Instituto Tecnológico y de Estudios Superiores de Occidente

Reconocimiento de validez oficial de estudios de nivel superior según acuerdo secretarial 15018, publicado en el Diario Oficial de la Federación del 29 de noviembre de 1976.

Departamento de Matemáticas y Física

Maestría en Ciencia de Datos

Pronóstico de temperatura en un invernadero usando algoritmos de aprendizaje supervisado

---

TESIS para obtener **GRADO** de
MAESTRO EN CIENCIA DE DATOS

Tesis presentada por:
Jesús Rodrigo Ponce González

Asesor de Tesis:
Dr. Héctor Alonso Guerrero Osuna

Tlaquepaque, Jalisco, Mayo, 2023

Instituto Tecnológico y de Estudios Superiores de Occidente
Reconocimiento de validez oficial de estudios de nivel superior según acuerdo secretarial 15018, publicado en el Diario Oficial de la Federación del 29 de noviembre de 1976.

Departamento de Matemáticas y Física

Formulario de Aprobación Maestría en Ciencia de Datos

Título de Tesis: **Pronóstico de temperatura en un invernadero usando algoritmos de aprendizaje supervisado**

Autor: **Jesús Rodrigo Ponce González**

Tesis aprobada para completar todos los requisitos de grado para la Maestría en Ciencias en Ciencia de Datos.

Tutor de Tesis, **Dr. Héctor Alonso Guerrero Osuna**

Sinodal, **Dr. Luis Fernando Luque Vega**

Sinodal, **Dr. Carlos Alberto Olvera Olvera**

Coordinadora Académico, **Dra. Rocío Carrasco Navarro**

Tlaquepaque, Jalisco, Mayo, 2023

Pronóstico de temperatura en un invernadero usando algoritmos de aprendizaje supervisado

Jesús Rodrigo Ponce González

Resumen

El microclima de un invernadero favorece el crecimiento de los cultivos y evita el desarrollo de patógenos dañinos. Dentro de las actividades claves se encuentran la recolección, monitoreo y control de las variables del proceso como la temperatura, humedad, ventilación, entre otras. Este trabajo presenta el pronóstico de la temperatura interna del invernadero con base en las condiciones climáticas externas e internas, en diferentes tiempos, con el fin de poder realizar ajustes anticipados y un sistema de control que ayude a mantener un microclima adecuado; a través de un algoritmo de aprendizaje automático.

Distintos métodos fueron utilizados inicialmente para la comprensión de la complejidad de los datos, así mismo, se encontró que no es necesario considerar todos los parámetros dentro del invernadero para obtener un pronóstico aceptable. Modelos más robustos como la máquina de vector de soporte para regresión (SVR, por sus siglas en inglés) y refuerzo de gradientes extremo (XGBoost, por sus siglas en inglés) son capaces de realizar pronósticos, para SVR con un error medio absoluto promedio de $1.282°C$, $1.747°C$, $2.209°C$ para 30, 45 y 60 minutos, respectivamente; mientras que XGBoost obtuvo resultados en mucho menor tiempo, el modelo logró pronosticar un error medio absoluto promedio de $1.420°C$, $1.968°C$, $2.458°C$ para 30, 45 y 60 minutos, respectivamente. Analizando los resultados, se encontró que el modelo SVR es una herramienta efectiva para el pronóstico de temperatura interna de un invernadero en diferentes tiempos.

Palabras clave: Agricultura, Invernadero, Temperatura Interna, Pronóstico, Aprendizaje Automático

Abstract

The microclimate of a greenhouse favors crop growth and prevents the development of harmful pathogens. Among the key activities are the collection, monitoring and control of process variables such as temperature, humidity, ventilation, among others. This work presents the forecast of the internal temperature of the greenhouse based on external and internal climatic conditions, at different times, in order to be able to make early adjustments and a control system that helps to maintain an adequate microclimate; through a machine learning algorithm.

Different methods were initially used to understand the complexity of the data and it was found that it is not necessary to consider all parameters within the greenhouse to obtain an acceptable forecast. More robust models such as support vector regression (SVR) and extreme gradient boosting (XGBoost) are able to make forecasts, for SVR with an average mean absolute error of $1.282°C$, $1.747°C$, $2,209°C$ for 30, 45 and 60 minutes, respectively; while XGBoost obtained results in much less time, the model was able to forecast an average mean absolute error of $1,420°C$, $1,968°C$, $2,458°C$ for 30, 45 and 60 minutes, respectively. Analyzing the results, it was found that the SVR model is an effective tool for forecasting internal temperature of a greenhouse at different times.

Keywords: Agriculture, Greenhouse, Internal Temperature, Forecast, Machine Learning

Agradecimientos

Estoy agradecido principalmente con mi asesor y tutor de tesis el Dr. Héctor Alonso Guerrero Osuna. Su guía, sus consejos y su mentoría me llevaron a través de todas las etapas de la redacción de mi tesis hasta la conclusión. Es también de mi placer agradecer a mis sinodales, el Dr. Luis Fernando Luque Vega, y el Dr. Carlos Alberto Olvera Olvera, quienes gracias a sus orientación logré mejorar la calidad y la conclusión del trabajo.

Me gustaría también dar gracias especiales a mis compañeros de maestría Ángel Wong, Elisa Vaca, Alejandra Galindo, Luis Guerrero, Alex Medina y Aldo Villareal. Además, a mis maestros que gracias a sus conocimientos compartidos y consejos me brindaron las herramientas para realizar este trabajo.

A mi madre Delta González y mis hermanos Delta Ponce y David Ponce quienes me brindaron su apoyo y entendimiento continuo mientras realizaba la investigación, experimentación y escritura de la tesis.

Estoy agradecido con la coordinadora académica la Dra. Rocío Carrasco que además de brindarme su guía, conocimiento y disponibilidad, creyó en mi y en mis capacidades. Por último, estoy totalmente agradecido con el ITESO por darme la oportunidad de continuar con mi educación en el nivel de maestría con el apoyo de una beca académica así como de permitirme ser parte de su comunidad.

Índice general

Índice de figuras

Page

Índice de tablas

Page

Dedicado a mi madre Delta Cecilia González Belmont y mis hermanos Delta Ponce y David Ponce.

1 Introducción

Contenido

La agricultura en México es uno de los sectores generadores más importantes de la demanda para la industria de invernaderos y sistemas de riego. Este sector representó el 2.9 % del PIB nacional en el 2021[1] siendo México el 11vo productor y 8vo exportador mundial, por lo que la agricultura mexicana tiene un impacto a nivel internacional.

Cultivar a cielo abierto se ha vuelto cada vez más difícil debido a las condiciones climáticas, malezas y enfermedades de cultivos, menor control del consumo de agua, luz que se encuentran expuestos los cultivos, entre otros factores. Una de las opciones para manejar estos problemas es el uso de invernaderos controlados. Los invernaderos facilitan el mantenimiento de un microclima adecuado para el crecimiento óptimo de los cultivos y plantas en su interior. Los elementos del microclima como lo son la temperatura y la humedad se deben mantener en niveles adecuados para el tipo de cultivo que se producirá en su interior lo que permitirá a la planta aprovechar las mejores condiciones e intensificar la producción.

Las tecnologías de automatización han facilitado el control del invernadero[2], el uso de estas permite la posibilidad de proporcionar las condiciones óptimas para los cultivos que se siembren en su interior. Otra de las ventajas de los invernaderos es que se puede cultivar todo el año y no solo en la etapa estacional cuando el clima fuera del invernadero lo permite[3]. Debido al auge que ha tenido la automatización de los invernaderos, el complementarlo con algoritmos de pronóstico inteligentes, resulta una estrategia competitiva al proveer

[1] INEGI. COMUNICADO DE PRENSA NÚM. 130/22, 2022. URL https://www.inegi.org.mx/contenidos/saladeprensa/boletines/2022/pib_pconst/pib_pconst2022_02.pdf. Accesado en: 2022-11-08

[2] R. Ramin Shamshiri, F. Kalantari, K. C. Ting, K. R. Thorp, I. A. Hameed, C. Weltzien, D. Ahmad, and Z. Mojgan Shad. Advances in greenhouse automation and controlled environment agriculture: A transition to plant factories and urban agriculture. *International Journal of Agricultural and Biological Engineering*, 11:1–22, 2018. DOI: https://doi.org/10.25165/j.ijabe.20181101.3210

[3] A. Abdullah, S. Al Enazi, and I. Damaj. Agrisys: A smart and ubiquitous controlled-environment agriculture system. In *2016 3rd MEC International Conference on Big Data and Smart City (ICBDSC)*, pages 1–6, 2016. DOI: 10.1109/ICBDSC.2016.7460386

datos confiables para guías y advertencias anticipadas ayudando a los agricultores en la toma de decisiones[4].

El monitoreo y control de las condiciones óptimas del invernadero influye directamente con el rendimiento de los cultivos. Cualquier nivel fuera de los adecuados, puede generar pérdidas de hortalizas, plagas y enfermedades que se propagan de forma acelerada con consecuencias devastadoras en el invernadero. Un ajuste necesario en los factores que influyen al microclima son esenciales para prevenir problemas y maximizar la productividad del invernadero.

Modelos y simulaciones como los realizados por Zheng Hui, Qin Lin-lin y Wu Gang[5] promueven la mejora continua de la predicción y, al mismo tiempo, las diversas aplicaciones de los modelos pueden proveer el desarrollo de algoritmos de control inteligentes que a su vez proporcionan condiciones favorables para el control del ambiente dentro del invernadero resultando conveniente el desarrollo de un algoritmo de pronóstico de temperatura para los invernaderos.

Este trabajo presenta un acercamiento con diferentes modelos matemáticos para pronosticar temperatura interna de un invernadero con el fin de obtener un algoritmo capaz de realizar pronósticos aceptables para ser utilizados en el invernadero de manera preventiva.

[4] A. Elanchezhian, J. K. Basak, J. Park, F. Khan, F. G. Okyere, Y. Lee, A. Bhujel, D. Lee, T. Sihalath, and H. T. Kim. Evaluating different models used for predicting the indoor microclimatic parameters of a greenhouse. *Applied Ecology and Environmental Research*, pages 2141–2161, 2020. DOI: https://doi.org/10.15666/aeer/1802_21412161

[5] Z. Hui, Q. Lin-lin, and W. Gang. Modeling and simulation of greenhouse temperature hybrid system based on armax model. In *2017 36th Chinese Control Conference (CCC)*, pages 2237–2241, 2017. DOI: 10.23919/ChiCC.2017.8027690

1.1 *Planteamiento del Problema*

Distintos problemas pueden llegar a presentarse en el proceso de producción de hortalizas bajo invernadero que van desde lo más elemental, como el desconocimiento de técnicas de cultivo[6], hasta costos elevados iniciales de establecimiento y explotación, suministro inadecuado de agua, enfermedades, etc. Algunos de estos problemas se encuentran relacionados con la temperatura y la humedad interna y externa, las cuales juegan un rol importante en el desarrollo de los cultivos dentro del invernadero[7].

La gran mayoría de los invernaderos automatizados tienen un sistema de control, actuadores y sensores que monitorean los cambios de temperatura, humedad, radiación, flujo de aire, etc. El sistema de control brinda información actual del invernadero lo que permite el uso de técnicas correctivas. Estas técnicas, usualmente, se llevan a cabo con conocimiento previo y en caso de no ser implementadas en el momento, puede llegar a generar más problemas dentro del invernadero.

Un pronóstico de temperatura oportuno, permite ajustar el sistema de calefacción, ventilación, y el suministro de fertilización carbónica, mediante actuadores instalados para tal propósito y así mantener el microclima adecuado para el cultivo. Cualquier cambio conlleva un consumo de energía, pronósticos erróneos o mal calibrados, se traducen en un mayor consumo de energía[8] y condiciones en las cuales el cultivo

[6] R. O. Wayua, V. Ochieng, V. Kirigua, and L. Wasilwa. Challenges in greenhouse crop production by smallholder farmers in kisii county, kenya. *African Journal of Agricultural Research*, 16:1411–1419, 10 2020. DOI: 10.5897/AJAR2020.15086

[7] R. Bessin, Lee. H. Townsend, R. Extension Entomologists & G. Anderson, and Extension Horticulturist University of Kentucky College of Agriculture. Greenhouse insect management, 2007. URL https://entomology.ca.uky.edu/ent60. Accesado en: 2023-01-15

[8] T. Hans-Juergen. Energy saving potential of greenhouse climate control. *Mathematics and Computers in Simulation (MATCOM)*, 48:245–251, 1998. URL https://ideas.repec.org/a/eee/matcom/v48y1998i1p93-101.html

no se desarrollará aprovechando todo su potencial.

Una carencia de control preventivo automático integrado en los sistemas del invernadero, influye en el proceso de mantener la temperatura en los niveles adecuados para cada cultivo, ya que el salirse de estos niveles podría ocasionar daños en la morfología y en los distintos procesos fisiológicos de los cultivos, como son la formación floral, quemadura de hojas, mala calidad del fruto, exceso de transpiración, acortamiento de la vida del cultivo, entre otros.

Las temperaturas extremas en el cultivo también puede afectar de manera positiva o negativa a cualquier cultivo, las bajas temperaturas, por ejemplo, dentro de un invernadero pueden afectar proporcionalmente el crecimiento y desarrollo de las plantas, esta afectación depende principalmente del tipo de cultivo. En ocasiones, los cultivos no están listos para el mercado por lo que se puede disminuir la temperatura para reducir su tasa de crecimiento por un tiempo. La cantidad de humedad que el aire contiene decrementa con la temperatura, esto es de mucha utilidad debido a que valores de humedad alta puede generar que se desarrollen patógenos en los cultivos ya que estas no se secan tan rápido. Por último, una baja temperatura dentro de un invernadero puede resultar en el colapso de tallos y hojas. Generalmente cultivos jóvenes son mucho más susceptibles a daños por baja temperatura a diferencia de los cultivos maduros, sin embargo ambos cultivos pueden generar lesiones por frío que después afectan al cultivo al ser expuestos a temperaturas altas[9].

[9] E. Runkle. The perils of low (greenhouse) temperature, Feb 2020. URL https://gpnmag.com/article/the-perils-of-low-greenhouse-temperature/. Accesado en: 2023-02-21

1.2 *Antecedentes*

Distintas técnicas han sido desarrolladas para intentar modelar el comportamiento térmico al interior de los invernaderos, utilizando el flujo y pérdida de calor, así como las cubiertas poliméricas y la densidad del aire[10,11]. Inicialmente en los casos de estudio no se contemplaba un pronóstico con variables climáticas de invernaderos, por lo que la implementación de un control en esas etapas del desarrollo de estudios era limitada a controles correctivos y no preventivos, lo cual no solía satisfacer completamente las necesidades de los productores bajo invernadero[12]. Estas técnicas se enfocan en la interpretabilidad a través del análisis cualitativo dentro de los cuales se tienen los sistemas de inferencia mediante conocimiento previo de la dinámica del sistema para determinar las predicciones sobre su comportamiento y cuentan con la ventaja de brindar una comprensión relativamente sencilla en la forma de hacer ajustes.

Albright y Scott[13] presentaron una solución mediante series exponenciales de Fourier para la variación periódica y constante de la temperatura del aire de un invernadero con paredes opacas

[10] G. Leonidopoulos. Greenhouse dimensions estimation and short time forecast of greenhouse temperature based on net heat losses through the polymeric cover. *Polymer Testing*, 19: 801–812, 2000. ISSN 0142-9418. DOI: https://doi.org/10.1016/S0142-9418(99)00050-1. URL https://www.sciencedirect.com/science/article/pii/S0142941899000501

[11] J. Leal Iga, J. Leal Iga, C. Leal Iga, and R. A. Flores. Effect of air density variations on greenhouse temperature model. *Mathematical and Computer Modelling*, 47: 855–867, 2008. ISSN 0895-7177. DOI: https://doi.org/10.1016/j.mcm.2007.05.011. URL https://www.sciencedirect.com/science/article/pii/S0895717707002130

[12] G.P.A. Bot. Physical modeling of greenhouse climate. *IFAC Proceedings Volumes*, 24:7–12, 1991. ISSN 1474-6670. DOI: https://doi.org/10.1016/B978-0-08-041273-3.50006-9. URL https://www.sciencedirect.com/science/article/pii/B9780080412733500069. IFAC/ISHS Workshop on Mathematical and Control Applications in Agriculture and Horticulture, Matsuyama, Japan, 30 Septembe

homogéneas, M.K. Selcuk[14] empleando un balance térmico en estado inestable y soluciones numéricas, presentó un análisis de invernaderos incluyendo efectos de evaporación de la humedad del suelo, la transpiración de las plantas y la condensación de la cubierta.

Algunas otras técnicas están centradas en asegurar la mayor precisión posible a través del análisis cuantitativo de los datos, López-Cruz[15] usó diversos tipos de redes neuronales y modelos autorregresivos (ARX) que otorgan resultados adecuados en cuanto al pronóstico, pero carecen de un modelo que sea representativo e interpretable y cuya dinámica no es fácil de distinguir. Litago, J., Bailey, B., y Sánchez-Girón, V.[16], desarrollaron dos modelos de series de tiempo usando cointegración y métodos de corrección de errores los cuales relativamente pueden estimar y pronosticar condiciones climáticas internas.

[14] M. K Selcuk. Use of digital computers for the heat and mass transfer of controlled environment greenhouses. *Environmental Research Laboratory, University of Arizona, Tucson, Arizona*, 1970

[15] I. L. López-Cruz, A. Rojano-Aguilar, W. Ojeda-Bustamante, and R. Salazar-Moreno. Arx models for predicting greenhouse air temperature: A methodology. *Agrociencia*, 41:181–192, 02 2007

[16] J. Litago, F. Baptista, J. Meneses, L. Navas, B. Bailey, and V. Sánchez-Girón. Statistical modelling of the microclimate in a naturally ventilated greenhouse. *Biosystems Engineering*, 92: 365–381, 2005. ISSN 1537-5110. DOI: https://doi.org/10.1016/j.biosystemseng.2005.07.015. URL https://www.sciencedirect.com/science/article/pii/S1537511005001625

1.3 *Justificación*

El pronóstico se vuelve una medida de prevención de alta relevancia en el uso de invernaderos, esto debido a la cantidad de sensores y sistemas que permiten tomar mediciones y evaluaciones adecuadas presentes dentro del microclima en cuestión de segundos. Las acciones posibles por tomar permiten el crecimiento, mantenimiento y control óptimo de los cultivos al interior de los invernaderos. Es necesario tener claridad de qué variables tienen mayor influencia en el sistema a la hora de determinar un algoritmo de aprendizaje automatizado. La automatización con un pronóstico permite entender y lidiar con datos complejos, obtener una precisión adecuada, realizar ajustes y la obtención de un modelo suficientemente robusto para pronosticar.

Las variables climáticas en la agricultura protegida son factores esenciales por considerar en el manejo de invernaderos; la temperatura a la que se expone el cultivo es una de las principales, donde su pronóstico se vuelve una medida de prevención de alta relevancia en el uso de invernaderos[17]. Mantener temperaturas adecuadas en los invernaderos es un factor muy importante para conservar la salud del cultivo, mejorar la productividad debido a la alta demanda de energía para la ambientación del microclima y llevar un control óptimo de este al interior del invernadero. Las heladas principalmente significan un gran reto tecnológico si se quiere asegurar la sostenibilidad del cultivo, ya que una reacción nula o tardía, podría significar la pérdida de la producción en menor a mayor escala.

[17] W. O. Baudoin. *Good agricultural practices for greenhouse vegetable crops principles for Mediterranean climate areas*, pages 21,22. Food and Agricultural Organization of the United Nations (FAO), 2013

Un pronóstico de temperatura interna provee información relevante para la toma de decisiones y acciones para proporcionar un ambiente estable para las hortalizas. El pronóstico puede considerar diferentes intervalos de tiempo y debe de ser apropiado para cada tipo de cultivo. Este puede ser obtenido con el uso de modelos matemáticos simples

o complejos desarrollados y entrenados con una gran cantidad de datos obtenidos en campo para mantener coherencia de los pronósticos con el ambiente en el que son adquiridos. Los ajustes posibles para realizarse una vez se que se obtiene el pronóstico usualmente son de manera anticipada ajustando los sistemas integrados del sistema de invernadero.

1.4 Objetivos

1.4.1 General

Realizar el análisis y procesamiento de datos climáticos de un ambiente controlado de invernadero para la obtención de un modelo capaz de comprender la estructura de los datos con el fin de realizar un pronóstico adecuado de la temperatura al interior de dicho invernadero.

1.4.2 Objetivos específicos

- Realizar análisis de datos, correlación de variables para así determinar significancia y relación entre las mismas.
- Aplicar distintos algoritmos de aprendizaje supervisado para entender el comportamiento de los datos con temperatura actual.
- Realizar pronósticos con modelos optimizados con mejor entendimiento del comportamiento de la temperatura interna.
- Analizar resultados con el propósito de elegir el mejor modelo y presentar conclusiones.

1.5 Hipótesis

Realizando un análisis de datos, modelando el comportamiento a través del tiempo y comprendiendo sus limitaciones en cuestión de dimensiones, un algoritmo de aprendizaje automatizado es capaz de generar un pronóstico adecuado de temperatura interna en un invernadero apto para proporcionar información a futuro acertada a fin de ser procesada en un control preventivo y realizar ajustes anticipados a diferencia de un control correctivo.

2 *Estado del Arte*

Un invernadero es un tipo de estructura donde se mantienen condiciones ambientales adecuadas como la temperatura, humedad, y otros factores para el crecimiento de un cultivo. El origen del invernadero se puede remontar a la época romana en el siglo XV, estos tenían forma de camas móviles donde se plantaban cultivos por orden del emperador romano Tiberio Julio César Augusto (Figura 2.1). Los invernaderos estaban fabricados con láminas de mica y alabastro utilizados regularmente en momentos donde el clima no era apropiado para el cultivo y gracias a su movilidad se trasladaban hacia el interior o exterior y luego en inverno, se resguardaban bajo una cubierta de láminas de piedra transparente y de mica.

Figura 2.1: Winter-plaats in den Hoff van d'Academie Tot Leyden, 1676. *Imagen cortesía de Princeton Architectural Press obtenida de https://www.archdaily.mx*

En el siglo XVI se construían invernaderos con materiales poco especializados como la madera, bambú, paneles de vidrio y papeles aceitados, su fin era el suplir la necesidad de proteger los cultivos. En el siglo XVII comenzaron a utilizar los invernaderos en Europa conocidos como los *orangeries*, en donde se cultivaban naranjos. Estas nuevas estructuras tenían ventanas amplias de cristal en su lado sur para permitir la entrada de la luz solar. Cuando no se contaba con luz

solar en el invernadero, se colocaban estufas alrededor de este para calentarlo. Conforme pasaron los años y con la expansión del uso de paneles de cristal para invernaderos, el recinto fue modificado con el fin de poder aprovechar mucha más la luz solar durante el día[1].

A partir de 1737 se empezó a tener un gran interés por conservar energía en los invernaderos con el uso de cortinas y se comenzó a cultivar con aportes de CO_2, calor y nutrientes procedentes del estiércol. A mediados del siglo XIX la demanda de invernaderos incrementó debido a la búsqueda de adaptar las necesidades de plantas exóticas[2], por lo que en el siglo XX los invernaderos fueron recibiendo mejoras progresivas y se desarrolló una amplia información sobre calefacción, riego y fertilización. En Países Bajos se desarrollaron gradualmente invernaderos con mejor uso de los cultivos y fue en 1937 cuando se construyó el invernadero Venlo a base de acero y cristal que podía ser utilizado para diferentes cultivos.

Fue entonces que a partir de la Primera Guerra Mundial se inventaron los plásticos y con ellos su uso en la agricultura con sus variaciones de polietileno, poliestireno o PVC. El uso del plástico en los invernaderos provocó una masificación de estos donde al ser mucho más efectivos que el vidrio, permitían tener ventajas como un mejor manejo del clima minimizando efectos de factores abióticos y bióticos[3] así como el cultivo de más productos agrícolas en países donde se tienen climas muy fríos o muy calientes.

La estructura básica de los invernaderos actuales tienen forma de *A*. Estos recintos tienen diferentes estilos dependiendo de su funcionalidad y objetivo, entre los más destacables se encuentran los invernaderos con marcos rígidos (*a*), barraca quonset o túnel (*b*), marco-A (*c*), los estilos góticos (*d*), y los de poste y viga (*e*) (Figura 2.2).

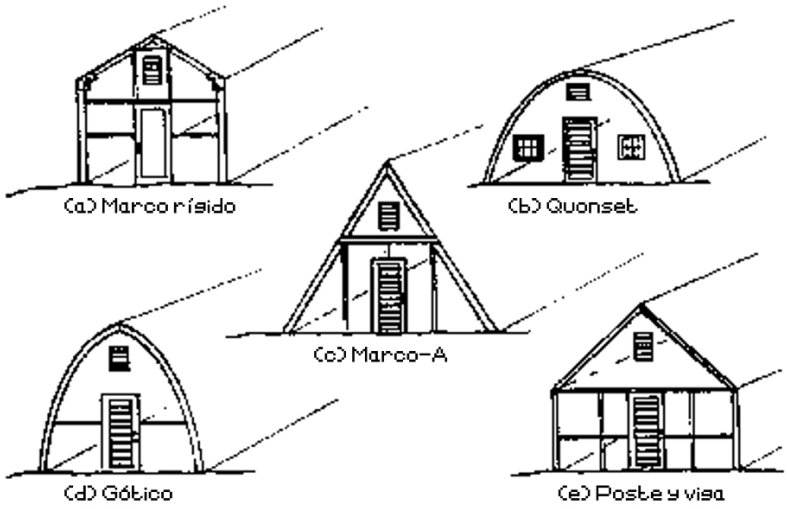

Figura 2.2: Estructuras básicas de invernaderos. *Basado en: Principle of greenhouse structures construction, John Ilu [2011]*

A veces se combinan dos o más tipos de invernaderos para tener

[1] J. Pérez-Parra and J. C. López-Hernández. Evolución de las estructuras de invernadero, 2007. URL https://www.publicacionescajamar.es/publicacionescajamar/public/pdf/series-tematicas/centros-experimentales-las-palmerillas/evolucion-de-las-estructuras.pdf. Accesado en: 2023-03-21

[2] Nightingale Garden Company Limited. Exotic plants in nineteenth century gardens, 2007. URL https://www.gardenvisit.com/history_theory/library_online_ebooks/ml_gothein_history_garden_art_design/exotic_plants_planting. Accesado en: 2023-01-30

[3] Firman D.M. and E.J. Allen. Chapter 33 - agronomic practices. In D. Vreugdenhil, J. Bradshaw, C. Gebhardt, F. Govers, D. Mackerron, M. Taylor, and H. Ross, editors, *Potato Biology and Biotechnology*, pages 719–738. Elsevier Science B.V., 2007. ISBN 978-0-444-51018-1. DOI: https://doi.org/10.1016/B978-044451018-1/50075-0. URL https://www.sciencedirect.com/science/article/pii/B9780444510181500750

menos paredes exteriores lo cual aumenta su coste de calefacción pero reduce el costo de construcción. Entre las combinaciones más comunes de invernaderos se encuentran capilla (*a*), capilla modificada (*b*), multitúnel (*c*), los tipo Venlo (origen Holandés) (*d*), entre otros (Figura 2.3).

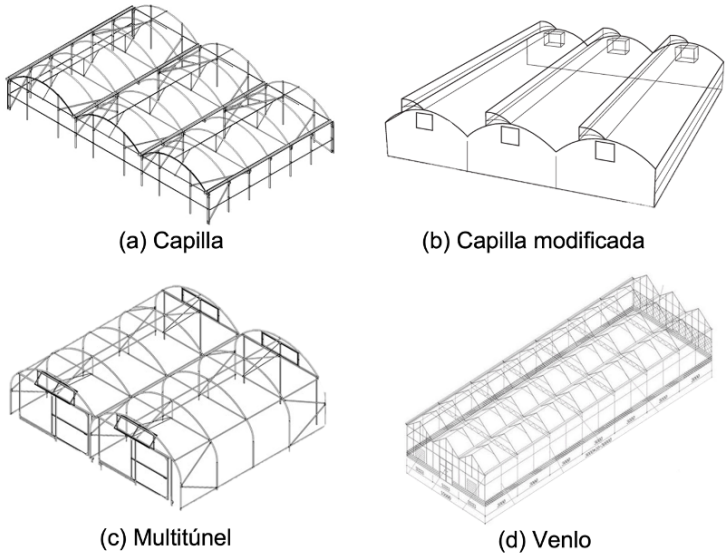

Figura 2.3: Tipos de invernaderos múltiples.

El plástico, como polietileno o polivinilo, y la fibra de vidrio son utilizados como materiales tradicionales de construcción en un invernadero, mientras que el aluminio, acero galvanizado, cedro o ciprés son usados para mantener la estructura firme[4,5]. La superficie acristalada en las paredes laterales y el techo permite que los cultivos estén expuestos a luz natural durante gran parte del día.

La ubicación y orientación del invernadero son de suma importancia. Es necesario considerar un lugar donde se tenga mayor luz solar, o en caso de tener un cultivo que no requiera mucha exposición a radiación solar, que se encuentre cerca de árboles de sombra; sin embargo algunos invernaderos incluyen un control de malla sombra sobre el techo para accionar cuando no se requiera de exposición continua a la luz solar. V.P. Sethi[6] realizó un estudio sobre la orientación y ubicación de invernaderos, donde concluyó que el patrón y la cantidad de radiación solar disponible en diferentes latitudes puede ser distinto para los mismos tipos de invernaderos. A una latitud de 10°N, cualquier tipo recibe más radiación solar en invierno pero menos en verano. Mientras que a 31°N reciben menos radiación solar en invierno pero más en verano, por lo tanto su orientación debe de ser de norte a sur debido a que el ángulo del sol es mucho mayor[7]. Estas diferencias aumentan aún más en la latitud 50°N y la radiación solar recibida en los meses de invierno es mucho menor en comparación con la cantidad recibida en

[4] The Editors of Encyclopedia Britannica. greenhouse, 2019. URL https://www.britannica.com/topic/greenhouse. Accesado en: 2023-01-30

[5] Shakuntala Pandey and Anil Pandey. Greenhouse technology. *International Journal of Research -GRANTHAALAYAH*, page 1–3, 2015. DOI: 10.29121/granthaalayah.v3.i9se.2015.3176

[6] V.P. Sethi. On the selection of shape and orientation of a greenhouse: Thermal modeling and experimental validation. *Solar Energy*, 83:21–38, 2009. DOI: 10.1016/j.solener.2008.05.018

[7] I. Jaisankar, A. Velmurugan, and C. Sivaperuman. Chapter 19 - biodiversity conservation: Issues and strategies for the tropical islands. In C. Sivaperuman, A. Velmurugan, I. Jaisankar, and A. K. Singh, editors, *Biodiversity and Climate Change Adaptation in Tropical Islands*, pages 525–552. Academic Press, 2018. ISBN 978-0-12-813064-3. DOI: https://doi.org/10.1016/B978-0-12-813064-3.00019-3. URL https://www.sciencedirect.com/science/article/pii/B9780128130643000193

los meses de verano, estos invernaderos deben tener orientación que corra de este a oeste para permitir luz solar de ángulo bajo que entre desde los dos lados.

Los diferentes tipos de invernaderos reciben diferente exposición solar es por eso que a 10°N se debería de preferir uno de tipo quonset o túnel, ya que recibe una cantidad mínima de radiación solar durante todo el año. En latitud 31°N se debe preferir uno de tipo capilla o capilla modificada ya que permite recibir más radiación en inverno pero menos en verano. A una latitud de 50°N, es preferible el tipo de Venlo ya que recibe la mayor radiación solar durante todo el año. Por último, independientemente del tipo, la orientación este-oeste es la más adecuada para aplicaciones de invernaderos durante todo el año en todas las latitudes, ya que esta orientación recibe más radiación en invierno pero menos en verano a excepción del ecuador.

Además de la ubicación y orientación cuando se construye un invernadero, se consideran las fuentes de calor, agua y electricidad, así como la protección del viento, materiales a utilizar, el cultivo que se crecerá dentro del recinto[8]. Es necesario considerar que el piso donde se estará plantando sea adecuado para el cultivo, se necesitan mínimo una hectárea para las instalaciones, la zona de cultivo exterior, el acceso, el aparcamiento y las zonas de amortiguación, en caso de requerirlo, un permiso debido a que las estructuras pueden solo estar ubicadas en ciertos espacios. En ocasiones, también es necesario considerar la ubicación con respecto a las autopistas, para un comercio minorista, una ubicación de carretera con mucho tráfico o cerca de una gran zona residencial puede aumentar el negocio. Para empresas mayoristas, el acceso a autopistas interestatales es deseable para gestionar el tráfico de camiones pesados. Por último, y no menos importante, la mano de obra debe de considerarse, ya que conforme el invernadero y el negocio crece se puede necesitar mano de obra adicional, de tiempo completo o parcial.

Considerando la importancia de la ubicación y orientación del invernadero, este se calienta, en parte, con la radiación recibida por los rayos solares. Cada cultivo requiere un ambiente diferente, para algunos es necesario mantener temperaturas entre ciertos rangos para su buen crecimiento, sin embargo en algunas ocasiones los rayos de luz no son suficientes para calentar el invernadero, por lo tanto se utilizan métodos artificiales[9,10] como vapor o por circulación de agua caliente o aire caliente a través de ductos de metal. Las ventanas de un invernadero son importantes para el flujo de aire, no contar con un mecanismo de ventilación puede llegar a desestabilizar el ambiente interno, por lo que también se necesita algún tipo de sistema de aireación eficiente que suele consistir en ventanas mecánicas o automáticas en el techo y ventanales o ventilas en los extremos de las

[8] J. W. Bartok Jr. Selecting and building a commercial greenhouse, Apr 2017. URL https://ag.umass.edu/greenhouse-floriculture/fact-sheets/selecting-building-commercial-greenhouse. Accesado en: 2023-03-21

[9] J. Zhang, S. Zhao, A. Dai, P. Wang, Z. Liu, B. Liang, and T. Ding. Greenhouse natural ventilation models: How do we develop with chinese greenhouses? *Agronomy*, 12, 2022. ISSN 2073-4395. DOI: 10.3390/agronomy12091995. URL https://www.mdpi.com/2073-4395/12/9/1995

[10] A. Ganguly and S. Ghosh. A review of ventilation and cooling technologies in agricultural greenhouse application. *Iranian (Iranica) Journal of Energy & Environment*, 2011. ISSN 2079-2115. URL https://www.ijee.net/article_64325.html

paredes. Los invernaderos más tecnificados se apoyan con ventiladores eléctricos que hacen circular el aire (Figura 2.4), también se puede tener un sistema de maya sombra la cual es una técnica de refrigeración basada en la disminución del porcentaje de radiación fotoactiva, en donde se puede tener dos tipos de sistemas principales; el estático que es utilizado para el encalado del invernadero y mallas de sombreo, y los dinámicos que permiten el control de cortinas móviles y riego de la cubierta.

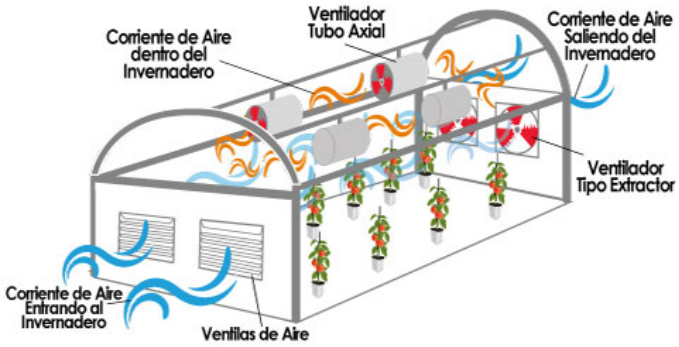

Figura 2.4: Climatización de un invernadero. *Imagen cortesía de Ventilación en un invernadero, Hydro Environment [2011]*

Es común tener un sistema de monitoreo con sensores para un invernadero, la información recolectada del exterior así como dentro del mismo es de suma importancia para su control óptimo. Usualmente dentro del invernadero se tienen sensores para medición del dióxido de carbono (CO_2), humedad relativa, radiación solar, déficit de la presión de vapor, y temperatura; mientras que algunos parámetros externos como la velocidad y dirección del viento, humedad y lluvia son medidos continuamente. La mayor ventaja de utilizar un invernadero para cultivar es la capacidad de proporcionar una temperatura deseable para el crecimiento y desarrollo de la planta, por lo que el medir y controlar la temperatura del aire es muy habitual en este tipo de sistemas porque es la que más influye en la temperatura del cultivo. Midiendo la temperatura a la que están expuestas la plantas y del ambiente se puede determinar si estas necesitan un ajuste ya que algunas requieren más calor o más frío. La temperatura se puede sensar con el uso de un termostato, el cual, en conjunto con una bomba de circulación de agua o un ventilador extractor, es capaz de controlar la temperatura dentro del invernadero. El control de temperatura de un invernadero depende del tipo de invernadero, su ubicación, tamaño del invernadero, la energía que se desee utilizar, el cultivo, la fase del cultivo, entre otros[11]. La refrigeración por evaporación es uno de los métodos utilizados que se realiza mediante el suministro mecánico de gotas al aire del invernadero. La evaporación de estas gotas convierte el calor sensible en calor latente y por ende enfría el invernadero. Existen tres principales métodos de refrigeración prácticos de bajo

[11] K. Nemali. Temperature control in greenhouses. *Horticulture and Landscape Architecture*, Feb 2021. URL https://www.extension.purdue.edu/extmedia/HO/HO-327-W.pdf

costo, aspersión, almohadilla y ventilador y nebulización fina[12]. La aspersión consiste en rociar grandes gotas de agua sobre la cubierta vegetal o el techo. El sistema de almohadilla y ventilador consiste en introducir aire exterior en el invernadero a través de una almohadilla húmeda colocada en las paredes del invernadero.

Los métodos descritos anteriormente son usualmente los más tradicionales pero tienen algunas desventajas[13]. El sistema de almohadilla y ventilador tiene la desventaja de gasto de agua y falta de homogeneidad mientras que un mal control del sistema de aspersión puede llevar a enfermedades por hongos. Por otro lado, se encuentra un método usando neblina, en donde finas gotas de agua son mecánicamente servidas a través de boquillas de alta presión en el interior del invernadero[14]. Estas gotas proporcionan una gran superficie de contacto con el aire suministrado a través de las ventanas del techo y bombeado al interior por ventiladores instalados a lo largo de las paredes laterales en el invernadero. La falta de estrategia de control robusta, eficacia, y de alto rendimiento hace el sistema de neblina menos popular que el de aspersión y almohadilla y ventilador. Agmail, W. I. R., Linker, R. y Arbel, A.[15], presentaron un control lineal de múltiples entradas y múltiples salidas (MIMO, por sus siglas en inglés) con el uso de neblina para la estabilización de temperatura y humedad de un invernadero el cual obtuvo resultados estables de $\pm 2°C$ y $\pm 10\,\%$ para temperatura y humedad respectivamente.

Considerando la importancia del control de la temperatura dentro de un invernadero, es necesario tener un pronóstico acertado además de un monitoreo climático y sistema de ventilación en el invernadero para el crecimiento saludable de los cultivos. Distintos métodos se han desarrollado a lo largo de los años para esto; Kittas, Bartzanas y Jaffrin[16] desarrollaron un modelo climático sencillo incorporando efectos de ventilación, sombreado del techo y transpiración de los cultivos. Su modelo se calibró a partir de mediciones realizadas en un invernadero comercial equipado con ventiladores y paneles evaporativos. Los datos mostraron que el sistema de refrigeración era capaz de mantener la temperatura del aire del invernadero controlado a niveles bajos, sin embargo, debido al tamaño del invernadero (50m ancho x 60m largo), se observaron diferencias de temperatura $\pm 8°C$ entre la almohadilla y los ventiladores. La calibración de este modelo consideró temperaturas del centro del invernadero y los extremos, por lo que después se logró (a pesar de la simplicidad) mejorar el diseño y gestión del sistema de refrigeración.

Teitel M., Atias M. y Barak M.[17] realizaron varios modelos para calcular los cambios de temperatura, proporción de humedad y concentración de CO_2 del aire en la dirección del flujo de aire predominantes a lo largo de un invernadero aireado por ventiladores.

[12] E. Reyes. Sistemas de climatización en invernaderos, 2014. URL https://www.mundohvacr.com.mx/2014/05/sistemas-de-climatizacion-en-invernaderos/. Accesado en: 2023-01-30

[13] C. Duarte-Galvan, I. Torres-Pacheco, R. G. Guevara-Gonzalez, R. J. Romero-Troncoso, L. M. Contreras-Medina, M. A. Rios-Alcaraz, and J. R. Millan-Almaraz. Review. advantages and disadvantages of control theories applied in greenhouse climate control systems. *Spanish Journal of Agricultural Research*, 10:926–938, Oct 2012. DOI: 10.5424/sjar/2012104-487-11. URL https://revistas.inia.es/index.php/sjar/article/view/2196

[14] A. Arbel, O. Yekutieli, and M. Barak. Performance of a fog system for cooling greenhouses. *Journal of Agricultural Engineering Research*, 72:129–136, 1999. DOI: 10.1006/jaer.1998.0351

[15] W. I. R. Agmail, R. Linker, and A. Arbel. Robust control of greenhouse temperature and humidity. *IFAC Proceedings Volumes*, 42:138–143, 2009. ISSN 1474-6670. DOI: https://doi.org/10.3182/20090616-3-IL-2002.00024. URL https://www.sciencedirect.com/science/article/pii/S147466701540391X

[16] C. Kittas, T. Bartzanas, and A. Jaffrin. Temperature gradients in a partially shaded large greenhouse equipped with evaporative cooling pads. *Biosystems Engineering*, 85:87–94, 2003. ISSN 1537-5110. DOI: https://doi.org/10.1016/S1537-5110(03)00018-7. URL https://www.sciencedirect.com/science/article/pii/S1537511003000187

[17] M. Teitel, M. Atias, and M. Barak. Gradients of temperature, humidity and co2 along a fan-ventilated greenhouse. *Biosystems Engineering*, 106:166–174, 2010. ISSN 1537-5110. DOI: https://doi.org/10.1016/j.biosystemseng.2010.03.007. URL https://www.sciencedirect.com/science/article/pii/S1537511010000590

Ellos tomaron en cuenta los gradientes de temperatura, proporción de la humedad y concentración de CO_2 que se desarrollaron a lo largo de la dirección del flujo de aire. Sus modelos se verificaron mediante mediciones experimentales en un invernadero sometido a una tasa de ventilación constante. Sus resultados mostraron que los gradientes de la temperatura cambian a lo largo del día, donde son más pronunciados hacia el mediodía y cuando la radiación solar es mayor.

Otro método por Hesham A. Ahmed, et al.[18], realizaron una comparación entre dos invernaderos. Uno de los invernaderos tenía sombreado exterior utilizando una red móvil de plástico negro con 30 % de transmisividad mientras que el otro invernadero se mantuvo sin sombrear; en ambos invernaderos se cultivaron fresas. Los resultados mostraron que la distribución espacial de la temperatura interna y la humedad interna se veían afectadas significativamente por la radiación solar exterior y el funcionamiento de la refrigeración evaporativa. Junto con un análisis de regresión se encontró que cuando la intensidad de la radiación solar aumentaba de 200 a 800 W m^{-2}, el estaño aumentaba 4.5°C en el invernadero sombreado y 2°C en el invernadero sin sombra, mientras que la humedad interna disminuía un 15 % en el invernadero sombreado y un 5 % en el invernadero sin sombra. El sombreado exterior mejoró la distribución espacial de la temperatura y la humedad y mejoró la eficacia de refrigeración del sistema de enfriamiento evaporativo en un 12 %, ya que la radiación solar transmitida y la energía térmica acumulada en el invernadero se redujeron significativamente.

Los diferentes métodos presentados anteriormente utilizaron características del invernadero para determinar y controlar la temperatura[19]. Usualmente cualquiera de estos se presentan cuando existe algún cambio climático drástico de temperatura, como mayor radiación solar, algún corte de energía, enfermedad de cultivos, entre otros; por lo que su uso es de corrección del ambiente interno del invernadero. Una alternativa a los métodos descritos anteriormente son los preventivos, los cuales tienen como objetivo principal el ajuste de los controles internos del invernadero para la previsión de algún evento no contemplado. Para este estudio, se investigaron diferentes métodos preventivos para el pronóstico de temperatura que se han realizado.

Uno de los métodos para el pronóstico de temperatura fue realizado por Kazuhisa Ito y Tsubasa Tabei[20] que obtuvieron un modelo matemático describiendo la dinámica de dos salidas de un calefactor, humidificador y ventilador. Su modelo fue comparado con otras dos estrategias simples, un control de umbrales y un control de modulación por ancho de impulsos (PWM, por sus siglas en inglés). El modelo propuesto generó tres señales de entrada que minimizó la suma de los cuadrados de los errores de déficit de temperatura y humedad y

[18] H. A. Ahmed, Y. X. TONG, Q. C. YANG, A. A. Al-Faraj, and A. M. Abdel-Ghany. Spatial distribution of air temperature and relative humidity in the greenhouse as affected by external shading in arid climates. *Journal of Integrative Agriculture*, 18: 2869–2882, 2019. ISSN 2095-3119. DOI: https://doi.org/10.1016/S2095-3119(19)62598-0. URL https://www.sciencedirect.com/science/article/pii/S2095311919625980

[19] A. J. Udink ten Cate. *Modeling and (adaptive) control of greenhouse climates*. PhD thesis, Technische Hogeschool Twente, 1983

[20] K. Ito and T. Tabei. Model predictive temperature and humidity control of greenhouse with ventilation. *Procedia Computer Science*, 192: 212–221, 2021. ISSN 1877-0509. DOI: https://doi.org/10.1016/j.procs.2021.08.022. URL https://www.sciencedirect.com/science/article/pii/S187705092101509X

llevó el déficit de la temperatura y humedad a rangos deseados con gran precisión. El modelo matemático obtuvo resultados favorables en raíz del error cuadrático medio (RMSE, por sus siglas en inglés) con un decremento del 78 % y 87 % de decremento a comparación de ambos modelos y un decremento de error medio absoluto (MAE, por sus siglas en inglés) del 79 % y 86 %.

Weidong Zou, et al.[21], utilizaron un método de aprendizaje automatizado conocido como máquina convexa bidireccional de aprendizaje extremo (CB-ELM, por sus siglas en inglés) el cual está basado en un método convexo de optimización para predecir temperatura y humedad. El método recalcula los pesos de las salidas de los nodos existentes después de otro nodo oculto es agregado. La ventaja de este algoritmo es la velocidad de convergencia y el tamaño de la red manteniendo la simplicidad y eficiencia de una máquina de aprendizaje extremo incremental (IELM, por sus siglas en inglés). Su algoritmo fue comparado con una red neuronal de retropropagación (BPNN, por sus siglas en inglés), una función de base radial (RBF, por sus siglas en inglés), una máquina de vector de soporte (SVM, por sus siglas en inglés) y máquina convexa de aprendizaje extremo (B-ELM, por sus siglas en inglés) donde la suma de los cuadrados de los errores para las predicciones de temperatura se redujeron en 1.27°C, 2.56°C, 1.33°C, 1.26°C y 2.26 %, 3.96 % 2.05 %, 2.32 % para los errores de humedad por cada algoritmo y la validez del modelo mejoró de 0.06, 0.15, 0.08, 0.05 a 0.12, 0.24, 0.1, 0.06 respectivamente con optimización.

La regresión lineal es uno de los algoritmos más utilizados para la predicción de valores reales, sin embargo carece de complejidad al modelar datos no lineales en N-dimensiones, una de las alternativas utilizadas en modelos de pronóstico de temperatura de un invernadero es SVM. El trabajo de Dingcheng Wang, Maohua Wang y Xiaojun Qiao[22] buscó una alternativa a las redes neuronales y estudios presentados anteriormente con el uso de máquina de vector de soporte para regresión (SVR, por sus siglas en inglés). Su modelo utilizó la luz solar del exterior, la capacidad del aire termal, la pérdida de calor del aire del invernadero al aire exterior, la temperatura externa del invernadero y la energía utilizada para calentar el invernadero. Para la definición del kernel se realizó validación cruzada sin embargo resultó ser muy lento el proceso por lo que también se utilizó una regresión en línea por mínimos cuadrados con máquinas de vectores soporte (OS-LSSVMR, por sus siglas en inglés). Los resultados fueron muy buenos con un error error cuadrático medio (MSE, por sus siglas en inglés) de 0.11. El tiempo de ejecución también fue tomado en consideración con tres variaciones de SVM; OSLSSVMR, regresión codiciosa en línea de vectores de apoyo dispersos (SOG-SVR, por sus siglas en inglés) y regresión secuencial por vectores de soporte (Seq-SVMR, por sus

[21] W. Zou, F. Yao, B. Zhang, C. He, and Z. Guan. Verification and predicting temperature and humidity in a solar greenhouse based on convex bidirectional extreme learning machine algorithm. *Neurocomputing*, 249:72–85, 2017. ISSN 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2017.03.023. URL https://www.sciencedirect.com/science/article/pii/S0925231217305180

[22] D. Wang, M. Wang, and X. Qiao. Support vector machines regression and modeling of greenhouse environment. *Computers and Electronics in Agriculture*, 66:46–52, 2009. ISSN 0168-1699. DOI: https://doi.org/10.1016/j.compag.2008.12.004. URL https://www.sciencedirect.com/science/article/pii/S0168169908002305

siglas en inglés) fueron utilizados y comparados. Los tres tuvieron el mismo MSE pero el diferente tiempo de ejecución favoreció al algoritmo OS-LSSVMR.

Kalavathi Devi Thangavel, et al.[23], utilizaron un método diferente utilizando SVM. En su trabajo se utilizó lógica difusa (*fuzzy logic* en inglés) junto con un modelo de SVM para clasificación con kernel radial para obtener clases para la temperatura, humedad e hidratación. Determinaron clases para los diferentes rangos y utilizando lógica difusa la cual a diferencia de una lógica binaria, obtuvieron clases intermedias; desde completamente negativo hasta completamente positivo. En relación a los valores de porcentaje medio de error para resultados, el porcentaje de error de la temperatura fue de 11.25, hidratación de 5.29 y humedad de 6.75.

Además del poder predictivo que un algoritmo puede brindar, uno de los beneficios es el anticipo del costo por uso de los controles para mantener estable el ambiente interno del invernadero. Ouazzani Chahidi, L., et al.[24] realizaron un estudio con el uso de cuatro modelos; Redes Neuronales (ANN, por sus siglas en inglés), Regresión de procesos gaussianos (GPR, por sus siglas en inglés), SVM y potenciación de gradiente (GB, por sus siglas en inglés). El objetivo principal fue medir la energía utilizada por el sistema de enfriamiento considerando el resultado del modelo en un plazo de 6 meses. Los cuatro modelos tuvieron buenos resultados con $10\,\% < \mathrm{n}RMSE < 30\,\%$, para los 6 meses. A mediados de agosto e inicios de septiembre los modelos obtuvieron mejores resultados debido a que su predictor principal fue la radiación solar y la temperatura externa del invernadero mientras que en marzo no fueron tan significativos para la predicción del modelo. Los dos mejores modelos fueron ANN y GPR debido a que la base de datos utilizada fue pequeña, mientras que el peor fue GB por el sobre ajuste.

[23] K. Devi Thangavel, U. Seerengasamy, S. Palaniappan, and R. Sekar. Prediction of factors for controlling of green house farming with fuzzy based multiclass support vector machine. *Alexandria Engineering Journal*, 62: 279–289, 2023. ISSN 1110-0168. DOI: https://doi.org/10.1016/j.aej.2022.07.016. URL https://www.sciencedirect.com/science/article/pii/S1110016822004732

[24] L. Ouazzani Chahidi, M. Fossa, A. Priarone, and A. Mechaqrane. Evaluation of supervised learning models in predicting greenhouse energy demand and production for intelligent and sustainable operations. *Energies*, 14, 2021. ISSN 1996-1073. DOI: 10.3390/en14196297. URL https://www.mdpi.com/1996-1073/14/19/6297

3 Metodología

Contenido

En el análisis presentado se utilizaron distintos métodos para la estimación y pronóstico de la temperatura interna del invernadero controlado. Inicialmente se describió el origen y preprocesado de los datos necesarios, posteriormente se utilizaron métodos como la regresión lineal, la regresión por mínimos cuadrados parciales, una máquina de vector de soporte para regresión y el algoritmo de aumento extremo del gradiente para describir el comportamiento de los datos y por último se compararon los modelos con el fin de determinar el mejor para su uso del pronóstico de la temperatura interna del invernadero.

3.1 Obtención de datos

Los datos utilizados para la estimación de la temperatura interna fueron obtenidos por una estación meteorológica de un invernadero con techumbre curva (165m^2 de área, 27.5m de largo, 6m de ancho) localizado en la Mezquitera Sur, Juchipila, Zacatecas, México con latitud y longitud (21.42624033959812, -103.10935313358475) y con orientación 21°25'34.5"N 103°06'33.8"W. Este tipo de invernadero con techumbre curva es de uso tradicional con ningún tipo de control climático en su interior y tiene ventilación natural.

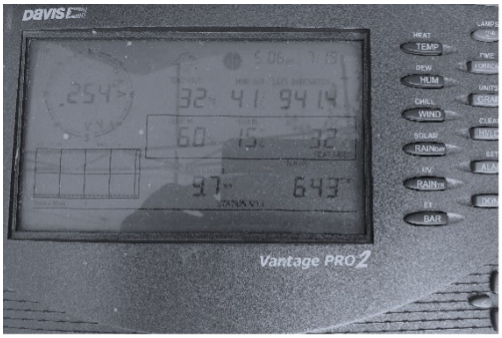

Figura 3.1: Sistema central meteorológico Davis Vantage Pro 2

El invernadero fue equipado con nueve diferentes sensores dentro y fuera del mismo. Todos éstos son parte del sistema central meteorológico Davis Vantage Pro 2 (Figura 3.1); por fuera se encuentran los sensores de temperatura, humedad, radiación solar, presión barométrica, lluvia en milímetros, y velocidad y dirección del viento (Figura 3.2), mientras que por dentro se encuentran los sensores de humedad y temperatura; en el centro del invernadero considerado como un sistema homogéneo. A partir de los sensores se obtuvieron las variables principales para la estimación; Humedad Relativa Externa, Radiación Solar, Temperatura Exterior, Temperatura Interna, Rocío Interno, Humedad Relativa Interna.

Figura 3.2: Estación Climatológica Davis Vantage Pro 2, colocado fuera del invernadero

La toma de los datos se realizó del 12 de julio del 2020 al 24 de junio del 2021 con un muestreo en intervalos de 5 minutos. Cada toma de datos incluye variables principales las cuales son parámetros de variables climáticos externos e internos. Se obtuvieron un total de 85,989 muestras para el entrenamiento y prueba de los modelos de predicción. Durante el tiempo de toma de datos se cultivó una verdura en el invernadero el cual fue el chile morrón desde febrero 2020 hasta el fin de los datos. La temperatura participa en distintas funciones de la hortaliza como el cierre o apertura de estomas, transpiración y respiración de la planta. Es por esto que por debajo o por encima de la

temperatura ideal u óptima del cultivo este no consigue desarrollarse adecuadamente. Para el chile morrón se considera temperatura ideal entre $18°C$ y $26°C$ y temperatura óptima entre $20°C$ y $25°C$.

3.2 *Preprocesamiento de datos*

Las variables para la estimación son (Tabla 3.1):

Nomenclatura	Tipo de variable climática	Unidad de medida
Temp_In	Temperatura Interna	°C
Temp_Out	Temperatura Externa	°C
Hum_Out	Humedad Externa	%
Hum_In	Humedad Interna	%
Dew_In	Punto de Rocío	%
Solar_Rad	Radiación Solar	W/m^2

Tabla 3.1: Variables climáticas de la base de datos

Los datos se encuentran divididos en las cuatro estaciones del año, cada estación tiene diferente tendencia y número de muestras por lo que se realizó un análisis descriptivo (Tabla 3.2) con respecto a la temperatura interna para poder identificar las diferencias. Para los datos en verano y primavera, la media es más alta a comparación de la media para los datos en otoño en invierno, para primavera y verano se tiene 29.57 y 31.00 respectivamente, mientras que para otoño e invierno la media es 25.10 y 23.73 respectivamente.

Estación	Conteo	Media	Desviación Std.	Mínimo	Máximo
Primavera	28684	29.57	15.03	2.60	66.49
Verano	14508	31.00	13.31	16.20	71.00
Otoño	24870	25.10	13.62	6.20	63.20
Invierno	17927	23.73	16.41	-0.61	67.70

Tabla 3.2: Análisis descriptivo de la temperatura interna por estación del año

La división de los datos por temperatura permite un mejor análisis y modelado para la temperatura interna[1]. Al utilizar todos los datos para modelar el comportamiento y generar un pronóstico de la temperatura interna es muy difícil capturar la tendencia completa y su correlación con las demás variables. Es por eso que se vuelve necesario entender el origen de los datos para así poder iniciar con candidatos de modelos viables para el pronóstico.

El ciclo de la ciencia de datos (Figura 3.3) inicia desde la comprensión del negocio, se busca comprender qué se busca resolver mediante el uso de datos y modelos de aprendizaje automático y cómo se estarán implementando en el ambiente, es por esto que el entendimiento de donde se llevará a cabo este proyecto es importante. En este análisis, se conoció la ubicación, los cultivos, la posición de los sensores dentro y fuera del invernadero, temporadas de cultivos, mejores prácticas, así como las posibles ventajas de la implementación de

[1] Y. Shao, Q.J. Wang, A. Schepen, and D. Ryu. Going with the trend: Forecasting seasonal climate conditions under climate change. *Monthly Weather Review*, 149:2513–2522, 2021. DOI: https://doi.org/10.1175/MWR-D-20-0318.1. URL https://journals.ametsoc.org/view/journals/mwre/149/8/MWR-D-20-0318.1.xml

un sistema de pronóstico de temperatura interna. Posteriormente, la limpieza y exploración se vuelve un proceso vital y necesario para la consideración de los diferentes modelos candidatos a usar para analizar el comportamiento de la temperatura interna. Las series de tiempo presentaron discontinuidad en diferentes estaciones; verano del 21 de junio 2020 hasta el 12 de julio 2020 e invierno del 21 de diciembre 2020 hasta el 3 de enero del 2021 y del 8 de marzo 2021 hasta el 21 de marzo del 2021.

Considerando el ciclo de la ciencia de datos, es importante destacar que esto se vuelve un proceso iterativo debido a que el modelado así como la comprensión de datos pueden ser alternados conforme pasa el ciclo. Un buen modelado permite comprender mejor los datos para seleccionar el mejor modelo que explique la temperatura interna, con el fin de llevar a cabo esto, se realizaron diferentes modelos para predecir el comportamiento de los datos de las series de tiempo así como para analizar de mejor manera los datos a través del tiempo.

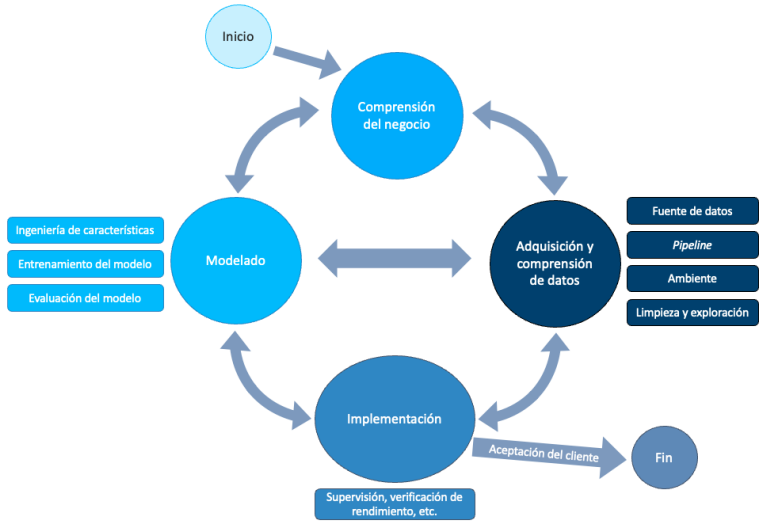

Figura 3.3: Ciclo de ciencia de datos.

Cada estación del año tiene una tendencia diferente, considerando esto, se realizó una diagrama de correlación por cada estación para comprender las relaciones de las variables predictoras con la temperatura interna (Figura. 3.4, 3.5).

Los diagramas de correlación indicaron relaciones de variables similares, para las estaciones del año donde tienden ser más calurosas la radiación solar tiene una mayor correlación positiva con la temperatura interna con 0.75 y 0.66 para verano y primavera respectivamente, mientras que para invierno y otoño, 0.24 y 0.28.

La humedad interna, rocío interno y la temperatura externa presentan valores altos de correlación con la temperatura interna, sin

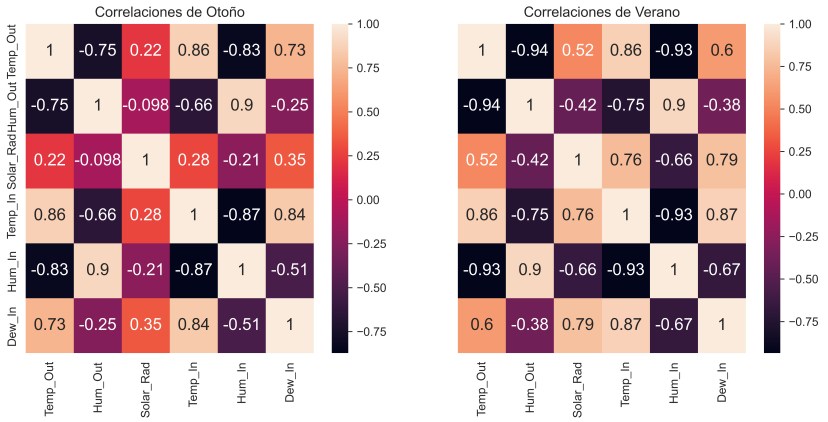

Figura 3.4: Diagrama de correlación para otoño y verano.

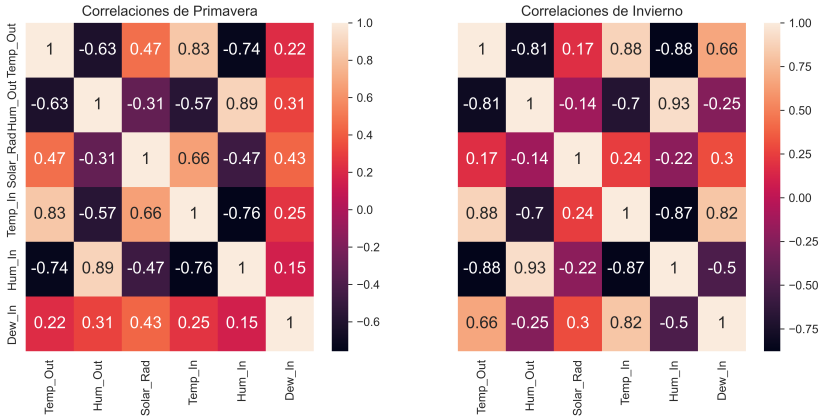

Figura 3.5: Diagrama de correlación para primavera e invierno.

especular causalidad, estas variables podrían explicar la temperatura interna del invernadero. Esto no es el caso para los datos de primavera, donde el rocío interno no tiene correlación significativa con la temperatura interna, en cambio, tiene correlaciones significativas con temperatura externa, radiación solar y humedad interna.

La temperatura externa tiene valores significativos de correlación con la humedad externa e interna, es notable el valor debido a la selección inicial de predictores en los modelos, éstas correlaciones entre variables independientes indican multicolinealidad, la cual también se presenta en la humedad externa e interna.

Continuando con el análisis de correlación, se tomaron todos los datos de todas las estaciones para generar un conjunto de datos principal, el cual fue utilizado para la realización de un diagrama de correlación el cual permitió determinar multicolinealidad entre variables predictoras así como las variables con mayor correlación con la temperatura interna del invernadero.

Considerando las diferencias entre las correlaciones por cada estación del año se generó un diagrama de correlación para el conjunto principal de los datos. Para los datos anuales, destacan las variables de humedad interna con correlación negativa y temperatura externa, mientras que la radiación solar, humedad externa y el rocío interno no mostraron valores significativos de correlación con la temperatura interna. Existe multicolinealidad entre la variable humedad interna con la temperatura y humedades externas (Figura 3.6).

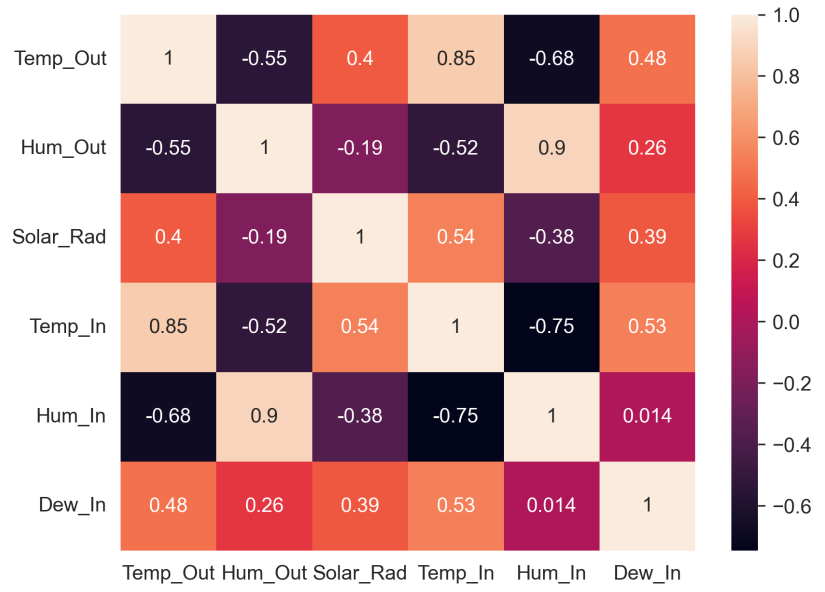

Figura 3.6: Diagrama de correlación para datos anuales.

Una parte esencial en el desarrollo de un modelo matemático para predicción es el ajuste de los hiper parámetros del algoritmo[2]. El ajuste de los hiperparámetros es un elemento importante debido a que controla el comportamiento de un modelo de aprendizaje automático. En caso de no ajustarlos, éstos pueden producir resultados no óptimos ya que no minimizan la función de costo, lo cual hace que el modelo genere más errores. La selección de los hiperparámetros puede recurrirse a valores por defecto que se han especificado en los paquetes de software de aplicación o por configuración manual, la cual puede ser obtenida por recomendaciones bibliográficas, experiencia o por ensayo y error.

Adicionalmente, se pueden utilizar estrategias de optimización de hiperparámetros[3] las cuales son dependientes de los datos y son procedimientos de segundo grado de optimización, que intentan minimizar el error esperado del modelo utilizando una búsqueda de candidatos para la configuración de hiperparámetros. Ésta búsqueda se evalúa con las predicciones que se generan con esa configuración o bien

[2] P. Probst, A. Boulesteix, and B. Bischl. Tunability: Importance of hyperparameters of machine learning algorithms, 2018

[3] L. Yang and A. Shami. On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, 415:295–316, 2020. DOI: 10.1016/j.neucom.2020.07.061

se puede evaluar con un esquema de remuestreo como la validación cruzada. Algunas de estas estrategias son *grid search* o el *random search*[4] los cuales son proceso iterativos y exhaustivos donde se especifica una lista de parámetros candidatos y escoge un modelo que tenga un error mínimo especificado. Así mismo, se encuentran otras estrategias más complejas que de igual manera son iterativas, entre ellas se encuentra la optimización bayesiana e *iterated F-racing*[5]

[4] J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.*, 13:281–305, feb 2012. ISSN 1532-4435

[5] M. López-Ibáñez, J. Dubois-Lacoste, L. P. Cáceres, M. Birattari, and T. Stützle. The irace package: Iterated racing for automatic algorithm configuration. *Operations Research Perspectives*, 3:43–58, 2016. DOI: 10.1016/j.orp.2016.09.002

3.3 Optimización Bayesiana

La optimización bayesiana fue utilizada para la obtención del mejor conjunto de hiperparámetros del modelo, su uso permitió minimizar la función no convexa de una forma la cual algún otro método no permite. Los métodos de *random search* y *grid search* fueron utilizados inicialmente para el análisis de datos de esta investigación, para el pronóstico se usó la optimización bayesiana para el ajuste del mejor candidato debido a su superioridad a comparación con los métodos antes dichos[6].

[6] R. Turner, D. Eriksson, M. McCourt, J. Kiili, E. Laaksonen, Z. Xu, and I. Guyon. Bayesian optimization is superior to random search for machine learning hyperparameter tuning: Analysis of the black-box optimization challenge 2020, 2021

Como en otros tipos de optimización, el método bayesiano busca encontrar el mínimo de una función $f(x)$ en algún conjunto acotado X, tomado de un subconjunto de $\mathbb{R} \rightarrow D$. La diferencia de la optimización bayesiana a los otros métodos es que construye un modelo probabilístico para $f(x)$ y después reutiliza este modelo para tomar decisiones en donde posteriormente se debe de evaluar la función en X, al mismo tiempo que integra la incertidumbre. Al ser un método bayesiano[7], utiliza toda la información disponible de las evaluaciones previas de $f(x)$ y no simplemente confía en las aproximaciones locales del gradiente o el Hessiano. Esto da lugar a un procedimiento donde puede encontrar mínimos de funciones no convexas con relativamente pocas evaluaciones, a costa de realizar más cálculos para determinar el siguiente punto a probar. Usualmente las evaluaciones en el entrenamiento de un algoritmo de aprendizaje automático son costosas por lo que es fácil justificar un cálculo adicional para tomar mejores decisiones.

[7] A. Gelman and C. Rohilla-Shalizi. Philosophy and the practice of bayesian statistics. In *Mathematical and Statistical Psychology*, volume 66, pages 8–38, 2013

Cuando se realiza optimización bayesiana hay que tomar dos decisiones importantes[8]. La primera es seleccionar un priori sobre las funciones que expresarán las hipótesis de la función que se estará optimizando. La segunda decisión es elegir una función de adquisición que se estará utilizando para construir una función de utilidad a partir del modelo posterior, lo que permite determinar el siguiente punto a evaluar (Figura 3.7).

[8] Cornell University. Gaussian processes and bayesian optimization. https://www.cs.cornell.edu/courses/cs4787/2019sp/notes/lecture16.pdf, 2019. Accesado en: 2023-03-21

El proceso Gaussiano a priori (GP, por sus siglas en inglés) es una distribución a priori conveniente y potente para las funciones[9]. GP se define por la propiedad que cualquier conjunto finito de N puntos

[9] K. Wabersich. Gaussian processes and bayesian optimization. https://www.kimpeter.de/wp-content/uploads/2016/12/project_GBO.pdf, 2016. Accesado en: 2023-03-21

$\{x_n \in X\}_n^N = 1$ induce una distribución Gaussiana multivariable en $\mathbb{R}^N$. El enésimo de estos puntos se toma como el valor de la función $f(x_n)$ y las propiedades de marginación de la distribución permiten calcular los marginales y condicionales de forma cerrada.

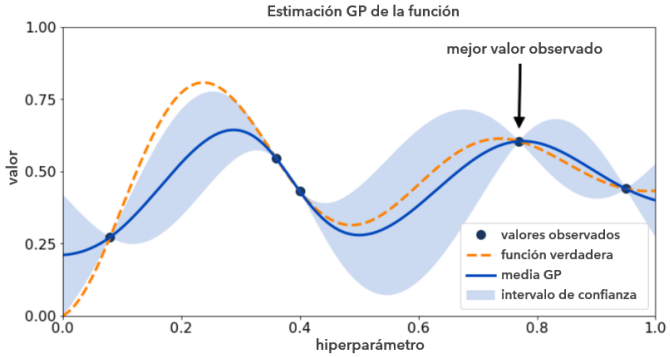

Figura 3.7: Primer estimación GP de la función. Imagen de cortesía *Bayesian Optimization for machine learning algorithms in the context of Higgs searches at the CMS experiment* en *https://doi.org/10.48550/arXiv.1911.02501*

La función de adquisición para la optimización bayesiana supone que la función $f(x)$ es extraída de un proceso Gaussiano a priori y que las observaciones son de la forma $\{x_n, y_n\}_n^N = 1$, donde $y_n \sim \mathcal{N}(f(x_n), v)$ y v es la varianza del ruido introducido en las observaciones. La priori y los datos inducen una posterior sobre la función de adquisición, la cual determina qué punto de X debe de ser evaluado a continuación con optimización $x_{next} = \arg\max_x a(x)$, donde se han propuesto varias funciones diferentes (Figura 3.8) y en general éstas funciones dependen de observaciones anteriores. Existen distintas opciones populares de funciones de adquisición las cuales dentro del proceso Guassiano a priori dependen completamente del modelo a utilizar.

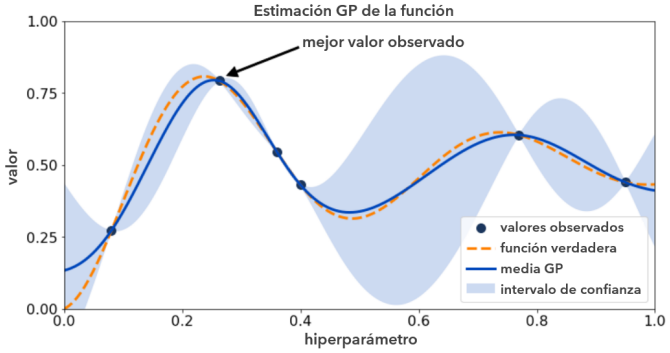

Figura 3.8: Segunda estimación GP de la función. Imagen de cortesía *Bayesian Optimization for machine learning algorithms in the context of Higgs searches at the CMS experiment* en *https://doi.org/10.48550/arXiv.1911.02501*

Cualquier modelo requiere ser evaluado para determinar su funcionalidad en el pronóstico de la temperatura interna del invernadero. Debido a ser un problema de estimación de valores continuos, regresión, las métricas utilizadas son diferentes a un problema de clasificación. La calidad del modelo debe evaluarse en relación con el objetivo, esta evaluación pretende saber si es adecuado o apto para el pronóstico de la temperatura. Este punto de vista puede

contrastarse con otro en el que la calidad del modelo es solo una cuestión de precisión y totalidad con que un modelo representa el objetivo general, donde el límite ideal es un pronóstico perfecto y completo[10]. En este estudio se utilizaron cuatro diferentes métricas para la evaluación y prueba de cada modelo matemático. A continuación se hace una descripción de cada uno.

[10] W. S. Parker. Model evaluation: An adequacy-for-purpose view. *Philosophy of Science*, 87:457–477, 2020. DOI: 10.1086/708691

3.4 Métricas

Predecir un objetivo de valores continuos es una tarea difícil no sólo en términos prácticos sino también en nivel conceptual. La base teórica del uso de un algoritmo para determinar valores continuos abarca varios aspectos que revelan conexiones posibles entre los datos, dependientes e independientes. La elección de la métrica para evaluar un modelo es de suma importancia debido a que gracias a este es posible explicar la relación y el objetivo principal del fenómeno. El presente estudio utiliza cuatro diferentes métricas para evaluar el pronóstico de temperatura interna dentro del invernadero controlado. La R-cuadrada[11] (R^2) es una métrica muy utilizada en el área de predicciones de valores continuos. La formulación original cuantifica en qué medida la variable dependiente está determinada por las variables independientes en términos de proporción de varianza, queda de la forma (3.1). Un valor cercano a uno indica una buena predicción independientemente de la escala en la que se midan las variables.

[11] Wright S. Correlation and causation. *Journal of agricultural research*, pages 557–585, 1921. ISSN 0095-9758

$$\begin{aligned} R^2 &= 1 - \frac{RSS}{TSS} \\ &= 1 - \frac{\sum(y_i - \hat{y}_1)^2}{\sum(y_i - \overline{y}_1)^2} \end{aligned} \tag{3.1}$$

donde *RSS* es la suma residual de cuadrados y *TSS* es la suma total de cuadrados.

La raíz del error cuadrático medio (RMSE) es una derivación natural del error cuadrático medio (MSE) utilizado para estandarizar sus unidades de medida. El MSE[12], a diferencia de R^2 que es determinada por la varianza, busca evaluar la calidad del ajuste en términos de distancia del regresor a los puntos de entrenamiento reales. El MSE se puede utilizar si es necesario detectar valores atípicos. Gracias a la norma L_2, el MSE es ideal par atribuir pesos más elevados a los puntos, esto significa que si la predicción es muy mala, el cuadrado de la función aumentará el error, la formulación queda de la forma (3.2). Un valor cercano a cero indica una buena predicción y una peor cuando el valor tiende a infinito.

[12] C. Sammut and G. I. Webb. *Encyclopedia of Machine Learning*, chapter Mean Squared Error, pages 653–653. Springer US, 2010b. ISBN 978-0-387-30164-8

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(X_i - Y_i)^2} \tag{3.2}$$

El error medio absoluto[13] (MAE) al igual que el MSE busca evaluar el resultado en términos de distancias desde el regresor a los puntos reales. La formulación del MAE considera una norma L_1 por lo que se puede utilizar si los valores atípicos representan partes corruptas de los datos la cual queda de la forma (3.3). MAE no penaliza demasiado los valores atípicos, esto debido a su norma que suaviza todos los errores de los posibles valores atípicos. Es por esto que MAE proporciona una medida de rendimiento genérica y acotada para el modelo. En caso de tener muchos valores atípicos el rendimiento del modelo será mediocre. Un valor cercano a cero indica una buena predicción y una peor cuando el valor tiende a infinito.

$$MAE = \frac{\sum_{i=1}^{n} \mid y_i - x_i \mid}{n} \tag{3.3}$$

[13] C. Sammut and G. I. Webb. *Encyclopedia of Machine Learning*, chapter Mean Absolute Error, pages 652–652. Springer US, 2010a. ISBN 978-0-387-30164-8

Dentro de la misma familia de métricas, el error porcentual medio absoluto[14] (MAPE) se centra en una interpretación porcentual del MAE, siendo así la métrica usada cuando existen variaciones que tienen un mayor impacto en la estimación en lugar de los valores absolutos. Debido a su formulación, está muy sesgado hacia los pronósticos bajos, por lo que no es adecuado para evaluar tareas donde se esperan errores de grandes magnitudes donde queda de la forma (3.4). Al igual que el MAE, un valor cercano a cero indica una buena predicción y una peor cuando el valor tiende a infinito.

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \frac{\mid y_i - x_i \mid}{y_i} \tag{3.4}$$

[14] A. De Myttenaere, B. Golden, B. Le Grand, and F. Rossi. Mean absolute percentage error for regression models. *Neurocomputing*, 192:38–48, 2016. DOI: 10.1016/j.neucom.2015.12.114

Las métricas utilizadas en este análisis tienen un objetivo específico. El MAE al mostrar un error medio absoluto en términos de los datos fue usado para servir como intervalo de aceptabilidad considerando una histéresis de $\pm 2°C$, mientras que el RMSE denota la variación de los errores considerando los picos de temperatura interna, se busca un RMSE igual o cercano al error medio absoluto donde se busca que grandes errores sean particularmente indeseables. El error MAPE permite la comparación entre los distintos modelos utilizados en el análisis gracias a su representación porcentual. Por último, la R^2 muestra qué tan bueno el modelo se ajusta a la variable dependiente en términos de varianza, el valor de ésta métrica indica la variación explicada por la variable dependiente por las variables independientes en el modelo, sin embargo no indica si el pronóstico es particularmente bueno debido a que puede ser afectada por sobre ajuste o multicolinealidad, es por eso que se consideraron otras métricas.

Debido a la existente multicolinealidad, múltiples combinaciones de variables para explicar la variable dependiente así como el comportamiento de los datos a través del tiempo, se planteó el uso de una regresión lineal simple para el pronóstico actual de la temperatura

interna del invernadero. Gracias a los coeficientes y la simplicidad de un modelo de regresión lineal es posible determinar predictores para su uso en un modelo más complejo. El uso de un modelo interpretable y reproducible para su integración en un microcontrolador es uno de los objetivos principales del proyecto magno no particularmente de este trabajo, por lo que la regresión lineal provee de información sobre correlación entre variables independientes con la variable dependiente lo cual es de suma importancia para la comprensión del crecimiento o disminución de la temperatura dentro del invernadero.

3.5 *Detalles del software utilizado*

Todos los algoritmos de modelado de datos fueron ejecutados usando el *software* editor de código *Visual Studio Code*[15] y utilizando el lenguaje de programación *Python*[16]. Los datos fueron extraídos del sistema central meteorológico e importados al *software* utilizando *pandas*[17] con el fin de manipular, analizar y utilizar los datos; mientras que para la ejecución de operaciones matemáticas complejas en vectores con el fin de optimizar el rendimiento del equipo donde se ejecutan los algoritmos se utilizó *NumPy*[18]. Se utilizaron las librerías de *MatPlotLib*[19] y *Seaborn*[20] para la representación gráfica de las series de tiempo así como de los pronósticos. *bayesopt*[21] fue utilizado para la optimización de hiperparámetros y las paqueterías de *xgboost*[22] y *sklearn*[23] para el desarrollo de los algoritmos matemáticos para el pronóstico de la temperatura interna.

3.6 *Regresión Lineal*

La regresión lineal es una de las técnicas relativamente simples y más utilizadas. La razón se debe a las ventajas que proporciona para el entendimiento de los datos así como la capacidad de representar fenómenos complejos de forma clara y sencilla; esto se obtiene mediante una fórmula matemática fácil de interpretar y reproducir.

Como fue mencionado, el análisis subyacente de la técnica de la regresión lineal posibilita medir la relación entre las variables, lo cual permite obtener conocimiento sobre las mismas en el campo; esto quiere decir que una vez conociendo estas relaciones se puede determinar si las variables tienen importancia para la predicción del modelo[24]. Esto también se conoce como un modelo capaz de analizar la variabilidad de una determinada variable en función a información que le proporcionan una o más variables[25].

El término regresión se debe a Sir Francis Galton cuando estudiaba la relación y su fascinación entre padres e hijos[26]. Galton en su publicación *Regression Towards Mediocrity in Heriditary Stature* observó que los hijos

[15] Microsoft Corporation. Visual studio code. https://code.visualstudio.com/docs, 2015. Accesado en: 2023-03-24

[16] G. Van Rossum and Fred L. D. *Python 3 Reference Manual*. CreateSpace, 2009. ISBN 1441412697

[17] Wes McKinney et al. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56, 2010

[18] Charles R. Harris et al. Array programming with NumPy. *Nature*, page 357–362, 2020. DOI: 10.1038/s41586-020-2649-2

[19] John D Hunter. Matplotlib: A 2d graphics environment. *Computing in science & engineering*, pages 90–95, 2007

[20] Michael Waskom et al. mwaskom/seaborn: v0.8.1 (september 2017). *Zenodo*, 2017. DOI: 10.5281/zenodo.883859. URL https://doi.org/10.5281/zenodo.883859

[21] R. Martinez-Cantin. Bayesopt: A bayesian optimization library for nonlinear optimization, experimental design and bandits. *Journal of Machine Learning Research*, pages 3915–3919, 2014. URL http://jmlr.org/papers/v15/martinezcantin14a.html

[22] T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016b. ISBN 978-1-4503-4232-2. DOI: 10.1145/2939672.2939785

[23] F. Pedregosa et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12:2825–2830, 2011

[24] S. Weisberg. *Applied linear regression*. John Wiley & Sons, Inc, 2005. ISBN 9780471704096

[25] C. Camacho. Regresión lineal simple. Accesado en: 2023-01-18, December 2019. URL https://personal.us.es/vararey/regresion-simple.pdf

[26] J. M. Stanton. Galton, pearson, and the peas: A brief history of linear regression for statistics instructors. *Journal of Statistics Education*, 9, 2001. DOI: 10.1080/10691898.2001.11910537. URL https://doi.org/10.1080/10691898.

altos tenían padres altos pero no tan altos como sus progenitores. Y de igual manera, los hijos que tenían padres de baja estatura tendían a tener abuelos de mayor estatura; esto indicaba una tendencia a una estatura media, la estatura regresaba al promedio[27]. Es por esto, la razón por la que la palabra regresión se empezó a utilizar para fenómenos de este tipo, y rápidamente fue adaptada al presente y empleándose hoy en día.

[27] F. Galton. Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263, 1886. ISSN 09595295. URL http://www.jstor.org/stable/2841583

La estructura general del modelo de regresión lineal se puede caracterizar principalmente por simple y múltiple. En cuestión al modelo de regresión lineal simple su estructura considera una variable dependiente Y y una variable independiente X quedando una expresión de la siguiente forma:

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{3.5}$$

La variable Y es dependiente, ésta es la que el modelo quiere predecir; también conocida como la variable *respuesta* o de *salida*. La regresión busca estimar la relación que tiene Y con X y predecir -con el uso de la ecuación- valores no muestrados. Las variables β_0 y β_1 corresponden al intercepto y la pendiente, respectivamente; los cuales son los coeficientes de regresión o parámetros del modelo. Por último, la variable ϵ representa el error en la predicción de la variable *respuesta* debido a la relación estocástica entre Y y X.

El valor del error usualmente se debe fundamentalmente a factores de medición incorrecta de la *respuesta* o algunas otras variables que no se tomaron en consideración que tienen relación con la variable Y. Es importante destacar que aunque se tenga el supuesto de relación lineal entre Y y X la realidad es que la ecuación de la forma lineal simple puede proveer una aproximación bastante aceptable a la estimación de Y, esto gracias a las desviaciones del modelo que serán recogidas por ϵ.

El modelo de regresión lineal se puede extender a ser múltiple, esto permite tomar k variables como predictores para el modelo. El modelo de regresión lineal múltiple entonces quedaría de la forma:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \epsilon \tag{3.6}$$

Es destacable que se incluye el intercepto como es en el modelo simple, sin embargo, las variables β_1, β_2 y β_k no representan la pendiente de la recta, sino al ahora ser un espacio de k dimensiones representa las pendientes del hiperplano que se forma con los predictores[28].

[28] D. C. Montgomery, E. A. Peck, and G. G. Vining. *Introduction to Linear Regression Analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., 2015. ISBN 9781119180173. URL https://books.google.com.mx/books?id=27kOCgAAQBAJ

Conociendo, entonces, que los parámetros β se deben de estimar con la muestra de los datos, las inferencias de la estimación deben ser adecuados tomando en consideración las siguientes características estadísticas (igualmente conocidos como supuestos)[29]:

- Linealidad: El supuesto de linealidad en términos básicos indica que la recta o hiperplano para la estimación de Y está determinada por

[29] C. Camacho. Regresión lineal simple. Accesado en: 2023-01-18, December 2019. URL https://personal.us.es/vararey/regresion-simple.pdf

los valores medios de Y para cada valor de X. En consecuencia, la esperanza matemática de los errores será cero, teniendo en cuenta que el error es:

$$\epsilon = Y - \hat{Y} \tag{3.7}$$

- Homocedasticidad: El supuesto de homocedasticidad indica que la dispersión de la variable Y a todo lo largo de la recta de regresión es constante. La ventaja de tener un supuesto como este es la simplicidad del modelo, donde se puede utilizar un valor de X para todo el recorrido para estimar Y (Figura 3.9).

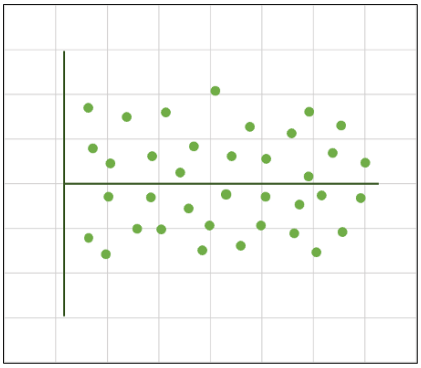

Figura 3.9: Representación gráfica de datos cumpliendo el supuesto de homocedasticidad

- Ausencia de autocorrelación: El supuesto de ausencia de autocorrelación establece que las variables independientes Y son independientes entre sí, esto quiere decir que no existe correlación entre ellas, covarianza cero.

- Normalidad: El supuesto de normalidad establece que la distribución de Y para cada valor de X se distribuye de manera normal.

Como fue mencionado anteriormente, β es un parámetro el cual es necesario ser estimado. La estimación de estos parámetros se obtienen con los datos obtenidos de la muestra. Para el caso de regresión lineal, se utiliza el método de mínimos cuadrados ordinarios o conocido *ordinary least squares*. El método consiste en minimizar la suma de los cuadrados del error, esto se puede visualizar fácilmente en (Figura 3.10) el cual es un diagrama de dispersión con líneas verticales que representan la distancia de los puntos a la línea recta; distancias verticales a los puntos partiendo de la línea recta de regresión.

La ecuación de mínimos cuadrados ordinarios para regresión lineal simple es, entonces:

$$S = \sum_{i=0}^{n} \epsilon_i^2 = \sum_{i=0}^{n} (y_i - \beta_i x_i - \beta_0)^2 \tag{3.8}$$

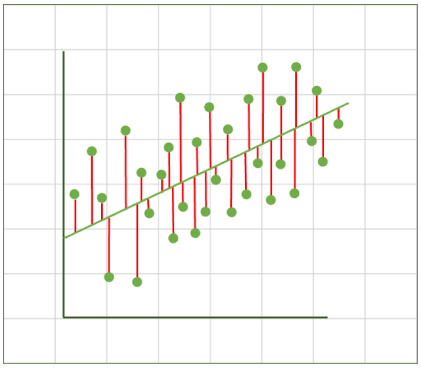

Figura 3.10: Representación gráfica del proceso de mínimos cuadrados ordinarios

Mínimos cuadrados ordinarios tiene como fin el encontrar las estimaciones de los coeficientes de regresión β que minimicen S. La ecuación de mínimos cuadrados ordinarios simple se puede extender de igual manera para el modelo de regresión múltiple, por lo que ayuda en encontrar el aporte de cada variable al modelo[30]; ésta entonces se vuelve la ventaja principal para el uso de la regresión lineal para el análisis presentado, sin embargo cabe destacar que este coeficiente no indica causalidad.

[30] J. Hernández-Lalinde, J. Espinosa-Castro, V. Bermudez, and D. García Álvarez. Sobre el uso adecuado de la regresión lineal: conceptualización básica mediante un ejemplo aplicado a las ciencias de la salud. *Archivos Venezolanos de Farmacologia y Terapeutica*, 38:608–614, 12 2020

El uso de la regresión lineal para este análisis brindó información relevante para la comprensión del comportamiento de los datos. Cada estación posee un comportamiento distinto, por lo que es de suma importancia capturar las diferencias para llegar a utilizar un modelo capaz de pronosticar la temperatura interna. El siguiente uso de la regresión lineal fue dividido en las estaciones del año con las 5 variables sin la consideración de las correlaciones entre cada una. Los datos fueron divididos en 80 % entrenamiento y 20 % prueba.

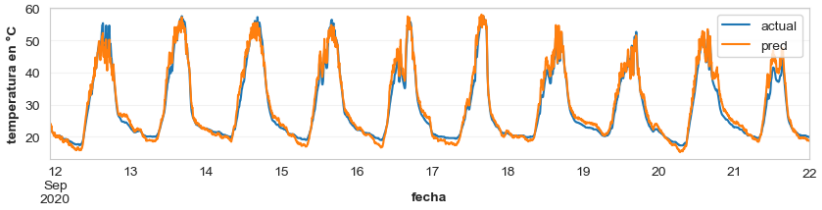

Figura 3.11: Predicción de temperatura interna con el uso de una regresión lineal múltiple para datos de verano.

El uso de la regresión lineal múltiple es capaz de realizar un acercamiento del comportamiento de la temperatura interna con resultados buenos para los datos de prueba en verano con $RMSE$: 1.9574, R^2: 0.9680, MAE: 1.5139 y $MAPE$: 0.0520. Una de las características principales de los datos de verano es el límite de temperaturas bajas en $20°C$. El comportamiento de la predicción para temperatura interna actual se mantiene y tiene dificultad para picos así como temperaturas internas mayores a $50°C$ (Figura 3.13). La regresión lineal múltiple permite la replicación sencilla del modelo mediante una

ecuación (3.9):

$$T_{interna} = 36.6276 - 0.6039x_1 - 0.2938x_2 - 0.0010x_3 + 2.2838x_4 - 0.2380x_5 \quad (3.9)$$

Considerando la estructura del modelo de regresión lineal múltiple, el primer término de la ecuación corresponde al intercepto β_0 mientras que los demás corresponde a cada variable independiente x utilizado como variable de entrada en el modelo. Los coeficientes de la regresión lineal permiten determinar el peso de cada variable de entrada para la regresión, indicando la aportación de cada variable.

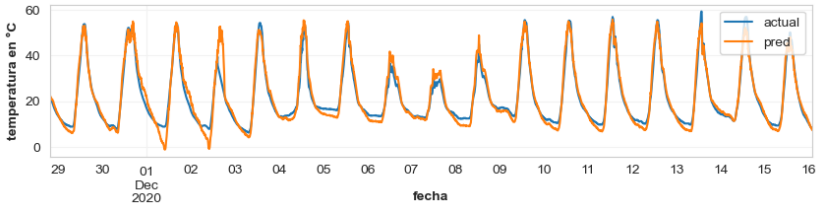

Figura 3.12: Predicción de temperatura interna con el uso de una regresión lineal múltiple para datos de otoño.

La siguiente regresión lineal múltiple fue utilizada para la predicción del comportamiento de temperatura interna para los datos de otoño. La temperatura interna dentro del invernadero es capaz de disminuir a menos de 20°C, es importante notar este comportamiento debido a que la regresión lineal múltiple para los datos de verano no considera temperaturas internas menores a 20°C, por lo que no es posible su uso para la predicción de datos en otoño. El modelo de otoño (Figura 3.12) tiene dificultades en capturar temperaturas menores a 20°C, el modelo realiza acercamientos de temperaturas más bajas que lo observado, sin embargo los resultados siguen siendo buenos con $RMSE$: 2.9244, R^2: 0.9462, MAE: 2.0621 y $MAPE$: 0.1163. De igual manera se puede obtener la ecuación para el modelo de regresión lineal (3.10):

$$T_{interna} = 38.8066 - 0.3622x_1 + 0.0201x_2 - 0.004x_3 + 1.4841x_4 - 0.4852x_5 \quad (3.10)$$

Los datos de invierno llegan a tener temperaturas desde 0°C hasta 70°C, con base en los resultados previos, el modelo de regresión lineal múltiple puede tener problemas en capturar los picos de temperatura interna. Debido a tener temperaturas menores a 10°C el modelo no es capaz de capturar temperaturas mayores a 50°C. El modelo de regresión lineal múltiple para datos de invierno (Figura 3.13) es el que tiene resultados desfavorables a comparación de los demás con $RMSE$: 4.6499, R^2: 0.9244, MAE: 3.0969 y $MAPE$: 0.1066.

La ecuación utilizada para el uso del modelo de regresión lineal múltiple para datos de invierno entonces queda de la forma (3.11):

$$T_{interna} = 41.4071 - 0.2003x_1 - 0.0498x_2 - 0.0105x_3 + 1.6687x_4 - 0.5285x_5 \quad (3.11)$$

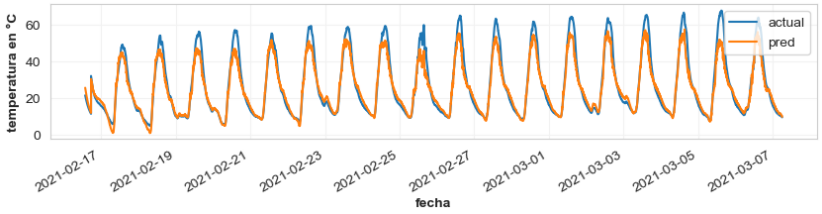

Figura 3.13: Predicción de temperatura interna con el uso de una regresión lineal múltiple para datos de invierno.

Por último, se realizó una regresión lineal múltiple para los datos de primavera (Figura 3.14). A diferencia de los otros modelos, este modelo fue capaz de capturar mejor el comportamiento de la serie de tiempos para temperaturas mayores a $40°C$ con excepción de tres días donde la temperatura se elevó hasta $60°C$, sin embargo la regresión tiene dificultad en realizar predicciones acertadas para temperaturas bajas, donde destaca el 13 de junio en adelante que es cuando la predicción no es buena. El modelo, no obstante, tuvo resultados buenos con $RMSE$: 3.6264, R^2: 0.9077, MAE: 2.9595 y $MAPE$: 0.1096

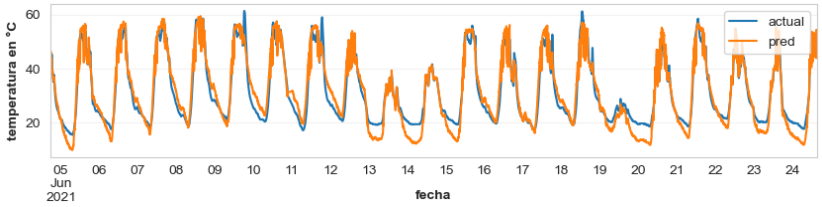

Figura 3.14: Predicción de temperatura interna con el uso de una regresión lineal múltiple para datos de primavera.

La ecuación utilizada para el uso del modelo de regresión lineal múltiple para datos de primavera entonces queda de la forma (3.12):

$$T_{interna} = 11.2897 + 0.8994x_1 + 0.1920x_2 + 0.0136x_3 + 0.0550x_4 - 0.3849x_5 \quad (3.12)$$

La regresión lineal múltiple permite mostrar una ecuación capaz de reproducir el fenómeno de temperatura interna del invernadero. Al ser un modelo simple, los resultados mostraron un buen acercamiento al comportamiento de la serie de tiempo por lo que es necesario buscar un modelo capaz de realizar un mejor acercamiento al comportamiento y con capacidades de pronóstico. Existe multicolinealidad entre variables

(Figura 3.6) por lo que es necesario considerar un modelo que pueda lidiar con esto así como mejorar el tiempo de ejecución. El modelo de regresión lineal múltiple no tomó tiempo significante de ejecución, sin embargo es posible reducirlo al utilizar menos de cinco variables de entrada. La regresión con mínimos cuadrados parciales es un acercamiento posible para lidiar con problemas de multicolinealidad y valores nulos. Este análisis permitirá usar un menor número de variables utilizando una combinación lineal de las variables independientes para así reducir su carga computacional así como su tiempo de ejecución. En el siguiente apartado se explicará de una regresión con mínimos cuadrados parciales con regresión (PLSR).

3.7 *Regresión con mínimos cuadrados parciales*

El análisis de mínimos cuadrados parciales o también conocido como *Partial Least Squares* (PLS) es una técnica que permite la comparación entre variables de entrada y variables de salida. La técnica es uno de diferentes métodos existentes para lidiar con la covarianza de una base de datos[31]. Originalmente fue diseñada para lidiar con problemas de regresión múltiple donde se no se tienen suficientes datos, existen muchos valores nulos y multicolinealidad. Su uso ha sido muy popular en ciencias como química y quimiometría donde existe una alta cantidad de variables correlacionadas y un límite de datos para ser utilizados[32].

Herman O. A. Wold en 1975 desarrolló la regresión PLS originalmente para el uso en el área de econometría para el modelado de base de datos complicadas en términos de cadenas de matrices (en bloques), sin embargo fue prontamente utilizado por el área de química con un uso analítico, físico y químico clínico. El método utiliza un manera simple y eficiente para estimar los parámetros conocida como NIPALS (*Non-linear Iterative PArtial Least Squares*). El acrónimo PLS se relaciona entonces, a la parte central de la estimación del algoritmo NIPALS; donde cada parámetro del modelo es estimado iterativamente como la pendiente de una regresión simple bivariante (mínimos cuadrados) entre una columna matriz o fila como la variable y y otro vector parámetro como la variable x[33].

Una alta multicolinealidad aumenta el riesgo que, teóricamente, la variable a predecir salga rechazada por el modelo como una variable no significativa. El objetivo de PLS es predecir la variable dependiente con la variable independiente y describir la estructura que exista entre ambas. Este es un método de regresión que permite la identificación de factores subyacentes, los cuales son combinaciones lineales de las variables dependientes que mejor expliquen la variable independiente.

La regresión PLS se puede comparar con el método de análisis de

[31] M. Tenenhaus. *La régression PLS: Théorie et pratique*. Technip, 1998

[32] Y. Chiba, K. Okada, Y. Hayashi, S. Kumada, and Y. Onuki. Usefulness of applying partial least squares regression to t_2 relaxation curves for predicting the solid form content in binary physical mixtures. *Journal of Pharmaceutical Sciences*, 112: 1041–1051, 2023. ISSN 0022-3549. DOI: https://doi.org/10.1016/j.xphs.2022.11.028. URL https://www.sciencedirect.com/science/article/pii/S0022354922005391

[33] S. Wold, M. Sjöström, and L. Eriksson. Pls-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58: 109–130, 2001. ISSN 0169-7439. DOI: https://doi.org/10.1016/S0169-7439(01)00155-1. URL https://www.sciencedirect.com/science/article/pii/S0169743901001551. PLS Methods

componentes principales (PCA), análisis canónico y mínimos cuadrados alternados, sin embargo el método de PLS es una mejor alternativa debido a que provee un modelo mucho más robusto de parámetros con una nueva calibración de variables que considera la variable dependiente. PCA obtiene las variables compuestas que expliquen al máximo la variabilidad de la matriz de datos X[34]. Adicionalmente, PLS es una mejora a la técnica de PCA debido a que la solución dada por PLS está restringida a la parte de la matriz de covarianza que está directamente relacionada a la manipulación experimental de los datos de entrada.

[34] S. Maitra and J. Yan. Principle component analysis and partial least squares: Two dimension reduction techniques for regression. *Casualty Actuarial Society*, 01 2008

La parte de matriz de correlación o covarianza que el PLS quiere entrenar es la estrategia de correlación entre las variables X y las variables Y. PLS mide la covarianza entre dos o más variables y crea un nuevo conjunto de variables utilizando combinaciones lineales optimizados para obtener su máxima covarianza utilizando el menor número de dimensiones, esto quiere decir que el número de variables de entrada se disminuye utilizando la covarianza de los datos de entrada e información de la variable dependiente[35]. El método PLS es una técnica lineal la cual es preferida como de uso predictivo y no como una técnica de interpretación, sin embargo se puede utilizar para esos propósitos en la parte de análisis y exploración de datos antes de hacer uso de algun otro modelo para predicción. Las relaciones lineales óptimas son calculadas entre las variables de entrada y pueden ser interpretadas como la mejor combinación de variables de predicción disponibles para el estudio tomando en cuenta todas sus limitaciones.

[35] A. McIntosh, F Bookstein, J. Haxby, and C. Grady. Spatial pattern analysis of functional brain images using partial least squares. *NeuroImage*, 3: 143–157, 1996. ISSN 1053-8119. DOI: https://doi.org/10.1006/nimg.1996.0016. URL https://www.sciencedirect.com/science/article/pii/S1053811996900166

El método PLS asume matrices de la variable X y combinaciones lineales de las mismas junto con la variable Y. Asumiendo que X es una matriz $n \times p$ y Y es una matriz $n \times q$, la técnica funciona extrayendo satisfactoriamente factores de ambas variables X y Y donde la covarianza de los factores extraídos es la máxima. A pesar que el método pueda también funcionar con múltiples variables de respuesta, para esta demostración solo se estará asumiendo una variable de respuesta; Y es $n \times 1$ y X es $n \times p$.

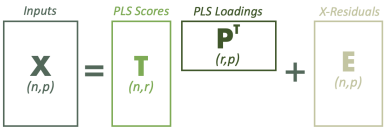

Figura 3.15: Forma de X: $X = TP^T + E$

Lo que se busca en el método es encontrar una descomposición lineal de X y Y que satisfaga $X = TP^T + E$ y para $Y = UQ^T + F$; donde:

$$T_{n\times r} = X - scores \qquad U_{n\times r} = Y - scores$$
$$P_{p\times r} = X - loadings \qquad Q_{1\times r} = Y - loadings$$
$$E_{n\times p} = X - residuals \qquad F_{n\times 1} = Y - residuals$$

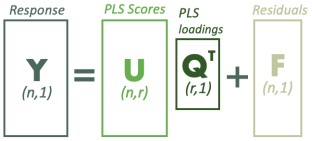

Figura 3.16: Forma de Y: $Y = UQ^T + F$

La descomposición se logra cuando se encuentra la covarianza máxima entre T y U. Para lograr resolver el problema de optimización, los algoritmos utilizados usan un proceso iterativo para extraer $X - scores$ y $Y - scores$. Estos factores o *scores* para X y Y son extraídos satisfactoriamente y el número *(r)* depende del rango de X y Y.

Cada $x - score$ es una combinación lineal de X. Por ejemplo, el primero *(t)* está de la forma $t = Xw$, donde w es un vector propio correspondiente al primer valor propio de $X^T YY^T X$. De igual manera, el primer $y - score$ *(u)* está de la forma $u = Yc$ donde c es el vector propio correspondiente al primer valor propio de $Y^T XX^T Y$. X^T denota la covarianza entre X y Y.

Una vez el primer factor se extrae, se deflectan los valores originales de X y Y con $X_1 = X - tt^T X \quad and \quad Y_1 = Y - tt^T Y$. Este proceso se repite hasta obtener los siguientes factores. Por último, el proceso de extracción se termina hasta que todos los posibles factores latentes de t y u se obtienen; esto quiere decir que termina cuando X se reduce a una matriz nula; el número de factores depende del rango de X. Usualmente de 3 a 7 factores contienen el 99 % de la varianza.

Obteniendo los factores disponibles utilizando PLS, se aplica una regresión lineal considerando los factores como variables de entrada. Ésta regresión tiene como objetivo el transformar un conjunto de variables correlacionadas en un nuevo conjunto de variables no correlacionadas. Los coeficientes de la regresión están derivadas por la relación directa entre las variables de entrada con la variable de salida.

La regresión por PLS permite crear un modelo donde se tienen distintas variables dependientes así como variables independientes. Considerando la multicolinealidad, el modelo es capaz de crear nuevos factores los cuales no se encuentran correlacionados y ayuda a crear un conjunto robusto a pesar de datos faltantes y con ruido. Al considerar la variable de salida cuando se crean los factores de X la predicción suele ser más acertada sin mucho riesgo a sobreajuste. Debido a que los factores son combinaciones lineales, es difícil interpretarse y no se puede determinar la importancia de cada variable[36].

[36] D. M. Pirouz. An overview of partial least squares. *SSRN Electronic Journal*, 10 2006. DOI: 10.2139/ssrn.1631359

La regresión con PLS juega un papel importante para eliminar la multicolinealidad de las variables independientes. Además, es posible utilizar la regresión con PLS para disminuir el número de variables de entrada al modelo así reduciendo el tiempo de ejecución del mismo. Al igual que con los modelos de regresión lineal múltiple, se dividieron los datos por estaciones del año y al final se hizo una comparación con el modelo con todos los datos.

Los modelos utilizados tienen cinco variables de entrada, teniendo en consideración que al utilizar un análisis de PLS se pueden crear componentes y debido a que para determinar el mejor número de componentes es un proceso de prueba y error, para todas las estaciones se realizó un diagrama de codo los cuales utilizaron la métrica MSE. El diagrama de codo tiene como objetivo el crear una regresión con PLS y considerar su error MSE.

Los datos de verano mostraron que el error se estabiliza entre tres y cuatro componentes. Utilizar cuatro componentes no muestra mucha

diferencia que el usar tres, por lo que se decidió utilizar solo tres componentes para la regresión (Figura 3.17).

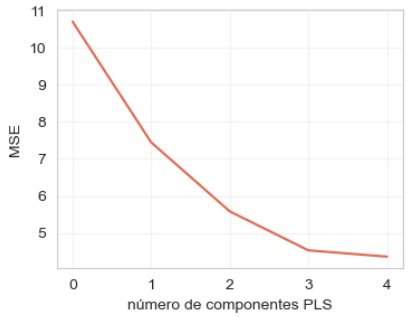

Figura 3.17: Diagrama de codo para componentes PLS para datos de verano.

Posteriormente, después de elegir el número de componentes, se realizó la regresión con los tres componentes creados por la combinación lineal (Figura 3.18).

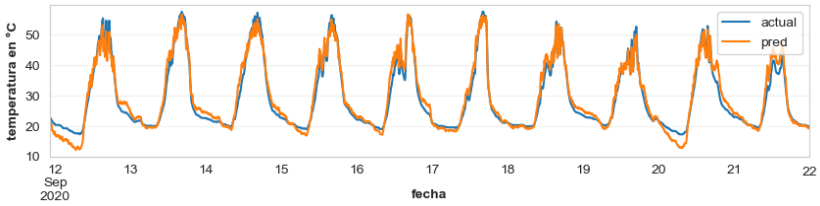

Figura 3.18: Pronóstico de temperatura interna con el uso de una regresión PLS para los datos de verano.

En relación a los errores de este pronóstico se encontró $RMSE$: 2.2454, R^2: 0.9578, MAE: 1.7678 y $MAPE$: 0.0643. Los resultados son buenos para un acercamiento del comportamiento con solo tres componentes, a comparación del modelo de regresión lineal los resultados son muy similares, sin embargo es notable la mejora de los errores en todas las métricas por muy pocos puntos porcentuales.

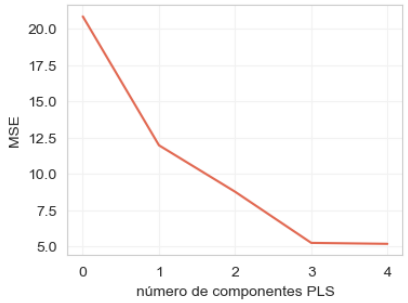

Figura 3.19: Diagrama de codo para componentes PLS para datos de otoño.

La gráfica de codo utilizada para los datos de otoño (Figura 3.19) igualmente mostró poca diferencia en MSE al utilizar cuatro componentes, por lo que se decide utilizar tres de igual manera. El pronóstico de la regresión con PLS entonces es generada con sus tres componentes (Figura 3.20).

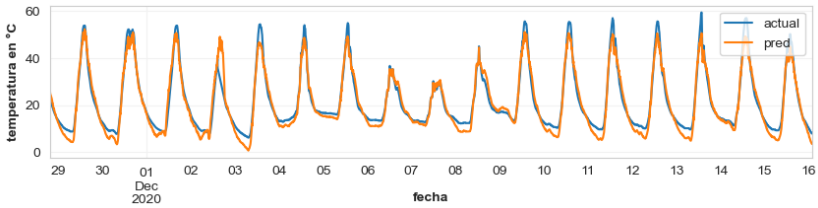

Figura 3.20: Pronóstico de temperatura interna con el uso de una regresión PLS para los datos de otoño.

Los errores del pronóstico fueron $RMSE$: 3.0444, R^2: 0.9417, MAE: 2.3403 y $MAPE$: 0.1383. Ambos modelos lineales (regresión lineal múltiple y regresión con PLS) tienen problemas para el pronóstico de los datos de otoño. De igual manera que el modelo de regresión lineal múltiple, el modelo con PLS tuvo dificultad con las temperaturas menores a $10°C$ y con las temperaturas mayores a $50°C$. Este modelo PLS obtuvo peores resultados que su contraparte, a excepción de la R^2 y MAPE donde tuvo dos más puntos porcentuales y casi 1 punto porcentual menos, respectivamente.

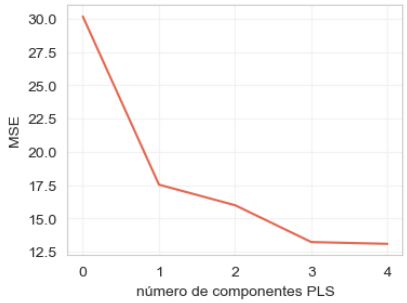

Figura 3.21: Diagrama de codo para componentes PLS para datos de invierno.

Al igual que las demás estaciones del año, se genera una gráfica de codo para determinar el número de componentes principales que se estarán utilizando para la regresión (Figura 3.21). Los resultados mostrados en la gráfica de codo indican que se puede utilizar tres componentes para su regresión por lo que se genera la regresión con esta información (Figura 3.22).

Figura 3.22: Pronóstico de temperatura interna con el uso de una regresión PLS para los datos de invierno.

El pronóstico para la regresión con PLS parece tener los mismos problemas que el modelo con la regresión lineal múltiple. Los datos a tener muchos picos iniciando desde $0°C$ hasta casi $70°C$ tiene problemas

en crear una regresión que pueda cubrir todos los picos de los datos. Los errores del pronósticos $RMSE$: 4.7876, R^2: 0.9199, MAE: 3.4890 y $MAPE$: 0.1321 fueron peores a comparación de la regresión lineal.

Para la última estación, en primavera también se considera un diagrama de codo el cual al igual que todas las estaciones indica que el mejor número de componentes es tres (Figura 3.23). Posteriormente se vuelve a generar la regresión con PLS (Figura 3.24).

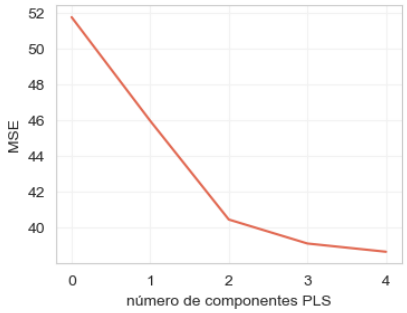

Figura 3.23: Diagrama de codo para componentes PLS para datos de primavera.

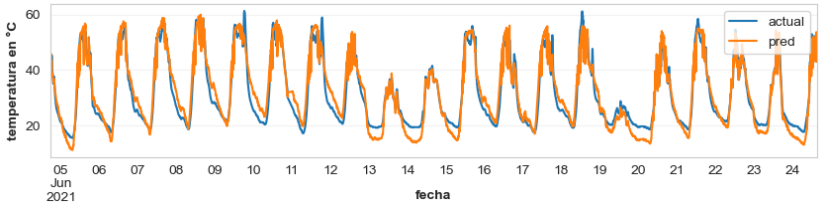

Figura 3.24: Pronóstico de temperatura interna con el uso de una regresión PLS para los datos de primavera.

Los errores para el pronóstico de primavera $RMSE$: 3.5875, R^2: 0.9096, MAE: 2.9543 y $MAPE$: 0.1057 fue mucho peor para RMSE pero fueron muy similares para las demás métricas.

La regresión con PLS tiene la facilidad de poder utilizar todos los predictores con combinaciones lineales de manera que es posible reducir la multicolinealidad existente entre las variables de entrada. Ambos modelos fueron utilizados como primer acercamiento, por su simplicidad, factibilidad, solución a la multicolinealidad, determinación de número de variables de entrada. Desafortunadamente y a pesar de que los resultados preliminares de estos modelos no son del todo malos, cuando se busca hacer una predicción en futuro, es necesario hacer un ajuste en los datos de entrada y al obtener errores como se obtuvieron en estos modelos lineales, es necesario hacer una búsqueda de otro modelo que sea capaz de capturar la tendencia de los datos para el pronóstico de la temperatura interna actual de manera casi perfecta para próximamente utilizarlo en su pronóstico futuro. Uno de los modelos que permite el uso de datos no lineales es el algoritmo de máquinas de vector de soporte para regresión (SVR)[37].

[37] J. a. K Suykens. Support vector machines: A nonlinear modelling and control perspective. *European Journal of Control*, 7:311–327, 2001. DOI: 10.3166/ejc.7.311-327

3.8 Máquina de Vector de Soporte para regresión

La máquina de vector de soporte (SVM, por sus siglas en inglés) es un algoritmo que posee capacidades de aprender patrones en los datos para así generar una predicción adecuada. SVM fue originalmente desarrollado para el reconocimiento de similitudes; para esto se utilizan las observaciones para crear una frontera de decisión utilizando vectores de soporte. SVM tiene resultados reproducibles y con una precisión balanceada, debido a ser un modelo convexo. SVM busca un hiperplano óptimo que permita dividir las observaciones en una clase utilizando patrones de información sobre las mismas[38].

El algoritmo de SVM fue primero propuesto por Vapnik en 1992 en su artículo *A training algorithm for optimal margin classifiers* en los laboratorios de AT&T. Inicialmente se utilizó un método para construir un hiperplano capaz de separar los datos en N-dimensiones[39], posteriormente entre 1992 y 1995 se utilizó la construcción del hiperplano para la construcción de funciones no lineales[40] y en 1997 se utilizó para construir funciones de valores reales[41].

Entre las grandes innovaciones del uso de SVM fue el uso explícito de optimización convexa, la teoría de aprendizaje estadístico y funciones de *kernel*. Para construir el hiperplano que separe los datos se necesita evaluar los productos punto entre dos vectores de los datos de entrenamiento. Como en el espacio de Hilbert la forma general de un producto punto tiene una representación de *kernel*, la evaluación del producto punto no solo depende de la dimensión del espacio.

El SVM produce funciones de predicción que se expanden en un subconjunto de vectores de soporte. SVM puede generalizar estructuras complicadas con sólo unos pocos vectores de soporte lo que permite también ser utilizado en el área de compresión de imágenes[42]. SVM puede dividirse en dos grandes categorías, clasificación *Support Vector Classification* (SVC) y regresión *Support Vector Regression* (SVR). El método utilizado para este análisis fue SVR.

El modelo SVC solo depende de un subconjunto de datos de entrenamiento, porque la función de costo para el desarrollo del modelo no considera puntos de entrenamiento fuera del margen; el modelo SVR también depende de un subconjunto de datos de entrenamiento debido a que la función de costo para el desarrollo del modelo ignora cualquier dato de entrenamiento que se encuentre dentro de la predicción del modelo considerando un umbral ϵ (Figura 3.25).

La idea básica de SVR es que el vector de datos x sea transformado a una dimensión mayor F con el uso de una transformación no lineal ψ, y después la regresión lineal se realiza de la forma:

$$f(x) = (w \cdot \Phi(x)) + b(\Phi : R^n \rightarrow F, w \in F), \tag{3.13}$$

[38] D. A. Pisner and D. M. Schnyer. Chapter 6 - support vector machine. In Andrea Mechelli and Sandra Vieira, editors, *Machine Learning*, pages 101–121. Academic Press, 2020. ISBN 978-0-12-815739-8. DOI: https://doi.org/10.1016/B978-0-12-815739-8.00006-7. URL https://www.sciencedirect.com/science/article/pii/B9780128157398000067

[39] V. Vapnik and A. Chervonenkis. A note on one class of perceptrons. *Automation and Remote Control*, 25, 1964

[40] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20: 273–297, Sep 1995. ISSN 1573-0565. DOI: 10.1007/BF00994018. URL https://doi.org/10.1007/BF00994018

[41] V. Vapnik, S. Golowich, and A. Smola. Support vector method for function approximation, regression estimation and signal processing. In *Proceedings of the 9th International Conference on Neural Information Processing Systems*, page 281–287. MIT Press, 1996

[42] R. Jiao, Y. Li, Q. Wang, and B. Li. Svm regression and its application to image compression. In H. De-Shuang, Z. Xiao-Ping, and H. Guang-Bin, editors, *Advances in Intelligent Computing*, pages 747–756. Springer Berlin Heidelberg, 2005. ISBN 978-3-540-31902-3

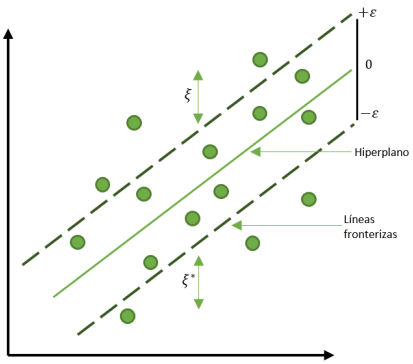

Figura 3.25: Umbral ϵ de SVR.

donde b es el valor del umbral. La regresión resultante en un espacio de características de alta dimensión corresponde a una regresión no lineal en el espacio de entrada de baja dimensión, por lo que se evita el calculo del producto punto de $w, \Phi(x)$, en un espacio de alta dimensión. Debido a que Φ es un mapa, el valor de w puede ser obtenido con los datos minimizando la suma del riesgo empírico R_{emp} y un término de complejidad $\| w \|^2$ que impone la planitud en el espacio de características. Esto se vuelve un problema de optimización con restricciones, las cuales aplicando multiplicadores de Lagrange queda de la forma,

$$R(w) = R_{emp} + \lambda \| w \|^2 = \sum_{i=1}^{l} e(f(x_i) - y_i) + \lambda \| w \|^2, \tag{3.14}$$

donde l es el número de ejemplos, λ es un término de regularización, y $e(.)$ es una función de costo. La función de costo $e(.)$ se puede representar de la forma[43,44]:

(1) Función de costo lineal ϵ-insensible:

$$e(f(x) - y) = max(0, \mid f(x) - y \mid -\epsilon) \tag{3.15}$$

(2) Función de costo cuadrática:

$$e(f(x) - y) = (f(x) - y)^2 \tag{3.16}$$

(3) Función de costo Huber:

$$e(f(x) - y) = \begin{cases} \mu \mid f(x) - y \mid -\frac{\mu 2}{2}, & \text{si } \mid (fx) - y \mid > \mu. \\ \frac{1}{2} \mid f(x) - y \mid^2, & \text{otro.} \end{cases} \tag{3.17}$$

Se busca minimizar $R(w)$ de (3.14), por lo que se tiene que obtener $\alpha_i - \alpha_i^*$ quedando de la forma,

$$w = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) \Phi(x_i), \tag{3.18}$$

donde $\alpha_i - \alpha_i^*$ son la solución que minimiza $R(w)$. Resolviendo para α_1 y α_i^* del problema de optimización, los valores para x_i por cada valor

[43] V. Vapnik, S. Golowich, and A. Smola. Support vector method for function approximation, regression estimation and signal processing. In *Proceedings of the 9th International Conference on Neural Information Processing Systems*, page 281–287. MIT Press, 1996

[44] V. N. Vapnik. *The Nature of Statistical Learning Theory*, chapter Methods of Function Estimation, pages 181–224. Springer New York, 2000. ISBN 978-1-4757-3264-1

respectivo de α_i, α_i^* diferente de cero se les denominan (3.13) se puede reescribir como:

$$\begin{aligned} f(x) &= \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)(\Phi(x_i)\Phi(x)) + b \\ &= \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)K(x_i, x) + b \end{aligned} \tag{3.19}$$

donde $K(x_i, x) + b = \Phi(x_i) \cdot \Phi(x)$ se denomina como la función *kernel*, que es cualquier función de *kernel* simétrico que cumple las condiciones de Mercer y corresponde a un producto punto en un espacio de características. b se puede obtener eligiendo un punto en el margen utilizando la nueva ecuación reescrita (3.19). Usualmente se recomienda tomar el promedio de todos los puntos en el margen,

$$b = \text{average}_k\{\delta_k + y_k - \sum_{i=1}^{l}(\alpha_i - \alpha_{i_i}^*)K(x_i, x_k)\} \tag{3.20}$$

donde δ_k es un error de predicción para la función de costo ϵ-insensible $\delta_k = \epsilon\text{sign}(\alpha_k - \alpha_k^*)$ y para la función de costo de Huber $\delta_k = (1/C)(\alpha_k - \alpha_k^*)$.

Una desventaja muy marcada de SVM es la selección de los parámetros, éstos dependen de la transformación de *kernel* que se va a utilizar así como los del mismo algoritmo; la complejidad afecta el tiempo de entrenamiento cuando se tienen conjuntos con muchos datos. Para algunas bases de datos, el rendimiento del SVM es muy sensible a la selección de los parámetros de costo y *kernel*.

La estrategia del *kernel* es tomar los datos de entrada y transformarlos en la forma requerida para su procesamiento. Este truco permite al algoritmo de SVM proyectar los datos de una baja dimensión a un espacio de alta dimensión. Una función de *kernel* K puede ser definida por un producto punto en el espacio de Hilbert quedando de la forma,

$$\langle x_1 \cdot x_2\rangle \leftarrow K(x_1, x_2) = \langle\Phi(x_1) \cdot \Phi(x_2)\rangle \tag{3.21}$$

dado en un mapa de características $\phi : \mathbb{R}^d \to V$.

El uso del truco *kernel* en la ecuación (3.19) equivale al mapeo de los datos en un espacio de características V que permite al algoritmo de SVM computar y resolver problemas donde los datos no son linealmente separables. Es importante notar que el problema de programación cuadrática requerida para encontrar el hiperplano óptimo es convexo solo si la función *kernel* cumple con las condiciones de Mercer[45], por lo que el kernel debe de satisfacer $K : S \times S \to \mathbb{R}$:

$$\int_S \int_S g(x)k(x, x')g(x')dxdx' \geq 0 \tag{3.22}$$

para cada función integral cuadrada $g(x)$.

[45] J. Shawe-Taylor, University of Southampton, N. Cristianini, and University of California. *Kernel Methods for Pattern Analysis*, pages 47–84. Cambridge University Press, 2011. ISBN 9780511809682. DOI: 10.1017/CBO9780511809682

Si k satisface la ecuación (3.22), entonces la matriz M, donde $m_{ij} = k(x_i, x_j), \forall x_1, ..., x_n \in S$ y $\forall n \in \mathbb{N}$, es
(1) simétrica, $(M = M^T)$ y
(2) matriz positiva semi-definida (PSD)

Una matriz es positiva semi-definida si $\mathbf{u}M\mathbf{u}^T \geq 0$ para cada uno de los vectores reales $\mathbf{u} \in \mathbb{R}^n$, en otras palabras, que todos los valores propios son no-negativos.

Considerando la efectividad del *kernel* para datos no lineales, la elección del *kernel* apropiado se vuelve vital y no trivial.Esto quiere decir que también se vuelve un parámetro más a considerar como problema de optimización convexa dentro de las condiciones generales. Una función óptima de *kernel* puede ser elegida de un conjunto fijo de *kernel* de forma estadísticamente rigurosa usando validación cruzada, sin embargo es importante notar que este método consume mucho tiempo y no garantiza que algún *kernel* no elegido genere mejores resultados.

En términos generales, algunas funciones *kernel* más populares son:
(1) Kernel Lineal:

$$K(x_i, x_k) = x_i^T x_j \tag{3.23}$$

(2) Kernel Polinomial:

$$K(x_i, x_k) = (\gamma x_i^T x_j + r)^d, \quad \gamma > 0 \tag{3.24}$$

(3) Kernel Función de base radial (RBF):

$$K(x_i, x_k) = exp(-\gamma \parallel x_i - x_j \parallel^2), \quad \gamma > 0 \tag{3.25}$$

(4) Kernel Sigmoidal:

$$K(x_i, x_k) = tanh(\gamma x_i^T x_j + r) \tag{3.26}$$

El agregar alguna de estas funciones significa un aumento en los parámetros a considerar al entrenar el algoritmo. γ, r y d se vuelven estos nuevos parámetros. Entre los más utilizados, el RBF es mucho más utilizado para algoritmos SVM[46,47] debido a:

- El *kernel* RBF tiene menor número de parámetros que el polinomial.
- El *kernel* RBF mapea de manera no lineal muestras en una alta dimensión a diferencia de uno lineal
- El *kernel* RBF tiene menor dificultades numéricas

El *kernel* es uno de los hiper parámetros más importantes a la hora de realizar un modelo de SVM; tomando en consideración los resultados de los modelos lineales, se concluyó que el *kernel* lineal no es apto, entonces se utilizaron RBF y polinomial. El modelo utilizó los datos de

[46] C. X. Xue, R. S. Zhang, H. X. Liu, M. C. Liu, Z. D. Hu, and B. T. Fan. Support vector machines-based quantitative structure-property relationship for the prediction of heat capacity. *Journal of Chemical Information and Computer Sciences*, 44:1267–1274, 2004. DOI: 10.1021/ci049934n. URL https://doi.org/10.1021/ci049934n

[47] H. Liu, X. Yao, C. Xue, R. Zhang, M. Liu, Z. Hu, and B. Fan. Study of quantitative structure-mobility relationship of the peptides based on the structural descriptors and support vector machines. *Analytica Chimica Acta*, 542:249–259, 2005. ISSN 0003-2670. DOI: https://doi.org/10.1016/j.aca.2005.04.006. URL https://www.sciencedirect.com/science/article/pii/S000326700500591X

temperatura interna actual para determinar el poder de predicción de comportamiento con SVR.

Utilizando la misma división de datos de 80 % para entrenamiento y 20 % para prueba, se realizaron dos modelos de SVR con ambos *kernel*. Los hiper parámetros usados fueron determinados con el uso de una malla de hiper parámetros y una búsqueda exhaustiva.

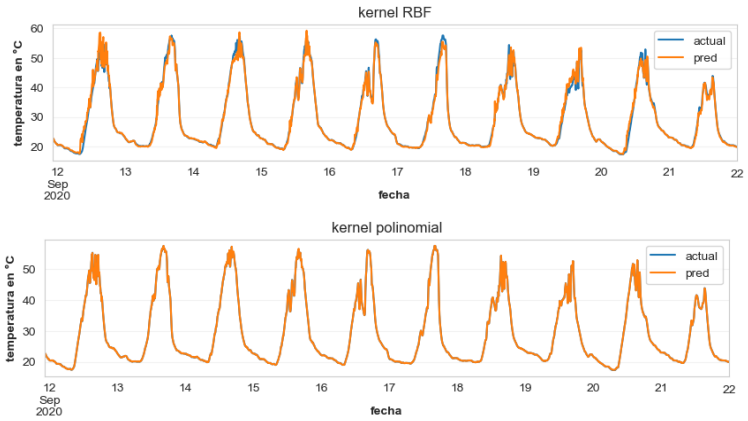

Figura 3.26: Predicción de temperatura interna con el uso de SVR RBF y polinomial para los datos de verano.

La predicción para los datos de verano con el modelo SVR (Figura 3.26) resultó mejor a diferencia de los resultados con los modelos lineales; para los datos de verano destacó una casi perfecta predicción en *kernel* RBF con resultados de $RMSE$: 1.0977, R^2: 0.9899, MAE: 0.5755 y $MAPE$: 0.0170 y para polinomial $RMSE$: 0.0549, R^2: 0.9999, MAE: 0.04222 y $MAPE$: 0.0015.

Los modelos SVR que fueron utilizados para los datos de otoño (Figura 3.27) también mostraron resultados excelentes a comparación de los modelos no lineales (Figura 3.27); estos modelos presentaron resultados buenos $RMSE$: 2.1047, R^2: 0.9721, MAE: 0.8269 y $MAPE$: 0.0491 para *kernel* RBF y $RMSE$: 1.7431, R^2: 0.9808, MAE: 0.2856 y $MAPE$: 0.0124 para el *kernel* polinomial.

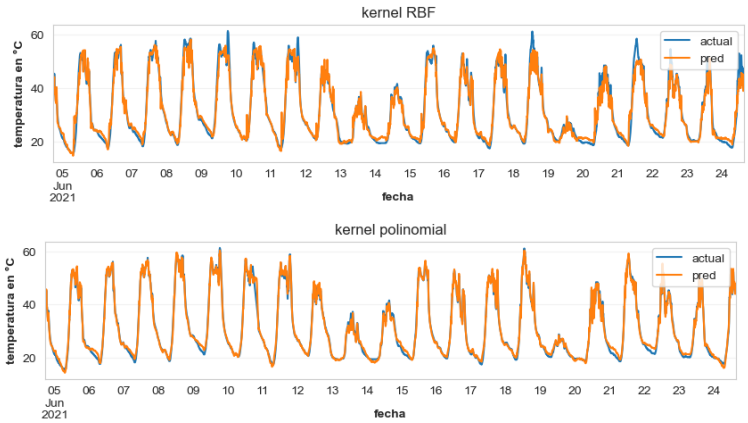

Figura 3.27: Pronóstico de temperatura interna con el uso de SVR RBF y polinomial para los datos de otoño.

El uso del modelo SVR con *kernel* RBF para los datos de invierno presentó dificultades con temperaturas altas (Figura 3.28) en donde se puede notar que cuando la temperatura interna del invernadero se eleva hasta $50°C$ el modelo no es capaz de captar el comportamiento, sin embargo el modelo con *kernel* polinomial fue apto para capturar los picos obteniendo resultados buenos con $RMSE$: 0.6760, R^2: 0.9984, MAE: 0.2929 y $MAPE$: 0.0123 mientras que en RBF: $RMSE$: 4.171, R^2: 0.9392, MAE: 2.6414 y $MAPE$: 0.1096.

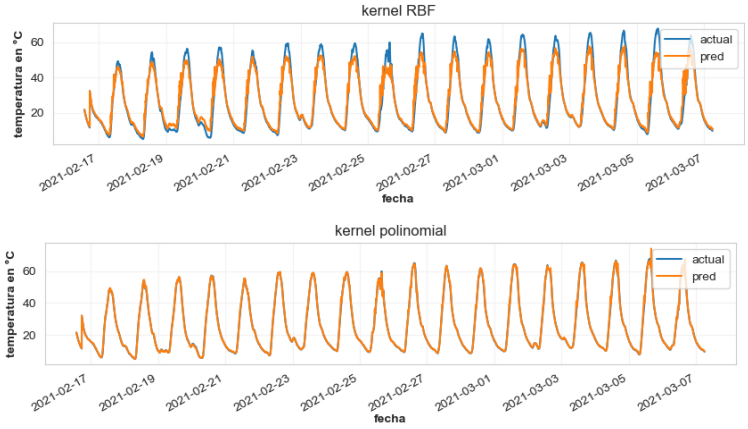

Figura 3.28: Predicción de temperatura interna con el uso de SVR RBF y polinomial para los datos de invierno.

Por último, la predicción de los modelos para los datos de primavera (Figura 3.29) es mucho mejor que cualquier modelo lineal. El modelo con *kernel* polinomial nuevamente obtuvo una mejor predicción de comportamiento con $RMSE$: 1.0360, R^2: 0.9924, MAE: 0.7443 y $MAPE$: 0.0270 a comparación del RBF $RMSE$: 2.2533, R^2: 0.9643, MAE: 1.5117 y $MAPE$: 0.0485, sin embargo los resultados siguen siendo superiores a los modelos antes presentados.

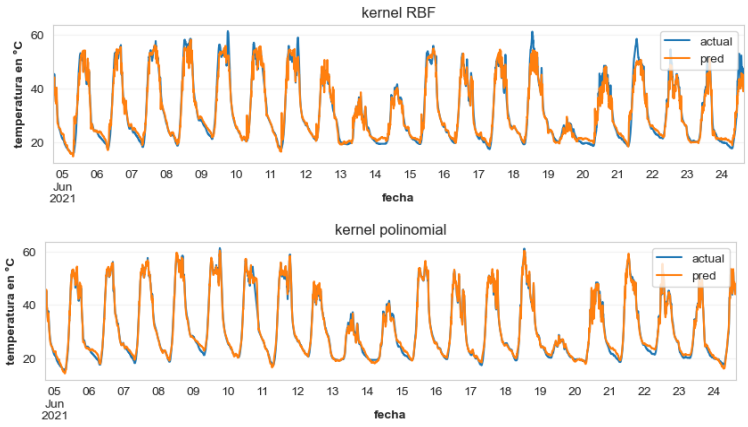

Figura 3.29: Predicción de temperatura interna con el uso de SVR RBF y polinomial para los datos de primavera.

Como parte del proceso de elección de un modelo capaz de predecir el comportamiento de los datos de la temperatura interna

del invernadero, se consideró el algoritmo de aumento extremo del gradiente (XGBoost, por sus siglas en inglés). XGBoost es un modelo de ensamble con *boosting* basado en árboles de decisión; si bien se utilizan más para problemas de clasificación, los árboles de decisión también pueden predecir valores continuos y resolver problemas de regresión. Wentao Cai, et al.[48] utilizaron varios modelos incluyendo XGBoost para la predicción de la temperatura interna con buenos resultados. XGBoost fue utilizado debido a su flexibilidad al tratar discontinuidad en datos, valores nulos, rendimiento, uso de recursos, entre otros beneficios. A continuación se hizo una introducción del algoritmo de XGBoost así como su implementación para la predicción de este análisis.

[48] W. Cai, R. Wei, L. Xu, and X. Ding. A method for modelling greenhouse temperature using gradient boost decision tree. *Information Processing in Agriculture*, 9:343–354, 2022. ISSN 2214-3173. DOI: https://doi.org/10.1016/j.inpa.2021.08.004. URL https://www.sciencedirect.com/science/article/pii/S2214317321000743

3.9 XGBoost

El algoritmo de XGBoost nació de un proyecto de investigación por Tianqi Chen[49] en 2014, donde buscaba alternativas de *boosting* para árboles de decisión. La primer versión del algoritmo fue escrita en un sistema Linux con el fin de poder utilizar todos los recursos de una manera más sencilla y eficiente. Tianqi, entonces, decidió utilizar el modelo en retos de *Kaggle*[50], donde el reto fue la identificación del Bosón de Higgs, el cual obtuvo el primer lugar. Debido a su popularidad, Tianqin junto con Bing Xu decidieron implementarlo como paquetería en Python para posteriormente agregarlo como paquetería en el lenguaje de programación R con la ayuda de Tong He.

[49] T. Chen. Story and lessons behind the evolution of xgboost, 2014. URL https://sites.google.com/site/nttrungmtwiki/home/it/data-science---python/xgboost/story-and-lessons-behind-the-evolution-of-xgboost. Accesado en: 2023-03-29

[50] URL https://www.kaggle.com/competitions/higgs-boson

XGBoost fue diseñado principalmente para ser un paquete cerrado que tomaba datos de entrada y resultaba un modelo de predicción, sin embargo aprovechando su integración con los sistemas de interfaz común, XGBoost se convirtió a lo que hoy es; una paquetería capaz de utilizar los recursos computacionales para mejorar su rendimiento así como para producir una predicción adecuada en menor tiempo computacional.

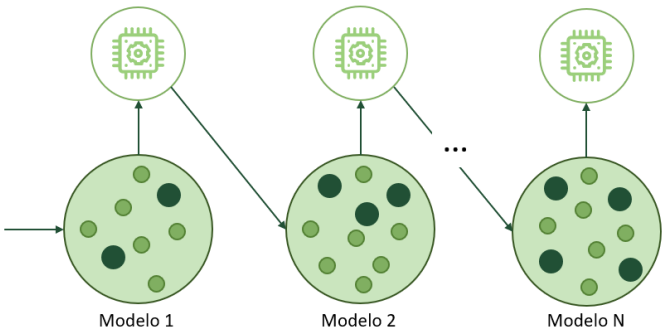

Figura 3.30: Algoritmo *boosting* secuencial.

El *boosting* es un método de ensamble el cual combina diferentes modelos débiles en un modelo fuerte para minimizar errores de

entrenamiento (Figura 3.30). Usualmente los modelos débiles tienden al sobre ajuste por lo que tener varios modelos aditivos permite crear un mejor modelo. En el *boosting*, se selecciona una muestra aleatoria de datos, se aplica un modelo y se van entrenando secuencialmente, en donde cada modelo intenta compensar los puntos débiles del anterior.

XGBoost es una alternativa al algoritmo[51] de *boosting* de refuerzo de gradiente (*Gradient Boosting* en inglés) el cual funciona añadiendo secuencialmente predictores a un conjunto y cada uno corrige los errores del anterior, *gradient boosting* se basa en los errores residuales del predecesor. *Gradient boosting* corresponde a la combinación del descenso de gradiente y el método de *boosting*.

[51] IBM. What is boosting?, 2022. URL https://www.ibm.com/topics/boosting. Accesado en: 2023-03-29

XGBoost utiliza la existente idea de *gradient boosting*, sin embargo tiene algunas modificaciones[52] las cuales son mejorías mínimas en la función de objetivo de regularización. Considerando una base de datos D con $\boldsymbol{m}$ características y $\boldsymbol{n}$ número de ejemplos $D = \{(x_i, y_i)\}(D = n_i \in \mathbb{R}^m, y_i \in \mathbb{R}^n)$, un modelo de ensamble de árboles que usa K funciones aditivas para la predicción de la salida y sea $\hat{y}_i$ la predicción de salida del modelo.

[52] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. page 785–794. Association for Computing Machinery, 2016a. ISBN 9781450342322. DOI: 10.1145/2939672.2939785

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F}, \tag{3.27}$$

donde $\mathcal{F} = \{f(x) = w_{q(x)}\}(q : \mathbb{R}^m \to T, w \in \mathbb{R}^T)$ es el espacio de árboles de regresión y clasificación (CART, por sus siglas en inglés). Aquí q representa la estructura de cada árbol que asigna un ejemplo al índice de la hoja correspondiente. T es el número de hojas en cada árbol. Cada f_k corresponde a una estructura independiente de árbol q y pesos de hojas w. K representa el número de árboles en el modelo utilizados. Para resolver la ecuación (3.27) es necesario encontrar el mejor conjunto de funciones minimizando la función de regularización y costo (3.28).

$$\begin{aligned} \mathcal{L}(\phi) &= \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \\ \textit{donde} \quad \Omega(f) &= \gamma T + \frac{1}{2}\lambda \parallel w \parallel^2 \end{aligned} \tag{3.28}$$

Aquí l representa la función de costo la cual es la diferencia entre la predicción $\hat{y}_i$ y el objetivo y_i.Ω penaliza la complejidad del modelo, el cual contribuye en evitar el sobre ajuste del modelo suavizando los pesos finales aprendidos. Cuando el término de regularización es cero, la función de objetivo regresa al tradicional *gradient tree boosting*.

El modelo ensamble de árboles en (3.27) incluye funciones como parámetros y no puede ser optimizada con métodos tradicionales de optimización en un espacio Euclidiano. Por lo que es entrenado de manera aditiva agregando una nueva función f conforme el modelo

sigue entrenándose. Entonces en la iteración t-ésima se añade una nueva función (árbol):

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{3.29}$$

Entonces utilizando una aproximación de segundo orden para optimizar el objetivo de forma general donde $g_i = \delta_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$ y $h_i = \delta^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$ y eliminando términos constantes se puede obtener la siguiente función objetivo simple en el paso t.

$$\tilde{\mathcal{L}}^{(t)} = \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \tag{3.30}$$

Definiendo $I_j = \{i|q(x_i) = j\}$ como una instancia de hoja j. Se puede reescribir la ecuación (3.30) como,

$$\tilde{\mathcal{L}}^{(t)} = \sum_{j=1}^{T} [(\sum_{i \in I_j} g_i) w_j + \frac{1}{2} (\sum_{i \in I_j} h_i + \lambda) w_j^2] + \gamma T \tag{3.31}$$

Para una estructura fija de $q(x)$ se obtener el valor óptimo del peso w_j^* de la hoja j con,

$$\tilde{\mathcal{L}}^{(t)}(q) = -\frac{1}{2} \sum_{j=1}^{T} \frac{(\sum_{i \in I_j} g_i)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T \tag{3.32}$$

Usualmente es imposible enumerar todas las posibles estructuras de árbol q. Por lo que un algoritmo codicioso inicia desde una sola hoja e iterativamente agrega ramas al árbol. Entonces asumiendo que I_L y I_R son conjuntos de las instancias de los nodos izquierdos y derechos después de una división y $I = I_L \cup I_R$ entonces la función de costo después de la división es,

$$\mathcal{L}_{split} = \frac{1}{2} [\frac{(\sum_{i \in I_L} g_i)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{(\sum_{i \in I_R} g_i)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{(\sum_{i \in I} g_i)^2}{\sum_{i \in I} h_i + \lambda}] - \gamma \tag{3.33}$$

Uno de los problemas en el aprendizaje de árboles es el encontrar la mejor división, por lo que el algoritmo de XGboost enumera todas las posibles separaciones para características continuas. Para realizarlo eficientemente, el algoritmo primero ordena los datos de acuerdo a los valores de la característica y visita los datos ordenados para acumular las estadísticas del gradiente para la puntuación en (3.33).

Cuando no se tiene suficiente espacio en memoria para los datos, XGBoost utiliza una estrategia de algoritmo codicioso aproximado, el cual divide la base de datos en diferentes cuantiles ponderados, en donde se considera la división para la puntuación de los nodos. Además, XGBoost es capaz de ejecutar un aprendizaje en paralelo en

donde permite utilizar diferentes porciones o también llamados bloques (*blocks*, en inglés) de datos en diferentes computadoras para ejecutar el algoritmo y así reducir el costo del ordenado.

En muchos fenómenos, es bastante común que la entrada x sea dispersa. Esto quiere decir que x puede tener presencia de valores perdidos en los datos, valores nulos, entradas nulas frecuentes en las estadísticas y artefactos de la codificación de un solo punto (*one-hot encoding*, por su término en inglés). Es por eso que el algoritmo tiene que ser capaz de la detección de los datos dispersos.

Mientras que la estructura de *blocks* ayuda a optimizar la complexidad computacional de la búsqueda de la división de nodos, el nuevo algoritmo requiere una búsqueda indirecta de gradientes y Hessianos por cada renglón, debido a que estos valores son accedidos por característica en orden. XGBoost lidia con estos problemas utilizando la memoria caché del procesador. XGBoost coloca los gradientes y Hessianos en la memoria caché para que así pueda calcular los puntajes de similitud y valores de salida, así mejorando el tiempo computacional de entrenamiento del modelo.

XGBoost fue un algoritmo con resultados buenos para todas las estaciones del año. Para los modelos presentados, se dividieron los datos en 80 % entrenamiento y 20 % de prueba y para la búsqueda de hiper parámetros se utilizó una malla con una búsqueda exhaustiva. El modelo para los datos de verano obtuvo errores $RMSE$: 0.2764, R^2: 0.9994, MAE: 0.1943 y $MAPE$: 0.01. Estos errores indican una buena aproximación al comportamiento de la temperatura interna del invernadero, además para el modelo con datos de otoño los errores $RMSE$: 2.3534, R^2: 0.9652, MAE: 0.9566 y $MAPE$: 0.0550 (Figura 3.31).

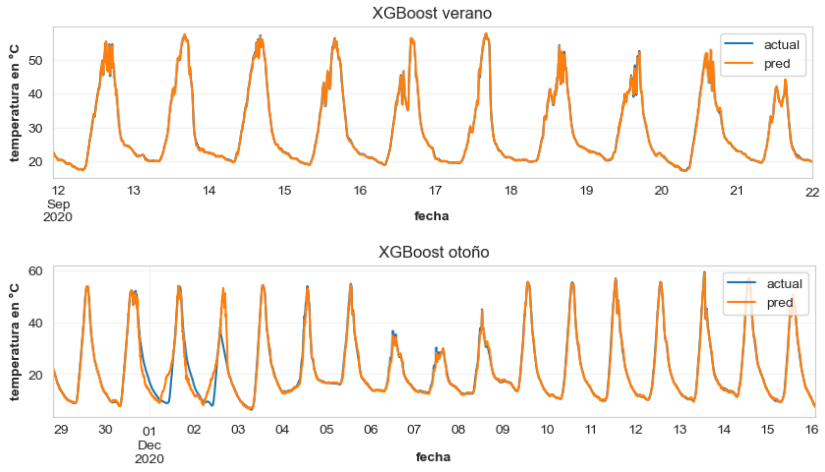

Figura 3.31: Predicción de comportamiento de temperatura interna con el uso de XGBoost para datos de verano y otoño.

De igual manera, los modelos para datos de invierno y primavera (Figura 3.32) presentaron resultados buenos para la predicción del

comportamiento de la temperatura interna, los resultados para el modelo de invierno fueron $RMSE$: 0.5514, R^2: 0.9989, MAE: 0.3910 y $MAPE$: 0.0207 y para el modelo de primavera $RMSE$: 1.5701, R^2: 0.9830, MAE: 1.0500 y $MAPE$: 0.0350.

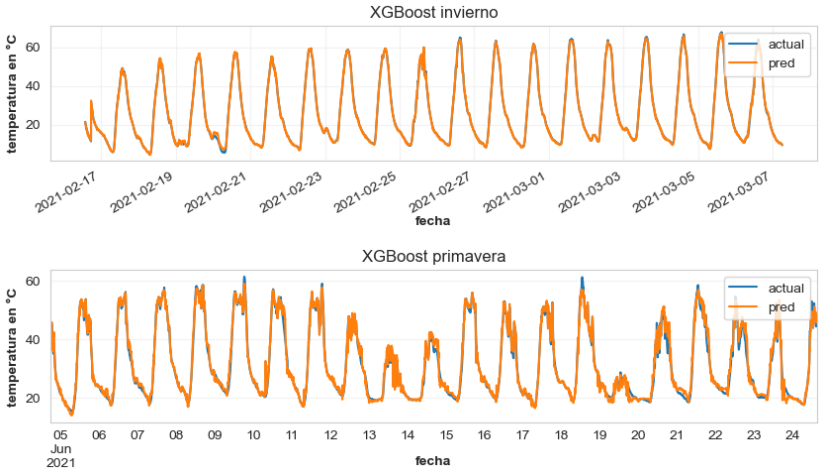

Figura 3.32: Predicción de comportamiento de temperatura interna con el uso de XGBoost para datos de invierno y primavera.

3.10 *Candidatos de modelos para pronóstico*

Los modelos utilizados en este análisis presentaron aproximaciones del comportamiento de la temperatura interna del invernadero. Un modelo lineal no es capaz de capturar el comportamiento de la temperatura interna, esto se debe principalmente a la estructura de los datos de la serie de tiempo, los cuales usualmente son no lineales. La regresión lineal supone que la relación entre las variables es constante a lo largo del tiempo, por lo que para pronóstico de temperatura muy corta puede ser utilizada, sin embargo para este análisis se busca hacer un pronóstico de mínimo 30 minutos.

Los modelos SVR y XGBoost presentaron mejores acercamientos y aproximaciones al comportamiento de la temperatura interna para la ejecución de un pronóstico a largo plazo. Ambos algoritmos fueron diseñados para tratar con datos no lineales, así como para adaptarse a patrones cambiantes como pueden ser las variables climáticas a lo largo del tiempo. El pronóstico de la temperatura se basa en considerar las variables climáticas de los datos para realizar una predicción aproximada.

4 Resultados

Los modelos que presentaron mejores resultados para el pronóstico de la temperatura interna fueron XGBoost y SVR. Para el pronóstico se dividieron los datos por estación del año y se realizó un pronóstico de 30 minutos, 45 minutos y 1 hora, esto debido al alcance de ambos modelos y el objetivo de mantener histéresis. Se utilizaron dos diferentes combinaciones *Hi-Id-To* y *Hi-Id-Rs* generando un total de 72 modelos a comparar y optimizar.

Considerando las dos diferentes combinaciones posibles, las estaciones del año así como los pronósticos posibles por cada estación, se realizaron 3 modelos por cada uno, de los cuales inicialmente se encontró que la combinación *Hi-Id-Rs* obtuvo mejores resultados de pronósticos a comparación de *Hi-Id-To* (Tabla 4.1).

Modelo	**Combinación**	**R^2**	**RMSE**	**MAE**	**MAPE**
SVR Radial	*Hi-Id-To*	0.9786	2.5379	1.8165	0.0641
SVR Radial	***Hi-Id-Rs***	**0.9871**	**1.9678**	**1.1558**	**0.0353**
SVR Poly	*Hi-Id-To*	0.9785	2.5418	1.8095	0.0642
SVR Poly	***Hi-Id-Rs***	**0.9764**	**2.6619**	**1.6548**	**0.0511**
XGBoost	*Hi-Id-To*	0.9779	2.5792	1.7505	0.0603
XGBoost	***Hi-Id-Rs***	**0.9821**	**2.3188**	**1.4416**	**0.0456**

Tabla 4.1: Comparación de pronósticos de 30 minutos para modelos con el uso de dos diferentes combinaciones para datos de invierno.

La combinación con mejores resultados de pronósticos Humedad Interna, Rocío Interno y Radiación Solar (*Hi-Id-Rs*) fue utilizada con los modelos que realizaron un mejor acercamiento al comportamiento de la temperatura interna. Para cada modelo se realizó un pronóstico de temperatura interna de 30 minutos, 45 minutos y 1 hora, así para decidir qué modelo obtuvo mejores resultados de prueba. A pesar que el modelo SVR con *kernel* polinomial obtuvo muy buenos resultados para la aproximación del comportamiento de los datos, cuando se utilizó para el pronóstico este obtuvo los resultados menos favorables a comparación de SVR con *kernel* radial y XGBoost (Tabla 4.2), uno de las mayores limitantes del modelo SVR polinomial fue el grado del polinomio, mientras que para el acercamiento solía usar un polinomio de grado 2 o 3, para el pronóstico este número solía ser 1 y 2 en

ocasiones, así imitando el comportamiento de un SVR con *kernel* lineal el cual fue descartado desde el inicio.

Tabla 4.2: Comparación de pronósticos de 30 minutos con la combinación Hi-Id-Rs.

Modelo	Estación	R^2	RMSE	MAE	MAPE
SVR Radial	Verano	0.9493	2.4632	1.3476	0.0371
SVR Poly	**Verano**	**0.9256**	**2.9852**	**1.6644**	**0.0451**
XGBoost	Verano	0.9426	2.6213	1.4888	0.0418
SVR Radial	Primavera	0.9664	2.1873	1.3255	0.0374
SVR Poly	**Primavera**	**0.9554**	**2.5200**	**1.6798**	**0.0489**
XGBoost	Primavera	0.9620	2.3276	1.4353	0.0409
SVR Radial	Otoño	0.9697	2.1952	1.3002	0.0526
SVR Poly	**Otoño**	**0.9648**	**2.3655**	**1.5361**	**0.0618**
XGBoost	Otoño	0.9694	2.2069	1.3123	0.0506
SVR Radial	Invierno	0.9871	1.9679	1.1558	0.0353
SVR Poly	**Invierno**	**0.9765**	**2.6619**	**1.6548**	**0.0511**
XGBoost	Invierno	0.9821	2.3188	1.4417	0.0456

Los modelos SVR Radial y XGBoost fueron utilizados principalmente para comparar pronósticos de la temperatura interna, además se utilizó la combinación de variables de entrada *Hi-Id-Rs*. En todas las estaciones del año, el modelo con mejores resultados fue SVR Radial (Tabla 4.3), XGBoost obtuvo resultados muy similares en R^2, sin embargo esta métrica no es suficiente para la evaluación de un modelo. La combinación de métricas así como el MAE que una de las más significantes para este análisis, mostró que el SVR Radial obtuvo resultados favorables y estables para el pronóstico de temperatura hasta 1 hora.

Tabla 4.3: Comparación de modelos SVR Radial y XGBoost con pronóstico de 45 minutos para todas las estaciones del año.

Modelo	Estación	R^2	RMSE	MAE	MAPE
SVR Radial	**Verano**	**0.9149**	**3.1926**	**1.7880**	**0.0497**
SVR Radial	**Otoño**	**0.9400**	**3.0895**	**1.7506**	**0.0664**
SVR Radial	**Invierno**	**0.9717**	**2.9208**	**1.7179**	**0.0509**
SVR Radial	**Primavera**	**0.9434**	**2.8410**	**1.7298**	**0.0486**
XGBoost	Verano	0.9025	3.4181	1.9748	0.0562
XGBoost	Otoño	0.9405	3.0771	1.8538	0.0711
XGBoost	Invierno	0.9618	3.3947	2.1190	0.0642
XGBoost	Primavera	0.9359	3.0223	1.9230	0.0557

Un modelo SVR Radial fue utilizado para el pronóstico de temperatura interna para cada estación del año. Los hiper parámetros utilizados para los pronósticos de cada modelo por estación del año fueron obtenidos por optimización Bayesiana los cuales fueron mejor que el uso de una búsqueda exhaustiva. Los pronósticos de datos de verano fueron los que mostraron menor exactitud para los datos de prueba, esto se debió principalmente a los picos inestables notables

(Figura 4.1), a pesar de esto, los resultados son aceptables para el análisis debido al MAE el cual se encuentra por dentro del intervalo aceptado (Tabla 4.4).

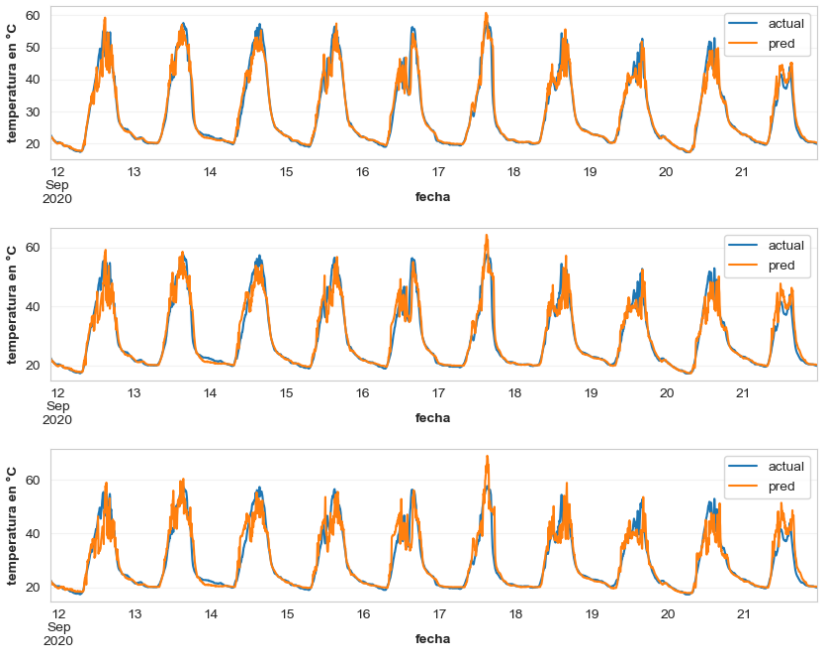

Figura 4.1: Pronóstico de temperatura interna con SVR Radial para 30 minutos, 45 minutos y 1 hora respectivamente para los datos de verano.

Modelo	Estación	Pronóstico	R^2	RMSE	MAE	MAPE
SVR Radial	Verano	30 min	0.9493	2.4632	1.3476	0.0371
SVR Radial	Verano	45 min	0.9149	3.1926	1.7880	0.0497
SVR Radial	Verano	1 hora	0.8692	3.9588	2.2309	0.0623

Tabla 4.4: Métricas de pronósticos de temperatura interna para los datos de verano con SVR Radial.

Los modelos de SVR Radial para los datos de otoño obtuvieron un pronóstico bueno y aceptable de igual manera, el modelo es capaz de adaptarse a los cambios repentinos de los datos de otoño y predecir de la mejor manera en el tiempo (Figura 4.2). De igual manera, se notó el buen pronóstico de los modelos con las métricas obtenidas (Tabla 4.5).

Modelo	Estación	Pronóstico	R^2	RMSE	MAE	MAPE
SVR Radial	Otoño	30 min	0.9697	2.1952	1.3002	0.0526
SVR Radial	Otoño	45 min	0.9400	3.0895	1.7506	0.0664
SVR Radial	Otoño	1 hora	0.9037	3.9143	2.2564	0.0852

Tabla 4.5: Métricas de pronósticos de temperatura interna para los datos de otoño con SVR Radial.

Los pronósticos para los datos de invierno fueron los mejores (Figura 4.3). En cuestión de la métrica de la R^2 fueron las más altas mientras que para el MAE de igual manera fueron lo más bajos (Tabla 4.6). Para

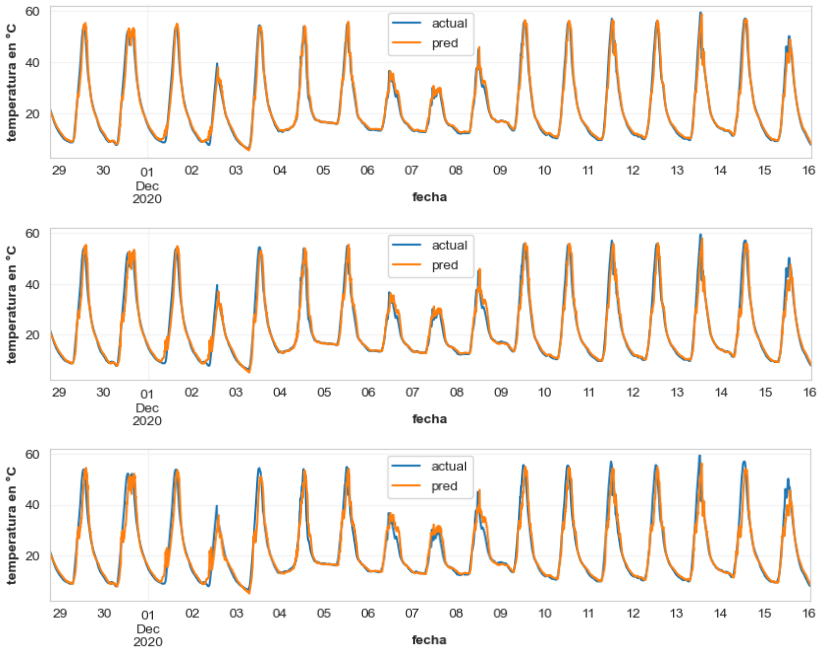

Figura 4.2: Pronóstico de temperatura interna con SVR Radial para 30 minutos, 45 minutos y 1 hora respectivamente para los datos de otoño.

los pronósticos de más de 45 minutos, se puede notar que el modelo tiene dificultad para temperaturas internas mayor a 50°, sin embargo es capaz de pronosticar temperaturas entre 0° y 20° sin problemas.

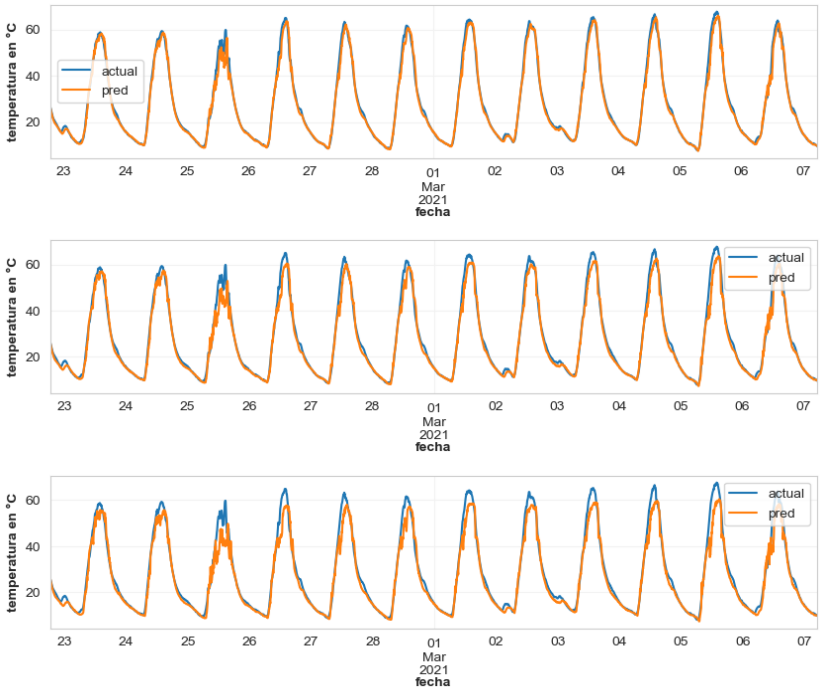

Figura 4.3: Pronóstico de temperatura interna con SVR Radial para 30 minutos, 45 minutos y 1 hora respectivamente para los datos de invierno.

Modelo	Estación	Pronóstico	R^2	RMSE	MAE	MAPE
SVR Radial	Invierno	30 min	0.9871	1.9679	1.1558	0.0353
SVR Radial	Invierno	45 min	0.9717	2.9208	1.7179	0.0509
SVR Radial	Invierno	1 hora	0.9521	3.7987	2.2425	0.0665

Tabla 4.6: Métricas de pronósticos de temperatura interna para los datos de invierno con SVR Radial.

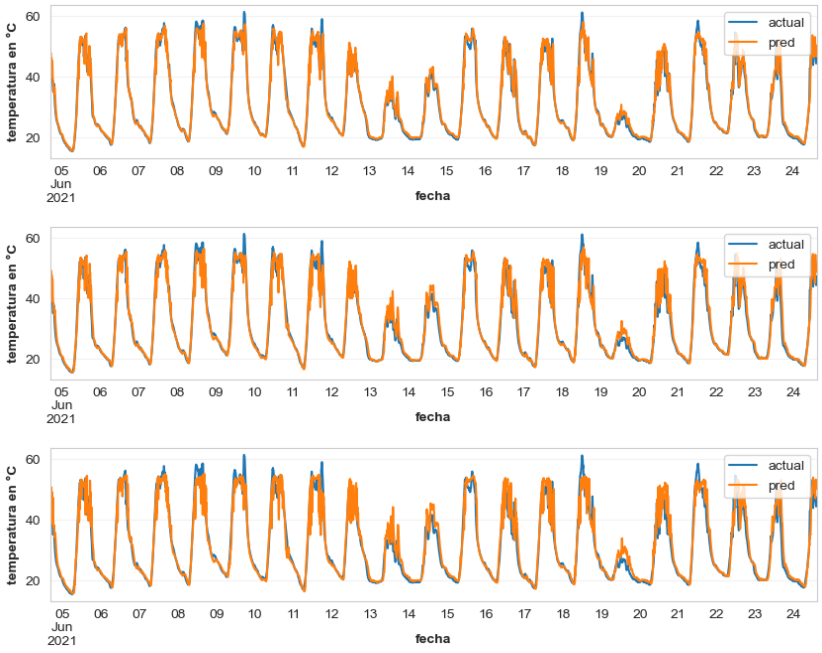

Figura 4.4: Pronóstico de temperatura interna con SVR Radial para 30 minutos, 45 minutos y 1 hora respectivamente para los datos de primavera.

Por último, el modelo SVR Radial para los datos de primavera de igual manera lograron tener buenos resultados (Tabla 4.7). El modelo es capaz de comprender el comportamiento de los datos a través del tiempo así para realizar un pronóstico de la temperatura manteniendo estabilidad (Figura 4.4).

Modelo	Estación	Pronóstico	R^2	RMSE	MAE	MAPE
SVR Radial	Primavera	30 min	0.9664	2.1873	1.3255	0.0374
SVR Radial	Primavera	45 min	0.9434	2.8410	1.7298	0.0486
SVR Radial	Primavera	1 hora	0.9201	3.3748	2.1054	0.0607

Tabla 4.7: Métricas de pronósticos de temperatura interna para los datos de primavera con SVR Radial.

Cada modelo utilizado para el pronóstico de la temperatura interna del invernadero consideró los datos de manera individual debido a que cada estación del año tiene una tendencia, máximo, mínimo diferente (Tabla 4.8). Es por esto que para la optimización de los hiper parámetros se consideró cada estación y cada pronóstico diferente, los cuales brindaron resultados muy aceptables para este análisis. A pesar de usar la optimización Bayesiana, la búsqueda de estos hiper parámetros fue exhausta utilizando diferentes combinaciones de parámetros de entrada así como diferentes alteraciones del método de la optimización.

Modelo	Estación	Pronóstico	C	epsilon	gamma
SVR Radial	Verano	30 min	42.88	0.97	0.10
SVR Radial	Verano	45 min	36.43	0.97	0.10
SVR Radial	Verano	1 hora	94.11	0.96	0.09
SVR Radial	Otoño	30 min	22.70	0.76	0.03
SVR Radial	Otoño	45 min	24.16	0.53	0.04
SVR Radial	Otoño	1 hora	94.36	0.43	0.02
SVR Radial	Invierno	30 min	70.46	0.58	0.15
SVR Radial	Invierno	45 min	96.81	0.37	0.10
SVR Radial	Invierno	1 hora	71.10	0.12	0.10
SVR Radial	Primavera	30 min	46.18	0.63	0.06
SVR Radial	Primavera	45 min	74.74	0.26	0.04
SVR Radial	Primavera	1 hora	43.89	0.48	0.07

Tabla 4.8: Hiperparámetros empleados para SVR Radial para todas las estaciones

5 Discusión

Considerando que la mejor combinación utilizada en el modelo contiene la variable de radiación solar, un sensor para esta variable es de suma importancia. Un sensor de radiación solar suele ser más costoso que alguno de humedad u otro, por ende algún invernadero puede no contar con alguno para su uso en algún modelo, sin embargo, los resultados de los modelos mostraron errores aceptables con otra combinación de humedad interna, rocío interno y temperatura externa, por lo que puede ser utilizada como alternativa en invernaderos sin un sensor de radiación solar.

El costo computacional es un factor a considerar de igual manera, XGBoost presentó un pronóstico aceptable a menor costo computacional y menor complejidad con el uso de árboles de decisión. SVR, por otro lado, es un modelo complejo que permitió obtener mejores pronósticos con un error aceptable para este análisis.

El modelo con mayor capacidad de pronóstico debe de ser capaz de ser implementado en el invernadero, por lo tanto se vuelve importante poder usar el modelo complejo sin aumentar costos del invernadero. Uno de los objetivos generales del proyecto magno y no particularmente de este trabajo, es el uso del modelo dentro de un microcontrolador, por lo que la librería *TinyML*[1] permite la exportación de modelos entrenados en lenguaje C el cual puede ser cargado a un dispositivo como Arduino o *Raspberry Pi* evitando costos adicionales por un equipo especializado.

El modelo obtenido para el pronóstico de la temperatura interna en este análisis, tiene resultados parecidos a los existentes. Weidong Zou, et al.[2] presentó resultados inferiores con SVM, sin embargo su modelo CB-ELM, el cual es un modelo con mayor complejidad y sin oportunidad de ser usado en un microcontrolador, obtuvo mejor resultados. El acercamiento de Kalavathi Devi Thangavel, et al.[3] utilizando una lógica *fuzzy* con un modelo de SVC obtuvo resultados parecidos al de este análisis, sin embargo al usar lógica *fuzzy* la complejidad de los datos y del modelo suele aumentar, por lo que su uso suele no ser implementado particularmente en un microcontrolador (Tabla 5.1).

[1] P. Warden and D. Situnayake. *TinyML: Machine Learning with TensorFlow Lite on Arduino and Ultra-Low-Power Microcontrollers*. O'Reilly Media, 2019. ISBN 9781492051992

[2] W. Zou, F. Yao, B. Zhang, C. He, and Z. Guan. Verification and predicting temperature and humidity in a solar greenhouse based on convex bidirectional extreme learning machine algorithm. *Neurocomputing*, 249:72–85, 2017. ISSN 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2017.03.023. URL https://www.sciencedirect.com/science/article/pii/S0925231217305180

[3] K. Devi Thangavel, U. Seerengasamy, S. Palaniappan, and R. Sekar. Prediction of factors for controlling of green house farming with fuzzy based multiclass support vector machine. *Alexandria Engineering Journal*, 62: 279–289, 2023. ISSN 1110-0168. DOI: https://doi.org/10.1016/j.aej.2022.07.016. URL https://www.sciencedirect.com/science/article/pii/S1110016822004732

Autor	**Modelo**	**MAE**	**RMSE**
Weidong Zou, et al.	SVM		2.77
Kalavathi Devi Thangavel, et al.	*Fuzzy Multiclass* SVM	0.609	1.07
Jesús Ponce	**SVR Radial**	**1.28**	**2.20**
Jesús Ponce	**XGBoost**	**1.42**	**2.37**

Tabla 5.1: Comparación de modelos SVM

6 Conclusiones y trabajo futuro

Contenido

6.1 Conclusiones

El análisis y procesamiento de datos realizado para obtener un pronóstico en diferentes tiempos presentó resultados con MAE 1.282°C, 1.747°C, 2.209°C para 30, 45 y 60 minutos, respectivamente. Los diferentes algoritmos utilizados mostraron errores aceptables con $\pm$2.5°C desde 30 hasta 60 minutos.

El modelo de regresión lineal contribuyó con la comprensión de la relación de las variables climáticas con la temperatura interna del invernadero, para así ser usada como referencia para la decisión de candidatos para modelos más complejos. La relación entre la temperatura interna y los demás factores mostró una tendencia lineal a muy corto plazo.

El modelo complejo XGBoost fue eficiente con la gran cantidad de datos y obtuvo resultados aceptables, mientras que SVR radial, a pesar de su mayor tiempo de ejecución, obtuvo mejores resultados por pocos puntos porcentuales. XGBoost fue capaz de generar un pronóstico rápido sin necesidad de afinar con exactitud sus hiper parámetros, sin embargo después de optimizar, SVR radial generó mejores resultados para todas las métricas.

El modelo SVR con *kernel* polinomial se consideró de igual manera, no obstante a comparación de los otros candidatos, el tiempo de ejecución así como los errores fue mayor. Es también importante notar que el SVR radial tiende a ser lento al procesar muchos datos, por lo que la división por estaciones resolvió el problema de tiempo de ejecución para este.

Desde el aspecto de precisión de pronóstico, el modelo SVR Radial obtuvo los mejores resultados en comparación con los demás algoritmos,

mientras que XGBoost obtuvo resultados similares a un menor tiempo de ejecución. Con un gran poder de predicción, adaptabilidad en el tiempo, capacidad de modelar con datos en N dimensiones y complejidad, el modelo SVR con *kernel* radial tiene gran aplicación para el pronóstico de temperatura en el invernadero.

6.2 *Trabajo futuro*

Poseer un modelo preventivo para la temperatura interna del invernadero, extiende las posibilidades de uso como determinar la humedad interna o la radiación solar con algún otro sensor de bajo costo reduciendo la inversión sustituyendo sensores.

Se busca realizar un modelo de pronóstico para humedad interna apoyado de la experiencia adquirida con el pronóstico de temperatura interna y radiación solar realizando un análisis de correlación entre un sensor barato con uno de mayor costo con el fin de reemplazar el uso de un sensor demasiado caro.

Es necesario poder utilizar el modelo dentro del invernadero, es por esto que gracias a librerías de *Python* como *TinyML*, el mejor modelo se estará compilando en un microcontrolador para su uso dentro del invernadero. Con el uso de un microcontrolador en lugar de una computadora, el sistema de control será capaz de realizar ajustes a menor costo con anticipación.

Bibliografía

URL https://www.kaggle.com/competitions/higgs-boson.

A. Abdullah, S. Al Enazi, and I. Damaj. Agrisys: A smart and ubiquitous controlled-environment agriculture system. In *2016 3rd MEC International Conference on Big Data and Smart City (ICBDSC)*, pages 1–6, 2016. DOI: 10.1109/ICBDSC.2016.7460386.

W. I. R. Agmail, R. Linker, and A. Arbel. Robust control of greenhouse temperature and humidity. *IFAC Proceedings Volumes*, 42:138–143, 2009. ISSN 1474-6670. DOI: https://doi.org/10.3182/20090616-3-IL-2002.00024. URL https://www.sciencedirect.com/science/article/pii/S147466701540391X.

H. A. Ahmed, Y. X. TONG, Q. C. YANG, A. A. Al-Faraj, and A. M. Abdel-Ghany. Spatial distribution of air temperature and relative humidity in the greenhouse as affected by external shading in arid climates. *Journal of Integrative Agriculture*, 18:2869–2882, 2019. ISSN 2095-3119. DOI: https://doi.org/10.1016/S2095-3119(19)62598-0. URL https://www.sciencedirect.com/science/article/pii/S2095311919625980.

L.D. Albright and N.R. Scott. An analysis of steady periodic building temperature variation in warm weather - part i a mathematical model - transactions of the asae. 1974.

A. Arbel, O. Yekutieli, and M. Barak. Performance of a fog system for cooling greenhouses. *Journal of Agricultural Engineering Research*, 72: 129–136, 1999. DOI: 10.1006/jaer.1998.0351.

J. W. Bartok Jr. Selecting and building a commercial greenhouse, Apr 2017. URL https://ag.umass.edu/greenhouse-floriculture/fact-sheets/selecting-building-commercial-greenhouse. Accesado en: 2023-03-21.

W. O. Baudoin. *Good agricultural practices for greenhouse vegetable crops principles for Mediterranean climate areas*, pages 21,22. Food and Agricultural Organization of the United Nations (FAO), 2013.

J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.*, 13:281–305, feb 2012. ISSN 1532-4435.

R. Bessin, Lee. H. Townsend, R. Extension Entomologists & G. Anderson, and Extension Horticulturist University of Kentucky College of Agriculture. Greenhouse insect management, 2007. URL https://entomology.ca.uky.edu/ent60. Accesado en: 2023-01-15.

G.P.A. Bot. Physical modeling of greenhouse climate. *IFAC Proceedings Volumes*, 24:7–12, 1991. ISSN 1474-6670. DOI: https://doi.org/10.1016/B978-0-08-041273-3.50006-9. URL https://www.sciencedirect.com/science/article/pii/B9780080412733500069. IFAC/ISHS Workshop on Mathematical and Control Applications in Agriculture and Horticulture, Matsuyama, Japan, 30 September-3 October 1992.

W. Cai, R. Wei, L. Xu, and X. Ding. A method for modelling greenhouse temperature using gradient boost decision tree. *Information Processing in Agriculture*, 9:343–354, 2022. ISSN 2214-3173. DOI: https://doi.org/10.1016/j.inpa.2021.08.004. URL https://www.sciencedirect.com/science/article/pii/S2214317321000743.

C. Camacho. Regresión lineal simple. Accesado en: 2023-01-18, December 2019. URL https://personal.us.es/vararey/regresion-simple.pdf.

T. Chen. Story and lessons behind the evolution of xgboost, 2014. URL https://sites.google.com/site/nttrungmtwiki/home/it/data-science---python/xgboost/story-and-lessons-behind-the-evolution-of-xgboost. Accesado en: 2023-03-29.

T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. page 785–794. Association for Computing Machinery, 2016a. ISBN 9781450342322. DOI: 10.1145/2939672.2939785.

T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016b. ISBN 978-1-4503-4232-2. DOI: 10.1145/2939672.2939785.

Y. Chiba, K. Okada, Y. Hayashi, S. Kumada, and Y. Onuki. Usefulness of applying partial least squares regression to t_2 relaxation curves for predicting the solid form content in binary physical mixtures. *Journal of Pharmaceutical Sciences*, 112:1041–1051, 2023. ISSN 0022-3549. DOI: https://doi.org/10.1016/j.xphs.2022.11.028. URL https://www.sciencedirect.com/science/article/pii/S0022354922005391.

Microsoft Corporation. Visual studio code. https://code.visualstudio.com/docs, 2015. Accesado en: 2023-03-24.

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, Sep 1995. ISSN 1573-0565. DOI: 10.1007/BF00994018. URL https://doi.org/10.1007/BF00994018.

A. De Myttenaere, B. Golden, B. Le Grand, and F. Rossi. Mean absolute percentage error for regression models. *Neurocomputing*, 192:38–48, 2016. DOI: 10.1016/j.neucom.2015.12.114.

K. Devi Thangavel, U. Seerengasamy, S. Palaniappan, and R. Sekar. Prediction of factors for controlling of green house farming with fuzzy based multiclass support vector machine. *Alexandria Engineering Journal*, 62:279–289, 2023. ISSN 1110-0168. DOI: https://doi.org/10.1016/j.aej.2022.07.016. URL https://www.sciencedirect.com/science/article/pii/S1110016822004732.

Firman D.M. and E.J. Allen. Chapter 33 - agronomic practices. In D. Vreugdenhil, J. Bradshaw, C. Gebhardt, F. Govers, D. Mackerron, M. Taylor, and H. Ross, editors, *Potato Biology and Biotechnology*, pages 719–738. Elsevier Science B.V., 2007. ISBN 978-0-444-51018-1. DOI: https://doi.org/10.1016/B978-044451018-1/50075-0. URL https://www.sciencedirect.com/science/article/pii/B9780444510181500750.

C. Duarte-Galvan, I. Torres-Pacheco, R. G. Guevara-Gonzalez, R. J. Romero-Troncoso, L. M. Contreras-Medina, M. A. Rios-Alcaraz, and J. R. Millan-Almaraz. Review. advantages and disadvantages of control theories applied in greenhouse climate control systems. *Spanish Journal of Agricultural Research*, 10:926–938, Oct 2012. DOI: 10.5424/sjar/2012104-487-11. URL https://revistas.inia.es/index.php/sjar/article/view/2196.

A. Elanchezhian, J. K. Basak, J. Park, F. Khan, F. G. Okyere, Y. Lee, A. Bhujel, D. Lee, T. Sihalath, and H. T. Kim. Evaluating different models used for predicting the indoor microclimatic parameters of a greenhouse. *Applied Ecology and Environmental Research*, pages 2141–2161, 2020. DOI: https://doi.org/10.15666/aeer/1802_21412161.

F. Galton. Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15: 246–263, 1886. ISSN 09595295. URL http://www.jstor.org/stable/2841583.

A. Ganguly and S. Ghosh. A review of ventilation and cooling technologies in agricultural greenhouse application. *Iranian (Iranica)*

Journal of Energy & Environment, 2011. ISSN 2079-2115. URL https://www.ijee.net/article_64325.html.

A. Gelman and C. Rohilla-Shalizi. Philosophy and the practice of bayesian statistics. In *Mathematical and Statistical Psychology*, volume 66, pages 8–38, 2013.

T. Hans-Juergen. Energy saving potential of greenhouse climate control. *Mathematics and Computers in Simulation (MATCOM)*, 48:245–251, 1998. URL https://ideas.repec.org/a/eee/matcom/v48y1998i1p93-101.html.

Charles R. Harris et al. Array programming with NumPy. *Nature*, page 357–362, 2020. DOI: 10.1038/s41586-020-2649-2.

J. Hernández-Lalinde, J. Espinosa-Castro, V. Bermudez, and D. García Álvarez. Sobre el uso adecuado de la regresión lineal: conceptualización básica mediante un ejemplo aplicado a las ciencias de la salud. *Archivos Venezolanos de Farmacologia y Terapeutica*, 38: 608–614, 12 2020.

Z. Hui, Q. Lin-lin, and W. Gang. Modeling and simulation of greenhouse temperature hybrid system based on armax model. In *2017 36th Chinese Control Conference (CCC)*, pages 2237–2241, 2017. DOI: 10.23919/ChiCC.2017.8027690.

John D Hunter. Matplotlib: A 2d graphics environment. *Computing in science & engineering*, pages 90–95, 2007.

IBM. What is boosting?, 2022. URL https://www.ibm.com/topics/boosting. Accesado en: 2023-03-29.

INEGI. COMUNICADO DE PRENSA NÚM. 130/22, 2022. URL https://www.inegi.org.mx/contenidos/saladeprensa/boletines/2022/pib_pconst/pib_pconst2022_02.pdf. Accesado en: 2022-11-08.

K. Ito and T. Tabei. Model predictive temperature and humidity control of greenhouse with ventilation. *Procedia Computer Science*, 192:212–221, 2021. ISSN 1877-0509. DOI: https://doi.org/10.1016/j.procs.2021.08.022. URL https://www.sciencedirect.com/science/article/pii/S187705092101509X.

I. Jaisankar, A. Velmurugan, and C. Sivaperuman. Chapter 19 - biodiversity conservation: Issues and strategies for the tropical islands. In C. Sivaperuman, A. Velmurugan, I. Jaisankar, and A. K. Singh, editors, *Biodiversity and Climate Change Adaptation in Tropical Islands*, pages 525–552. Academic Press, 2018. ISBN

978-0-12-813064-3. DOI: https://doi.org/10.1016/B978-0-12-813064-3.00019-3. URL https://www.sciencedirect.com/science/article/pii/B9780128130643000193.

R. Jiao, Y. Li, Q. Wang, and B. Li. Svm regression and its application to image compression. In H. De-Shuang, Z. Xiao-Ping, and H. Guang-Bin, editors, *Advances in Intelligent Computing*, pages 747–756. Springer Berlin Heidelberg, 2005. ISBN 978-3-540-31902-3.

C. Kittas, T. Bartzanas, and A. Jaffrin. Temperature gradients in a partially shaded large greenhouse equipped with evaporative cooling pads. *Biosystems Engineering*, 85:87–94, 2003. ISSN 1537-5110. DOI: https://doi.org/10.1016/S1537-5110(03)00018-7. URL https://www.sciencedirect.com/science/article/pii/S1537511003000187.

J. Leal Iga, J. Leal Iga, C. Leal Iga, and R. A. Flores. Effect of air density variations on greenhouse temperature model. *Mathematical and Computer Modelling*, 47:855–867, 2008. ISSN 0895-7177. DOI: https://doi.org/10.1016/j.mcm.2007.05.011. URL https://www.sciencedirect.com/science/article/pii/S0895717707002130.

G. Leonidopoulos. Greenhouse dimensions estimation and short time forecast of greenhouse temperature based on net heat losses through the polymeric cover. *Polymer Testing*, 19:801–812, 2000. ISSN 0142-9418. DOI: https://doi.org/10.1016/S0142-9418(99)00050-1. URL https://www.sciencedirect.com/science/article/pii/S0142941899000501.

Nightingale Garden Company Limited. Exotic plants in nineteenth century gardens, 2007. URL https://www.gardenvisit.com/history_theory/library_online_ebooks/ml_gothein_history_garden_art_design/exotic_plants_planting. Accesado en: 2023-01-30.

J. Litago, F. Baptista, J. Meneses, L. Navas, B. Bailey, and V. Sánchez-Girón. Statistical modelling of the microclimate in a naturally ventilated greenhouse. *Biosystems Engineering*, 92:365–381, 2005. ISSN 1537-5110. DOI: https://doi.org/10.1016/j.biosystemseng.2005.07.015. URL https://www.sciencedirect.com/science/article/pii/S1537511005001625.

H. Liu, X. Yao, C. Xue, R. Zhang, M. Liu, Z. Hu, and B. Fan. Study of quantitative structure-mobility relationship of the peptides based on the structural descriptors and support vector machines. *Analytica Chimica Acta*, 542:249–259, 2005. ISSN 0003-2670. DOI: https://doi.org/10.1016/j.aca.2005.04.006. URL https://www.sciencedirect.com/science/article/pii/S000326700500591X.

I. L. López-Cruz, A. Rojano-Aguilar, W. Ojeda-Bustamante, and R. Salazar-Moreno. Arx models for predicting greenhouse air temperature: A methodology. *Agrociencia*, 41:181–192, 02 2007.

M. López-Ibáñez, J. Dubois-Lacoste, L. P. Cáceres, M. Birattari, and T. Stützle. The irace package: Iterated racing for automatic algorithm configuration. *Operations Research Perspectives*, 3:43–58, 2016. DOI: 10.1016/j.orp.2016.09.002.

J. M. Stanton. Galton, pearson, and the peas: A brief history of linear regression for statistics instructors. *Journal of Statistics Education*, 9, 2001. DOI: 10.1080/10691898.2001.11910537. URL https://doi.org/10.1080/10691898.2001.11910537.

S. Maitra and J. Yan. Principle component analysis and partial least squares: Two dimension reduction techniques for regression. *Casualty Actuarial Society*, 01 2008.

R. Martinez-Cantin. Bayesopt: A bayesian optimization library for nonlinear optimization, experimental design and bandits. *Journal of Machine Learning Research*, pages 3915–3919, 2014. URL http://jmlr.org/papers/v15/martinezcantin14a.html.

A. McIntosh, F Bookstein, J. Haxby, and C. Grady. Spatial pattern analysis of functional brain images using partial least squares. *NeuroImage*, 3:143–157, 1996. ISSN 1053-8119. DOI: https://doi.org/10.1006/nimg.1996.0016. URL https://www.sciencedirect.com/science/article/pii/S1053811996900166.

Wes McKinney et al. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56, 2010.

D. C. Montgomery, E. A. Peck, and G. G. Vining. *Introduction to Linear Regression Analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., 2015. ISBN 9781119180173. URL https://books.google.com.mx/books?id=27kOCgAAQBAJ.

K. Nemali. Temperature control in greenhouses. *Horticulture and Landscape Architecture*, Feb 2021. URL https://www.extension.purdue.edu/extmedia/HO/HO-327-W.pdf.

The Editors of Encyclopedia Britannica. greenhouse, 2019. URL https://www.britannica.com/topic/greenhouse. Accesado en: 2023-01-30.

L. Ouazzani Chahidi, M. Fossa, A. Priarone, and A. Mechaqrane. Evaluation of supervised learning models in predicting greenhouse

energy demand and production for intelligent and sustainable operations. *Energies*, 14, 2021. ISSN 1996-1073. DOI: 10.3390/en14196297. URL https://www.mdpi.com/1996-1073/14/19/6297.

Shakuntala Pandey and Anil Pandey. Greenhouse technology. *International Journal of Research -GRANTHAALAYAH*, page 1–3, 2015. DOI: 10.29121/granthaalayah.v3.i9se.2015.3176.

W. S. Parker. Model evaluation: An adequacy-for-purpose view. *Philosophy of Science*, 87:457–477, 2020. DOI: 10.1086/708691.

F. Pedregosa et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12:2825–2830, 2011.

D. M. Pirouz. An overview of partial least squares. *SSRN Electronic Journal*, 10 2006. DOI: 10.2139/ssrn.1631359.

D. A. Pisner and D. M. Schnyer. Chapter 6 - support vector machine. In Andrea Mechelli and Sandra Vieira, editors, *Machine Learning*, pages 101–121. Academic Press, 2020. ISBN 978-0-12-815739-8. DOI: https://doi.org/10.1016/B978-0-12-815739-8.00006-7. URL https://www.sciencedirect.com/science/article/pii/B9780128157398000067.

P. Probst, A. Boulesteix, and B. Bischl. Tunability: Importance of hyperparameters of machine learning algorithms, 2018.

J. Pérez-Parra and J. C. López-Hernández. Evolución de las estructuras de invernadero, 2007. URL https://www.publicacionescajamar.es/publicacionescajamar/public/pdf/series-tematicas/centros-experimentales-las-palmerillas/evolucion-de-las-estructuras.pdf. Accesado en: 2023-03-21.

R. Ramin Shamshiri, F. Kalantari, K. C. Ting, K. R. Thorp, I. A. Hameed, C. Weltzien, D. Ahmad, and Z. Mojgan Shad. Advances in greenhouse automation and controlled environment agriculture: A transition to plant factories and urban agriculture. *International Journal of Agricultural and Biological Engineering*, 11:1–22, 2018. DOI: https://doi.org/10.25165/j.ijabe.20181101.3210.

E. Reyes. Sistemas de climatización en invernaderos, 2014. URL https://www.mundohvacr.com.mx/2014/05/sistemas-de-climatizacion-en-invernaderos/. Accesado en: 2023-01-30.

E. Runkle. The perils of low (greenhouse) temperature, Feb 2020. URL https://gpnmag.com/article/

the-perils-of-low-greenhouse-temperature/. Accesado en: 2023-02-21.

Wright S. Correlation and causation. *Journal of agricultural research*, pages 557–585, 1921. ISSN 0095-9758.

C. Sammut and G. I. Webb. *Encyclopedia of Machine Learning*, chapter Mean Absolute Error, pages 652–652. Springer US, 2010a. ISBN 978-0-387-30164-8.

C. Sammut and G. I. Webb. *Encyclopedia of Machine Learning*, chapter Mean Squared Error, pages 653–653. Springer US, 2010b. ISBN 978-0-387-30164-8.

M. K Selcuk. Use of digital computers for the heat and mass transfer of controlled environment greenhouses. *Environmental Research Laboratory, University of Arizona, Tucson, Arizona*, 1970.

V.P. Sethi. On the selection of shape and orientation of a greenhouse: Thermal modeling and experimental validation. *Solar Energy*, 83:21–38, 2009. DOI: 10.1016/j.solener.2008.05.018.

Y. Shao, Q.J. Wang, A. Schepen, and D. Ryu. Going with the trend: Forecasting seasonal climate conditions under climate change. *Monthly Weather Review*, 149:2513–2522, 2021. DOI: https://doi.org/10.1175/MWR-D-20-0318.1. URL https://journals.ametsoc.org/view/journals/mwre/149/8/MWR-D-20-0318.1.xml.

J. Shawe-Taylor, University of Southampton, N. Cristianini, and University of California. *Kernel Methods for Pattern Analysis*, pages 47–84. Cambridge University Press, 2011. ISBN 9780511809682. DOI: 10.1017/CBO9780511809682.

J. a. K Suykens. Support vector machines: A nonlinear modelling and control perspective. *European Journal of Control*, 7:311–327, 2001. DOI: 10.3166/ejc.7.311-327.

M. Teitel, M. Atias, and M. Barak. Gradients of temperature, humidity and co2 along a fan-ventilated greenhouse. *Biosystems Engineering*, 106:166–174, 2010. ISSN 1537-5110. DOI: https://doi.org/10.1016/j.biosystemseng.2010.03.007. URL https://www.sciencedirect.com/science/article/pii/S1537511010000590.

M. Tenenhaus. *La régression PLS: Theéorie et pratique*. Technip, 1998.

R. Turner, D. Eriksson, M. McCourt, J. Kiili, E. Laaksonen, Z. Xu, and I. Guyon. Bayesian optimization is superior to random search for machine learning hyperparameter tuning: Analysis of the black-box optimization challenge 2020, 2021.

A. J. Udink ten Cate. *Modeling and (adaptive) control of greenhouse climates*. PhD thesis, Technische Hogeschool Twente, 1983.

Cornell University. Gaussian processes and bayesian optimization. https://www.cs.cornell.edu/courses/cs4787/2019sp/notes/lecture16.pdf, 2019. Accesado en: 2023-03-21.

G. Van Rossum and Fred L. D. *Python 3 Reference Manual*. CreateSpace, 2009. ISBN 1441412697.

V. Vapnik and A. Chervonenkis. A note on one class of perceptrons. *Automation and Remote Control*, 25, 1964.

V. Vapnik, S. Golowich, and A. Smola. Support vector method for function approximation, regression estimation and signal processing. In *Proceedings of the 9th International Conference on Neural Information Processing Systems*, page 281–287. MIT Press, 1996.

V. N. Vapnik. *The Nature of Statistical Learning Theory*, chapter Methods of Function Estimation, pages 181–224. Springer New York, 2000. ISBN 978-1-4757-3264-1.

K. Wabersich. Gaussian processes and bayesian optimization. https://www.kimpeter.de/wp-content/uploads/2016/12/project_GBO.pdf, 2016. Accesado en: 2023-03-21.

D. Wang, M. Wang, and X. Qiao. Support vector machines regression and modeling of greenhouse environment. *Computers and Electronics in Agriculture*, 66:46–52, 2009. ISSN 0168-1699. DOI: https://doi.org/10.1016/j.compag.2008.12.004. URL https://www.sciencedirect.com/science/article/pii/S0168169908002305.

P. Warden and D. Situnayake. *TinyML: Machine Learning with TensorFlow Lite on Arduino and Ultra-Low-Power Microcontrollers*. O'Reilly Media, 2019. ISBN 9781492051992.

Michael Waskom et al. mwaskom/seaborn: v0.8.1 (september 2017). *Zenodo*, 2017. DOI: 10.5281/zenodo.883859. URL https://doi.org/10.5281/zenodo.883859.

R. O. Wayua, V. Ochieng, V. Kirigua, and L. Wasilwa. Challenges in greenhouse crop production by smallholder farmers in kisii county, kenya. *African Journal of Agricultural Research*, 16:1411–1419, 10 2020. DOI: 10.5897/AJAR2020.15086.

S. Weisberg. *Applied linear regression*. John Wiley & Sons, Inc, 2005. ISBN 9780471704096.

S. Wold, M. Sjöström, and L. Eriksson. Pls-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58: 109–130, 2001. ISSN 0169-7439. DOI: https://doi.org/10.1016/S0169-7439(01)00155-1. URL https://www.sciencedirect.com/science/article/pii/S0169743901001551. PLS Methods.

C. X. Xue, R. S. Zhang, H. X. Liu, M. C. Liu, Z. D. Hu, and B. T. Fan. Support vector machines-based quantitative structure-property relationship for the prediction of heat capacity. *Journal of Chemical Information and Computer Sciences*, 44:1267–1274, 2004. DOI: 10.1021/ci049934n. URL https://doi.org/10.1021/ci049934n.

L. Yang and A. Shami. On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, 415: 295–316, 2020. DOI: 10.1016/j.neucom.2020.07.061.

J. Zhang, S. Zhao, A. Dai, P. Wang, Z. Liu, B. Liang, and T. Ding. Greenhouse natural ventilation models: How do we develop with chinese greenhouses? *Agronomy*, 12, 2022. ISSN 2073-4395. DOI: 10.3390/agronomy12091995. URL https://www.mdpi.com/2073-4395/12/9/1995.

W. Zou, F. Yao, B. Zhang, C. He, and Z. Guan. Verification and predicting temperature and humidity in a solar greenhouse based on convex bidirectional extreme learning machine algorithm. *Neurocomputing*, 249:72–85, 2017. ISSN 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2017.03.023. URL https://www.sciencedirect.com/science/article/pii/S0925231217305180.