UAEM Universidad Autónoma del Estado de México

UNIVERSIDAD AUTÓNOMA DEL ESTADO DE

MÉXICO

TESIS

Predicción de Radiación Solar Mediante Técnicas de Inteligencia Artificial

---

Tesis que presenta

Omar Arturo Cruz Rios

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

Abstract

In recent years, solar energy has emerged as a very important alternative in the transition toward sustainable energy systems. However, its efficiency depends on the ability to accurately predict solar radiation levels, especially when using solar batteries as a storage medium.

This thesis implements and compares different machine learning models to predict solar radiation based on meteorological variables such as wind direction, wind speed, humidity, and temperature.

In the experimental results, different supervised algorithms were implemented, including linear regression, decision trees, random forests, and neural networks, and their performance was evaluated using statistical metrics. The results obtained allow the identification of the most appropriate model for solar radiation prediction. This offers a valuable tool for planning and managing the use of solar systems under variable climatic conditions.

Resumen

En los últimos años, la energía solar se ha posicionado como una alternativa muy importante en la transición hacia sistemas energéticos sostenibles. Sin embargo, su eficiencia depende de la capacidad para anticipar de manera precisa los niveles de radiación solar, especialmente cuando se utilizan baterías solares como medio de almacenamiento.

En esta Tesis se implementan y comparan diferentes modelos de aprendizaje automático para predecir la radiación solar a partir de variables meteorológicas como dirección del viento, velocidad del viento, humedad y temperatura.

En los resultados experimentales, se implementaron distintos algoritmos supervisados, incluyendo regresión lineal, árboles de decisión, bosques aleatorios y redes neuronales, evaluando su desempeño mediante métricas estadísticas. Los resultados obtenidos permiten identificar el modelo más adecuado para predicción de radiación solar. Esto ofrece una herramienta valiosa para la planificación y gestión del uso de sistemas solares en condiciones climáticas variables.

Índice general

Índice de figuras

Lista de Tablas

Capítulo 1

Introducción

1.1. Introducción

"La energía proveniente del sol, se denomina energía Solar, aunque se conoce como radiación solar"(Pareja Aparicio, 2010). Para poder comprender la interacción de la radiación con la Inteligencia artificial deben quedar en claro los factores que queremos relacionar, como lo es el tiempo, el contexto, y las herramientas que se utilizan para la predicción en tiempo real. Cuando hablamos de la predicción de radiación solar nos hace énfasis sobre todo en la parte del clima y su aprovechamiento en diferentes contextos, ya que el objetivo principal es saber que cantidad de energía llegara a la superficie terrestre en un tiempo determinado y en que momentos para su máximo aprovechamiento en recolección de energía a través de fotoceldas. En la actualidad ya necesitamos de otros métodos mas allá de lo tradicional para saber con una mejor exactitud y mayor velocidad las predicciones climatológicas y eso lo hacemos haciendo uso de técnicas de inteligencia artificial para tener una exactitud de datos y recopilarlos a una velocidad considerable y adaptable.

Lo que nos dará resultados optimizados teniendo la confiabilidad de que ya no deberán tener perdidas como algunos años atrás por la duda de no saber como actuar ante las situaciones climatológicas impredecibles. La predicción de la radiación solar es un área de estudio fundamental, particularmente cuando se cuenta con sistemas de almacenamiento

energético como baterías solares, cuya eficiencia y vida útil dependen del conocimiento anticipado de la disponibilidad de energía. La toma de decisiones inteligentes respecto al uso o almacenamiento de energía requiere modelos predictivos robustos y adaptables.

Esta Tesis tiene como objetivo desarrollar e implementar modelos de aprendizaje automático para predecir los niveles de radiación solar a corto plazo, utilizando para ello factores atmosféricos como la dirección del viento, la velocidad del viento, la humedad y la temperatura. El conjunto de datos utilizado contiene registros de los últimos cuatro meses, y la variable objetivo a predecir es la radiación solar.

En esta Tesis, se emplearán y compararán diversas técnicas de aprendizaje máquina, incluyendo modelos supervisados como regresión lineal, árboles de decisión, bosques aleatorios y redes neuronales artificiales. A través de la evaluación de su desempeño, se buscará determinar cuál técnica proporciona los mejores resultados para este tipo de datos y condiciones, con el fin de aportar una herramienta útil para usuarios e investigadores interesados en optimizar el uso de sistemas solares.

Este estudio no solo contribuirá a una mejor planificación del uso de energía solar, sino que también fortalecerá el enfoque hacia una transición energética más inteligente y sostenible.

1.2. Planteamiento del Problema

La generación de energía solar depende directamente de la radiación solar disponible, la cual está sujeta a variaciones constantes que se deben a factores meteorológicos como la nubosidad, ya que las nubes afectan la predicción al no saber cuanto tiempo y en que momento pueda hacer una interrupción del sol con la superficie ya que reduce la intensidad del sol de manera drástica y haciéndolo impredecible de esa misma forma otros factores que debemos tener en cuenta son la temperatura humedad y viento ya que afectan en el tiempo y movimiento de las mismas. Esta variabilidad representa un reto importante para la integración eficiente de sistemas solares en redes energéticas, especialmente cuando se cuenta con baterías de almacenamiento, cuyo uso debe planificarse con base en la disponibilidad energética.

En muchas regiones, incluidas aquellas con alta irradiación solar potencial, la falta de modelos predictivos precisos limita la toma de decisiones informadas sobre cuándo utilizar o almacenar energía, lo que puede derivar en pérdidas económicas y menor eficiencia en el aprovechamiento del recurso solar. Aunque existen modelos físicos y estadísticos tradicionales para este tipo de predicción, su precisión suele verse afectada por la complejidad y no linealidad de los datos climáticos.

Por ello, surge la necesidad de investigar y aplicar métodos de predicción más robustos, como los basados en técnicas de aprendizaje automático, que puedan modelar relaciones complejas entre múltiples variables atmosféricas y la radiación solar. La pregunta central que guía esta investigación es: ¿Qué modelo de aprendizaje automático ofrece el mejor desempeño para predecir la radiación solar a corto plazo, utilizando variables meteorológicas como entrada?

1.3. Justificación

En la actualidad una de las energías renovables a aprovechar es la solar, ya que la radiación solar es una de las energías que se puede aprovechar al máximo si se emplean las técnicas adecuadas para su almacenamiento y por supuesto que las técnicas adecuadas para su predicción como lo son las de la inteligencia artificial con el avance tecnológico que hemos estado teniendo actualmente podemos aprovechar dichas herramientas al máximo para optimizar la predicción de la radiación.

Tradicionalmente se han venido usando las técnicas de modelos tanto físicos como estadísticos que con el pasar de los años han demostrado una limitación para la precisión de los datos a recolectar debido a la complejidad de los factores atmosféricos que hacen que los resultados se vean afectados.

En esta Tesis se busca implementar diferentes técnicas de IA para una mejor predicción y esas limitaciones que se presentan en modelos tradicionales ya no sea un factor que afecte los resultados de las predicciones, convirtiéndolo en un modelo preciso y con resultados eficientes. Esto beneficiaria a empresas que recolectar energía solar, pero sobre todo estos resultados son de gran impacto para el planeta puesto que asi evitamos el uso

de energías que son altamente contaminantes para la atmosfera. Estos modelos tienen un gran impacto económico hasta nivel social siendo una de estas e mejorar la economía de muchas personas. Al aplicar estos modelos a la predicción de radiación solar, es posible mejorar la toma de decisiones relacionadas con la operación de sistemas solares y el uso de baterías, y esto se traduce en una mayor eficiencia energética y ahorro económico.

En esta Tesis se evaluan comparativamente distintas técnicas de aprendizaje máquina y se propone el modelo más adecuado para tareas de predicción solar. Asimismo, sienta las bases para el desarrollo de sistemas inteligentes de gestión energética, que pueden ser implementados en viviendas, industrias o comunidades que buscan adoptar soluciones energéticas sostenibles.

1.4. Objetivo general

Implementar y evaluar modelos de aprendizaje automático para la predicción de la radiación solar a corto plazo, utilizando variables meteorológicas como dirección del viento, velocidad del viento, humedad y temperatura, con el fin de optimizar el aprovechamiento de sistemas de energía solar con almacenamiento.

1.4.1. Objetivos específicos

1. Implementar diversos algoritmos de aprendizaje máquina supervisado, tales como regresión lineal, árboles de decisión, bosques aleatorios y redes neuronales, empleando como variables predictoras los datos climáticos disponibles.

2. Evaluar y comparar el desempeño de los modelos desarrollados mediante métricas de error y ajuste (como MAE, MSE y R^2), con el objetivo de identificar la técnica más precisa para la predicción de radiación solar.

3. Analizar la viabilidad de uso de los modelos predictivos en entornos reales con sistemas de baterías solares, considerando escenarios de planificación energética.

4. Proponer recomendaciones de selección de modelos de predicción solar, enfocadas

en aplicaciones prácticas en contextos domésticos o industriales.

1.5. Hipótesis

Los modelos de aprendizaje automático son capaces de predecir la radiación solar con un alto grado de precisión utilizando variables meteorológicas como dirección del viento, velocidad del viento, humedad y temperatura, superando en desempeño a los métodos estadísticos tradicionales.

1.6. Estado del arte

La predicción de la Radiación solar hoy en día es un tema importante por abarcar puesto que hablamos de una de las energías renovables más importantes que llega a la superficie terrestre y una de las cuales podemos obtener a diario con la tecnología que se ha desarrollado. Pero a pesar de los avances obtenidos con la tecnología tradicional que se hay usado tenemos algunas limitantes aun como la recolección de datos y "La predicción precisa de la radiación solar es de gran importancia para la utilización de la energía solar y su integración en la red, pero debido a la intrínseca no estacionaria y no linealidad de la radiación solar diaria, que está influenciada por muchos elementos, los modelos de predicción individuales pueden tener dificultades para obtener resultados con alta precisión." (Zhang, 2019). Estas limitantes son una de las cuales la precisión de la predicción de la radiación solar no es una de las mejores usando las técnicas físicas tradicionales por este motivo hoy en día tenemos una de las tecnologías más importantes del mundo que ha sido de un gran impacto hablamos de la Inteligencia Artificial haciendo varias de las técnicas de inteligencia Artificial podremos tener una de las predicciones más precisas. Con la gran ayuda que no puede brindar la Inteligencia Artificial hoy en día con este tema de la predicción en tiempo real y haciendo uso de sus técnicas "Métodos de IA basados en el aprendizaje automático –y especialmente basados en redes neuronales profundas– nos permiten modelar el clima y el tiempo, identificar patrones y hacer predicciones precisas de los cambios en la temperatura global a partir del análisis grandes

cantidades de datos meteorológicos y climáticos multidimensionales. Además de utilizarse para construir predicciones y modelos climáticos más precisos, los métodos de IA también se pueden aplicar para mejorar los sistemas de modelado meteorológico de última generación al permitir, por ejemplo, la detección y separación del ruido en las observaciones climáticas o el etiquetado automático de los datos climáticos." (Oliver, 2022) El uso de la herramienta de la "La IA también ha demostrado compatibilidad en el sector de las energías renovables y se ha utilizado ampliamente para predecir los recursos energéticos dependientes del clima y la producción de energía de centrales eléctricas intermitentes como la eólica y la solar." (Faisal Nawab, 2023) Para poder comprender mejor las redes neuronales que son utilizadas en esta investigación para comprender mejor el propósito puesto que "Las redes neuronales y el aprendizaje automático son bastante utilizados en aplicaciones de radiación solar, evaluando el desempeño del perceptrón multicapa (MLP) y los árboles de decisión potenciados mediante la combinación con regresión lineal para la estimación de la energía solar" (Luis Eduardo Ordoñez Palacios, 2020) gracias a esta que es una de las técnicas de IA es una de las maneras de obtener una mejor precisión de la radiación solar, pudiendo aprovecharlo en los campos ya vistos como lo son en la agricultura y la recolección de esta gran energía renovable, limpia he inagotable ya que ha existido con el ser humano por miles de años. "Sin embargo, sólo la utilización de modelos atmosféricos físicamente establecidos puede rellenar las amplias áreas terrestres donde no existen instrumentos, sobre todo en las vastas áreas sobre los océanos. La radiación solar ultravioleta en un determinado lugar geográfico depende de la distancia Tierra-Sol, la que varía a lo largo del año, y el ángulo cenital. Además del contenido vertical del ozono y por los aerosoles presentes en la atmósfera." (Wright, 2008) Todos estos son los factores físicos que se enfrentan los métodos tradicionales que con el paso del tiempo ha sido un estorbo para la precisión de la predicción de grandes datos para el mismo modelo físico el cual se busca quitara para una mejor predicción y no solo eso, si no también se busca que las técnicas de Inteligencia artificial sean adaptables dependiendo el entorno que se encuentre ya que mencionado anteriormente cada posición geográfica es diferente y el ángulo del sol actúa de una forma distinta.

En los últimos años, el uso de modelos de aprendizaje automático para la predicción

de variables meteorológicas y energéticas ha crecido significativamente. A continuación se presentan algunos enfoques destacados:

R. Mohandes et al. (2019) utilizaron redes neuronales artificiales y máquinas de soporte vectorial para la predicción de radiación solar en regiones áridas, obteniendo mejores resultados que modelos de regresión lineal tradicional.

F. Pedro et al. (2020) compararon el desempeño de modelos como Random Forest, XGBoost y KNN para la predicción solar en estaciones ubicadas en Brasil, concluyendo que los modelos ensemble ofrecían mayor precisión.

L. Jiang y Z. Zhang (2021) propusieron un modelo híbrido basado en redes LSTM y preprocesamiento por wavelets para la predicción horaria de radiación solar, destacando su capacidad para capturar patrones temporales complejos.

Investigaciones recientes han demostrado que la inclusión de variables meteorológicas locales mejora sustancialmente el rendimiento de los modelos ML, especialmente en regiones con alta variabilidad climática.

A pesar de estos avances, aún existen oportunidades para validar y adaptar estos modelos en contextos específicos, con datos reales y recientes, lo cual justifica la necesidad de investigaciones aplicadas como la que se presenta en este trabajo.

Capítulo 2

Preliminares

Esta Tesis se sustenta en conceptos clave relacionados con la energía solar, la predicción meteorológica y el aprendizaje automático. En este Capítulo, se describen los fundamentos teóricos más relevantes

2.1. Radiación Solar

Cuando hablamos de radiación solar debemos comprender de primer idea que es una forma de energía producida por el sol que llega a el planeta Tierra en un tiempo estimado de 8 minutos con 20 segundos en forma de ondas electromagnéticas, es la primordial fuente de energía de este, debido a que es quien regula la temperatura en la Tierra y así mismo quien ayuda en el proceso natural de la conservación de muchas especies de vida desde microorganismos que son parte de la cadena alimentaria de otras especies para subsistir hasta de células vegetales que son parte de la fotosíntesis para la reproducción de estas formas de vida para el crecimiento de vida natural en la Tierra.

2.1.1. Definición y tipos de radiación solar

Definamos la radiación solar como uno de los tipos de energía que nos proporciona el sol ya que es una de las energías limpias y renovables. También tiene otros tipos de radiaciones como son los siguientes.

1. Radiación solar directa. Esta radiación es aquella proveniente del sol que no tiene ninguna alteración en su forma o estructura, es la que llega hasta la tierra de manera más limpia y pura.

2. Radiación solar difusa (o dispersa). Esta radiación es aquella que llega hasta la tierra, que viene directamente desde el sol pero que esta radiación es interrumpida por elementos existentes en la atmosfera terrestre y otros elementos que provocan un cambio en su llegada sufriendo un cambio de dirección y esta no llegue en su forma más limpia, este tipo de radiación es capaz de aumentar dependiendo la nubosidad que exista en el entorno.

3. Radiación solar reflejada. Esta radiación es aquella la cual proviene del sol y llega hasta la tierra, pero con la diferencia de que esta no es absorbida por la superficie del planeta Tierra si no que esta es reflejada y revuelta hacia el espacio o entorno donde pueda manifestarse, esta misma podemos verla reflejada en gran cantidad como en el hielo por ejemplo las zonas de los polos de la tierra, en las nubes si son más claras las nubes se manifestara en gran manera, en el agua como son los ríos, lagos, océanos, etc. Por supuesto que también en algunos tipos de arenas sobre todo las arenas más claras como son en los desiertos, este tipo de energía es una pequeña fracción de la radiación directa ya que al ser reflejada deja de ser limpia en su forma, podemos encontrarla reflejada en más de otras superficies ya mencionadas y este dependerá de la superficie a este efecto se le conoce como efecto albedo.

2.1.2. Factores que afectan la radiación solar

Latitud y fecha

La latitud es un factor a nivel global que afecta el cómo recibe la tierra la radiación ya que debido a el movimiento de la traslación de la tierra y la posición geográfica en la que se pueda encontrar una persona en ese momento recibirá de manera distinta radiación, esto da pauta a las estaciones del año pues no es lo mismo recibir una radiación Solar el verano o primavera que en el Invierno de tal forma que si nos enfocamos a las posiciones

geográficas el ejemplo más claro son los polos o los desiertos la diferencia entre ambos es contraria ya que se encontrara mayor llegada de radiación en un desierto que una zona polar, si llega la radiación pero con distinta intensidad una en menor cantidad que es en el polo norte o sur debido a la misma inclinación de la tierra puesto que los rayos del sol tiene que cruzar la atmosfera con una mayor densidad provocando que se filtre y pierda calor siendo que esta tenga una menor intensidad esto debido al ángulo que se encuentra y la otra en una mayor cantidad como lo son los más cercanos al eje ecuador ya que es la que más se aproxima al sol y tiene una llegada de radiación más cercana y recta en un ángulo de 0° grados, pero esto no quiere decir que en esta zona no pueda tener menor radiación ya que con el pasar del día la posición del sol cambia haciendo que por la mañana o tarde reduzca y la radiación disminuya, esto también dependerá de las épocas del año en que se encuentre.

Altitud

Otro factor para tomar en cuenta es la altitud pues dependiendo a la altura que se encuentre una superficie esta recibirá en mayor o menor medida la Radiación, tomando como referencia el nivel del mar ya que existen en el mundo superficies terrestres por encima del nivel del mar, en zonas más elevadas sobre todo zonas montañosas que a nivel del mar existirá una menor dispersión de radiación y una mayor absorción de la misma y esta será más limpia, debido que tiene menos obstrucción con la atmosfera.

Nubosidad

La cantidad de nubes existentes en el cielo afecta a la superficie en su temperatura ya que al existir mayor cantidad de nubes le temperatura es menor pero al contrario si existen en menor cantidad la temperatura es mayor, los días nublados es el factor que interviene en la llegada de la radiación solar directa ya que es un factor natural al ser un fenómeno de evaporación y dispersión de partículas de agua y humedad, estas pueden depender debido a la densidad de las mismas nubes entre más densas sean menor cantidad de radiación y calor llegara y la luz será más tenue debido al mismo efecto del reflejo en partículas

de agua y hielo que las nubes puedan contener, esta misma energía se ve reflejada y enviada en todas direcciones del entorno aumentando la radiación difusa y disminuyendo la directa en grandes cantidades. Existen ocasiones en que el sol aún sigue atravesando esa nubosidad y las nubes son parciales en estos casos puede reflejar de la superficie a las nubes y de las nubes a la superficie haciendo que aumente el nivel de irradiancia en la superficie.

Albedo

Cuando existe la reflexión de la radiación solar en una superficie o cuerpos sobre todo en lo que es la nieve donde esté más se manifiesta se le conoce como albedo esta característica de reflexión se da mucho sobre todo es superficies claras y puede presentarse hasta un 90 % de la radiación solar en la nieve o zonas polares, este efecto se da sobre todo en colores más claros como lo son el blanco que este tiende a reflejar la radiación y tiende a tener una absorción de calor menor en comparación del color negro el cual este absorbe mucho calor pero el efecto de reflexión es mínimo, como se mencionó anteriormente en la nieve que es de color blanco en contacto con la radiación solar tiende a tener un albedo mucho grande por sus propiedades en color y en reflexión por las partículas de agua que contiene este, por otra parte en menor cantidad como lo es el agua ya que este tienden a reflejar un mínimo la radiación y a absorber un máximo la energía en forma de calor, este porcentaje de albedo puede variar dependiendo el movimiento ya que si el agua esta solida o quieta en un atardecer sobre el horizonte este puede presentar una gran reflexión o si las olas presentan espuma también .

Contaminación y Aerosoles

La radiación tiende a tener obstáculos para llegar a la superficie terrestre algunos son naturales como lo son las nubes, partículas de sal y agua entro otros, pero actualmente se tiene factores que has sido creados por la mano del hombre que son los contaminantes combustibles, los aerosoles, gases, entre muchos otros que han sido fabricados por el ser humano, estas partículas muchas veces son visibles a el ojo humano, como la combustión

de los líquidos combustibles de motores de todo tipo como automóviles, aviones, barcos, maquinaria pesada de las fábricas, etc. Todas estas al ser combustión en el medio ambiente llegan a acumularse en el espacio quedando en suspensión creando una capa más por la que la radiación es obstaculizada y puede ser absorbida por estas partículas, dispersando y reduciendo la cantidad de radiación directa, viéndose como una gran capa de neblina café a simple vista que cubre ciudades enteras.

Humedad y polvo en el aire

La humedad y el polvo en el aire que quedan suspendidas son factores más naturales como lo son las erupciones volcánicas o la evaporación del agua, el derretimiento del hielo algunas partículas de polvo son carbono o hollín en pequeñas partículas que estas mismas absorben la radiación mientras que las demás partículas afectan y dispersan la radiación.

2.1.3. Métodos tradicionales de medición y predicción

Los métodos de medición de la radicación solar han ido cambiando y evolucionando a través de tiempo para poder obtener una lectura con mejor precisión y una predicción de los rayos UV con mejor lectura.

Pirheliómetro

El pirheliómetro es una herramienta de medición de la irradiancia solar que es utilizada en la meteorología y climatología para saber cuánta radiación directa llega a la superficie ya que este se concentra específicamente en los rayos directos del disco solar este instrumento tiene un tubo largo por el cual cruzan la luz solar pero con la diferencia de que este solamente obtiene lectura de los rayos directos y limpios dejando de lado la radiación difusa que llega como es la que pasa por las nubes, este mismo obtiene la lectura a través de un sensor termoeléctrico el cual convierte esta energía directa en una señal proporcional a la radiación recibida, esta lectura se mide en $fracWm^2$ lo que se puede leer como vatios por metros cuadrados, otra característica de este instrumento es que debe tener un seguimiento del sol para poder obtener lecturas de alta precisión .

Piranómetro

Por otra parte, el piranómetro también es una herramienta de medición de irradiancia solar con la diferencia de que esta herramienta puede tener lecturas de toda la radiación no importa cuál sea su origen ya que no solo se concentra en el disco solar sino que también se concentra en el ambiente tomando, hasta puede obtener lecturas de radiación difusa lo que hace diferente la lectura con la herramienta pirheliómetro, este al igual que el pirheliómetro convierte los rayos de llegada en una lectura de vatios por metro cuadrado W/m2 a través de un sensor termoeléctrico o mejor conocido como termopila que genera una señal eléctrica igual que el calor que absorbe, esta herramienta es usara estaciones meteorológicas y plantas solares fotovoltaicas, en investigaciones climáticas y por supuesto en la agricultura, esta herramienta a deferencia de la otra es que este no debe tener un seguimiento del sol directamente ya que las lecturas las toma de mismo ambiente.

Heliógrafo de Campbell-Stokes

El heliógrafo es otra de las herramientas utilizadas en el área meteorológica cual funcionamiento es el medir la intensidad lumínica solar y la duración de la insolación de un lugar determinado de forma diaria, este proporciona datos importantes en el campo de la meteorología, este instrumento es una esfera de vidrio con un diámetro aproximado de 10 centímetros el cual tiene como función proyectar la luz de llegada sobre una cartulina el cual funcionara como una lupa quemando el rastro de la misma formando un camino conforme va moviéndose la trayectoria del sol, este proporciona información como las horas de salida y puesta del sol asi mismo como la intensidad del sol a cierta hora del día y los momentos en que el sol queda cubierto por las nubes este fenómeno también queda marcado en la cartulina. Los métodos de predicción de radiación solar que se usan para poder predecir la radiación solar en el planeta y sobre la misma superficie son los siguientes.

Métodos empíricos

Los métodos empíricos son los primeros en comenzar la predicción ya que estos están basados en simple estadística y fórmulas matemáticas, todos estos datos recolectados a través del tiempo, estos datos estadísticos son fácilmente de medir no son detallados de manera física solo es una estadística, algunos de estos son la temperatura, la nubosidad, latitud, altitud, humedad y horas de sol. Un ejemplo de uno de los métodos empíricos es el modelo del Amströng Prescott uno de los más conocidos para saber la radiación global conforme a las horas del sol reales y las más disponibles, para la temperatura hacen uso de amplitud térmica diaria ya sea máxima o mínima para saber la nubosidad, entre más cielos despejados existan más radiación solar existirá en la superficie, para la nubosidad de hace uso de 8 partes imaginarias en el cielo conocidas como octavos, los cuales dependiendo las partes que tengan más nubosidad será el porcentaje que ira aumentando o disminuyendo determinando que tan nublado está en el día cada uno utiliza una ecuación para determinar ya sea la nubosidad, radiación, temperatura, etc.

Tablas climatológicas

Las tablas climatológicas son grandes bases de datos promedio de capturas de cada uno de los datos climatológicos recolectados por varios años generalmente, normalmente se presentan en meses desde enero hasta diciembre en cada tabla se presentan datos específicos, asi como los máximos y mínimos, incluyendo los climas extremos quedando en el historial marcado sea el años que se presentó , el fin de trabajar y recolectar datos del clima ayudan mucho a los campos de la agricultura ya que a pesar de ser un proceso lento, pero con el pasar del tiempo ha sido de mucha ayuda por países ya que estas tablas y datos recolectados pueden ser por regiones específicas, cada una teniendo mejores resultados para la agricultura y para el estudio del clima ya que estas son presentadas por meses, años y décadas mostrando resultados que son bastantes predecibles puesto que por estación de año y región se hacer un promedio de manera estadística, dando a conocer en qué momento del año puede ser soleado, en que tiempo y que factores le pueden afectar, asi como otras estaciones en el año, como la lluvia, los días nublados y por supuesto días

ventosos, esto dependerá de cada región.

Mapas solares

Los mapas solares de uno de los métodos que se han utilizado durante mucho tiempo en el campo de la climatología pues esta toma lecturas de la radiación solar dándonos la información en el mapa global, este mapa nos indica donde hay más o menos radiación en la su superficie ya sean en mapas de 2 dimensiones o 3 dimensiones, de igual manera muestra estos datos de todas las regiones de manera diaria, mensual o anual.

2.2. Algoritmos de Aprendizaje

El aprendizaje automático es una rama de la inteligencia artificial que se enfoca en desarrollar algoritmos sofisticados capaces de aprender patrones a partir de datos. En el contexto de predicción, los modelos supervisados permiten estimar una variable de salida (como la radiación solar) a partir de un conjunto de entradas o variables conocidas (temperatura, humedad, viento, etc.).

Dentro de los modelos supervisados, las técnicas de regresión se utilizan para predecir valores continuos. Ejemplos comunes incluyen la regresión lineal, árboles de decisión, bosques aleatorios y redes neuronales artificiales. Estos modelos pueden capturar relaciones lineales y no lineales entre las variables.

2.2.1. Regresión lineal

El modelo de *regresión lineal* es una de las técnicas más simples y ampliamente utilizadas para tareas de predicción continua. Su objetivo es modelar la relación entre una variable dependiente y y un conjunto de variables independientes $\mathbf{x} = (x_1, x_2, \ldots, x_p)$, asumiendo que esta relación es lineal. La función de predicción toma la forma:

$$\hat{y} = \beta_0 + \sum_{j=1}^{p} \beta_j x_j = \mathbf{x}^\top \boldsymbol{\beta} + \beta_0$$

donde β_0 es el término independiente y $\boldsymbol{\beta}$ es el vector de coeficientes asociados a cada predictor. Los parámetros del modelo se estiman típicamente mediante el método de *mínimos cuadrados ordinarios* (OLS, por sus siglas en inglés), que consiste en minimizar la suma de los errores cuadráticos entre las predicciones del modelo y los valores reales:

$$\min_{\boldsymbol{\beta},\beta_0} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Este modelo es fácil de interpretar, computacionalmente eficiente y proporciona una base sólida para modelos más complejos. Su simplicidad permite identificar la influencia de cada variable sobre la salida, lo cual resulta especialmente útil en análisis exploratorios y contextos donde la interpretabilidad es prioritaria [14].

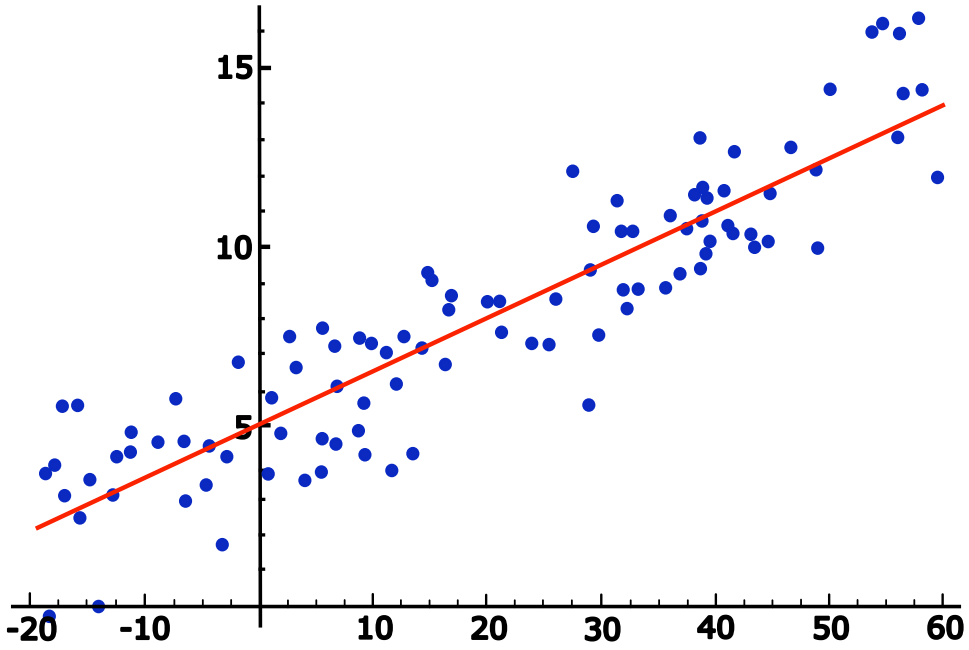


Figura 2.1: Modelo de Regresión Lineal

Sin embargo, la regresión lineal también presenta limitaciones importantes. La suposición de linealidad puede ser poco realista en muchos problemas reales, y el modelo es sensible a la presencia de multicolinealidad y valores atípicos (*outliers*). Además, no captura interacciones ni relaciones no lineales entre variables sin modificaciones adicionales [10]. Para superar estas limitaciones, pueden emplearse extensiones como la regresión

polinómica o técnicas de regularización como *Ridge Regression* y *Lasso* [17].

2.2.2. Árboles de decisión para regresión

El algoritmo de *árboles de decisión para regresión* (Regression Trees) construye un modelo predictivo basado en una estructura jerárquica de decisiones. A diferencia de los modelos lineales, los árboles no requieren suposiciones sobre la distribución de los datos ni la linealidad de las relaciones. El enfoque consiste en dividir recursivamente el espacio de atributos en regiones homogéneas en cuanto a la variable de salida, mediante la minimización del error cuadrático medio (MSE) en cada partición [1].

El proceso de construcción se basa en seleccionar, en cada nodo, el atributo x_j y el umbral s que minimicen la función de pérdida:

$$\min_{j,s} \left[\sum_{x_i \in R_1(j,s)} (y_i - \bar{y}_{R_1})^2 + \sum_{x_i \in R_2(j,s)} (y_i - \bar{y}_{R_2})^2 \right]$$

donde $R_1(j,s) = \{x \mid x_j \leq s\}$ y $R_2(j,s) = \{x \mid x_j > s\}$, y $\bar{y}_{R_k}$ representa la media de los valores de salida en la región R_k. El árbol se expande hasta que se alcanza una condición de parada, como una profundidad máxima o un número mínimo de muestras por nodo. Posteriormente, puede aplicarse una poda para reducir el sobreajuste.

Una de las principales ventajas de los árboles de decisión para regresión es su *interpretabilidad*: el modelo puede visualizarse fácilmente como un conjunto de reglas "si-entonces", lo que facilita su comprensión por parte de expertos no técnicos. Además, pueden manejar automáticamente atributos categóricos y numéricos, y no requieren normalización previa de los datos [15].

No obstante, los árboles de regresión también presentan limitaciones, como su *alta varianza*: pequeños cambios en los datos de entrenamiento pueden generar árboles significativamente distintos. También son susceptibles al sobreajuste si no se realiza una adecuada poda o regularización [10]. Para mitigar estos problemas, suelen utilizarse métodos basados en conjuntos como *Random Forests* o *Gradient Boosted Trees*, que mejoran la precisión y robustez del modelo a costa de una menor interpretabilidad.

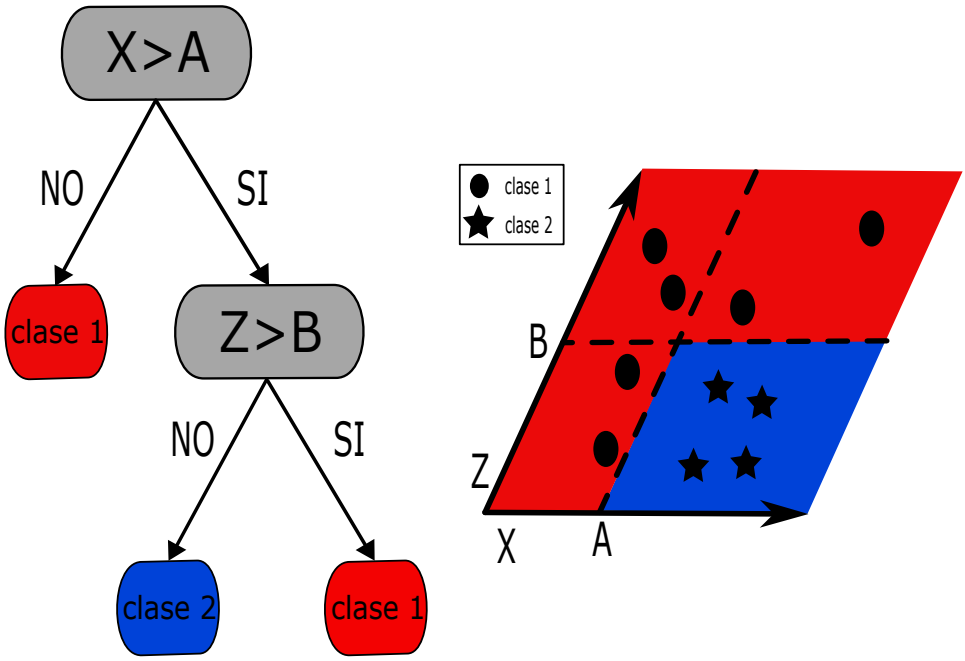


Figura 2.2: Modelo de árboles de decisión

2.2.3. Redes Neuronales para regresión

Las *Redes Neuronales Artificiales* (ANN, por sus siglas en inglés) son modelos computacionales inspirados en el funcionamiento del cerebro humano. Para tareas de regresión, las redes neuronales modelan relaciones complejas y no lineales entre las variables de entrada $\mathbf{x} = (x_1, x_2, \ldots, x_p)$ y una salida continua y, mediante una arquitectura compuesta por capas de nodos (neuronas) interconectados [11].

Una red neuronal típica de tipo *feedforward* con una capa oculta realiza la predicción $\hat{y}$ como:

$$\hat{y} = f\left(\sum_{j=1}^{H} w_j^{(2)} \cdot \sigma\left(\sum_{i=1}^{p} w_{ji}^{(1)} x_i + b_j^{(1)}\right) + b^{(2)}\right)$$

donde:

- $w_{ji}^{(1)}$ y $w_j^{(2)}$ son los pesos de las conexiones entre capas,
- $b_j^{(1)}$ y $b^{(2)}$ son los términos de sesgo (bias),

- $\sigma(\cdot)$ es la función de activación (por ejemplo, ReLU, tanh o sigmoide),

- $f(\cdot)$ es una función de salida (usualmente la identidad para regresión),

- H es el número de neuronas en la capa oculta.

El entrenamiento se realiza mediante el algoritmo de *retropropagación* (backpropagation), que ajusta los pesos minimizando una función de pérdida, comúnmente el error cuadrático medio (MSE):

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

donde $\boldsymbol{\theta}$ representa todos los pesos y sesgos del modelo.

Una de las principales ventajas de las redes neuronales es su capacidad para **aproximar funciones no lineales complejas** con alta precisión, especialmente en presencia de relaciones altamente no lineales entre variables [12]. Además, pueden escalarse fácilmente con el número de entradas y datos de entrenamiento.

Sin embargo, también presentan desventajas: son **difíciles de interpretar**, requieren una gran cantidad de datos para generalizar adecuadamente, y su entrenamiento puede ser computacionalmente costoso. También son susceptibles a **sobreajuste** si no se aplican técnicas de regularización como la detención temprana, *dropout* o penalización L_2 [9].

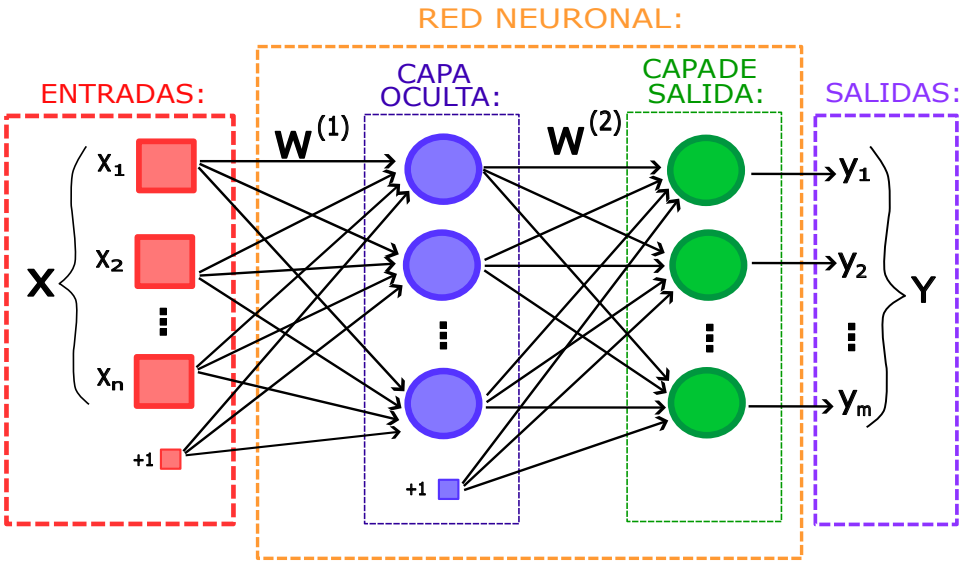


Figura 2.3: Modelo de redes neuronales

2.2.4. SVM para regresión

Las *Máquinas de Vectores de Soporte para regresión* (SVR, por sus siglas en inglés) son una extensión de los algoritmos SVM diseñados originalmente para clasificación, adaptadas para predecir variables continuas. El objetivo principal de SVR es encontrar una función $f(\mathbf{x})$ que se desvíe lo menos posible de los valores reales y_i, dentro de un margen de tolerancia ε, minimizando al mismo tiempo la complejidad del modelo [18] [2] [5].

El modelo toma la forma:

$$f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$$

y se entrena resolviendo el siguiente problema de optimización:

$$\min_{\mathbf{w},b,\xi_i,\xi_i^*} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)$$

sujeto a:

$$\begin{cases} y_i - \mathbf{w}^\top \mathbf{x}_i - b \leq \varepsilon + \xi_i \\ \mathbf{w}^\top \mathbf{x}_i + b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases}$$

donde:

- $\mathbf{w}$ es el vector de pesos del modelo,
- b es el sesgo,
- ξ_i, ξ_i^* son variables de holgura que permiten errores mayores a ε,
- C es el parámetro de penalización que controla el equilibrio entre la complejidad del modelo y la tolerancia a errores fuera del margen,
- ε define una banda donde los errores no son penalizados.

SVR puede extenderse a funciones no lineales mediante el uso de *núcleos* (kernel trick), permitiendo proyectar los datos de entrada a espacios de mayor dimensión donde una función lineal puede aproximar la salida con mayor precisión [3] [4].

Entre sus ventajas destacan su capacidad para modelar relaciones no lineales complejas con alta precisión y control de sobreajuste mediante el parámetro C y el margen ε [8]. Es especialmente útil en problemas donde se desea mantener la complejidad del modelo baja.

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 asociados [16][6] .

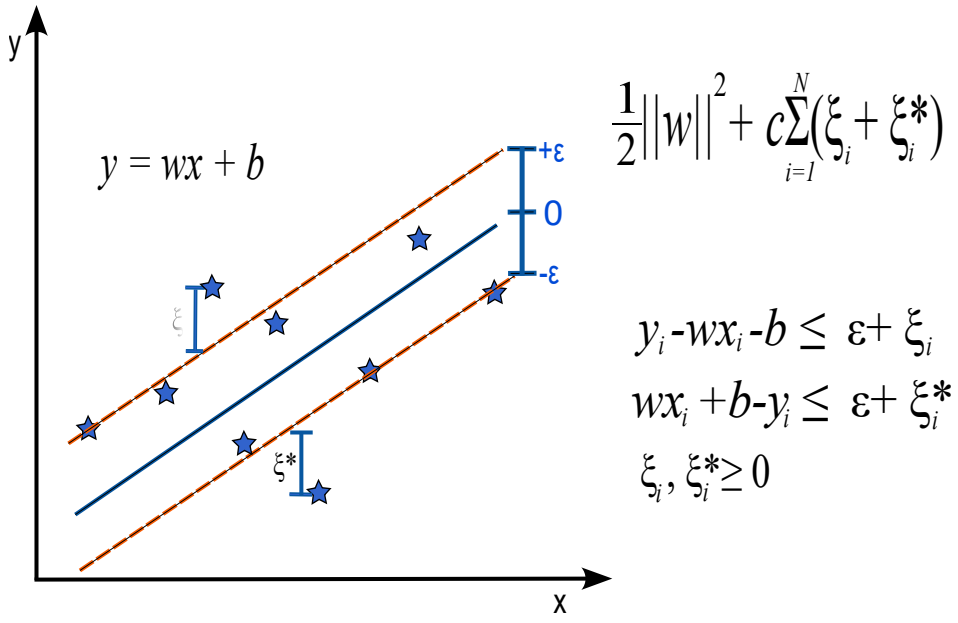


Figura 2.4: Modelo de SVM

2.3. Métricas de evaluación para modelos de regresión

Para evaluar el desempeño de los modelos de regresión es fundamental utilizar métricas que cuantifiquen la diferencia entre los valores predichos y los valores reales. A continuación se describen las métricas más utilizadas en la literatura [7, 19, 13]:

- **Error Absoluto Medio (MAE)**: mide el promedio del valor absoluto de los errores, y se expresa como:

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

  Es una métrica fácil de interpretar y robusta ante valores atípicos moderados.

- **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*.

- **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.

- **Error Absoluto Porcentual Medio (MAPE)**: expresa el error en términos porcentuales respecto al valor real:

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

Aunque intuitiva, esta métrica no es recomendable cuando existen valores cercanos a cero.

- **Coeficiente de Determinación (R^2)**: mide la proporción de la varianza total explicada por el modelo:

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

Un valor cercano a 1 indica un buen ajuste; un valor cercano a 0 sugiere un modelo poco explicativo.

- **Error Cuadrático Medio Relativo (RRSE)**:

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

Mide la efectividad del modelo comparada con una predicción constante usando la media.

- **Error Absoluto Relativo (RAE)**:

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

  Al igual que RRSE, compara con un modelo base constante. Valores menores a 1 indican mejor desempeño.

Cada una de estas métricas aporta una perspectiva diferente sobre el comportamiento del modelo. En la práctica, es recomendable utilizar múltiples métricas para obtener una evaluación más completa del desempeño de un modelo de regresión [7, 13].

2.4. Validación de Modelos de Predicción

La validación de modelos de predicción es el proceso que permite evaluar la capacidad de generalización de un modelo entrenado a partir de un conjunto de datos. Su propósito principal es estimar el desempeño del modelo sobre información no utilizada durante el entrenamiento, garantizando así que las predicciones no dependan exclusivamente de patrones específicos del conjunto de entrenamiento.

Formalmente, sea un conjunto de datos $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, donde $x_i \in \mathbb{R}^n$ representa el vector de características y $y_i \in \mathbb{R}$ (para problemas de regresión) o $y_i \in \{1, \ldots, K\}$ (para problemas de clasificación) representa la variable objetivo. El conjunto $\mathcal{D}$ se particiona en dos subconjuntos disjuntos:

$$\mathcal{D} = \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{test}}, \quad \mathcal{D}_{\text{train}} \cap \mathcal{D}_{\text{test}} = \emptyset$$

donde $\mathcal{D}_{\text{train}}$ se emplea para ajustar los parámetros del modelo $f_\theta(x)$, y $\mathcal{D}_{\text{test}}$ se utiliza para estimar su rendimiento mediante una métrica de error o precisión.

La función de pérdida o error se define como:

$$L(y, f_\theta(x))$$

y el error esperado del modelo se aproxima mediante la media empírica sobre el conjunto de validación:

$$\hat{E}[L] = \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{(x_i, y_i) \in \mathcal{D}_{\text{test}}} L(y_i, f_\theta(x_i))$$

Este valor representa una estimación del *riesgo generalizado*, el cual cuantifica la capacidad del modelo para predecir valores futuros o no observados.

Dependiendo de la naturaleza de los datos, se utilizan distintos esquemas de validación:

- **Validación Hold-Out:** se reserva una fracción del conjunto total para la prueba, entrenando el modelo con el resto de los datos.
- **Validación Cruzada (K-Fold Cross-Validation):** el conjunto de datos se divide en k subconjuntos; el modelo se entrena k veces, dejando uno distinto para prueba en cada iteración, y se promedian los resultados.
- **Validación para Series Temporales:** se preserva el orden temporal de las observaciones, entrenando con los datos pasados y validando con los futuros, de forma acumulativa o deslizante.

El proceso de validación permite identificar problemas de *sobreajuste* (*overfitting*) o *subajuste* (*underfitting*), y proporciona una base sólida para la selección de hiperparámetros, la comparación entre modelos y la estimación de la incertidumbre asociada a las predicciones.

Capítulo 3

Metodología

En esta Capítulo se muestra la metodología llevada a cabo al realizar los experimentos. La Figura 3.1 muestra la metodología utilizada. En nuestros experimentos utilizamos 1 conjunto de datos. Una vez seleccionado el conjunto de datos se implementaron varios proceso para mejorar los datos. Una vez procesado el conjunto de datos, se seleccionaron los modelos de aprendizaje máquina para predecir la radiación solar. El siguiente paso recae en el entrenamiento y la validación de los resultados. Finalmente se realiza la comparacion de los resultados con las diferentes métricas de desempeño. En este Capítulo se muestran en detalle cada uno de los pasos de la metodología propuesta.

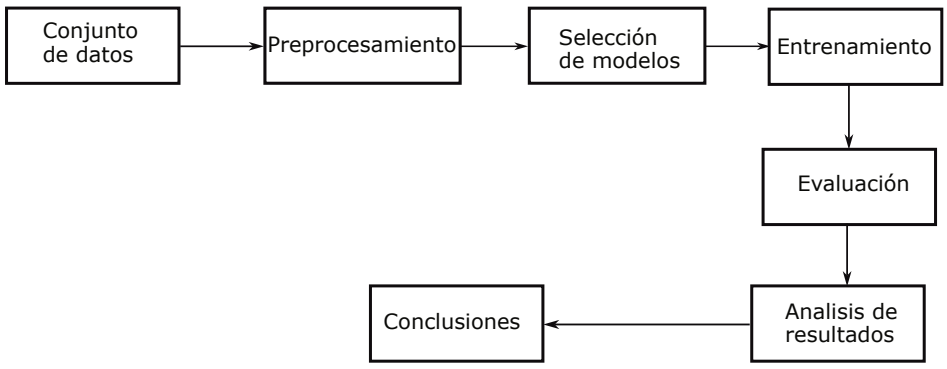
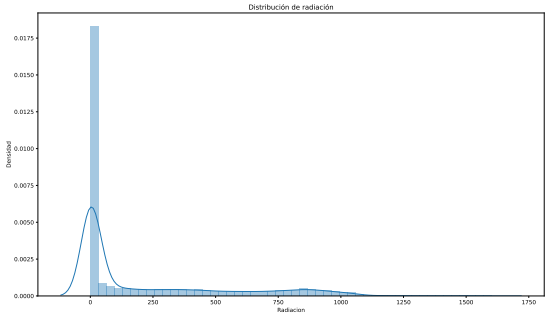
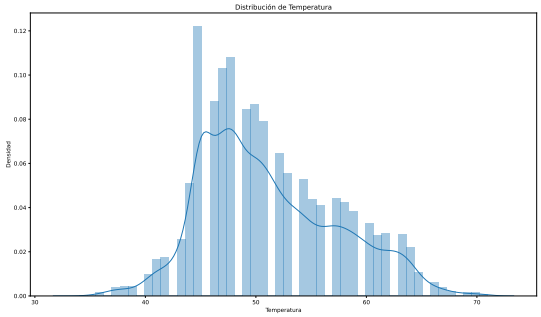


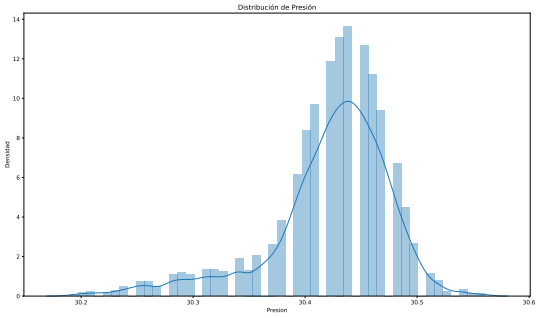
Figura 3.1: Metodología propuesta



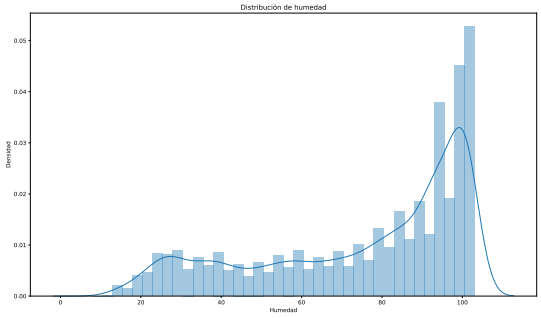
(a) Distribución de Radiación



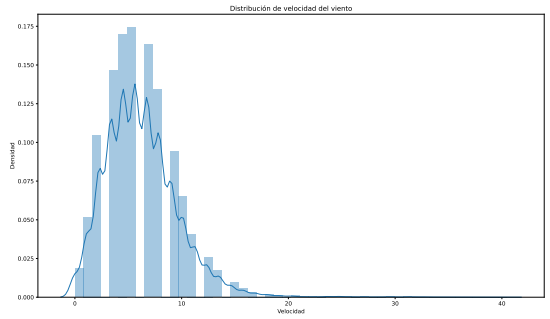
(b) Distribución de Temperatura



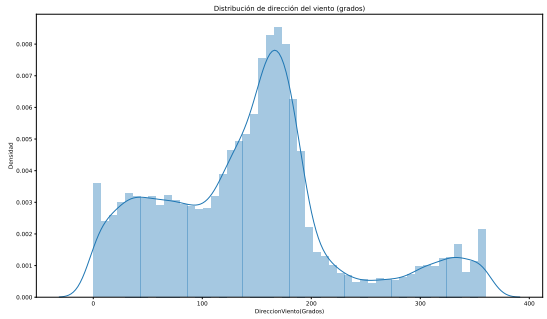
(c) Distribución de Presión



(d) Distribución de Humedad



(e) Distribución de Velocidad del viento



(f) Distribución de Dirección del viento

Figura 3.2: Distribuciones de cada una de las variables del conjunto de datos

3.1. Conjunto de datos

En los experimentos realizados, se utilizó un conjunto de datos que contiene registros meteorológicos. Los autores capturaron diversas variables durante un periodo de cuatro meses. Las variables independientes son: dirección del viento, velocidad del viento, humedad y temperatura, mientras que La variable dependiente es la radiación solar.

El conjunto de datos original fué obtenido de Kaggle. Este conjunto de datos contiene 32686 instancias, cada una con 10 atributos cuyos rangos de valores varian con diferentes máximos y mínimos. Las gráficas mostradas en las Figuras 4.1 muestrtan los diferentes intervalos en los que varian cada uno de los atributos. Estas Figuras permiten visualizar posibles valores atípicos e identificar tendencias.

Las 10 variables que definen a cada dato son: UNIXTime, Data, Time, Temperature, Pressure, Humidity, WindDirection, Speed, timeSunRise y TimeSunSet y la clase o la variable que necesitamos predecir: Radiation. Sin embargo, en la Figura solo se muestra la distribución de algunas de las variables del conjunto de datos.

3.2. Pre-procesamiento

Como primer paso de pre-procesamiento, para mejorar la confiabilidad y desempeño de las técnicas de predicción, se llevó a cabo un preprocesamiento y limpieza de datos. En cuanto a la limpieza, los datos estan completos y el conjunto de datos no contiene valores perdidos.

Sin embargo, como se menciono anteriormente el conjunto de datos original contiene 10 variables. Para eliminar algunas variables que no contribuyen en la predicción y que por el contrario solo introducen ruido en los modelos se analizó la correlación entre las variables.

La Tabla 3.1 y la Figura 3.3 muestran la correlación entre las variables del conjunto de datos. En los experimentos solo se utilizaron las variables más correlacionadas. El uso de las variables más correlacionadas facilita a las técnicas de IA encontrar o construir el modelo adecuado para predecir la radiación y ayuda significativamente en la creación de

modelos más precisos.

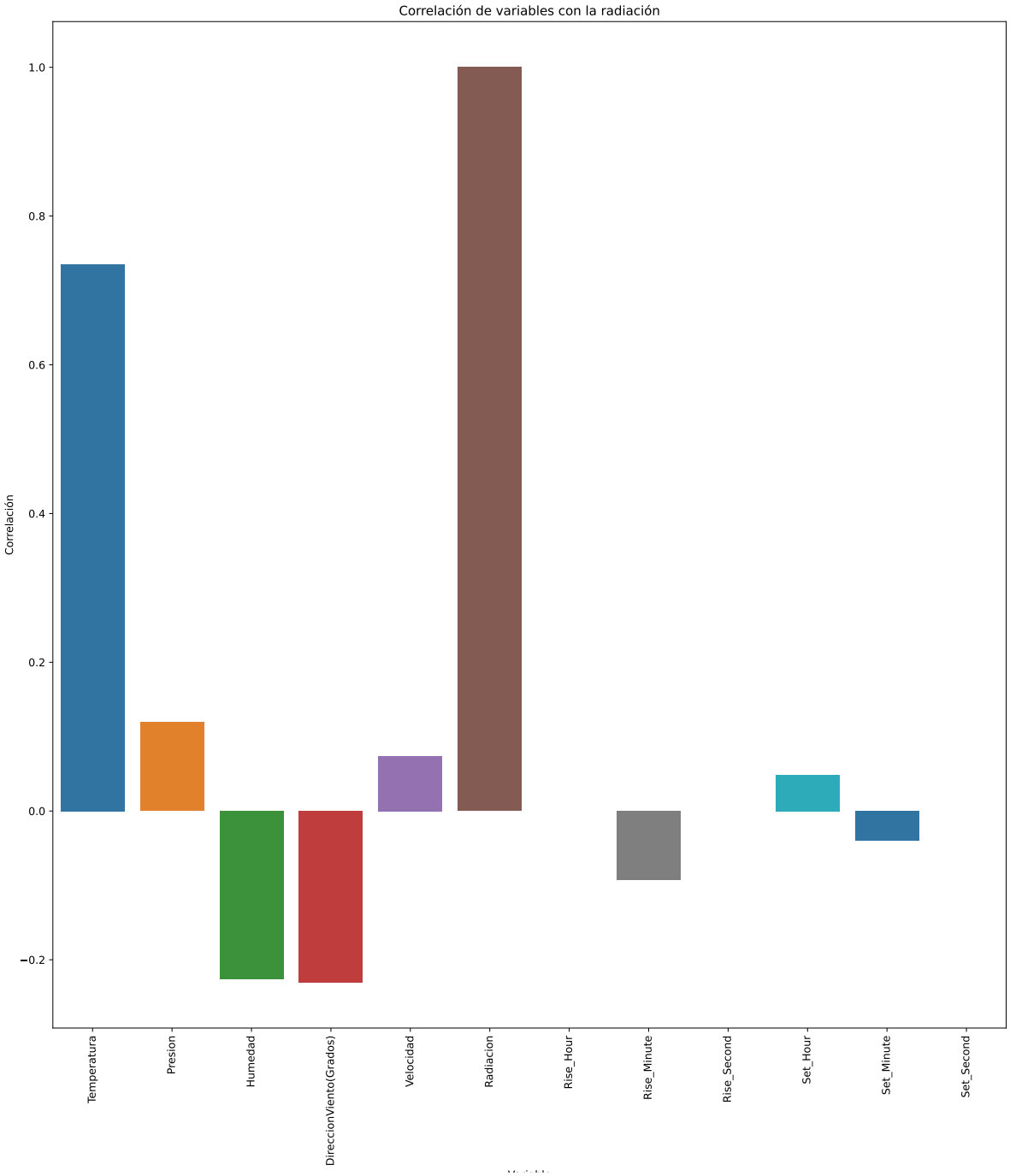


Figura 3.3: Correlación entre las variables del conjunto de datos

Índice	Variable	Correlación
0	Temperatura	0.734955
1	Presión	0.119016
2	Humedad	-0.226171
3	DirecciónViento(Grados)	-0.230324
4	Velocidad	0.073627
5	Radiación	1.000000
6	Rise_Hour	NaN
7	Rise_Minute	-0.092850
8	Rise_Second	NaN
9	Set_Hour	0.048719
10	Set_Minute	-0.039816
11	Set_Second	NaN

Tabla 3.1: Correlación de variables con la radiación

En el segundo paso de pre-procesamiento, las variables fueron normalizadas, en este proceso se ajustan los valores de cada variable a una media de 0 y una desviación estándar de 1. Esta técnica permite homogeneizar las escalas entre distintas variables, facilitando la comparación entre los modelos. Aunado a lo anterior, el uso detécnicas de normalización mejorar el rendimiento y desempeño de los algoritmos de aprendizaje automático al reducir la varianza no deseada.

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.

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. Redes Neuronales para regresión

La ventaja principal de las*Redes Neuronales Artificiales* podría ser su capacidad para aproximar funciones no lineales complejas. Su principal desventaja posiblemente sea su difícultad para interpretarse, además, requieren una gran cantidad de datos para generalizar adecuadamente, y su entrenamiento puede ser computacionalmente costoso. Al igual que los árboles de decisión también son susceptibles a sobreajuste.

3.3.4. 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

El conjunto de datos se dividió en subconjuntos de entrenamiento y prueba. La proporción que se utilizó en nuestros experimentos para conjuntos de entrenamiento y prueba fue de 75 % y 25 % respectivamente.

Para la obtención de cada modelo se optimizaron los parámetros y el modelo que se presenta es el mejor después de varias pruebas realizadas.

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.

3.6. Análisis de Resultados

En esta sección se mostrará un análisis de los resultados obtenidos, dentro del análisis que se llevará a cabo se mostrarán:

- Los modelos que mostraron un mejor desempeño en general en términos de las métricas seleccionadas.
- Los modelos que peor se desempeñaron
- Importancia del pre-procesamiento en el desempeño de los modelos
- Se obtendran las gráficas de desempeno de cada modelo

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 5 técnicas de inteligencia artificial. En cada uno de los modelos entrenados fueron optimizados los parámetros propios decada modelo. La primera sección muestra los resultados de la predicción de los 5 modelos implementados, estos son mostrados utilizando 4 métricas distintas. La segunda Sección muestra el análisis de las distribuciones de error de cada modelo. La tercera Sección muestra los resultados del Análisis Comparativo del error absoluto por muestra. La cuarta Sección muestra el Análisis Comparativo de la predicción Vs el valor real. La quinta Sección muestra Análisis Comparativo de la Residuos Vs Predicción. La sexta Seccion muestra la importancia de las variables, Finalmente la última Sección muestra las métricas de desempeño para los diferentes modelos utilizados.

4.1. Análisis Comparativo de Modelos de Regresión

En esta sección se muestra el desempeño de los cinco modelos de regresión utilizados. Como se describio en el Capítulo anterior los cinco modeos empleados, fueron: Random Forrest, Redes Neuronales, Árboles de Decisión, Regresión Lineal y Máquinas de Vectores de Soporte (SVR). En los experimentos llevados a cabo se utilizaron cuatro métricas de desempeño. La Tabla 4.1 muestra en sus columnas las métricas utilizadas.

Tabla 4.1: Desempeño de modelos de regresión

Modelo	**MSE**	**RMSE**	**RMSE Normalizado**	R^2
Redes Neuronales	29,466.94	171.66	0.13338	0.70353
Árboles de Decisión	31,835.64	178.43	0.13863	0.67970
Random Forrest	25,811.44	160.65	0.12483	0.74031
SVR	33,534.88	183.13	0.14228	0.66260
Regresión Lineal	44,228.84	210.31	0.16340	0.55500

La Tabla muestra las cuatro métricas utilizadas: MSE, RMSE, RMSE Normalizado y R^2. MSE mide el error cuadrático medio, es decir mientras más pequeña o más cercana a cero sea la medición es mejor.

La Tabla 4.1 muestra los resultados obtenidos. En la Tabla puede verse que de forma general el método que obtuvo el mejor desempeño es Random Forrest.

En la Tabla puede apreciarse que Random Forrest presenta el menor error cuadrático medio (MSE), asi como el menor RMSE, esto es una muestra que las predicciones realizadas por el modelo se encuentran en promedio más cercanas a los valores reales. En el caso de su RMSE normalizado, el modelo obtuvo un valor de 0.12483. El RMSE Normalizado muestra el desvio con respecto a la variable objetivo, en este caso, nos dice que el modelo solo se desvia un 12.48 % de la variable objetivo.

La métrica R^2 mide la variabilidad de los datos, en este caso al obtener $R^2 = 0{,}74031$, nos dice que el modelo explica el 74 % de la variabilidad de los datos.

Los resultados en la Tabla tambien muestran que el segundo modelo mejor evaluado son las redes neuronales. En los resultados puede verse que MSE y RMSE son ligeramente superiores a los obtenidos por Random Forrest.

El modelo con el peor desempeño es regresión lineal, casi en todas las métricas se puede ver la misma tendencia.

4.2. Análisis Comparativo de la distribución de residuos

En cuanto a la distribución del error, estos son mostrados en las graficas 4.1. En estas gráficas se muestran los resultados obtenidos con los cinco modelos empleados. En las Figuras es posible observar mediante la forma de la distribución que el ajuste de cada modelo varia en cada modelo. Una buena distribución del error sería una campana de Gauss con la mayoria de los errores agrupados alrededor de cero.

4.2.1. Árboles de decisión

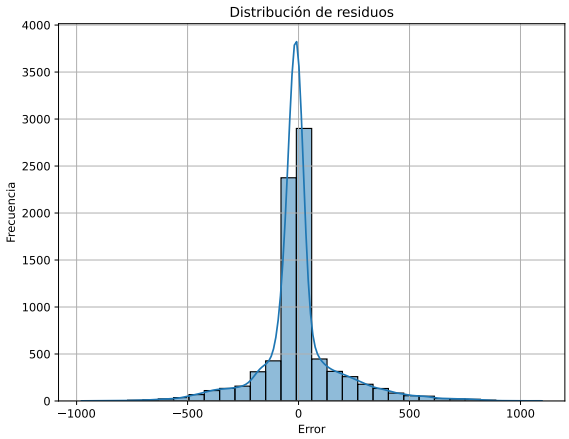
La forma de distribución de residuos de este modelo 4.1a tiene un pico muy agudo en cero. Esto significa que el modelo de árbol de decisión tiene una gran cantidad de predicciones acertadas (residuos cercanos a cero), pero también significa que el modelo comete errores más grandes frecuentemente. En muchos casos, los modelos sobreentrenados muestran este tipo de comportamiento.

4.2.2. Regresión lineal

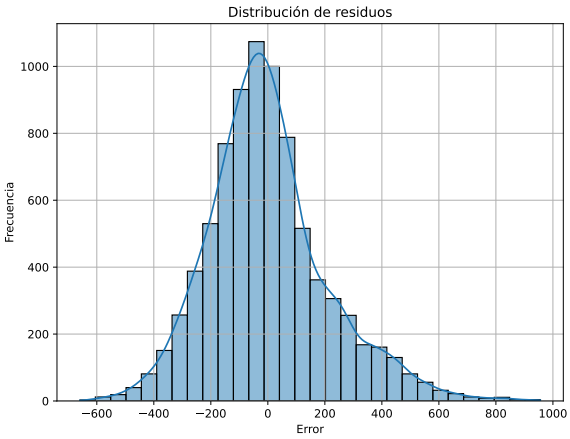
La distribución de residuos obtenida con el modelo de regresion lineal 4.1b es la que más se asemeja a una distribución normal o gaussiana. Esto significa que el modelo de regresión lineal se ajusta de forma general y que los errores que comete el modelo se distribuyen de forma aleatoria.

4.2.3. SVR (Support Vector Regression)

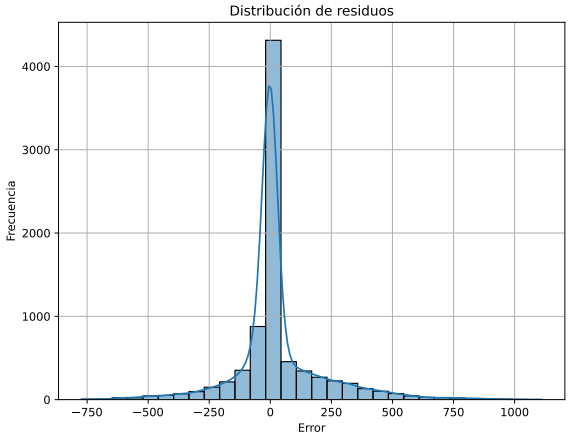
La distribución de residuos para SVR 4.1c logra un buen ajuste, En la gráfica es posible observar que los errores se concentran cerca a a cero. En la gráfica se puede observar que no existen errores de gran magnitud.



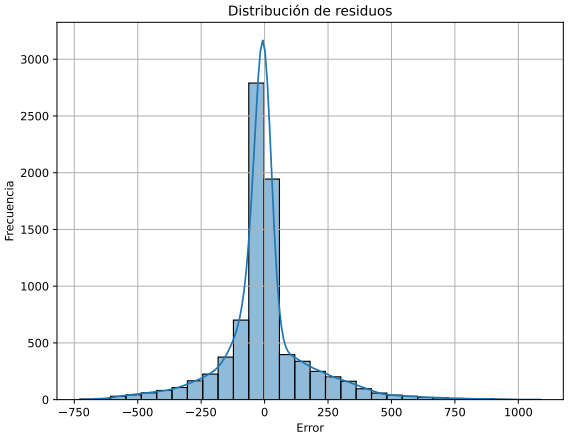
(a) Árboles de decisión



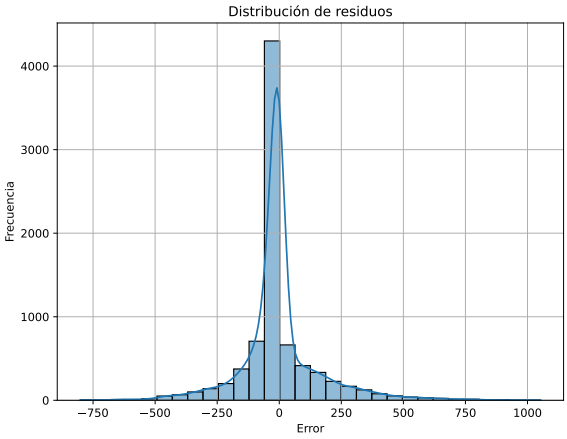
(b) Regresión lineal



(c) SVR



(d) Redes Neuronales



(e) Random Forrest

Figura 4.1: Distribuciones de Error para cada modelo

4.2.4. Redes neuronales

Como se puede obserevar en las graficas 4.1d la distribución que más se parece a una distribución ideal (gaussiana) de los cinco modelos mostrados son las redes neuronales. Esto significa que el modelo de redes neuronales tiene un excelente ajuste.

4.2.5. Random Forest

Similar a los árboles de decisión, la gráfiica de Random Forest 4.1e también tiene una distribución de residuos con un pico agudo en el centro, lo que sugiere una alta precisión. Los modelos basados en Random Forrest son modelos de ensamble que promedian las predicciones de múltiples árboles, reduciendo el sobreentranamiento.

En resumen, los modelos con los mejores rendimientos en términos de distribución de residuos son las redes neuronales y el Random Forest, ya que sus residuos se distribuyen de manera más parecida a la distribución normal ideal o tienen errores de gran magnitud menos frecuentes.

4.3. Análisis Comparativo del error absoluto por muestra

A partir de las Figuras 4.2 que muestran el error absoluto por muestra para cada modelo, se puede realizar un análisis visual de su desempeño. En este tipo de gráficas, un modelo ideal tendría una línea horizontal en cero, indicando que no hay errores. Sin embargo, es calor que esto es imposible en la practica, ya que un buen modelo mostrará la mayoría de los errores cerca de cero y tambien mostrará una baja dispersión.

4.3.1. Árboles de decisión

El gráfico de árboles de decisión 4.2a muestra una distribución de errores con una gran dispersión, con algunos picos muy altos en ciertas muestras. Esto significa que el modelo no solo tiene una variabilidad significativa en sus errores, sino que también comete errores

absolutos muy grandes y esporádicos. Esto puede ser un indicio de que el modelo es demasiado sensible a datos atípicos o que está sobreajustando el conjunto de entrenamiento, lo que provoca que no generalice bien a nuevos datos.

4.3.2. Regresión lineal

El modelo de regresión lineal 4.2b muestra un patrón de errores más uniforme y acotado. Aunque los errores absolutos son consistentemente altos para muchas muestras. Los picos de error son menos frecuentes, lo que sugiere que el modelo tiene un rendimiento más predecible y no está cometiendo errores masivos con tanta frecuencia.

4.3.3. SVR (Support Vector Regression)

El gráfico de SVR 4.2c muestra una mayor consistencia en los errores. Los valores absolutos de los errores se mantienen dentro de un rango más estrecho, con menos picos extremos. En otras palabras el modelo SVR es mas robusto a valores atípicos y controla mejor el error en todo el conjunto de datos.

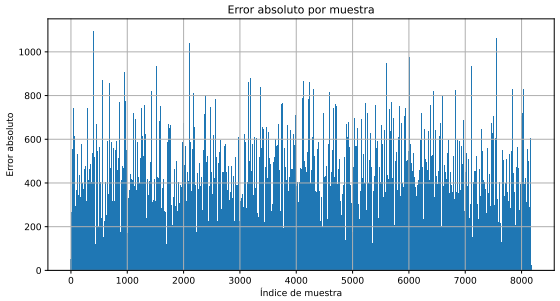
4.3.4. Redes neuronales

La gráfica de redes neuronales 4.2d muestra la menor magnitud de errores en general. La mayoría de los errores se agrupan dentro en un rango bajo, y la frecuencia y altura de los picos son significativamente menores en comparación con los otros modelos. Esto significa que el modelo de redes neuronales tiene el mejor rendimiento de los cinco, logrando predicciones más precisas de manera consistente en la mayoría de las muestras.

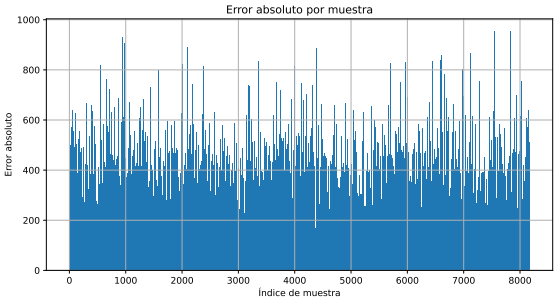
4.3.5. Random Forest

El gráfico de Random Forest 4.2e muestra un comportamiento similar al de SVR y regresión lineal, con un rango de errores absolutos relativamente uniforme. En resumen, el análisis visual del error absoluto por muestra sugiere que las redes neuronales y SVR

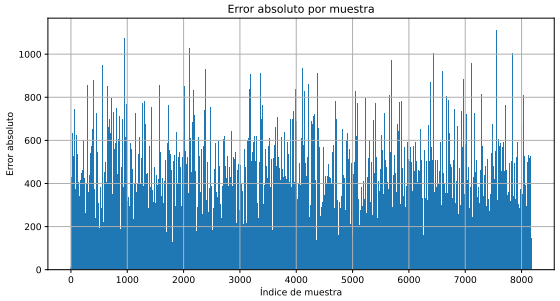
son los modelos con el mejor rendimiento, ya que logran mantener los errores en un rango más bajo y estable.



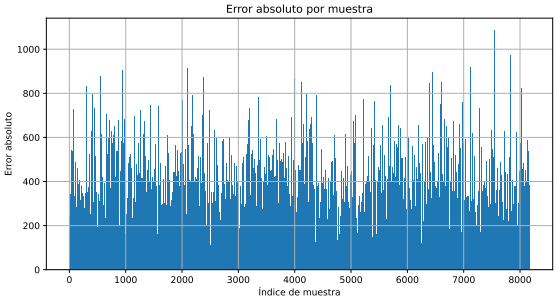
(a) Árboles de decisión



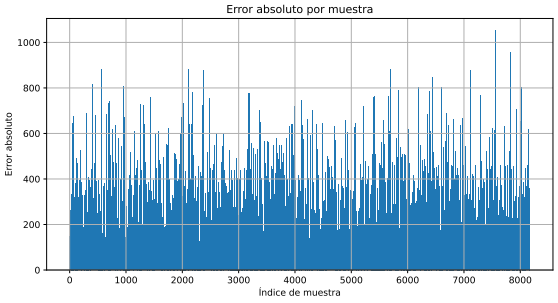
(b) Regresión lineal



(c) SVR



(d) Redes Neuronales



(e) Random Forrest

Figura 4.2: Resultados de Error Absoluto para cada modelo

4.4. Análisis Comparativo de la predicción Vs el valor real

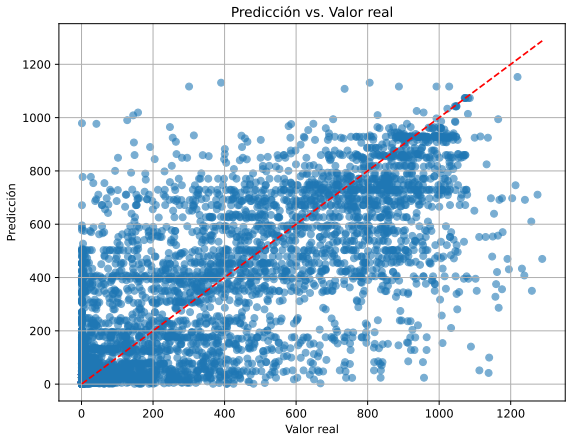
Para evaluar rigurosamente la eficacia de los modelos de regresión, es crucial comparar las predicciones generadas por un modelo con los valores reales contenidos en el conjunto de datos. En las representaciones gráficas de la predicción frente al valor real 4.3, el objetivo principal es que todos los puntos de datos converjan lo más cerca posible de la línea diagonal roja, que representa la condición óptima en la que la predicción es exactamente equivalente al valor real (y=x).

4.4.1. Árboles de decisión

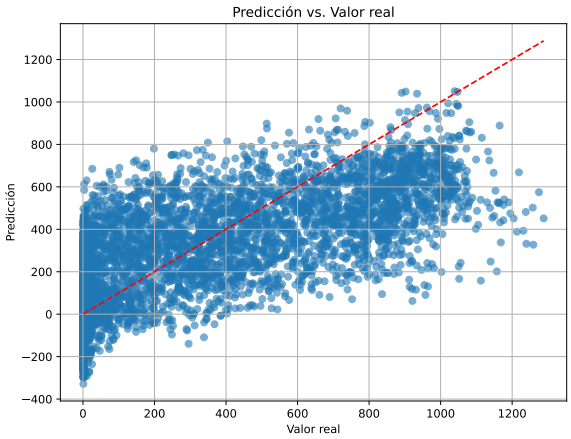
La gráfica de árboles de decisión 4.4a muestra una dispersión muy grande de los puntos. La mayoría de los datos se agrupan en grupos horizontales, indicando que el modelo tiende a predecir valores discretos y repetitivos, en lugar de un rango continuo de valores. Esta es una característica de los árboles de decisión, que dividen los datos en segmentos y asignan un valor constante a cada segmento. El alto grado de dispersión muestra un bajo rendimiento del modelo.

4.4.2. Regresión lineal

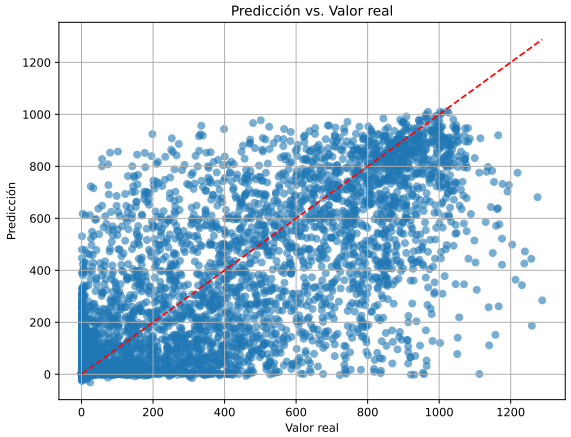
En la gráfica de regresión lineal 4.4b, los puntos se distribuyen de forma más uniforme alrededor de la línea que es ideal. Sin embargo, la dispersión es considerable, especialmente en los valores más altos. Esto indica que el modelo logra capturar la tendencia lineal general de los datos, pero su precisión es moderada, y tiene dificultades para predecir con exactitud los valores extremos.



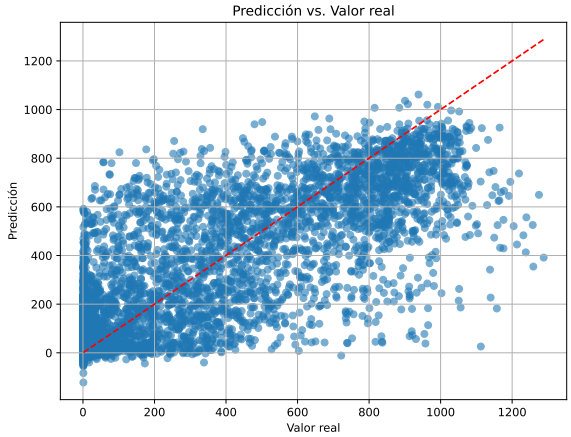
(a) Árboles de decisión



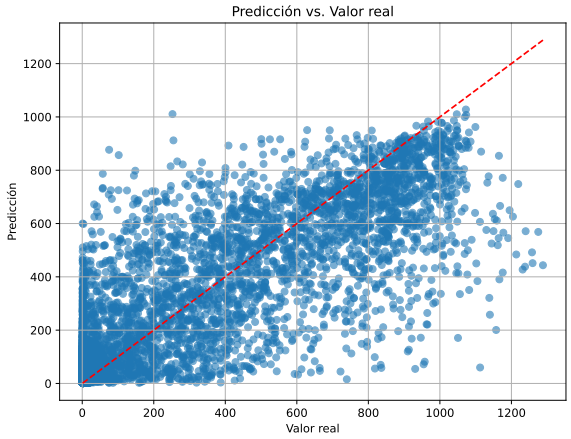
(b) Regresión lineal



(c) SVR



(d) Redes Neuronales



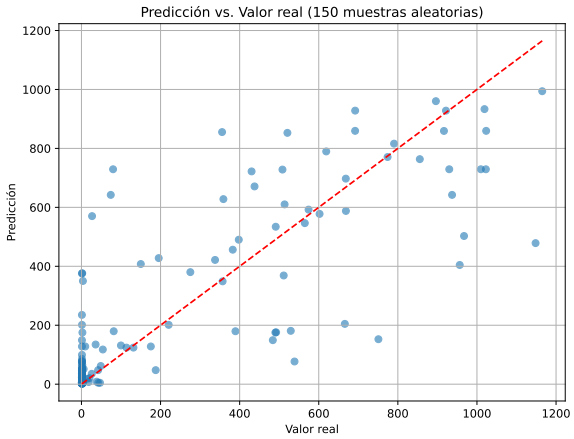
(e) Random Forrest

Figura 4.3: Resultados de prediccion del valor real para cada modelo

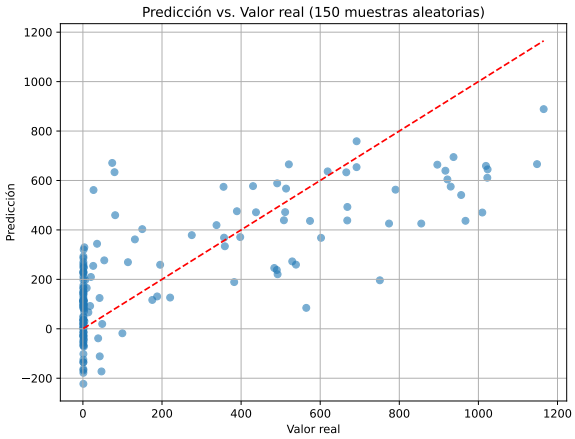
La Figura 4.4 presenta los resultados de la comparación entre los valores reales y los valores predichos por los clasificadores, utilizando un subconjunto aleatorio de 150 muestras del conjunto de prueba. Esta reducción en la cantidad de datos permite visualizar de forma más clara el comportamiento individual de las predicciones, así como las zonas en las que los clasificadores presentan un mejor o peor desempeño.

En la gráfica, la línea diagonal discontinua en rojo representa la línea ideal de predicción, es decir, el caso en el que el valor predicho coincide exactamente con el valor real. Por otro lado, los puntos azules representan las predicciones realizadas por el modelo. Cuanto más cerca se encuentren estos puntos de la línea ideal, menor será el error de predicción; mientras que una mayor distancia respecto a la línea refleja un mayor error cometido por el clasificador.

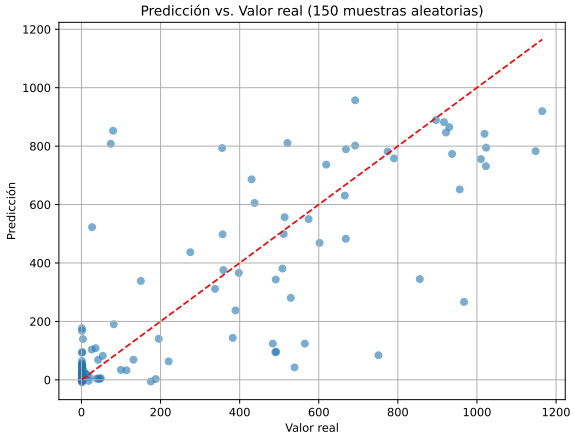
Al centrarse únicamente en 150 muestras, esta Figura permite identificar con mayor precisión las regiones en las que el modelo es más preciso, así como aquellas en las que tiende a subestimar o sobreestimar los valores reales. Esto facilita una evaluación más detallada del comportamiento del clasificador y su desempeño frente a diferentes rangos de valores de salida.



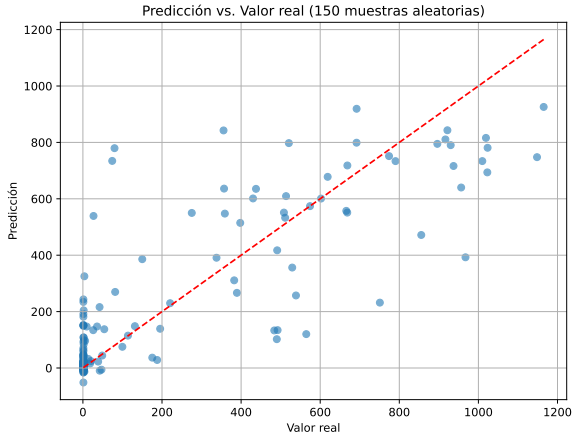
(a) Árboles de decisión



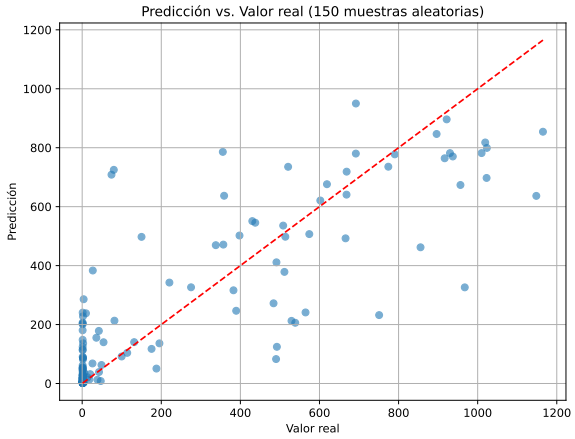
(b) Regresión lineal



(c) SVR



(d) Redes Neuronales



(e) Random Forrest

Figura 4.4: Resultados de prediccion del valor real para cada modelo

4.4.3. SVR (Support Vector Regression)

La gráfica de SVR 4.4c muestra que los puntos se agrupan de manera más densa alrededor de la línea ideal que en los modelos anteriores. La dispersión es menor y más uniforme en todo el rango de valores. Esto indica que SVR es un modelo más preciso y robusto, y logra un ajuste general que se acerca consistentemente a los valores reales.

4.4.4. Redes neuronales

La gráfica de redes neuronales 4.4d muestra la mejor alineación de todos los modelos. Los puntos se agrupan de forma muy compacta y densa alrededor de la línea ideal, con una dispersión mínima. Los errores son pequeños en la mayoría de las muestras, lo que se traduce en un rendimiento superior y una alta precisión en las predicciones. Este modelo no solo sigue la tendencia, sino que también captura las variaciones con mayor exactitud.

4.4.5. Random Forest

Similar a SVR y las redes neuronales, la gráfica de Random Forest 4.4e muestra un agrupamiento más cercano de los puntos a la línea diagonal que los modelos de regresión lineal y árboles de decisión. A diferencia del árbol de decisión individual, el modelo de ensamble reduce la varianza y la dispersión, lo que resulta en un rendimiento mucho más consistente y preciso. En general, los modelos de redes neuronales y SVR muestran el mejor ajuste, con predicciones muy cercanas a los valores reales.

4.5. Análisis Comparativo de la Residuos Vs Predicción

El análisis de las gráficas de Residuos vs. Predicciones es fundamental para evaluar la calidad de un modelo de regresión. Una distribución ideal de los puntos en estas gráficas es una nube aleatoria de puntos centrada en cero y sin ningún patrón discernible. Un patrón o una dispersión irregular de los residuos indican problemas en el modelo, como falta de linealidad o varianza no constante.

4.5.1. Árboles de decisión

La gráfica de árboles de decisión 4.5a muestra un patrón de bandas horizontales en los residuos. Esto se debe a que este tipo de modelo hace predicciones discretas, lo que se traduce en que los errores también se agrupen en bandas horizontales. La gran cantidad de puntos dispersos es muy grande, esto indica un ajuste del modelo muy pobre y con presencia de errores de gran magnitud. El modelo de árbol de decisión parece tener un rendimiento deficiente, especialmente con los valores más altos, donde la dispersión de los residuos es mayor.

4.5.2. Regresión lineal

La gráfica de regresión lineal 4.5b muestra una dispersión de los residuos en forma de abanico. A medida que los valores de las predicciones aumentan, la dispersión de los residuos también aumenta. Este patrón, conocido como heterocedasticidad, indica que la varianza de los errores no es constante en todo el rango de predicciones. Esto sugiere que el modelo no tiene un ajuste uniforme en todas las predicciones y que su precisión disminuye a medida que los valores predichos son más altos.

4.5.3. SVR (Support Vector Regression)

La gráfica de SVR 4.5c muestra una dispersión más uniforme y concentrada de los puntos alrededor de la línea cero, especialmente en el centro de las predicciones. Sin embargo, a medida que los valores de las predicciones se acercan a los extremos (especialmente en el lado derecho), la dispersión de los residuos parece aumentar ligeramente. Aunque este modelo es más robusto que los árboles de decisión y la regresión lineal, la leve heterocedasticidad sugiere que su rendimiento puede no ser uniforme en todo el rango de datos.

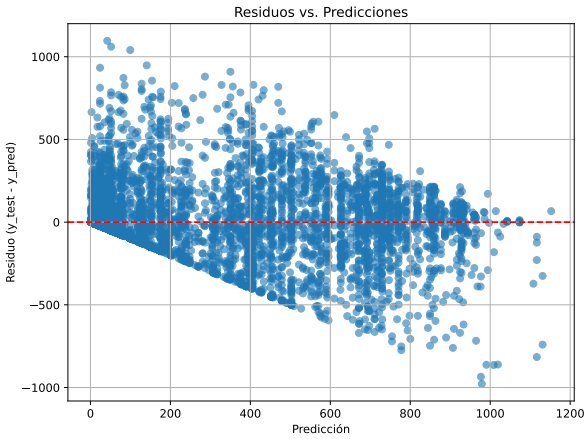
4.5.4. Redes neuronales

Las redes neuronales 4.5d muestran la distribución de residuos más cercana a la ideal. Los puntos están dispersos de manera aleatoria y se agrupan de forma compacta alrededor de la línea cero. No se observa ningún patrón evidente, lo que indica que el modelo no tiene sesgos y que la varianza de los errores es constante. Esto es un signo de un modelo de alto rendimiento y gran precisión, que logra un ajuste uniforme en todo el rango de predicciones.

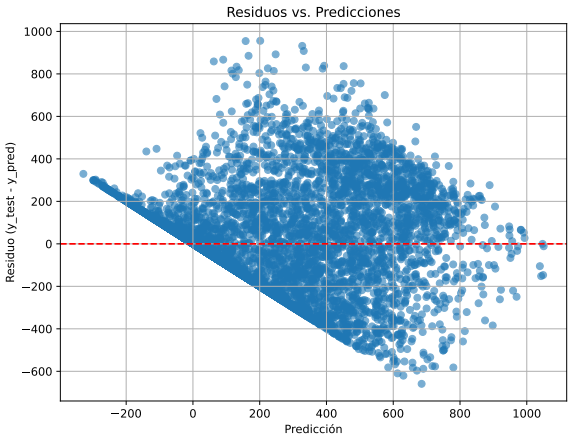
4.5.5. Random Forest

El gráfico de Random Forest 4.5e exhibe una dispersión de residuos similar a la del modelo de SVR, pero con un patrón de bandas horizontales menos pronunciado que en el árbol de decisión individual. A medida que las predicciones aumentan, la dispersión de los residuos también aumenta. Sin embargo, el ensamble de árboles (Random Forest) reduce la varianza y mejora la generalización, lo que resulta en un mejor rendimiento general comparado con un solo árbol de decisión.

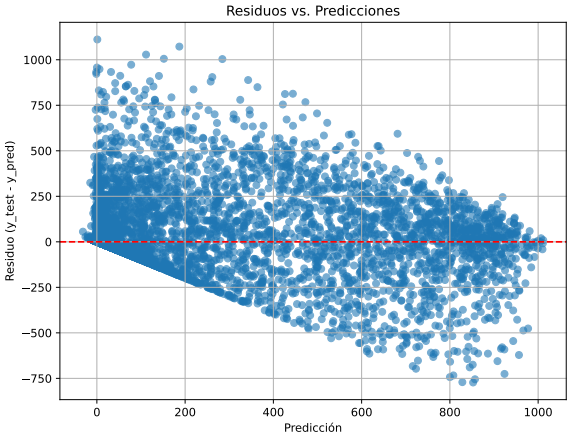
En resumen, el análisis de los residuos vs. las predicciones muestra que los modelos de redes neuronales y SVR tienen el mejor rendimiento, con residuos más aleatorios y una varianza más constante. La regresión lineal y el Random Forest exhiben heterocedasticidad, y el árbol de decisión presenta un patrón de predicciones discretas.



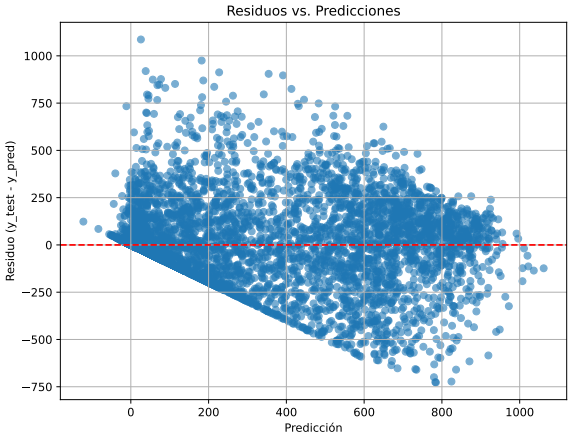
(a) Árboles de decisión



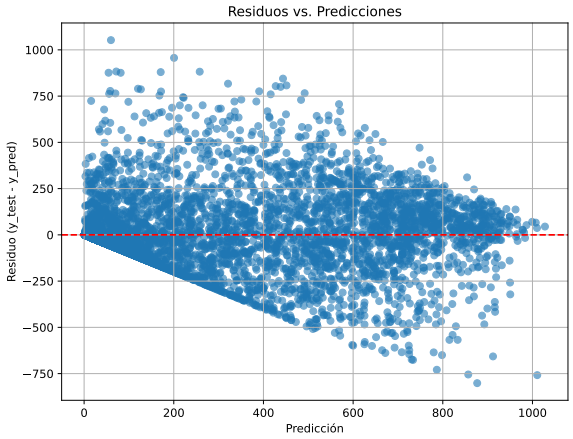
(b) Regresión lineal



(c) SVR



(d) Redes Neuronales



(e) Random Forrest

Figura 4.5: Resultados de residuo Vs Predicción

4.6. Análisis Comparativo de la importancia de las variables

La importancia de las variables es indispensable para comprender su influencia en el modelo de predicción. Cada modelo aprende de forma distinta las relaciones entre variables y la variable objetivo. Por lo tanto, la importancia no es una propiedad de los datos sino del modelo al ser entrenado sobre los datos. Los siguientes pasos describen como se calcula la importancia de las variables

Pasos para calcular importancia de las variables

1. Dado un modelo ya entrenado $f(x)$ y un conjunto de prueba $(Xtest, ytest)$

2. Se calcula el desempeño original con alguna métrica de desempeño

3. Para cada varible específica x_j se tomas esa columna del conjunto de prueba y se barajan aleatoriamente sus valores (se permutan las filas de esa variable). Es decir, cada observación mantiene todas las demás variables iguales, excepto x_j. Esto rompe la correspondencia real entre x_j y y, que ahora tiene el valor de otra muestra al azar, sin alterar la distribución de x_j ni su escala.

4. Se evalua el modelo con este conjunto modificado. Si el modelo dependía mucho de x_j su desempeño caerá significativamente.La importancia se calcula como la diferencia entre el desempeño original y el desempeño permutado

Tabla 4.2: Comparación del cálculo de importancia de variables entre diferentes modelos

Modelo	**Cómo aprende**	**Cómo influye en la importancia de las variables**
Random Forest / Árboles de decisión	Divide el espacio de entrada en regiones según umbrales de las variables.	Las variables que permiten realizar cortes que reducen más la varianza del modelo obtienen mayor importancia.
SVR (Support Vector Regression)	Usa una función *kernel* (lineal, RBF, etc.) para proyectar los datos en un espacio de alta dimensión.	Si una variable afecta considerablemente la posición de los vectores soporte en el espacio transformado, su permutación alterará más las predicciones.
Regresión lineal	Asigna coeficientes proporcionales al impacto lineal de cada variable sobre la salida.	La importancia depende del tamaño absoluto de los coeficientes y de la escala de las variables.
Redes Neuronales (ANN / Deep Learning)	Aprenden representaciones jerárquicas y no lineales de los datos mediante múltiples capas de neuronas conectadas.	La importancia de una variable depende de cómo sus valores afectan las activaciones internas y la salida final del modelo.

Según las gráficas de importancia de las características para cada modelo 4.6, se puede determinar qué variables tienen un mayor impacto en las predicciones. La importancia se mide de forma diferente para cada tipo de modelo, pero en general, un valor más alto indica una mayor influencia.

4.6.1. Árbol de decisión y SVR

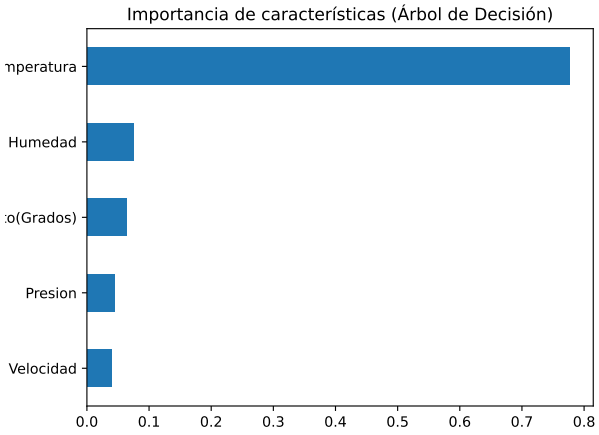
En las gráficas de los modelos de Árbol de decisión 4.6a y SVR , la temperatura es la variable más importante. En las gráficas, se muestra la temperatura con un valor de importancia superior a 0.7 en ambos casos. Esto significa que las predicciones de estos modelos dependen principalmente de la temperatura, mientras que las otras variables como la humedad, presión, velocidad y dirección del viento (grados) tienen una importancia mucho menor.

4.6.2. Regresión lineal

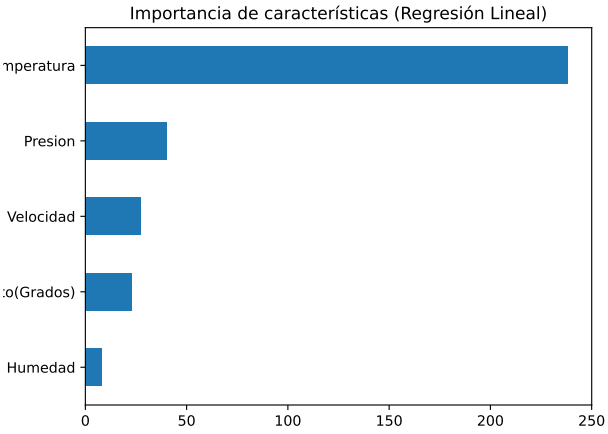
En la regresión lineal 4.6b, la temperatura también es la variable más importante, con un valor de importancia que supera los 200, mucho más alto que las demás. La humedad y la presión tienen una importancia moderada, mientras que la velocidad y la dirección del viento tienen un impacto mucho menor.

4.6.3. Redes neuronales

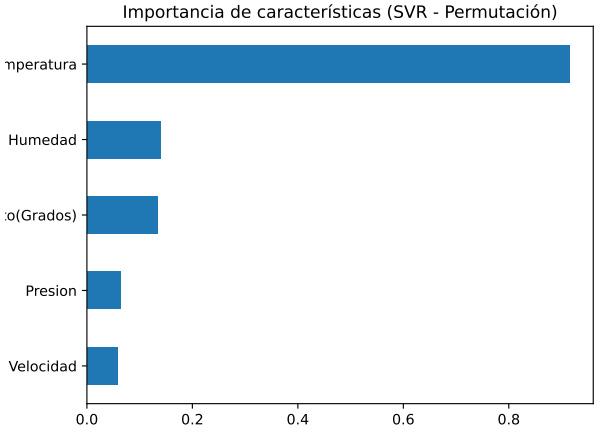
En el modelo de Redes neuronales 4.6d, la temperatura es nuevamente la característica más influyente, con un valor de importancia que se acerca a 0.8. La humedad y la dirección del viento (grados) son las siguientes variables más importantes, mientras que la presión y la velocidad tienen un impacto mínimo.



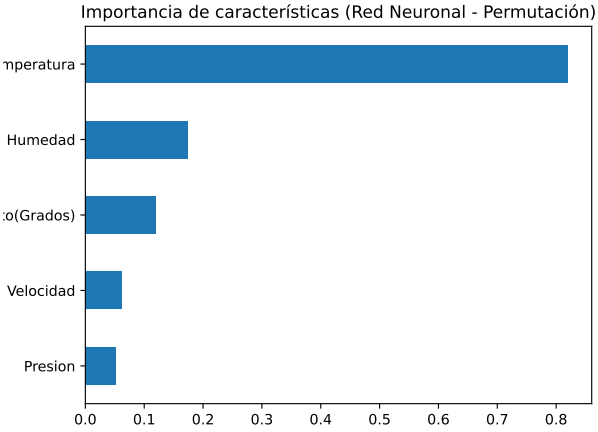
(a) Árboles de decisión



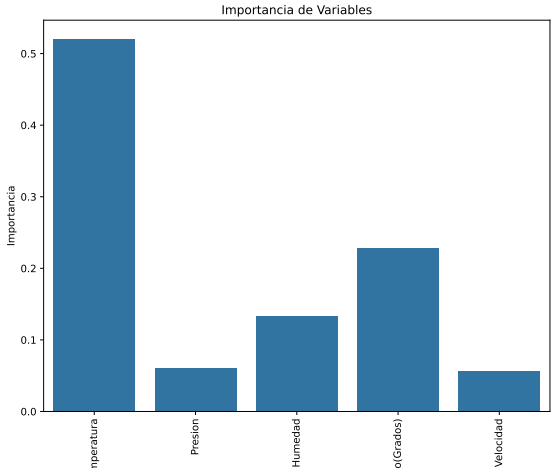
(b) Regresión lineal



(c) SVR



(d) Redes Neuronales



(e) Random Forrest

Figura 4.6: Importancia de variables utilizadas

4.6.4. Random Forest

Similar a los otros modelos, en el Random Forest 4.6e la temperatura es la variable más importante. Sin embargo, en este modelo la importancia de la temperatura (casi 0.4) es comparativamente más baja que en los demás modelos. Esto se debe a que el modelo de Random Forest considera las interacciones entre las variables, por lo que la importancia se distribuye de manera más uniforme entre ellas. La presión, humedad, y la dirección del viento (grados) tienen una importancia relevante en este modelo.

En todos los modelos, la temperatura es consistentemente la variable más importante para las predicciones. La humedad y la presión también tienen un impacto significativo, mientras que la velocidad y la dirección del viento son generalmente las menos relevantes, con la excepción de las redes neuronales, donde la dirección del viento tiene una importancia ligeramente mayor. La importancia de las variables varía entre los modelos, lo que indica que cada algoritmo utiliza las características de manera diferente para hacer sus predicciones.

4.7. Análisis Comparativo de las métricas para cada modelo

Para comparar el rendimiento de los modelos, analizaremos las métricas de error (MSE, RMSE, RMSE Normalizado) y la bondad de ajuste (R2), junto con los gráficos de distribución de residuos, error por muestra y predicción vs. valor real.

4.7.1. Análisis de Métricas

Los gráficos de barras de MSE, RMSE y RMSE Normalizado muestran que Regresión Lineal tiene los valores más altos, indicando que este modelo comete los errores más grandes en promedio. Por otro lado, Random Forest y las Redes Neuronales tienen los valores más bajos en estas métricas de error, sugiriendo que son los modelos con mejor desempeño en términos de precisión.

El gráfico de R^2 mide la proporción de la varianza en la variable dependiente que es predecible a partir de la(s) variable(s) independiente(s). Un valor más cercano a 1 indica un mejor ajuste. Aquí, Random Forest y las Redes Neuronales tienen los valores más altos (>0.85), lo que significa que explican una gran parte de la variabilidad en los datos. Regresión Lineal tiene un valor de R^2 significativamente más bajo (alrededor de 0.25), lo que confirma su bajo rendimiento.

4.7.2. Análisis de Residuos

Distribución de Residuos: La distribución ideal de residuos se asemeja a una campana de Gauss, centrada en cero. Las Redes Neuronales y Random Forest muestran distribuciones que se ajustan mejor a este patrón ideal, lo que indica que sus errores son aleatorios y no sesgados. Por el contrario, la Regresión Lineal tiene una distribución más dispersa, y el Árbol de Decisión muestra una distribución de Laplace con un pico agudo y colas pesadas.

Residuos vs. Predicciones: Las gráficas de residuos contra predicciones revelan si el modelo tiene sesgos o si la varianza de los errores es constante. Las Redes Neuronales muestran una nube de puntos aleatoria y centrada en cero, lo cual es el comportamiento ideal. La Regresión Lineal presenta una forma de abanico (heterocedasticidad), lo que indica que la precisión disminuye a medida que los valores predichos aumentan. Los Árboles de Decisión muestran bandas horizontales de errores, un patrón no deseado.

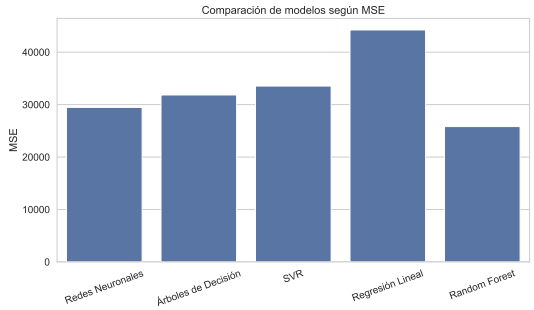
4.7.3. Análisis Visual del Error

Error Absoluto por Muestra: Las gráficas de error por muestra muestran la magnitud de los errores para cada punto de datos. Las Redes Neuronales y Random Forest muestran los errores más bajos y consistentes a lo largo de todas las muestras. La Regresión Lineal y los Árboles de Decisión tienen errores más grandes y una mayor variabilidad, con picos de error más frecuentes y de mayor magnitud.

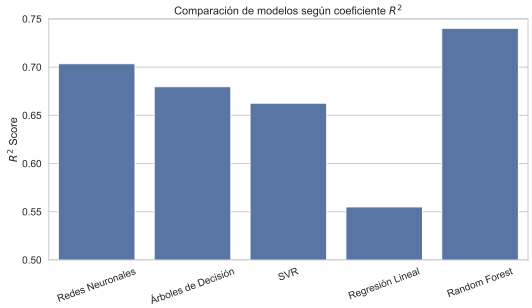
Predicción vs. Valor Real: En estas gráficas, los puntos deben alinearse lo más cerca posible de la línea diagonal. Las Redes Neuronales muestran la mejor alineación, con

una dispersión mínima de los puntos, lo que indica una alta precisión en las predicciones. Random Forest y SVR también muestran una buena alineación, mientras que la Regresión Lineal y el Árbol de Decisión muestran una dispersión mucho mayor y patrones indeseables (como la agrupación en el caso del árbol de decisión).

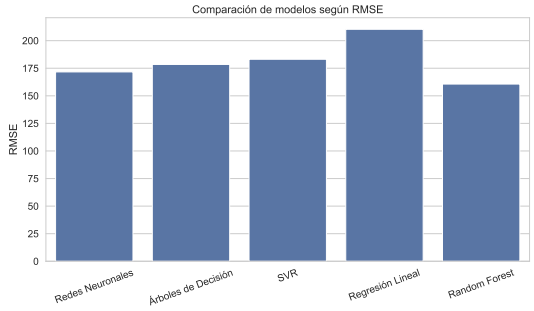
Basado en todas las métricas y visualizaciones, Redes Neuronales y Random Forest son los modelos con el mejor rendimiento. Logran la mayor precisión (menor error) y una mejor bondad de ajuste (R2), con distribuciones de residuos aleatorias y sin patrones indeseables. La Regresión Lineal tiene el peor desempeño, con altos errores, baja capacidad explicativa y problemas de heterocedasticidad. SVR y Árbol de Decisión tienen un rendimiento intermedio, pero con patrones de error menos ideales en comparación con los modelos de mayor rendimiento.



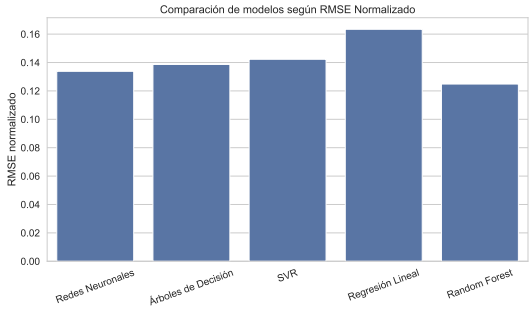
(a) MSE



(b) R^2



(c) RMSE



(d) RMSE Normalizado

Figura 4.7: Resultados de Comparación de las métricas de desempeño

Capítulo 5

Conclusiones

5.1. Conclusiones

Esta tesis muestra el análisis comparativo de varios modelos para predecir la radiación solar mediante técnicas de inteligencia artificial. Las técnicas de inteligencia artificial comparadas fueron: Árboles de decision, regresión lineal, Redes neuronales, Máquinas de vectores de soporte para regresión y random Forrest. En la tesis se muestran los modeos y los resultados obtenidos para cada uno de ellos.

Los resultados obtenidos muestran que para el conjunto de datos utilizado los modelos que mejor se desempeñan en este conjunto de datos son: redes neuronales y random Forrest. Aunque el algoritmo Random Forest mostró un buen desempeno al predecir la radiación solar.

Bibliografía

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

[2] Jair Cervantes, Xiaoou Li y Yu Wen. "Support Vector classification for large data sets by reducing training data with change of classes". En: *2008 IEEE International Conference on Systems, Man and Cybernetics*. IEEE, 2008. DOI: 10.1109/icsmc.2008.4811689.

[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] Ian Goodfellow, Yoshua Bengio y Aaron Courville. *Deep Learning*. MIT Press, 2016.

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

[11] Simon Haykin. *Neural Networks: A Comprehensive Foundation*. 2nd. Prentice Hall, 1999.

[12] Kurt Hornik, Maxwell Stinchcombe y Halbert White. "Multilayer feedforward networks are universal approximators". En: *Neural Networks* 2.5 (1989), págs. 359-366.

[13] 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.

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

[15] J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

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

[17] Robert Tibshirani. "Regression Shrinkage and Selection via the Lasso". En: *Journal of the Royal Statistical Society: Series B* 58.1 (1996), págs. 267-288.

[18] Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

[19] 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.