuales de generación eléctrica desde el año 2010 hasta 2022, desglosados por tipo de fuente energética (carbón, gas natural, hidroeléctrica, solar, eólica, entre otras), expresados en gigavatios-hora (GWh).

La calidad, amplitud temporal y nivel de desagregación de este dataset lo hacen especialmente adecuado para la construcción y validación de modelos de predicción basados en aprendizaje automático. Además, su estructura permite realizar una exploración inicial de tendencias, estacionalidades y correlaciones entre distintas fuentes de energía, lo que enriquece el análisis y favorece una interpretación más integral del fenómeno de la demanda.

1.2. Planteamiento del problema

La predicción precisa de la demanda eléctrica es un componente esencial para la operación eficiente y segura de los sistemas eléctricos modernos. Sin embargo, diversos factores como las variaciones climáticas, los cambios en los patrones de consumo, el crecimiento de la población y la incorporación de nuevas tecnologías incrementan la complejidad de estimar con exactitud el comportamiento de la demanda eléctrica. Una predicción errónea puede generar inestabilidad entre oferta y demanda de energía, provocando sobrecargas en la red, pérdidas económicas y un uso ineficiente de los recursos de generación. Los

modelos tradicionales de predicción, basados en métodos estadísticos o lineales, generan limitaciones al intentar capturar la naturaleza no lineal y dinámica del consumo eléctrico, ante la influencia de múltiples variables como la temperatura, hora del día, estación del año o actividad económica. Esto puede dar paso a errores significativos en la creación de pronósticos y afectar la planificación del sistema eléctrico. Ante esta problemática los modelos tradicionales de predicción y demanda eléctrica no son suficiente para capturar la complejidad de estos fenómenos, ya que no consideran factores climáticos, el comportamiento del mercado eléctrico y la combinación de diversas fuentes de energía. En este caso, el uso de técnicas de aprendizaje automático (machine learning) y aprendizaje profundo (Deep learning) se ha vuelto una alternativa para mejorar la precisión de predicciones de este tipo. Estos modelos son capaces de identificar patrones complejos y relaciones no lineales en grandes volúmenes de datos históricos y meteorológicos para generar estimaciones más confiables y adaptarlas a las condiciones volátiles del sistema eléctrico. En este caso, la aplicación de estas herramientas es fundamental para la gestión de la demanda, ayudar a la toma de decisiones y mejorar la eficiencia energética del sistema.

1.3. Justificación

La predicción de la demanda eléctrica es un reto importante debido a la variabilidad de los patrones de consumo y la influencia de factores externos, como las condiciones meteorológicas, la población y la economía. La mala predicción puede generar inestabilidad entre generar y consumo, incremento de costos energéticos. En este contexto, es importante el uso de aprendizaje automático y el aprendizaje profundo para mejorar la predicción y la adaptabilidad de la demanda. Los algoritmos como Random Forest, XGBoost y redes neuronales (RNN, LSTM) pueden generar grandes volúmenes de datos históricos y meteorológicos, identificando patrones complejos y relaciones no lineales que los métodos tradicionales no logran capturar. Esta investigación es relevante porque busca mejorar la predicción de la demanda eléctrica mediante el uso de modelos inteligentes para adaptarse dinámicamente al comportamiento del sistema. Al mejorar la eficiencia en la planificación del suministro eléctrico, reducir costos asociados a la operación del sistema eléctrico y

favorecer al uso sostenible de la energía, mejorara la estabilidad y confiabilidad del servicio eléctrico.

1.4. Objetivos

1.4.1. Objetivo general

Desarrollar un modelo predictivo de demanda eléctrica utilizando técnicas de aprendizaje automático (machine learning) y aprendizaje profundo (deep learning), con el propósito de mejorar la precisión en las estimaciones, optimizar la planificación del sistema eléctrico y contribuir a una gestión eficiente y sostenible de la energía.

1.5. Objetivos específicos

1. Integrar los datos de consumo eléctrico, generación y condiciones meteorológicas.
2. Limpiar y normalizar los datos para su uso en modelos de IA.
3. identificar patrones y correlaciones entre las variables (consumo, generación y clima).
4. Estudiar el impacto de las condiciones meteorológicas en la demanda eléctrica.
5. Implementar y comparar modelos de aprendizaje automáticos (regresión lineal, Random, Forest, Gradient Boosting) y aprendizaje profundo (redes neuronales recurrentes – RNN, LSTM).
6. Incorporar variables meteorológicas como predictores claves para mejorar la precisión.
7. Proponer estrategias para equilibrar la oferta y la demanda basadas en las predicciones.
8. Comparar el rendimiento de los modelos utilizando métricas como el error absoluto medio (MAE), el error cuadrático medio (MSE) y el coeficiente de determinación (R^2).

9. Validar los resultados con datos históricos y escenarios simulados.

1.6. Hipótesis

La aplicación de modelos de aprendizaje automático y profundo, que integran variables meteorológicas y datos del mercado eléctrico, mejora significativamente la precisión en la predicción de demanda eléctrica en el mercado. Esta mejora se vuelve evidente en comparación con los modelos tradicionales de predicción, que no logran capturar adecuadamente la variabilidad del sistema eléctrico. Por lo tanto, la mayor capacidad predictiva de estos modelos permite una planificación operativa del sistema eléctrico, reduciendo la dependencia de fuentes no renovables y contribuye a una transición energética más estable, eficiente y sostenible.

1.7. Estado del arte

Para comprender el enfoque de esta investigación, es necesario entender primero algunos conceptos fundamentales, como el aprendizaje automático. "Herramienta que busca mejorar el análisis de datos, en pro de una predicción futura, ya sea por la implementación de nuevos sistemas o simplemente el mejoramiento de los ya existentes, mediante el uso de algoritmos basados en información antigua o reciente que permita el funcionamiento optimo del sistema a trabajar" (RAMÍREZ, 2018). Desde una perspectiva más formal, "El 'machine learning' (aprendizaje automático) es una rama de la inteligencia artificial que permite que las máquinas aprendan ciertas tareas sin ser programadas de manera específica para ellas. Para ello, utilizan estadísticas para predecir y reconocer patrones, por ello funcionan mejor en grandes conjuntos de datos." (BBVA, 2024). En el contexto de demanda eléctrica, "El Mercado Eléctrico Marginalista que se emplea en diversos países, principalmente en la Unión Europea, fue implementado en México a partir de la Reforma Eléctrica de 2013. En este tipo de mercados, la última central de generación en entregar energía eléctrica al Sistema es quien determina el precio que se pagará por la energía a todas las demás centrales que hayan generado para satisfacer la demanda,

esto con independencia del precio ofertado por cada generador o la fuente de energía. Por lo cual, el orden del despacho de las centrales de generación se realiza en función de los costos variables, los cuales están determinados principalmente por el costo de los combustibles. Como consecuencia de lo anterior, en la práctica las primeras centrales de generación en ser despachadas son las energías renovables que presentan un costo variable cero, sin embargo, estas centrales reciben como pago el precio de la oferta más alta del mercado "la central más cara", esto sin importar que sean consideradas como "baratas". Lo anterior implica una enorme rentabilidad económica para los generadores privados con "costos variables bajos" y ningún beneficio a los usuarios finales de la energía eléctrica." (Electricidad, 2021). En los últimos años, diferentes estudios basados en modelos de aprendizaje automático para predecir los precios y demanda eléctrica en el mercado energético renovable toma en cuenta, "Uno de los factores más importantes a tener en cuenta en relación con la energía renovable, es el hecho de que la naturaleza es impredecible. Por poco sorprendente que parezca, esto es clave, ya que puede dificultar la generación de la cantidad de energía necesaria en un día determinado debido a las condiciones naturales. No poder predecir cuánta energía será generada por, por ejemplo, por un panel solar o una turbina eólicas puede suponer un coste considerable. Este coste es económico, pero también operacional, y puede llevar a la desestabilización, aunque sea momentánea, de la red eléctrica." (Guillard, 2022). Los factores necesarios para la predicción meteorológica con mayor precisión, "está basado en leyes físicas como la conservación de masa, conservación de energía y la ley de gas ideal, por mencionar algunos.

Se genera una malla y en cada una de estas se realiza el cálculo de algunos factores físicos: el Movimiento del aire (vientos); Transferencia de calor (termodinámica); Radiación (solar y terrestre); Contenido de humedad (humedad relativa); e hidrología superficial (precipitación, evaporación, deshielo y escorrentía). Estas variables se pueden utilizar como punto de partida conocidas como de entrada y obtener de resultado estas mismas, pero en otro periodo de tiempo para conocer como se ve afectado el sistema climático. Antes de emplear un modelo, obtener las proyecciones y los cambios en el futuro, los modelos son probados por científicos y programadores. Este proceso consiste en ejecutar el modelo desde el tiempo presente hacia atrás es decir al pasado. Después, los resultados del modelo se

comparan con el clima y las condiciones meteorológicas observadas en ese tiempo, para ver qué tan bien coinciden. Cuando funcionan bien estas pruebas, entonces los resultados para simular el clima futuro son válidos." (CLIMATICO, 2021) Algunos instrumentos que utilizan los meteorólogos para obtener datos del clima, "Destacan los termómetros que miden la temperatura, los barómetros que informan de la presión atmosférica, los anemómetros que calculan la velocidad del tiempo, posteriormente se lanzan a la atmósfera globos sonda con estos instrumentos y así conocer la condición de la atmósfera." (snovit, 2021) Dado lo anterior, "se dispone de un conjunto de ecuaciones físico-matemáticas que representan los procesos dinámicos y las relaciones energéticas del sistema tierra-atmósfera, de tal forma que aglutinan la información que permite diagnosticar y pronosticar el comportamiento de éste." (a, 2008) Existen algoritmos que toman todos estos resultados y generan, "la posibilidad de prever o predecir los valores futuros que toman sus variables principales: demanda eléctrica y precio eléctrico, está generando mayor interés. El objetivo principal de obtener previsiones a futuro es aumentar los beneficios y/o reducir los costes. En cuanto a la demanda eléctrica, la previsión de su valor a futuro es muy útil, por ejemplo, para aportar información a los agentes implicados en la cadena de suministro energético donde destaca la generación, el almacenamiento y la distribución de energía eléctrica. Por otro lado, la predicción del precio eléctrico implica tratar un problema de naturaleza volátil, no lineal y sin componente estacional, en la mayoría de los casos. Los productores y consumidores hacen uso de las previsiones de precio eléctrico del día anterior para desarrollar sus propias estrategias. Por ejemplo, si se predice de forma precisa y correcta las horas de precio alto, éstas podrán evitarse y centrarse en las horas en las que se prediga un precio menor" (GARCIA, 2024) En este contexto, destacan las técnicas de predicción basadas en modelos tipo árbol, "Este tipo de técnicas son modelos explicativos que se consiguen con la partición recursiva del espacio de variables de manera que queda dividido en numerosas regiones más simples denominadas nodos. El algoritmo consiste en lo siguiente: las regiones se subdividen en dos nuevas regiones mediante una variable explicativa. A esta subdivisión se le denomina corte binario y cada uno de ellos crea dos nuevos nodos que explican de manera más precisa el comportamiento de los datos. Las particiones se hacen de manera que se disminuya la impureza de los dos nuevos nodos creados a partir

del anterior y denominados nodo hijo o mediante test estadísticos." (Remon, 2017) Un factor clave en el sector energético es la demanda de energía eléctrica, "ya que permite al ente productor contar con un conocimiento más refinado de los mercados y de los usuarios del sistema, así como con una mejor posición a través de la reducción de la incertidumbre para la toma de decisiones. Existen diferentes horizontes de predicción relevantes, dependiendo de cuáles sean las decisiones estratégicas que se deban tomar" (Collazo, 2015) La problemática para dar un paso a las energías renovables es, "Abandonar la generación de energía basada en el carbono controlada por el ser humano y optar por la generación de energía controlada por la naturaleza, como la solar o la eólica. Esta variabilidad genera costos y complejidad en los modelos de pronóstico de la oferta y la demanda. Además, el aumento de la generación de energía por parte de los prosumidores y la proliferación de pequeñas empresas energéticas también exige que las empresas de servicios públicos desarrollen sistemas analíticos y estratégicos más complejos para anticipar y compensar esta pérdida de control centralizado." (públicos, 2022) En conclusión, la predicción de precios y demanda eléctrica enfrentan desafíos con la incorporar energías renovables, debido a la constante variabilidad climática que afecta al sistema haciéndolo más volátil. A diferencia de las fuentes tradicionales, estas no pueden ser controladas directamente complicando la planificación y estabilidad del mercado eléctrico. Para hacer frente a este problema, es fundamental utilizar modelos como arboles de decisiones, métodos estadísticos tradicionales y redes neuronales que permitan mejorar la precisión de las predicciones.

Capítulo 2

Marco Teórico

2.1. Fundamentos de la demanda eléctrica

La demanda eléctrica es la cantidad de energía eléctrica requerida por los usuarios (hogares, comercios, industrias, entre otros) en un momento especifico. Esta demanda se monitorea a través de medidores, los cuales permiten determinar la infraestructura necesaria y calcular el costo que el usuario paga durante el periodo de facturación. La demanda varia de manera continua y esta influenciada por factores tecnológicos, estacionales, socioeconómicos y climatológicos. Debido a que la energía eléctrica producida en gran escala no puede almacenarse de forma directa, la generación de electricidad debe ajustarse en tiempo real a la demanda actual requerida por los consumidores.

2.2. Factores que afectan la predicción de la demanda eléctrica

2.2.1. Factores climáticos

Los cambios en el clima aumentan el consumo eléctrico dependiendo de la estación del año (en verano e invierno), las variaciones de temperatura provocan el aumento o disminución del uso de sistemas de climatización para regular la temperatura y otros

dispositivos eléctricos, tanto en los hogares como en la industria [23].

2.2.2. Factores económicos

El crecimiento de la actividad económica, medida a través del Producto Interno Bruto (PIB), incrementa la demanda eléctrica. Cuando la producción industrial aumenta, el consumo de energía también lo hace, de forma contraria, una desaceleración económica reduce la demanda. Otro aspecto relevante es el papel de los combustibles fósiles, estos permiten el funcionamiento de centrales generadoras de electricidad. Estos combustibles permiten cubrir la demanda eléctrica actual y contribuyen a una predicción más precisa de los precios sobre la electricidad, en comparación con las fuentes de energía renovable, que son más intermitentes.

2.2.3. Factores sociales y demográficos

Estos factores son importantes porque el crecimiento de la población aumenta el número de hogares y negocios, lo que genera un cambio en la infraestructura de las centrales generadoras de electricidad para satisfacer la demanda actual.

2.2.4. Factores económicos

El nivel socio económico permite la adquisición de dispositivos eléctricos y electrónicos. Esto quiere decir: a mayor cantidad de dispositivos, mayor consumo de energía eléctrica. Este factor está ligado al uso responsable de la energía generando nuevos hábitos para el ahorro de energía. Algunas encuestas realizados por el INEGI permiten conocer algunos patrones de consumo en viviendas conocido como “Encuesta Nacional sobre Consumo de Energéticos en Viviendas Particulares (ENCEVI), La encuesta se levantó durante el primer semestre de 2018, con el objetivo de generar información estadística que permita conocer los patrones de consumo de las distintas fuentes de energía utilizadas para el consumo en las viviendas y, conjuntamente, conocer sobre los hábitos y las prácticas en el manejo de energéticos.” Esta información permite a las empresas generadoras de energía eléctrica

obtener un panorama general de las regiones con mayor consumo eléctrico, permitiendo ajustar su infra estructura y potencia eléctrica.

2.2.5. Factores tecnológicos

El avance tecnológico también impacta en el aumento de la demanda eléctrica. El desarrollo de la inteligencia artificial y los centros de datos requieren servidores con alto consumo eléctrico. Así mismo, la electrificación del transporte (vehículos eléctricos y sus diferentes tipos de carga), representan un desafío para gestionar la demanda eléctrica global.

2.3. Tipo de demanda

La demanda eléctrica se puede clasificar el nivel y variabilidad a lo largo del tiempo. Los principales tipos son los siguientes:

2.3.1. Demanda base

Corresponde al mínimo nivel de demanda eléctrica constante en un intervalo de tiempo relativamente corto dentro de un periodo de tiempo determinado. Representa el consumo permanente de electricidad necesario para mantener en operación los servicios esenciales, como alumbrados públicos, hospitales, servidores, entre otros.

2.3.2. Demanda media

corresponde al promedio de la demanda en un periodo dado tal como un día, un mes o un año. Este tipo de permite evaluarla eficiencia general del sistema eléctrico y planificar la generación promedio necesaria para satisfacer la mayor parte del consumo.

Demanda Promedio=(Consumo en el periodo)/(Número de horas del periodo)

2.3.3. Demanda máxima

Es el mayor valor de consumo registrado en un intervalo de tiempo determinado. Se presenta en horarios o épocas específicas, en las horas de mayor actividad industrial o las estaciones del año en los cuales se presenta un pico de demanda. “La demanda máxima representa para un instante dado, la máxima coincidencia de cargas eléctricas operando al mismo tiempo, es decir la demanda máxima corresponde a un valor instantáneo en el tiempo. No es igual encender una línea de motores al mismo tiempo que hacerlo en arranque escalonado. El medidor de energía almacenara únicamente, la lectura corresponde al máximo valor registrado de demanda, en cualquier intervalo de tiempo de cualquier día del ciclo de lectura. Los picos por demanda máxima se pueden controlar evitando el arranque y la operación simultanea de cargas eléctricas.” En consecuencia, la demanda máxima es un parámetro clave para dimensionar la capacidad de generación, transmisión y distribución del sistema eléctrico, determinado la infraestructura necesaria para cubrir los momentos de mayor exigencia energética.

2.4. previsión de carga

Tomando en cuenta este aspecto “La previsión de carga es el proceso de predecir cuánta electricidad se necesitará en un momento dado y cómo afectará esa demanda a la red de servicios públicos. Se utiliza para garantizar que se dispone de energía suficiente para satisfacer las necesidades de consumo, evitando al mismo tiempo el despilfarro y la ineficacia.” (Amanda McGrath, 2024).

2.4.1. Plazos de previsión de carga a corto, mediano y largo plazo

Estos términos se refieren a la predicción y planeación de la demanda eléctrica, analizando los datos históricos y otros datos relevantes para prevenir las distintas variaciones de la demanda eléctrica.

2.4.2. Previsión de carga a corto plazo

Abarca un lapso de hasta una semana, se utiliza principalmente para la operación diaria del sistema eléctrico, tomando en cuenta los datos meteorológicos y de carga reciente. Esto permite a los operadores del sistema tomar decisiones sobre cuanta energía generar y hacia dónde dirigirla. este aspecto es crucial, ya que pequeños errores en la previsión pueden provocar una carencia de energía o la sobrecarga de las líneas eléctricas.

2.4.3. Previsión de carga a mediano plazo

Comprende un periodo de tiempo que abarca desde una semana hasta un año. Se utiliza para planificar el mantenimiento de plantas eléctrica, las compras de combustibles y la programación de contratos de suministro eléctricos. Este tipo de previsión ayuda a mantener la estabilidad operativa y de los recursos energéticos.

2.4.4. Previsión de carga a largo plazo

Suele abarcar un periodo de más de un año tiene en cuenta factores como los cambios demográficos, el crecimiento económico y las repercusiones de la política energética. Su objetivo es la planificación estratégica del sistema eléctrico: expansión de redes, construcción de nuevas centrales, integración de energías renovables y políticas de sostenibilidad energética.

2.5. Importancia de la predicción de demanda en la planificación eléctrica

Es un elemento esencial en la planificación y operación de los sistemas eléctricos, porque permite: Garantizar la confiabilidad del suministro, anticipando el nivel de generación necesario para evitar apagones o sobrecargas. Optimiza los recursos energéticos, evitando el uso innecesario de combustibles fósiles para reducir la emisión de contaminantes. Reducción de costos operativos, mediante una planificación del despacho eléctrico y la

compra eficiente de combustibles para el funcionamiento de las plantas generadoras de electricidad. Promover la sostenibilidad energética, promoviendo la integración de energías renovables, cuya generación depende de condiciones intermitentes como el sol o el viento. Planificación de la infraestructura futura, plantea nuevas propuestas de redes de transmisión, distribución y generación que corresponden al crecimiento de la población y de la economía.

2.6. Modelos tradicionales de predicción

Estos modelos toman como base distintos bancos de datos que permiten la predicción sobre la demanda eléctrica actual, estos se basan en métodos estadísticos y econométricos que comprenden el comportamiento histórico de la demanda eléctrica para evaluar su futura evolución. Estos métodos suponen que las relaciones entre variables son estables y lineales, lo cual permite obtener resultados interpretables, aunque con ciertas limitaciones frente a fenómenos más complejos o no lineales. La predicción estadística se basa en el análisis de datos históricos para identificar patrones que permiten estimar valores futuros. A diferencia del aprendizaje automático, es una rama de la informática que se centra en el desarrollo de algoritmos y modelos los cuales son introducidos en computadoras que pueden aprender y tomar decisiones basándose en los datos [14] [19] [1] [17] [22].

2.7. Uso de datos históricos y series temporales

Los modelos estadísticos utilizan datos histórico-organizados en series temporales esta "es una técnica estadística para analizar puntos de datos registrados a intervalos de tiempo regulares. Ayuda a identificar patrones, tendencias y variaciones estacionales, de modo que es útil para proyectar resultados a lo largo del tiempo. Los equipos de ingeniería y ciencias que trabajan con datos de series temporales pueden utilizar el análisis de series temporales para monitorizar, modelar y predecir comportamientos de sistemas, lo que optimiza los sistemas y mejora la predicción de las proyecciones." .

2.8. Componentes de series temporales

Tendencia: Toma como base la dirección general de los datos a lo largo del tiempo, como aumento, disminución o constante. Estacionalidad: se refiere a patrones de datos que se repiten durante un conjunto de periodos de tiempo (ya sea diariamente, mensual o anual). Variaciones irregulares (ruido): componente aleatorio o residual dentro de un modelo de series temporales. Como los que se presentan en la predicción de demanda eléctrica.

2.9. Pasos del análisis de series temporales

El análisis de datos de series temporales integra distintas técnicas para comprender, modelar y realizar proyecciones de puntos de datos recopilados a lo largo del tiempo.

2.10. Análisis exploratorio de datos

El análisis exploratorio de datos (EDA) consiste en la recopilación y examinación de datos no procesados, que posteriormente serán procesados y serán visualizados para un análisis profundo. Este proceso incluye la recopilación de datos a lo largo de un periodo especifico, para luego realizar un preprocesamiento y visualización, lo que permite detectar patrones, relaciones y posibles anomalías, preparando la información para su análisis y modelado más profundo.

2.11. Descomposición

Es una técnica utilizada para separar los datos de series temporales ya sea por tendencia, estacionalidad, ciclos y residual, lo cual facilita el análisis de patrones y la interpretación de datos.

2.12. Selección y ajuste del modelo

Selecciona el modelo para capturar los patrones en función de estacionalidad, tendencia y estacionariedad. El ajuste del modelo se realiza durante el entrenamiento para reducir la diferencia de datos observados y predicciones. De modo que el modelo pueda generar nuevos datos. Predicciones y proyección de modelos Se basa en los datos obtenidos del paso anterior para generar nuevos datos futuros que se basan en patrones históricos.

2.13. Evaluación del modelo

En este paso se determina cuan bien se desempeña un modelo y la precisión de sus predicciones. Consta de tres pasos esenciales: Para evaluar y optimizar el ajuste se emplean métricas de error como el Error cuadrático medio (MSE) o el Error Absoluto Medio (MAE), y se valida la capacidad de generación mediante técnicas como la validación cruzada temporal.

2.13.1. Métricas de rendimiento

Se emplean métricas de error como el Error cuadrático medio (RMSE), calcula la diferencia entre valores previstos y reales, permitiendo medir la precisión de un modelo.

2.13.2. Técnicas de validación

la validación cruzada, backtesting permiten evaluar la fiabilidad y capacidad del modelo para realizar predicciones sobre conjunto de datos nuevos, garantizando la precisión y robustez ante nuevos escenarios.

2.13.3. Métodos de interpretabilidad

“las técnicas de LIME (explicaciones locales interpretables independientes del modelo) y SHAP (explicación de Shapley) ayudan a comprender las predicciones del modelo

y permiten que sus decisiones sean más transparentes." (MathWorks, Análisis de series temporales, 2025)

2.13.4. Estacionariedad

La estacionariedad en una serie temporal se refiere a que sus propiedades estadísticas, como la media, la varianza y la autocorrelación, permanecen constantes a lo largo del tiempo.

Detección de estacionariedad

Para detectar si una serie es estacionaria, se utilizan tanto métodos visuales como pruebas estadísticas.

Métodos visuales

- **Gráfico de la serie en el tiempo:** observar si la media parece cambiar o si la varianza se amplía o contrae.
- **Gráficos de autocorrelación (ACF) y autocorrelación parcial (PACF):** para una serie estacionaria la ACF tiende a caer relativamente rápido; para una no estacionaria, la ACF decae lentamente.

Pruebas estadísticas

- **Augmented Dickey-Fuller test (ADF):** prueba de raíz unitaria. Hipótesis nula H_0: la serie tiene raíz unitaria (es decir, no es estacionaria).
- **KPSS (Kwiatkowski-Phillips-Schmidt-Shin):** enfoque opuesto al ADF. Hipótesis nula H_0: la serie es estacionaria.

Métodos para lograr la estacionariedad

Cuando una serie no es estacionaria, se aplican técnicas de transformación para estacionarizarla antes de aplicar modelos que requieren esta propiedad.

Diferenciación

- **Diferenciación de primer orden:** $\Delta y_t = y_t - y_{t-1}$. Esto elimina un cambio constante en el nivel.
- **Diferenciación de orden más alto:**

$$\Delta^2 y_t = \Delta(\Delta y_t) = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}),$$

útil cuando existe una tendencia polinómica.

Transformación de la variable

- **Transformación logarítmica:** $y' = \log(y)$, apropiada cuando la varianza crece con el nivel de la serie.

2.14. Ventajas de los modelos estadísticos

Tabla 2.1: Ventajas de los modelos estadísticos tradicionales

Criterio	**Descripción**
Simplicidad y facilidad de implementar	Modelos sencillos de aplicar, con fundamentos matemáticos establecidos.
Interpretabilidad	Permite observar de forma clara la relación entre las variables (por ejemplo, temperatura o el PIB).
Bajo costo computacional	Requiere de menor uso de recursos de procesamiento y tiempo de entrenamiento.
Buen desempeño con datos lineales y estacionarios	Funciona adecuadamente cuando los datos presentan comportamientos regulares o patrones bien definidos.
Base teórica sólida	Están sustentados en métodos estadísticos clásicos que facilitan la validación y comparación de resultados.

2.15. Modelado de series temporales: enfoque clásico vs. enfoque de aprendizaje automático

El modelo de series temporales puede abordarse principalmente desde dos perspectivas: el enfoque clásico o estadístico y el enfoque basado en aprendizaje automático (ML) y aprendizaje profundo (DL). Ambos métodos buscan identificar patrones temporales para realizar predicciones, pero difieren en sus fundamentos teóricos y capacidades predictivas. en síntesis, los modelos clásicos son mas adecuados para series temporales estables, lineales y de bajo dimensionalidad, mientras que los modelos de aprendizaje automático destacan en entornos dinámicos, no lineales y con gran cantidad de datos y con variables de origen externo.

2.16. Desventajas de los modelos estadísticos

Los modelos tradicionales no capturan bien relaciones no lineales ni comportamientos complejos (como los cambios bruscos de consumo). Por ejemplo, el aumento de la temperatura no siempre produce un incremento proporcional en la demanda eléctrica. Ejemplo: Supongamos que en una ciudad la demanda promedio es de 2000 MW a 25°c y aumenta ligeramente a 2050 MW cuando la temperatura sube a 28°C. sin embargo, si la temperatura alcanza 36°C, el uso de sistemas de aire acondicionado puede disparar la demanda hasta 2400 MW. Por lo tanto, este comportamiento no lineal y los picos repentinos no son capturados correctamente por modelos lineales o econométricos tradicional ya que estos asumen cambios proporcionales y continuos. Este motivo llevo a la incorporación de técnicas de aprendizaje automático e inteligencia artificial, que permiten manejar datos complejos y no necesariamente lineales.

Tabla 2.2: Limitaciones de los modelos lineales en predicción de series temporales

Limitación	**Descripción**
Supuestos estrictos	Requiere que los datos sean lineales, presenten normalidad y estacionariedad, lo que limita su aplicación a series más complejas.
Baja capacidad para capturar relaciones no lineales	No modelan completamente bien los patrones complejos o interacciones no lineales.
Sensibilidad a valores atípicos	Los datos anómalos pueden distorsionar significativamente las predicciones.
Dependencia de variables históricas	Se basa en el comportamiento pasado sin adaptarse a cambios estructurales.
Limitada capacidad predictiva a largo plazo	Su desempeño baja cuando se proyecta a horizontes de tiempo extensos.

2.17. Ejemplo de aplicación en el sector energético

En la predicción de demanda eléctrica, los modelos ARIMA se empleados para pronosticar el consumo horario o diario y los modelos SARIMA se utilizan para incorporar la variabilidad estacional y predecir un fenómeno en la demanda anual. Asimismo, la regresión lineal es común para estimar la demanda mensual o anual en función de variables como la temperatura o el crecimiento económico.

2.18. Modelos ARIMA Y SARIMA

Las series temporales son un reto clave en estadística y ciencia de datos. Un dato se convierte en una serie temporal cuando se muestrea según un atributo temporal, como día, meses y años. El pronóstico consiste en tomar los datos y predecir nuevos datos a futuro. Los modelos ARIMA Y SARIMA son algoritmos de pronóstico. El modelo ARIMA

(Media Móvil Integrada Autorregresiva), considera los valores pasados de la demanda y predecir valores futuros con base en ellos combinado tres componentes: Autor regresión (AR): utiliza la dependencia entre un valor y sus observaciones anteriores. Integración (I): aplica diferenciaciones sucesivas que se aplican a una serie temporal para convertirla en estacionaria. Media Móvil (MA). Incorpora los errores de la predicción anterior para mejorar el ajuste del modelo. Utiliza las diferencias entre los valores reales y los pronósticos en periodos previos para corregir futuras estimaciones (por ejemplo, picos diarios o anuales en la demanda eléctrica). Ventajas: tiene buena precisión cuando los datos son estacionarios y tiene parámetros repetitivos. Limitaciones: requieren una gran preparación de datos y no consideran variables externas.

2.18.1. ARIMA (AutoRegressive Integrated Moving Average)

Promedio Móvil Integrado Autorregresivo. Integra tres componentes:

- **AR(p):** Orden de autorregresión de tendencia.

$$X_t = c + \sum_{i=1}^{p} \phi_i X_{t-i} + \varepsilon_t$$

- **I(d):** Orden de diferenciación para hacer estacionaria la serie.

- **MA(q):** Orden de media móvil de tendencia.

$$X_t = \mu + \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}$$

El modelo ARIMA(p, d, q) se obtiene aplicando d diferencias a la serie original para remover la no estacionariedad, y posteriormente se ajusta un modelo ARMA(p, q) sobre la serie resultante. La identificación de los parámetros (p, d, q) se realiza siguiendo la metodología de Box-Jenkins: inspección de gráficos ACF/PACF y pruebas de estacionariedad (ADF, KPSS).

2.18.2. SARIMA $(p, d, q)(P, D, Q)_s$

Promedio Móvil Integrado Autorregresivo Estacional.

La forma general del modelo es:

$$\phi_P(L^s)\,\phi_p(L)\,(1-L)^d\,(1-L^s)^D\,\gamma_t = \theta_Q(L^s)\,\theta_q(L)\,\varepsilon_t$$

Donde:

- γ_t: valor de la serie temporal en el tiempo t.
- ε_t: término de error (ruido blanco).
- L: operador de rezago, tal que $L\gamma_t = \gamma_{t-1}$.
- p, d, q: órdenes del componente no estacional (AR, diferenciación, MA).
- P, D, Q: órdenes del componente estacional.
- s: periodo estacional.

Elementos de tendencia:

- AR(p): autorregresión.
- I(d): diferenciación.
- MA(q): media móvil.

Elementos estacionales:

- P: orden autorregresivo estacional.
- D: orden de diferenciación estacional.
- Q: orden de media móvil estacional.
- s: periodo de la estacionalidad.

Cuando la serie presenta estacionalidad marcada (ciclos diarios, semanales o anuales), se utiliza SARIMA, que extiende ARIMA añadiendo los componentes estacionales (P, D, Q) y el periodo s (por ejemplo, $s = 12$ meses o $s = 24$ horas).

2.18.3. SARIMAX

(Modelo Autorregresivo Integrado de Media Móvil Estacional con regresores exógenos) La forma general es:

$$\phi_P(L^s)\,\phi_p(L)\,(1-L)^d\,(1-L^s)^D\,\gamma_t = \beta X_t + \theta_Q(L^s)\,\theta_q(L)\,\varepsilon_t$$

Donde:

- X_t: vector de variables exógenas (por ejemplo, temperatura, hora del día, etc.).
- β: vector de coeficientes asociados a las variables exógenas.

SARIMAX es una extensión del modelo SARIMA que permite incluir variables externas X_t, como temperatura, día de la semana, días feriados, carga industrial, entre otras.

2.19. Aprendizaje automático (ML)

El aprendizaje automático ha tomado un papel muy importante en los últimos años, permitiendo el uso de herramienta capaces de manejar grandes volúmenes de datos que organizan y procesan para descubrir patrones e identificar tendencias. Podemos definir que el aprendizaje automático (ML) "es una rama de la inteligencia artificial que permite que las máquinas aprendan ciertas tareas sin ser programadas de manera específica para ellas. Para ello, utilizan estadísticas para predecir y reconocer patrones, por ello funcionan mejor en grandes conjuntos de datos." (BBVA, 2024) "En resumen, podemos decir que todo aprendizaje automático es IA, pero no toda IA es aprendizaje automático." (Chen, 2024)

Diferencia entre IA, ML y Deep Learning La inteligencia artificial engloba cualquier sistema que simula la inteligencia humana. El aprendizaje automático es una rama de la IA que permite a los sistemas aprender atreves de la exploración de datos. Dentro del Machine Learning existe el Deep Learning representa un enfoque basado en redes neuronales profundas capaces de capturar relaciones altamente complejas en grandes volúmenes de información.

2.19.1. Tipos de aprendizaje

Principales tipos de aprendizaje Aprendizaje supervisado: utiliza datos etiquetados, los cuales permiten a cada entrada ser asocia con una salida. Ejemplo, reconocimiento de spam.

1. Aprendizaje no supervisado: trabaja con datos sin etiquetas explicitas. Pretende descubrir patrones o relación entre datos, es decir, el sistema aprende por sí mismo. Es usual mente usada para agrupación de datos similares y encontrar relación entre variables.

2. Aprendizaje por refuerzo: modelo que es entrenado mediante retroalimentación en forma de recompensas o penalizaciones. El objetivo de este modelo es descubrir la mejor política de acciones para maximizar la recompensa a largo plazo. Es usualmente utilizada en robótica o videojuegos.

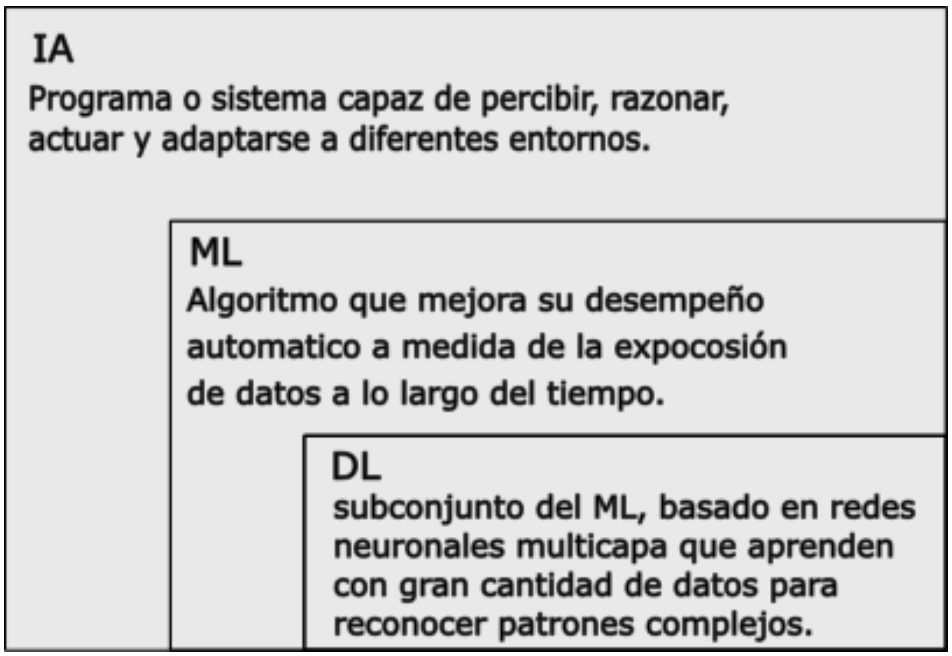

Figura 2.1: Metodología propuesta

2.19.2. Ventajas del aprendizaje automático frente a modelos tradicionales

El aprendizaje automático (ML) ofrece múltiples ventajas en modelos estadísticos o econométricos tradicionales, hablado en el contexto de demanda eléctrica, donde los datos suelen ser complejos, no lineales y de alta variabilidad. Capacidad para capturar relaciones no lineales Adaptabilidad y mejora continua Manejo de grandes volúmenes de datos Menor dependencia de supuestos estadísticos Automatización Integración con sistemas inteligentes

2.20. Técnicas de Aprendizaje Supervisado para Regresión

En este apartado "se pueden utilizar técnicas de regresión en el aprendizaje supervisado para comprender la relación entre la respuesta y las variables de entrada. Son útiles para conjuntos de datos con un rango de valores o cuando la respuesta es un número real, como la temperatura o el tiempo transcurrido hasta la falla del equipo. Es decir, los modelos de regresión predicen respuestas continuas. Entre sus aplicaciones típicas se incluyen la predicción de la carga eléctrica, la predicción del ciclo de vida restante de las baterías, el trading algorítmico, la predicción de los precios de las acciones." (MathWorks, Aprendizaje supervisado, 2025). Estas técnicas son ampliamente utilizadas en el sector energético debido a su capacidad para modelar relaciones complejas y no lineales entre factores técnicos, económicos y ambientales.

2.20.1. Regresión lineal

La regresión es un conjunto de métodos estadísticos que permiten estimar la relación entre una variable dependiente y una o más variables independientes [15]. El objetivo es predecir el valor de la variable dependiente a partir de las independientes. Un ejemplo de la forma simple de la regresión lineal simple es:

$$Y = a + bX + \varepsilon$$

Donde: Y: variable dependiente, X: variable independiente, a: intercepto, b: pendiente, ε: error o residual.

Tipos de regresión

Regresión lineal (simple)

Consiste en utilizar una única variable independiente X para predecir la variable dependiente Y.

La fórmula general es:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Donde: β_0: intercepto, β_1: pendiente.

Este método requiere ciertas condiciones, como que la relación sea lineal, que el error tenga varianza constante (homocedasticidad), y que los errores sean independientes, entre otras.

Regresión múltiple

Es una extensión de la regresión simple: varias variables independientes $X_1, X_2, \ldots, X_p$ se utilizan para explicar Y.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

Este modelo permite capturar varios factores que influyen simultáneamente en el resultado, lo cual es especialmente útil en escenarios complejos. Sin embargo, se debe prestar atención a la multicolinealidad entre las variables independientes.

Regresión no lineal

Se utiliza cuando la relación entre las variables independientes y la dependiente no puede describirse adecuadamente mediante una línea recta.

- La relación entre X y Y puede adoptar forma curvilínea.

- Ofrece mayor flexibilidad para modelar comportamientos complejos.

- Implica también mayor dificultad en el ajuste, interpretación y validación del modelo.

2.20.2. Arboles de decisión y Random Forest

Los árboles de decisión, o arboles de clasificación y arboles de regresión, predicen respuestas a los datos. Un árbol de decisiones comienza en el nodo raíz y desciende hacia a un nodo hoja, estos nodos contienen la respuesta. Una de sus ventajas radica en su interpretabilidad y facilidad para manejar tanto variables numéricas como categóricas. Sin embargo, los árboles individuales tienden a sobre ajustarse a los datos (overfitting), lo que puede reducir su capacidad de realizar predicciones precisas [16]. Para mitigar esta limitación, el modelo Random Forest combina múltiples árboles de decisión entrenados sobre subconjuntos aleatorios de los datos y de las variables. La agregación de las predicciones individuales permite reducir la varianza y mejorar la precisión del modelo. En la predicción de demanda, Random Forest ha demostrado un desempeño robusto para estimar el consumo diario o semanal, considerando factores como la temperatura, el tipo de día o el comportamiento histórico de la carga [2] [9].

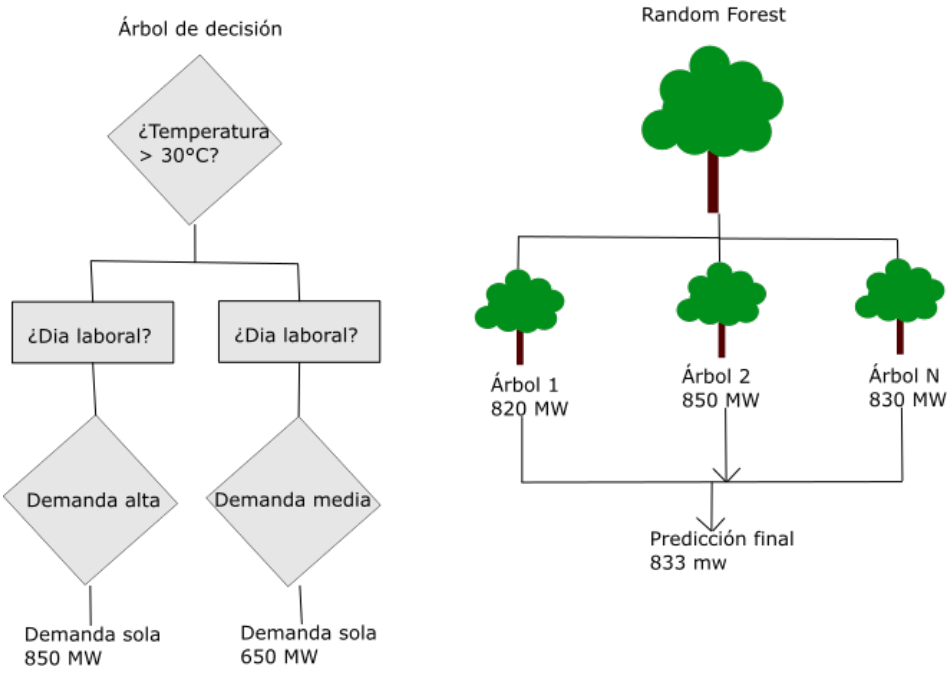

Figura 2.2: Predictor Random Forrest

Estructura

- **Nodo raíz:** representa todos los datos.
- **Nodos internos:** aplican una división en base a alguna característica.
- **Hojas o nodos terminales:** predicen un valor numérico (particionamiento recursivo).

Criterios de partición

Error cuadrático medio (MSE)

Es la reducción de la varianza del valor objetivo dentro de los nodos hijos; es decir, busca minimizar el error cuadrático medio (MSE).

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\gamma_i - \bar{y})^2$$

Donde: γ_i: valores observados, $\bar{y}$: promedio dentro del nodo.

Error absoluto medio (MAE)

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\gamma_i - \bar{y}|$$

El MAE mide el promedio de las desviaciones absolutas entre los valores predichos y los valores reales.

- Un MAE más bajo indica un modelo más preciso.
- A diferencia del MSE, el MAE no eleva los errores al cuadrado, por lo que no penaliza tanto los errores grandes y es más robusto ante valores atípicos.

Bootstrap Aggregation (Bagging)

Consiste en entrenar múltiples modelos (arboles de decisión) sobre muestras de datos obtenidas mediante muestreo con reemplazo (bootstrap) para cada árbol. Esto genera conjuntos de entrenamiento ligeramente distintos para cada árbol, lo cual genera diversidad. La predicción final del ramdom forest es el promedio de las predicciones de todos los árboles. Beneficios clave del bagging • Se reduce la varianza del modelo, comparado con un solo árbol. • Menor sobre ajuste. • Se mantiene la capacidad de modelar relaciones no lineales y complejas gracias a los árboles como base.

Importancia de variables en Random Forest El random forest permite cuantificar que tan influyebtes son las variables.ejemplo los dos métodos más comunes: Reducción promedio de impurezas (MDI) Cuando random forest entrena muchos árboles, cada árbol divide los datos una y otra vez en nodos usando distintas variables.Cada división se realiza para minimizar la impureza del conjunto de datos en ese nodo. • En regresión, la impureza es la varianza de valor objetivo (Y)dentro del nodo. • En clasificación, es el índice Gini o la entropía. En cada división, se calcula cuando disminuye la impureza al usar cierta variable. Esa reducción se suma para cada variable a lo largo de todos los nodos y árboles, luego se promedia [20].

Permutacion

Este método mide que tanto empeora el modelo si se rompe la relación entre una variable y las variables objetivo. En lugar de mirar dentro de los árboles, analiza el modelo

ya entrenado. Los pasos que lleva a cabo son los siguientes:

1. Se mide el rendimiento mediante (MAE O RMSE).
2. Se barajan (permutan) aleatoriamente los valores de una variable, dejando todas las demás igual.
3. Se calcula de nuevo el rendimiento del modelo con los datos permutados.
4. La diferencia de esa variable es la diferencia entre ambos errores.
5. Se repite en todas las variables.

2.20.3. Support Vector Regression (SVR)

Las Máquinas de Soporte Vectorial (SVM) son algoritmos de aprendizaje supervisado utilizados primordialmente para clasificación y, en su extensión (SVR), para regresión [4] [11] [13]. Su principio fundamental es encontrar un hiperplano óptimo que separe las clases o ajuste los datos con el máximo margen posible.

Normalización o escalado de variables

Debido a que los modelos SVM son sensibles a la escala de los datos, es necesario aplicar procesos de normalización o estandarización.

División del conjunto de datos

Una vez procesados los datos, estos se dividen en subconjuntos de entrenamiento y prueba (por ejemplo, 70/30 o 80/20). El conjunto de entrenamiento se utiliza para ajustar los parámetros del modelo, mientras que el de prueba se emplea para evaluar su capacidad de generalización sobre datos no vistos.

Validación cruzada

Se utiliza esta técnica para evaluar de manera más robusta el rendimiento del modelo y optimizar los hiperparámetros (como C y γ). Esto permite reducir la varianza en los

resultados y seleccionar la configuración más adecuada.

Técnica: Validación cruzada *k*-fold Consiste en dividir los datos en k subconjuntos (o *folds*) de tamaño aproximadamente igual. Un subconjunto se utiliza para validar el modelo entrenado con los subconjuntos restantes. El proceso se repite k veces, de modo que cada subconjunto se emplea exactamente una vez como validación. El error promedio en las k particiones se reporta como ε. Esta es una de las técnicas más populares, aunque puede tardar en ejecutarse debido al entrenamiento repetido del modelo. (MathWorks, 2025)

Entrenamiento y evaluación final

El modelo SVM se entrena con el conjunto de datos procesados y se evalúa usando métricas de desempeño como MAE, RMSE o R^2, verificando su capacidad predictiva sobre datos nuevos.

2.21. Función del Kernel

Cuando los datos no son linealmente separables, las SVM utilizan una función *kernel* que transforma los datos a un espacio de mayor dimensión donde la separación es posible [3] [12]. Los tipos de kernel y sus características se describen a continuación

Tipo de kernel	**Ecuación**	**Características**
Lineal	$K(x_i, x_j) = x_i^T x_j$	Rápido cuando los datos son linealmente separables o casi lineales.
Polinomial	$K(x_i, x_j) = (x_i^T x_j + 1)^d$	Captura relaciones no lineales.
RBF (Gaussiano)	$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$	Muy utilizado y flexible para patrones complejos.

2.22. Parámetros C y γ

En las Máquinas de Soporte Vectorial, los parámetros C y γ desempeñan un papel fundamental en el rendimiento del modelo [8] [21] [5].

- **Parámetro C:** Controla el equilibrio entre la complejidad del modelo y el grado de error permitido. Funciona como un término de regularización.
- **Parámetro γ:** Determina el alcance de influencia de cada punto de datos en los modelos con *kernel* no lineales, afectando la forma y suavidad de la función de decisión.

La correcta elección de estos parámetros es clave para obtener un modelo con alta capacidad predictiva.

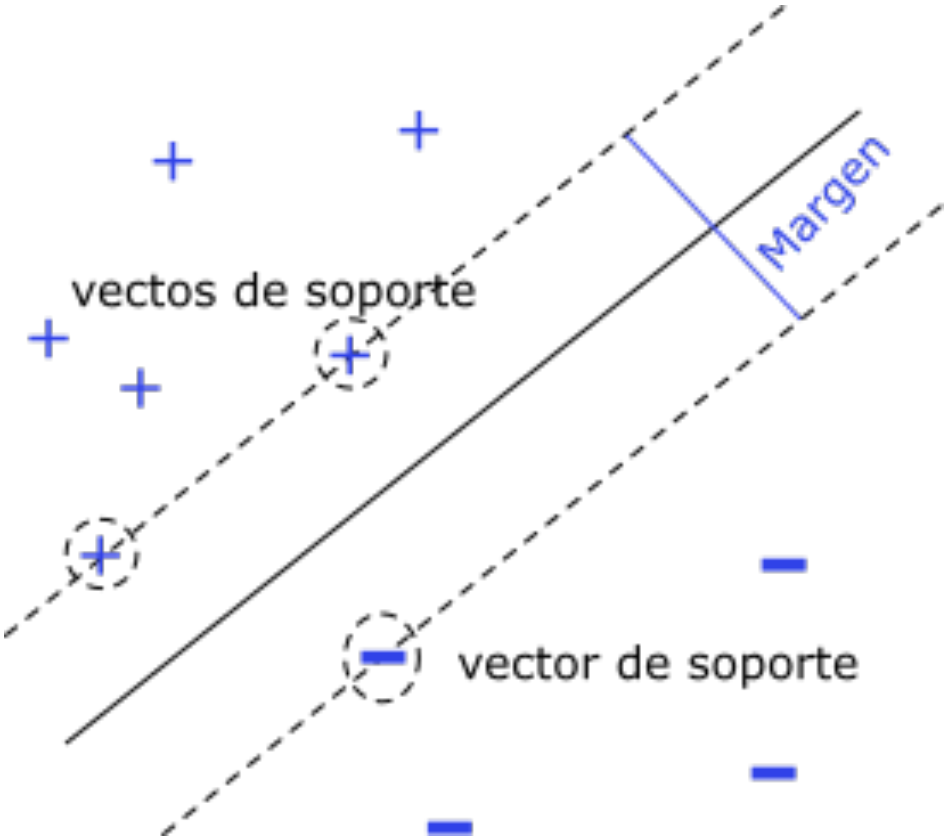

Figura 2.3: Metodología propuesta

2.23. Preprocesamiento y Preparación de Datos

El preprocesamiento constituye una etapa fundamental en el desarrollo de modelos de aprendizaje automático, ya que garantiza la calidad y coherencia de la información antes del entrenamiento del modelo.

2.23.1. División de datos (train–test split): entrenamiento, validación y prueba

En un modelo predictivo, la división de datos permite evaluar el rendimiento real del modelo. Su objetivo es separar los datos en diferentes subconjuntos que faciliten el entrenamiento, ajuste y evaluación sin sesgos.

En el caso de series temporales, este proceso presenta características particulares: los datos están ordenados cronológicamente y no deben mezclarse aleatoriamente (no se utiliza *shuffle*). En su lugar, se aplican cortes temporales que respetan la secuencia temporal.

Los conjuntos se definen de la siguiente manera:

- **Entrenamiento (Train):** contiene los datos más antiguos y se utiliza para ajustar el modelo.
- **Validación (Validation):** corresponde a un periodo posterior al entrenamiento y sirve para ajustar hiperparámetros o seleccionar el mejor modelo.
- **Prueba (Test):** incluye los datos más recientes y se utiliza para evaluar el desempeño final con información no vista.

2.23.2. Técnicas comunes de división en series temporales

- **Simple Split:** división única respetando el orden temporal.
- **TimeSeriesSplit (Expanding Window):** el modelo se entrena con una ventana de tiempo que se expande progresivamente, añadiendo nuevos datos en cada iteración.

- **Rolling Window:** mantiene una ventana de tamaño fijo que se desplaza en el tiempo, reemplazando datos antiguos por nuevos.

2.23.3. Escalado de variables: normalización Min–Max y estandarización Z-score

El escalado consiste en ajustar la magnitud de las variables numéricas para que todas tengan una escala comparable. Este paso es crucial en modelos sensibles a la escala, como las máquinas de soporte vectorial (SVM) y las redes neuronales (NN).

Los métodos de escalado más comunes son los siguientes:

Min–Max Scaling (Normalización)

Transforma los valores a un rango definido, generalmente $[0, 1]$, mediante la expresión:

$$x' = \frac{x - x_{\text{mín}}}{x_{\text{máx}} - x_{\text{mín}}}$$

Este método conserva la relación entre los valores originales, pero puede ser afectado por la presencia de valores atípicos (*outliers*).

Z-score Scaling (Estandarización)

Ajusta los datos para que tengan media 0 y varianza 1, calculándose como:

$$x' = \frac{x - \mu}{\sigma}$$

donde μ es la media y σ la desviación estándar de la variable. Este enfoque es más robusto ante valores extremos.

2.23.4. Transformación temporal: ventanas deslizantes (sliding windows)

La técnica de ventanas deslizantes es fundamental para transformar series temporales en un formato adecuado para modelos de aprendizaje supervisado, como redes neuronales

o modelos de regresión.

El objetivo principal es predecir valores futuros utilizando un número fijo de observaciones pasadas. Esta técnica permite que los modelos aprendan relaciones temporales y patrones en los datos.

Se define una ventana temporal de tamaño L (*lag window*) que contiene las observaciones pasadas, junto con un horizonte de predicción h que indica cuántos pasos hacia adelante se desean pronosticar.

La transformación se expresa como:

$$X_t = [y_{t-L+1}, y_{t-L+2}, \ldots, y_t]$$

$$y_{t+h} = \text{valor objetivo a predecir}$$

De esta manera, la serie original se convierte en un conjunto de pares (X_t, y_{t+h}), donde X_t contiene las observaciones pasadas y y_{t+h} representa el valor futuro correspondiente.

2.24. Evaluación de Modelos

2.24.1. Validación cruzada (Cross-Validation): K-fold y Time Series Split

La validación cruzada es una técnica utilizada para evaluar el rendimiento y la capacidad de generalización de un modelo predictivo. Consiste en dividir el conjunto de datos en varios subconjuntos (*folds*) y entrenar el modelo repetidamente, utilizando parte de los datos para entrenamiento y otra parte para validación.

K-fold Cross-Validation

La validación cruzada k-fold divide los datos en k particiones de igual tamaño. En cada iteración, una de las particiones se utiliza como conjunto de validación y las restantes como conjunto de entrenamiento.

La fórmula general del error de validación cruzada es:

$$E_{\text{cv}} = \frac{1}{k} \sum_{i=1}^{k} E_i$$

donde: k: número de divisiones o *folds*. E_i: error del modelo en el i-ésimo fold (por ejemplo, RMSE, MAE o R^2). E_{cv}: error promedio de validación cruzada.

El algoritmo primero divide los datos en k particiones iguales. En cada iteración, un subconjunto se utiliza como validación y los restantes como entrenamiento. El proceso se repite k veces. Finalmente, se promedian los errores obtenidos para generar una medida global del rendimiento.

Time Series Split

En series temporales, el orden cronológico no puede alterarse, por lo que se utiliza una variante de k-fold denominada *TimeSeriesSplit*, la cual respeta la secuencia temporal al realizar los cortes.

La fórmula es análoga:

$$E_{\text{tscv}} = \frac{1}{k} \sum_{i=1}^{k} E_i$$

La diferencia radica en cómo se realizan los cortes. En cada iteración, el modelo se entrena con los datos anteriores en el tiempo y se valida con los datos más recientes y no se mezclan observaciones pasadas con futuras.

Este algoritmo es ideal para modelos predictivos dependientes del tiempo (por ejemplo, predicción de demanda eléctrica) y permite evaluar cómo se comporta el modelo al predecir datos futuros basándose en información histórica.

2.24.2. Overfitting y Underfitting

Estos fenómenos describen problemas relacionados con la capacidad de generalización de un modelo predictivo.

Overfitting (Sobreajuste)

El sobreajuste ocurre cuando un modelo aprende demasiado bien los datos de entrenamiento, incluyendo ruido o patrones irrelevantes. Como consecuencia, muestra un alto rendimiento en entrenamiento, pero un bajo desempeño en datos nuevos.

Las causas principales suelen ser los modelos excesivamente complejos (demasiados parámetros o capas), conjuntos de datos pequeños o con ruido. En otras ocasiones se debe a un entrenamiento prolongado sin aplicar regularización y tambien debiso a una falta de validación cruzada adecuada.

Para prevenir este problema se deben implementar técnicas de regularización (por ejemplo, *dropout* o penalización de complejidad). Otra solución podría ser aumentar el conjunto de datos y aplicación de una validación cruzada robusta.

Underfitting (Subajuste)

El subajuste se presenta cuando el modelo no logra aprender los patrones relevantes de los datos, evidenciándose en un bajo desempeño tanto en entrenamiento como en prueba.

Las causas de un subajuste se deben principalmente a modelos demasiado simple para la complejidad del problema, entrenamiento insuficiente, falta de variables o características representativas y a un preprocesamiento deficiente o selección de atributos inadecuada.

Para prevenir este problema se debe incrementar la complejidad del modelo, mejorar el entrenamiento (más iteraciones o mejor optimización), incorporar más características o variables relevantes y revisar el preprocesamiento y la representación de los datos.

2.25. Métricas de Desempeño

Las métricas de desempeño permiten evaluar la precisión y calidad de un modelo predictivo. En problemas de regresión, donde se predicen valores continuos (como la demanda eléctrica), se emplean diversas medidas para cuantificar la diferencia entre los valores reales y los predichos.

2.25.1. Error Absoluto Medio (MAE)

El MAE mide el promedio de los errores absolutos entre las predicciones y los valores reales. Indica, en promedio, cuánto se equivoca el modelo sin considerar la dirección del error [7] [24].

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Un MAE más pequeño implica mejor desempeño. Es intuitivo, aunque no penaliza fuertemente los errores grandes.

2.25.2. Error Cuadrático Medio (MSE)

El MSE calcula el promedio del cuadrado de las diferencias entre los valores reales y los predichos. Penaliza más los errores grandes debido al cuadrado.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Un MSE pequeño indica mejor ajuste y es útil cuando se desea castigar con mayor fuerza las desviaciones severas.

2.25.3. Raíz del Error Cuadrático Medio (RMSE)

El RMSE es la raíz cuadrada del MSE y tiene la ventaja de estar en las mismas unidades que la variable predicha [7].

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

La RMSE refleja la magnitud promedio del error. valores más bajos representan un mejor ajuste.

2.25.4. Porcentaje de Error Absoluto Medio (MAPE)

El MAPE mide el error en términos porcentuales, comparando el error absoluto con el valor real [10].

$$\text{MAPE} = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

El MAPE expresa el error promedio como porcentaje, lo que facilita la comparación entre modelos o conjuntos de datos distintos. No es recomendable cuando existen valores reales iguales o cercanos a cero.

2.25.5. Coeficiente de Determinación (R^2)

El coeficiente R^2 indica qué proporción de la variabilidad de los datos reales es explicada por el modelo.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

donde $\bar{y}$ es el promedio de los valores reales.

El coeficiente R^2 puede ser interpretado como sigue:

1. $R^2 = 1$: ajuste perfecto.
2. $R^2 = 0$: el modelo no explica la variabilidad.
3. $R^2 < 0$: el modelo es peor que una predicción constante.

2.25.6. Error Promedio Porcentual (MPE)

El MPE evalúa el sesgo promedio del modelo, considerando el signo del error. Un MPE positivo indica subestimación del modelo; uno negativo, sobreestimación [10].

$$\text{MPE} = \frac{100}{n} \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{y_i} \right)$$

El MPE ayuda a determinar si el modelo tiende a predecir valores por encima o por debajo de los reales.

2.26. Optimización de hiperparámetros

La optimización de hiperparámetros es el proceso de encontrar la mejor combinación de parámetros que maximice el rendimiento de un modelo de aprendizaje automático. A diferencia de los parámetros internos, los hiperparámetros se definen antes del entrenamiento y su correcta selección mejora la precisión, la capacidad de generalización y la estabilidad del modelo.

Los hiperparámetros controlan aspectos como:

- Complejidad del modelo,
- Regularización,
- Tasa de aprendizaje,
- Tamaño de ventana.

Métodos de optimización

Búsqueda en rejilla (Grid Search) Explora todas las combinaciones posibles de hiperparámetros definidos por el usuario. Garantiza encontrar el mejor conjunto dentro del rango especificado, aunque puede ser costosa computacionalmente.

Este métodos es exhaustivo y sistemático, pero requiere un alto tiempo de cómputo.

Búsqueda aleatoria (Random Search) Selecciona aleatoriamente un número limitado de combinaciones dentro de los rangos definidos. Es más eficiente en tiempo y suele funcionar bien en la práctica.

Encuentra una solución de forma rápida y escalable. Sin embargo, no garantiza encontrar el mejor conjunto absoluto.

Validación temporal (Time Series Cross-Validation) En series temporales no es posible mezclar los datos debido a la importancia del orden cronológico. En este método, los datos se dividen en ventanas crecientes: se entrena con un bloque pasado y se valida con el siguiente, repitiendo el proceso varias veces.

Este método es adecuado para datos secuenciales y evita fugas temporales, aunque suele ser más lento que la validación tradicional.

2.26.1. Visualización y análisis de resultados

La visualización de series temporales permite comprender el comportamiento de los datos, validar modelos predictivos y detectar patrones como tendencias, estacionalidades o anomalías. En la predicción de demanda eléctrica, los gráficos permiten observar cómo varía el consumo a lo largo del tiempo y evaluar la precisión de los modelos.

Gráficos de series temporales

Gráficos de tendencias Muestran el comportamiento a largo plazo de la serie, indicando si los valores aumentan, disminuyen o se mantienen.

Las técnicas más utilizadas son gráfico de líneas, suavizado y descomposición de series (tendencia, estacionalidad y residuo).

2.26.2. Correlogramas

Un correlograma es una representación gráfica de la autocorrelación de una serie temporal en distintos retardos o rezagos. Mide el grado de similitud entre una serie y una versión desplazada de sí misma. Su valor oscila entre -1 y $+1$.

$$r_k = \frac{\sum_{t=k+1}^{n}(y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{n}(y_t - \bar{y})^2}$$

Donde r_k: autocorrelación en el retardo k, y_t: valor en el tiempo t, $\bar{y}$: media de la serie y n: número total de observaciones.

Capítulo 3

Metodología

En este capítulo se presenta la metodología llevada a cabo al realizar los experimentos. La Figura 3.1 ilustra de forma general el procedimiento metodológico. En los experimentos se utilizó un conjunto de datos, el cual fue preprocesado. Una vez preparado el conjunto de datos, se seleccionaron los modelos de aprendizaje automático empleados para la predicción de demanda electrica. Posteriormente, se llevó a cabo el entrenamiento y la validación de los modelos, evaluando su desempeño mediante diferentes métricas. Finalmente, se compararon los resultados obtenidos con el fin de determinar el modelo con mejor rendimiento.

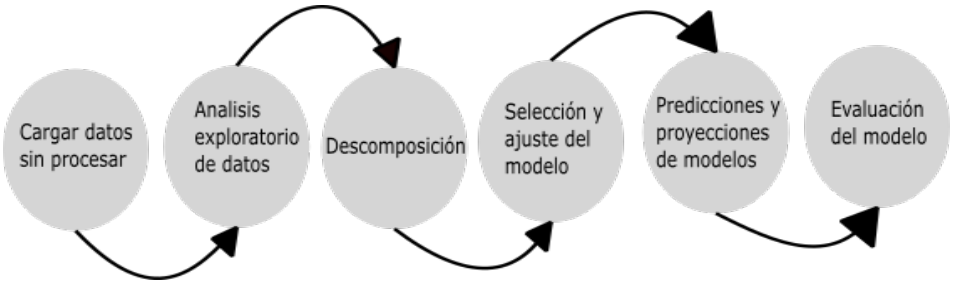

Figura 3.1: Metodología propuesta

3.1. Conjunto de datos

El conjunto de datos se obtuvó del sitio web de la Agencia Internacional de Energía (AIE), específicamente de su herramienta de Estadísticas Mensuales de Electricidad, dis-

ponible en *iea.org/data−and−statistics/data−tools/monthly−electricity−statistics*.

La siguiente liga muestra la metodología y espacios de tiempo utilizados para recopilar los datos *github.com/ccan*23/*iea_electricity_generation_data_scraper*.

Los datos incluyen información sobre la producción de energía en varios países mensualmente, de 2010 a 2022. La producción de energía se mide en gigavatios-hora (GWh) y abarca varios medios para obtener energia, como la hidroeléctrica, la eólica, la solar, la geotérmica, la nuclear, los combustibles fósiles y otros.

Los paises donde se obtuvo información para el conjunto de datos son los siguientes: Argentina, Australia, Austria, Bélgica, Brasil, Bulgaria, Canadá, Chile, Colombia, Costa Rica, Croacia, Chipre, República Checa, Dinamarca, Estonia, Finlandia, Francia, Alemania, Grecia, Hungría, Total AIE, Islandia, India, Irlanda, Italia, Japón, Corea, Letonia, Lituania, Luxemburgo, Malta, México, Países Bajos, Nueva Zelanda, Macedonia del Norte, Noruega, OCDE Américas, OCDE Asia Oceanía, OCDE Europa, Total OCDE, República Popular China, Polonia, Portugal, República de Turquía, Rumania, Serbia, República Eslovaca, Eslovenia, España, Suecia, Suiza, Reino Unido, Estados Unidos

Las fuentes de obtención de energia donde se obtuvieron datos son las siguientes: Hidroeléctrica, Eólica, Solar, Geotérmica, Otras renovables, Nuclear, Total de combustibles, Carbón, Petróleo, Gas natural, Combustibles renovables, Otros combustibles no renovables, No especificado, Producción neta de electricidad, Importaciones totales, Exportaciones totales, Electricidad suministrada, Utilizada para almacenamiento por bombeo, Pérdidas de distribución, Consumo final, Comercio de electricidad, Renovables, No renovables, Otros, Otras renovables agregadas, Bajo Carbono, combustibles fósiles

Las columnas del conjunto de datos incluyen:

1. **PAÍS**: Nombre del país
2. **CODE_TIME**: Código que representa el mes y el año (p. ej., ENE2010 para enero de 2010)
3. **TIME**: Mes y año en un formato más legible (p. ej., enero de 2010)
4. **YEAR**: Año del punto de datos

5. **MONTH**: Mes del punto de datos como número (1-12)

6. **MONTH_NAME**: Mes del punto de datos como cadena (p. ej., enero)

7. **PRODUCT**: Tipo de producto energético (p. ej., hidroeléctrico, eólico, solar)

8. **VALUE**: Cantidad de electricidad generada en gigavatios-hora (GWh)

9. **DIPLAY_ORDER**: Orden en que se deben mostrar los productos

10. **yearToDate**: Cantidad de electricidad generada del año en curso hasta el mes actual en GWh

11. **previousYearToDate**: Cantidad de electricidad generada del año anterior hasta el mes actual mes en GWh

12. **Participación**: La participación del producto en la generación total de electricidad del país en formato decimal.

Este conjunto de datos contiene 181915 instancias, cada una con los atributos antes mencionados. Sin embargo, para propositos de esta tesis, solo se utilizaron los datos para México. Para este subconjunto el número de datos es de 3958 entradas. Las gráficas mostradas en las Figuras **??** muestran los diferentes intervalos en los que varian cada uno de los atributos. Estas Figuras permiten visualizar posibles valores atípicos e identificar tendencias. La frecuencia del conjunto de datos o la granulidad de este es mensual.

Las Figuras 3.2 y 3.3 muestran los porcentajes de electricidad por fuentes de energia y su evolución desde 2010 a 2022 respectivamente. La Figura 3.4 muestra la proporción de producción de energía eléctrica divididas en energia renovables y no renovables.

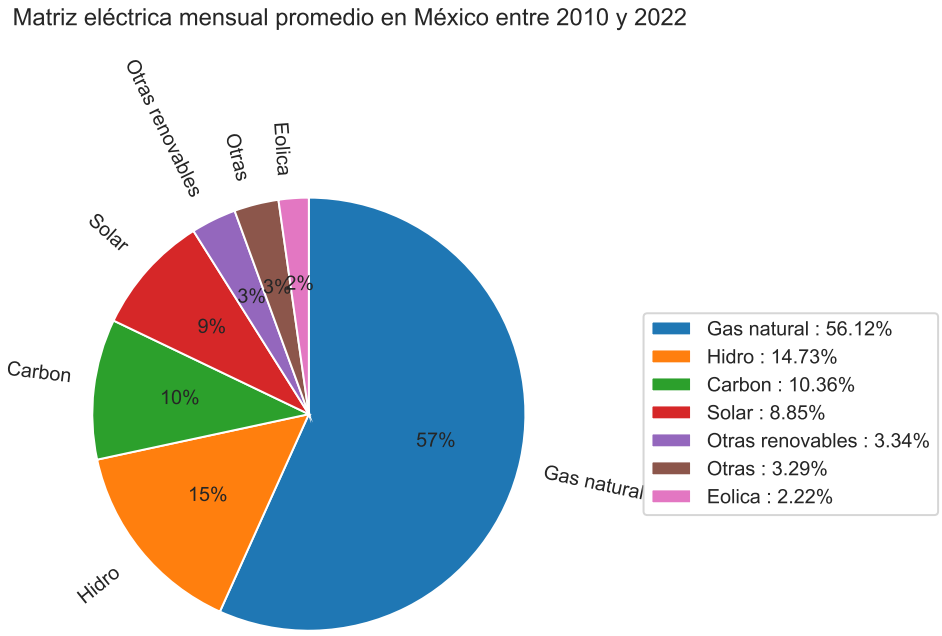

Figura 3.2: Porcentajes de generación de electricidad por fuentes de energia en México

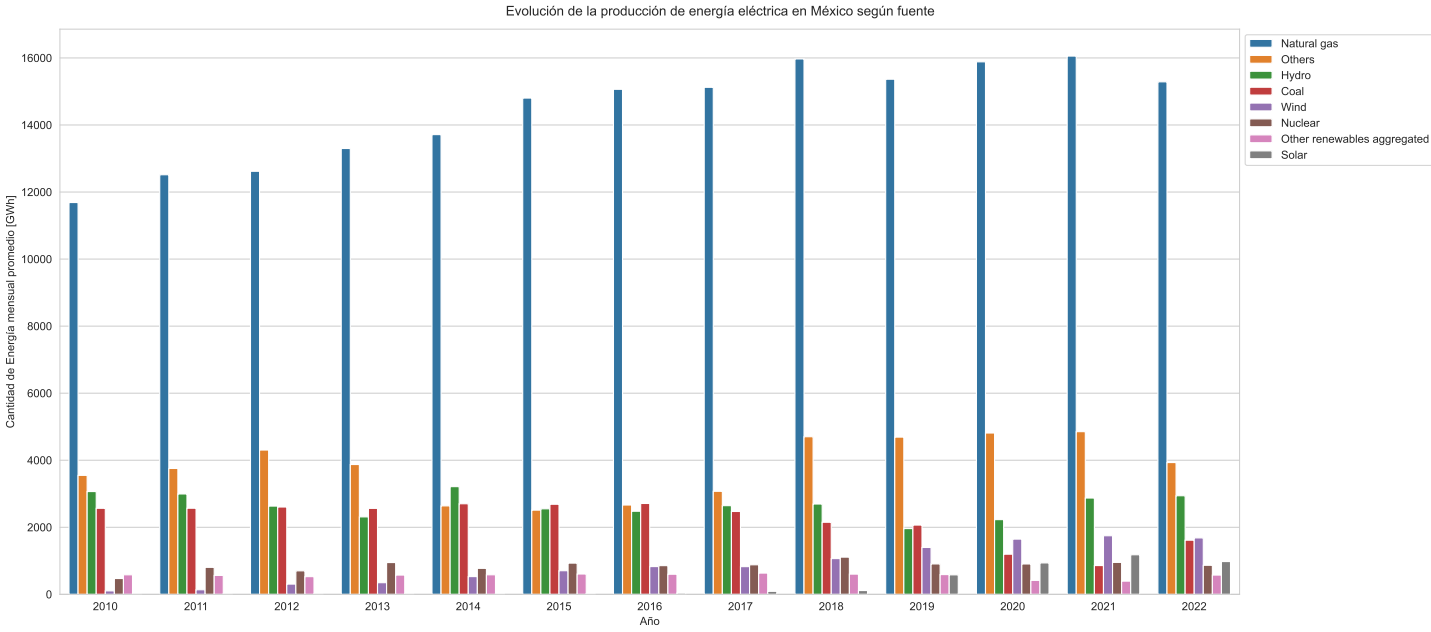

Figura 3.3: Evolución de la producción de energia en Mexico (2010-2022)

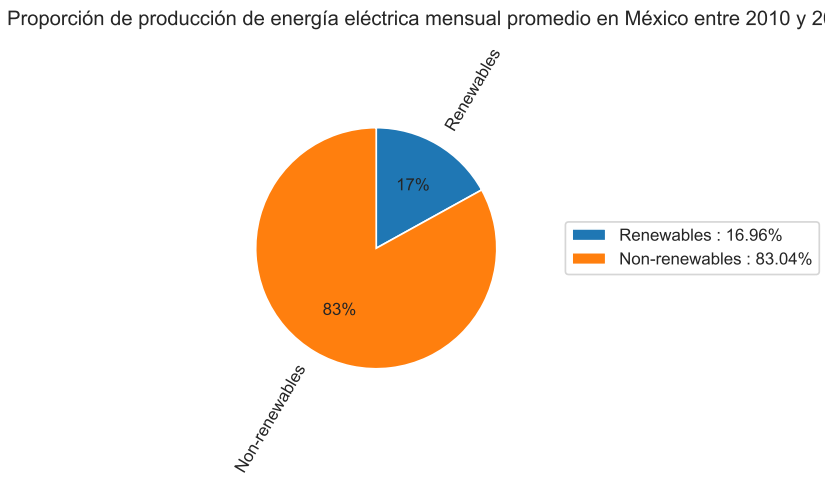

Figura 3.4: Proporción de producción de energía eléctrica divididas en energia renovables y no renovable

3.2. Pre-procesamiento

Como primer paso del proceso de preprocesamiento de datos, se llevó a cabo una etapa de limpieza y verificación de integridad con el objetivo de mejorar la confiabilidad y el desempeño de las técnicas de predicción empleadas. Esta fase es muy importante, ya que la calidad de los datos influye directamente en la capacidad del modelo para aprender patrones significativos y generar predicciones precisas.

Durante esta etapa se verificó la completitud y consistencia del conjunto de datos, confirmándose que no existen valores faltantes ni registros anómalos. Por lo tanto, no fue necesario aplicar técnicas de imputación o eliminación de observaciones incompletas.

En este caso particular, el conjunto de datos está conformado por una única variable de entrada, correspondiente a la producción total de energía eléctrica expresada en gigavatios (GW). Dicha variable representa la cantidad total de energía generada, independientemente de la fuente de origen (por ejemplo, solar, eólica, hidroeléctrica o térmica). En consecuencia, el análisis se centra exclusivamente en la magnitud de la producción energética, sin distinguir entre las distintas fuentes generadoras.

Como segundo paso del preprocesamiento, se procedió a la normalización de las variables. Este procedimiento consiste en ajustar los valores de las observaciones a una

media igual a cero y una desviación estándar igual a uno, empleando la técnica conocida como escalado tipificado (standard scaling). La normalización permite homogeneizar las escalas de las variables, facilitando la comparación y el entrenamiento de los modelos, especialmente en aquellos algoritmos sensibles a la magnitud de los datos.

Además, la aplicación de técnicas de normalización contribuye a reducir la varianza no deseada y a mejorar la estabilidad numérica durante el proceso de optimización de los modelos de aprendizaje automático, lo que se traduce en un mayor rendimiento y generalización de los resultados obtenidos.

3.3. Selección de Modelos

Para la predicción de radiación se seleccionaron varias técnicas de IA. Los modelos seleccionados son descritos a continuación, sin embargo para más detalle fueron descritos en los preliminares:

3.3.1. Regresión lineal

Esta técnica se seleccionó debido a su simplicidad, facilidad para implementación, además de ello, es muy facil de interpretar y computacionalmente muy económico. A pesar de las ventajas mencionadas anteriormente, este método a menudo presenta limitaciones significativas debido a que la mayoría de los problemas del mundo real son no lineales y el uso de este algoritmo puede no tener un desempeno adecuado en muchos problemas reales.

3.3.2. Árboles de decisión para regresión

Esta técnica se seleccionó debido a su interpretabilidad. Esta es una de sus principales ventajas, ya que el modelo puede visualizarse fácilmente como un conjunto de reglas "si-entonces", facilitando su comprensión incluso por parte de personas sin experiencia. El algoritmo puede manejar automáticamente atributos categóricos y numéricos, y a diferencia de otras técnicas de IA no requiere una normalización previa de las variables

[18].

Sin embargo, a pesar de estas grandes ventajas, los árboles de decisión pueden sufrir debido a que pequeños cambios en los datos de entrenamiento pueden generar árboles significativamente distintos. Otra desventaja de estos radica en que son susceptibles al sobreajuste si no se realiza una adecuada poda o regularización.

3.3.3. SVM para regresión

Entre las ventajas de SVM para regresión estan su capacidad para modelar relaciones no lineales complejas. Estas tienen una gran capacidad para controlar el sobreajuste mediante el parámetro C y el margen ε.

Sin embargo, las principales desventajas incluyen su alto costo computacional en conjuntos de datos grandes y la dificultad para interpretar el modelo resultante. Además, su desempeño puede depender significativamente de la correcta selección del kernel y de sus parámetros[6].

3.4. Entrenamiento

Para validar los resultados, se empleó un esquema de validación de series de tiempo "walk-forward validation". En este método, el conjunto de datos se divide respetando el orden cronológico, de modo que los modelos se entrenan con los datos disponibles hasta un punto en el tiempo y se validan con el siguiente intervalo. El método de validación utilizado pretende evitar la aleatorización lo que induce fugas de información en las fases de entrenamiento y prueba. La ventana de entrenamiento se amplía o desplaza en cada iteración de forma progresiva hacia adelante en el tiempo, repitiendo el proceso en múltiples iteraciones. De esta manera, se simula un escenario realista en el que el modelo se entrena únicamente con información disponible hasta un instante determinado y se valida con datos futuros no observados.

Finalmente, las métricas de desempeño obtenidas en cada iteración se promedian para obtener una estimación global de la capacidad predictiva del modelo, El uso del método

de validación proporciona una evaluación más robusta y esta es realmente representativa del comportamiento de los datos ante nuevas observaciones temporales.

3.5. Evaluación de desempeño

Para evaluar los modelos y las técnicas de balanceo de datos se utilizaron varias metricas de desempeño que se describen en esta sección.

1. **Error Cuadrático Medio (MSE)**: calcula el promedio de los errores elevados al cuadrado:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

   Penaliza con mayor fuerza los errores grandes, lo cual lo hace sensible a *outliers*.

2. **Raíz del Error Cuadrático Medio (RMSE)**: es la raíz cuadrada del MSE, expresada en las mismas unidades que la variable de salida:

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

   Es ampliamente usada por su interpretabilidad y capacidad de resaltar errores importantes.

3. **Raíz del Error Cuadrático Medio Normalizado (Normalized RMSE).** La raíz del error cuadrático medio normalizado (Normalized RMSE) permite evaluar el desempeño del modelo de regresión en relación con la escala de la variable objetivo. Se define como:

$$\text{RMSE}_{\text{normalizado}} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}}{\text{máx}(y) - \text{mín}(y)}$$

   donde y_i representa el valor real, $\hat{y}_i$ es la predicción del modelo, y n es el número total de observaciones. Esta métrica expresa el error promedio de predicción relativo

al rango de los valores reales, por lo que es útil para comparar el rendimiento del modelo entre diferentes conjuntos de datos. Valores cercanos a 0 indican un mejor ajuste del modelo.

4. **Coeficiente de Determinación (R^2).** El coeficiente de determinación, conocido como R^2, mide la proporción de la varianza de la variable dependiente que puede ser explicada por el modelo. Su fórmula es:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

donde $\bar{y}$ es la media de los valores reales y_i. El valor de R^2 se encuentra en el intervalo $(-\infty, 1]$, donde un valor de 1 indica un modelo con ajuste perfecto, y valores cercanos o menores a 0 indican que el modelo no mejora con respecto a una predicción basada únicamente en la media. Esta métrica es ampliamente utilizada por su capacidad para resumir el desempeño general del modelo en un solo valor.

Intervalos de confianza e incertidumbre
(tendencia/estacionalidad).
Autocorrelación (ACF) y autocorrelación parcial (PACF) para ver lags relevantes.
d) Tendencia explícita

Capítulo 4

Resultados experimentales

Los resultados experimentales fueron realizados sobre el conjunto de datos que se describe en el Capítulo anterior. En este Capítulo se muestran los resultados obtenidos utilizando 4 técnicas clásicas de inteligencia artificial y 3 técnicas de estadística clásica. En cada uno de los modelos entrenados fueron optimizados los parámetros para cada modelo. La primera sección muestra los resultados de la predicción de los modelos implementados, estos son mostrados utilizando 4 métricas distintas. La primera Sección muestra el análisis de los modelos estadísticos. La segunda Sección muestra el análisis de los modelos de aprendizaje máquina. La Tercera Sección muestra un análisis comparativo de ambos.

4.1. Modelos estadísticos de series temporales

En esta sección se muestra el desempeño de 3 modelos estadísticos de series temporales. La Figura 4.1 muestra los datos reales mes a mes de demanda electrica del año 2014 al 2022. Como se describio en el Capítulo anterior los tres modeos empleados son ARIMA,SARIMA y SARIMAX. En los experimentos llevados a cabo se utilizaron cinco métricas de desempeño. La Tabla 4.1 muestra en sus columnas las métricas utilizadas.

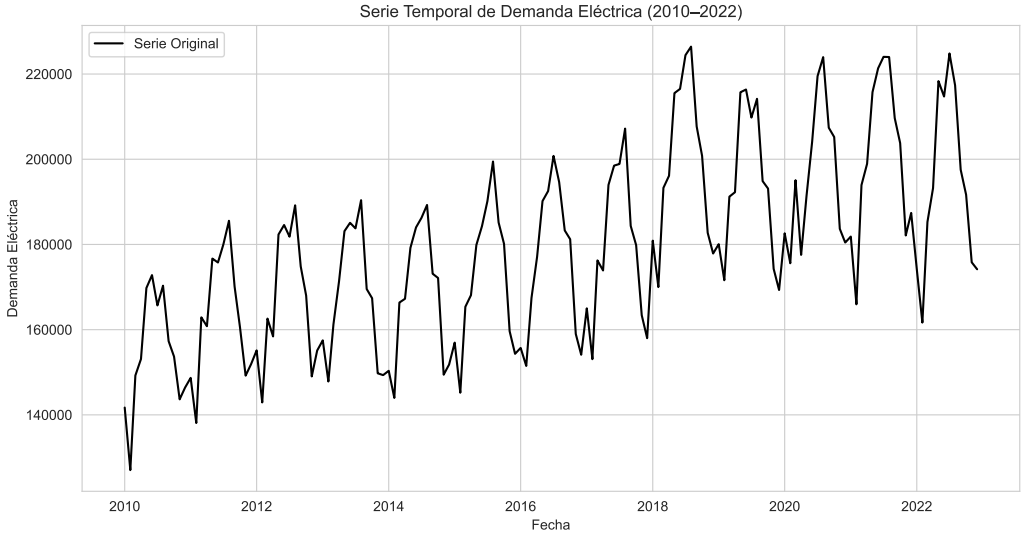

Figura 4.1: Grafica de demanda de electricidad en México por mes

En esta Sección se muestra un análisis comparativo entre los modelos ARIMA, SARIMA y SARIMAX durante el periodo 2019–2022 que permite evaluar su desempeño en la predicción de series temporales, considerando las métricas de error y ajuste más relevantes.

Como se puede apreciar en los resultados de la Tabla 4.1 Los tres modelos presentan valores negativos del coeficiente de determinación (R^2), esto muestra que los modelos poseen un bajo poder predictivo y además que las predicciones no logran seguir adecuadamente la tendencia real de los datos.

De los tres modelos observados, es claro que SARIMAX muestra un desempeño ligeramente superior en la mayoría de los años Aunque como lo muestra la figura 4.2 los resultados obtenidos con el modelo SARIMA son muy cercanos a este. Sin embargo los mejores resultados son obtenidos en 2019, donde obtiene un R^2 = -0.2240, que es el mejor resultados dentro de los valores negativos, lo que refleja una menor desviación respecto a las observaciones reales.

Tabla 4.1: Métricas de Desempeño de Modelos ARIMA, SARIMA y SARIMAX (2019–2022)

Año	**Modelo**	**MAE**	**MSE**	**RMSE**	**MAPE (%)**	**MPE (%)**	**R^2**
2019	ARIMA	18900.25	5.32×10^8	23056.94	9.28	-7.54	-0.9138
2019	SARIMA	18572.25	3.67×10^8	19160.64	9.65	9.65	-0.3217
2019	SARIMAX	17839.77	3.40×10^8	18439.08	9.27	9.27	-0.2240
2020	ARIMA	18269.06	5.68×10^8	23831.42	8.79	-8.59	-1.2984
2020	SARIMA	32284.68	1.12×10^9	33489.64	16.62	16.62	-3.5388
2020	SARIMAX	31260.02	1.06×10^9	32500.64	16.10	16.10	-3.2747
2021	ARIMA	25037.80	8.63×10^8	29384.23	11.92	-10.75	-1.6141
2021	SARIMA	41626.01	1.75×10^9	41778.93	20.96	20.96	-4.2846
2021	SARIMAX	40074.99	1.62×10^9	40235.10	20.18	20.18	-3.9012
2022	ARIMA	20545.62	6.68×10^8	25842.05	9.97	-7.49	-0.6786
2022	SARIMA	63280.63	4.02×10^9	63431.36	33.09	33.09	-9.1134
2022	SARIMAX	61321.49	3.78×10^9	61478.40	32.07	32.07	-8.5002

En la Tabla 4.1 se puede observar que os valores de MAE (Error Absoluto Medio) y RMSE (Raíz del Error Cuadrático Medio) son muy altos en todos los modelos, esto muestra que existen errores significativos en magnitud.

Sin embargo es claro tanto en la Tabla 4.1 como en la Figura 4.2 que SARIMAX presenta los menores valores de error en 2019, con un MAE de 17,839.77 y un RMSE de 18,439.08, estos resultados muestran que SARIMAX tiene una mayor precisión relativa en comparación con ARIMA y SARIMA en ese año.

Por otro lado, la Tabla muestra que el año 2022 muestra un deterioro generalizado en el desempeño de todos los modelos, con errores superiores a 60,000 en los modelos SARIMA y SARIMAX, esto muestra en todos los modelos una disminución en la estabilidad temporal o un cambio estructural en los datos.

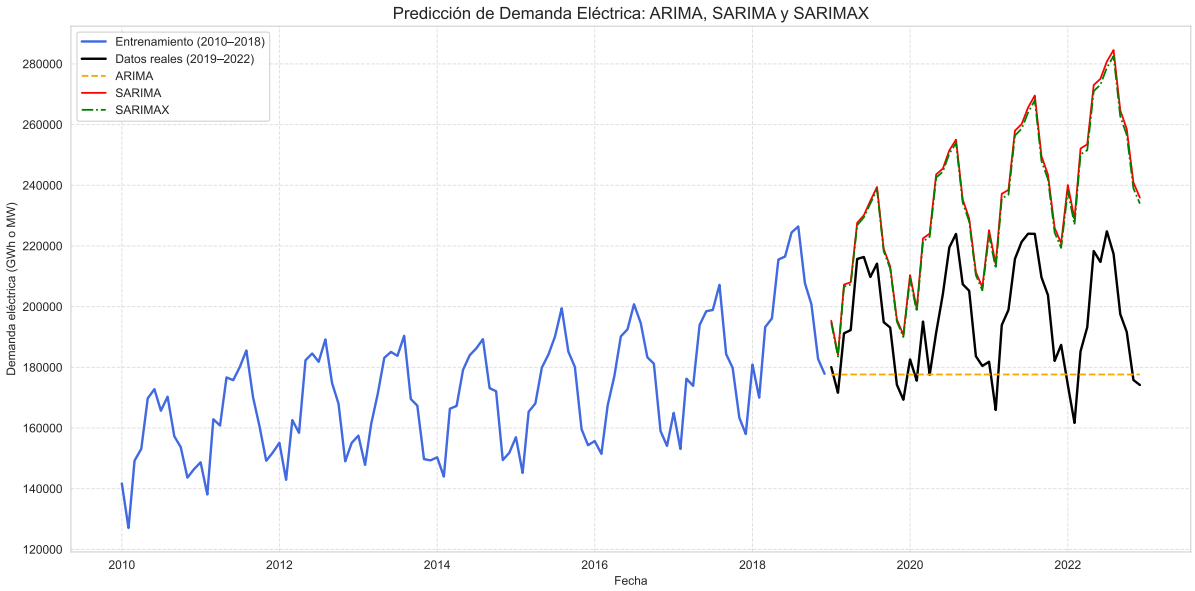

Figura 4.2: Predicción de modelos clásicos

REspecto a los valores de MAPE, estos varían entre 8,7 % y 33 % lo cual representa una precisión de moderada a baja dependiendo del año y el modelo.

La métrica MPE muestra tendencias de sobreestimación o subestimación, destacando valores negativos en los modelos SARIMA y SARIMAX, que indican una tendencia a subestimar los valores reales.

Para todos los modelos evaluados, el año 2019 presenta los mejores resultados generales, esto sugiere una mayor estabilidad en la serie temporal o una mejor capacidad de los modelos para ajustarse a los patrones presentes. Sin embargo, a partir del año 2020, los errores crecen de forma considerable, especialmente en SARIMA, donde el RMSE supera los 33,000, esto podría asociarse a mayor variabilidad o presencia de anomalías en los datos durante ese periodo.

Finalmente, los resultados indican que, aunque SARIMAX ofrece un desempeño ligeramente superior a ARIMA y SARIMA en la mayoría de los años, ninguno de los modelos logra una predicción precisa o estable en todo el periodo analizado.

4.2. Modelos de Aprendizaje Máquina Usados para Predicción

En esta Sección se muestran los resultados obtenidos con modelos de aprendizaje máquina. La Tabla 4.2 resume el los resultados de desempeño comparativo de los cuatro modelos de aprendizaje automático empleados: Regresión Lineal, Árbol de Decisión, Random Forest y SVR en la predicción de la demanda eléctrica durante el periodo 2019–2022. La Tabla 4.2 muesra las métricas de error y ajuste empleadas: MAE (Error Absoluto Medio), MSE (Error Cuadrático Medio), RMSE (Raíz del Error Cuadrático Medio), MAPE (Porcentaje de Error Absoluto Medio), MPE (Porcentaje de Error Medio) y R^2 (Coeficiente de Determinación).

Tabla 4.2: Métricas de Desempeño por Año (2019–2022)

Año	**Modelo**	**MAE**	**MSE**	**RMSE**	**MAPE (%)**	**MPE (%)**	**R^2**
2019	SVR	5205.44	3.96E+07	6292.47	2.73	1.04	0.857
2019	Regresión Lineal	5618.80	4.11E+07	6409.12	2.91	-0.28	0.852
2019	Random Forest	5382.90	4.39E+07	6627.44	2.81	2.28	0.842
2019	Árbol de Decisión	6093.48	6.29E+07	7933.38	3.15	2.98	0.773
2020	Regresión Lineal	7558.01	9.09E+07	9533.21	3.92	0.86	0.632
2020	Árbol de Decisión	6648.77	9.78E+07	9888.81	3.48	1.95	0.604
2020	Random Forest	7084.55	9.87E+07	9934.31	3.68	1.27	0.601
2020	SVR	14066.88	3.74E+08	19340.30	7.24	6.29	-0.514
2021	Árbol de Decisión	2622.58	1.33E+07	3641.36	1.34	-0.67	0.960
2021	Random Forest	4502.32	2.90E+07	5388.20	2.30	-1.33	0.912
2021	Regresión Lineal	4591.08	3.99E+07	6313.80	2.44	0.40	0.879
2021	SVR	23769.46	7.27E+08	26965.16	11.57	11.49	-1.201
2022	Random Forest	5212.03	3.63E+07	6022.01	2.82	2.17	0.909
2022	Árbol de Decisión	5824.72	4.40E+07	6633.63	3.10	2.86	0.889
2022	Regresión Lineal	11555.77	1.79E+08	13382.80	6.32	6.15	0.550
2022	SVR	51621.13	2.95E+09	54356.67	26.15	26.15	-6.427

Los resultados de la Tabla 4.2 muestran que los modelos basados en árboles, especialmente Random Forest, presentaron un desempeño más estable y robusto durante los

años evaluados, manteniendo valores de R^2 superiores a 0.84 en la mayoría de los casos y errores relativamente bajos. La Figura 4.3 muestra claramente como la curva predicha se ajusta mejor a la curva real en comparación con las obtenidas con los otros modelos.

La Tabla muestra que en el año 2019, todos los modelos presentan un desempeño elevado (R^2 >0.77). Sin ambargo, el modelo SVR logra el mejor ajuste global (R^2 = 0.857) y el menor MAE (5205.44), lo que sugiere una adecuada capacidad para capturar la demanda eléctrica. Los errores porcentuales ($MAPE < 3\%$) confirman un nivel de precisión alto en las predicciones.

Sin embargo, aunque el modelo SVR mostró un comportamiento inconsistente, ya que a pesar de que en 2019 alcanzó un R^2 competitivo (0.857), su rendimiento se degradó significativamente en 2020 y 2022, con R^2 negativos, estos resultados sugieren un sobreajuste o mala generalización ante variaciones temporales de los datos.

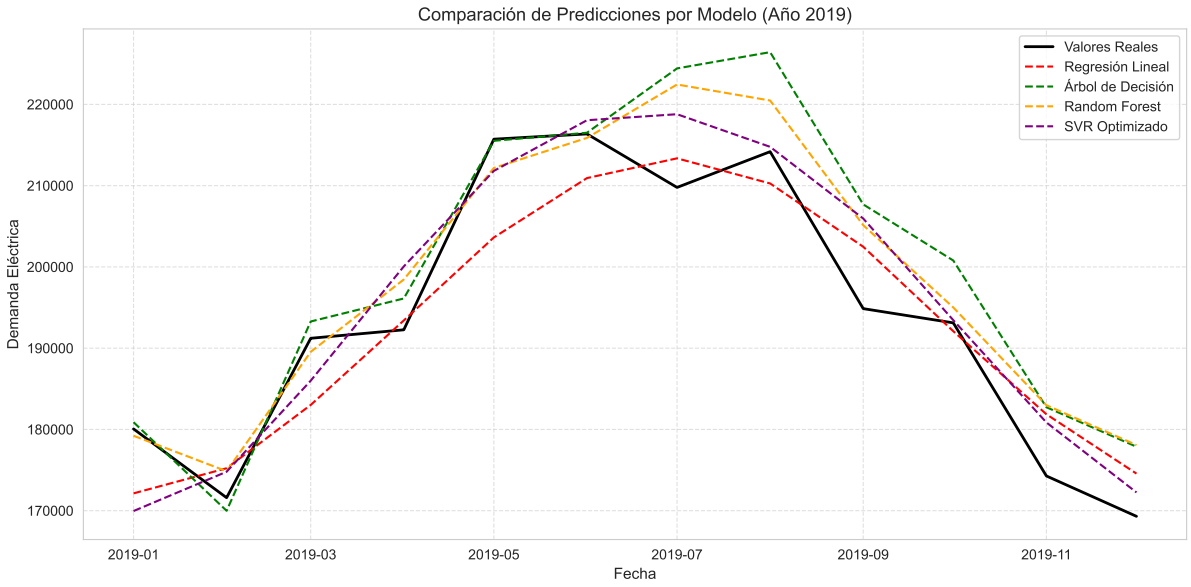

Figura 4.3: Predicción de modelos de aprendizaje máquina

En 2020 se observa un incremento generalizado de los errores, esto se debe posiblemente a fluctuaciones atípicas en la demanda energética (por ejemplo, efectos derivados de la pandemia de COVID-19). El Random Forest y el Árbol de Decisión muestran un aceptable desempeño con valores de cercanos a $R^2 = 0{,}60$, mientras que el SVR cae drásticamente ($R^2 = -0{,}51$), evidenciando falta de capacidad para generalizar bajo condiciones no

estacionarias.

En el año 2021, el desempeño mejora considerablemente para los modelos basados en árboles, destacando el Árbol de Decisión ($R^2 = 0{,}96$) y Random Forest ($R^2 = 0{,}91$) como los de mayor precisión y menor error (MAE entre 2600 y 4500). Por otro lado, el modelo SVR vuelve a mostrar una inestabilidad significativa ($R^2 = -1{,}20$), con un MAPE superior al 11 %.

Finalmente el último año analizado 2022, el modelo Random Forest mantiene un rendimiento sólido ($R^2 == 0{,}91, MAE5200$), seguido del Árbol de Decisión ($R^2 = 0{,}89$). El modelos de Regresión Lineal muestra una disminución notable de su capacidad predictiva ($R^2 = 0{,}55$), lo que indica limitaciones para modelar relaciones no lineales. El modelo SVR presenta nuevamente resultados deficientes ($R^2 = -6{,}43$), esto muestra que el modelo no logró adaptarse a las características de los datos recientes.

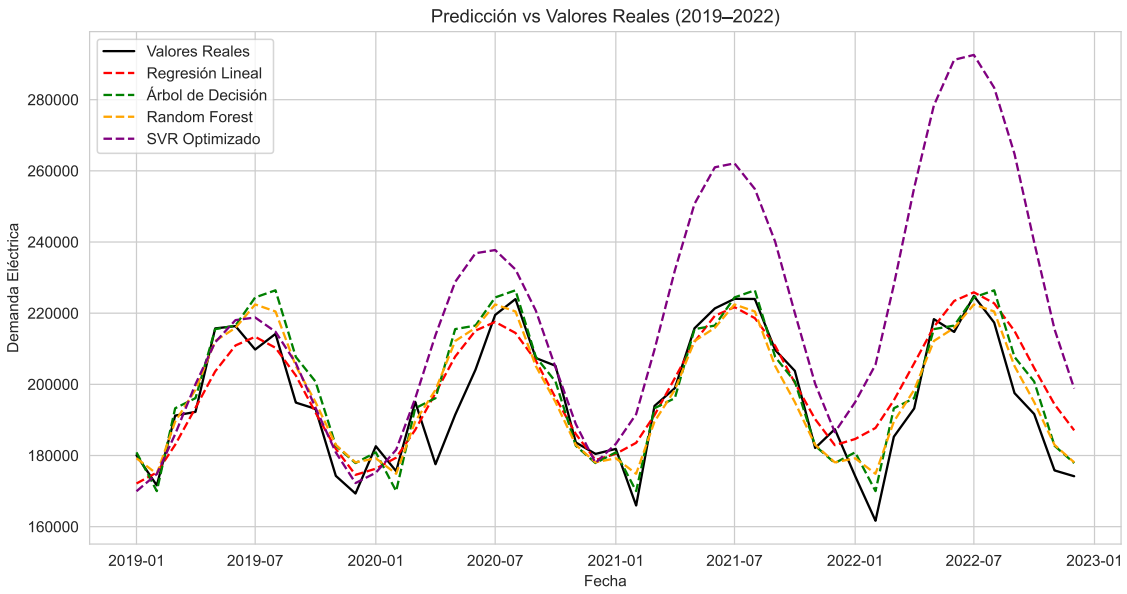

Figura 4.4: Predicción de modelos de aprendizaje máquina

En general, el modelo Random Forest se muestra en este estudio como el modelo más equilibrado y robusto frente a la variabilidad interanual, obteniendo una alta capacidad de generalización.

El modelo de Árbol de Decisión obtiene excelentes resultados en años específicos, aunque su rendimiento varía según la complejidad del patrón temporal.

El modelo de Regresión Lineal mantiene un desempeño aceptable, pero su naturaleza lineal limita su precisión frente a comportamientos no estacionarios.

El modelo SVR, aunque en un inicio (2019) tuvo los mejores resultados, resulta inestable y sensible a la distribución de los datos, requiriendo un ajuste de hiperparámetros más exhaustivo o una revisión de la estrategia de escalado de variables.

4.3. Comparación entre Modelos de ML y Series Temporales

Los modelos de Machine Learning (ML) mostraron un desempeño notablemente superior, con valores de R^2 positivos y altos (en algunos casos superiores a 0.90, como el Random Forest en 2021), mientras que los modelos ARIMA, SARIMA y SARIMAX obtuvieron R^2 negativos en todos los años, esto muestra que los modelos estadísticos tienen una baja capacidad de ajuste a los datos reales.

Lo anterior muestra que los modelos ML capturan mejor las relaciones no lineales y las interacciones complejas entre las variables, en contraste con los modelos clásicos de series temporales, que asumen una estructura lineal.

En términos de error absoluto (MAE) y raíz del error cuadrático medio (RMSE), los modelos ML presentaron valores mucho menores (alrededor de 5,000–10,000 en promedio) frente a los modelos ARIMA y sus variantes, cuyos errores superaron los 18,000 e incluso alcanzaron más de 60,000 en años como 2022.

Esto sugiere que los modelos ML ofrecen predicciones más consistentes y menos sensibles a variaciones o cambios abruptos en la serie temporal.

Por otro lado, los modelos ML mantuvieron un desempeño estable entre los distintos años, mostrando una menor degradación del rendimiento conforme avanzaba el periodo analizado (2019–2022).

Por el contrario, los modelos ARIMA/SARIMA/SARIMAX experimentaron un dete-

rioro significativo en su capacidad predictiva en los últimos años, lo cual podría atribuirse a cambios estructurales o no estacionariedad que estos modelos no lograron modelar adecuadamente.

Los modelos de Machine Learning, especialmente Random Forest y SVR, demostraron mejor capacidad de generalización, gracias a su naturaleza no paramétrica y su habilidad para capturar patrones no lineales y dependencias complejas.

Aunque los modelos ML son más precisos, tienden a ser menos interpretables. Los modelos ARIMA y sus variantes, en cambio, ofrecen mayor interpretabilidad estadística, siendo útiles para análisis exploratorios o pronósticos a corto plazo en contextos con alta estacionalidad regular. En este caso específico, los modelos de Machine Learning superan claramente a los modelos ARIMA, SARIMA y SARIMAX en todas las métricas de desempeño.

Esto evidencia que el comportamiento de la serie energética analizada no puede describirse de forma lineal o puramente estacional, sino que presenta patrones no lineales y dependencias complejas, las cuales son mejor capturadas por algoritmos de aprendizaje automático. Por tanto, para aplicaciones de predicción energética o monitoreo de rendimiento solar, los modelos ML —en particular Random Forest y SVR optimizado— constituyen la opción más adecuada por su precisión, estabilidad y capacidad adaptativa.

4.4. Correlogramas ACF y PACF

El correlograma es una herramienta fundamental en el análisis de series temporales, utilizada para examinar la dependencia lineal entre observaciones separadas en el tiempo. Representa gráficamente los coeficientes de autocorrelación o autocorrelación parcial en función del número de retardos (lags), permitiendo identificar patrones de persistencia temporal, tendencias o estacionalidad en los datos.

El correlograma de la función de autocorrelación (ACF) representa el grado de relación lineal existente entre los valores de la serie temporal y sus retardos (lags). En este caso, se analiza cómo la demanda eléctrica en México depende de sus valores pasados.

En el eje horizontal se muestran los retardos (lags) —meses anteriores—, mientras que

en el eje vertical aparecen los coeficientes de autocorrelación (entre 1 y 1). Cada barra indica la fuerza y dirección de la correlación de la serie con su versión desplazada en el tiempo. Las líneas azules horizontales representan los intervalos de confianza del 95 %; las barras que exceden estos límites son estadísticamente significativas.

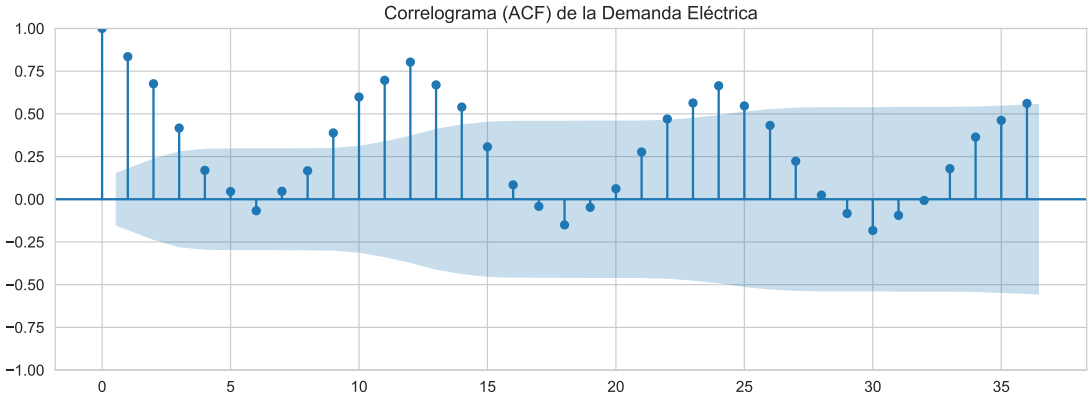

Figura 4.5: Predicción de modelos de aprendizaje máquina

Las primeras correlaciones (lags cercanos a 1–3) muestran valores altamente positivos, lo que indica que la demanda eléctrica presenta una fuerte dependencia temporal inmediata. En otras palabras, el consumo en un mes está fuertemente relacionado con el del mes anterior.

Conforme aumenta el retardo, la autocorrelación decrece gradualmente, pero mantiene un patrón suavemente oscilante, lo que sugiere la presencia de estacionalidad (probablemente anual, es decir, cada 12 meses). Este comportamiento es característico de series que siguen patrones cíclicos o estacionales, como la demanda eléctrica, que tiende a elevarse en meses calurosos o con mayor actividad económica.

La persistencia de autocorrelaciones positivas en varios lags indica que la serie no es puramente aleatoria (no es un ruido blanco), y por tanto contiene información predecible. Este tipo de estructura justifica el uso de modelos ARIMA, SARIMA o SARIMAX, ya que estos aprovechan precisamente las dependencias temporales y estacionales observadas.

El correlograma ACF de la demanda eléctrica evidencia que la serie presenta alta autocorrelación en los primeros retardos, confirmando su naturaleza temporal depen-

diente. Exhibe un patrón estacional significativo, lo que implica fluctuaciones cíclicas recurrentes en la demanda. No hay evidencia de comportamiento aleatorio puro, lo que valida el enfoque de modelado con métodos de series temporales paramétricos (ARIMA/SARIMA/SARIMAX).

En resumen, el análisis del correlograma indica que la demanda eléctrica posee persistencia, tendencia y estacionalidad, por lo que el pronóstico debe considerar dichos componentes para lograr una estimación robusta.

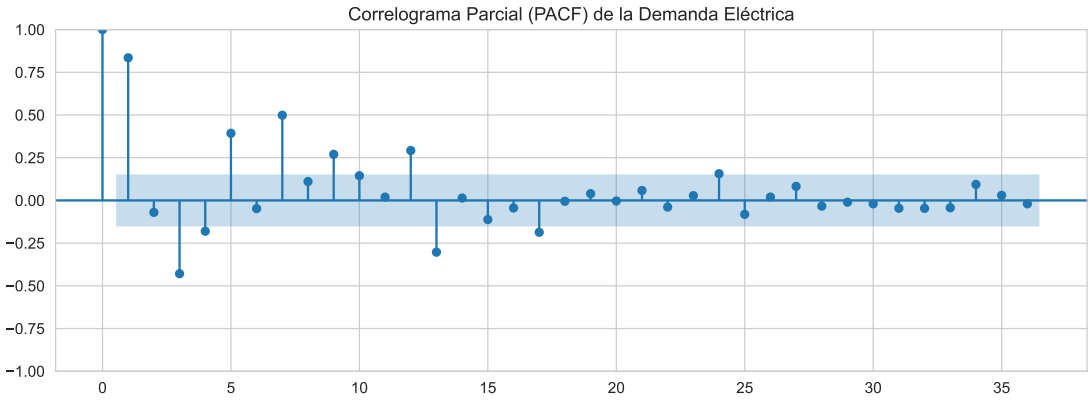

Figura 4.6: Predicción de modelos de aprendizaje máquina

El PACF (Partial Autocorrelation Function) muestra la correlación entre los valores actuales de la serie y sus retardos, eliminando el efecto de los retardos intermedios. Es una herramienta esencial para determinar el orden autorregresivo (p) en un modelo ARIMA o SARIMA.

En el eje horizontal se representan los retardos (lags) y en el vertical los coeficientes de autocorrelación parcial. Las líneas azules marcan los intervalos de confianza del 95 %; las barras que los superan son estadísticamente significativas. En los primeros retardos (lag 1 y posiblemente lag 2) se observan valores positivos significativos, lo cual sugiere que la demanda eléctrica tiene una dependencia directa e inmediata con sus valores previos.

Después del lag 2 o 3, los coeficientes tienden a disminuir rápidamente y se mantienen dentro del rango de confianza, lo que indica que las correlaciones de orden superior son débiles o inexistentes una vez controlado el efecto de los primeros retardos. A diferencia

del ACF, el PACF no muestra un patrón claramente oscilante, pero sí una disminución abrupta tras los primeros retardos, típica de un proceso autorregresivo (AR).

La presencia de un pico significativo en el lag 1 y la caída posterior de las correlaciones parciales sugiere que la serie puede describirse adecuadamente con un componente AR(1), es decir, un modelo autorregresivo de primer orden. Sin embargo, al combinar esta información con la del ACF —donde se observó un patrón oscilante y persistente—, se infiere que la serie podría ajustarse mejor mediante un modelo estacional SARIMA(p, d, q)(P, D, Q, 12), con:

1. p 1 (por la significancia en el PACF)
2. q 1 o 2 (por la atenuación gradual en el ACF)
3. P y Q 1 (debido a la periodicidad anual evidente)
4. s = 12, correspondiente a la estacionalidad mensual.

El PACF confirma que la serie no es un ruido blanco, ya que presenta correlaciones parciales significativas en los primeros retardos. El comportamiento indica una dependencia lineal estructurada, en la que los valores pasados influyen directamente sobre los futuros.

La rápida pérdida de significancia en retardos mayores implica que las dependencias más fuertes son de corto plazo, mientras que la estacionalidad detectada proviene de ciclos de mayor longitud (reflejados más claramente en el ACF).

El análisis del PACF revela que la demanda eléctrica presenta dependencias directas de corto alcance, concentradas principalmente en los primeros meses. El patrón sugiere un componente autorregresivo dominante (AR(1)) dentro del modelo global.

En conjunto con el ACF, se corrobora que la serie posee estructura temporal y estacionalidad, justificando el uso de modelos SARIMA o SARIMAX, que integren tanto efectos autorregresivos como estacionales y exógenos.

Capítulo 5

Conclusiones

5.1. Conclusión general

El análisis comparativo de los modelos clásicos de series temporales y los modelos de aprendizaje automático permite establecer diferencias claras en su capacidad para predecir la demanda eléctrica en México durante el periodo 2019–2022. Los resultados revelan que los modelos estadísticos (ARIMA, SARIMA y SARIMAX) presentan limitaciones importantes para capturar la dinámica real de la serie, lo cual se refleja en valores de R^2 negativos en todos los años evaluados y en errores de magnitud considerable. Aunque SARIMAX mantiene un desempeño ligeramente superior dentro de este grupo —especialmente en 2019— su precisión disminuye notablemente a partir de 2020, evidenciando sensibilidad ante cambios estructurales y variaciones no estacionarias en los datos.

En contraste, los modelos de aprendizaje automático muestran un comportamiento sustancialmente más robusto y preciso. Entre ellos, Random Forest destaca como el modelo con mayor estabilidad interanual, obteniendo consistentemente valores altos de R^2 y errores significativamente menores que los modelos estadísticos. El Árbol de Decisión también presenta resultados sobresalientes en ciertos años, mientras que la Regresión Lineal, aunque limitada por su naturaleza lineal, mantiene un desempeño aceptable. El modelo SVR evidencia un rendimiento muy variable: logra los mejores resultados en 2019, pero presenta degradaciones severas en años posteriores, lo que indica una sensibilidad marca-

da a la variabilidad temporal y una necesidad de ajustes más cuidadosos en el proceso de modelado.

En conjunto, estos resultados confirman que los modelos de aprendizaje máquina son más adecuados para este tipo de problema, ya que capturan de manera más efectiva relaciones no lineales, variabilidad temporal y patrones complejos propios de la demanda energética. Además, su capacidad de generalización resulta superior ante escenarios con cambios abruptos o fluctuaciones significativas, como los observados a partir de 2020.

Finalmente, se observa que ningún modelo mantiene un desempeño óptimo en todos los años, lo cual subraya la importancia de evaluar los modelos de manera anual y de considerar enfoques híbridos o modelos avanzados que integren información contextual adicional (clima, actividad económica, estacionalidad no lineal) para mejorar la precisión futura de las predicciones.

5.2. Discusiones

Los resultados obtenidos evidencian diferencias sustanciales entre los modelos estadísticos tradicionales y los modelos de aprendizaje automático para la predicción de la demanda eléctrica mensual en México. Los modelos ARIMA, SARIMA y SARIMAX, aunque históricamente utilizados en series temporales, mostraron dificultades para ajustarse a los patrones recientes de la serie, especialmente a partir de 2020. Este comportamiento puede explicarse por la sensibilidad inherente de estos modelos a cambios estructurales, variaciones abruptas y fenómenos no estacionarios, elementos que caracterizan el comportamiento energético en años recientes. La presencia de R^2 negativos confirma que estos enfoques no fueron capaces de capturar la tendencia real ni las fluctuaciones propias del consumo eléctrico moderno.

Por otro lado, los modelos de aprendizaje automático demostraron una capacidad significativamente mayor para modelar relaciones no lineales y dinámicas complejas. Random Forest emergió como el modelo más consistente, probablemente debido a su capacidad para mitigar el sobreajuste y capturar interacciones entre características sin requerir supuestos estrictos sobre la estructura estadística de los datos. El Árbol de Decisión, aunque más

simple, mostró un desempeño notable en varios años, evidenciando que incluso modelos no paramétricos de baja complejidad pueden superar a los enfoques clásicos en contextos con alta variabilidad.

El desempeño irregular del SVR, que logra resultados sobresalientes en 2019 pero se degrada en años posteriores, sugiere una fuerte dependencia de la selección de hiperparámetros, la escala de los datos y la estructura temporal del conjunto de entrenamiento. Este modelo podría beneficiarse de estrategias de ajuste más sofisticadas, como la optimización bayesiana o el uso de kernels adaptativos a series temporales.

En conjunto, la evidencia respalda el uso de métodos de aprendizaje automático como herramienta principal para este tipo de predicciones, especialmente en contextos donde la serie presenta comportamiento no lineal, efectos externos no modelados y cambios de tendencia a mediano plazo.

5.3. Limitaciones del Estudio

Es importante reconocer las limitaciones inherentes al estudio:

1. La predicción se basó únicamente en la variable demanda eléctrica mensual. La ausencia de variables explicativas como temperatura, actividad económica, consumo industrial, o factores estacionales específicos, limita la capacidad de los modelos para capturar relaciones exógenas.

2. Solo se analizaron datos de 2010 a 2022, lo que reduce la capacidad de los modelos para aprender patrones de largo plazo o ciclos energéticos más amplios.

3. El uso de datos mensuales impide capturar variaciones semanales o diarias que podrían enriquecer la modelación y mejorar la precisión predictiva.

4. El estudio se centró en comparar modelos por año calendario, sin realizar análisis complementarios como estabilidad temporal, sensibilidad a hiperparámetros o validación cruzada especializada para series temporales.

Bibliografía

[1] BBVA. "Machine learning: ¿qué es y cómo funciona el maestro en reconocer patrones?" En: (15 julio 2024). URL: https://www.bbva.com/es/innovacion/machine-learning-que-es-y-como-funciona/.

[2] Leo Breiman et al. *Classification and Regression Trees*. Wadsworth International Group, 1984.

[3] Jair Cervantes, Xiaoou Li y Wen Yu. "Support Vector Machine Classification Based on Fuzzy Clustering for Large Data Sets". En: *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2006, págs. 572-582. DOI: 10.1007/11925231_54.

[4] Jair Cervantes, Xiaoou Li y Wen Yu. "SVM Classification for Large Data Sets by Considering Models of Classes Distribution". En: *2007 Sixth Mexican International Conference on Artificial Intelligence, Special Session (MICAI)*. IEEE, 2007. DOI: 10.1109/micai.2007.27.

[5] Jair Cervantes et al. "A comprehensive survey on support vector machine classification: Applications, challenges and trends". En: *Neurocomputing* 408 (sep. de 2020), págs. 189-215. ISSN: 0925-2312. DOI: 10.1016/j.neucom.2019.10.118.

[6] Jair Cervantes et al. "Support vector machine classification for large data sets via minimum enclosing ball clustering". En: *Neurocomputing* 71.4-6 (2008), págs. 611-619. DOI: 10.1016/j.neucom.2007.07.028.

[7] T. Chai y R. R. Draxler. "Root mean square error (RMSE) or mean absolute error (MAE)?—Arguments against avoiding RMSE in the literature". En: *Geoscientific Model Development* 7.3 (2014), págs. 1247-1250.

[8] Harris Drucker et al. "Support Vector Regression Machines". En: *Advances in Neural Information Processing Systems* 9 (1997), págs. 155-161.

[9] Trevor Hastie, Robert Tibshirani y Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd. Springer, 2009.

[10] Rob J. Hyndman y Anne B. Koehler. "Another look at measures of forecast accuracy". En: *International Journal of Forecasting* 22.4 (2006), págs. 679-688.

[11] Xiaoou Li, Jair Cervantes y Wen Yu. "A Novel SVM Classification Method for Large Data Sets". En: *2010 IEEE International Conference on Granular Computing*. IEEE, 2010. DOI: 10.1109/grc.2010.46.

[12] Xiaoou Li, Jair Cervantes y Wen Yu. "Fast classification for large data sets via random selection clustering and Support Vector Machines". En: *Intelligent Data Analysis* 16.6 (2012), págs. 897-914. DOI: 10.3233/ida-2012-00558.

[13] Xiaoou Li, Jair Cervantes y Wen Yu. "Two-stage svm classification for large data sets via randomly reducing and recovering training data". En: *2007 IEEE International Conference on Systems, Man and Cybernetics*. IEEE, 2007. DOI: 10.1109/icsmc.2007.4413814.

[14] "Mercados Eléctricos Marginalistas, las Grandes Rentabilidades de los Generadores Privados y su Efecto en las Tarifas Eléctricas". En: *CFE* (2021). URL: https://app.cfe.mx/Aplicaciones/OTROS/Boletines/boletin?i=2408.

[15] Douglas C. Montgomery, Elizabeth A. Peck y G. Geoffrey Vining. *Introduction to Linear Regression Analysis*. 5th. Wiley, 2012.

[16] J.R. Quinlan. "Simplifying decision trees". En: *International Journal of Man-Machine Studies* 27.3 (sep. de 1987), págs. 221-234. ISSN: 0020-7373. DOI: 10.1016/s0020-7373(87)80053-6.

[17] Denniye Hinestroza Ramírez. "El Machine Learning a Traves de los Tiempos y los Aportes a la Humanidad". En: *Universidad Libre Seccional Pereira* (2018).

[18] María Camino; McWilliams José Manuel Remón Ugarte Adrián; González Fernández. "Predicción de precios de energía eléctrica utilizando árboles dinámicos". En: *E.T.S.I. Industriales (UPM)* (2017).

[19] Laura Melgar García ; José Francisco Torres Maldonado; Alicia Troncoso ; José Cristóbal Riquelme Santos. "Técnicas de Big Data Para la Predicción de la Demanda y Precio Eléctrico". En: *Economía industrial* 431 (2024), págs. 119-130. ISSN: 0422-2784.

[20] Ch. Ravi Sekhar, Minal y E. Madhu. "Mode Choice Analysis Using Random Forrest Decision Trees". En: *Transportation Research Procedia* 17 (2016), págs. 644-652. ISSN: 2352-1465. DOI: 10.1016/j.trpro.2016.11.119.

[21] Alex J. Smola y Bernhard Schölkopf. "A tutorial on support vector regression". En: *Statistics and Computing* 14.3 (2004), págs. 199-222.

[22] SPACEWELL. "Predicción del Consumo de Energía mediante el Machine Learning y la IA". En: *SPACEWELL* (2022). URL: https://spacewell.com/es/recursos/blog/prediccion-del-consumo-de-energia-mediante-el-machine-learning-y-la-ia/.

[23] Werner Stolz. "Predicción del clima y modelos numéricos". En: *Revista de Ciencias Ambientales* 35.1 (jun. de 2008), pág. 34. ISSN: 1409-2158. DOI: 10.15359/rca.35-1.7.

[24] Cort J. Willmott y Kenji Matsuura. "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance". En: *Climate Research* 30 (2005), págs. 79-82.