Desarrollo de modelo de machine learning y herramienta de visualización de estimaciones futuras de DSV para la prevención de la sigatoka negra en Belice

Oscar José Méndez Tello

UVG UNIVERSIDAD DEL VALLE DE GUATEMALA

UNIVERSIDAD DEL VALLE DE GUATEMALA

Facultad de Ingeniería

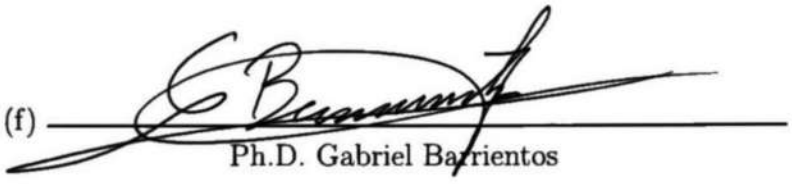
Desarrollo de modelo de machine learning y herramienta de visualización de estimaciones futuras de DSV para la prevención de la sigatoka negra en Belice

Trabajo de graduación presentado por Oscar José Méndez Tello para optar al grado académico de Licenciado en Ingeniería en Ciencia de los Datos

Guatemala,

2024

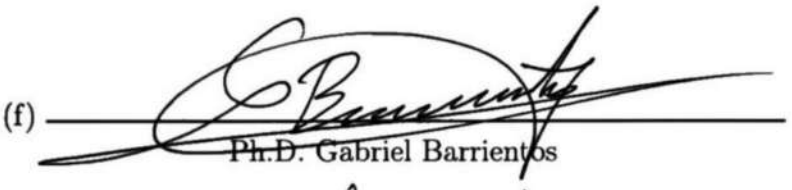
Vo.Bo.:



(f) _____

Ph.D. Gabriel Barrientos

Tribunal Examinador:

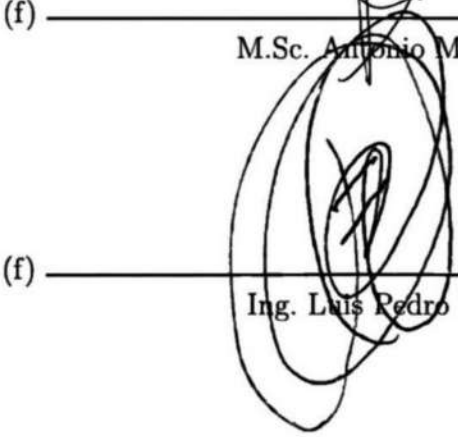


(f) _____

Ph.D. Gabriel Barrientos

(f) _____

M.Sc. Antonio Medrano



(f) _____

Ing. Luis Pedro Flores

Fecha de aprobación: Guatemala, 23 de enero de 2025.

Índice

Índice de figuras

Índice de cuadros

Resumen

La sigatoka negra es una enfermedad causada por el hongo *Mycosphaerella Fijiensis*, el cual se dispersa principalmente en condiciones de lluvia y humedad. Se encuentra en la mayoría de las zonas tropicales y subtropicales donde se cultivan plátanos y bananos, en condiciones favorables, puede reducir los rendimientos hasta en un 50 %, convirtiéndose en un problema para todo el continente americano.

Este trabajo se basa en el modelo de *disease severity value* (DSV) utilizado por la World Wide Fund for Nature (WWF) en Belice, donde se calcula diariamente un valor de riesgo de desarrollo de la enfermedad, tomando en cuenta variables como la humedad de las hojas, temperatura, precipitación y humedad relativa. El objetivo es poder estimar el DSV para los siguientes 3 días y desarrollar una herramienta de visualización con las predicciones que ayuden al manejo y control de la sigatoka negra.

Se integran datos meteorológicos de sensores y satélites para entrenar un modelo de transformador de fusión temporal para predecir la humedad de las hojas. Se optimizó la métrica MAE y se evaluó con *precision*, *recall* y F1-score. El F1-score de las regiones de Belice fue 0.82 en el primer día de predicción y 0.79 en el tercer día de predicción. Estos resultados positivos permiten una correcta estimación del $\mathrm{DSV_{LW}}$ con un R^2 de 0.63 en el primer día de predicción y 0.55 en el tercer día de predicción. Los datos se integran en un tablero que facilita la visualización y análisis de información reciente y pronosticada, proporcionando así una herramienta de control intuitiva para la toma de decisiones en el manejo de la sigatoka negra en Belice.

Palabras clave — agricultura; aprendizaje automático; sigatoka negra; transformador de fusion temporal; variables meteorológicas.

Abstract

The black sigatoka is a disease caused by the fungus *Mycosphaerella Fijiensis*, which spreads primarily under conditions of rain and humidity. It is found in most tropical and subtropical regions where bananas and plantains are cultivated, and under favorable conditions, it can reduce yields by up to 50 %, making it a problem across the entire American continent.

This work is based on the Disease Severity Value (DSV) model used by the World Wide Fund for Nature (WWF) in Belize, where a daily risk value for disease development is calculated, taking into account variables such as leaf wetness, temperature, precipitation, and relative humidity. The objective is to predict the DSV values for the following three days and develop a visualization tool with these predictions to aid in the management and control of black sigatoka.

Meteorological data from sensors and satellites are integrated to train a temporal fusion transformer model to predict leaf wetness. The model was optimized for the MAE metric and evaluated using precision, recall, and F1-score. The F1-score for Belize regions was 0.82 on the first prediction day and 0.79 on the third prediction day. These positive results allow an accurate estimation of $\mathrm{DSV_{LW}}$, achieving an R^2 of 0.63 on the first prediction day and 0.55 on the third day. The data is integrated into a dashboard that enbale the visualization and analysis of recent information and forecasts, providing an intuitive control tool for decision-making for managing black sigatoka in Belize.

Keywords — agriculture; black sigatoka; machine learning; meteorological variables; temporal fusion transformer.

CAPÍTULO 1

Introducción

La sigatoka negra es una enfermedad que afecta a las plantas de banano y plátano, comienza con manchas en las hojas, las cuales evolucionan a rayas marrones y negras, a veces con un tono púrpura [1]. Debido a su forma de reproducción, las condiciones de lluvia y humedad favorecen la dispersión del hongo, convirtiéndose en un problema para todo el continente americano [2]. La sigatoka negra se encuentra en la mayoría de las zonas tropicales y subtropicales donde se cultivan plátanos y bananos, en condiciones favorables, puede reducir los rendimientos hasta en un 50 % [3], [4]. El combate de esta enfermedad constituye uno de los principales rubros para la industria bananera, representando hasta un 27 % del costo total de producción [5].

Existen varios formas de controlar la enfermedad, pero el más utilizado son los controles químicos, en este caso, fungicidas. Según la WWF, estos fungicidas se clasifican en protectores, sistémicos y aceites. Los fungicidas protectores actúan en la superficie de la hoja, los sistémicos penetran en la hoja y se utilizan en infecciones más avanzadas, y los aceites se aplican junto con ambos tipos de fungicidas para mejorar su eficacia [6]. Si bien los fungicidas sistémicos son eficaces contra la sigatoka negra, sus efectos sobre el medio ambiente son motivo de preocupación [7].

Este proyecto se basa en el modelo de *disease severity value* (DSV), utilizado por la WWF en Belice, el cual estima el riesgo de desarrollo de la enfermedad utilizando datos climáticos y niveles de humedad en las hojas [6]. Sin embargo, el modelo actual presenta limitaciones al considerar únicamente datos actuales y pasados, sin considerar como estas variables van a estar en los días siguientes. Por este motivo se busca desarrollar un modelo de *machine learning*, específicamente un transformador de fusión temporal, y una herramienta

de visualización de estimaciones futuras de DSV para la prevención de la sigatoka negra en Belice.

Se opta por transformadores de fusión temporal para predecir la humedad de las hojas, ya que permiten la utilización de series de tiempo pasadas y futuras [8]; en este caso, las variables futuras hacen referencia a los pronósticos del clima. A través de una validación cruzada, se optimiza la métrica error absoluto medio (MAE, por sus siglas en inglés) y posteriormente, se evalúa de forma categórica utilizando *precision*, *recall* y F1-score. Aunque se reconoce que la precisión de los pronósticos climáticos introduce un margen de error en los resultados, el objetivo es mejorar la exactitud de las predicciones basándose en el modelo DSV existente. A pesar de los desafíos relacionados con la exactitud de los pronósticos climáticos, se busca ofrecer estimaciones futuras con la mayor certeza posible mediante el uso de técnicas de *machine learning*. Este trabajo no pretende mitigar el error inherente de los pronósticos climáticos, sino en cambio utilizarlos para poder predecir lo que estos pronósticos no consideran, es decir, la humedad de las hojas. Se espera que estas estimaciones, junto con el tablero de visualización desarrollado, puedan formar parte de una estrategia más eficaz para el control de la sigatoka negra, lo que permitirá a los agricultores y gestores de cultivos tomar decisiones más informadas y oportunas.

CAPÍTULO 2

Justificación

La sigatoka negra es una enfermedad que afecta a las musáceas, especialmente a las variedades del subgrupo cavendish. Para el control integrado de la enfermedad en las plantaciones de banano, la aplicación constante de fungicidas se ha convertido en una herramienta esencial. Sin embargo, sin un combate químico adecuado, la enfermedad puede causar pérdidas superiores al 50 %, afectando la fotosíntesis, respiración y transporte de nutrientes en las plantas, lo que reduce la vida verde de la fruta [9].

La sigatoka negra se encuentra en la mayoría de las zonas tropicales y subtropicales donde se cultivan musáceas. Las variedades de plátano y banano más importantes son susceptibles a esta enfermedad, que en condiciones favorables, puede reducir significativamente los rendimientos hasta en un 50 %[1], [3], [4]. El combate de esta enfermedad constituye uno de los principales rubros para la industria bananera, representando hasta un 27 % del costo total de producción [5].

En el continente americano y el Caribe, la enfermedad se ha dispersado ampliamente, causando epidemias y obligando a intensificar las medidas de combate ante esta enfermedad. Esto ha aumentado la necesidad de desarrollar estrategias de manejo integrado de la enfermedad para mitigar su impacto negativo en la industria bananera [2]. Los fungicidas sistémicos son bastante eficaces para combatir la sigatoka negra, sin embargo, tiene efectos negativos contra el medio ambiente [7].

El alto impacto negativo que tiene este hongo en la industria bananera, así como el efecto negativo causado al medio al medio ambiente por los fungicidas, justifican el por qué es importante conocer en qué días es necesario aplicar fungicidas. En Belice, la WWF utiliza un sistema de pronóstico de enfermedades que, junto con el monitoreo de niveles de enfer-

medad y condiciones climáticas, ayuda a los administradores de plantaciones a programar las aplicaciones de fungicidas de manera óptima, reduciendo la cantidad de químicos necesarios. Este modelo estima el riesgo que tiene de desarrollarse la enfermedad en plantaciones de banano, basado en las condiciones microclimáticas. A este valor de riesgo diario es lo que define la WWF como el *disease severity value* (DSV) [6].

Este trabajo busca mejorar el modelo utilizado por la WWF al complementarlo con una predicción del DSV en días futuros por medio de un modelo de *machine learning*, así como el desarrollo de un herramienta visual que permitan llevar el control. Con estas estimaciones se puede conocer si las variables climatológicas van a favorecer al desarrollo de la enfermedad en un futuro cercano, y en base a esta información tomar medidas preventivas. Las limitaciones del modelo actual es la utilización de variables climatológicas de los últimos días, y no considera el clima de los siguientes. Es importante reconocer que existen diferentes tipos de fungicidas, y que no todos son utilizados para prevención. Según la WWF, los fungicidas utilizados en el control de la sigatoka negra se dividen en protectores, sistémicos y aceites. Los fungicidas protectores no penetran la superficie de la hoja y se utilizan en etapas tempranas de la enfermedad o como preventivos. Los fungicidas sistémicos, en cambio, penetran la hoja y se usan después de que las infecciones han ocurrido.

Los modelos de *machine learning* han demostrado ser efectivos en múltiples aplicaciones de gestión de cultivos, principalmente en la predicción de rendimientos y la detección de enfermedades [10]. La implementación de modelos de *machine learning* y sistemas de inteligencia artificial en agricultura ofrecen mejores recomendaciones para la toma de decisiones, mejorando la producción y la calidad de los productos biológicos. Por este motivo, se busca aprovechar las técnicas de *machine learning* para la predicción del DSV en las plantaciones de banano en Belice. Esto permite optimizar la aplicación de fungicidas, reducir costos y minimizar el impacto ambiental, además de mejorar la eficiencia en la gestión de la sigatoka negra en las plantaciones de banano en Belice.

CAPÍTULO 3

Objetivos

3.1. Objetivo general

Predecir los valores futuros del *disease severity value* (DSV) para mejorar la toma de decisiones en el control de la sigatoka negra en Belice, mediante el desarrollo de un modelo de *machine learning* que utilice variables climatológicas de las plantaciones de banano en Belice y una herramienta de visualización de las predicciones del modelo.

3.2. Objetivos específicos

- Extraer datos históricos y pronósticos del clima para calcular las variables y el DSV mediante fuentes meteorológicas.
- Crear el dataset para el entrenamiento mediante la definición de la unidad de análisis, creación de variables, cálculo del DSV histórico y limpieza de datos.
- Desarrollar y optimizar un modelo de *machine learning* supervisado de regresión para estimar el DSV a futuro mediante la utilización de un *time split cross validation* como técnica de evaluación.
- Desarrollar una herramienta de visualización para presentar los resultados mediante un tablero interactivo que permita al usuario ver y analizar los pronósticos de DSV en los días futuros, así como los valores históricos.

CAPÍTULO 4

Alcance

Este proyecto aborda diversas áreas de ciencia de datos. Comienza con procesos de extracción y transformación de datos, asegurado que estos estén listos para su análisis. Posteriormente, se realiza un análisis de datos detallado de las variables obtenidas por diversas fuentes, para comprender patrones, tendencias y datos atípicos. Además, se desarrollan modelos predictivos utilizando transformadores de fusión temporal para predecir los valores de humedad de las hojas, que permiten estimar los valores futuros de DSV. Finalmente, el proyecto incluye la visualización de datos mediante un tablero interactivo, el cual facilita la interpretación y el seguimiento de los resultados para los usuarios interesados.

Los resultados del modelo dependen en gran medida de la disponibilidad y calidad de los datos proporcionados por la WWF. Dado que los datos de entrada necesarios para el modelo son administrados por esta organización, cualquier limitación en la cantidad, calidad o frecuencia de actualización de estos datos impactará directamente en el rendimiento y precisión del modelo. Además, el modelo no será puesto en producción, lo que limita la evaluación a un entorno de prueba y demostración. Esto debido a que la WWF aún no ha decidido qué herramienta utilizará para automatizar el procesamiento de datos. Actualmente, no disponen de ninguna herramienta o sistema para automatizar la ingesta de datos, la ejecución del modelo y la actualización del tablero de visualización. En consecuencia, el tablero tampoco mostrará predicciones en tiempo real, ya que, por el momento, los datos de la WWF solo están disponibles a través de descargas manuales de archivos en formato csv.

Otra limitación importante es que el error inherente a los pronósticos climáticos no se busca mitigar en este proyecto, lo que implica que este error se propagará a los resultados del modelo. Esto es relevante, dado que los pronósticos de variables climáticas son una parte

crucial de los insumos para el modelo. El modelo fue entrenado exclusivamente con datos de Centroamérica y está diseñado para predecir los valores futuros de DSV específicamente en Belice. Como resultado, su efectividad en otras regiones no se considera en el alcance de este trabajo.

CAPÍTULO 5

Marco teórico

5.1. ¿Qué es la sigatoka negra?

La sigatoka negra es una enfermedad que afecta a las plantas de banano y plátano, comienza con pequeñas manchas cloróticas en la parte inferior de las hojas jóvenes. Estas manchas evolucionan a rayas marrones y negras, a veces con un tono púrpura, y los tejidos adyacentes pueden parecer encharcados en condiciones de alta humedad [1].

Las enfermedades de sigatoka en bananos son causadas por dos hongos, *Mycosphaerella Fijiensis* (que causa la sigatoka negra) y *M. Musicola* (que causa la sigatoka amarilla). Se pueden distinguir morfológicamente por las características de sus conidios y conidióforos. *M. Fijiensis* tiene conidios más largos y flexuosos, y conidióforos alargados, mientras que *M. Musicola* tiene conidios más cortos y conidióforos con forma de botella. También se han desarrollado métodos moleculares para identificarlos [7].

El hongo *Mycosphaerella Fijiensis* se dispersa principalmente mediante ascosporas y, en menor medida, conidios. Los conidios, que se forman en condiciones húmedas y en la presencia de una capa delgada de agua en las hojas, se dispersan durante la lluvia. Los pseudotecios, formados en hojas muertas saturadas de agua, producen ascosporas que se dispersan a largas distancias en condiciones de alta humedad. Se reproduce de manera asexual en lesiones tempranas, con conidios dispersos por salpicaduras de lluvia, y de manera sexual en lesiones más maduras, donde los pseudotecios liberan ascosporas que se dispersan a largas distancias por el aire durante períodos de alta humedad [5], [11]-[13].

5.2. Impacto de la sigatoka negra en la agricultura

La sigatoka negra es altamente perjudicial para la agricultura de bananos y plátanos. Bajo condiciones favorables para la enfermedad, la necrosis de las hojas puede reducir los rendimientos en un 35-50 %. Muchas de las variedades más importantes y comúnmente cultivadas son susceptibles a esta enfermedad. En 1995, el costo promedio para controlar la sigatoka negra era de US$1500 por hectárea al año, con 38-50 aplicaciones de fungicidas necesarias anualmente. Estos fungicidas pueden representar aproximadamente el 30 % de los costos de producción. En Centroamérica, la sigatoka negra puede constituir hasta el 27 % del costo total de producción, en comparación con solo el 3-5 % para otras enfermedades y plagas [1].

Para mantener la calidad del fruto durante el transporte, es esencial que la planta tenga al menos cinco hojas en el momento de la cosecha. Los frutos de plantas gravemente afectadas tienden a madurar de manera prematura e irregular, lo que significa un problema para los exportadores que deben cumplir con los estándares de calidad de los mercados en países desarrollados.

5.3. Métodos de control contra la sigatoka negra

Las plantaciones grandes dependen de fungicidas como el mancozeb y el clorotalonil, a menudo combinados o alternados con fungicidas sistémicos como las morfolinas, DMI y estrobilurinas. Sin embargo, existe resistencia a estos fungicidas en muchas áreas de producción, lo que reduce su eficacia. Es crucial seguir las recomendaciones del Comité de Acción Contra la Resistencia a Fungicidas (FRAC) para evitar la resistencia [13].

Además de los controles químicos existen otras alternativas que también ayudan a combatir esta enfermedad, entre ellas se encuentran el control biológico, cultivos resistentes y manejo cultural. El control biológico de la sigatoka negra es complicado debido a la naturaleza continua de la enfermedad. Se han probado bacterias epifíticas como *Pseudomonas*, *Bacillus* y *Serratia spp.*, pero la investigación aún está en etapas preliminares [1]. Desarrollar cultivos resistentes a la enfermedad es otra alternativa, sin embargo, el desarrollo de cultivos resistentes que al mismo tiempo sean aceptables es complicado porque muchos no cumplen con los gustos locales. Desarrollar cultivos resistentes y aceptables es una prioridad importante en los centros de investigación internacionales.

5.4. Modelo de *disease severity value*

El modelo utilizado por la WWF para el control de la sigatoka negra (*Mycosphaerella Fijiensis*) en plantaciones de banano se basa en un modelo en línea que estima el riesgo diario de la enfermedad. Este riesgo, representado como el *disease severity value* (DSV), considera factores microclimáticos que influyen en la germinación de ascosporas y el crecimiento del hongo.

El DSV se desarrolló utilizando una metodología de matriz, similar a la empleada por Wallin en 1962 [14] para la roya de la papa. Se crearon matrices que evaluaban la humedad de las hojas, la humedad relativa y la presencia lluvia en tres rangos de temperatura [6].

Estos valores ayudan a determinar cuándo aplicar fungicidas sistémicos y protectores. Durante los meses más secos, un promedio semanal de DSV puede justificar la no aplicación de fungicidas protectores, lo que representa un ahorro significativo en costos de producción. La determinación del DSV se basa en matrices que combinan rangos de temperatura y horas con un índice de favorabilidad ambiental (EFI, por sus siglas en inglés).

5.4.1. Matriz LW (humedad de las hojas y temperatura)

La matriz LW genera un valor $\mathrm{DSV_{LW}}$ basado en la humedad de las hojas y la temperatura, utilizando datos de estaciones meteorológicas. Se cuenta el total de horas con humedad en las hojas y se multiplica por el EFI correspondiente al rango de temperatura. A continuación, se presentan los valores por hora y temperatura:

Cuadro 1.

Matriz DSV de humedad de las hojas

Horas	1-3	4-7	8-11	12
$T < 78°F$	10	14	18	22
$78°F \leq T \leq 82°F$	14	18	22	26
$T > 82°F$	12	16	20	24

Nota. Adaptada de [6].

5.4.2. Matrices RH (humedad relativa y temperatura)

La matrices RH generan un valor $\mathrm{DSV_{RH}}$ considerando la humedad relativa y la temperatura. Se cuenta el total de horas con humedad relativa en los rangos especificados y se multiplica por el EFI correspondiente al rango de temperatura. Este riesgo se divide en dos matrices, cada una con un rango de humedad relativa distinto. La matriz del Cuadro 2 genera un valor de riesgo basado en el rango de temperatura y en la cantidad de horas con humedad relativa mayor o igual al 80 % y menor al 98 %. En cambio, la segunda matriz del Cuadro 3 calcula el valor de riesgo considerando únicamente las horas con humedad relativa superior al 98 %. La suma de ambos valores dan como resultado el $\mathrm{DSV_{RH}}$.

Cuadro 2.
Matriz DSV de humedad relativa 80-98

Horas	1-3	4-7	8-11	12
$T < 78°F$	2	6	10	14
$78°F \leq T \leq 82°F$	6	10	14	18
$T > 82°F$	4	8	12	16

Nota. Adaptada de [6].

Cuadro 3.
Matriz DSV de humedad relativa 98-100

Horas	1-3	4-7	8-11	12
$T < 78°F$	6	10	14	18
$78°F \leq T \leq 82°F$	10	14	18	22
$T > 82°F$	8	12	16	20

Nota. Adaptada de [6].

5.4.3. Matriz free water (Lluvia y temperatura)

Esta matriz genera un valor $\mathrm{DSV_{Free-Water}}$ basado en la cantidad de horas de lluvia y la temperatura. Se cuenta el total de horas con condiciones de lluvia y se multiplica por el EFI correspondiente al rango de temperatura. Los valores por hora y temperatura son los siguientes:

Cuadro 4.
Matriz DSV de lluvia

Horas	1-3	4-7	8-11	12
$T < 78°F$	10	14	18	22
$78°F \leq T \leq 82°F$	14	18	22	26
$T > 82°F$	12	16	20	24

Nota. Adaptada de [6].

5.5. Aprendizaje supervisado

Machine learning (ML) se define como un campo de la inteligencia artificial que permite a las computadoras aprender y mejorar a partir de la experiencia sin ser explícitamente programadas para lo que hacen. Una definición ampliamente aceptada de ML, propuesta por Tom Mitchell, establece que "un programa de computadora se dice que aprende de la experiencia E en relación con alguna clase de tareas T y una medida de desempeño P, si su desempeño en las tareas en T, medido por P, mejora con la experiencia E." [15]. Existen diferentes tipos de aprendizaje automático, dependiendo de la naturaleza de las tareas se define la medida de desempeño y la experiencia proporcionada al sistema. Los modelos de ML se dividen en tres categorías: aprendizaje supervisado, no supervisado y de refuerzo.

En el aprendizaje supervisado, el modelo se entrena utilizando un conjunto de datos etiquetados, es decir, que se tienen datos de entrenamiento para la salida esperada del modelo. El objetivo es aprender una función que mapee las entradas a las salidas con la mayor precisión posible [16]. El aprendizaje supervisado se define formalmente como la tarea de aprender una función f que mapea entradas $x \in X$ a salidas $y \in Y$, donde x representa las características de los datos y y los resultados esperados. Matemáticamente, se considera un conjunto de entrenamiento $D = \{(x_n, y_n)\}_{n=1}^{N}$, donde N es el tamaño de la muestra, es decir, que es una secuencia de características con su respectiva salida [15].

El modelo se ajusta a este conjunto de datos mediante la minimización de una función de pérdida que mide el error que existe entre las predicciones del modelo y los valores reales. Este proceso de ajuste continúa hasta que el modelo alcanza un rendimiento óptimo.

5.6. Métricas de regresión

Para los modelos de regresión existen varias métricas para evaluar qué tan bueno es su rendimiento, cada una de estas métricas tiene sus ventajas y desventajas. Una de las métricas más comunes es el *mean squared error* (MSE), como se menciona en la sección de regresión,

se tiende a utilizar para el proceso de optmización de modelos [15]. Sin embargo, para la evaluación de modelos, también se utilizan otras métricas como lo son el *root mean squared error* (RMSE), *mean absolute error* (MAE), *mean absolute percentage error* (MAPE) y el coeficiente de determinación R^2.

El RMSE se calcula obteniendo la raíz cuadrada del MSE. Ambas métricas son sensibles a grandes errores debido a que estos errores se elevan al cuadrado, lo que significa que errores grandes tienen un impacto muy alto en el valor final, por lo que es útil cuando se desea enfatizar grandes errores en el modelo [17]. El MAE mide el promedio de los errores absolutos entre las predicciones y los valores reales. A diferencia del MSE, no penaliza tanto los errores grandes, lo que lo hace más robusto en presencia de valores atípicos. Tanto el RMSE como el MAE aseguran que la diferencia entre la predicción y el valor real sea positiva, el RMSE (y el MSE) lo hacen al elevarla al cuadrado, mientras que el MAE lo hace al calcular el valor absoluto. La fórmula del MAE es la siguiente:

$$\text{MAE} = \frac{1}{N}\sum_{n=1}^{N}|y_n - f(x_n)| \tag{1}$$

Otra medida que se utiliza es el MAPE, que mide el error porcentual medio absoluto. Esto permite interpretar los errores como porcentajes, haciendo que sea una métrica más interpretable que el RMSE o MAE, sin importar el conjunto de datos utilizado para el entrenamiento. Sin embargo, tiene la limitación de que puede ser inestable si los valores reales se acercan a cero, dado que el porcentaje de error se dispara. Su cálculo es similar al del MAE, pero la resta entre el valor real y la predicción se divide por el valor real para que se interprete como un porcentaje. Su fórmula es la siguiente:

$$\text{MAPE} = \frac{1}{N}\sum_{n=1}^{N}\left|\frac{y_n - f(x_n)}{y_n}\right| \tag{2}$$

Otra de las métricas utilizadas para regresión es el R^2, también conocido como coeficiente de determinación, que compara la variabilidad de las predicciones del modelo con la variabilidad de la variable dependiente. Su fórmula es la siguiente:

$$R^2 = 1 - \frac{\sum_{n=1}^{N}(y_n - f(x_n))^2}{\sum_{n=1}^{N}(y_n - \bar{y})^2} \tag{3}$$

donde $\bar{y}$ es la media de la variable a predecir. El valor máximo del R^2 es 1, lo cual indica que el modelo predice correctamente el resultado exacto para todas las observaciones. Un valor 0 en esta métrica puede interpretarse como que el modelo tiene el mismo error que un

modelo que predice la media para todas las observaciones, y un valor negativo significa que el modelo tiene un rendimiento inferior a este modelo hipotético.

También existe la métrica R^2 ajustado, que toma en cuenta la cantidad de variables independientes incluidas. A diferencia del R^2, esta métrica penaliza la inclusión de variables innecesarias [18]. Esto se logra ajustando la fórmula para tener en cuenta el número de predictores d y el tamaño de la muestra N, cuya fórmula es la siguiente:

$$R^2_{\text{ajustado}} = 1 - \frac{\left(\frac{\sum_{n=1}^{N}(y_n - f(x_n))^2}{N-d-1}\right)}{\left(\frac{\sum_{n=1}^{N}(y_n - \bar{y})^2}{N-1}\right)} \tag{4}$$

La idea detrás del R^2 ajustado es que, después de haber incluido todas las variables relevantes, la adición de variables adicionales solo contribuiría a un pequeño descenso en el numerador. Dado que el denominador de la fórmula incluye el número de predictores, esto podría resultar en una disminución del R^2 ajustado. En teoría, el modelo con el mayor R^2 ajustado incluirá todas las variables relevantes y ninguna irrelevante.

5.7. Métricas de clasificación

Existe una gran cantidad de métricas para modelos de clasificación, algunos utilizan las probabilidades del modelo como entrada y otras utilizan la clasificacion (0 o 1) para medir el rendimiento del modelo. Entre las métricas más utilizadas son *precision*, *recall* y *F1-score*. Estas métricas se derivan de la matriz de confusión y miden distintos aspectos del rendimiento del modelo sobre la clase positiva (la clase de interés) [19].

El ***precision*** es la proporción de observaciones correctamente clasificadas como positivas sobre el total de observaciones asignadas como positivas por el modelo. La fórmula del *precision* se define como el cociente entre los verdaderos positivos (TP), y la suma de verdaderos positivos y falsos positivos (FP):

$$precision = \frac{TP}{TP + FP} \tag{5}$$

Este valor indica que dado de que el modelo de una predicción positiva, la probabilidad que realmente sea positiva esa observación.

El ***recall***, también conocido como sensibilidad, es la proporción de observaciones correctamente clasificadas como positivas sobre el total de observaciones positivas, es la proporción de observaciones correctamente clasificadas como positivas sobre el total de observaciones

asignadas como positivas por el modelo. Se calcula como el cociente entre los verdaderos positivos y la suma de verdaderos positivos y falsos negativos (FN):

$$recall = \frac{TP}{TP + FN} \tag{6}$$

Esta métrica indica que dado de que es positivo, cual es la probabilidad que el modelo lo reconozca correctamente como positivo.

El **F1 score** es la media armónica de la *precision* y el *recall*, combina ambas para proporcionar un único valor que refleje el rendimiento general del modelo, especialmente útil en situaciones de desbalance de clases. Se define como:

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{7}$$

La media armónica en el F1 score pondera los valores bajos de *precision* o *recall* más significativamente que la media aritmética, resaltando las debilidades del modelo si alguna de las dos métricas es baja.

5.8. *Cross validation*

El proceso de optimización de ML requiere de tomar una serie de decisiones hasta llegar al modelo óptimo, para lo cual es importante evaluar todas las iteraciones del modelo para saber qué decisiones tomar para la optimización. Para evaluar el rendimiento del modelo en la vida real, se debe evaluar con un subconjunto de datos distinto que no contenga datos con los que el modelo se entrenó, y que tampoco se hayan utilizado para la toma de decisiones en el proceso de optimización, a este subconjunto de datos se le conoce como el dataset de prueba [20]. Adicionalmente, en el proceso de optimización se requiere evaluar el rendimiento de las iteraciones del modelo con datos que no se hayan utilizado para entrenar, y que tampoco se vayan a utilizar en el dataset de prueba, esto para poder estimar el error en el dataset de prueba.

Una de estas técnicas implica dividir aleatoriamente las observaciones en un conjunto de entrenamiento y un conjunto de validación [18]. El modelo se ajusta con los datos de entrenamiento y su desempeño se evalúa con los datos de validación, utilizando la métrica que se adapte al problema que se quiere solucionar. Sin embargo, este enfoque presenta algunas limitaciones, principalmente el hecho de que solo una parte de los datos se utiliza para evaluar el modelo, lo que puede llevar a un sobreajuste a esos datos seleccionados, y obtener un modelo que solo es el mejor con ese dataset de validación específico.

Para abordar estas limitaciones, se puede utilizar la técnica de *cross validation*, la cual mejora la estimación del error de prueba y mitiga los riesgos asociados con la selección de un solo subconjunto de datos de validación. Una forma de implementar esta técnica es el k-fold *cross validation*, el cual divide el conjunto de observaciones en k subconjuntos con aproximadamente la misma cantidad de observaciones [18]. El método se ajusta en k-1 subconjuntos y se evalúa en el subconjunto de datos restante, repitiendo este proceso k veces y promediando los resultados para obtener una estimación más precisa del error de prueba. También existe el método *time split cross validation* (validación cruzada con división temporal), el cual sigue el mismo proceso que el k-folds, a diferencia que los subconjuntos de datos se separan por orden de tiempo y no de forma aleatoria [21]. De esta forma siempre se evalúa con datos futuros a modelos que únicamente se entrenan con datos previos, simulando así el entorno de producción, donde las observaciones nuevas pertenecen a un espacio de tiempo distinto al de entrenamiento y validación.

5.9. Redes neuronales

Las redes neuronales feedforward son un tipo de red neuronal donde la información va en una sola dirección, empezando desde la capa de entrada, pasando por las capas ocultas, hasta llegar a la capa de salida. En estas redes, cada neurona de una capa está conectada a todas las neuronas de la capa siguiente, pero no hay conexiones entre neuronas de la misma capa ni con las capas anteriores. Cada neurona en la capa de entrada representa una variable del modelo. En las capas ocultas y en la capa de salida, las neuronas son combinaciones lineales de las salidas de las neuronas de la capa anterior, a las que se les aplica una función de activación no lineal, como tanh o ReLU, con el objetivo de añadir no linealidad al modelo. La red se entrena minimizando una función de costo mediante algoritmos de optimización como el descenso de gradiente estocástico, *RMSProp*, *AdaBoost*, entre otros [22].

Por otro lado, las redes neuronales recurrentes (RNNs, por sus siglas en inglés) están diseñadas para procesar secuencias de datos, como lo pueden ser las series de tiempo o texto. A diferencia de las redes feedforward, las RNNs tienen conexiones de retroalimentación que permiten a las neuronas mantener una memoria de los estados anteriores en el tiempo. Esto permite que la red capture dependencias de una misma variable a lo largo del tiempo, es decir, que se utilizan los estados anteriores para predecir los estados futuros de dicha variable. El hecho de que las RNNs compartan parámetros para diversos inputs asegura que el modelo pueda generalizar a secuencias de diferentes longitudes y posiciones, mejorando así su capacidad para manejar datos secuenciales. La capacidad de las RNNs para mantener información a lo largo de una secuencia es una herramienta importante para tareas como el procesamiento de lenguaje natural y el análisis de series temporales [22].

5.10. LSTM

El artículo original de las LSTM, *Long Short-Term Memory* de Sepp Hochreiter y Jürgen Schmidhuber (1997), introduce una arquitectura de red neuronal recurrente (RNN) que supera las limitaciones que tenían las RNNs tradicionales en cuanto a la retención de información a largo plazo. Las RNNs convencionales tienen dificultades para aprender dependencias temporales a medida que el error del *backpropagation* tiende a volverse muy pequeño, lo que impide que se mantenga información relevante de eventos pasados en secuencias largas [23]. Las LSTM resuelven este problema mediante una estructura de celdas de memoria que regulan el flujo de información a través de tres componentes: la puerta de entrada, la puerta de olvido y la puerta de salida. Estos componentes permiten que el modelo decida qué información conservar y que información descartar en cada paso de la serie de tiempo, optimizando así tanto la memoria a corto como a largo plazo, de ahí obtiene el nombre.

Las principales ventajas de las LSTM incluyen su capacidad para manejar dependencias a largo plazo, lo cual es de gran importancia en tareas como el modelado de secuencias de texto o series temporales [23]. A diferencia de las RNNs tradicionales, las LSTM logran preservar la información relevante durante largos periodos, evitando que el gradiente se vuelva muy pequeño y mejorando el rendimiento en tareas complejas[23]. Según el artículo original, las LSTM representan un avance significativo al permitir la memoria a largo plazo, lo que facilita el aprendizaje en dominios como el procesamiento de lenguaje natural o predicciones de series de tiempo.

5.11. Transformadores

En 2017, el artículo *Attention is All You Need* de Vaswani et al. introduce la arquitectura de transformadores, que ofrece una alternativa a las RNNs tradicionales y LSTM, al emplear un mecanismo de atención para procesar secuencias. La atención permite que el modelo determine qué partes de la secuencia de entrada son más relevantes en cada momento, asignando pesos específicos a cada elemento [24]. Este mecanismo elimina la necesidad de procesar las secuencias de manera secuencial, lo que facilita la paralelización y mejora la eficiencia computacional. Un componente de estas redes es el *multi-head attention*, el cual permite al modelo enfocarse en diferentes relaciones entre palabras en la secuencia de forma simultánea, capturando dependencias complejas a largo plazo de manera más efectiva que las RNNs o las LSTM.

Los transformadores constan de dos componentes principales: el *encoder* (codificador) y el *decoder* (decodificador) [24]. El *encoder* procesa la secuencia de entrada y crea representaciones vectoriales en base al contexto, mientras que el *decoder* utiliza estas representaciones

para generar la salida del modelo. Ambas partes utilizan capas de atención (*multi-head attention*) y capas densas de redes neuronales [24].

5.12. Temporal fusion transformers

El modelo *temporal fusion transformer* (TFT) es una arquitectura basada en transformadores aplicada para pronósticos (*multi-horizon*), es decir que se adapta a distintos horizontes de predicción. Los TFT son capaces de utilizar diversos tipos de entradas, y tratar los datos dependiendo e su tipo de dato, ya sean variables estáticas, variables desconocidas del futuro, o variables conocidas del futuro [8]. Otra de las motivaciones para diseñar esta arquitectura fue poder proporcionar interpretabilidad a modelos complejos, lo cual es sí es posible en los TFT. A continuación, se describen los principales componentes que conforman el TFT y sus funciones.

5.12.1. *Gating mechanisms*

Este componente permite que el modelo omita partes innecesarias de la arquitectura, ajustando la profundidad y complejidad según las necesidades de los datos [8]. Los *gated residual networks* (GRN) utilizan (*Gated linear units*, GLU) para calcular la contribución de cada capa. Dado un vector de entrada γ, el GLU se calcula como:

$$GLU(\gamma) = \sigma(W_4\gamma + b_4) \odot (W_5\gamma + b_5)$$

donde σ es la función sigmoide y $\odot$ representa el producto elemento a elemento [8]. Esto le permite al TFT suprimir capas no necesarias, optimizando la eficiencia y adaptándose a la complejidad de los datos.

5.12.2. *Variable selection networks*

Las redes de selección de variables ayudan al TFT a enfocarse en las variables más relevantes en cada paso temporal [8]. Esto se logra mediante la asignación de pesos de selección para cada variable en función de su importancia, utilizando un GRN seguido de una función *softmax*. La red de selección de variables filtra el ruido y enfoca el aprendizaje en las características más relevantes, mejorando el rendimiento del modelo.

5.12.3. *Static covariate encoders*

Este componente integra características estáticas invariantes en el tiempo mediante el uso de GRN para generar vectores de contexto [8]. Estos vectores de contexto condicionan dinámicas temporales dentro del modelo, influyendo en la selección de variables y en el procesamiento temporal. De esta forma, características estáticas como la ubicación geográfica afectan las predicciones a lo largo de toda la serie temporal, permitiendo al modelo capturar información global.

5.12.4. Procesamiento temporal

El procesamiento temporal combina capas *sequence-to-sequence* y un *multi-head attention* interpretable [8]. La capa *sequence-to-sequence* captura relaciones locales a corto plazo en los datos, mientras que el bloque de *multi-head attention* permite aprender dependencias a largo plazo en la secuencia. El (*multi-head attention*) se define como:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_{attn}}}\right)V$$

donde Q, K, y V son las matrices de *queries*, *keys* y *values*, respectivamente. Este mecanismo permite que el TFT se enfoque en diferentes partes de la secuencia en cada *head*, capturando patrones temporales más complejos.

5.12.5. Intervalos de predicción con pronósticos de cuantiles

El TFT genera intervalos de predicción en lugar de valores puntuales mediante pronósticos de cuantiles, produciendo diferentes percentiles (por ejemplo, 15 %, 50 % y 85 %) para cada horizonte de predicción [8]. Estos pronósticos con cuantiles se calculan aplicando una transformación lineal sobre la salida del *temporal fusion decoder*, lo cual permite estimar un rango probable para los valores futuros y facilita el análisis de riesgos en diversas aplicaciones.

CAPÍTULO 6

Metodología

6.1. Extracción de datos

La primera fase del proyecto fue la extracción de datos, la cual comenzó con su descarga de las estaciones climatológicas de la WWF. De todas las estaciones con datos disponibles, se seleccionaron únicamente las que contaban con información sobre temperatura, humedad de las hojas, humedad relativa y precipitación, ya que son variables necesarias para el cálculo del $\mathrm{DSV}_{\mathrm{Total}}$. Para cada una, se determinó el rango de fechas de actividad de los sensores, así como el país en el que se encontraban, y sus coordenadas. Con esta información, se descargaron datos satelitales históricos del clima a nivel hora, los cuales incluían variables como temperatura, humedad relativa, precipitación, entre otras. Estos fueron obtenidos a través de un API climático disponible en Python, utilizando la plataforma Open-Meteo.com, la cual ofrece datos históricos desde 1940. Para simular el ambiente productivo, también se descargaron pronósticos climáticos históricos a nivel hora, disponibles únicamente a partir de 2022 con la misma plataforma Open-Meteo.com. Dado que los pronósticos históricos del clima no están disponibles para datos más antiguos, fue necesario simular el ambiente productivo únicamente con las estaciones con datos de sensores a partir de 2022, y con esas estaciones se evaluó en qué medida el uso de pronósticos del clima afectaba los resultados del modelo.

6.2. Análisis y limpieza de datos

El proceso de análisis y limpieza de datos se centró en asegurar la calidad y la coherencia de la información. En primer lugar, se eliminaron los valores nulos y los periodos de inactividad de los sensores de forma que se requirió que todas las variables no fueran nulas para considerarlas en el modelo. Una vez teniendo solo los datos de interés, se asignaron periodos de actividad a cada estación, de forma que cada periodo cumple con tener todos los días y horas en su rango de fechas. También se detectaron periodos en los que la media móvil de 2 meses estaba por encima del límite superior $LS = Q3 + IQR$ o por debajo del límite inferior $LI = Q1 - IQR$, donde $Q1$ y $Q3$ representan el primer y tercer cuartil de la diferencia entre los datos de los sensores y los satelitales, respectivamente, y IQR es el rango intercuartil. Este método es importante antes de calibrar los datos de las estaciones ya que el error puede cambiar a lo largo del tiempo, por lo que el proceso de calibración se realiza por periodo. Los periodos con menos del 20 % de los datos de la estación y que no alcanzaban un mínimo de 3 meses, se eliminaron con el objetivo de reducir el ruido en el modelo.

Una vez asignados los periodos de cada región, se identificaron problemas de calibración en la información proporcionada por los sensores, comparando estos datos con los satelitales. Se asumió que la relación entre los datos de los sensores y los satélites podía modelarse como $y = x + \epsilon$, donde y representa la temperatura medida por el sensor, x es la temperatura proporcionada por los satélites y ϵ es un error aleatorio. En los casos donde este supuesto no se cumplió, se entrenó una regresión lineal para encontrar los coeficientes m y b, que permitieron ajustar las mediciones de los sensores transformando y a $\frac{y-b}{m}$. Los periodos con un R^2 inferior a 0.60 después de la calibración fueron eliminados, ya que esto es un indicador de que el sensor no tenía mucha correlación con los datos satelitales, siendo estos periodos atípicos que introducirían ruido al modelo en caso se incluyeran. Finalmente, se detectaron y corrigieron datos atípicos en cada una de las variables, usando la diferencia entre los sensores y los satélites como referencia. Los valores atípicos fueron aquellos que superaban el límite superior $LS = Q3 + 1.5(IQR)$ o que estaban por debajo del límite inferior $LI = Q1 - 1.5(IQR)$, esto utilizando la diferencia $y - x$ como variable a monitorear.

6.3. Construcción de los datasets

En la fase de construcción de los datasets, se seleccionó toda la información de Belice como el conjunto de datos de prueba (*test*), dado que este país es el objetivo principal del modelo. Luego, se eligieron estaciones con datos de sensores posteriores a 2022 para crear un segundo conjunto de prueba, denominado *prod*, que sirvió para comparar los resultados del modelo utilizando información histórica satelital con aquellos utilizando pronósticos climá-

ticos históricos, ya que estos datos solo estaban disponibles a partir de 2022. Las estaciones sin información a partir de este año y que no pertenecían a Belice fueron seleccionadas para formar el conjunto de entrenamiento y validación, denominado *train*. Al finalizar la validación cruzada, estos datos fueron utilizados para entrenar el modelo final.

Se creó una tabla base sobre la cual aparecen cada una de las estaciones en todas las horas de rango de fecha activo. A esta tabla base se le agregaron las variables climáticas de los satélites y los datos limpios de los sensores para los datasets *train*, *test* y *prod*. También se creó el conjunto de datos denominado *prod_forecasts* que contiene los mismos datos que *prod*, a diferencia que para este se utilizan pronósticos del clima históricos en lugar de la información real satelital. Esto con el objetivo de poder comparar los resultados de los datasets *prod* y *prod_forecasts* para analizar si los resultados cambian significativamente a la hora de utilizar pronósticos del clima.

6.4. Desarrollo del modelo

El modelo fue desarrollado utilizando un transformador de fusión temporal (TFT, por sus siglas en inglés), que fue implementado con la librería PyTorch Forecasting. El modelo tenía como objetivo predecir la humedad de las hojas en las próximas 72 horas, el resto de las variables climáticas pueden ser obtenidas a través de los pronósticos del clima. Para la optimización de los hiperparámetros se utilizó una validación cruzada con división temporal (*time split cross validation*), para asegurar que siempre se evaluara con información posterior a la de entrenamiento. De esta forma se realizaron 3 *splits*, en donde en cada uno de estos se seleccionó un subconjunto del dataset *train* para entrenar y otro para validar. Como en este caso se utilizó un *time split cross-validation*, esto significa que en el primer *split* se ignoraron las últimas observaciones de cada estación, las cuales no se utilizan para entrenar ni para validar. En el segundo *split* se acumuló la validación del primer *split* como entrenamiento, y se validó con los siguientes datos. Finalmente, en el *split* 3 se acumuló la validación del *split* 2 para entrenar, y se validó con los siguientes datos.

Se seleccionó la combinación de hiperparámetros con el mejor MAE, ya que en este caso era importante únicamente que el modelo acierte en que sea mayor a 3 o no, por lo tanto, no era adecuado aplicar penalizaciones más severas a los errores grandes. No se trató como un problema de clasificación debido a ser un modelo de series de tiempo, y si se hubiera reducido a una representación de ceros y unos se hubiera perdido información importante para predecir las siguientes horas. Después de la selección de los hiperparámetros, se entrenó el modelo con todos los datos del dataset *train*. Para evaluar el rendimiento del modelo, se calculó el MAE y R^2 sobre las predicciones y la humedad de las hojas real. También se calculó el *precision*, *recall* y *F1-score* sobre las versiones categóricas para cada uno de los datasets.

También se realizaron gráficas de series tiempo para las últimas observaciones disponibles de cada estación, comparando las predicciones con los valores reales de humedad de las hoja, con el objetivo de entender de forma visual el comportamiento del modelo. Adicionalmente, se obtuvieron las importancias de las variables, lo cual ayuda para entender el comportamiento del modelo TFT final.

6.5. Cálculo de DSV sobre predicciones

Con las predicciones de humedad de las hojas generadas por el modelo y los valores históricos obtenidos de los satélites, se procedió al cálculo de los valores de DSV_{LW} para cada estación. Las métricas MAE y R^2 fueron calculadas comparando los valores de DSV obtenidos de las predicciones con aquellos calculados usando datos de los sensores. Finalmente, los resultados a nivel hora fueron exportados para ser utilizados en Power BI, con el fin de realizar análisis más detallados y visualizaciones interactivas.

6.6. Tablero de visualización

Los resultados del modelo fueron simulados para los últimos días de cada región, y luego concatenados con los valores históricos para ser exportados a Power BI. Se desarrolló un tablero interactivo en Power BI para presentar los resultados de las predicciones del DSV. Se incluyeron filtros para seleccionar la plantación de interés y los rangos de fecha que se quieren considerar. Además, se calcularon métricas de Power BI para cada uno de los valores de DSV y para el valor total, $\text{DSV}_{\text{Total}}$. Se incorporaron gráficos que muestran la información de DSV en los últimos días, las predicciones para los próximos tres días, y también las medias móviles de 7 días, tanto hacia atrás como centradas (4 días hacia atrás y los 3 días del futuro). También se incluyó una selección dinámica para que la gráfica muestre los datos del DSV que le interese, ya sea el DSV_{LW}, $\text{DSV}_{\text{Total}}$, etc. Se creó otra página del tablero, específicamente para mostrar datos climáticos a nivel hora, incluyendo la humedad de las hojas, temperatura, humedad relativa y precipitación, así como los pronósticos para las siguientes 72 horas.

CAPÍTULO 7

Resultados

7.1. Extracción de datos

La extracción de datos de la fuente de WWF, filtrando únicamente las regiones con las variables de interés, resultó con información de estaciones de Belice, Guatemala y Honduras. En los Cuadros 7, 6, 5 y 8, se encuentran todas las estaciones utilizadas para el entrenamiento y evaluación del modelo TFT.

Cuadro 5.
Estaciones climatológicas de Belice

Estación	País	Fecha mínima	Fecha máxima	Días disponibles
41997	Belice	2011-07-14	2021-03-16	3394
510979	Belice	2013-04-13	2021-03-17	2638
41929	Belice	2008-04-05	2015-09-16	2359
39738	Belice	2008-05-30	2017-06-07	2720
862331	Belice	2017-06-28	2021-03-17	1359

Nota. Elaboración propia.

Cuadro 6.

Estaciones climatológicas de Guatemala

Estación	País	Fecha mínima	Fecha máxima	Días disponibles
858158	Guatemala	2015-07-19	2022-05-12	1389
378690	Guatemala	2008-08-28	2015-04-12	2307
38080	Guatemala	2008-08-28	2015-04-12	2277
860343	Guatemala	2016-04-06	2019-02-22	1053
41933	Guatemala	2009-10-29	2017-03-07	2481
378691	Guatemala	2008-08-28	2015-04-12	2257
37940	Guatemala	2008-08-28	2013-09-06	1724
41928	Guatemala	2010-12-07	2019-03-13	2090
41932	Guatemala	2010-12-07	2016-04-06	1928
41916	Guatemala	2011-01-25	2013-11-28	1001

Nota. Elaboración propia.

Cuadro 7.

Estaciones climatológicas de Honduras

Estación	País	Fecha mínima	Fecha máxima	Días disponibles
41954	Honduras	2012-11-20	2013-06-22	215
41913	Honduras	2008-02-29	2022-09-15	3808
857749	Honduras	2019-05-17	2023-12-01	37
853161	Honduras	2012-12-21	2013-12-02	346
853158	Honduras	2012-12-22	2023-05-09	3681
853156	Honduras	2012-12-21	2020-06-02	2720
853153	Honduras	2013-01-30	2019-09-02	2406
49941	Honduras	2012-12-20	2021-05-30	2876
48045	Honduras	2015-06-18	2024-10-08	3007
41917	Honduras	2011-07-12	2020-06-01	848
44165	Honduras	2010-04-29	2016-11-30	2235
44163	Honduras	2010-06-09	2018-07-11	2179
41985	Honduras	2008-05-13	2016-11-01	1899
49868	Honduras	2012-12-20	2019-09-02	2253
41910	Honduras	2008-04-19	2016-12-11	2361

Nota. Elaboración propia.

Cuadro 8.

Otras estaciones climatológicas

Estación	País	Fecha mínima	Fecha máxima	Días disponibles
45019	El Salvador	2011-05-05	2012-01-20	94
48031	Costa Rica	2011-07-28	2017-08-07	1086

Nota. Elaboración propia.

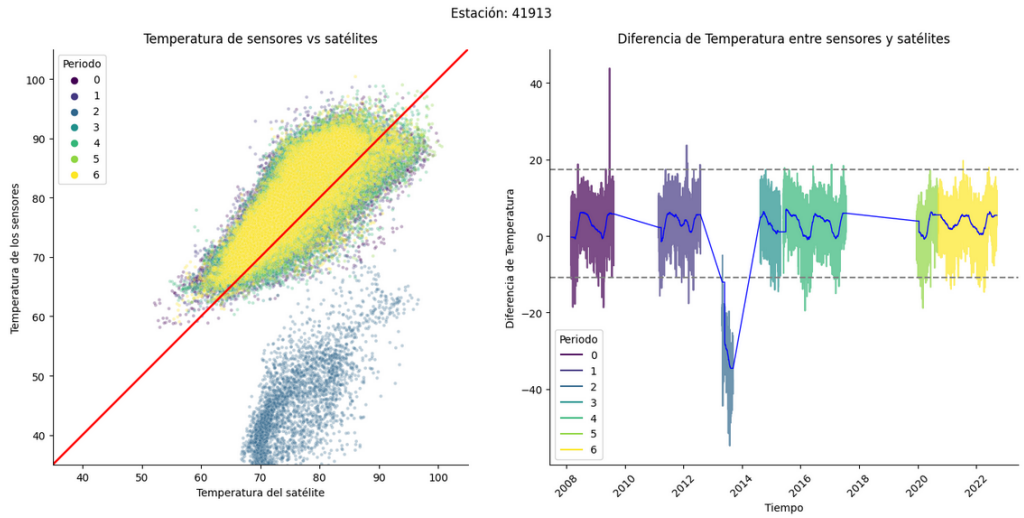
7.2. Análisis y limpieza de datos

Las Figuras 1, 18, 7 muestran los datos de temperatura luego de haber detectado los periodos, en base a la actividad de los sensores y la media móvil, específicamente de las estaciones 41913, 41997 y 49868. En la gráfica izquierda de cada figura se muestran la comparación entre los datos de sensores y los datos satelitales, mostrando con una línea roja la recta $y = x$, donde y es la temperatura de los sensores y x la de los satélites. En las gráficas de la derecha se muestra la diferencia entre ambas fuentes a lo largo del tiempo de la forma $y - x$.

Los resultados de las figuras mencionadas en el párrafo anterior aún no contemplan la calibración ni el tratamiento de datos atípicos. En las Figuras 2, 5 y 8, se muestran los resultados después de realizar la calibración, y en las Figuras 3, 19 y 9, se encuentran los resultados luego de haber realizado el tratamiento de datos atípicos. Los resultados de la limpieza de datos de las estaciones de Belice se encuentran en el Anexo.

Figura 1.

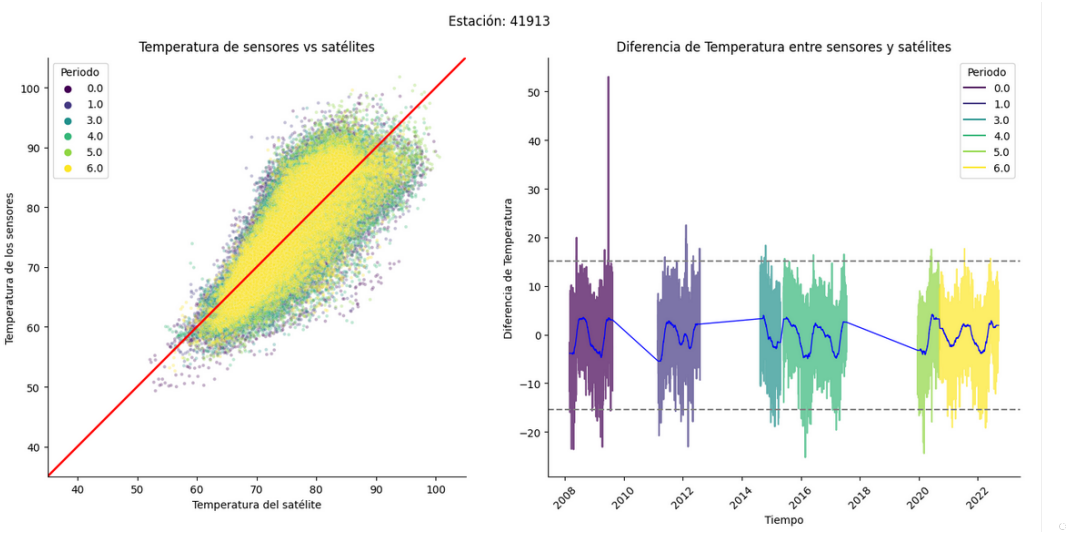
Estación 41913 - sin calibrar



Nota. Elaboración propia.

Figura 2.

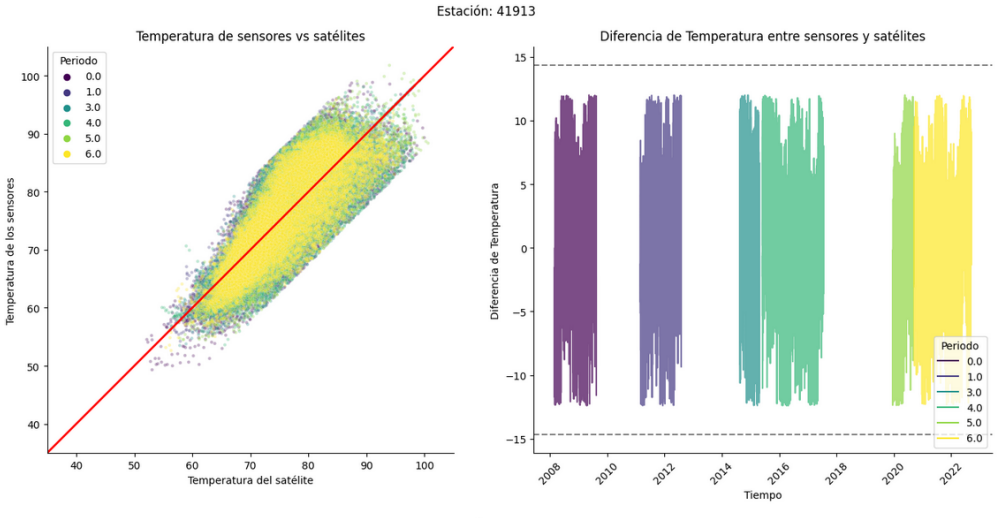
Estación 41913 - calibrado



Nota. Elaboración propia.

Figura 3.

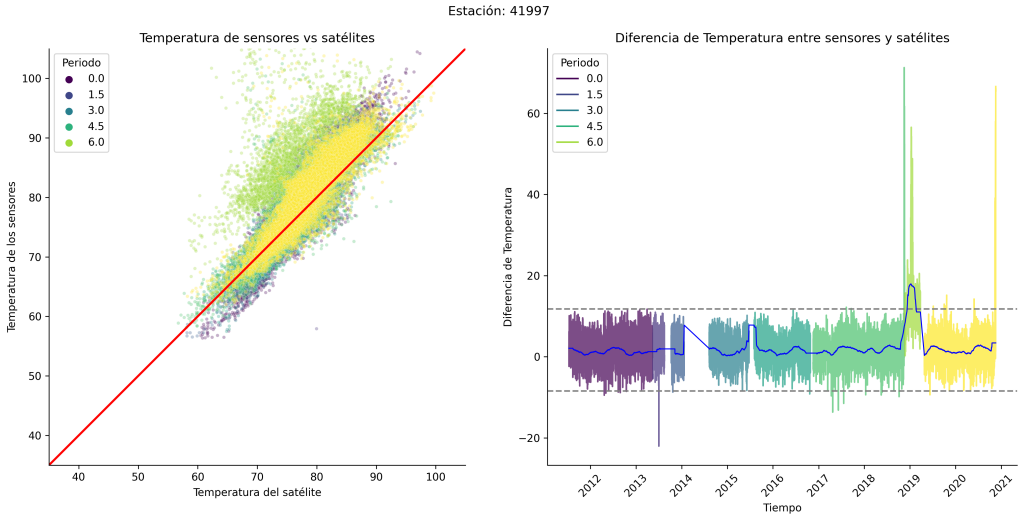
Estación 41913 - calibrado y sin datos atípicos



Nota. Elaboración propia.

Figura 4.

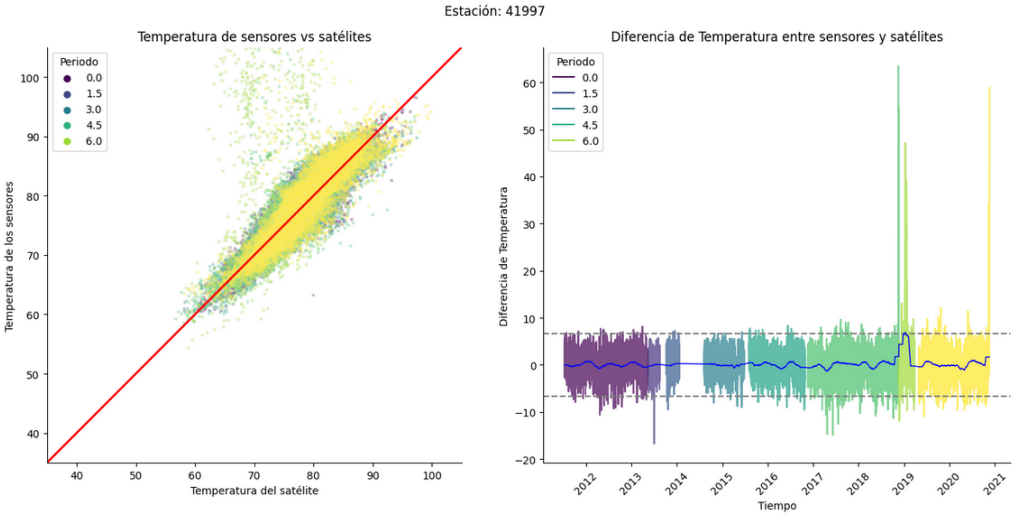
Estación 41997 - sin calibrar



Nota. Elaboración propia.

Figura 5.

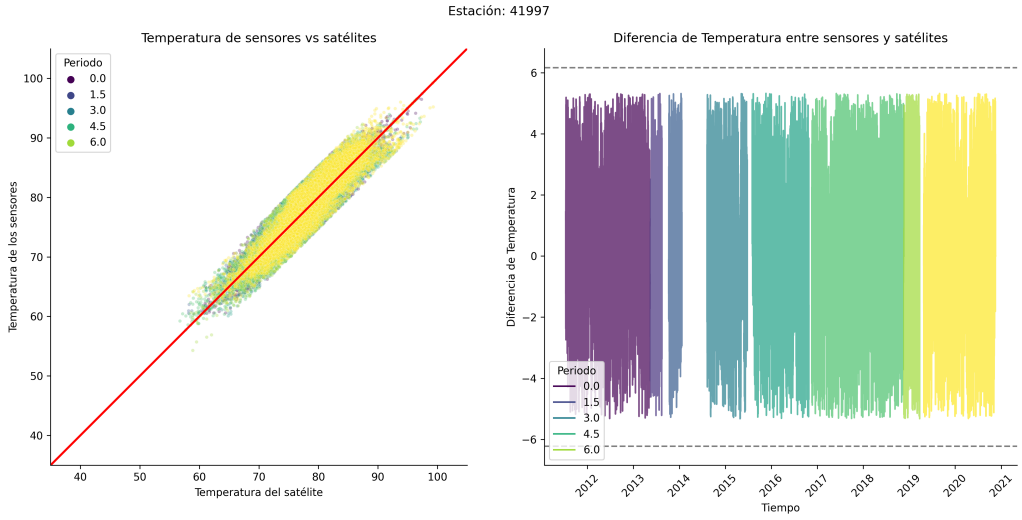
Estación 41997 - calibrado



Nota. Elaboración propia.

Figura 6.

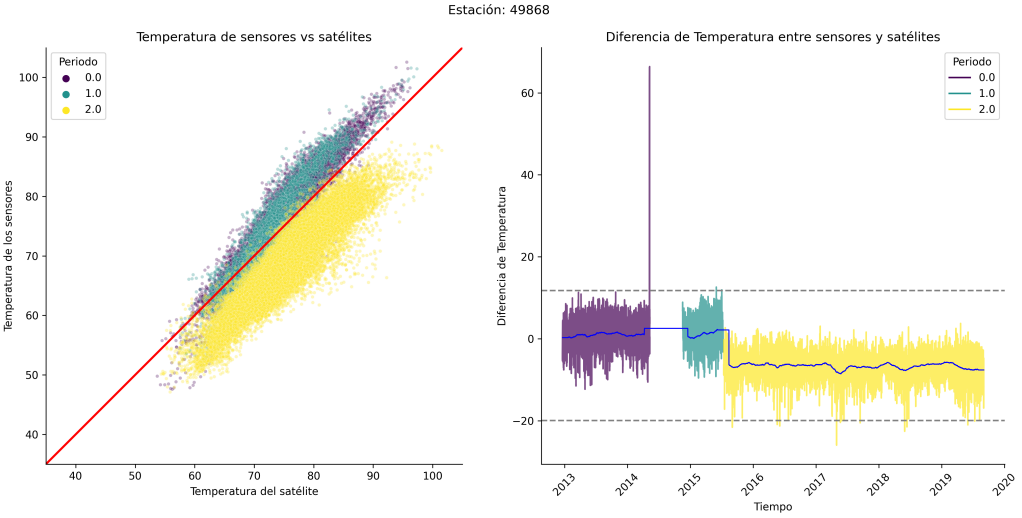
Estación 41997 - calibrado y sin datos atípicos



Nota. Elaboración propia.

Figura 7.

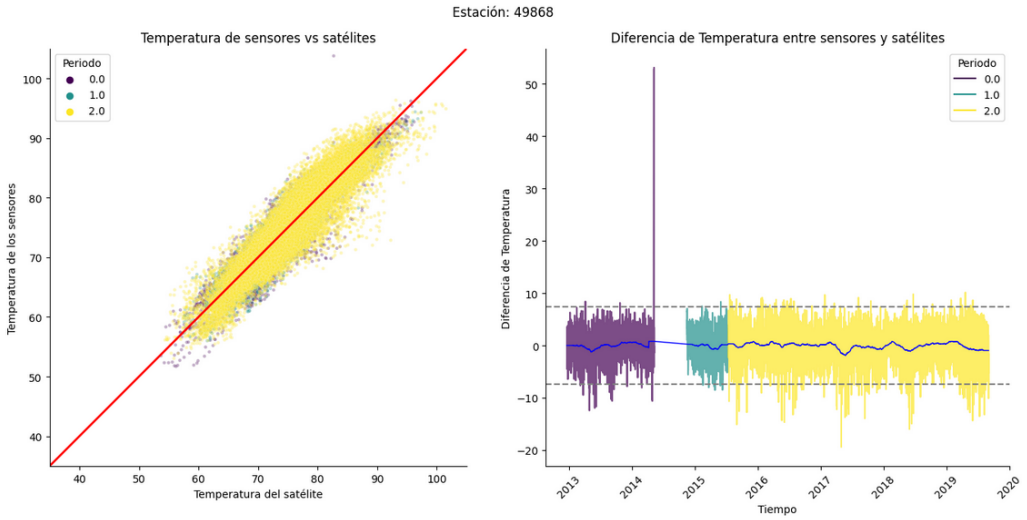
Estación 49868 - sin calibrar



Nota. Elaboración propia.

Figura 8.

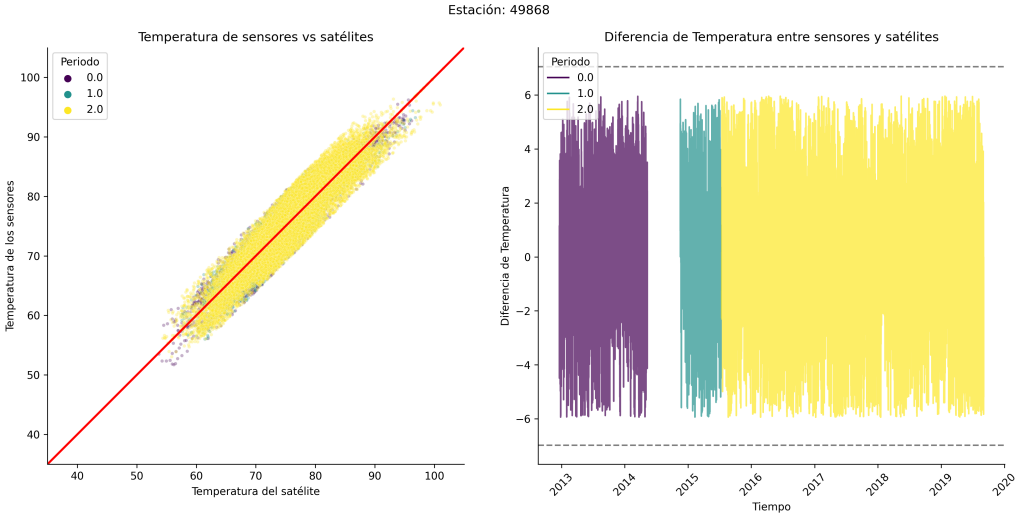
Estación 49868 - calibrado



Nota. Elaboración propia.

Figura 9.

Estación 49868 - calibrado y sin datos atípicos



Nota. Elaboración propia.

7.3. Resultados del modelo

Con los datos ya procesados y limpios, se realizó la validación cruzada con los modelos TFT. En el Cuadro 9, se muestran las 6 diferentes combinaciones de hiperparámetros que se evaluaron, con sus respectivos resultados en cada uno de los splits. El split 0 hace referencia a la validación mas antigua y el split 2 a la validación mas reciente. En total, se entrenaron 18 TFT para obtener los siguientes resultados.

Cuadro 9.

Resultados de la validación cruzada

Combinación	Learning rate	Hidden size	MAE split 0	MAE split 1	MAE split 2
0	0.005	16	0.568683	1.781424	0.751261
1	0.005	32	0.523909	0.516008	0.716703
2	0.010	16	0.573949	0.581538	0.693118
3	0.010	32	1.806597	1.781424	1.891493
4	0.030	16	0.972573	1.171928	1.253337
5	0.030	32	1.179593	1.434953	1.891465

Nota. Elaboración propia.

Se eligió la combinación 1 para entrenar el modelo con todos los datos del *train*. En el Cuadro 10, se muestran los resultados luego de haber entrenado el modelo con los hiperparámetros seleccionados. Los resultados se dividen por dataset con base en cuantos días pasaron desde la predicción, las predicciones realizadas para las primeras 24 horas están categorizadas como 0 días desde a predicción. Las siguientes 24 horas como 1 y las últimas 24 horas pronosticadas como 2.

Cuadro 10.

Resultados del modelo TFT

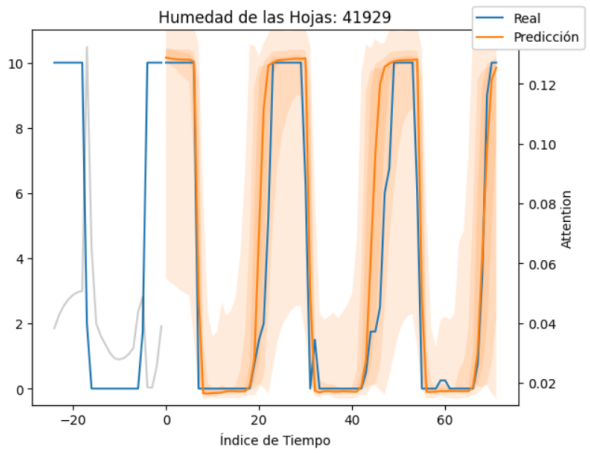
Días desde la predicción	Dataset	R^2	MAE	Recall	Precision	F1-Score
0	train	0.812891	0.510995	0.872523	0.862330	0.867396
0	test	0.593106	1.343110	0.798705	0.842340	0.819942
0	prod	0.851462	0.775055	0.953542	0.952068	0.952805
0	prod_forecast	0.854088	0.769045	0.948509	0.957405	0.952937
1	train	0.773758	0.583963	0.870658	0.840300	0.855209
1	test	0.532604	1.485209	0.800845	0.813448	0.807097
1	prod	0.740779	1.086109	0.936895	0.914934	0.925784
1	prod_forecast	0.742432	1.091430	0.935346	0.915845	0.925493
2	train	0.768415	0.593796	0.864943	0.836154	0.850305
2	test	0.500064	1.561976	0.794749	0.801309	0.798016
2	prod	0.732515	1.097535	0.939853	0.912929	0.926195
2	prod_forecast	0.732685	1.106193	0.938688	0.911454	0.924870

Nota. Elaboración propia.

Para las estaciones de Belice (*test*), se obtuvo un *recall* mayor a 0.79 en todos los días de predicción y un *precision* mayor a 0.80 (las métricas de clasificación indican si la humedad de la hoja es mayor a 3 o no). En las Figuras 10-13, se muestran algunos ejemplos de predicciones para estas estaciones. La línea azul muestra los valores reales de humedad de las hojas mientras que la anaranjada, las predicciones hechas en el índice 0. También, se presentan los resultados reales 24 horas antes de la predicción. En una línea gris, se muestra la proporción de atención que tiene cada momento en el tiempo, es decir, el peso que cada momento pasado de la humedad de la hoja toma en el modelo.

Figura 10.

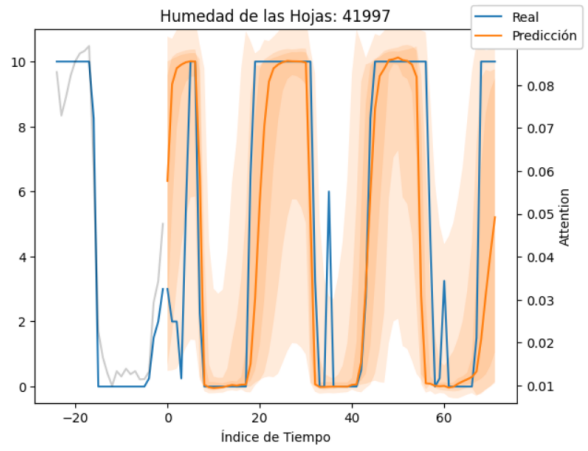
Últimas predicciones de la estación 41929



Nota. Elaboración propia.

Figura 11.

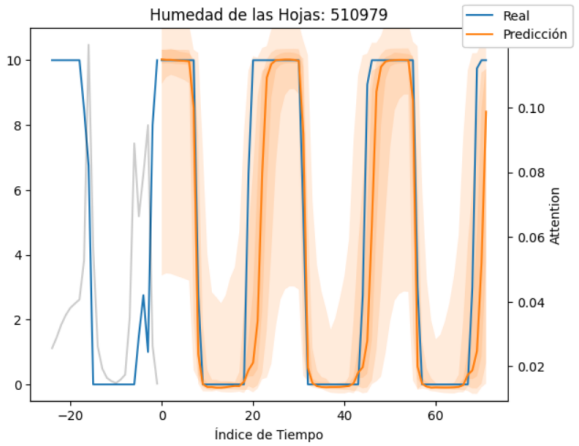
Últimas predicciones de la estación 41997



Nota. Elaboración propia.

Figura 12.

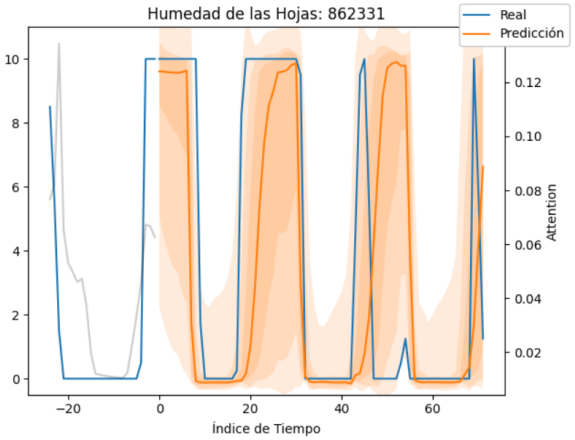
Últimas predicciones de la estación 510979



Nota. Elaboración propia.

Figura 13.

Últimas predicciones de la estación 862331



Nota. Elaboración propia.

Con las predicciones de humedad de las hojas y los datos de temperatura satelital se obtuvieron los resultados de DSV_{LW} presentados en el Cuadro 11. Para las estaciones de Belice se obtuvo un R^2 de 0.68, 0.57 y 0.55 para los días desde predicción 0, 1, y 2 respectivamente.

Cuadro 11.

Resultados del DSV de humedad de hoja

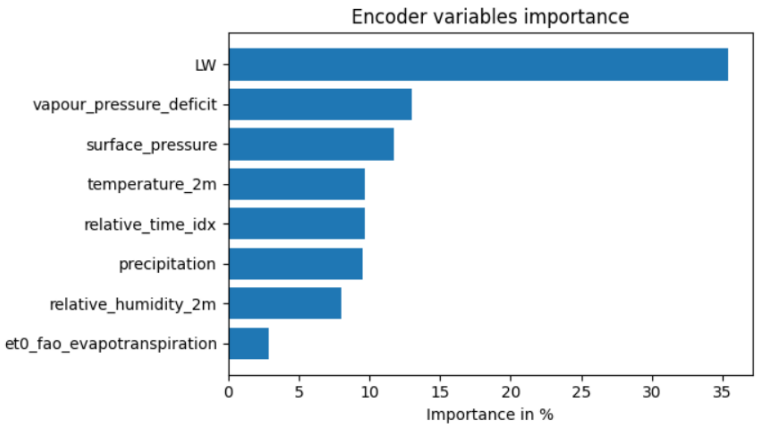
Días desde la predicción	Dataset	R^2	MAE
0	train	0.824480	25.419057
0	test	0.575219	58.925000
0	prod	0.939010	29.170000
0	prod_forecast	0.932639	32.230000
1	train	0.814079	25.962090
1	test	0.502023	64.725000
1	prod	0.914377	36.230000
1	prod_forecast	0.908144	39.590000
2	train	0.821063	25.260625
2	test	0.491159	66.270000
2	prod	0.907031	36.950000
2	prod_forecast	0.893420	41.760000

Nota. Elaboración propia.

Los TFT también permiten ver las importancias de las variables, tanto en el *encoder*, como en el *decoder*. En las Figuras 14 y 15, se presentan las importancias de variables.

Figura 14.

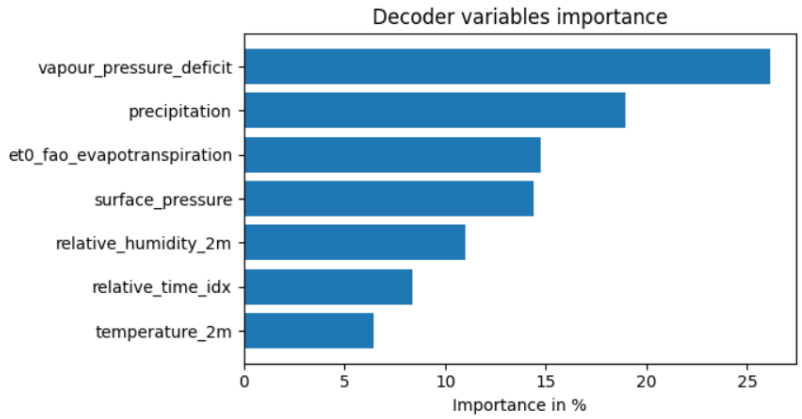
Importancias del encoder del TFT



Nota. Elaboración propia.

Figura 15.

Importancias del decoder del TFT



Nota. Elaboración propia.

7.4. Tablero de visualización

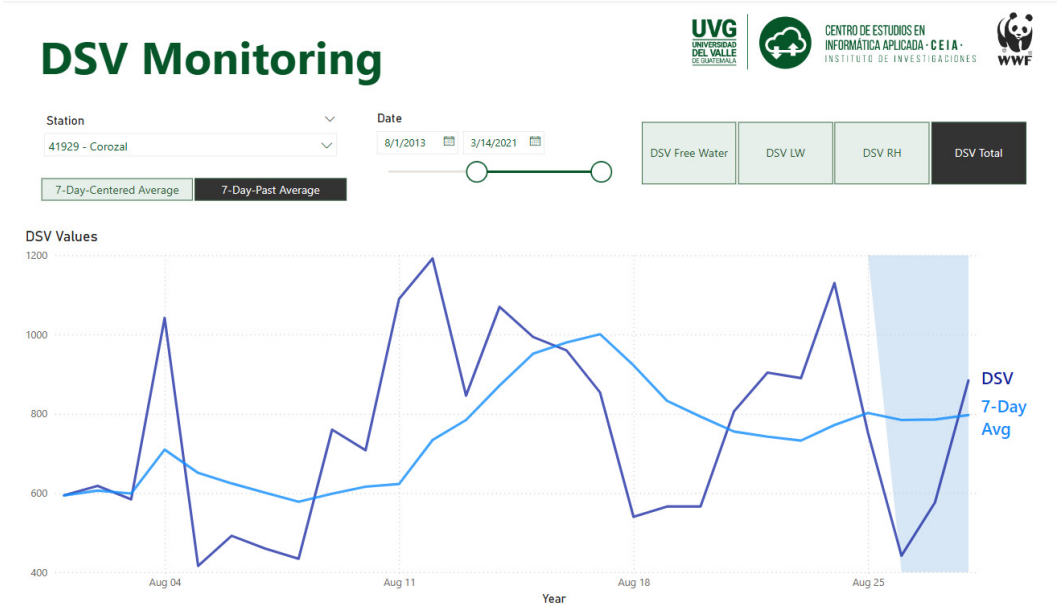
Los resultados de los últimos días de cada región se exportaron a Power BI, y se desarrolló el tablero de visualización presentado en las Figuras 16 y 17.

La Figura 16 presenta la página principal del tablero, la cual tiene el objetivo de monitorear el DSV, tanto de los últimos valores históricos como de las predicciones de los próximos 3 días. En la parte superior se encuentran los filtros donde se puede elegir la estación y el rango de fechas que el usuario desea visualizar. Con botones interactivos se puede elegir si visualizar el $\text{DSV}_{\text{Total}}$ o ver uno en específico. También se incluye la opción de visualizar el promedio de 7 días de DSV hacia atrás o centrado (4 valores hacia atrás más los 3 valores siguientes).

Por último, la Figura 17, muestra la segunda página del tablero, cuyo objetivo es visualizar las variables climatológicas a nivel de hora. Al igual que en la página anterior, esta también permite ver los pronósticos de cada variable. Para la temperatura, humedad relativa y precipitación, se presentan los pronósticos climáticos, mientras que para la humedad de la hoja se muestran los resultados del modelo. Se da la opción de elegir cuál de las cuatro variables se desea visualizar, además de incluir filtros por estación y rango de fechas.

Figura 16.

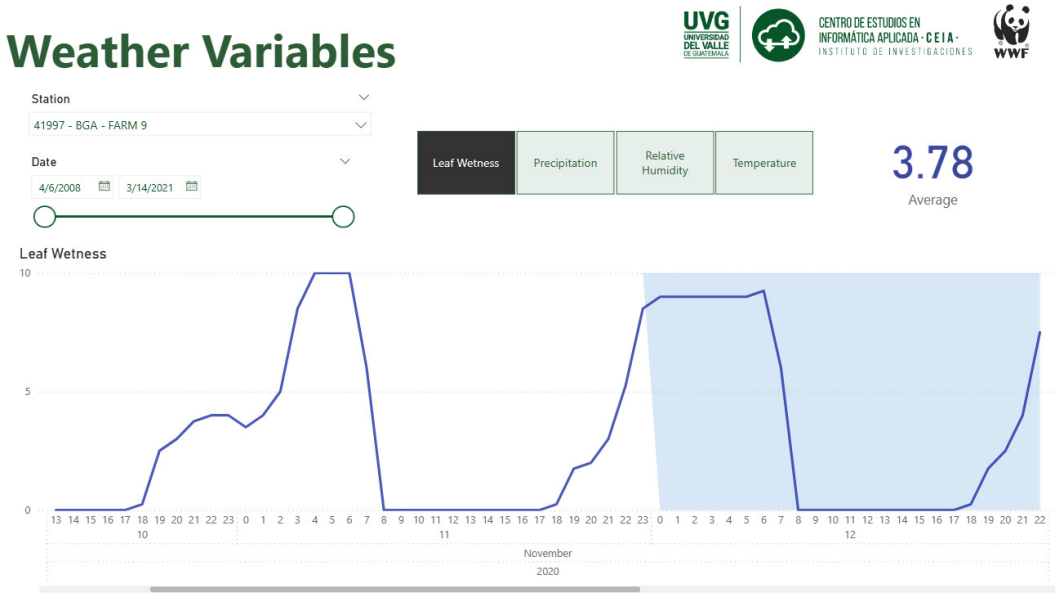
Tablero de visualización - monitoreo de DSV



Nota. Elaboración propia.

Figura 17.

Tablero de visualización - variables del clima



Nota. Elaboración propia.

CAPÍTULO 8

Discusión

Los datos proporcionados por la WWF incluyen regiones de Belice, Guatemala, El Salvador, Honduras y Costa Rica. Cada plantación cuenta con información sobre temperatura, humedad de las hojas, humedad relativa y precipitación. Al separar los datos de cada estación en períodos, se observó que había muchos períodos de inactividad, así como otros en los que la diferencia entre la información satelital y la de los sensores de temperatura presentaba desviaciones atípicas. Se asume que dicho error no debería mostrar una tendencia o un componente determinista, por lo que, en promedio, las diferencias deberían estar alrededor de 0.

En la Figura 1, se observa claramente cómo se desvía la media móvil en el período 2, detectando de manera efectiva los períodos de desviación atípica de duración corta, lo que probablemente se debe a problemas de calibración del sensor o alguna otra falla técnica. Sin embargo, la media móvil no logra detectar períodos largos con desviación de 0, por lo que se realizó una calibración de datos como un paso adicional de limpieza. Se entrenó un modelo de regresión lineal para cada período de cada estación, y si el modelo alcanzaba un R^2 mayor o igual a 0.60, se mantenía esa región. Los periodos que no cumplían con este mínimo se eliminaron, ya que no existía una relación lineal entre la temperatura del sensor y la satelital. En las Figuras 2, 5 y 8, se puede observar cómo el proceso de calibración ajustó los datos para que no se desvíen de las temperaturas de los sensores. Este proceso solo se aplica en el entrenamiento, dado que en producción no siempre se contarán con suficientes datos para realizarlo. Por lo tanto, sería necesario analizar una medida constante de la media móvil para

asegurar que se mantenga alrededor de 0, de modo que se pueda monitorear continuamente para evitar que lo sensores estén descalibrados.

Una vez ajustados los períodos, se realizó el tratamiento de datos atípicos, ya que, como se observa en las Figuras 2, 5 y 8, aún existen datos atípicos que podrían deberse a problemas más inmediatos de los sensores. En lugar de eliminar los datos atípicos, se decidió rellenarlos con la información satelital para evitar interrumpir los períodos y de esta forma no generar más subperiodos, permitiendo conservar los datos. Los demás sensores no mostraron tendencias marcadas de estar descalibrados, por lo que se rellenaron los datos atípicos con la información satelital.

Con los datos ya limpios se entrenaron los *transformadores de fusión temporal*, utilizando la validación cruzada por series temporales para determinar el *learning rate* y el *hidden size* adecuados. Los resultados del Cuadro 9, mostraron que había casos en los que la red no convergía, mientras que en otros casos sí lo hacía, ya que los valores estaban por encima de 1.15 o por debajo de 0.80. Solo hubo dos combinaciones de hiperparámetros que lograron un en el MAE inferior a 0.80 en todos los splits de validación. En este caso, la combinación 1 fue la que en promedio logró los mejores resultados, eligiendo un *learning rate* de 0.005 y un *hidden size* de 32.

Luego de entrenar el modelo final con todos los datos de entrenamiento y los hiperparámetros elegidos, se muestra en el Cuadro 10 que los resultados fueron muy buenos, tanto desde la perspectiva de regresión como cuando se analizan desde una perspectiva categórica (mayor a 3 o no de humedad de las hojas). En los datasets *train*, *test* y *prod*, se alcanzó un F1-score superior a 0.80, con *recall* y *precision* bastante altos también. El conjunto de producción fue el que obtuvo los mejores resultados. Este conjunto, al igual que el de *test*, no se utilizó para el entrenamiento y, aun así, obtuvo muy buenos resultados. Sin embargo, se observó una caída considerable en el *test* en comparación con el entrenamiento, lo que sugiere que un reentrenamiento o ajuste del modelo podría ser adecuado a medida que se incorporen datos de más regiones al modelo. Además, como ocurre con la mayoría de modelos de *deep learning*, los *transformadores de fusión temporal* tienden a aprender mejor con más datos [8]. En cuanto a por qué el conjunto de producción tuvo un mejor desempeño, esto podría deberse a que los datos de producción fueron seleccionados a partir de 2022, esto para poder compararlos con pronósticos climáticos. Esto sugiere que los datos satelitales más recientes pueden ser más precisos que los más antiguos, que los sensores de esas estaciones sean más precisos, o que tengan un comportamiento diferente. Al ser pocas regiones las utilizadas en *test* y *prod*, es posible que la diferencia de desempeño se deba a la calidad de datos de los sensores de humedad de las hojas, ya que a diferencia del resto de variables, esta no puede validarse con información satelital. De cualquier forma, los resultados fueron positivos para los tres conjuntos de datos.

Se observó que el dataset *prod_forecast* mostró un rendimiento casi idéntico al de *prod*, con algunas métricas incluso insignificativamente mejores, lo que es un buen indicio de que, si pone en producción el modelo y se utilizan los pronósticos climáticos, el modelo continuará ofreciendo muy buenos resultados. Incluso, si se usan datos recientes, podría incluso mejorar, como ocurrió en estos casos. Los resultados 10 están segmentados por la cantidad de días transcurridos desde la predicción. Es importante recordar que el modelo predice para las siguientes 72 horas, es decir, los próximos tres días. Esta segmentación en días desde la predicción tiene como objetivo medir si la precisión de las predicciones disminuye con el tiempo. Basándonos en el R^2, se observa una leve caída conforme pasan los días desde la predicción. Asimismo, al comparar los *F1-scores*, se nota una caída ligera, pero los resultados siguen siendo buenos, lo que indica que incluso para futuras iteraciones, podría evaluarse la viabilidad de generar predicciones para un horizonte de predicción más largo, como una semana, quincena o mes, dependiendo del rendimiento del modelo.

Para un análisis más detallado de lo que está sucediendo, en las Figuras 10-13, se muestran ejemplos de pronósticos para las regiones de Belice (*test*). Se puede observar que el comportamiento tiende a ser más suavizado que la realidad, pero que es muy efectivo para detectar las subidas y bajadas de la humedad de las hojas. Se nota que aún hay margen de mejora en el modelo, especialmente en los pequeños picos que superan el umbral de humedad de la hoja, que podrían pasar desapercibidos para el modelo. También se puede observar una sobre-estimación en el tiempo de humedad, como se observa en la Figura 13 entre los índices de tiempo 40 y 60. Es importante recordar que el modelo predice para las siguientes 72 horas, por lo que las líneas anaranjadas de predicción en estas gráficas se generaron con una sola corrida en el índice de tiempo 0, que es el momento en que se realizaron estas predicciones. Las líneas grises indican las partes de la serie temporal en las que el modelo concentra mayor atención a la hora de hacer predicciones, y en estos casos siempre la mayor atención se pone en los valores más altos de humedad de las hojas en las últimas 24 horas.

Una de las motivaciones para el desarrollo de los transformadores es su capacidad de ofrecer interpretabilidad en modelos complejos de pronóstico, así como este mecanismo de atención mencionado, también se puede analizar las importancias de variables. En este caso, como se observa en las Figuras 14 y 15, el TFT muestra las importancias tanto del *encoder* como del *decoder* de esta arquitectura. En el *encoder*, que utiliza únicamente información pasada, se destaca la humedad de las hojas (LW) como la variable más importante, indicando que si existe una correlación muy fuerte entre sus valores pasados y los siguientes. Para el *decoder*, que utiliza únicamente información futura, la variable más importante es el déficit de presión de vapor (*vapor_pressure_deficit*), lo que tiene sentido ya que, según Open-Meteo.com, este índice indica que cuando es alto, la transpiración de las plantas aumenta, y cuando es bajo, la transpiración disminuye. Esto es coherente con el objetivo del modelo, que es predecir la humedad. Las variables de evapotranspiración y precipitación también son

importantes en el *decoder*, lo que refuerza la idea de que el modelo puede combinar información relevante del pasado con información futura. Todas las variables del *decoder* pueden ser extraídas exactamente de la misma manera a través de Open-Meteo.com en forma de pronósticos, y como se observa en la comparación de los resultados de *prod* y *prod_forecast*, estos resultados parecen indicar que este modelo seguiría funcionando adecuadamente incluso si se utilizan pronósticos climáticos.

Una vez obtenidos los resultados del modelo para predecir la humedad de las hojas, es posible comparar los DSVs de humedad de las hojas calculados de las predicciones con los calculados a partir de los datos de sensores. En el Cuadro 11, se observa que en los datasets *train*, *prod* y *prod_forecast*, se obtienen resultados con un R^2 bastante alto, y aunque el *test* tiene resultados más bajos, la correlación sigue siendo bastante alta. Esto demuestra que el DSV calculado a partir de las predicciones es en un estimador útil. Como se mencionó anteriormente, esto puede indicar que un re-entrenamiento con nuevas regiones podría ser beneficioso, ya que el modelo podría llegar a comprender mejor las relaciones entre diversas regiones.

Como se observa en las Figuras 10 y 12, el tablero de visualización permite identificar de manera clara y dinámica los factores climáticos que están provocando un riesgo alto o bajo de desarrollo de la sigatoka negra, ofreciendo una visión detallada de cómo las condiciones actuales y las predicciones meteorológicas afectan el pronóstico. Gracias a la transparencia en los resultados, los usuarios pueden comprender fácilmente las variables que influyen en el riesgo, aumentando la confianza en las decisiones tomadas. Este sistema puede ser utilizado en la página principal para monitorear de manera diaria el riesgo de desarrollo de la enfermedad, esencial para los productores que necesitan tomar decisiones rápidas y fundamentadas. Además, el tablero proporciona pronósticos para los siguientes tres días, lo que permite anticipar posibles riesgos y actuar preventivamente, eligiendo el fungicida adecuado o tomando otras medidas. Con esta información no solo se puede gestionar el riesgo en tiempo real, sino que también se puede planificar de manera proactiva, ajustando las acciones a las condiciones futuras y optimizando los recursos disponibles para la prevención de la sigatoka negra.

CAPÍTULO 9

Conclusiones

- La implementación del modelo de *machine learning* para predecir el DSV, en combinación con el tablero interactivo permiten una mejora en la toma de decisiones para el control de la sigatoka negra en Belice. Al integrar datos climatológicos específicos de las plantaciones de banano, tanto de sensores como de datos satelitales, el modelo ofrece predicciones precisas que permiten a los productores anticiparse a posibles incrementos en el riesgo de desarrollo de la enfermedad.
- Las fuentes meteorológicas utilizadas, tanto de sensores como de satélites, ofrecen una base de datos funcional para el entrenamiento de modelos predictivos, combinando datos de un punto en específico con datos satelitales disponibles en internet. Ambas fuentes demuestran ser altamente efectivas para los resultados del modelo.
- La creación del dataset de entrenamiento, mediante una unidad de análisis bien definida y un exhaustivo proceso de limpieza de datos, ha permitido obtener un conjunto de datos de alta calidad, apto para el entrenamiento del modelo de *machine learning*. Si bien las fuentes cumplen con la información requerida, la limpieza de datos fue necesaria para garantizar que estuvieran bien calibrados y que el modelo no entrenara con datos atípicos.
- La utilización del *time split cross validation* como técnica de evaluación permitió encontrar la combinación de hiperparámetros con un *lerning rate* de 0.05 y un *hidden size* de 32, los cuales logran que el modelo se desempeñe bien a lo largo de diferentes periodos de tiempo, dando como resultado una combinación óptima para el entrenamiento del modelo final de transformador de fusión temporal.

- El modelo final de *machine learning*, con la arquitectura de transformador de fusión temporal para predecir la humedad de las hojas, obtuvo un *recall* superior a 0.79 y un *precision* superior a 0.80 para las estaciones de Belice en todos los días de predicción. Estos resultados, junto a los pronósticos del clima disponibles por medio de un API del clima, recolectan toda la información necesaria para el cálculo de DSV.
- La creación de un tablero interactivo de visualización permite a los usuarios visualizar tanto los pronósticos futuros del DSV como los valores históricos, proporcionando un análisis claro y accesible del riesgo de la sigatoka negra. Esta herramienta intuitiva facilita la interpretación de los datos generados por el modelo, fortaleciendo la capacidad de los productores para tomar decisiones informadas y oportunas en la gestión de la enfermedad.

CAPÍTULO 10

Recomendaciones

- Monitorear que los sensores estén calibrados mediante un promedio móvil de error comparado con los datos satelitales para detectar posibles desviaciones. Este control de calibración aseguraría que los datos de los sensores se mantengan dentro de los rangos esperados.
- Evaluar el uso de múltiples sensores y aplicar validación cruzada para identificar datos atípicos, especialmente en la medición de humedad de las hojas, dado que esta variable no se puede validar con datos satelitales. Además, aplicar este método a otras variables puede mejorar la detección de datos atípicos.
- Evaluar el modelo con datos de regiones más recientes para medir la efectividad de los pronósticos actuales. Es importante considerar que entre mejor sea la resolución y precisión de los pronósticos del clima se pueden obtener mejores resultados.
- En caso de disponer de más datos, evaluar el reentrenamiento del modelo para mejorar su capacidad de generalización conforme se vayan agregando mas estaciones de clima a los datos.
- Parar futuras iteraciones del modelo, seguir optimizando los hiperparámetros del modelo y evaluar si se puede ajustar aún más para detectar picos de humedad de corta duración en las hojas, lo cual es una limitante del modelo presentado.
- Almacenar los datos de los sensores en una herramienta que permita su descarga automática mediante el uso de una nube pública para generar predicciones en tiempo real.

- Realizar pruebas de campo para definir umbrales del DSV con el fin de categorizar el riesgo en categorías (por ejemplo: muy bajo, bajo, medio, alto y muy alto). Esta clasificación proporcionaría una forma mas sencilla de interpretar los resultados.

CAPÍTULO 11

Referencias

[1] R. Bennett y P. A. Arneson, "Black sigatoka of bananas and plantains," *The Plant Health Instructor*, vol. 3, 2003. [En línea]. Disponible: https://doi.org/10.1094/PHI-I-2003-0905-01.

[2] M. Guzmán, "Estado actual y perspectivas futuras del manejo de la Sigatoka negra en América Latina," en *XVII Reunión ACORBAT*, vol. 1, Joinville, Santa Catarina, BR, 2006, págs. 83-91.

[3] A. Riveros, "Estudio del potencial antifúngico y de inducción de resistencia de extractos de origen vegetal para el control de la Sigatoka negra en plátano," *Memorias, Primer Encuentro de Investigadores en Agricultura Orgánica, Programa de Investigación y Transferencia de Tecnología Agrícola (PITTA)*, págs. 14-16, 2001.

[4] E. Álvarez, A. Pantoja, L. Guarín y G. Ceballos, *Estado del arte y opciones de manejo del moko y la Sigatoka negra en América Latina y el Caribe*. Cali, Colombia: Centro Internacional de Agricultura Tropical (CIAT), 2006, Publicación CIAT No. 346. [En línea]. Disponible: https://cgspace.cgiar.org/server/api/core/bitstreams/d7aa77a9-dbdd-489b-aa96-ded3f9fbf454/content.

[5] D. H. Marín, R. A. Romero, M. Guzmán y T. B. Sutton, "Black Sigatoka: An increasing threat to banana cultivation," *Plant Disease*, vol. 87, n.º 3, págs. 208-222, 2003.

[6] World Wildlife Fund, *Weather based disease forecast system for Black Sigatoka (Mycosphaerella fijensis) control in banana farms*, s.f.

[7] X. Mourichon, J. Carlier y E. Fouré, *Enfermedades de Musa: Hoja Divulgativa N° 8*, 1997. [En línea]. Disponible: https://agritrop.cirad.fr/314356/7/314356_ES.pdf.

[8] B. Lim, S. O. Arik, N. Loeff y T. Pfister, "Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting," 2019. [En línea]. Disponible: https://doi.org/10.48550/arXiv.1912.09363.

[9] L. F. Benavides, M. Camacho-Calvo y M. Muñoz Fonseca, "Relación entre factores climáticos y la infección foliar de Sigatoka negra (Pseudocercospora fijiensis) en plantas de banano (Musa AAA) con y sin la aplicación de fungicida," *Revista AgroInnovación en el Trópico Húmedo*, vol. 3, n.º 1, págs. 1-13, 2020. [En línea]. Disponible: https://doi.org/10.18860/rath.v3i1.650.

[10] K. Liakos, P. Busato, D. Moshou, S. Pearson y D. Bochtis, "Machine Learning in Agriculture: A Review," *Sensors*, vol. 18, n.º 8, pág. 2674, 2018. [En línea]. Disponible: https://doi.org/10.3390/s18082674.

[11] R. H. Stover, "Sigatoka leaf spot of banana and plantains," *Plant Disease*, vol. 64, págs. 750-756, 1980.

[12] L. Pérez, "Morfología de las especies de Mycosphaerella asociadas a manchas de las hojas en Musa spp.," *Fitosanidad*, vol. 6, n.º 2, págs. 3-9, 2002.

[13] I. Martínez, R. Villalta, E. Soto, G. Murillo y M. Guzmán, "Manejo de la Sigatoka negra en el cultivo del banano," *OJA Divulgativa*, n.º No. 2-2011, 2011. [En línea]. Disponible: https://www.corbana.co.cr/wp-content/uploads/HD-n.%C2%B0-2-2011-Manejo-de-la-Sigatoka-negra.pdf.

[14] J. Wallin, "Summary of recent progress in predicting late blight epidemics in United States and Canada," *American Potato Journal*, vol. 39, págs. 306-312, 1962. DOI: 10.1007/BF02862155.

[15] K. P. Murphy, *Probabilistic machine learning: An introduction*. MIT Press, 2022. [En línea]. Disponible: https://lccn.loc.gov/2021027430.

[16] IBM, *Machine learning supervisado*, 2024. [En línea]. Disponible: https://www.ibm.com/es-es/topics/machine-learning.

[17] D. Chicco, M. J. Warrens y G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Computer Science*, vol. 7, e623, 2021. [En línea]. Disponible: https://doi.org/10.7717/peerj-cs.623.

[18] G. James, D. Witten, T. Hastie y R. Tibshirani, *An Introduction to Statistical Learning with Applications in Python*. Springer, 2023.

[19] J. D. Kelleher, B. M. Namee y A. D'Arcy, *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*. Cambridge, Massachusetts; London, England: The MIT Press, 2015.

[20] T. Hastie, R. Tibshirani y J. Friedman, *The elements of statistical learning: Data mining, inference, and prediction*, 2nd. Springer Series in Statistics, 2017.

[21] R. P. Sheridan, "Time-split cross-validation as a method for estimating the goodness of prospective prediction," *Journal of Chemical Information and Modeling*, vol. 53, págs. 783-790, 2013. [En línea]. Disponible: https://doi.org/10.1021/ci400084k.

[22] I. Goodfellow, Y. Bengio y A. Courville, *Deep Learning*. MIT Press, 2016. [En línea]. Disponible: http://www.deeplearningbook.org.

[23] S. Hochreiter y J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, n.º 8, págs. 1735-1780, 1997.

[24] A. Vaswani et al., "Attention Is All You Need," *Proceedings of NeurIPS*, 2017.

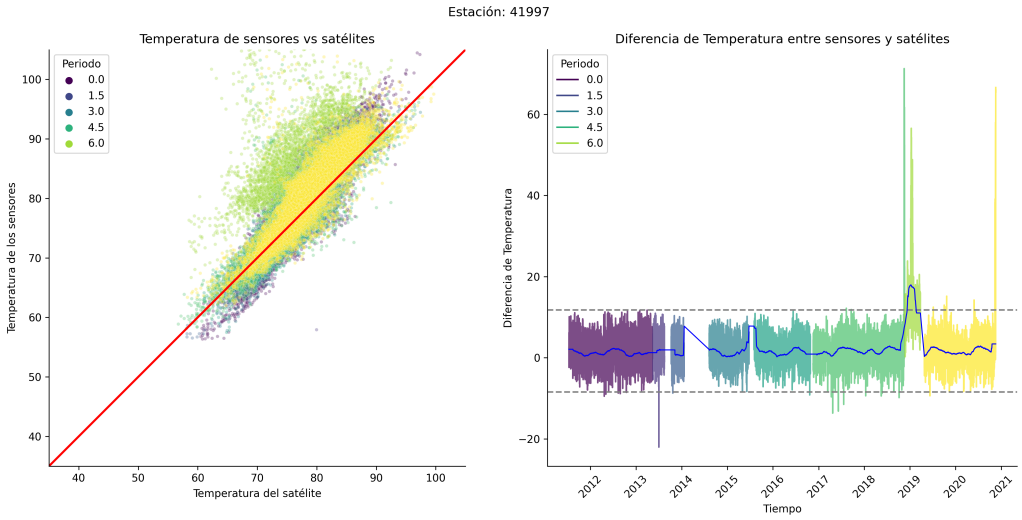
CAPÍTULO 12

Anexos

Limpieza de datos para las estaciones de Belice

Figura 18.

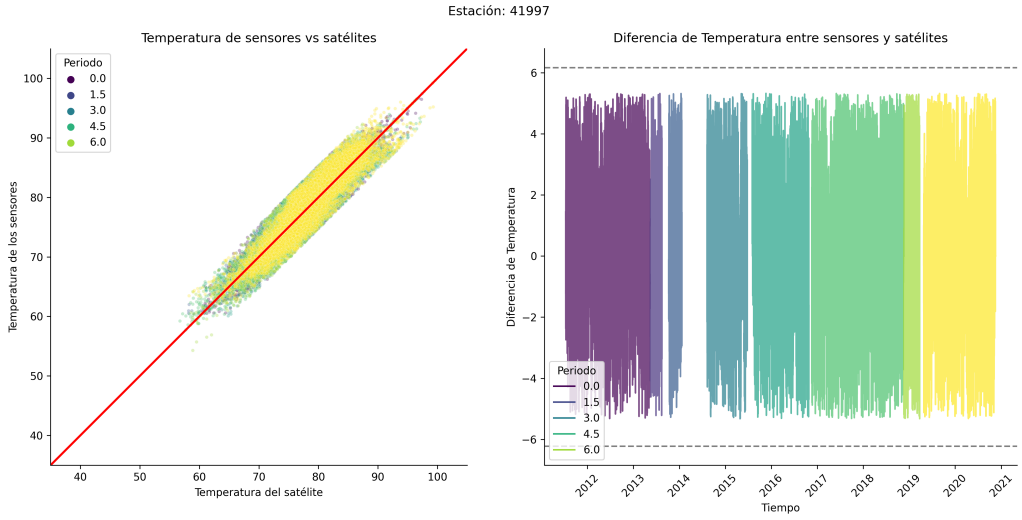
Estación 41997 - sin calibrar



Nota. Elaboración propia.

Figura 19.

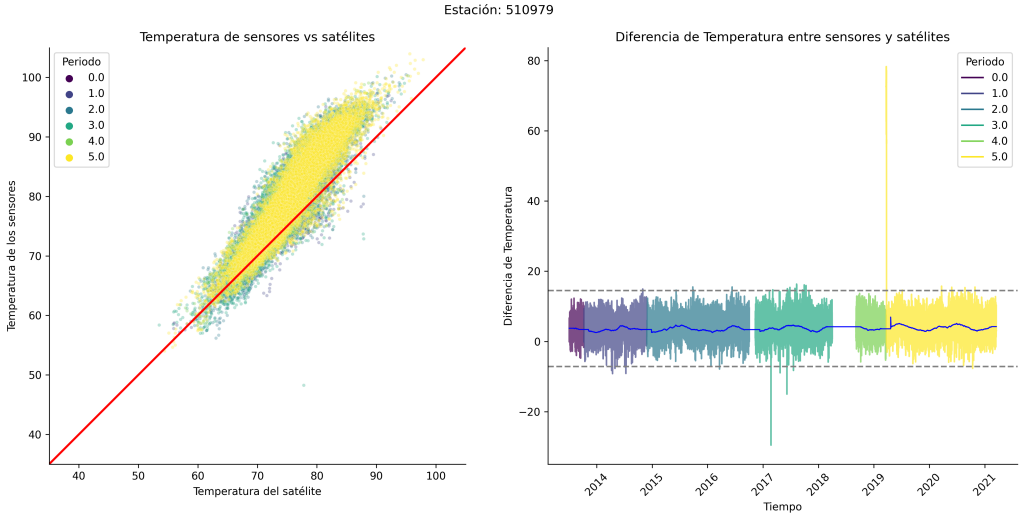
Estación 41997 - calibrado y sin datos atípicos



Nota. Elaboración propia.

Figura 20.

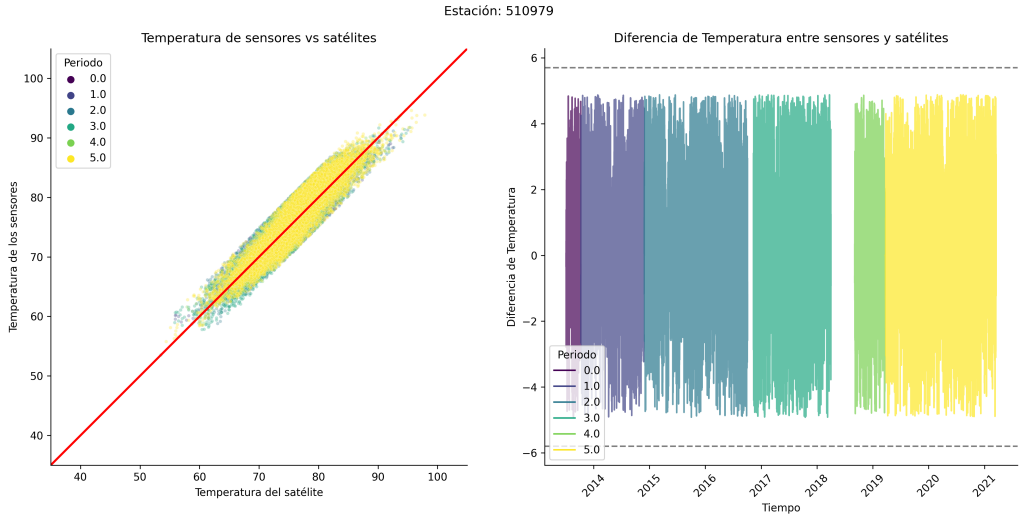
Estación 510979 - sin calibrar



Nota. Elaboración propia.

Figura 21.

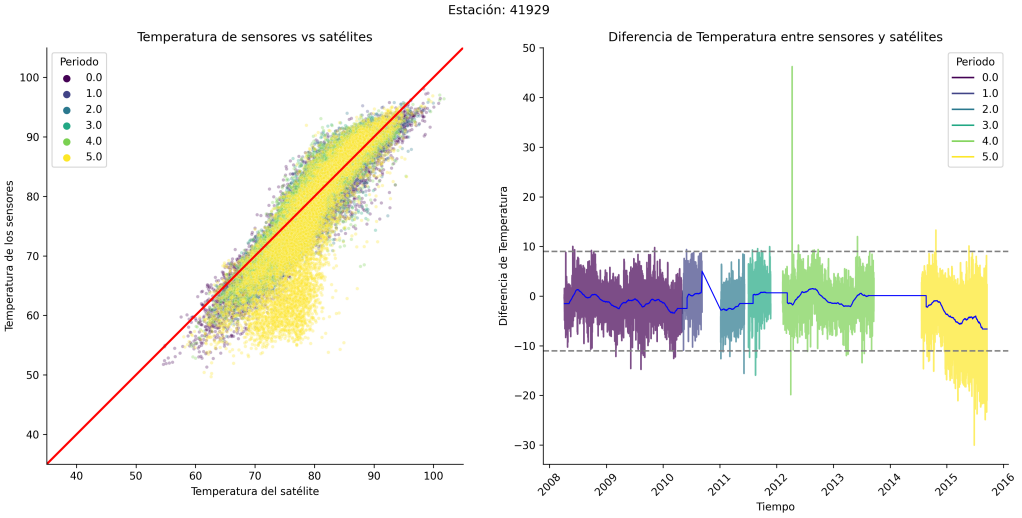
Estación 510979 - calibrado y sin datos atípicos



Nota. Elaboración propia.

Figura 22.

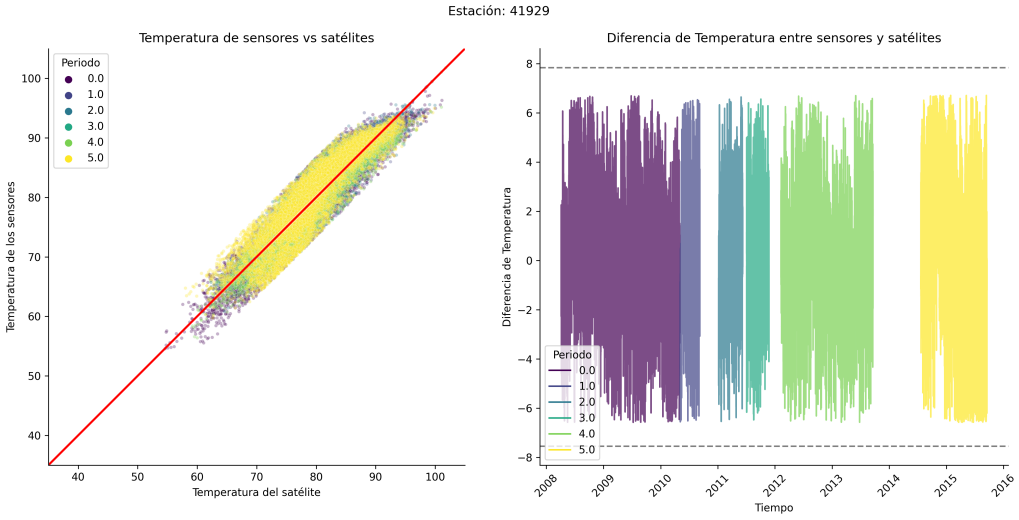
Estación 41929 - sin calibrar



Nota. Elaboración propia.

Figura 23.

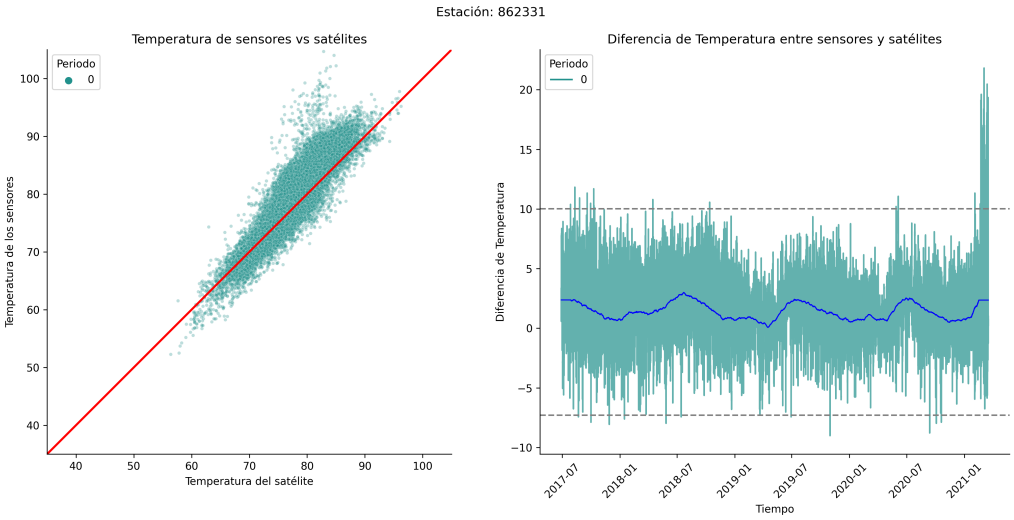
Estación 41929 - calibrado y sin datos atípicos



Nota. Elaboración propia.

Figura 24.

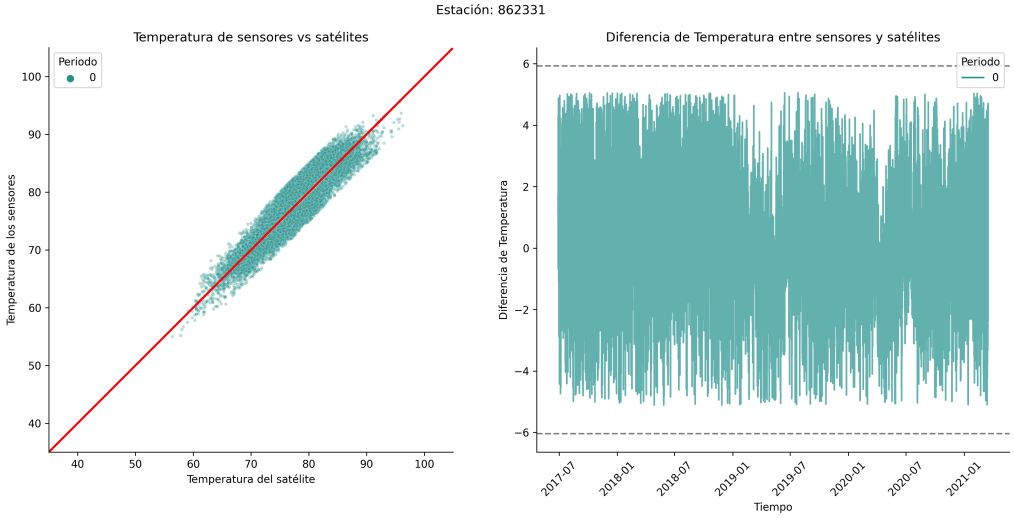
Estación 862331 - sin calibrar



Nota. Elaboración propia.

Figura 25.

Estación 862331 - calibrado y sin datos atípicos



Nota. Elaboración propia.