Detección temprana de estrés biótico y abiótico usando modelos de clasificación de datos de espectroscopía de reflectancia VIS/NIR: Aplicación en plantas de Banano Gros Michel

Cristian Camilo Díaz Herrera

Universidad Nacional de Colombia
Facultad de Ciencias, Departamento de Estadística
Bogotá D,C, Colombia
2024

Detección temprana de estrés biótico y abiótico usando modelos de clasificación de datos de espectroscopía de reflectancia VIS/NIR: Aplicación en plantas de Banano Gros Michel

Cristian Camilo Díaz Herrera

Trabajo final de grado presentado como requisito parcial para optar al título de:
Maestria en Estadística

Directora:
Ph.D. Verónica Botero Fernández

Universidad Nacional de Colombia
Facultad de Ciencias, Departamento de Estadística
Bogotá, Colombia
2024

Lema

El hombre es el más inteligente de los animales y también el más tonto.

Diógenes de Sinope

Agradecimientos

Primero, deseo expresar mi profundo agradecimiento a mis padres, mi esposa y mis hijos por su incondicional apoyo. En segundo lugar, extiendo el reconocimiento a mi directora Verónica Botero Fernández, al profesor Juan Carlos Marín Ortiz y a la profesora Lilliana María Hoyos Carvajal, así como al Centro de Investigaciones para el Banano (Cenibanano). Además, agradezco a la naturaleza por ofrecernos su diversidad y belleza, un recordatorio constante de la importancia de preservar nuestro entorno.
Finalmente, quiero expresar mi sincero agradecimiento a la Universidad Nacional de Colombia por brindarme la oportunidad de acceder a conocimientos tan valiosos. Su acogida y apoyo han sido fundamentales en mi camino académico y personal.

Resumen Detección temprana de estrés biótico y abiótico usando modelos de clasificación de datos de espectroscopía de reflectancia VIS/NIR: Aplicación en plantas de Banano Gros Michel

La detección temprana de enfermedades y estrés hídrico (EH) en las plantas es crucial para la agricultura y la soberanía alimentaria de los países latinoamericanos. En este contexto, se han utilizado métodos de espectroscopía de reflectancia electromagnética visible (VIS) e infrarroja (NIR), que son no invasivos y han demostrado ser prometedores para identificar el estrés biótico y abiótico en las plantas incluso en su fase asintomática. Un ejemplo relevante de esto es la infección de las plantas de banano por enfermedades devastadoras, como la marchitez vascular causada por el hongo *Fusarium oxysporum f.sp. cubense Raza 1* (FOCR1) y por la bacteria *Ralstonia solanacearum Raza 2* (RSR2), que pueden resultar en pérdidas de hasta el 100 % en las plantaciones.

Para abordar este desafío, se llevó a cabo un estudio en el que se analizaron datos de reflectancia de 240 plantas de banano en el municipio de Carepa, ubicado en el departamento de Antioquia, Colombia. Estos datos incluyeron plantas sanas, aquellas sometidas a EH, infectadas con FOCR1, contagiadas con RSR2 y sus interacciones. El análisis se realizó utilizando un espectrómetro portátil ASD FieldSpec. Inicialmente, se aplicaron diversas técnicas de preprocesamiento a los datos de reflectancia en el rango espectral de 350-2500 nm, estas incluyen 2 de tratamiento de datos atípicos y 5 de suavizamiento. Luego, se llevaron a cabo diferentes enfoques para la selección de características, identificando las longitudes de onda que mejor discriminaban entre los diferentes tratamientos mediante la metodología RELIEF.

Se emplearon métodos de clasificación supervisada, como Análisis Lineal Discriminante (ALD), Análisis Cuadrático Discriminante (ACD), Bosques Aleatorios (BA), Bayes ingenuo (BI), Máquinas de Soporte Vectorial (MSV), K vecinos más cercanos (KVC) y Perceptrón Multicapa (PM) con el objetivo de optimizar la exactitud de clasificación de los tratamientos, para esta medición se tuvo en cuenta una división de las plantas en el 75 % de entrenamiento y 25 % de prueba. Los resultados mostraron que, a pesar de que el período asintomático de las plantas de banano es de 20 días, con el ALD se logró un porcentaje de clasificación correcto del 86 % en el día 3 con métodos de preprocesamiento de Mínimos Cuadrados Asimétricos (MCAS) y la gestión de datos atípicos mediante el Método de la Bolsa (MB). Sin tener en cuenta las interacciones, la mejor metodología se obtiene al emplear un ALD con una precisión similar. Al día 6 post-inoculación, se obtienen precisiones similares con el ALD, siendo el más óptimo al usar los métodos de preprocesamiento como el de Corrección de Dispersión Multiplicativa (CDM) al tratar los datos atípicos con el MB. Estos resultados sugieren que

la detección temprana de FOCR1, RSR2 y el EH en plantas de banano, mediante el uso de la espectroscopía de reflectancia, puede mejorar significativamente con la elección adecuada de metodologías de preprocesamiento, selección de características y clasificación de datos.

Palabras clave: detección temprana, VIS/NIR, clasificación, estrés hídrico, Fusarium oxysporum, Ralstonia solanacearum, banano.

Abstract

Early detection of biotic and abiotic stress using VIS/NIR reflectance spectroscopy data classification models: Application in Gros Michel Banana plants

The early detection of diseases and water stress (WS) in plants is crucial for the agriculture and food sovereignty of Latin American countries. In this context, non-invasive methods such as visible-near infrared (VIS) and infrared (NIR) reflectance spectroscopy have been employed and proven promising for identifying biotic and abiotic stress in plants, even in asymptomatic phases. A relevant example is the infection of banana plants by devastating diseases like vascular wilt caused by the fungus *Fusarium oxysporum f.sp. cubense Race 1* (FOCR1) and bacterial wilt caused by *Ralstonia solanacearum Race 2* (RSR2), which can lead to losses of up to 100 % in plantations.

To address this challenge, a study was conducted analyzing reflectance data from 240 banana plants in municipality of Carepa, located in the department from Antioquia, Colombia. The data included healthy plants, those subjected to WS, plants infected with FOCR1, contaged plants with RSR2, and their interactions. The analysis was performed using a portable ASD FieldSpec spectrometer. Initially, various preprocessing techniques were applied to the reflectance data in the spectral range of 350-2500 nm, including two for outlier treatment and five for smoothing. Different approaches for feature selection were then employed, identifying the wavelengths that best discriminated between the different treatments using the RELIEF methodology.

Supervised classification methods such as Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Random Forests (RF), Naïve Bayes (NB), Support Vector Machines (SVM), k-Nearest Neighbors (KNN) and Multilayer Perceptron (MLP) were employed to optimize the classification accuracy of treatments. For this, a division of plants into 75 % training and 25 % testing was considered. Results showed that despite the asymptomatic period of 20 days for banana plants, LDA achieved a correct classification rate of 86 % on day 3 with asymmetric least squares (ALS) preprocessing and outlier management using the Bag method. Excluding interactions, the best methodology was obtained using LDA with similar accuracy. On day 6 post-inoculation, similar accuracies were achieved with LDA being optimal when using preprocessing methodologies such as multiplicative scatter correction (MSC) and outlier treatment with the Bag method. These results suggest that early detection of FOCR1, RSR2, and WS in banana plants through reflectance

spectroscopy can significantly improve with the appropriate choice of preprocessing, feature selection, and data classification methodologies.

Keywords: early detection, VIS/NIR, classification, water stress, Fusarium oxysporum, Ralstonia solanacearum, banana.

Lista de Figuras

6-1 Promedios de los espectros de reflactancia según longitud de onda por tratamiento para el día 3. 46
6-2 Promedios de los espectros de reflactancia según longitud de onda por tratamiento para el día 6. 47
6-3 Detección de atípicos por tratamiento con el MB para el día 3. 49
6-4 Detección de atípicos por tratamiento con el método RAD para el día 3. . . 50
6-5 Detección de atípicos por tratamiento con el MB para el día 6. 51
6-6 Detección de atípicos por tratamiento con el método RAD para el día 6. . . 52
6-7 Diagramas de regiones discriminantes según longitudes de onda seleccionadas para el ALD parsimonioso para el día 3. 65
6-8 Diagramas de regiones discriminantes según longitudes de onda seleccionadas para el ALD parsimonioso para el día 6 73

Lista de Tablas

3-1 Relación de investigaciones en estrés vegetal basados en espectroscopia VIS/NIR que tienen como objetivo clasificar presencia o ausencia del estrés (parte 1) . 16
3-2 Relación de investigaciones en estrés vegetal basados en espectroscopia VIS/NIR que tienen como objetivo clasificar presencia o ausencia del estrés (parte 2) . 17

6-1 Tabla de exactitud sin división en conjunto de entrenamiento y prueba para el día 3 58
6-2 Tabla de exactitud con división en conjunto de entrenamiento y prueba para el día 3 59
6-3 Probabilidades a priori para el ALD en el conjunto de datos MCAS_MB para el día 3 60
6-4 Coeficientes de las ecuaciones lineales discriminantes para el conjunto de datos MCAS_MB para el día 3 61
6-5 Matriz de confusión para el ALD con el conjunto de datos MCAS_MB para el día 3 62
6-6 Probabilidades a priori para el ALD parsimonioso en el conjunto de datos MCAS_MB para el día 3 63
6-7 Coeficientes de las ecuaciones lineales discriminantes para el conjunto de datos MCAS_MB para el modelo parsimonioso en el día 3 63
6-8 Matriz de confusión para el ALD parsimonioso con el conjunto de datos MCAS_MB para el día 3 64
6-9 Tabla de exactitud sin división en conjunto de entrenamiento y prueba para el día 6 66
6-10 Tabla de exactitud con división en conjunto de entrenamiento y prueba para el día 6 68
6-11 Probabilidades a priori para el ALD en el conjunto de datos MCS_MB para el día 6 69
6-12 Coeficientes de las ecuaciones lineales discriminantes para el conjunto de datos MCS_MB para el día 6 69
6-13 Matriz de confusión para el ALD en el conjunto de datos MCS_MB para el día 6 70
6-14 Probabilidades a priori para el ALD parsimonioso en el conjunto de datos MSC_MB para el día 6 71

6-15 Coeficientes de las ecuaciones lineales discriminantes del modelo ALD parsimonioso para el conjunto de datos MCS_MB para el día 6 71
6-16 Matriz de confusión del para el ALD en el conjunto de datos MCS_MB del modelo ALD parsimonioso para el día 6 72

Contenido

Agradecimientos vii

Resumen ix

Lista de figuras xiii

Lista de tablas xv

1 Introducción 1

2 Planteamiento del problema 4

3 Antecedentes investigativos 7
3.1 Detección enfermedad en plantas . . . 7
3.2 Estrés biótico y abiótico en plantas de banano Gros Michel . . . 8
3.2.1 *Fusarium oxysporum* . . . 8
3.2.2 *Ralstonia solanacearum* . . . 9
3.2.3 Estrés hídrico . . . 10
3.3 Espectroscopía usada en enfermedad de plantas . . . 11
3.3.1 Técnicas de detección basadas en espectroscopía en el visible (VIS) e infrarrojo cercano (NIR) . . . 12
3.3.2 Quimiometría . . . 13
3.4 Metodologías de preprocesamiento, selección de características y análisis de clasificación . . . 15

4 Marco teórico 18
4.1 Métodos de Preprocesamiento . . . 18
4.1.1 Tratamiento de atípicos . . . 19
4.1.2 Métodos de Suavizamiento . . . 21
4.2 Métodos de selección de características . . . 25
4.2.1 RELIEF . . . 25
4.3 Métodos de Clasificación . . . 26
4.3.1 Conjuntos de prueba y entrenamiento . . . 26
4.3.2 Análisis Lineal Discriminante (ALD) . . . 27
4.3.3 Análisis Cuadrático Discriminante (ACD) . . . 30

4.3.4 Bosques Aleatorios (BA) . . . 31
4.3.5 Bayes Ingenuo (BI) . . . 32
4.3.6 Máquinas de Soporte Vectorial (MSV) . . . 33
4.3.7 K vecinos más cercanos (KVC) . . . 34
4.3.8 Perceptrón Multicapa (PM) . . . 35
4.4 Métricas de Clasificación . . . 37

5 Metodología 40
5.1 Metodología . . . 40
5.2 Población y muestra . . . 40
5.2.1 Inoculación . . . 41
5.3 Metodología de espectroscopía . . . 41
5.4 Metodología de preprocesamiento . . . 41
5.5 Metodología de selección de longitudes de onda . . . 42
5.6 Metodología estadística . . . 43

6 Resultados 44
6.1 Descripción de los espectros . . . 44
6.1.1 Descripción de los espectros en el tercer día . . . 45
6.1.2 Descripción de los espectros en el sexto día . . . 46
6.2 Detección de atípicos . . . 47
6.2.1 Detección de atípicos en el tercer día . . . 48
6.2.2 Detección de atípicos en el sexto día . . . 50
6.3 Selección de longitudes de onda . . . 52
6.3.1 Selección de características en el tercer día . . . 53
6.3.2 Selección de características en el sexto día . . . 54
6.4 Análisis de Clasificación . . . 56
6.4.1 Análisis de clasificación en el tercer día . . . 56
6.4.2 Análisis de clasificación en el sexto día . . . 65

7 Conclusiones y recomendaciones 74
7.1 Conclusiones . . . 74
7.2 Recomendaciones . . . 75

Bibliografía 77

Lista de símbolos

Abreviaturas

Abreviatura	Término
ACD	Análisis Cuadrático Discriminante
ACP	Análisis de Componentes Principales
ACV	Análisis de componentes de vecinos
AD	Árboles de decisión
$ADMCP$	Análisis discriminante de mínimos cuadrados parciales
ALD	Análisis Lineal Discriminante
$ANOVA$	Análisis de varianza
APD	Agar Papa Dextrosa
APS	Algoritmo de proyecciones sucesivas
$ASBD$	Análisis de sensibilidad basado en datos
$AUGURA$	Asociación de bananeros de Colombia
BA	Bosques Aleatorios
BI	Bayes Ingenuo
C	Carbono
CDM	Corrección de dispersión multiplicativa
$CDME$	Corrección de dispersión multiplicativa extendida
CH	Grupo de un átomo de carbono (C) unido a un átomo de hidrógeno (H)
CM	Centrado medio
DO	Datos Originales
DR	Diferencia de reflectancia
EC	Eliminado Continuo
$ELISA$	Ensayo Inmunosorbente Ligado a Enzimas
EN	Estandarización Normal
ET	Eliminación de tendencias
$EVNI$	Eliminación de variables no informativas
FBR	Función de base radial
FM	Filtro mediano
FN	Falsos Negativos
FN	Falsos negativos

Abreviatura	**Término**
$FOCR1$	*Fusarium oxysporum f.sp. cubense Raza 1*
FP	Falsos Positivo
FP	Falsos positivos
$Frog$	Frog aleatorio
GBE	Gradiente Boosting estocástico
H	Hidrógeno
KVC	K vecinos más cercanos
$MAEBK$	Máquina de aprendizaje extremo basada en kernel
$MBRNCU$	Modelo basado en red neuronal convolucional unidimensional
$MCAR$	Muestreo competitivo adaptativo reponderado
$MCAS$	Mínimos cuadrados Asimétricos
MSV	Máquinas de Soporte Vectorial
$MSVMC$	Máquina de vectores de soporte de mínimos cuadrados
N	Nitrógeno
NH	Grupo de un átomo de nitrógeno (N) unido a un átomo de hidrógeno (H)
NIR	Espectroscopía del infrarrojo cercano (750 – 2.500nm)
nm	nanómetros
O	Oxígeno
OH	Grupo funcional hidroxilo
OM	Ondas electromagnéticas
ORC	Clasificador One-vs
P	Fósforo
PCR	Reacción en Cadena de la Polimerasa
PM	Perceptrón Multicapa
$PMDRA$	Perceptrón multicapa con determinación de relevancia automatizada
RAD	Método de detección de atípicos representación de alta dimensión
RCP	Redes de creencias profundas
$RFXY$	Red de fusión XY
RNA	Red neuronal artificial
RNC	Red neuronal convolucional
RNR	Red neuronal de retropropagación
RPD	Reflectancia de primera derivada
RPT	Radar de penetración terrestre
$RSR2$	*Ralstonia solanacearum Raza 2*
S	Azufre
SG	Savitzky-Golay
$SG2$	Savitzky-Golay de segundo orden derivativo
$SIMCA$	Modelado independiente suave de analogías de clase
SPM	Suavizado de promedio móvil

Abreviatura	**Término**
SR	Sensibilidad de reflectancia
$SRAM$	Spline de regresión adaptativa multivariante
UCC	Umbral de correlación cruzada
UV	Ultra Violeta
VIS	Espectroscopía Visible (400 – 750nm)
VN	Verdaderos Negativos
VNE	Variación normal estándar
VP	Verdaderos Positivos

1 Introducción

Los datos estequiométricos desempeñan un papel crucial en los sistemas de detección temprana de enfermedades en las plantas. La relación entre el desarrollo de enfermedades en las plantas y el estrés es innegable, ya que la enfermedad puede inducir estrés en las plantas [Tunsagool et al., 2019]. Los factores de estrés en las plantas se pueden categorizar en dos tipos principales: estrés biótico, que abarca patógenos y plagas (especies biológicas que comparten el entorno y tienen interacciones con las plantas), y estrés abiótico, que incluye factores como luz, agua, salinidad y temperatura [Gull et al., 2019] [Mosa et al., 2017].

En el contexto de la detección temprana de enfermedades en las plantas, se han empleado diversos métodos, que incluyen la inspección visual, pruebas de laboratorio y técnicas espectroscópicas. Históricamente, la inspección visual ha sido ampliamente utilizada con el propósito de identificar, por ejemplo, el crecimiento de cuerpos fructíferos de hongos en los tallos de árboles de robles rojos [Luana et al., 2015]. Sin embargo, este enfoque, que requiere inspecciones continuas por parte de expertos, conlleva altos costos y puede ser subjetivo en su interpretación.

En contraste, se han desarrollado métodos más objetivos, como las pruebas de laboratorio, que incluyen pruebas serológicas y moleculares como el Ensayo Inmunosorbente Ligado a Enzimas (ELISA) [Klap et al., 2020] y la Reacción en Cadena de la Polimerasa (PCR) [Dupas et al., 2019] [Madihah et al., 2014]. Por ejemplo, en 2020 se realizó un estudio para detectar el virus de la mancha de hoja clorótica de manzana mediante la prueba ELISA [Koc et al., 2020]. A pesar de su utilidad, estas técnicas requieren tiempo para su análisis y no son eficientes en la detección temprana de enfermedades, especialmente en cultivos extensos.

Por su parte, las técnicas espectroscópicas representan un enfoque prometedor. Son métodos no destructivos que no demandan mucho tiempo, son confiables y altamente efectivos en la detección temprana de enfermedades en las plantas. Entre las técnicas espectroscópicas y de imágenes más utilizadas en este campo se encuentran la espectroscopía de fluorescencia, la espectroscopía visible e infrarroja, y las imágenes de fluorescencia. [Farber et al., 2019a] [Farber et al., 2019b] [Sankaran et al., 2010].

La detección temprana de enfermedades en plantas, como el banano Gros Michel, es esen-

cial para la seguridad alimentaria y la agricultura. Para abordar este desafío, se emplean análisis de preprocesamiento de datos, la selección de características y el análisis de clasificación supervisada. Estos análisis han demostrado ser efectivos y confiables en la identificación temprana de enfermedades, como la marchitez vascular causada por el hongo *Fusarium oxysporum* y la infección bacteriana *Ralstonia solanacearum*.

El preprocesamiento de datos implica la preparación y limpieza de los datos de reflectancia, lo que permite un análisis más preciso. Investigaciones previas han destacado la importancia de este paso en la detección temprana de enfermedades en plantas.[Rinnan et al., 2009]

La selección de características, en particular las longitudes de onda específicas en el VIS y en el NIR, desempeña un papel crucial. Estas se eligen cuidadosamente para identificar diferencias significativas entre tratamientos y condiciones, lo que mejora la precisión de la detección. El análisis de clasificación supervisada es fundamental para asignar categorías a los datos. Se han empleado métodos como el ALD , el ACD y otros, que han demostrado su eficacia en la identificación de enfermedades en plantas, incluso en etapas asintomáticas .

El banano Gros Michel es particularmente vulnerable a enfermedades como la marchitez vascular y la infección por *Ralstonia solanacearum*. La detección temprana es crítica para prevenir pérdidas devastadoras en las plantaciones. El análisis de clasificación supervisada y su aplicación específica en el contexto del banano Gros Michel representan un enfoque prometedor en la detección temprana de enfermedades, podría tener un impacto significativo en la industria agrícola.

El propósito central de este estudio es la comparación y evaluación de metodologías de clasificación, para determinar cuál de estas técnicas de clasificación supervisada es más eficiente en la detección temprana de plantas sometidas a estrés biótico y abiótico, utilizando como base la espectroscopía de reflectancia VIS/NIR. Esta investigación se centra en abordar los desafíos relacionados con el diagnóstico temprano de enfermedades en plantas, particularmente en el contexto del banano Gros Michel, que es vulnerable a patógenos como el FOCR1 y la RSR2 , lo que subraya la relevancia de esta investigación para la industria agrícola.

Es importante señalar que este estudio tiene sus limitaciones, ya que se enfoca en experimentos controlados dentro de un diseño de experimentos específico. Aunque los resultados y las metodologías proporcionarán información valiosa sobre la detección temprana de estrés en plantas sometidas a condiciones controladas, es esencial reconocer que la agricultura práctica involucra factores externos significativos, como las condiciones del suelo y el clima. Estos factores pueden variar ampliamente en el entorno real de la agricultura, lo que limita la generalización de los resultados a situaciones fuera de un entorno experimental controlado. A pesar de estas limitaciones, este estudio ofrece una base sólida para futuras investigaciones

y presenta avances prometedores en la detección temprana de estrés en plantas, lo que puede tener aplicaciones significativas en la agricultura y la seguridad alimentaria.

2 Planteamiento del problema

La problemática de las pérdidas económicas, culturales, de mano de obra y seguridad alimentaria asociadas a enfermedades en plantas, en particular en los cultivos de banano, es un tema de gran relevancia a nivel mundial, en Latinoamérica y, específicamente, en Colombia. Estas afecciones tienen un impacto significativo en la producción agrícola y la seguridad alimentaria, y el FOCR1 y la RSR2 son de las plagas más perjudiciales para el cultivo de banano.

El FOCR1 es el agente causante de una variante devastadora de la marchitez vascular denominada "Panamá disease". Esta enfermedad ha conllevado a la destrucción de extensas zonas de cultivos de banano. La propagación del FOCR1 puede acarrear pérdidas económicas significativas, debido a la disminución de la producción y a la necesidad de erradicar y reemplazar plantaciones afectadas.

Junto con las implicaciones económicas, estas enfermedades pueden tener un impacto cultural importante en las comunidades que dependen de la producción de banano. La pérdida de cultivos con frecuencia se traduce en detrimentos de empleos y de medios de vida para numerosas personas en las regiones productoras de banano de todo el mundo. Además, la seguridad alimentaria se ve amenazada, ya que el banano representa una fuente esencial de alimento en muchas zonas.

En 2019 la producción mundial de banano superó las 115 millones de toneladas [FAO, 2021]. Específicamente, la marchitez vascular provocada por el FOCR1, en particular la Raza 4 Tropical (Foc R4T), se presenta como una amenaza importante para la producción de banano en diversas regiones, y se ha estimado que sus efectos pueden ocasionar pérdidas de miles de millones de dólares [FAO, 2019]. Las enfermedades del banano pueden impactar considerablemente a países exportadores ubicados en América Latina, África y Asia causando notables pérdidas económicas.

En América Latina la producción de banano desempeña un papel fundamental en las economías de diversos países, entre los que se incluyen Ecuador, Costa Rica, Colombia. Las enfermedades que afectan los cultivos de banano, como el FOCR1 y la RSR2, tienen el potencial de ocasionar pérdidas económicas significativas en la región, dada la importancia de estos cultivos como fuente principal de ingresos por exportaciones. La propagación de estas

enfermedades repercute negativamente en la competitividad de los productores latinoamericanos en los mercados internacionales, afectando así su presencia global.

Aunado a las consecuencias económicas, la industria bananera en América Latina enfrenta desafíos adicionales relacionados con la sostenibilidad y la salud de los trabajadores agrícolas, ya que las medidas de control de enfermedades pueden requerir el uso de productos químicos. Esto plantea inquietudes relacionadas con la salud ocupacional y la seguridad de los trabajadores en el sector agrícola [FAO, 2017].

Ecuador, uno de los principales productores de banano en la región, ha experimentado las repercusiones de enfermedades del banano, como la "Panamá disease", que han resultado en la pérdida de miles de hectáreas de cultivos [Montoya Rios et al., 2022]. Costa Rica, otro destacado exportador de banano en América Latina, también se ha visto enfrentado con problemáticas vinculadas a enfermedades en los cultivos de banano [el Financiero, 2019]. Estos ejemplos subrayan la importancia crítica de abordar y gestionar de manera efectiva las enfermedades de las plantas en la industria bananera de América Latina.

En Colombia el cultivo de banano es una de las principales actividades agrícolas. Se ha estimado que las pérdidas causadas por enfermedades como *Fusarium oxysporum* y *Ralstonia solanacearum* pueden alcanzar millones de dólares anuales [Espectador, 2019]. Las pérdidas económicas debido a la marchitez vascular por Fusarium han sido significativas, y la infección ha llevado a la erradicación de miles de hectáreas de plantaciones [(ICA), 2020].
En resumen, las enfermedades en plantas, como el FUCR2 y el RSR2 representan una problemática global que afecta negativamente la economía, la cultura, la seguridad alimentaria y la mano de obra en la producción de banano. Esto plantea la necesidad de abordar el estudio de estas enfermedades de manera efectiva a nivel mundial, regional y local para garantizar la sostenibilidad de la industria bananera y la seguridad alimentaria de las comunidades.
Se ha usado espectroscopía con distintos métodos para detección temprana pero no se ha usado para comparar cuales métodos son más eficientes, es por esto que la pregunta de investigación es: ¿Cuáles son la metodologías de preprocesamiento y de clasificación más eficiente para detección temprana de plantas sometidas a estrés biótico y abiótico con base en Espectroscopía de reflectancia VIS/NIR? Para responder a esta pregunta, se plantean los siguientes objetivos: Principal: Comparar las metodologías de clasificación con el fin de establecer cuál de estas es más eficiente para detección temprana de plantas sometidas a estrés biótico y abiótico con base en Espectroscopía de reflectancia VIS/NIR. Específicos: (i) Describir el comportamiento de la reflectancia de los espectros VIS/NIR de plantas sanas, sometidas a EH, infectadas con: FOCR1 o Ralstonia solanacearum, y sus interacciones. (ii) Identificar las longitudes de onda que generan mayor discriminación entre los distintos grupos experimentales. (iii) Comparar entre las metodologías de preprocesamiento y clasificación supervisada. (vi) Identificar las mejores combinaciones de técnicas para realizar una

casificación óptima de las plantas según sus enfermedades.

3 Antecedentes investigativos

La espectroscopía aplicada a la detección de enfermedades en plantas representa una herramienta poderosa en la agricultura y la seguridad alimentaria. La capacidad de analizar la firma espectral de las plantas en el rango de espectro visible (VIS) e infrarrojo cercano (NIR) ha revolucionado la detección temprana de estrés biótico y abiótico en la vegetación. La espectroscopía VIS/NIR se ha convertido en un enfoque primordial en la identificación de enfermedades, permitiendo la adquisición de datos rápidos y no invasivos que revelan información valiosa sobre el estado de salud de las plantas.

El éxito de esta técnica de espectroscopía radica en la quimiometría, que comprende una serie de análisis estadísticos y matemáticos diseñados para procesar y analizar los datos espectrales. La quimiometría es esencial para extraer información significativa de los espectros y transformarla en conocimiento práctico. Esto incluye la selección de características relevantes, el preprocesamiento de datos para mejorar la calidad de la información y la aplicación de algoritmos de clasificación y detección. La combinación de espectroscopía VIS/NIR y quimiometría ha brindado avances notables en la identificación de enfermedades en plantas, lo que tiene un impacto significativo en la agricultura y la seguridad alimentaria.

3.1. Detección enfermedad en plantas

Los datos estequiométricos desempeñan un papel fundamental en los sistemas de detección temprana de enfermedades en plantas. El desarrollo de enfermedades en las plantas está estrechamente vinculado al estrés, ya que la enfermedad puede inducir estrés en la planta [Tunsagool et al., 2019]. Los factores de estrés en las plantas se pueden clasificar en estrés biótico (causado por patógenos y plagas, es decir, otras especies biológicas que comparten el entorno e interactúan con las plantas) y estrés abiótico (como luz, agua, salinidad y temperatura) [Gull et al., 2019] [Mosa et al., 2017].

Para la detección temprana de enfermedades en plantas, se han utilizado métodos como la inspección visual, pruebas de laboratorio y técnicas espectroscópicas [Rinnan et al., 2009]. Aunque las técnicas de inspección visual han sido ampliamente empleadas con este propósito, un estudio realizado en 2015 utilizó esta metodología para identificar el crecimiento de cuerpos fructíferos de hongos en el tallo de los robles rojos [Luana et al., 2015]. Dado que

este método requiere una inspección continua por parte de expertos, puede generar altos costos económicos y no es muy confiable, ya que está sujeto a opiniones subjetivas.

Por otro lado, se han desarrollado métodos más objetivos, como las pruebas de laboratorio para la detección temprana de enfermedades mediante técnicas serológicas y moleculares, las cuales incluyen el Ensayo Inmunoabsorbente Ligado a Enzimas (ELISA) [Klap et al., 2020] y la Reacción en Cadena de la Polimerasa (PCR) [Dupas et al., 2019] [Madihah et al., 2014]. En el caso del ELISA, un estudio realizado en 2020 buscaba detectar el virus de la mancha de hoja clorótica de manzana mediante este método [Koc et al., 2020]. Aunque estas técnicas son efectivas, requieren mucho tiempo para su análisis y no son ideales para la detección temprana de enfermedades, especialmente en cultivos extensos.

Las técnicas espectroscópicas son métodos no destructivos para las plantas, requieren poco tiempo, son confiables y altamente efectivas en la detección temprana de enfermedades. Entre estas se encuentran la espectroscopía de fluorescencia, la espectroscopía visible e infrarroja, así como las imágenes de fluorescencia [Farber et al., 2019a] [Farber et al., 2019b] [Sankaran et al., 2010].

3.2. Estrés biótico y abiótico en plantas de banano Gros Michel

La producción de banano Gros Michel, una variedad altamente cultivada en el mundo, se ve constantemente desafiada por diversos factores que afectan la salud y el rendimiento de las plantas. [Manzo-Sánchez et al., 2014] Entre estos desafíos, el estrés biótico y abiótico emerge como un componente crítico que puede influir significativamente en la productividad de los cultivos. En este contexto, se destacan tres elementos clave: *Fusarium oxysporum*, *Ralstonia solanacearum* y el estrés hídrico.

3.2.1. Fusarium oxysporum

Una de las amenazas más importantes para las plantas de banano Gros Michel es la infección por *Fusarium oxysporum*, un hongo que causa la marchitez vascular. Esta enfermedad, también conocida como "Panama disease", puede resultar en pérdidas catastróficas al provocar el colapso del sistema vascular de la planta, afectando su capacidad para absorber agua y nutrientes. La detección temprana de esta infección es esencial para implementar medidas preventivas y de control que preserven la salud de las plantaciones.

La enfermedad conocida como marchitez por Fusarium en bananos, o fusariosis, es causada por *Fusarium oxysporum f. sp. cubense raza 1*, la cual ha sido intensificada más recientemente por la raza tropical 4 de Foc (Foc TR4) en el banano Cavendish [Wang et al., 2020]. Recientemente, esta enfermedad ha sido clasificada como Fusarium odoratissimum, quienes consideran que actualmente representa la mayor amenaza para la producción global de Musaceae en general. Se trata de una enfermedad en la cual el patógeno invade, coloniza y obstruye los vasos xilema de las raíces, interrumpiendo la translocación de agua y nutrientes, lo que provoca una marchitez severa [Li et al., 2014]. Sus síntomas típicos incluyen amarillamiento y marchitez de las hojas, decoloración vascular en el rizoma y pseudotallo, y la muerte de la planta infectada [Ploetz, 2006]. No se conoce un control químico efectivo para tratar este hongo y, dado que es un patógeno vascular, su detección y diagnóstico son complejos[Macias-Echeverri et al., 2022].

Las enfermedades del tipo vascular, como la marchitez causada por Fusarium, son difíciles de detectar en las primeras etapas. Otros tipos de enfermedades, como las localizadas o superficiales, pueden detectarse mediante la observación de síntomas; sin embargo, en el caso de enfermedades vasculares, los síntomas visibles indican que el hongo u organismo patógeno ya ha invadido el tejido vascular [García-Bastidas et al., 2020]. Al observar las interacciones entre el patógeno y los hospedadores, se pueden identificar una variedad de síntomas y daños en las plantas, proporcionando una base para el monitoreo mediante teledetección [Zhang et al., 2019]. No obstante, la detección temprana requiere la capacidad de identificar aquellas características que hacen única cada infección, incluso en ausencia de síntomas externos. El comportamiento de las plantas saludables debe compararse con el de las plantas enfermas para lograr diagnósticos exitosos [Zhang et al., 2012].[Macias-Echeverri et al., 2022].

3.2.2. Ralstonia solanacearum

Es una bacteria fitopatógena causante de enfermedades vasculares en diversas plantas, siendo particularmente relevante en el cultivo del banano. La *Ralstonia solanacearum* pertenece al complejo de especies que incluye varias subespecies y cepas que pueden afectar una amplia gama de hospederos vegetales. La cepa que afecta al banano es conocida como *Ralstonia solanacearum* Raza 2 (RSR2)[Genin and Denny, 2012].

La RSR2 es responsable de inducir enfermedades como la marchitez bacteriana, siendo un patógeno vascular que invade los sistemas de conducción de agua y nutrientes de la planta, conocidos como los vasos xilemáticos. Una vez dentro la RSR2 provoca obstrucciones en estos conductos, impidiendo el flujo normal de agua y nutrientes. La consecuencia directa de esta infección es la marchitez de la planta, ya que se ve privada de los recursos esenciales para su desarrollo y supervivencia [Ploetz, 2015].

En el caso específico del banano, esta bacteria puede ocasionar enfermedades como la llamada Moko o "Mal de Panamá", que es causada por la cepa RSR2. Esta enfermedad es altamente destructiva y puede tener consecuencias devastadoras en las plantaciones de banano [Fegan and Prior, 2006]. Las plantas afectadas por RSR2 muestran síntomas como amarillamiento de las hojas, marchitez, necrosis y, en casos severos, la muerte de la planta.

La transmisión de RSR2 puede ocurrir a través del suelo contaminado, agua de riego, herramientas agrícolas contaminadas y otros vectores. Una vez que el suelo o el sistema radicular de una planta está contaminado, la bacteria puede propagarse rápidamente en condiciones propicias.

La importancia de comprender y controlar RSR2 radica en su capacidad para causar pérdidas significativas en la producción de banano, uno de los cultivos alimentarios más importantes a nivel mundial. Los esfuerzos de investigación y manejo están dirigidos a desarrollar estrategias eficaces para prevenir, controlar y gestionar las infecciones por *Ralstonia solanacearum*, contribuyendo así a la sostenibilidad y seguridad alimentaria en las regiones donde se cultiva el banano.

3.2.3. Estrés hídrico

Se refiere a la situación en la cual la demanda de agua por parte de un organismo o un sistema supera la cantidad disponible [Reyes-Matamoros et al., 2014]. Este fenómeno puede afectar tanto a organismos vivos como a sistemas ecológicos, y puede tener consecuencias negativas en el crecimiento, desarrollo y funcionamiento normal de las plantas y otros seres vivos.

Cuando las plantas experimentan EH las condiciones de escasez de agua pueden afectar su capacidad para realizar funciones esenciales como la fotosíntesis, el transporte de nutrientes y la regulación térmica[Reyes-Matamoros et al., 2014]. Este estrés puede manifestarse de diversas maneras, como la reducción del crecimiento, marchitez de las hojas, cierre de estomas, y en casos extremos, incluso la muerte de la planta.

Las plantas de banano son particularmente sensibles al EH, ya que requieren una cantidad adecuada de agua para su desarrollo óptimo. Aquí hay algunos aspectos específicos relacionados con el EH como la necesidad de Agua, pues estas plantas necesitan un suministro constante de agua para su crecimiento y producción de frutos. La falta de agua puede afectar la calidad y cantidad de la cosecha.

El estrés hídrico en las plantas de banano, generado por una insuficiencia de agua en relación con la demanda, afecta negativamente su crecimiento, producción y salud [Salazar et al., 2014]. Para las regiones tropicales que dependen económicamente del cultivo de banano, la detección temprana de este estrés reviste gran importancia.

3.3. Espectroscopía usada en enfermedad de plantas

La espectroscopía se destaca como uno de los métodos más ampliamente empleados en la detección de enfermedades de las plantas debido a sus notables ventajas, que incluyen su no destructividad, no invasividad, alta velocidad, y alta sensibilidad y especificidad para identificar enfermedades específicas [Khaled et al., 2018b] [Khaled et al., 2018a] [Li et al., 2008].

En términos generales, la espectroscopía se centra en la medición y el análisis de los espectros generados a raíz de la interacción entre la materia y la radiación electromagnética, que se propaga en forma de ondas electromagnéticas (EM) [Pavia et al., 2014]. Esta técnica se puede categorizar en espectroscopia molecular, que abarca regiones como el visible, el infrarrojo, la resonancia magnética nuclear, la espectroscopia de masas y la impedancia eléctrica, y espectroscopia atómica, que incluye la espectroscopia de fluorescencia, dependiendo de la aplicación y la naturaleza de la interacción [Svanberg, 2012].

La espectroscopía VIS-NIR ha demostrado ser una herramienta efectiva y no invasiva en la identificación y análisis de una amplia gama de variables en productos agrícolas, lo que la convierte en una elección destacada para la detección temprana de enfermedades y estrés en plantas [Rumpf et al., 2010]. Este método ha demostrado su capacidad para discriminar entre condiciones de salud y enfermedad en las plantas, lo que ha generado un interés creciente en su aplicación para la identificación temprana de enfermedades y estrés antes de que los síntomas sean visibles. La información proporcionada por los datos espectrales en las regiones visible e infrarroja cercana permite la detección precoz de anomalías debidas a enfermedades, infecciones o estrés. Esta detección temprana, en particular antes de que los síntomas sean evidentes, es de vital importancia para implementar estrategias efectivas de prevención y mitigación de pérdidas en la industria agrícola.

Asimismo, se ha comprobado que el estudio de la interacción entre el agua y la luz, conocido como acuafotómica, proporciona información valiosa sobre la funcionalidad de los sistemas. Este enfoque se ha revelado fiable para evaluar el estado de diversos sistemas que contienen moléculas de agua, desde alimentos hasta biomateriales y organismos biológicos, incluyendo bacterias y animales. La acuafotómica ha demostrado que cuando las moléculas de agua interactúan con la luz, generan patrones espectrales específicos que brindan información precisa en diversas condiciones. Este método, especialmente desarrollado en el espectro visible

e infrarrojo cercano debido a su capacidad de penetración y absorbancia de la luz, se ha aplicado en estudios de patología, incluyendo el seguimiento del crecimiento de plantas como el frijol mungo en diferentes contextos [Tjandra Nugraha et al., 2021].

3.3.1. Técnicas de detección basadas en espectroscopía en el visible (VIS) e infrarrojo cercano (NIR)

La espectroscopia VIS-NIR, que abarca desde el espectro visible (400-750 nm) hasta el infrarrojo cercano, se destaca como una técnica no destructiva y de alta velocidad que ofrece la capacidad de predecir la composición química y biológica de un sistema. Mientras la espectroscopia VIS se enfoca en el análisis de color y pigmentos, la espectroscopia NIR se emplea principalmente para medir macrocomponentes, con un énfasis particular en el agua [Walsh et al., 2020].

En la región visible, la cual abarca desde 400 hasta 750 nm, se obtiene información valiosa sobre las características espectrales de los pigmentos que son fundamentales en el proceso de fotosíntesis de las plantas. Pigmentos como la clorofila, antocianina y carotenos influyen en la coloración visual de las plantas y pueden servir como indicadores cruciales para la detección de enfermedades y estrés en ellas. Cada pigmento exhibe un espectro de absorción específico; por ejemplo, la antocianina muestra una absorbancia en la región verde (530-550 nm), mientras que los carotenos absorben en el rango de longitud de onda de 420-503 nm [Zahir et al., 2022].

Por otro lado, la región del infrarrojo cercano se asocia con la medición de armónicos y vibraciones moleculares combinadas. Por ejemplo, la intensidad de una vibración asimétrica en el infrarrojo cercano podría indicar la elongación de enlaces de hidrógeno, como CH, OH y NH [Gomez et al., 2008].

El proceso de espectroscopía VIS-NIR involucra la irradiación de luz en la muestra, resulta en la vibración de los enlaces entre los átomos de C, N, H, O, P y S. Estas vibraciones provocan cambios en la longitud y el ángulo de los enlaces, lo que, a su vez, afecta la cantidad de luz absorbida, transmitida o reflejada en la muestra. La cantidad de luz absorbida por las muestras depende en gran medida de la interacción de la luz con las moléculas, esto puede usarse para estimar directamente el contenido de sustancias como el agua y los nutrientes [de Carvalho et al., 2015].

La comparación de los modos de reflectancia y transmitancia se ha estudiado en el contexto de la detección de patógenos en plantas de cebada, demostrando que los datos espectrales preprocesados a través de técnicas como el suavizado Savitzky-Golay (SG) y analizados

mediante el análisis de componentes principales (ACP) pueden proporcionar información esencial [Zahir et al., 2022]. En este estudio, la reflectancia se destacó al permitir la detección temprana de la infección, incluso antes de que los síntomas sean visibles, mientras que la transmitancia mostró cambios distintos, pero dos días después de la aparición de los síntomas. En consecuencia, ambos métodos resultan complementarios, ya que la reflectancia facilita la detección anticipada de enfermedades en las plantas, mientras que los datos de transmisión enriquecen la comprensión de las interacciones planta-patógeno.

Las mediciones espectroscópicas involucran cuatro pasos fundamentales: la iluminación de las muestras con luz; la reflexión y absorción de la luz incidente; la detección de la luz transmitida y reflejada y la interpretación de los datos obtenidos [Zahir et al., 2022]. La absorción tiene lugar cuando la frecuencia de la luz utilizada para la irradiación es igual a la diferencia de energía entre el estado fundamental y el estado excitado de una molécula [Zahir et al., 2022]. Este principio, central en la espectroscopia, puede describirse mediante la ecuación de Planck (3-1):

$$E = hv = \frac{hc}{\lambda}, \tag{3-1}$$

donde,
E es la energía requerida para cambiar el estado del electrón del estado fundamental al estado de excitación, h es la constante de Planck, v es el número de onda, c es la velocidad de la luz y λ es la longitud de onda [Zahir et al., 2022]. Según esta ecuación, menor será la energía necesaria durante el estado de transición del electrón desde el estado fundamental al estado excitado, mayor será la longitud de onda de la banda de absorción. La ecuación de Planck muestra que la energía de una onda electromagnética es directamente proporcional a su frecuencia y que la constante de Planck establece la proporción. En otras palabras, a medida que la frecuencia aumenta, la energía también aumenta. Esta ecuación es fundamental para entender conceptos clave en la física cuántica, como la cuantización de la energía y la dualidad onda-partícula de la luz. Además, es esencial en la comprensión de cómo los sistemas cuánticos intercambian energía con la radiación electromagnética.
En la espectroscopÍa, los espectros originales pueden ser complicados debido al ruido de fondo y las interferencias. Este ruido y las señales no deseadas dificultan la identificación precisa de las características relevantes en los espectros.

3.3.2. Quimiometría

Es una disciplina que se basa en la aplicación de métodos matemáticos y estadísticos para extraer información valiosa de los datos espectrales. En el contexto de la espectroscopía, esta

disciplina se utiliza para establecer una relación entre los atributos químicos y/o físicos de un conjunto de muestras y los espectros correspondientes que han sido previamente medidos y sometidos a un proceso de preprocesamiento. Durante el proceso de calibración, se utiliza la ley de Beer para identificar y cuantificar la correlación entre las propiedades de las muestras y la absorbancia de la radiación electromagnética [Basa, 2022].

La ley de Beer, también conocida como la ley de Beer-Lambert, es un principio fundamental en la espectroscopia que establece una relación entre la concentración de una sustancia química en una muestra y la absorbancia de la radiación electromagnética que pasa a través de la muestra. Esta ley se utiliza para cuantificar la cantidad de sustancia presente en una muestra y se aplica a una amplia variedad de técnicas espectroscópicas, incluida la espectroscopia UV-visible, la espectroscopia infrarroja y muchas otras.

La ley de Beer se expresa matemáticamente de la siguiente manera [Zahir et al., 2022]:

$$A = e * c * l, \tag{3-2}$$

donde,

A es la absorbancia de la muestra. e es el coeficiente de absorción molar de la sustancia (también conocido como coeficiente de extinción molar). Este valor es específico para cada sustancia y para una longitud de onda dada. c es la concentración de la sustancia en la muestra. Se expresa en moles por litro (Molar, M). l es la longitud del camino de la radiación a través de la muestra. Se mide en metros.
La ley de Beer establece que la absorbancia es directamente proporcional a la concentración de la sustancia y a la longitud del camino óptico de la radiación. En otras palabras, a medida que aumenta la concentración de la sustancia o la longitud del camino, la absorbancia también aumenta. Esta relación lineal es válida siempre que se cumplan ciertas condiciones, como que la radiación incidente sea monocromática y que no ocurran interacciones moleculares significativas que afecten la absorción.

La quimiometría se ha convertido en una herramienta esencial en diversas industrias, como la química, farmacéutica y alimentaria, para abordar tanto problemas descriptivos como predictivos [Zahir et al., 2022]. En las aplicaciones descriptivas, la quimiometría se emplea para modelar las propiedades de sistemas químicos con el objetivo de estudiar las relaciones subyacentes y las estructuras presentes en dichos sistemas. Por otro lado, en las aplicaciones predictivas, esta disciplina se utiliza para desarrollar modelos que permiten predecir las propiedades de sistemas químicos y, en consecuencia, anticipar el comportamiento de nuevas muestras de interés.

Dada la complejidad y la alta dimensionalidad de los conjuntos de datos espectrales, la quimiometría ha desempeñado un papel fundamental en la simplificación y la interpretación de estos datos. Los avances en quimiometría y la evolución de la instrumentación han dado como resultado la creación de métodos rápidos y precisos para el análisis de datos espectrales, lo que ha mejorado significativamente la capacidad de obtener información relevante a partir de complejos espectros y ha ampliado las aplicaciones de la espectroscopia en diversas disciplinas científicas y técnicas.

3.4. Metodologías de preprocesamiento, selección de características y análisis de clasificación

En esta sección se realiza la revisión que se centra en investigaciones basadas en espectroscopía de reflectancia VIS-NIR aplicadas en la detección temprana de estrés vegetal desarrolladas en la última década, identificándose un total de 34 estudios. Se incluyeron investigaciones que abordaban el propósito de clasificar la presencia o ausencia de estrés vegetal.

Este capítulo se enfoca en tres aspectos fundamentales: los métodos de preprocesamiento, las técnicas de selección de características (longitudes de onda) y los métodos de análisis. Durante el período de estudio, se identificaron 13 metodologías distintas de preprocesamiento de datos, que incluyen: estandarización normal (EN), Baseline centrado medio (BCM), filtro mediano (FM), suavizado de promedio móvil (SPM), Savitzky-Golay (SG), corrección de dispersión multiplicativa (CDM), variación normal estándar (VNE), reflectancia de primera derivada (RPD), estándar normal variable (ENV), datos originales (DO), eliminado continuo (EC) y eliminación de tendencias (ET). Además, se encontraron 11 métodos diferentes para la selección de características o longitudes de onda, que incluyen: muestreo competitivo adaptativo ponderado (MCAP), frog aleatorio (Frog), eliminación de variables no informativas (EVNI), bosques aleatorios (BA), algoritmo de proyecciones sucesivas (APS), análisis de componentes de vecindario (ACV), indices de vegetación (IV), análisis de sensibilidad basado en datos (ASBD), umbral de correlación cruzada (UCC), diferencia de reflectancia (DR) y sensibilidad de reflectancia (SR). Por último, se identificaron 36 métodos diferentes de análisis, que abarcan desde técnicas basadas en aprendizaje automático hasta métodos de análisis quimiométrico.

En las Tablas **3-1** y **3-2** se presentan los estudios que se centraron en el propósito de clasificar la presencia o ausencia de estrés vegetal y se proporciona información general como el año de publicación, la especie de planta estudiada, el estrés biótico y abiótico investigado, el área de la planta afectada por el estrés, las mediciones espectrales y las longitudes de onda relacionadas con cada tipo de estrés. Además, se detallan los métodos de preproce-

Tabla 3-1: Relación de investigaciones en estrés vegetal basados en espectroscopia VIS/NIR que tienen como objetivo clasificar presencia o ausencia del estrés (parte 1)

Artículo	Año	Especie	E. Biótico	E. Abiótico	Área infectada	Medición (nm)
[Shin et al., 2023]	2023	Papa	Hongo Verticillium			1596-2396
[Tu et al., 2022]	2022	Tomate	Hongo (Solanum lycopersicum)	Hídrico		348–1052
[Hou et al., 2022]	2022	Papa	oomicetos (Tizón tardío)			350–1000
[Ignat et al., 2022]	2022	Tomate		Salinidad		350-2500
[Muncan et al., 2022]	2022	Haba de Soya		Clima Frio		588–1025
[Zhang et al., 2021]	2021	Tomate		Hídrico	Hojas	325-1075
[Marín-Ortiz et al., 2020]	2020	Tomate	Hongo (Fusarium oxysporum)	Hídrico	Hojas	380-1000
[Morellos et al., 2020]	2020	Tomate	Virus (Virus de la clorosis del tomate)		Hojas	310-1100
[Yu et al., 2021]	2020	Tabaco		Metal pesado Hg	Hojas	380-1080
[Bienkowski et al., 2019]	2019	Papa	Hongo (Solanum tuberosum)		Hojas	400-1000
[R. Beghi and Guidetti, 2017]	2017	Uva	Hongo y Bacteria (Botrytis cinerea)		Frutas	400-1650
[Abdulridha et al., 2016]	2016	Aguacate	Hongo (Laurel marchitamiento y phytophthora)	Sal	Hojas, raiz	400-950
[Abu-Khalaf, 2015]	2015	Tomate	Hongo y Bacteria (h. oxysporum)		Frutas	550-1100
[Kaliramesh et al., 2013]	2013	Frijol mungo	Insecto (Gorgojo del arroz)		Núcleos	1000-1600

samiento utilizados y cuál de estos arrojó los mejores resultados, las técnicas de selección de características (longitudes de onda) empleadas, el método de selección de características óptimo y las metodologías de análisis utilizadas, indicando cuál de ellas presenta los mejores resultados en términos de precisión y clasificación exitosa.

A pesar de haber llevado a cabo una revisión del objetivo (*ii*), es importante destacar que la clasificación del estadío de la enfermedad no constituye un aspecto de interés primordial para el presente estudio. En consecuencia, la información obtenida durante la revisión de dicho objetivo no será considerada en el análisis y desarrollo de la investigación. Este enfoque se basa en la decisión deliberada de centrarse en otros aspectos más específicos y relevantes para los objetivos y alcances del estudio en cuestión.

Tabla 3-2: Relación de investigaciones en estrés vegetal basados en espectroscopia VIS/NIR que tienen como objetivo clasificar presencia o ausencia del estrés (parte 2)

Artículo	L.O estrés	Tranf.	T. óptima	Selección	S. óptimo	Análisis	A. óptimo
[Shin et al., 2023]						ANN	ANN
[Tu et al., 2022]		SNV	NS			CNN-1D, PLS-DA, RF	1D-SP-NET
[Hou et al., 2022]		Baseline, MC, MF, MA, SG, MSC, SNV	MF	CARS, Frog, UVE, RF	Frog	PLS, SVM, KNN, DT, ANOVA	SVM
[Ignat et al., 2022]	500-600, 986	SG	SG			PLS-DA	PLS-DA
[Muncan et al., 2022]	799–803, 827, 868–874, 880, 900, 908, 918–922, 928, 934, 943–946, 959, 973, 985, 995–996	Baseline		PCA	PCA	SIMCA	
[Zhang et al., 2021]	483, 557, 674, 783, 869, 964	FDR	FDR	SPA	SPA	MLPC, ORC, CNN	MLPC
[Marín-Ortiz et al., 2020]	510, 560, 658, 694, 750	SNV	SNV	PCA	PCA	ANOVA, LDA	LDA
[Morellos et al., 2020]	550, 670,750	SG	SG	NCA	NCA	XY-F, MLP-ARD	MLP-ARD
[Yu et al., 2021]	510, 680, 740, 940	OD, SNV	OD, SNV	CARS	CARS	PCA, PLS-DA, LS-SVM	LS-SVM
[Bienkowski et al., 2019]	550,740,970	FDR	FDR			PLS-R, BPNN	BPNN
[R. Beghi and Guidetti, 2017]	550,680,840,970	MA	MA	PCA	PCA	PLS-DA	PLS-DA
[Abdulridha et al., 2016]	550,750,845	OD	OD			STEPDISC, MLPC, RBF	MLPC
[Abu-Khalaf, 2015]	580-670, 675-690, 700-950, 1000-1050	SG, Baseline	Baseline	PCA	PCA	PCA, SVM	SVM
[Kaliramesh et al., 2013]	1100, 1290, 1450	PCA	PCA	PCA	PCA	LDA, QDA	QDA

4 Marco teórico

En este capítulo se exploran diversas metodologías de preprocesamiento, selección de características y clasificación supervisada. En la fase de preprocesamiento, se abordan métodos para tratar atípicos, como el MB y el método de representación de alta dimensión (RAD), que buscan optimizar la detección y manejo de datos atípicos en la información espectroscópica. Asimismo, se analizan métodos de suavizamiento, entre ellos, el método de SPM, el SG, el CDM, y el MCA, con el objetivo de reducir el ruido inherente a las mediciones.

En la etapa de selección de características, se explora la aplicación del método RELIEF, una técnica que identifica las longitudes de onda más relevantes para discriminar entre los diferentes tratamientos espectroscópicos. Este enfoque permite optimizar la eficiencia del análisis al centrarse en las características más informativas, contribuyendo a una clasificación más precisa.

Finalmente, en el ámbito de la clasificación supervisada, se examinan diversos métodos: ALD, ACD, BA, BI, MSV, KVC, y PM. Estos enfoques ofrecen distintas aproximaciones para asignar eficientemente los tratamientos a categorías específicas, destacando la importancia de elegir la metodología más adecuada para maximizar la precisión en la clasificación de datos espectroscópicos asociados a plantas de banano Gros Michel.

4.1. Métodos de Preprocesamiento

En el análisis de espectroscopía, varios factores pueden introducir información irrelevante en los espectros, lo que incluye ruido de fondo, dispersión de la luz y variaciones en las geometrías de las muestras. Para obtener resultados precisos, es crucial realizar una serie de pasos de preprocesamiento para eliminar el ruido, corregir los efectos de la línea de base y ajustar las diferencias en la geometría de las muestras.

Al seleccionar un método de preprocesamiento en el análisis de datos espectrales, es esencial encontrar un equilibrio entre la eliminación del ruido no deseado y el exceso de suavizado que podría conducir a un sobreajuste de los datos. Un modelo que se ajusta en exceso con polinomios de alto grado o un suavizado excesivo puede introducir ruido innecesario y errores

engañosos durante la validación del modelo. Según Rinnan et al. 2009, los métodos de preprocesamiento más comúnmente utilizados se pueden clasificar en dos categorías principales: derivados espectrales y métodos de corrección de dispersión [Zahir et al., 2022].

Los métodos de derivados espectrales, como el filtrado SG, aplican suavizado a los espectros antes de calcular derivadas para mejorar la relación señal-ruido. Estos métodos son eficaces para eliminar los efectos de la línea de base en los datos.

Por otro lado, los análisis de corrección de dispersión, que incluyen la CDM reducen los efectos de dispersión que ocurren dentro de las muestras. Estos se centran en minimizar los efectos de dispersión en los datos espectrales.

Algunos análisis específicos utilizados en el preprocesamiento de datos espectrales incluyen el SPM, el SG, las derivadas de primer orden (DPO) y segundo orden (DSO), la transformación de CDM y el método MCA. Estos métodos desempeñan un papel fundamental en la mejora de la calidad de los datos espectrales y permiten un análisis más preciso y confiable.

4.1.1. Tratamiento de atípicos

El tratamiento de atípicos, es una etapa esencial en el análisis de datos espectroscópicos. Estos representan observaciones que difieren significativamente del patrón general de los datos y pueden distorsionar la interpretación de los resultados. Estos tratamientos de atípicos contribuyen a la robustez y fiabilidad del análisis posterior, asegurando que los resultados de la clasificación estén basados en información precisa y representativa. Las técnicas que se presentan a continuación son muy usadas en el contexto de datos funcionales.

Criterios del método de la bolsa (MB)

Este método se basa en la detección de valores que se alejan significativamente de la distribución típica de los datos funcionales, los cuales son una extensión de los datos multivariados donde cada observación se representa como una función, generalmente una curva suave [Ramsay and Silverman, 2002].

El MB es una herramienta gráfica utilizada para identificar observaciones atípicas en datos funcionales. Se compone de dos regiones: interior, es una región que rodea la mediana de las observaciones funcionales, y el exterior qu es una región que se encuentra fuera del interior y es utilizada para definir los límites de lo que se considera normal en los datos funcionales.

El criterio de detección de datos atípicos en un método de la bolsa se basa en el principio de que cualquier observación que caiga fuera de los límites del exterior se considera un valor atípico. Estos límites se definen en términos de la mediana y la dispersión de las observaciones funcionales.El MB también se puede utilizar para identificar patrones de datos atípicos que pueden no ser evidentes en otras representaciones gráficas. Por ejemplo, si varias observaciones funcionales caen fuera del exterior en una región específica, esto podría indicar una tendencia o patrón de datos inusual en esa región [Rousseeuw et al., 1999].

En resumen, el criterio de detección de datos atípicos para datos funcionales mediante el MB se basa en la identificación de observaciones que caen fuera de los límites definidos por el fence en relación con la mediana y la dispersión de las observaciones. Estas observaciones se consideran atípicas y pueden indicar patrones inusuales en los datos funcionales.

Representación de Alta Dimesión (RAD)

El Criterio RAD es una técnica utilizada en la detección de datos atípicos o anómalos en datos funcionales. Está relacionado con la representación gráfica de proyecciones de datos funcionales en un espacio de alta dimensión. La idea es evaluar si los datos funcionales tienen una estructura específica en un espacio de alta dimensión y si algunos puntos se desvían significativamente de esta estructura. Si se desvían, se consideran atípicos [Ramsay and Silverman, 2002].

El RAD se basa en el análisis de regresión y ajuste de modelos a los datos funcionales. Para detectar datos atípicos, se siguen estos pasos: [Choi and Marron, 2019]

1. Se ajusta un modelo de regresión en el espacio de alta dimensión de los datos funcionales. Este modelo puede ser lineal o no lineal, según el contexto del problema. Para un conjunto de datos funcionales X_i con $i = 1, 2, \ldots, n$, donde n es el número de observaciones, se ajusta un modelo de regresión que representa la estructura de los datos funcionales como una función f en función de un conjunto de covariables Y_i en un espacio de alta dimensión:

   $$X_i = f(Y_i) + \varepsilon_i, \tag{4-1}$$

   donde,
   ε_i es el término de error fucional.

2. Se evalúa cuán bien se ajusta el modelo a los datos funcionales originales, aquí se calculan las proyecciones de los datos funcionales originales y las proyecciones del modelo:

$$P_{orig} - X_i - \widehat{f}(Y_i), \tag{4-2}$$

donde,
P_{orig} es la proyección de los datos originales y $\widehat{f}(Y_i)$ es la proyección del modelo ajustado.

3. Se comparan las proyecciones del modelo con las proyecciones de los datos originales para cada punto de datos. Los puntos que tienen una gran discrepancia entre las proyecciones del modelo y las proyecciones originales se consideran atípicos.

4.1.2. Métodos de Suavizamiento

Los métodos de suavizamiento desempeñan un papel crucial en el procesamiento de datos espectroscópicos, contribuyendo a mejorar la calidad de la información al reducir el ruido y realzar las características relevantes.

Método de Suavizado de Promedio Movil (SPM)

El SPM es una técnica de suavizado utilizada en el procesamiento de datos espectrales que se basa en un algoritmo similar al del vecino más cercano [Buja et al., 1989]. Es uno de los métodos más simples de suavizado y se utiliza para promediar un conjunto de muestras adyacentes en los datos espectrales con el fin de eliminar datos atípicos o valores que están fuera de rango.
Este método es especialmente útil cuando se tratan a datos espectrales ruidosos o se quieren eliminar variaciones espurias en los datos. La operación básica del SPM implica el cálculo de un nuevo valor promediando un conjunto de muestras que se encuentran en una ventana definida que se desplaza a lo largo del espectro.
Matemáticamente, el cálculo del Running Mean Smoother se expresa de la siguiente manera: Dado un conjunto de datos espectrales X con N puntos espectrales, el valor suavizado $X_{suavizado}$ en la posición i se calcula promediando los valores en una ventana de tamaño W que se desplaza a lo largo del espectro:

$$X_{suavizado}[i] = \frac{1}{W} \sum_{j=i-\frac{W}{2}}^{i+\frac{W}{2}} X[j], \tag{4-3}$$

donde,
$X_{suavizado}[i]$ es el valor suavizado después de aplicar el SPM en la posición i del espectro.
W es el tamaño de la ventana utilizada para el promedio.

j es un índice que recorre los valores dentro de la ventana centrada en la posición i. $X[j]$ es el valor en la posición j del espectro original.
La elección del tamaño de la ventana W es crucial, ya que determina el nivel de suavizado aplicado a los datos. Un valor pequeño de W mantendrá más detalles del espectro, pero no reducirá el ruido significativamente, mientras que un valor grande de W proporcionará un suavizado más efectivo, pero puede eliminar detalles importantes.

El Método SPM es una técnica efectiva para reducir el ruido en los datos espectrales y eliminar valores atípicos. Sin embargo, al igual que con otros métodos de suavizado, es importante encontrar un equilibrio adecuado en la elección del tamaño de la ventana para satisfacer los objetivos específicos del análisis de datos espectrales. [Buja et al., 1989]

Método Savitzky-Golay (SG)

El Método SG es una técnica de suavizado ampliamente utilizada en el procesamiento de datos espectrales para mejorar la precisión de los datos sin distorsionar significativamente los espectros originales. Este método se basa en el concepto de convolución y se utiliza para ajustar una curva polinomial a un número específico de puntos en el espectro. La aplicación de este filtro es fundamental para mantener una relación señal/ruido aceptable.

La convolución, en el contexto del Método Savitzky-Golay, implica el deslizamiento de una ventana móvil a lo largo del espectro. En cada posición de la ventana, se ajusta un polinomio de cierto orden para representar los puntos dentro de la ventana. El resultado es una curva suavizada que refleja una versión más precisa de la señal subyacente, eliminando el ruido y las fluctuaciones no deseadas. [Savitzky and Golay, 1964]

El cálculo del Método Savitzky-Golay se describe mediante la siguiente ecuación:
Dado un conjunto de datos espectrales X con N puntos espectrales, el valor suavizado $X_{suavizado}$ en la posición i se calcula ajustando un polinomio de orden M en una ventana de tamaño W centrada en i:

$$X_{suavizado}[i] = \sum_{k=0}^{M} a_k X[j - \frac{W}{2} + k], \tag{4-4}$$

donde,
$X_{suavizado}[i]$ es el valor suavizado después de aplicar el método SG en la posición i,
W es el tamaño de la ventana utilizada para la convolución,
M es el orden del polinomio que se ajusta en cada ventana,
k es in índice que recorre los coeficientes del polinomio de orden M,

a_k son los coeficientes del polinomio que se ajustan para obtener la mejor aproximación de la señal de la ventana.

La elección del tamaño de la ventana W y el orden del polinomio M es crucial en el método Savitzky-Golay. Un valor adecuado de estas constantes deben determinarse según las características específicas de los datos espectrales y los objetivos del análisis. En espectroscopía se usa frecuentemente y los valores de M aconsejados en estos casos es de 0 el cual es Método Savitzky-Golay original y cuando $M = 2$ toma el nombre de Método Savitzky-Golay de segundo orden derivativo.

El Método Savitzky-Golay tiene varias ventajas, ya que no solo reduce el ruido de la señal, sino que también preserva características importantes de los espectros, como la posición y el ancho de los picos de absorbancia. Esto lo convierte en una herramienta valiosa para el análisis de datos espectrales en diversas aplicaciones, como la espectroscopía. [Brown et al., 2000].

Método de Corrección de Dispersión Multiplicativa (CDM)

El método CDM es una técnica ampliamente utilizada en el preprocesamiento de datos espectrales para eliminar los efectos de dispersión no deseados. Este se basa en dos pasos clave que implican la estimación de los coeficientes de corrección seguida del espectro registrado, como se describe en [Martens et al., 1983]. El CDM no es susceptible a datos atípicos.

Para aplicar CDM, es necesario disponer de un espectro de referencia que represente una condición ideal, es decir, un espectro sin efectos de dispersión. Para lograr esta condición, se puede promediar el espectro para reducir los efectos de dispersión causados por diferencias en el tamaño de partículas y la longitud de trayectoria. La corrección CDM se realiza ajustando el espectro medido utilizando una regresión lineal con el espectro de referencia, y esta corrección se logra mediante la pendiente y la intersección de la línea de regresión. El método original propuesto por Martens y Naes [Martens et al., 1983] sugería usar un espectro más pequeño sin línea de base o información química como referencia. Sin embargo, encontrarlo a menudo resulta difícil en la práctica. Por lo tanto, una alternativa común y preferible es utilizar el promedio como espectro de referencia. Esto permite la corrección eficaz de los efectos de dispersión no deseados y mejora la calidad de los datos espectrales. El proceso de corrección CDM se puede expresar matemáticamente como sigue:

La regresión lineal la referencia (X_{ref}) y el espectro medido (X) se define mediante la siguiente ecuación:

$$X = aX_{ref} + b, \tag{4-5}$$

donde,
X representa el espectro medido,
X_{ref} es el espectro de referencia,
a es la pendiente de la regresión,
b es el intercepto de la regresión.

El método CDM es eficaz para eliminar los efectos de dispersión no deseados y mejorar la calidad de los datos espectrales, lo que facilita un análisis más preciso y fiable [Martens et al., 1983].

Mínimos Cuadrados Asimétricos (MCAS)

El método MCAS se en el procesamiento de datos espectrales para abordar los problemas relacionados con la línea de base y las señales superpuestas. Su principal objetivo es estimar una curva a la línea de base, sin considerar las desviaciones por encima de esta curva o asignándoles un peso mínimo. Esto se logra mediante el uso de funciones de ponderación asimétrica, lo que permite la eliminación eficiente de la línea de base mientras se preserva la información de las señales de pico [Newey and Powell, 1987].

La estimación de la línea se basa en el principio de mínimos cuadrados, que busca minimizar la suma de los cuadrados de las desviaciones entre los valores observados y los valores ajustados por la curva de referencia. Sin embargo, a diferencia de otros métodos que tratan las desviaciones tanto por encima como por debajo de la curva, en el MCAS se da más importancia a las desviaciones negativas, es decir, las que están por debajo de la línea de base.

El método del MCAS se describe mediante la siguiente ecuación:

Dado un conjunto de datos espectrales Y con N puntos espectrales, el valor ajustado $Y_{ajustado}$ en la posición i se calcula como:

$$Y_{ajustado}[i] = \underset{Y_{ajustado}[i]}{\arg\min} \sum_{i=1}^{N} w_i (Y_{observado}[i] - Y_{ajustado}[i])^2, \tag{4-6}$$

donde,
$Y_{ajustado}[i]$ es el valor ajustado en la posición i del espectro,
$Y_{observado}[i]$ es el valor observado en la posición i del espectro,

w_i es la ponderación asignada a la desviación en la posición i.

El MCAS es especialmente útil cuando se trata de señales superpuestas en datos espectrales. Al asignar un peso mayor a las desviaciones negativas, el MCAS permite una corrección más eficiente de la línea de base, lo que a su vez conserva mejor la información de los picos de señal. Esto lo convierte en una herramienta valiosa en aplicaciones que requieren un procesamiento preciso de datos espectrales, como la espectroscopía.

4.2. Métodos de selección de características

La etapa de selección de características tiene como objetivo representar de manera significativa la señal original y comprimir los datos sin perder información relevante. Esta reducción del número de datos o variables de entrada es esencial para garantizar una etapa de clasificación efectiva [Bishop et al., 1995].

En el proceso de selección de características, se deben considerar dos aspectos fundamentales: los criterios de selección y los métodos de búsqueda. Los algoritmos de selección de subconjuntos seleccionados se basan en el filtrado, lo que significa que la elección de características no depende del clasificador utilizado. En otras palabras, este procedimiento evalúa los atributos según heurísticas basadas en características generales de los datos de manera independiente de la función de evaluación (es decir, el método de clasificación) que se utilizará más adelante [Roa Martínez and Loaiza Correa, 2011]. Este enfoque de filtrado es eficaz para reducir la dimensionalidad de los datos y resaltar las características más relevantes sin depender de un clasificador específico.

4.2.1. RELIEF

El algoritmo RELIEF se fundamenta en la idea de que una característica es considerada altamente relevante si es capaz de distinguir con facilidad entre dos instancias de diferentes clases. Se basa en esta lógica para asignar un peso a cada característica [Sun and Wu, 2008].

En su forma original, el algoritmo RELIEF se limita a problemas con únicamente dos clases. Por lo tanto, para determinar la relevancia de las características, se basa en la información proporcionada por el vecino más cercano de una clase opuesta, ya que solo trabaja con una clase en oposición. Sin embargo, esta restricción planteó la necesidad de ampliar el ámbito de aplicación del algoritmo [Urbanowicz et al., 2018].

En 1994, Kononenko propuso una versión extendida del algoritmo RELIEF denominada RELIEF-F [Kononenko, 1994]. RELIEF-F generaliza el comportamiento del algoritmo ori-

ginal y se aplica a problemas con más de dos clases. En esta versión, se busca un vecino más cercano por cada clase opuesta, lo que amplía significativamente su aplicabilidad.

El proceso de asignación de pesos en RELIEF-F se basa en la siguiente ecuación:

$$W_i = W_j - \frac{1}{k}\sum_{i=1}^{k}\left|x_j^A - x_j^B\right| - \frac{1}{k}\left|x_j^A - x_j^N\right|, \tag{4-7}$$

donde,

W_j es el peso de la característica j,
x_j^A es el valor de la característica j en la instancia actual,
x_j^B es el valor de la característica j en el vecino más cercano de la misma clase,
x_j^N es el valor de la característica j en el vecino más cercano de una clase opuesta,
k representa el número de vecinos cercanos considerados en el cálculo.

Esta ecuación refleja cómo se actualiza la valoración acumulada de las características en un vector de pesos. Con RELIEF-F, es posible evaluar la relevancia de las características en problemas de clasificación con más de dos clases, lo que lo convierte en una herramienta valiosa en la selección de características. [Kira and Rendell, 1992, Kononenko, 1994]

4.3. Métodos de Clasificación

En esta sección se exponen las estrategias de análisis de datos espectroscópicos aplicado a la clasificación de tratamientos en un contexto científico e industrial. Los datos espectroscópicos ofrecen una ventana única hacia la comprensión y clasificación de muestras en diversas áreas, desde la industria alimentaria y farmacéutica hasta la agricultura y la investigación médica. En este contexto, los métodos de análisis de clasificación supervisada juegan un papel fundamental [Monroy and Rivera, 2012].

4.3.1. Conjuntos de prueba y entrenamiento

Por lo general, el conjunto de observaciones se divide en dos grupos: la muestra de entrenamiento (L_1) y la muestra de prueba (L_2). Con la muestra de entrenamiento, se estima el modelo, denominado regla de asignación o de clasificación, y con la muestra de prueba se valida el modelo. Para la validación este mismo, se construye una matriz de clasificación o de confusión y se calcula el porcentaje de mala clasificación. Estos se han consolidado como herramientas esenciales en el campo de la quimiometría para lograr una clasificación precisa

de las muestras. En este caso, es esencial realizar esta división pues al estimar la precisión (accuracy) sobre la totalidad de los datos se incurre en un sesgo [Monroy and Rivera, 2012].

Conjunto de entrenamiento (L_1): Este subconjunto se utiliza para construir y entrenar el modelo de clasificación. En otras palabras, el modelo aprende de estos datos. En el contexto de clasificación, el conjunto de entrenamiento se compone de observaciones a las cuales se les ha asignado una etiqueta o categoría conocida. El modelo utiliza estas etiquetas para aprender a hacer predicciones.

Conjunto de prueba (L_2): Este subconjunto se reserva para evaluar el rendimiento del modelo. Las observaciones en este conjunto no se utilizan durante el entrenamiento, por lo que son datos nuevos o no vistos para el modelo. El objetivo es evaluar qué tan bien el modelo puede clasificar estas observaciones. Para ello, se aplican las reglas de clasificación aprendidas en el conjunto de entrenamiento a las observaciones del conjunto de prueba y se comparan las predicciones con las etiquetas reales.

La forma en que se realiza la división puede variar, pero generalmente se elige una proporción adecuada de datos para el entrenamiento y la prueba. Por ejemplo, es común utilizar el 75 % de los datos para entrenamiento y el 25 % para prueba. El rendimiento del modelo se evalúa en el conjunto de prueba através de la matriz de confusión o del porcentaje de mala clasificación. El objetivo principal de esta división es asegurarse de que el modelo pueda generalizar y realizar predicciones en datos que no ha visto antes. La validación del modelo es crucial para evaluar su rendimiento y su capacidad para clasificar nuevas observaciones con precisión.

4.3.2. Análisis Lineal Discriminante (ALD)

El ALD maneja fácilmente el caso donde las frecuencias dentro de la clase son desiguales y sus actuaciones se han examinado al azar en datos de prueba generados. Este método maximiza la relación de varianza entre clase y la varianza dentro en cualquier conjunto de datos en particular, lo que garantiza la separabilidad máxima. el ALD no cambia la ubicación, pero solo trata de proporcionar más separabilidad de clases y dibujar una región de decisión entre las clases dadas, esta región está separada por una función lineal [Balakrishnama and Ganapathiraju, 1998].

Este análisis supone que las matrices de varianzas y covarianzas de los grupos a clasificar son iguales. El estadístico principal que usa este método es la T^2 de Hotelling [Monroy and Rivera, 2012], esto implica que hay un supuesto de normalidad multivariada en cada uno de los grupos sobre las longitudes de onda.

Si se van a clasificar k grupos y se supone matrices de covarianzas iguales:

$$S = \frac{1}{n-k}\sum_{k=1}^{k}(n_k - 1)S_k, \tag{4-8}$$

Los puntajes discriminantes con los cuales se va a tomar la decisión del grupo a asignar son:

$$W_{jk} = x^{'}S^{-1}(\overline{x}_j - \overline{x}_k)^{'}S^{-1}(\overline{x}_j - \overline{x}_k,) \tag{4-9}$$

Se asigna x a la población j si $W_jk > 0$ para todo $j \neq k$,
Esta asignación también se puede hacer mediante la distancia de Mahalanobis:

$$D_k^2(x) = (x - \overline{x}_k)^{'}S^{-1}(x - \overline{x}_k), \tag{4-10}$$

La regla de decisión se asigna a la población k si:

$$D_k^2(x) = min_k \left\{D_1^2(x), ..., D_K^2(x)\right\}, \tag{4-11}$$

La realización de las pruebas de normalidad multivariada y de homogeneidad de varianzas deben ser evaluadas antes de aplicar el ALD para asegurarse de que las suposiciones subyacentes sean válidas. En caso de que se rechacen estas hipótesis, pueden ser necesarios otros enfoques de clasificación que no dependan de dichos supuestos. En resumen, los tests Henze-Zirkler y Box's M-test son herramientas cruciales para garantizar la validez y la eficacia del ALD en la clasificación de datos [Monroy and Rivera, 2012].

Prueba de normalidad multivariada

La prueba de normalidad multivariada Henze-Zirkler es una herramienta estadística utilizada para evaluar si un conjunto de datos multivariados sigue una distribución de probabilidad normal multivariada [Hanusz et al., 2018]. La normalidad multivariada es una suposición clave en muchos métodos estadísticos, incluido el ALD.

La prueba Henze-Zirkler se basa en la comparación de las funciones de densidad de probabilidad (fdp) empíricas de las diferentes clases o grupos en los datos multivariados. La hipótesis nula de esta prueba (H_0) establece que las distribuciones de probabilidad de todas las clases son iguales, es decir, que los datos siguen una distribución normal multivariada idéntica para todas las clases.

La prueba Henze-Zirkler se realiza mediante el cálculo del estadístico:

$$H = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,i\neq j}^{n}\frac{f_i(x_j)}{\widehat{f}(x_j)}, \tag{4-12}$$

donde,
n es el número de observaciones en el conjunto de datos,
$f_i(x_j)$ es la función de densidad de probabilidad empírica de la i-ésima clase evaluada en la j-ésima observación,
$\widehat{f}(x_j)$ es la función de densidad de probabilidad empírica global de todas las observaciones.

El estadístico H se compara con una distribución de referencia (por ejemplo, una distribución chi-cuadrado) para determinar si es significativamente diferente de cero. Si el valor calculado de H es mayor que el valor crítico de la distribución de referencia, se rechaza la hipótesis nula, lo que indicaría que las distribuciones de probabilidad de las clases son significativamente diferentes y que los datos no siguen una distribución normal multivariada idéntica.

En resumen, la prueba Henze-Zirkler se utiliza para verificar si los datos multivariados siguen una distribución normal multivariada en todas las clases. Esto es crucial en el contexto del ALD.

Pruebas de homogeneidad en matrices de covarianza

La prueba de Box M se basa en la comparación de las matrices de covarianza de las diferentes clases en los datos multivariados. La hipótesis nula (H_0) establece que todas las matrices de covarianza son iguales [Box, 1953].

La prueba de Box M lleva a cabo el estadístico de prueba denominado estadístico M de Box. definido como:

$$M = \frac{(N-1)ln|E|}{2} - \frac{1}{2}\sum_{i=1}^{g}(n_i - 1)ln|S_i|, \tag{4-13}$$

donde,
N es el número total de observaciones en el conjunto de datos,
g es el número de grupos o clases,
n_i es el número de observaciones en la i-ésima clase,
$|E|$ es el determinante de la matriz de covarianza conjunta de todos los datos,
$|S_i|$ es el determinante de la matriz de covarianza de la i-ésima clase.

El estadístico M de Box sigue una distribución chi-cuadrado con grados de libertad igual a $(g1)$. Se compara el valor calculado de M con la distribución chi-cuadrado para determinar si es significativamente diferente. Si el valor calculado de M es mayor que el valor crítico de la distribución chi-cuadrado, se rechaza la hipótesis nula, lo que indica que las matrices

de covarianza de las clases son significativamente diferentes y por tanto no se cumple la homogeneidad en la varianza.

4.3.3. Análisis Cuadrático Discriminante (ACD)

El ACD es una técnica estadística utilizada en problemas de clasificación que se basa en supuestos sobre la distribución de las observaciones en diferentes clases. A diferencia del Análisis Lineal Discriminante, el ACD no asume igualdad de varianzas o matrices de covarianza iguales entre las clases, sin embargo, se basa en supuestos (i) Normalidad multivariada: Se asume que las observaciones dentro de cada clase siguen una distribución normal multivariada. Esto significa que los datos en cada clase se distribuyen en forma de una elipse multidimensional en lugar de una línea recta y (ii) Homocedasticidad: A diferencia del ALD, no se asume igualdad de varianzas (homocedasticidad) entre las clases. En su lugar, se permite que las diferentes clases tengan diferentes matrices de covarianza. Cada clase tiene su propia matriz de covarianza, que se denota como Σ_i, donde i representa la clase [Anderson and Gupta, 2009].

La función de discriminante cuadrática para asignar una nueva observación a una clase se calcula mediante la regla de Bayes. La FDC para la clase j se define como:

$$D_j(x) = -\frac{1}{2}(x-\mu_j)^T\Sigma_j^{-1}(x-\mu_j) - \frac{1}{2}ln|\Sigma_j| + lnP(Y=j), \tag{4-14}$$

donde,
$D_j(x)$ es la función discriminante cuadrática para la clase j,
x es la nueva observación,
μ_j es el vector de medias de la clase j,
Σ_j es la matriz de covarianza de la clase j,
$P(Y=j)$ es la probabilidad a priori de la clase j.

Para asignar x a una de las clases, se calculan las funciones de discriminante cuadráticas para todas las clases y se elige la clase con la mayor función de discriminante:

$$y = \arg\max_j D_j(x), \tag{4-15}$$

donde,
y es la clase asignada,
$\arg\max_j$ es la clase con la función discriminante cuadrática maxima.

4.3.4. Bosques Aleatorios (BA)

El método BA es un enfoque de clasificación y regresión que se basa en un conjunto de árboles de decisión. Está diseñado para mejorar la precisión y reducir el sobreajuste. Los siguientes pasos muestran la estrategia [Pal, 2005]:

1. Construcción de árboles de decisión: Se basa en un conjunto de árboles de decisión. Cada árbol se construye a partir de un subconjunto aleatorio del conjunto de entrenamiento.

2. Muestreo Bootstrap: Para crear estos subconjuntos, se utiliza un muestreo bootstrap. En cada árbol, se seleccionan aleatoriamente N observaciones (con reemplazo) del conjunto de entrenamiento. Esto genera una variedad de subconjuntos de entrenamiento, lo que introduce diversidad en el modelo.

3. Selección de características aleatorias: Además del muestreo bootstrap, en cada división de nodo en un árbol, solo se considera un subconjunto aleatorio de las características. Esto evita que algunas características dominen el proceso de decisión.

4. Construcción de arboles: Cada árbol se construye de acuerdo con las reglas de un árbol de decisión. Se selecciona una característica en cada nodo y se divide el nodo en dos en función de algún criterio, como la ganancia de información o la impureza de Gini.

5. Votación o promedio: Una vez que se han construido todos los árboles (por lo general, se construyen cientos o miles), se utiliza un enfoque de votación (para clasificación) o promedio (para regresión) para tomar una decisión final. En la clasificación, cada árbol vota por una clase, y la clase con más votos se elige como la predicción final. En la regresión, se promedian las predicciones de todos los árboles.

En el caso de la clasificación, se puede definir la predicción final como:

$$\widehat{Y}(x) = Moda\left\{Y_1(x), ..., Y_B(x),\right\} \tag{4-16}$$

donde,
$\widehat{Y}(x)$ es la predicción para la observación x,
$Y_B(x)$ es la predicción del árbol B,
$Moda$ representa la moda (la clase más común) entre todas las predicciones de los árboles.

El método BA es efectivo porque aprovecha la diversidad y la robustez de múltiples árboles de decisión para lograr predicciones más precisas y reducir el sobreajuste [Pal, 2005]. Cada árbol aporta su perspectiva única al modelo, y la combinación de estas perspectivas mejora significativamente la capacidad de generalización del modelo [Breiman, 2001].

4.3.5. Bayes Ingenuo (BI)

El método BI es un algoritmo de clasificación supervisada en el campo del aprendizaje automático. Se basa en el teorema de Bayes y se utiliza comúnmente en tareas de clasificación de texto y minería de datos [Berrar, 2019]. Este enfoque asume una simplificación fuerte: la independencia condicional entre las características (variables predictoras) dadas las clases, lo que le da su nombre "naive"(ingenuo).

En un problema de clasificación, se tiene un conjunto de datos de entrenamiento etiquetado que consta de características denotadas como X y una variable de clase denotada como C. El objetivo es predecir la clase C de un nuevo ejemplo basado en sus características. El teorema de Bayes es fundamental en este método y se expresa de la siguiente manera:

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)}, \tag{4-17}$$

donde,
$P(CX)$ es la probabilidad condicional de que el ejemplo pertenezca a la clase C dado un conjunto de características X,
$P(XC)$ es la probabilidad condicional de las características X dado que el ejemplo pertenece a la clase C,
$P(C)$ es la probabilidad a priori de que un ejemplo pertenezca a la clase C,
$P(X)$ es la probabilidad a priori de las características X.

El BI asume que las características son independientes dadas las clases, lo que se expresa como:

$$C_{predicho} = \arg\max_{C} P(C) \prod_{i=1}^{n} P(X_i|C), \tag{4-18}$$

La estimación de $P(C)$ y $P(X_i|C)$ se realiza utilizando un conjunto de datos de entrenamiento y técnicas como la frecuencia relativa de ocurrencia en el conjunto de entrenamiento.

El método BI asume varios supuestos fundamentales:

1. Independencia condicional de las características: Se asume que todas las características son condicionalmente indpendiente, es decir, que no existe ninguna relación entre las características una vez que se conoce la clase. Esta es una simplificación fuerte que en la mayoría de los casos reales, rara vez se cumple por completo. Sin embargo, simplifica en gran medida el cálculo de probabilidades y hace que el algoritmo sea computacionalmente eficiente.

2. Distribución de probabilidades: BI también supone una distribución específica de probabilidad para las características. Dependiendo del tipo de datos, se pueden asumir una distribución de Bernoulli (para datos binarios), la distribución de Poisson (para datos de conteo) o una Gaussiana (normal) 8para datos continuos). La elección adecuada depende del tipo de datos.

3. Independencia de los ejemplos de entrenamiento: BI asume que los ejemplos de entrenamiento son independientes. Es decir que la ocurrencia de un ejemplo de entrenamiento no afecta la ocurrencia de otro. Este supuesto es común en muchos algoritmos de aprendizaje automático, pero en la práctica, la independencia completa de los ejemplos rara vez se cumple.

4. Probabilidad a priori constante: En BI se supone que las probabilidades a priori $P(C)$ son constantes para todas las clases. Esto significa que, antes de observar las características, todas las clases son igualmente probables. Este supuesto puede ser inavlidado en situaciones donde las clases no son igualmente probables en la población subyacente.

Es importante tener en cuenta que estos supuestos hacen que el método BI sea ingenuo en el sentido de que simplifica en exceso la complejidad de los datos. No obstante, BI a menudo funciona sorprendentemente bien en la práctica, especialmente en tareas de clasificación de texto y minería de datos, donde las características suelen ser términos independientes [Learning, 1997, Bishop, 2006].

4.3.6. Máquinas de Soporte Vectorial (MSV)

Se busca encontrar un hiperplano en el espacio de características que maximice la separación entre las clases. El MSV asume que los datos son linealmente separables en el espacio de características. Para casos en los que no lo son, se puede utilizar un kernel para proyectar los datos en un espacio de características de mayor dimensión donde sean linealmente separables

[Gold and Sollich, 2003]

El objetivo es encontrar un hiperplano de separación $wx + b = 0$ que maximiza el margen entre las dos clases. El margen es la distancia perpendicular desde el hiperplano a los ejemplos de entrenamiento más cercanos de ambas clases. esto es maximizar $\frac{2}{|w|}$ sugeto a $y_i(wx_i + b) \geq 1$ para todo i. donde w es el vector de pesos del hiperplano, x_i es el vector de características de un ejemplo de entrenamiento, b es el sesgo o término de sesgo, y y_i es la etiqueta de clase del ejemplo i.

El MSV busca minimizar la función de pérdida regularizada:

$$L(w, b) = \frac{1}{2}|w|^2 - \sum_{i=1}^{n} \alpha_i[y_i(wx_i + b) - 1], \tag{4-19}$$

donde, α_i son los multiplicadores de Lagrange que se introducen para imponer las restricciones del margen.

Una vez que se ha encontrado el hiperplano óptimo, la clasificación de un nuevo ejemplo se realiza evaluando la expresión $wx + b$. Si el resultado es positivo, se clasifica en una clase, y si es negativo, se clasifica en la otra clase.

El MSV es un algoritmo de clasificación que ha demostrado ser eficaz en una variedad de aplicaciones. La elección del kernel y la configuración de hiperparámetros son aspectos importantes en su implementación exitosa [Cortes and Vapnik, 1995, Schölkopf and Smola, 2002].

4.3.7. K vecinos más cercanos (KVC)

Este es un algoritmo de aprendizaje automático que se utiliza en tareas de clasificación. Su enfoque se basa en el principio de que los objetos que son similares en el espacio de características tienden a pertenecer a la misma clase.Se parte del supuesto de que los objetos se pueden representar en un espacio de características multidimensional, donde cada objeto se describe mediante un vector de características. Se supone que objetos que son cercanos en este espacio de características tienen una alta probabilidad de pertenecer a la misma clase [Guo et al., 2003].

Cuando se recibe una nueva observación que se desea clasificar, el algoritmo calcula su distancia a todos los ejemplos de entrenamiento en el espacio de características. La distancia puede calcularse utilizando diferentes métricas, como la distancia euclidiana o de Mahalanobis. Luego, el algoritmo selecciona las k observaciones de entrenamiento más cercanos (los

vecinos) a la nueva observación, donde k es un número entero definido por el usuario. Estos vecinos son determinados por sus distancias más pequeñas a la nueva observación. Finalmente, el algoritmo asigna la clase que es más común entre los k vecinos a la nueva observación. Esta asignación se realiza utilizando una regla de mayoría, donde la clase con más vecinos se convierte en la clase predicha para la nueva observación.

Dado un conjunto de datos de entrenamiento X_i, y_i donde X_i es un vector de características y y_i es la etiqueta de clase correspondiente, la predicción de la clase para un nuevo ejemplo Y_nuevo se realiza de la siguiente manera:

$$y_{nuevo} = Moda(\{y\}_i^k), \tag{4-20}$$

donde,
k es el número de vecinos más cercanos,
$Moda$ se refiere a la clase más común entre los k vecinos más cercanos al nuevo ejemplo.

La elección de k es un hiperparámetro crítico en KVC y puede afectar significativamente el rendimiento del algoritmo. Un valor de k más pequeño puede llevar a una clasificación más ruidosa y sensible a valores atípicos, mientras que un valor de k más grande suaviza la clasificación, pero puede perder detalles importantes en la frontera de decisión.

la ventaja de este método es simple de entender e implementar. Permite analizar datos no lineales usando su método no paramétricos. y las desventajas radican en que uede ser computacionalmente costoso para conjuntos de datos grandes y sensibles a la elección de k. También puede ser sensible a características irrelevantes. KVC es un algoritmo versátil y se utiliza en diversas aplicaciones, aunque es importante considerar sus ventajas y desventajas al aplicarlo a un problema específico [Gazalba et al., 2017].

4.3.8. Perceptrón Multicapa (PM)

Es una red neuronal artificial que se utiliza para tareas de clasificación y regresión en aprendizaje automático. Consta de múltiples capas de neuronas interconectadas y se caracteriza por su capacidad para aprender representaciones no lineales de datos. Se parte del supuesto de que los datos se pueden representar en un espacio de características multidimensional, donde cada ejemplo se describe mediante un vector de características. Se asume que los datos son intrínsecamente no lineales y requieren una representación más compleja que un modelo lineal.

Un PM consta de una capa de entrada, una o más capas ocultas y una capa de salida. Cada capa contiene múltiples neuronas o unidades, y cada neurona está conectada a todas las neuronas en las capas adyacentes. Las neuronas en cada capa aplican una función

de activación no lineal a la suma ponderada de las salidas de las neuronas en la capa anterior.

La entrada se propaga hacia adelante a través de la red neuronal, capa por capa. Cada neurona en una capa oculta calcula una combinación lineal de las salidas de las neuronas en la capa anterior y aplica una función de activación no lineal, como la función sigmoide o ReLU [LeCun et al., 1989]. La capa de salida produce una salida que se utiliza para la clasificación o regresión.

La salida de una neurona en una capa oculta se calcula de la siguiente manera:

$$a_j = f \sum_{i=1}^{n} w_{ij} x_i + b_j, \tag{4-21}$$

donde,
a_j es la salida de la neurona j,
$f(.)$ es la función de activación,
w_{ij} son los pesos entre la neurona i en la capa anterior y la neurona j en la capa actual,
x_i es la salida de la neurona i en la capa anterior,
b_j es el sesgo de la neurona j.

La salida de la capa de salida se calcula de manera similar. En tareas de clasificación, se utiliza una función de activación como la sigmoide para producir probabilidades de pertenencia a clases. En tareas de regresión, la capa de salida puede producir una estimación numérica.

El entrenamiento del PM implica la propagación hacia atrás (backpropagation) para ajustar los pesos de las conexiones. Se utiliza una función de pérdida, como el error cuadrático medio, para calcular el error entre las predicciones y las etiquetas de entrenamiento.Luego, se aplican algoritmos de optimización, como el descenso de gradiente, para minimizar la función de pérdida y ajustar los pesos de la red.

Para evitar el sobreajuste, se pueden aplicar técnicas de regularización, como la penalización L1 o L2 en los pesos. El PM es una herramienta poderosa en aprendizaje automático, capaz de modelar relaciones complejas en los datos. Su estructura y capacidad de representación lo hacen adecuado para una amplia gama de aplicaciones, pero también requiere una cantidad significativa de datos de entrenamiento y ajuste de hiperparámetros para un rendimiento óptimo [Bishop, 2006].

4.4. Métricas de Clasificación

Desempeñan un papel crítico en la evaluación de modelos de clasificación y proporcionan información valiosa sobre su capacidad para tomar decisiones en función de las observaciones y grupos.

Cuando se entrena un modelo de clasificación supervisada, su objetivo principal es aprender a predecir las etiquetas de clase de nuevos ejemplos en función de las características de entrada. Para evaluar qué tan bien cumple esta tarea, se requiere un análisis objetivo y cuantitativo. Aquí es donde entran en juego las métricas de clasificación.

El núcleo del modelo radica en comparar las predicciones generadas por el modelo con las etiquetas verdaderas en un conjunto de datos de prueba. Estos generalmente se mantienen separados de los datos utilizados para entrenar el modelo y se utilizan para simular situaciones del mundo real [Müller and Guido, 2016].

Las métricas de clasificación supervisada ofrecen diversas formas de medir el rendimiento del modelo y cada una de ellas proporciona una perspectiva diferente. Por ejemplo, la precisión mide la proporción de predicciones correctas en general, lo que es útil para obtener una visión general del rendimiento. La exactitud y la recuperación se centran en la calidad de las predicciones de una clase específica, lo que es crucial en aplicaciones donde los errores pueden ser costosos.

Las métricas también permiten explorar el equilibrio y el compromiso entre diferentes aspectos del rendimiento del modelo. Por ejemplo, el F1-Score combina precisión y recuperación [Lipton et al., 2014], lo que es útil cuando se necesita un equilibrio entre la minimización de falsos positivos y falsos negativos. El valor F permite ajustar el equilibrio mediante el parámetro β.

Es importante destacar que el rendimiento de un modelo puede variar según la métrica utilizada y el contexto. Algunas aplicaciones pueden requerir un alto énfasis en la minimización de falsos positivos, mientras que otras pueden priorizar la minimización de falsos negativos. La elección de la métrica adecuada depende de los objetivos y restricciones específicos del problema [Murphy, 2012].

Las métricas de clasificación supervisada permiten una evaluación cuantitativa y objetiva del rendimiento de los modelos de clasificación, ayudando a tomar decisiones informadas sobre la idoneidad de un modelo para una tarea dada. Estas métricas son esenciales para la evaluación y comparación de algoritmos y para ajustar modelos apropiados.

Matriz de Confusión

Esta matriz de confusión es una tabla que se utiliza para resumir el rendimiento de un modelo al comparar sus predicciones con las etiquetas verdaderas en un conjunto de datos de prueba. Consta de cuatro componentes principales [Visa et al., 2011]:

- Verdaderos Positivos: Representa el número de instancias positivas que el modelo ha clasificado correctamente como positivas. En otras palabras, son las predicciones positivas que son realmente correctas de acuerdo con las etiquetas verdaderas.
- Falsos Positivos: Estos son los casos en los que el modelo ha clasificado incorrectamente instancias negativas como positivas. Es decir, el modelo ha predicho que algo es positivo cuando en realidad no lo es.
- Verdaderos Negativos: Representa el número de instancias negativas que el modelo ha clasificado correctamente como negativas. Son las predicciones negativas que son verdaderas según las etiquetas verdaderas.
- Falsos Negativos: En este caso, el modelo ha clasificado erróneamente instancias positivas como negativas. Esto significa que el modelo ha predicho incorrectamente que algo es negativo cuando en realidad es positivo.

La matriz de confusión permite evaluar varias métricas de clasificación, como la precisión, la recuperación, la especificidad, el F1-Score y otras métricas, al calcular diferentes combinaciones de los valores VP, FP, VN y FN. Estas métricas proporcionan una comprensión detallada del rendimiento del modelo y su capacidad para discriminar entre las clases de interés.

La matriz de confusión es una herramienta valiosa para la evaluación de modelos, especialmente en problemas de clasificación donde el desequilibrio de clases o la importancia relativa de los errores pueden variar. Permite una evaluación cuantitativa y un análisis detallado del rendimiento del modelo en función de los objetivos específicos del problema.

Exactitud

La métrica de exactitud o accuracy es una de las métricas de clasificación más fundamentales y se utiliza para evaluar el rendimiento global de un modelo de clasificación en un conjunto de datos de prueba. Esta métrica se refiere a la proporción de predicciones correctas que el modelo ha realizado en comparación con el número total de predicciones. En otras palabras,

mide la capacidad del modelo para clasificar correctamente tanto las instancias positivas como las negativas. La fórmula para calcular la exactitud es la siguiente [Visa et al., 2011]:

$$Exactitud = \frac{No.\,de\,Predicciones\,Correctas}{No.\,de\,Predicciones\,Totales}, \tag{4-22}$$

donde,

Número de predicciones correctas: es la suma de verdaderos positivos (VP) y verdaderos negativos (VN). Representa todas las instancias que el modelo ha clasificado correctamente, ya sean positivas o negativas. Número total de predicciones: es la suma de verdaderos positivos (VP), falsos positivos (FP), verdaderos negativos (VN) y falsos negativos (FN). Representa todas las predicciones realizadas por el modelo, sin importar si son correctas o incorrectas. La precisión es una métrica útil para evaluar el rendimiento general de un modelo de clasificación. Proporciona una visión general de cuántas de las predicciones del modelo son correctas en comparación con todas las predicciones realizadas. Sin embargo, la precisión puede ser engañosa en situaciones donde las clases están desequilibradas. En tales casos, un modelo que predice siempre la clase mayoritaria puede tener una alta precisión, pero no necesariamente ser útil.

5 Metodología

5.1. Metodología

En este capítulo se expone la población y muestra utilizada en la investigación. Acto seguido, se describen las metodologias implementadas como la de espectroscopía, preprocesamiento, selecciones de longitudes de onda y, finalmente, la estadística.

5.2. Población y muestra

Se aplicó un diseño experimental aleatorizado para investigar el impacto de ocho tratamientos diferentes en un conjunto de 240 plantas (30 plantas asignadas a cada uno de los tratamientos). Todas las plantas se mantuvieron bajo condiciones ambientales uniformes y controladas. El proceso de diseño se desarrolló de la siguiente forma:

Primero se seleccionaron aleatoriamente las 240 plantas de la población de estudio, asignándoles identificadores y garantizando que la selección fuera completamente aleatoria.

Luego, se asignaron aleatoriamente las 30 plantas seleccionadas a cada uno de los ocho tratamientos disponibles. Estos tratamientos incluían un grupo de control, así como tratamientos relacionados con la infección de Ralstonia solanacearum, Fusarium oxysporum, estrés hídrico y sus interacciones. La asignación aleatoria se empleó para asegurar que las plantas fueran comparables y las diferencias observadas se debieran exclusivamente a los tratamientos.

A lo largo del experimento, se registraron y analizaron los resultados y observaciones correspondientes a las 30 plantas en cada uno de los ocho tratamientos. Este diseño totalmente aleatorizado garantiza la validez de las comparaciones entre tratamientos y proporciona resultados confiables y científicamente sólidos para la investigación.

Se emplearon plantas de banano de la variedad Gros Michel que fueron cultivadas in vitro y adquiridas en un vivero comercial. La investigación se llevó a cabo en el invernadero de AUGURA, ubicado a 4 km de Carepa, Antioquia (Colombia), con una altitud aproximada de 30 metros sobre el nivel del mar. A las plantas se les suministró un fertilizante edáfico

cada 8 días, en una dosis de 4 gramos por litro, hasta el momento de la medición el día anterior a la inoculación (día 0). Estas fueron sembradas en bolsas de vivero de 2 kg, con una proporción de 1:1 de cascarilla de arroz y suelo. Las condiciones climáticas del invernadero fueron las siguientes: temperatura mínima de alrededor de 22°C, temperatura máxima de 35°C, temperatura promedio de 28°C y una humedad relativa oscilante entre el 85 % y el 90 %.

5.2.1. Inoculación

La cepa Foc R1 Varonesa fue cultivada en un medio de cultivo de Agar Papa Dextrosa (PDA). Después de 7 días de crecimiento, la superficie de las colonias se lavó con agua destilada estéril para obtener una suspensión con una concentración de $1x10^6$ conidias/ml. Para llevar a cabo la infección, se empleó la metodología propuesta por [Jie et al., 2009]: cuando las plantas alcanzaron el estadio de desarrollo 1030 según la escala BBCH, se practicó un corte en la base de la raíz de cada planta con un bisturí, seguido de la inoculación inmediata con 15 ml de la suspensión de conidias.[Macias-Echeverri et al., 2022]

5.3. Metodología de espectroscopía

Se obtuvieron los espectros de reflectancia mediante el uso de un espectrofotómetro portátil ASD FieldSpec 4 Hi-Res NG que utiliza resolución espectral AVRIS-NG y HySpex ODIN-1024, este incorpora detectores SWIR de fotodiodo InGaAs de índice graduado para proporcionar el intervalo de muestreo espectral de 1,875 longitudes de onda medidas en todo el rango espectral de 350 a 2500 nm. Las mediciones se realizaron cuando las plantas tenían 5 hojas funcionales, en la trecera hoja, con la fibra óptica en la cara adaxial obteniéndose tres espectros o pseudorepeticiones por cada hoja.

5.4. Metodología de preprocesamiento

En este caso se eliminan los atípicos encontrados con los métodos de la bolsa y RAD a los métodos de suavizamiento. A los que no son suceptibles a datos atípicos no se eliminan, pues el mismo método de suavizamiento lo hace indirectamente. En el caso de los datos crudos de hacen 3 grupos de métodos para comparar la precisión (accuracy), el primero se dejan los datos crudos, el segundo eliminando por el MB y el tercero eliminando atípicos con el método RAD.
Grupos de comparación:

- Datos Crudos
- Datos crudos eliminando atípicos por MB

- Datos crudos eliminando atípicos por RAD
- Método MB: se eliminan atípicos por MB.
- Método RAD: se eliminan atípicos por RAD
- Método Savitzky-Golay MB
- Método Savitzky-Golay RAD
- Método Savitzky-Golay de segundo orden derivativo MB
- Método Savitzky-Golay de segundo orden derivativo RAD
- Método Multiplicative MB
- Método Multiplicative RAD
- Método Asymmetric MB
- Método Asymmetric RAD

5.5. Metodología de selección de longitudes de onda

En el contexto de las metodologías previamente mencionadas, se lleva a cabo la identificación de longitudes de onda discriminativas mediante la aplicación de la metodología RELIEF. Este proceso se realiza de manera sistemática, dividiendo el espectro en segmentos de 100 nm, con el objetivo de evaluar y seleccionar las longitudes de onda más influyentes en cada tramo. Una vez identificadas las longitudes de onda significativas en cada segmento, se procede a seleccionar las 10 más relevantes entre ellas. Este procedimiento se repite para cada uno de los 13 grupos de metodologías consideradas en el estudio.

Posteriormente, se realiza una integración de todas las longitudes de onda seleccionadas en los diferentes grupos de metodologías. Esta integración tiene como propósito identificar las 10 longitudes de onda más importantes en los datos espectroscópicos crudos. Este enfoque de selección estratificada por tramos y consolidación de resultados permite abordar la complejidad inherente a la variabilidad espectral y a la identificación de las longitudes de onda más informativas en la clasificación de tratamientos en plantas de banano Gros Michel.

Por último, para la variedad en las longitudes de onda seleccionadas se realiza un procedimiento manual para elegir las longitudes de onda más relevantes de tal manera que no haya longitudes de onda adyacentes en un rango de 10nm.

5.6. Metodología estadística

La métrica central que guía la evaluación y selección de combinaciones metodológicas es la precisión, se usa para discernir qué combinaciones de metodologías ofrecen los resultados más confiables y precisos.

En el caso específico del Análisis Lineal Discriminante, se realiza una exhaustiva evaluación de supuesto de normalidad multivariada y la homogeneidad en la varianza para cada grupo. Este análisis es crucial para asegurar la validez de las inferencias.

En cuanto a los Bosques Aleatorios, se implementa una estrategia que implica la construcción de una grilla de 1 a 30 para evaluar diferentes números de árboles, comparándolos mediante la métrica de precisión (accuracy) para seleccionar la configuración óptima.

En el método de los KVC, se realiza una exploración sistemática de diferentes valores de k, desde 2 hasta 30, evaluando el rendimiento de cada configuración en términos de precisión y seleccionando el k que maximiza este criterio.

Para la red neuronal se establece una semilla para garantizar la reproducibilidad de los resultados. Se procede entrenar varios modelos con diferentes arquitecturas y funciones de activación para explorar las posibles configuraciones. Esto incluye modelos con distintas cantidades de capas y neuronas, así como la prueba de diversas funciones de activación. Este enfoque permite determinar la combinación más efectiva de capas y funciones de activación que maximiza la precisión en la clasificación de los tratamientos de las plantas de banano Gros Michel.

6 Resultados

En este estudio centrado en la detección temprana de estrés en plantas de Banano Gros Michel del municipio de Carepa, ubicado en el departamento de Antioquia, Colombia. se analizaron datos de reflectancia VIS/NIR de 240 plantas en Colombia. Estos incluyeron condiciones de salud, EH y la presencia de patógenos como FOCR1 y RSR2.

La descripción detallada de los espectros reveló patrones distintivos para cada condición, proporcionando una base visual para entender las variaciones en la reflectancia asociadas al estrés biótico y abiótico. Se aplicaron técnicas de preprocesamiento para abordar atípicos, garantizando la calidad de los datos.

La selección de características usando RELIEF permitió identificar longitudes de onda clave para discriminar entre tratamientos, optimizando la eficiencia de los modelos de clasificación supervisada.

La integración de resultados en los días 3 y 6 post-inoculación proporcionó una perspectiva temporal esencial. Se examinaron tendencias en los patrones espectrales, mejorando la comprensión de la dinámica de respuestas ante patógenos y estrés hídrico. Este enfoque contribuye a estrategias mejoradas de detección temprana en la agricultura.

6.1. Descripción de los espectros

Se presenta una descripción detallada de los espectros obtenidos mediante el espectrómetro portátil ASD FieldSpec. Se exploran las características espectrales en el rango de 350-2500 nm, revelando patrones distintivos asociados a plantas sanas, sometidas a EH, infectadas con FOCR1 y con RSR2, así como sus interacciones. Se destacan las variaciones en las curvas de reflectancia, proporcionando una base visual para comprender las diferencias espectrales entre los distintos estados de las plantas de banano tanto para el día 3 como el día 6 de medición.

6.1.1. Descripción de los espectros en el tercer día

La representación gráfica **6-1** se configura mediante la recopilación de promedios de reflectancias de las plantas en cada grupo experimental. Esta construcción se fundamenta en la síntesis cuidadosa de los datos espectrales, promediando las respuestas ópticas de las plantas dentro de cada condición experimental. Este enfoque proporciona una visión consolidada de las características espectrales distintivas que definen cada grupo, permitiendo una comparación visual clara y precisa entre las diferentes condiciones evaluadas.

En el espectro de las plantas analizadas en general, se observa una reflectancia aproximada del 5 % en el rango de 350 nm a 700 nm. Sin embargo, se destaca un pico significativo del 9 % aproximadamente a los 550 nm.

La reflectancia más baja en el rango inicial puede estar vinculada a la absorción de luz por pigmentos como la clorofila, que tiende a mostrar mínima reflectancia en ciertas longitudes de onda debido al proceso de fotosíntesis. No obstante, el pico del 9 % a los 550 nm sugiere una respuesta específica de la planta en esa longitud de onda. Este pico elevado podría indicar la presencia de pigmentos específicos, como la clorofila, que reflejan más luz en esa longitud de onda particular. También podría asociarse con fenómenos como la dispersión de la luz o características estructurales de las células vegetales en esa región del espectro. Este aumento en la reflectancia a los 550 nm podría tener implicaciones fisiológicas y bioquímicas importantes. Por ejemplo, podría señalar una fase específica del ciclo de crecimiento de la planta o indicar su capacidad para absorber eficientemente la luz en esa longitud de onda para la fotosíntesis. También hay que tener en cuenta que el color verde se evidencia entre los 497 nm y 570nm.

En el contexto de la detección temprana de estrés en las plantas de Banano Gros Michel, este tipo de análisis espectral detallado puede ser clave para identificar patrones que denotan condiciones particulares, como la presencia de patógenos o el impacto del estrés ambiental. Este enfoque proporciona información valiosa para comprender la salud y el estado fisiológico de las plantas.

Por otra parte, se destaca la observación de que el espectro vinculado a las plantas afectadas por Ralstonia exhibe porcentajes de reflectancia significativamente inferiores en comparación con los demás grupos experimentales en la mayor parte del espectro. Esta discrepancia en los niveles de reflectancia revela una respuesta espectral distintiva en las plantas que han sido afectadas por RSR2 en comparación con los otros tratamientos evaluados. Este fenómeno sugiere la existencia de cambios específicos en las propiedades ópticas de las plantas afectadas por la presencia de RSR2, proporcionando así un indicio valioso para la identificación y caracterización del estrés asociado con esta condición experimental.

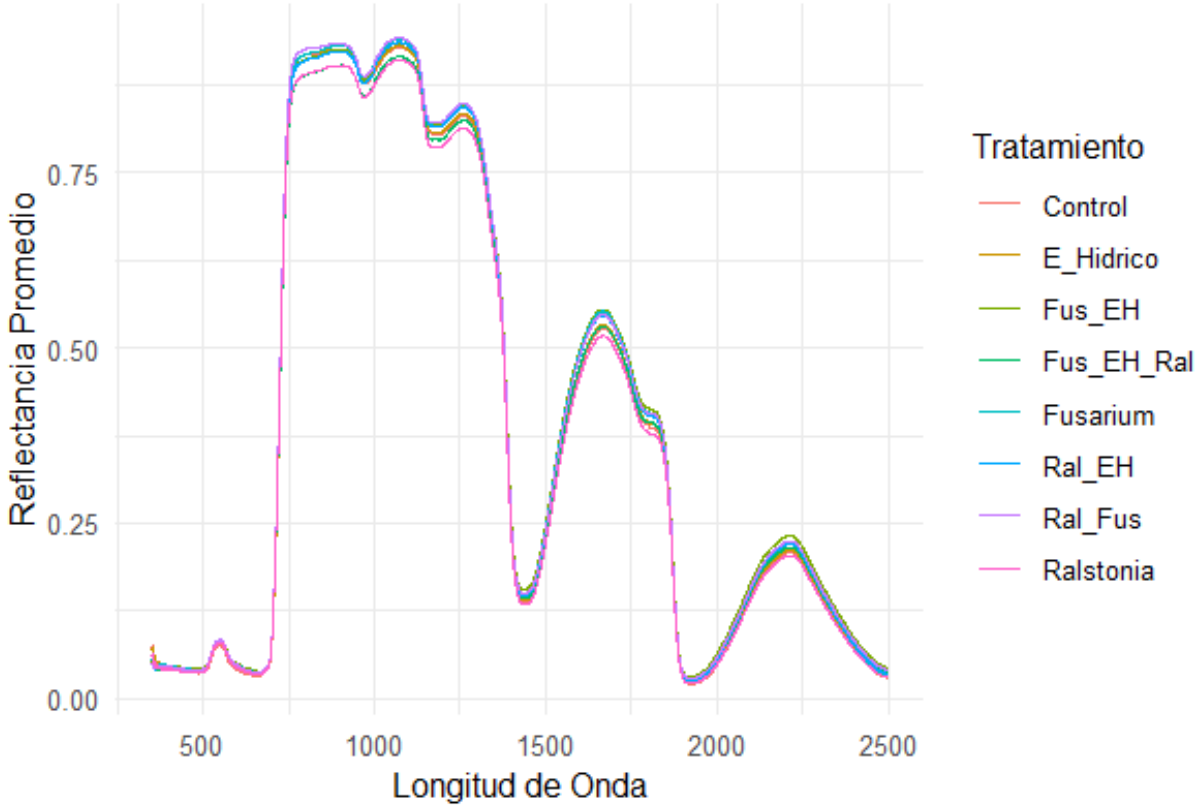

Figura 6-1: Promedios de los espectros de reflactancia según longitud de onda por tratamiento para el día 3.

6.1.2. Descripción de los espectros en el sexto día

En relación al sexto día de medición de reflectancia, se observa en la gráfica **6-2** que, en términos generales, los espectros de reflectancia de todas las plantas siguen una tendencia muy similar a la descrita en el día 3. No obstante, se evidencian variaciones notables al diferenciar los espectros según los distintos grupos experimentales. Este fenómeno se refleja particularmente en las plantas afectadas por Fusarium, donde la mayoría de las longitudes de onda, en el rango de 350 nm a 1000 nm, muestran valores de reflectancia inferiores en comparación con los demás grupos experimentales. Por otro lado, en el intervalo de 1000 nm a 1500 nm, se observa que la reflectancia de estas plantas se sitúa por encima de los demás grupos experimentales.

Estas diferencias en los patrones espectrales sugieren respuestas específicas en las plantas infectadas con Fusarium al sexto día post-inoculación, lo cual podría indicar cambios bioquímicos y estructurales en las hojas que son distintivos de la infección. Este análisis deta-

llado de los espectros a lo largo del tiempo proporciona una comprensión más profunda de la evolución de las respuestas espectrales y contribuye a la identificación precisa de patrones asociados a condiciones de estrés biótico en las plantas de Banano Gros Michel.

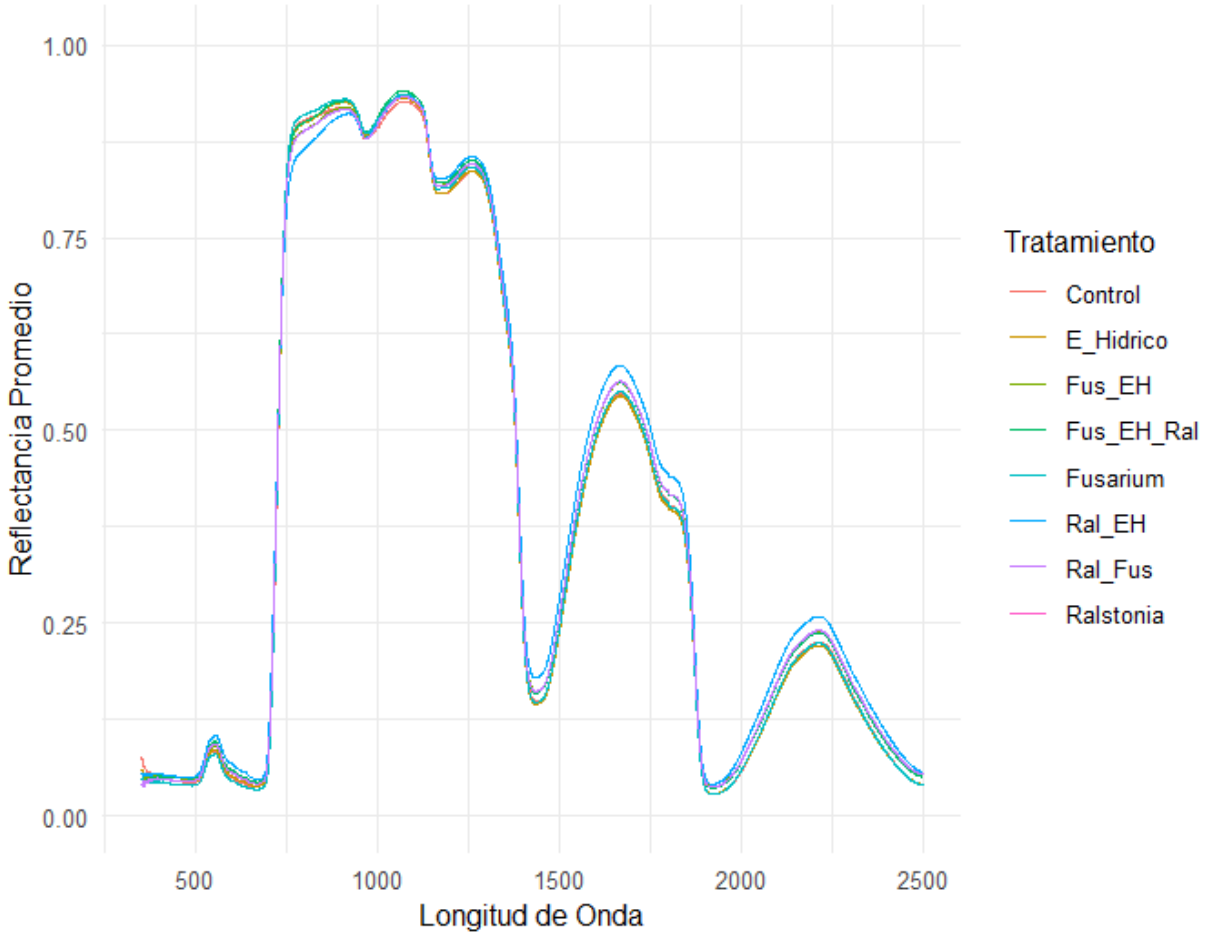

Figura 6-2: Promedios de los espectros de reflactancia según longitud de onda por tratamiento para el día 6.

6.2. Detección de atípicos

Los resultados presentados a continuación se basaron en los datos sin aplicar ningún método de suavizamiento. No obstante, es esencial destacar que el segundo subcapítulo se enfoca específicamente en la identificación y manejo de datos atípicos. A través de técnicas de preprocesamiento, se aborda de manera detallada la presencia de atípicos en los datos de reflectancia. La aplicación de métodos, como el tratamiento de datos atípicos y el suavizamiento, contribuye significativamente a la mejora de la calidad de los datos. Esta acción

asegura una representación más precisa de la realidad subyacente en los datos espectrales, facilitando así la interpretación de los resultados obtenidos en el análisis posterior. La inclusión de estos métodos de preprocesamiento refleja un enfoque riguroso para garantizar la robustez y la fiabilidad de los resultados presentados en este estudio.

6.2.1. Detección de atípicos en el tercer día

Método de la Bolsa en el tercer día

En el tercer día de medición, mediante la aplicación del MB para la detección de atípicos en datos funcionales, se revelaron resultados distintivos para cada grupo experimental. El grupo control exhibió 2 observaciones atípicas, mientras que el grupo sometido a EH no presentó atípicos. Por otro lado, el grupo inoculado con el hongo Fusarium mostró 2 atípicos, y el grupo inoculado con la bacteria Ralstonia destacó con 3 observaciones atípicas.

En los grupos que experimentaron combinaciones de condiciones, los resultados también fueron notables. El grupo sometido a FOCR1 y EH presentó 2 atípicos, al igual que el grupo sometido a Fusarium y Ralstonia. Asimismo, el grupo sometido a RSR2 y EH también exhibió 2 observaciones atípicas. Finalmente, el grupo que enfrentó EH, FOCR1 y RSR2 mostró la presencia de 3 atípicos. (ver gráfica **6-3**)

Estos hallazgos resaltan la capacidad del MB para identificar observaciones atípicas en cada grupo experimental, proporcionando información esencial sobre la variabilidad y posibles anomalías en los datos de reflectancia en el día 3 post-inoculación. Este análisis de atípicos contribuye significativamente a la comprensión de la robustez y consistencia de los conjuntos de datos evaluados en condiciones específicas.

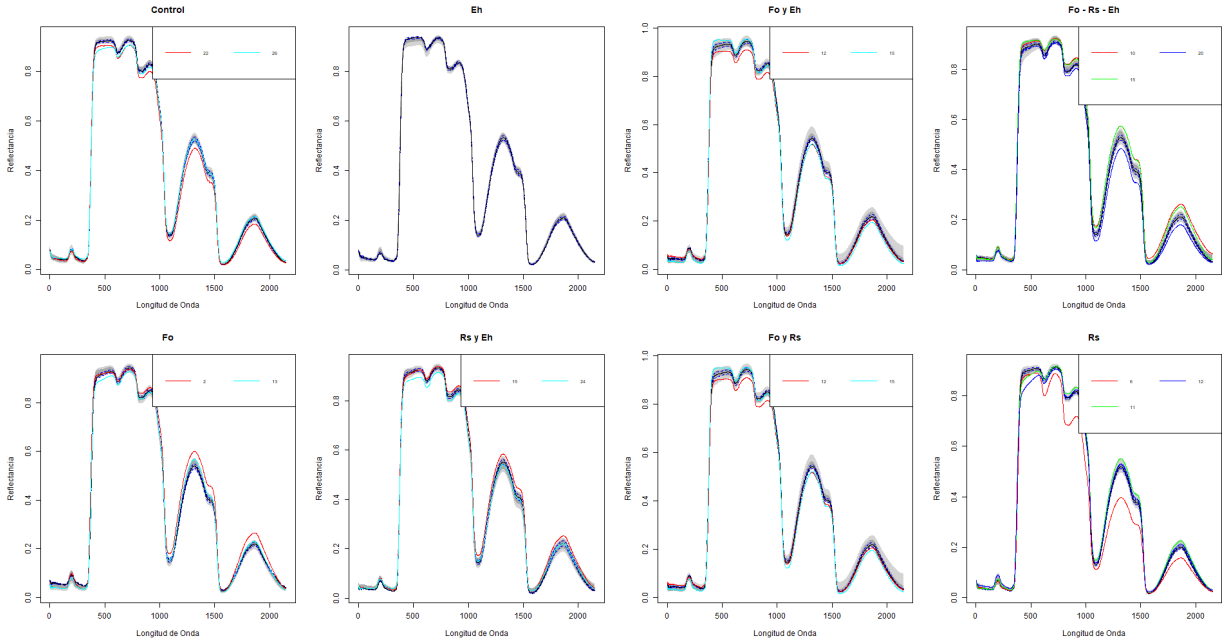

Figura 6-3: Detección de atípicos por tratamiento con el MB para el día 3.

Método RAD en el tercer día

En el tercer día de medición, al aplicar el método RAD para la detección de atípicos en datos funcionales, se observaron patrones distintivos en cada grupo experimental. En el grupo control, se identificaron 2 observaciones atípicas, mientras que el grupo sometido a EH mostró 2 atípicos. En paralelo, tanto el grupo inoculado con el hongo Fusarium como el grupo inoculado con la bacteria Ralstonia exhibieron 2 observaciones atípicas.

En los grupos que experimentaron combinaciones de condiciones, los resultados fueron consistentes. Tanto el grupo sometido a FOCR1 y EH como el grupo sometido a FOCR1 y Ralstonia presentaron 2 atípicos. De manera similar, el grupo sometido a RSR2 y estrés hídrico también mostró 2 observaciones atípicas. Finalmente, el grupo que enfrentó estrés hídrico, Fusarium y RSR2 presentó 2 atípicos.

Estos resultados destacan la capacidad del método RAD para identificar de manera eficaz y uniforme observaciones atípicas en diferentes grupos experimentales. La consistencia en la detección de atípicos sugiere la robustez de este enfoque en la evaluación de datos de reflectancia en el día 3 post-inoculación. Estos hallazgos fortalecen la confianza en la integridad de los conjuntos de datos analizados y su capacidad para revelar patrones significativos asociados a condiciones específicas.

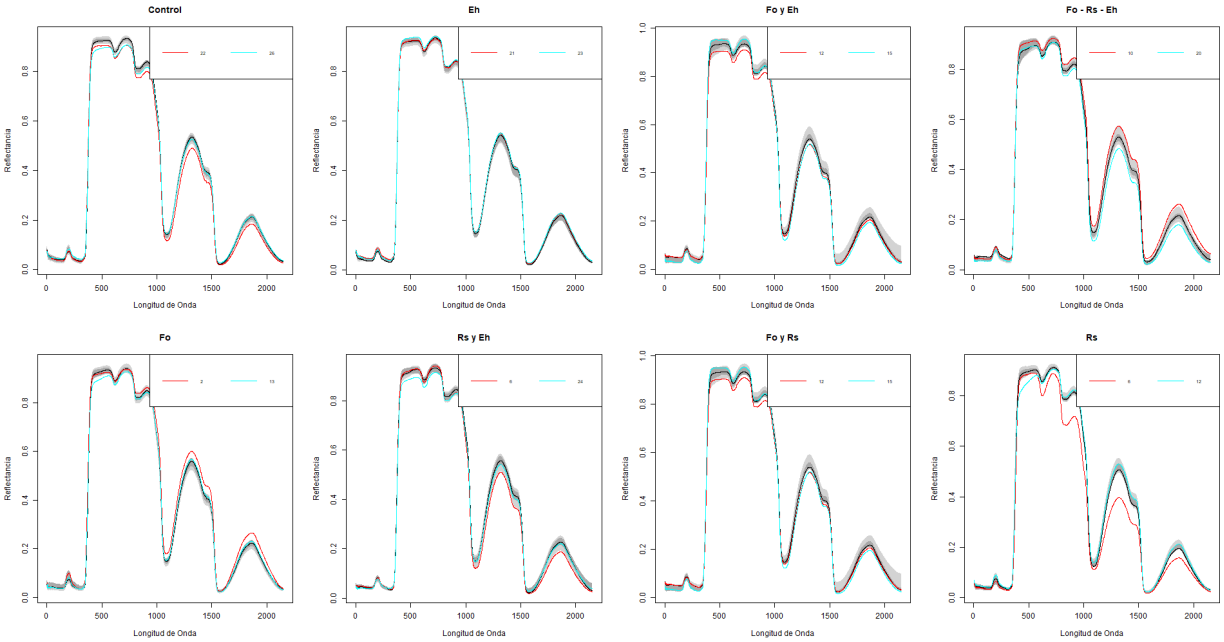

Figura 6-4: Detección de atípicos por tratamiento con el método RAD para el día 3.

6.2.2. Detección de atípicos en el sexto día

Método de la Bolsa (MB) en el sexto día

En el sexto día de medición, la aplicación del MB para la detección de outliers en datos funcionales reveló patrones significativos en los diferentes grupos experimentales. En particular, el grupo control presentó un notable aumento con la identificación de 6 observaciones atípicas, mientras que el grupo sometido a estrés hídrico también mostró 6 atípicos. De manera similar, el grupo inoculado con el hongo Fusarium exhibió 6 observaciones atípicas, evidenciando una mayor presencia de datos atípicos en comparación con el tercer día.

En contraste, el grupo inoculado con la bacteria Ralstonia tuvo 3 atípicos, y el grupo sometido a FOCR1 y RSR2 también presentó 3 observaciones atípicas. En el caso del grupo sometido a RSR2 y estrés hídrico, se identificó 1 outlier. Finalmente, el grupo expuesto a estrés hídrico, Fusarium y RSR2 mostró 3 atípicos.

Este método de detección de atípicos en el día 6 destaca un aumento en la identificación de datos atípicos en comparación con el día 3. Este fenómeno sugiere que, a medida que transcurre el tiempo, se incrementa la variabilidad en los datos espectrales, resultando en una mayor detección de observaciones atípicas. Este hallazgo subraya la importancia de considerar la evolución temporal al interpretar resultados y proporciona una perspectiva valiosa sobre la dinámica de las respuestas espectrales en las plantas de Banano Gros Michel.

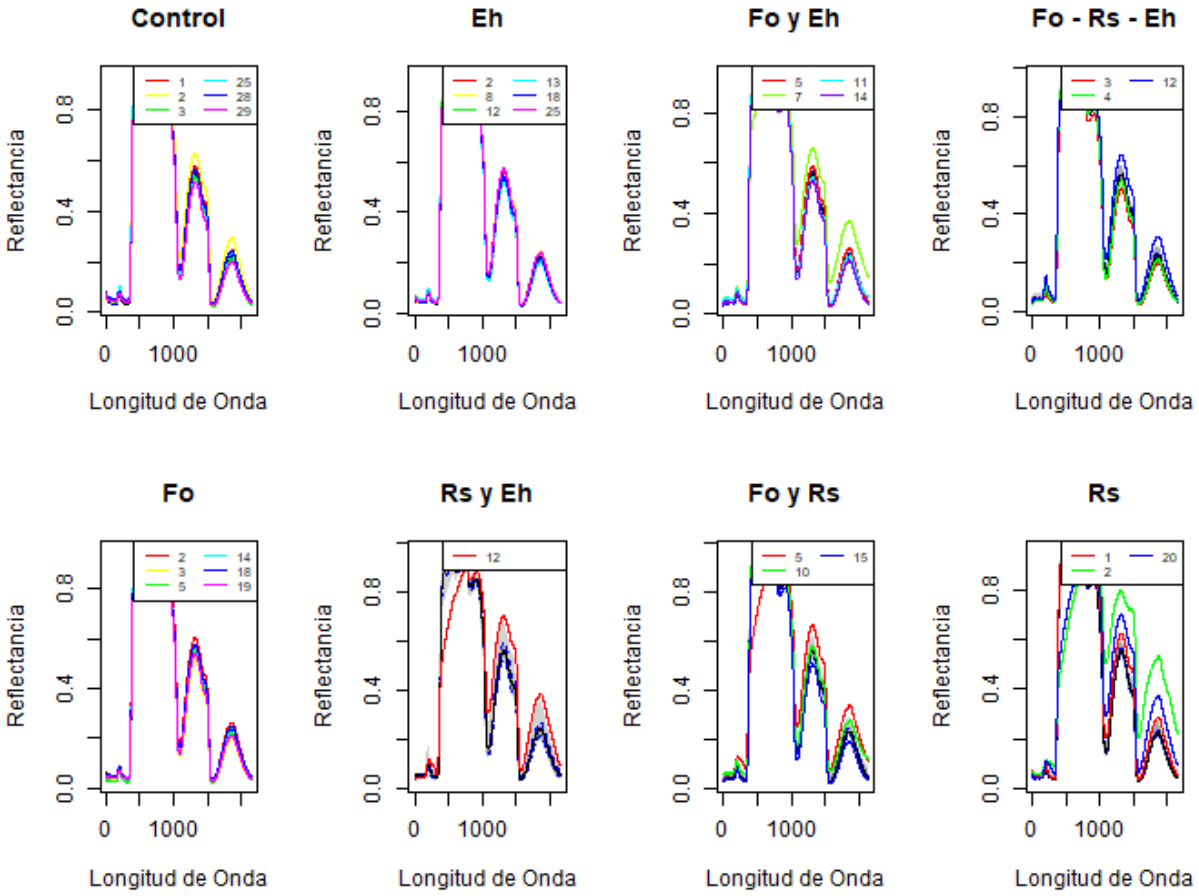

Figura 6-5: Detección de atípicos por tratamiento con el MB para el día 6.

Método RAD en el sexto día

En el sexto día de medición, la aplicación del método RAD para la detección de atípicos en datos funcionales reveló resultados distintivos en los diversos grupos experimentales. Se detectaron 2 observaciones atípicas en el grupo control, así como en el grupo sometido a estrés hídrico, el grupo inoculado con el hongo FOCR1, el grupo inoculado con la bacteria Ralstonia, el grupo sometido a Fusarium y estrés hídrico, el grupo sometido a FOCR1 y RSR2, el grupo sometido a Ralstonia y estrés hídrico, y, finalmente, en el grupo expuesto a estrés hídrico, Fusarium y Ralstonia.

En comparación con la aplicación del MB, se observa que el método RAD detectó menos datos atípicos en el sexto día de medición. Esta diferencia puede indicar una mayor robustez del método RAD ante la variabilidad de los datos espectrales, resaltando la importancia

de seleccionar cuidadosamente el enfoque de detección de atípicos según las características específicas de los conjuntos de datos. Este hallazgo subraya la necesidad de considerar múltiples metodologías para la detección de atípicos y resalta la importancia de elegir la más adecuada para garantizar resultados precisos y confiables en la interpretación de datos de reflectancia.

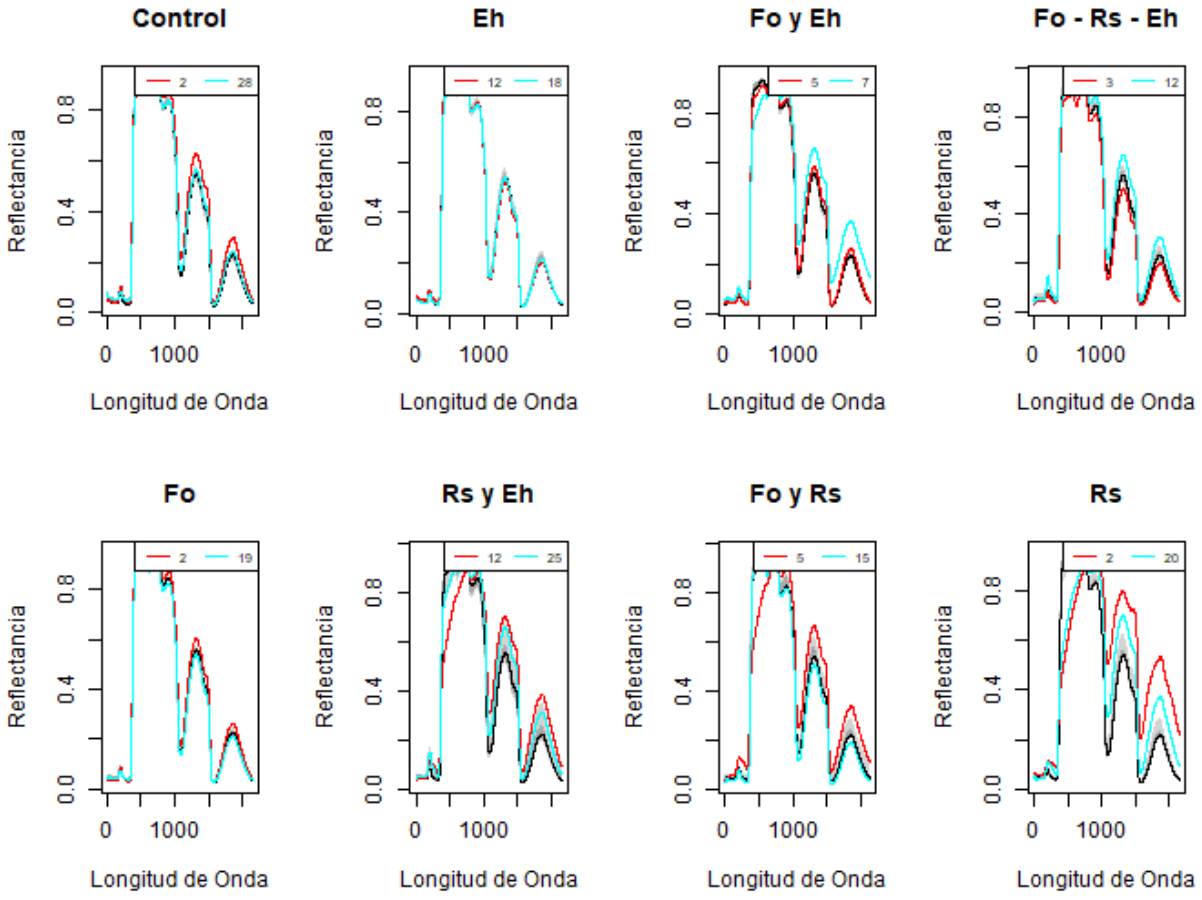

Figura 6-6: Detección de atípicos por tratamiento con el método RAD para el día 6.

6.3. Selección de longitudes de onda

En el tercer subcapítulo de resultados, se detalla el proceso de selección de características a través de la metodología RELIEF. Se identifican las longitudes de onda más relevantes que permiten una discriminación efectiva entre los distintos tratamientos. Este análisis contribuye a optimizar la eficiencia de los modelos de clasificación supervisada al enfocarse en las características espectrales más informativas para la detección temprana de estrés biótico

y abiótico en plantas de banano. Cabe destacar que, para garantizar una cobertura completa del espectro, se seleccionaron las longitudes de onda más relevantes en tramos de 100 nm desde los 350 nm hasta los 2500 nm. Esta estrategia integral busca abordar de manera exhaustiva las variaciones espectrales y asegurar una representación completa de la información relevante en el análisis de reflectancia.

6.3.1. Selección de características en el tercer día

A continuación, se presentan los resultados de la selección de longitudes de onda más discriminantes para cada grupo experimental, así como la combinación de todas ellas y la selección final de las 10 longitudes de onda más importantes según el algoritmo RELIEF en nanómetros:

- Datos Crudos: 350, 550, 578, 750, 750, 910, 1001, 1070, 1150, 1253, 1450, 1494, 1550, 1650, 1801, 1950, 1950, 2050, 2150, 2250, 2450.
- Datos crudos eliminando atípicos por MB: 350, 529, 636, 750, 850, 946, 996, 1090, 1250, 1257, 1450, 1512, 1635, 1650, 1801, 1922, 1954, 2109, 2250, 2255, 2369.
- Datos crudos eliminando atípicos por método RAD: 350, 525, 563, 705, 848, 910, 1048, 1050, 1250, 1253, 1350, 1462, 1650, 1651, 1850, 1919, 1986, 2050, 2199, 2286, 2419.
- Método SPM - MB: 353, 532, 567, 750, 850, 898, 1042, 1084, 1150, 1258, 1450, 1473, 1550, 1650, 1804, 1929, 2050, 2098, 2150, 2253, 2350.
- Método SPM - RAD: 353, 452, 632, 750, 825, 850, 1039, 1076, 1150, 1250, 1450, 1469, 1550, 1684, 1804, 1850, 2050, 2150, 2266, 2350.
- Método Savitzky-Golay - MB: 355, 544, 576, 750, 830, 911, 997, 1077, 1150, 1250, 1450, 1550, 1550, 1654, 1805, 1850, 1983, 2107, 2150, 2250, 2350.
- Método Savitzky-Golay RAD: 355, 550, 606, 750, 850, 943, 1001, 1050, 1150, 1252, 1350, 1475, 1649, 1650, 1804, 1940, 2015, 2115, 2150, 2253, 2350.
- Método Savitzky-Golay de segundo orden derivativo MB: 374, 473, 598, 670, 796, 909, 966, 1141, 1204, 1265, 1355, 1534, 1609, 1652, 1828, 1877, 1978, 2142, 2188, 2332, 2407.
- Método Savitzky-Golay de segundo orden derivativo RAD: 447, 473, 640, 670, 797, 908, 976, 1141, 1172, 1279, 1355, 1495, 1618, 1651, 1815, 1875, 2028, 2090, 2187, 2276, 2357.
- Método CDM - MB: 350, 550, 550, 750, 850, 943, 1050, 1150, 1150, 1314, 1450, 1550, 1624, 1650, 1801, 1881, 2050, 2097, 2150, 2349, 2350.

- Método CDM - RAD: 350, 450, 650, 750, 820, 950, 975, 1118, 1150, 1350, 1399, 1535, 1550, 1651, 1801, 1881, 1950, 2068, 2150, 2250, 2448.
- Método MCAS - MB: 350, 511, 634, 688, 850, 945, 1001, 1150, 1249, 1350, 1428, 1450, 1550, 1650, 1801, 1902, 1950, 2066, 2150, 2346, 2407.
- Método MCAS - RAD: 350, 495, 633, 683, 850, 950, 973, 1150, 1162, 1321, 1428, 1450, 1550, 1650, 1801, 1902, 1950, 2144, 2161, 2264, 2350.

Longitudes de onda más importantes según RELIEF (15 más relevantes, se unen todos y se seleccionan las 10 LO con más importancia según RELIEF más las 5 sugeridas por los expretos): 350, 353, 355, 1070, 1050, 1048, 1076, 1077, 1042, 1084, 1275, 1479, 1600, 1801, 2167.

La proximidad de las longitudes de onda identificadas como más relevantes por el algoritmo RELIEF sugiere que existe una cierta continuidad en la información espectral. Dada esta cercanía, se adopta un enfoque pragmático para la selección final de longitudes de onda, optando por aquellas que se encuentran más próximas entre sí. Este enfoque se basa en la premisa de que, aunque el algoritmo RELIEF destaca longitudes de onda específicas, la información contenida en rangos cercanos puede ser igualmente valiosa y complementaria.

Para potenciar aún más la robustez de la selección de características, se lleva a cabo una segunda iteración del algoritmo RELIEF, centrada en los rangos espectrales de 1200-1300 nm, 1375-1500 nm, 1600-1700 nm, 1750-1850 nm y 2150-2250 nm. Estos rangos fueron elegidos estratégicamente debido a su relevancia conocida para los expertos en el campo, incluso si no emergieron como prioritarios en la primera aplicación del algoritmo RELIEF.

La inclusión de estas nuevas longitudes de onda busca capturar información específica relacionada con procesos bioquímicos y fenómenos físicos relevantes en la reflectancia de las plantas. Aunque estos rangos pueden no haber sido inicialmente resaltados por el algoritmo RELIEF, su importancia contextual para la investigación y la experiencia previa de los expertos respalda su consideración en la selección final de características. Este enfoque integrado tiene como objetivo enriquecer la representación espectral, equilibrando la capacidad del algoritmo para identificar características específicas con el conocimiento especializado del dominio.

6.3.2. Selección de características en el sexto día

Los resultados del día 6 del experimento muestran la importancia de las longitudes de onda en la detección de estrés en plantas de banano. A continuación, se presentan los resultados para cada grupo de comparación:

- Datos Crudos: Las longitudes de onda relevantes incluyen 350, 550, 581, 703, 775, 850, 950, 1150, 1236, 1252, 1354, 1550, 1645, 1650, 1750, 1850, 2050, 2150, 2350.

- Datos crudos eliminando atípicos por MB: La eliminación de outliers destaca longitudes de onda como 352, 505, 619, 662, 750, 850, 1050, 1150, 1183, 1250, 1450, 1474, 1550, 1650, 1801, 1950, 2050, 2150, 2250, 2449.

- Datos crudos eliminando atípicos por método RAD: La metodología RAD resalta la importancia de longitudes de onda como 352, 506, 630, 707, 850, 855, 950, 1150, 1213, 1250, 1386, 1550, 1557, 1743, 1750, 1925, 1981, 2050, 2150, 2350, 2440.

- Método SPM- MB: La aplicación de suavizamiento destaca longitudes de onda como 353, 503, 619, 663, 750, 850, 950, 1150, 1183, 1250, 1450, 1474, 1550, 1650, 1804, 1950, 2050, 2150, 2250, 2450.

- Método SPM - RAD: El suavizamiento mediante RAD resalta longitudes de onda como 353, 508, 630, 713, 850, 857, 950, 1150, 1214, 1250, 1385, 1550, 1555, 1743, 1750, 1924, 1981, 2050, 2150, 2350, 2439.

- Método Savitzky-Golay - MB: La aplicación de este método destaca longitudes de onda como 355, 504, 619, 664, 750, 850, 950, 1150, 1183, 1250, 1450, 1474, 1550, 1650, 1805, 1950, 2050, 2150, 2250, 2450.

- Método Savitzky-Golay RAD: El suavizamiento mediante RAD resalta longitudes de onda como 355, 508, 630, 705, 850, 856, 950, 1150, 1214, 1250, 1386, 1550, 1556, 1743, 1796, 1925, 1981, 2050, 2150, 2350, 2440.

- Método Savitzky-Golay de segundo orden derivativo MB: La derivación de segundo orden mediante MB destaca longitudes de onda como 425, 459, 628, 661, 819, 902, 958, 1140, 1209, 1280, 1354, 1506, 1584, 1747, 1814, 1857, 1994, 2140, 2158, 2257, 2443.

- Método Savitzky-Golay de segundo orden derivativo RAD: La derivación de segundo orden mediante RAD destaca longitudes de onda como 441, 458, 609, 653, 835, 880, 953, 1140, 1209, 1280, 1354, 1489, 1584, 1737, 1816, 1875, 2004, 2089, 2166, 2343, 2431.

- Método CDM - MB: La corrección de dispersión multiplicativa mediante MB destaca longitudes de onda como 350, 458, 650, 744, 850, 950, 963, 1148, 1173, 1350, 1350, 1550, 1550, 1650, 1801, 1880, 2035, 2050, 2250, 2356.

- Método CDM - RAD: La corrección de dispersión multiplicativa mediante RAD resalta longitudes de onda como 352, 454, 650, 710, 815, 854, 956, 1150, 1198, 1342, 1352, 1534, 1550, 1650, 1801, 1883, 1955, 2076, 2150, 2250, 2450.

- Método MCAS - MB: La aplicación de este método destaca longitudes de onda como 352, 508, 650, 686, 750, 913, 963, 1150, 1150, 1303, 1418, 1450, 1550, 1729, 1801, 1936, 1950, 2050, 2150, 2349, 2443.

- Método MCAS - RAD: La aplicación de este método resalta longitudes de onda como 352, 550, 623, 678, 849, 850, 1008, 1145, 1150, 1250, 1440, 1450, 1550, 1650, 1800, 1937, 1950, 2050, 2150, 2349, 2405.

Se unen todos y se seleccionan las 10 longitudes de onda con más importancia según RELIEF: 350, 352, 353, 355, 703, 705, 707, 710, 581, 609. Al igual que en la selección de longitudes de onda del día 3, se aplicó el algoritmo RELIEF para elegir aquellas que mejor discriminan entre los grupos experimentales. Se seleccionó la longitud de onda más relevante en los siguientes intervalos: 1200-1300 nm, 1375-1500 nm, 1600-1700 nm, 1750-1850 nm y 2150-2250 nm. Las longitudes de onda identificadas en este último procedimiento, junto con las obtenidas previamente, son: 350, 352, 353, 355, 703, 705, 707, 710, 581, 609, 1300, 1383, 1643, 1750, 2150.

6.4. Análisis de Clasificación

Se presenta un análisis de los resultados obtenidos mediante diversos métodos de clasificación supervisada. Se exploran los desempeños de técnicas como ALD, ADC , BA, BI, MSV, KVC y PM. La evaluación se realiza en base a la exactitud de clasificación, destacando la eficacia de cada método para discriminar entre los tratamientos analizados en los días 3 y 6 post-inoculación de patógenos. Cabe destacar que estos análisis de clasificación se fundamentan en las longitudes de onda seleccionadas en el capítulo anterior, las cuales han demostrado ser relevantes para la detección temprana de estrés biótico y abiótico en plantas de banano.

6.4.1. Análisis de clasificación en el tercer día

En el análisis correspondiente al día 3 de medición, se lleva a cabo una exhaustiva serie de pruebas para validar los supuestos necesarios antes de aplicar técnicas de clasificación supervisada, como el ALD y el ADC. Se inicia con pruebas de normalidad multivariada, las cuales son cruciales para asegurar la validez de los resultados.

Adicionalmente, se examina el supuesto de homogeneidad en las matrices de covarianzas, un aspecto esencial para la aplicación exitosa de ALD y ACD. Para evaluar este supuesto, se utiliza la prueba de Box's M-test en cada uno de los 13 conjuntos de datos, los cuales resultan de la combinación de diferentes métodos de suavizamiento con la eliminación de

datos atípicos.

Los resultados indican que, con un nivel de significancia del 5 %, la hipótesis de normalidad multivariada no es rechazada para conjuntos de datos específicos, como SG_RAD, MSC_MB, MSC_RAD, MCAS_MB y MCAS_RAD, según la prueba de Royston. Por otro lado, las pruebas de homogeneidad en las matrices de covarianzas revelan que no hay evidencia estadísticamente significativa para afirmar que existe igualdad en estas matrices en ninguno de los 13 conjuntos de datos, manteniendo un nivel de significancia del 5 %.

Con base en las longitudes de onda presentadas en el anterior capítulo de resultados, se llevó a cabo un análisis integral que abarcó las siete técnicas de clasificación supervisada para cada uno de los ocho tratamientos y los trece conjuntos de datos, considerando todas las plantas del experimento. Es importante destacar que en esta etapa no se realizó la división del conjunto de datos en entrenamiento y prueba, lo que permitió evaluar el desempeño global de cada técnica en el conjunto completo. Entre los resultados obtenidos, se destaca que la técnica de Bosques Aleatorios exhibió la exactitud más óptima, oscilando entre el 99,1 % y el 100 %. Sin embargo, esta elevada exactitud plantea la posibilidad de sobreajuste en el modelo de clasificación, especialmente al considerar la elección del número óptimo de árboles en el bosque.

En contraste, la técnica de Perceptrón Multicapa mostró la menor exactitud, fluctuando entre el 23,7 % y el 52,7 %. Estos resultados sugieren que las técnicas basadas en redes neuronales, como el Perceptrón Multicapa, podrían requerir un mayor tamaño de muestra para alcanzar niveles óptimos de exactitud, incluso al explorar configuraciones variadas de capas ocultas y número de neuronas. La siguiente tabla presenta las precisiones (Accuracy) de las distintas técnicas de clasificación (columnas) aplicadas a cada conjunto de datos (filas), considerando la totalidad de las plantas y eliminando los atípicos en cada caso. (ver tabla **6-1**)

Tabla 6-1: Tabla de exactitud sin división en conjunto de entrenamiento y prueba para el día 3

	Preprocesamiento	ALD	ACD	BI	BA	MSV	KVC	PM
1	Crudos	0.87	0.99	0.50	1.00	0.57	0.68	0.40
2	MB	0.90	1.00	0.54	0.99	0.60	0.69	0.38
3	RAD	0.90	1.00	0.52	1.00	0.58	0.75	0.42
4	SPM_MB	0.85	0.98	0.51	1.00	0.60	0.77	0.38
5	SPM_RAD	0.84	0.98	0.49	1.00	0.58	0.75	0.39
6	SG_MB	0.84	0.96	0.46	1.00	0.58	0.73	0.33
7	SG_RAD	0.84	0.97	0.46	1.00	0.57	0.80	0.38
8	SG2_MB	0.12	0.99	0.86	1.00	0.95	0.69	0.53
9	SG2_RAD		0.99	0.87	1.00	0.95	0.66	0.39
10	MSC_MB	0.91	1.00	0.39	0.99	0.60	0.69	0.28
11	MSC_RAD	0.92	1.00	0.40	1.00	0.59	0.68	0.25
12	MCAS_MB	0.91	1.00	0.46	1.00	0.70	0.79	0.24
13	MCAS_RAD	0.91	1.00	0.49	1.00	0.69	0.78	0.35

Considerando la limitación inherente al modelar la clasificación supervisada basada en los datos espectrales sobre todas las plantas, podría generar sesgos, por lo tanto, se procedió a implementar los modelos sobre conjuntos de entrenamiento y prueba, dividiendo aleatoriamente la muestra en un 75 % de entrenamiento y un 25 % de prueba. Esta división reveló que el método de Bosques Aleatorios estaba experimentando sobreajuste, ya que, al corregir este sesgo, se observó una reducción en la exactitud, fluctuando entre el 38,6 % y el 5,31 % para diferentes conjuntos de datos.
En relación con los parámetros más óptimos identificados en esta fase, considerando tanto los supuestos como las precisiones obtenidas, se destacan las siguientes observaciones (ver tabla **6-2**):

- Análisis lineal discriminante (ALD): A pesar de no cumplir con los supuestos de normalidad multivariada y homogeneidad en las matrices de covarianzas en la mayoría de los subconjuntos de datos, se encontraron las precisiones más altas, oscilando entre el 71,9 % y el 86 %.

- Análisis cuadrático discriminante (ACD): Aunque las precisiones varían entre el 51,8 % y el 77,2 %, se destaca como la segunda metodología con mejores precisiones a pesar de no cumplir con el supuesto de normalidad multivariada, pues, sin tener en cuenta las precisiones obtenidas con el ALD, esta metodología presenta las mayores precisiones.

- Máquinas de Soporte Vectorial (MSV): Excluyendo las precisiones de ALD y ACD, en 5 de los 13 conjuntos de datos, MSV se posiciona como el tercer método con mejores precisiones, presentando valores entre el 32,1 % y 78,9 %.

- Naïve Bayes: Este método destaca al alcanzar la mejor exactitud en 4 de los 13 conjuntos de datos, este método se posiciona como el cuarto método con mejores precisiones, oscilando entre el 24,6 % y el 83,9 %.
- K Vecinos Más Cercanos (KVC) y Bosques Aleatorios (BA): Sin considerar las precisiones de ALD y ACD, en 3 de los 13 conjuntos de datos, cada uno, KVC y BA se ubican como el quinto método con mejores precisiones. Para KVC las precisiones oscilan entre el 28,6 % y el 50 %, el número óptimo de vecinos se encuentra entre 2 y 5. Por otro lado, Para BA las precisiones oscilan entre el 31,6 % y el 69,2 %, este resultado confirma la observación anterior sobre el sobreajuste cuando se evalúa el conjunto completo, el número de árboles óptimo para este método varía entre 5 y 30.
- Perceptrón Multicapa: Este método exhibe las precisiones más bajas entre todos los métodos, atribuibles a su mejor rendimiento con un número más extenso de muestras. Se observa un desempeño superior en modelos con 5 capas ocultas, cada una con 10 neuronas.

Tabla 6-2: Tabla de exactitud con división en conjunto de entrenamiento y prueba para el día 3

	Preprocesamiento	ALD	ACD	BI	BA	MSV	KVC	PM
1	Crudos	0.81	0.56	0.45	0.39	0.50	0.42	0.25
2	MB	0.77	0.60	0.47	0.37	0.46	0.35	0.27
3	RAD	0.77	0.57	0.39	0.43	0.41	0.39	0.36
4	SPM_MB	0.79	0.58	0.47	0.42	0.46	0.33	0.23
5	SPM_RAD	0.79	0.52	0.39	0.39	0.43	0.50	0.27
6	SG_MB	0.72	0.63	0.42	0.35	0.42	0.37	0.30
7	SG_RAD	0.75	0.52	0.38	0.43	0.34	0.32	0.20
8	SG2_MB		0.77	0.70	0.69	0.79	0.33	0.48
9	SG2_RAD		0.59	0.84	0.66	0.86	0.38	0.66
10	MSC_MB	0.75	0.54	0.25	0.32	0.35	0.39	0.27
11	MSC_RAD	0.84	0.55	0.27	0.32	0.32	0.29	0.29
12	MCAS_MB	0.86	0.58	0.42	0.35	0.39	0.35	0.39
13	MCAS_RAD	0.80	0.59	0.36	0.34	0.39	0.43	0.29

Mejor modelo en el tercer día

El modelo más efectivo identificado durante el análisis de clasificación supervisada fue el Análisis Lineal Discriminante (ALD), aplicando el método de suavizamiento mínimos cuadrados asimétricos (MCAS) y la técnica de eliminación de atípicos MB, logrando una destacable exactitud del 86 %. Es esencial someter este modelo a pruebas que evalúen la normalidad

multivariada y la homogeneidad en las matrices de covarianzas.

La prueba de normalidad multivariada, efectuada mediante la prueba de Royston, arrojó una estadística H de 5,569611 y un p-valor de 0,1383688. Esto indica que, con un nivel de significancia del 5 %, hay evidencia estadística para no rechazar la hipótesis nula de normalidad multivariada. En contraste, la prueba de Homogeneidad en la varianza de Box's M-test resultó en una estadística chi cuadrado aproximada de 1087,2 con 840 grados de libertad y un p valor de 1,508 e-08. En este caso, se rechaza la hipótesis nula de homogeneidad en las varianzas de los tratamientos con un nivel de significancia del 5 %. En consecuencia, solo se cumple la hipótesis de normalidad multivariada.

A pesar del rechazo de la hipótesis de igualdad en las matrices de covarianzas, la opción más recomendada sería utilizar Análisis Discriminante Cuadrático (ACD). Sin embargo, se debe destacar que la exactitud obtenida con ACD es significativamente menor que la alcanzada con ALD.

Posteriormente, se examinaron las probabilidades a priori de los diferentes tratamientos. Se observa que las variaciones son mínimas, lo cual se atribuye al número reducido de atípicos eliminados en cada caso, indicando una relativa estabilidad en las probabilidades a priori.(ver **6-3**

Tabla 6-3: Probabilidades a priori para el ALD en el conjunto de datos MCAS_MB para el día 3

Control	E_Hidrico	Fus_EH	Fus_EH_Ral	Fusarium	Ral_EH	Ral_Fus	Ralstonia
0,12	0,13	0,12	0,11	0,12	0,12	0,12	0,11

La tabla **6-4** presenta los coeficientes asociados a las 7 ecuaciones lineales discriminantes. Cada ecuación, expresada como una combinación lineal de las longitudes de onda correspondientes, contribuye a la toma de decisiones al ubicar los nuevos valores espectroscópicos calculados en regiones específicas. Estas ecuaciones lineales desempeñan un papel crucial al predecir el tratamiento más probable para una nueva observación, brindando un marco analítico que aprovecha la información espectral para realizar inferencias precisas y contextualizadas en función de los distintos tratamientos experimentales. El análisis detallado de estos coeficientes proporciona una comprensión más profunda de cómo cada longitud de onda contribuye al proceso de clasificación, facilitando una interpretación informada de los resultados obtenidos.

Tabla 6-4: Coeficientes de las ecuaciones lineales discriminantes para el conjunto de datos MCAS_MB para el día 3

Longitud de onda	LD1	LD2	LD3	LD4	LD5	LD6	LD7
350nm	-182,04	29,89	53,57	-74,39	11,82	218,00	-138,54
353nm	-116,62	-28,15	-92,65	141,79	-70,51	20,22	-7,78
355nm	-113,40	39,40	20,72	5,48	49,12	-276,95	211,64
1042nm	2561,73	-6871,84	-1520,84	-2247,92	-5522,65	-1783,56	-2949,60
1048nm	-1616,84	33699,68	-7038,64	3172,64	6802,49	8517,70	14930,44
1050nm	-986,14	-27480,59	10940,06	122,72	1757,23	-7270,03	-12178,07
1070nm	1933,37	808,70	2108,60	-9603,83	-11429,70	-173,85	4084,85
1076nm	-1407,63	-15237,41	-17373,97	30694,36	8815,67	14942,81	18059,14
1077nm	560,59	18619,43	6693,78	-19753,43	3611,87	-16917,08	-28140,26
1084nm	-917,11	-3633,49	6325,32	-2404,36	-4180,82	2674,55	6311,64
1275nm	54,26	162,81	-92,30	15,12	229,98	13,54	-156,99
1479nm	189,27	-12,17	-127,14	468,49	-159,59	182,36	95,91
1600nm	-210,75	37,45	174,34	8,37	-182,34	-13,13	-27,43
1801nm	92,49	-128,01	-295,08	-144,47	-52,14	202,76	104,31
2167nm	217,84	100,16	259,08	31,67	222,46	-224,44	-128,31

El rendimiento del modelo de clasificación se destaca con una exactitud del 85,96 %, respaldada por un intervalo de confianza del 95 % que varía desde el 74,21 % hasta el 93,74 %. Esta notable capacidad de clasificación se logra mediante la combinación estratégica del Análisis Lineal Discriminante (ALD) con los métodos de suavizamiento mínimos cuadrados asimétricos (MCAS) y la gestión de atípicos mediante el método MB.

Este modelo demuestra no solo una capacidad efectiva de clasificación, sino también una utilidad significativa en la detección temprana de estrés en plantas. En el análisis del día 3 de medición, se observa que se puede prever con gran certeza cada uno de los tratamientos, manteniendo un bajo error de clasificación del 14 %. Este hallazgo subraya la eficacia del modelo en identificar patrones tempranos de estrés, proporcionando así una herramienta valiosa para la toma de decisiones y la implementación de medidas correctivas en el ámbito de la salud de las plantas.

Tabla 6-5: Matriz de confusión para el ALD con el conjunto de datos MCAS_MB para el día 3

	Referencia							
Predicción	Control	E_Hidrico	Fus_EH	Fus_EH_Ral	Fus	Ral_EH	Ral_Fus	Ral
Control	6	0	0	0	0	0	0	0
E_Hidrico	1	7	0	0	0	0	0	0
Fus_EH	0	0	4	1	0	0	1	0
Fus_EH_Ra	0	0	0	5	0	0	0	0
Fusarium	0	0	1	0	7	0	0	0
Ral_EH	0	0	2	1	0	7	0	0
Ral_Fus	0	0	0	0	0	0	6	0
Ralstonia	0	1	0	0	0	0	0	7

Mejor modelo parsimonioso en el tercer día

Considerando la complejidad asociada con el análisis de siete tratamientos y quince longitudes de onda, se busca simplificar la interpretación del discriminante lineal presentado previamente. La estrategia adoptada implica la reducción tanto en el número de tratamientos como en el de variables, con el objetivo de mejorar la claridad y la comprensión del análisis.

En el proceso de reducción de tratamientos, se opta por focalizarse en los grupos fundamentales, excluyendo las interacciones. Esto se traduce en la consideración exclusiva del grupo control, el estrés hídrico, las plantas sometidas únicamente a la bacteria RSR2 y las plantas sometidas únicamente al hongo Fusarium. Este enfoque simplificado facilita la interpretación y permite un análisis más detallado de los tratamientos individuales.

Paralelamente, para abordar la reducción en el número de variables, se seleccionan las más significativas mediante el algoritmo RELIEF. Se eligen las cinco longitudes de onda más relevantes, a saber: 350 nm, 1042 nm, 1479 nm, 1600 nm y 1801 nm. Este paso contribuye a optimizar la eficiencia del modelo al centrarse en las características espectrales más informativas.

En el análisis resultante, se exploran las probabilidades a priori de los tratamientos seleccionados (ver tabla **6-6**). Se destaca la mínima variación observada, atribuible al reducido número de atípicos eliminados en cada caso. Este fenómeno refleja una relativa estabilidad en las probabilidades a priori, reforzando la confiabilidad y consistencia de los resultados obtenidos con este enfoque simplificado.

Tabla 6-6: Probabilidades a priori para el ALD parsimonioso en el conjunto de datos MCAS_MB para el día 3

Control	E_Hidrico	Fusarium	Ralstonia
0,25	0,26	0,25	0,23

A continuación, se presenta una descripción de los coeficientes vinculados a las tres ecuaciones lineales discriminantes en el modelo parsimonioso. Estas ecuaciones, formuladas como combinaciones lineales de las longitudes de onda respectivas, desempeñan un papel clave en la toma de decisiones al asignar ubicaciones específicas a los nuevos valores espectroscópicos calculados. Su función esencial radica en predecir el tratamiento más probable para una nueva observación, utilizando la información espectral para realizar inferencias precisas y contextualizadas en el contexto de los diversos tratamientos experimentales.

El análisis detallado de estos coeficientes proporciona una comprensión más profunda de la contribución de cada longitud de onda al proceso de clasificación. Esto facilita una interpretación informada de los resultados obtenidos, permitiendo discernir de manera más precisa cómo cada característica espectral influye en la asignación de tratamientos, enriqueciendo así la interpretación analítica de los datos. (ver **6-7**)

Tabla. Coeficientes de las ecuaciones lineales discriminantes del modelo ALD parsimonioso para el conjunto de datos MCAS_MB para el día 3

Tabla 6-7: Coeficientes de las ecuaciones lineales discriminantes para el conjunto de datos MCAS_MB para el modelo parsimonioso en el día 3

Longitud de Onda	LD1	LD2	LD3
350nm	263,91	76,98	66,68
1042nm	-78,33	74,87	-22,52
1479nm	-491,40	-26,94	42,75
1600nm	33,08	82,54	220,03
1801nm	-83,90	-21,59	-251,11

El desempeño del modelo de clasificación se distingue por su destacada exactitud del 86,21 %, respaldada por un intervalo de confianza del 95 % que oscila entre el 68,34 % y el 96,11 %. Este nivel excepcional de clasificación se logra mediante una estratégica combinación del Análisis Lineal Discriminante (ALD) con los métodos de suavizamiento mínimos cuadrados asimétricos (MCAS) y la gestión de atípicos mediante el método MB, aplicado a solo cuatro tratamientos y cinco variables.

Este modelo no solo exhibe una efectiva capacidad de clasificación, sino que también demuestra utilidad significativa en la detección temprana de estrés en plantas. Al analizar los resultados del día 3 de medición, se destaca la capacidad predictiva precisa para cada tratamiento, manteniendo un bajo error de clasificación del 13,7 %. Este hallazgo subraya la eficacia del modelo en identificar patrones tempranos de estrés, ofreciendo así una herramienta valiosa para la toma de decisiones y la implementación de medidas correctivas en el ámbito de la salud de las plantas.

Al compararse con el modelo anterior, se observan precisiones muy similares, pero con la ventaja de reducir tanto el número de variables como el número de tratamientos. Este enfoque más simplificado no compromete la eficacia del modelo, destacando su capacidad para lograr resultados robustos con mayor parsimonia.

Tabla 6-8: Matriz de confusión para el ALD parsimonioso con el conjunto de datos MCAS_MB para el día 3

	Referencia			
Predicción	Control	E_Hidrico	Fusarium	Ralstonia
Control	7	0	0	0
E_Hidrico	0	6	0	2
Fusarium	0	0	7	0
Ralstonia	0	2	0	5

En la gráfica **6-7**, se aprecia con claridad que cada par de longitudes de onda consideradas en el análisis lineal discriminante delinea de manera distintiva regiones específicas para cada uno de los tratamientos examinados. Destaca notablemente que en la mayoría de estos planos, las plantas de control se sitúan en un extremo, mientras que en el extremo opuesto se encuentran las plantas inoculadas con Fusarium. En el espacio intermedio se distribuyen tanto las plantas sometidas a estrés hídrico como aquellas inoculadas con Ralstonia. Este patrón visual subraya la capacidad del modelo para generar una separación clara y discernible entre los diferentes tratamientos, proporcionando una representación gráfica efectiva de la discriminación lograda a través de las longitudes de onda seleccionadas.

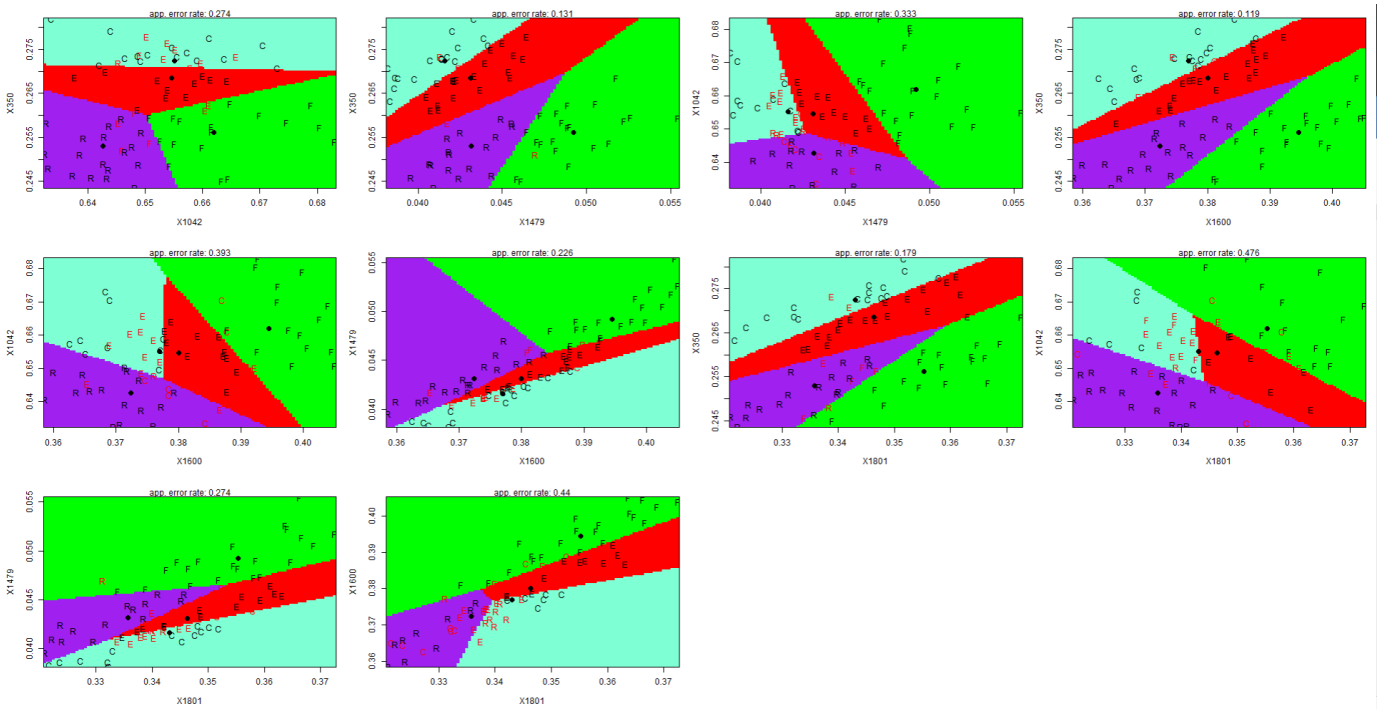

Figura 6-7: Diagramas de regiones discriminantes según longitudes de onda seleccionadas para el ALD parsimonioso para el día 3.

6.4.2. Análisis de clasificación en el sexto día

En el día 6 post-inoculación, se llevaron a cabo técnicas de análisis discriminante para clasificar los 8 tratamientos. En este proceso, se realizaron pruebas cruciales para validar los supuestos necesarios antes de aplicar técnicas de clasificación supervisada, como el Análisis Lineal Discriminante (ALD) y el Análisis Cuadrático Discriminante (ACD). Se comenzó con pruebas de normalidad multivariada, esenciales para garantizar la validez de los resultados.

Además, se evaluó el supuesto de homogeneidad en las matrices de covarianzas, un aspecto fundamental para el éxito de ALD y ACD. La prueba de Box's M-test se utilizó en cada uno de los 13 conjuntos de datos, que resultaron de la combinación de diferentes métodos de suavizamiento con la eliminación de datos atípicos.

Los resultados indicaron que, con un nivel de significancia del 5 %, la hipótesis de normalidad multivariada fue rechazada para los 13 subconjuntos de datos, según la prueba de Royston. Por otro lado, las pruebas de homogeneidad en las matrices de covarianzas mostraron que no había evidencia estadísticamente significativa para afirmar la igualdad en estas matrices en ninguno de los 13 conjuntos de datos, manteniendo un nivel de significancia del 5 %.

Con base en las longitudes de onda presentadas anteriormente, se realizó un análisis integral

que abarcó siete técnicas de clasificación supervisada para cada uno de los ocho tratamientos y los trece conjuntos de datos, considerando todas las plantas del experimento. Es crucial destacar que en esta etapa no se dividió el conjunto de datos en entrenamiento y prueba, lo que permitió evaluar el desempeño global de cada técnica en el conjunto completo. Entre los resultados obtenidos, se destaca que la técnica de Bosques Aleatorios exhibió la máxima exactitud, oscilando entre el 99,5 % y el 100 %. Sin embargo, esta elevada exactitud plantea la posibilidad de sobreajuste en el modelo de clasificación, especialmente al considerar la elección del número óptimo de árboles en el bosque.

En contraste, la técnica de Perceptrón Multicapa mostró la menor exactitud, fluctuando entre el 29,2 % y el 48,5 %. Estos resultados sugieren que las técnicas basadas en redes neuronales, como el Perceptrón Multicapa, podrían requerir un mayor tamaño de muestra para alcanzar niveles óptimos de exactitud, incluso al explorar configuraciones variadas de capas ocultas y número de neuronas. A diferencia del día 3, se observaron menores precisiones en el día 6. La siguiente tabla presenta las precisiones (Accuracy) de las distintas técnicas de clasificación (columnas) aplicadas a cada conjunto de datos (filas), considerando la totalidad de las plantas y eliminando los atípicos en cada caso (ver tabla **6-9**).

Tabla 6-9: Tabla de exactitud sin división en conjunto de entrenamiento y prueba para el día 6

	Preprocesamiento	ALD	ACD	BI	BA	MSV	KVC	PM
1	Crudos	0.84	0.98	0.53	1.00	0.69	0.71	0.37
2	MB	0.92	0.99	0.56	0.99	0.69	0.76	0.38
3	RAD	0.88	0.98	0.54	1.00	0.69	0.74	0.35
4	SPM_MB	0.85	0.97	0.48	1.00	0.64	0.72	0.41
5	SPM_RAD	0.82	0.97	0.47	1.00	0.65	0.72	0.36
6	SG_MB	0.81	0.95	0.43	1.00	0.63	0.68	0.41
7	SG_RAD	0.78	0.94	0.40	1.00	0.63	0.62	0.30
8	SG2_MB	0.12	0.96	0.64	0.99	0.81	0.70	0.48
9	SG2_RAD		0.96	0.65	1.00	0.83	0.71	0.40
10	MSC_MB	0.88	0.99	0.46	1.00	0.64	0.78	0.35
11	MSC_RAD	0.85	0.99	0.45	1.00	0.63	0.81	0.34
12	MCAS_MB	0.88	0.99	0.52	0.99	0.74	0.80	0.26
13	MCAS_RAD	0.88	0.99	0.50	1.00	0.75	0.76	0.29

Considerando la limitación al modelar la clasificación supervisada basada en todas las plantas, lo cual podría introducir sesgos, se optó por implementar los modelos mediante conjuntos de entrenamiento y prueba. Se dividió aleatoriamente la muestra en un 75 % de entrenamiento y un 25 % de prueba. Esta división reveló que el método de Bosques Aleatorios experimen-

taba sobreajuste, ya que, al corregir este sesgo, se observó una disminución en la exactitud, fluctuando entre el 31,6 % y el 69,2 % para diferentes conjuntos de datos.

En cuanto a los parámetros más óptimos identificados en esta fase, considerando tanto los supuestos como las precisiones obtenidas, se destacan las siguientes observaciones: (ver **6-10**)

- Análisis Discriminante Lineal (ALD): A pesar de no cumplir con los supuestos de normalidad multivariada y homogeneidad en las matrices de covarianzas en los 13 subconjuntos de datos, se observaron las precisiones más altas, oscilando entre el 64,2 % y el 86,8 %.

- Análisis Discriminante Cuadrático (ACD): Sin tener en cuenta las precisiones obtenidas con el ALD, los resultados con esta metodología presentaron las mayores precisiones en 7 de los 13 conjuntos de datos, variando entre el 43,4 % y el 66,7 %.

- Bosques Aleatorios (BA): Para el día 6 de medición y sin considerar los resultados obtenidos con el ALD, se encontró que con esta metodología se obtuvieron las mayores precisiones en 4 de los 13 conjuntos de datos, fluctuando entre el 35,1 % y 63,2 %. El número óptimo de árboles varió dependiendo del conjunto de datos y estuvo entre 10 y 35.

- Máquinas de Soporte Vectorial (MSV): Excluyendo las precisiones de ALD, en 3 de los 13 conjuntos de datos, MSV se posicionó como el cuarto método con mejores precisiones, presentando valores entre el 35,1 % y 61,4 %.

- Bayes Ingenuo (BI), K Vecinos Más Cercanos (KVC) y Perceptrón Multicapa (PM): Sin considerar las precisiones de ALD, ninguno de los 13 conjuntos de datos presentó las precisiones más óptimas. Para BI, las precisiones oscilaron entre el 39,6 % y el 63,2 %. Para KVC, las precisiones variaron entre el 30,2 % y el 56,1 %, con el número óptimo de vecinos entre 2 y 5. Finalmente, para el PM, se ubicó en último lugar con precisiones entre el 24,5 % y el 63,6 %, mostrando un rendimiento superior en modelos con 5 capas ocultas, cada una con 10 neuronas.

Tabla 6-10: Tabla de exactitud con división en conjunto de entrenamiento y prueba para el día 6

	Preprocesamiento	ALD	ACD	BI	BA	MSV	KVC	PM
1	Crudos	0.67	0.53	0.57	0.58	0.57	0.47	0.39
2	MB	0.77	0.51	0.40	0.53	0.45	0.42	0.32
3	RAD	0.77	0.61	0.54	0.47	0.53	0.44	0.33
4	SPM_MB	0.64	0.55	0.38	0.43	0.42	0.26	0.40
5	SPM_RAD	0.77	0.63	0.44	0.49	0.51	0.42	0.31
6	SG_MB	0.68	0.47	0.36	0.43	0.40	0.30	0.26
7	SG_RAD	0.79	0.65	0.39	0.35	0.44	0.37	0.35
8	SG2_MB		0.45	0.49	0.57	0.70	0.42	0.36
9	SG2_RAD		0.53	0.58	0.61	0.77	0.35	0.64
10	MSC_MB	0.87	0.43	0.40	0.51	0.36	0.47	0.25
11	MSC_RAD	0.81	0.60	0.46	0.63	0.54	0.46	0.40
12	MCAS_MB	0.75	0.53	0.40	0.45	0.53	0.42	0.40
13	MCAS_RAD	0.77	0.67	0.44	0.49	0.54	0.56	0.27

Mejor modelo en el sexto día

Durante el análisis de clasificación supervisada, el modelo más preciso identificado fue el Análisis Lineal Discriminante (ALD), que aplicó el método de suavizamiento corrección de dispersión multiplicativa (MSC) y la técnica de eliminación de atípicos MB, logrando una notable exactitud del 86,79 %. Sin embargo, es crucial someter este modelo a pruebas que evalúen la normalidad multivariada y la homogeneidad en las matrices de covarianzas.

La prueba de normalidad multivariada, llevada a cabo mediante la prueba de Royston, reveló una estadística H de 113,62 y un p-valor de 4,903e-25, indicando evidencia estadística, con un nivel de significancia del 5 %, para rechazar la hipótesis nula de normalidad multivariada. En contraste, la prueba de homogeneidad en la varianza de Box's M-test resultó en una estadística chi cuadrado aproximada de 1401,3 con 840 grados de libertad y un p-valor de 2,2e-16. En este caso, se rechaza la hipótesis nula de homogeneidad en las varianzas de los tratamientos, manteniendo un nivel de significancia del 5 %. Por lo tanto, no se cumplen los supuestos de normalidad ni de homogeneidad en las varianzas.

A pesar de la falta de cumplimiento de la hipótesis de matrices de covarianzas iguales y de normalidad multivariada, se destaca que las precisiones más altas se obtienen con el modelo ALD.

Posteriormente, se examinaron las probabilidades a priori de los diferentes tratamientos. Se

observa que las variaciones son mínimas, lo cual se atribuye al número reducido de atípicos eliminados en cada caso, indicando una relativa estabilidad en las probabilidades a priori.Tabla. probabilidades a priori para el ALD en el conjunto de datos MSC_MB para el día 6. (ver tabla **6-11**)

Tabla 6-11: Probabilidades a priori para el ALD en el conjunto de datos MCS_MB para el día 6

Control	E_Hidrico	Fus_EH	Fus_EH_Ral	Fusarium	Ral_EH	Ral_Fus	Ralstonia
0,11	0,12	0,10	0,12	0,11	0,14	0,12	0,13

La tabla **6-12** presenta los coeficientes vinculados a las 7 ecuaciones lineales discriminantes. Cada ecuación, formulada como una combinación lineal de las longitudes de onda correspondientes, desempeña un papel crucial en la toma de decisiones al situar los nuevos valores espectroscópicos calculados en regiones específicas. Estas ecuaciones lineales son esenciales para predecir el tratamiento más probable para una nueva observación, proporcionando un marco analítico que aprovecha la información espectral para realizar inferencias precisas y contextualizadas según los distintos tratamientos experimentales. Al examinar detenidamente estos coeficientes, se obtiene una comprensión más profunda de cómo cada longitud de onda contribuye al proceso de clasificación, lo que facilita una interpretación informada de los resultados obtenidos.

Tabla 6-12: Coeficientes de las ecuaciones lineales discriminantes para el conjunto de datos MCS_MB para el día 6

L. de Onda	**LD1**	**LD2**	**LD3**	**LD4**	**LD5**	**LD6**	**LD7**
350nm	-195,12	-138,63	-66,22	50,25	-24,91	35,78	-81,93
352nm	-171,70	-217,96	215,01	121,12	124,64	85,46	63,14
353nm	-23,97	159,40	-69,23	-109,64	-222,65	-16,93	15,92
355nm	193,25	310,42	-15,67	-201,04	160,20	-129,40	-5,63
581nm	286,17	322,42	591,88	353,47	632,34	412,62	-69,57
609nm	-232,46	-325,04	-1031,47	-37,10	-732,21	-668,04	-11,54
703nm	914,20	-4702,62	-3110,02	-178,71	10785,33	-4050,81	30,59
705nm	-323,95	15178,99	10789,85	2018,40	-26436,88	11544,92	-4070,73
707nm	-1811,61	-14626,71	-10349,74	-3862,96	19988,38	-9984,72	6435,22
710nm	1172,01	4136,73	2758,42	1834,36	-4195,46	2455,14	-2541,42
1300nm	92,61	194,66	279,39	145,77	52,59	65,66	-54,01
1383nm	-142,31	2,94	-247,64	-47,19	-136,43	-5,74	-126,76
1643nm	2,68	-215,17	12,93	-538,51	-148,65	138,71	-298,16
1750nm	105,54	240,76	24,89	731,71	207,34	-269,63	332,44
2150nm	-241,31	31,65	-63,62	-269,97	201,02	-243,51	-2,96

El rendimiento del modelo de clasificación destaca con una exactitud del 86,79 %, respaldada por un intervalo de confianza del 95 % que varía entre el 74,66 % y el 94,52 %. Esta destacada capacidad de clasificación se logra mediante la combinación estratégica del Análisis Lineal Discriminante (ALD) con los métodos de suavizamiento corrección de dispersión multiplicativa (MSC) y la gestión de atípicos mediante el MB.

Este modelo no solo demuestra una efectiva capacidad de clasificación, sino que también exhibe una utilidad significativa en la detección temprana de estrés en plantas. En el análisis del día 6 de medición, se observa una precisa predicción de cada tratamiento, manteniendo un bajo error de clasificación del 13,21 %. Este descubrimiento resalta la eficacia del modelo en identificar patrones tempranos de estrés, ofreciendo así una herramienta valiosa para la toma de decisiones y la implementación de medidas correctivas en el ámbito de la fitopatología.

Tabla 6-13: Matriz de confusión para el ALD en el conjunto de datos MCS_MB para el día 6

	Referencia							
Predicción	Control	E_Hidrico	Fus_EH	Fus_EH_Ral	Fus	Ral_EH	Ral_Fus	Rals
Control	7	0	0	0	0	0	0	0
E_Hidrico	0	6	0	0	0	0	0	1
Fus_EH	0	0	2	0	0	0	0	0
Fus_EH_Ral	0	0	1	5	0	0	0	0
Fusarium	0	0	0	0	6	1	0	0
Ral_EH	0	0	0	1	0	7	0	0
Ral_Fus	0	0	0	2	1	0	8	0
Ralstonia	0	0	0	0	0	0	0	5

Mejor modelo parsimonioso en el sexto día

Considerando la complejidad asociada a la interpretación de siete tratamientos y quince longitudes de onda, se busca simplificar el análisis discriminante lineal presentado anteriormente. La estrategia adoptada apunta a reducir tanto el número de tratamientos como el de variables con el fin de mejorar la claridad y comprensión del análisis.

En la reducción de tratamientos, se decide enfocarse en los grupos fundamentales, excluyendo las interacciones. Esto implica considerar exclusivamente el grupo control, el estrés hídrico, las plantas sometidas solo a la bacteria Ralstonia y las plantas sometidas solo al hongo Fusarium. Este enfoque simplificado facilita la interpretación y permite un análisis más detallado de cada tratamiento individual.

Simultáneamente, para abordar la reducción en el número de variables, se emplea el algoritmo RELIEF para seleccionar las más importantes. Se eligen las cinco longitudes de onda más relevantes: 350 nm, 609 nm, 705 nm, 1383 nm y 1750 nm. Este paso contribuye a optimizar la eficiencia del modelo al centrarse en las características espectrales más informativas.

En el análisis resultante, se exploran las probabilidades a priori de los tratamientos seleccionados (ver **6-14**. Se destaca la mínima variación observada, atribuible al reducido número de atípicos eliminados en cada caso. Este fenómeno refleja una relativa estabilidad en las probabilidades a priori, respaldando la confiabilidad y consistencia de los resultados obtenidos con este enfoque simplificado.

Tabla 6-14: Probabilidades a priori para el ALD parsimonioso en el conjunto de datos MSC_MB para el día 6

Control	E_Hidrico	Fusarium	Ralstonia
0,24	0,25	0,24	0,28

La tabla **6-15** presenta los coeficientes de las tres ecuaciones lineales discriminantes en el modelo parsimonioso. Estas ecuaciones, expresadas como combinaciones lineales de las longitudes de onda, son fundamentales para asignar ubicaciones a nuevos valores espectroscópicos y predecir tratamientos. Analizar detalladamente estos coeficientes proporciona una comprensión profunda de cómo cada longitud de onda influye en la asignación de tratamientos, enriqueciendo la interpretación de los datos en el contexto de los diversos tratamientos experimentales.
Tabla. Coeficientes de las ecuaciones lineales discriminantes del modelo ALD parsimonioso para el conjunto de datos MCS_MB para el día 6

Tabla 6-15: Coeficientes de las ecuaciones lineales discriminantes del modelo ALD parsimonioso para el conjunto de datos MCS_MB para el día 6

Longitudes de onda	LD1	LD2	LD3
350nm	240,90	-46,37	6,23
609nm	-61,81	390,76	-180,30
705nm	25,93	27,45	-18,25
1383nm	101,40	56,91	-19,75
1750nm	31,55	220,54	-10,01

El modelo de clasificación logra una destacada exactitud del 80,77 %, respaldada por un intervalo de confianza del 95 % entre el 60,65 % y el 93,45 %. Esta eficiencia se alcanza mediante la combinación estratégica de Análisis Lineal Discriminante (ALD) con métodos de suavizamiento mínimos cuadrados asimétricos (ALS) y la gestión de atípicos a través del

MB, aplicados a solo cuatro tratamientos y cinco variables. Aunque esta exactitud disminuye en comparación con el modelo parsimonioso del día 3, sigue siendo alta.

Este modelo no solo muestra una efectiva capacidad de clasificación, sino que también se destaca por su utilidad en la detección temprana de estrés en plantas. En el análisis del día 6 de medición, demuestra una capacidad predictiva precisa para cada tratamiento, manteniendo un bajo error de clasificación del 19,3 %. Este resultado resalta la eficacia del modelo para identificar patrones tempranos de estrés, proporcionando así una herramienta valiosa para la toma de decisiones y la implementación de medidas correctivas en la salud de las plantas.

Tabla 6-16: Matriz de confusión del para el ALD en el conjunto de datos MCS_MB del modelo ALD parsimonioso para el día 6

	Referencia			
Predicción	Control	E_Hidrico	Fusarium	Ralstonia
Control	7	0	0	1
E_Hidrico	0	5	1	2
Fusarium	0	0	6	0
Ralstonia	0	1	0	3

En el gráfico **6-8**, se observa claramente que cada par de longitudes de onda consideradas en el análisis lineal discriminante delinean regiones específicas de manera distintiva para cada uno de los tratamientos examinados. Es destacable que, en la mayoría de estos planos, las plantas de control se encuentran en un extremo, mientras que en el extremo opuesto se ubican las plantas inoculadas con Fusarium. En el espacio intermedio se distribuyen tanto las plantas sometidas a estrés hídrico como aquellas inoculadas con Ralstonia. Este patrón visual resalta la capacidad del modelo para generar una separación clara y discernible entre los diferentes tratamientos, ofreciendo así una representación gráfica efectiva de la discriminación lograda a través de las longitudes de onda seleccionadas.

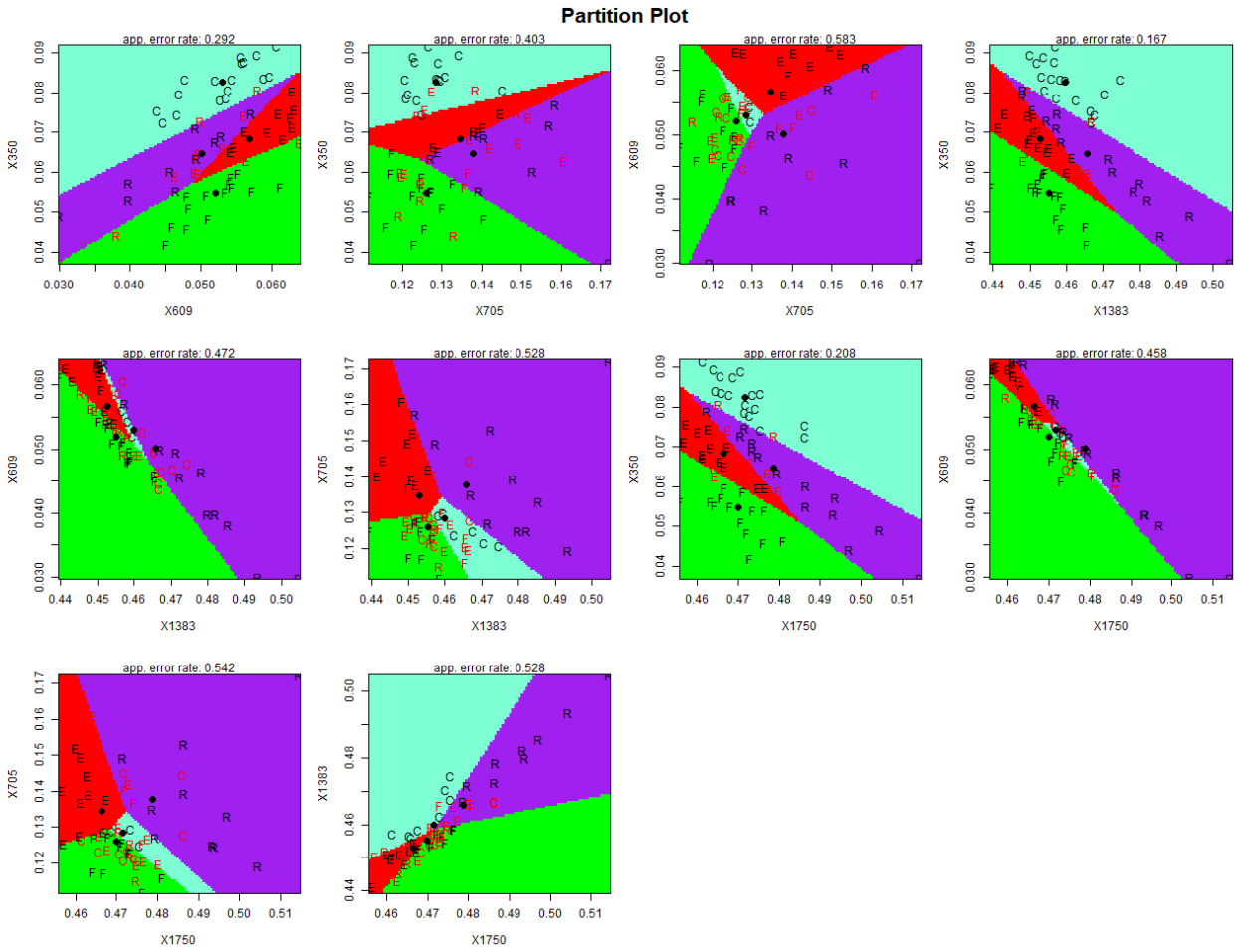

Figura 6-8: Diagramas de regiones discriminantes según longitudes de onda seleccionadas para el ALD parsimonioso para el día 6

Finalmente, se procedió a evaluar este análisis utilizando un conjunto específico de longitudes de onda: 350 nm, 609 nm, 705 nm, 1383 nm y 1750 nm, y haciendo referencia a los datos en su estado crudo. Los resultados exhibieron una destacada exactitud del 80,64 %, lo cual representa una mínima disminución del 0,13 % en comparación con la aplicación de la técnica de suavizamiento MCS parsimonioso, combinada con la eliminación de datos atípicos mediante la MB. Esta ligera variación indica que el procedimiento de suavizamiento y eliminación de datos atípicos tiene un impacto insignificante en este caso, sugiriendo que los datos pueden ser manejados sin modificar según la presentación original del espectrofotómetro.

7 Conclusiones y recomendaciones

7.1. Conclusiones

Estas conclusiones consolidan el valor del análisis espectral y las metodologías aplicadas para la detección temprana de estrés en plantas, proporcionando información esencial para la gestión de cultivos y la mitigación de enfermedades en el cultivo de Banano Gros Michel.

- El análisis descriptivo espectral de las plantas de Banano Gros Michel revela patrones distintivos de reflectancia. La baja reflectancia inicial, posiblemente relacionada con la absorción de luz por pigmentos como la clorofila, contrasta con un pico destacado alrededor de los 550 nm. Este pico sugiere respuestas específicas de la planta hacia el color verde, siendo crucial para la detección temprana de estrés en las plantas. Además, la observación de reflectancias más bajas en el espectro de las plantas afectadas por Ralstonia y por el Fusarium destaca una respuesta espectral distintiva, ofreciendo pistas valiosas para la identificación y caracterización del estrés asociado con esta condición experimental.

- Las longitudes de onda más distintivas entre grupos para el día 3 fueron 350, 353, 355, 1070, 1050, 1048, 1076, 1077, 1042, 1084, 1275, 1479, 1600, 1801 y 2167 nanómetros. Aunque, se realizaron análisis para cada uno de los 13 conjuntos de datos y se observa que estas longitudes de onda en cada conjunto son muy cercanas entre sí. En contraste, las longitudes de onda más discriminantes entre los tratamientos para el día 6 fueron 350, 352, 353, 355, 703, 705, 707, 710, 581, 609, 1300, 1383, 1643, 1750 y 2150 nanómetros. Estos resultados proporcionan información detallada sobre las características espectrales distintivas en diferentes días, contribuyendo a una comprensión más completa de las respuestas de las plantas a diversas condiciones experimentales.

- La metodología de preprocesamiento más óptima se encuentra mediante el suavizamiento con mínimos cuadrados asimétricos junto con la eliminación de atípicos utilizando el MB. Este preprocesamiento aplicado con un ALD resulta en una alta precisión del 86 % en el día 3 post-inoculación, demostrando ser precisa para detectar tempranamente el estrés hídrico y las enfermedades de *Ralstonia solanacearum raza 2* y *Fusarium oxysporum raza 4*. En el día 6, no se encuentran diferencias significativas al realizar dis-

tintos preprocesamientos, manteniendo los datos sin modificaciones como los entrega el espectrofotómetro.

- La metodología de preprocesamiento más óptima se encuentra mediante el suavizamiento con Mínimos Cuadrados Asimétricos junto con la eliminación de atípicos utilizando el método Bag. Este preprocesamiento aplicado con un ALD resulta en una alta precisión del 86 % en el día 3 post-inoculación, demostrando ser precisa para detectar tempranamente el estrés hídrico y las enfermedades de Ralstonia solanacearum raza 2 y Fusarium oxysporum raza 4. En el día 6, no se encuentran diferencias significativas al realizar distintos preprocesamientos, manteniendo los datos sin modificaciones como los entrega el espectrofotómetro.

- Se determina que el ALD es la mejor metodología estadística tanto en el día 3 como en el día 6, a pesar de que en algunos casos no se cumplen los supuestos de normalidad y homogeneidad en las varianzas. Otras técnicas, como los bosques aleatorios, presentan sobreajuste en el modelo de clasificación al elegir el número óptimo de árboles. Además, las técnicas de redes neuronales, como MPL, requieren un mayor tamaño de muestra, incluso al probar con distintas cantidades de capas ocultas y neuronas.

7.2. Recomendaciones

Como trabajos futuros se recomienda lo siguiente:

- Para mejorar la medida de exactitud se recomienda implementar métodos como K-Fold de validación cruzada, para reducir la dependencia de la división de conjuntos de entrenamiento y pruba, en el cálculo de esta medida.

- Análisis de Datos Funcionales:
  Dada la disponibilidad de 14 días de mediciones en el experimento, se sugiere realizar análisis de datos funcionales para discernir entre los 8 tratamientos. Este enfoque permitiría capturar las variaciones a lo largo del tiempo y proporcionar una comprensión más profunda del comportamiento experimental.

- Clasificación No Supervisada por Severidad:
  Recomendamos llevar a cabo análisis de clasificación no supervisada para evaluar la severidad de la infección en cada día de medición dentro de cada tratamiento. Esto sería especialmente beneficioso dada la variabilidad observada en el desarrollo de síntomas entre las plantas del mismo grupo experimental.

- Análisis Longitudinal de Severidad:
  Se propone realizar un análisis longitudinal para clasificar la severidad de la enfermedad en función de los días posteriores a la inoculación. Este enfoque proporcionaría

información valiosa sobre la evolución de la infección a lo largo del tiempo y permitiría identificar patrones temporales significativos.

- Establecimiento de Valores de Referencia:
  Para una detección temprana de anomalías, se sugiere establecer valores de referencia basados en las longitudes de onda más relevantes mediante un análisis estadístico multivariado. Utilizando la teoría de la normalidad, se puede realizar una prueba de hipótesis e intervalos de confianza, especialmente en los controles, para cada día, considerando la correlación de los datos. De esta manera, valores de reflectancia fuera del intervalo de confianza podrían indicar irregularidades en la planta medida.

Bibliografía

[Abdulridha et al., 2016] Abdulridha, J., Ehsani, R., and De Castro, A. (2016). Detection and differentiation between laurel wilt disease, phytophthora disease, and salinity damage using a hyperspectral sensing technique. *Agriculture*, 6(4):56.

[Abu-Khalaf, 2015] Abu-Khalaf, N. (2015). Sensing tomato's pathogen using visible/near infrared (vis/nir) spectroscopy and multivariate data analysis (mvda). *Palest. Tech. Univ. Res. J.*, 3(1):12–22.

[Anderson and Gupta, 2009] Anderson, H. S. and Gupta, M. R. (2009). Classifying linear system outputs by robust local bayesian quadratic discriminant analysis on linear estimators. In *2009 IEEE/SP 15th Workshop on Statistical Signal Processing*, pages 789–792. IEEE.

[Balakrishnama and Ganapathiraju, 1998] Balakrishnama, S. and Ganapathiraju, A. (1998). Linear discriminant analysis-a brief tutorial. *Institute for Signal and information Processing*, 18(1998):1–8.

[Basa, 2022] Basa, J. (2022). Big data para quimiometría: Distribución asintótica del estimador pls en alta dimensión.

[Berrar, 2019] Berrar, D. (2019). Bayes' theorem and naive bayes classifier.

[Bienkowski et al., 2019] Bienkowski, D., Aitkenhead, M. J., Lees, A. K., Gallagher, C., and Neilson, R. (2019). Detection and differentiation between potato (solanum tuberosum) diseases using calibration models trained with non-imaging spectrometry data. *Computers and Electronics in Agriculture*, 167:105056.

[Bishop, 2006] Bishop, C. (2006). Pattern recognition and machine learning. *Springer google schola*, 2:531–537.

[Bishop et al., 1995] Bishop, C. M. et al. (1995). *Neural networks for pattern recognition*. Oxford university press.

[Box, 1953] Box, G. E. (1953). Non-normality and tests on variances. *Biometrika*, 40(3/4):318–335.

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

[Brown et al., 2000] Brown, C. D., Vega-Montoto, L., and Wentzell, P. D. (2000). Derivative preprocessing and optimal corrections for baseline drift in multivariate calibration. *Applied Spectroscopy*, 54(7):1055–1068.

[Buja et al., 1989] Buja, A., Hastie, T., and Tibshirani, R. (1989). Linear smoothers and additive models. *The Annals of Statistics*, pages 453–510.

[Choi and Marron, 2019] Choi, H. Y. and Marron, J. (2019). Theory of high-dimensional outliers. *arXiv preprint arXiv:1909.02139*.

[Cortes and Vapnik, 1995] Cortes, C. and Vapnik, V. (1995). Support vector machine. *Machine learning*, 20(3):273–297.

[de Carvalho et al., 2015] de Carvalho, G. G. A., Moros, J., Santos Jr, D., Krug, F. J., and Laserna, J. J. (2015). Direct determination of the nutrient profile in plant materials by femtosecond laser-induced breakdown spectroscopy. *Analytica chimica acta*, 876:26–38.

[Dupas et al., 2019] Dupas, E., Legendre, B., Olivier, V., Poliakoff, F., Manceau, C., and Cunty, A. (2019). Comparison of real-time pcr and droplet digital pcr for the detection of xylella fastidiosa in plants. *Journal of microbiological methods*, 162:86–95.

[el Financiero, 2019] el Financiero, P. (2019). Fusarium raza 4 tropical mantiene en vilo a los bananeros. Fecha de acceso: 02/07/2023.

[Espectador, 2019] Espectador, P. E. (2019). Ica firma acuerdos con asociaciones bananeras para controlar hongo fusarium. Fecha de acceso: 27/08/2023.

[FAO, 2017] FAO (2017). Manual de seguridad y salud en la industria bananera. Fecha de acceso: 01/10/2023.

[FAO, 2019] FAO (2019). La marchitez del banano por fusarium raza 4 tropical: ¿una creciente amenaza al mercado mundial del banano? Fecha de acceso: 20/10/2022.

[FAO, 2021] FAO (2021). Análisis del mercado del banano, resultados preliminares 2020. Fecha de acceso: 28/10/2022.

[Farber et al., 2019a] Farber, C., Mahnke, M., Sanchez, L., and Kurouski, D. (2019a). Advanced spectroscopic techniques for plant disease diagnostics. a review. *TrAC Trends in Analytical Chemistry*, 118:43–49.

[Farber et al., 2019b] Farber, C., Shires, M., Ong, K., Byrne, D., and Kurouski, D. (2019b). Raman spectroscopy as an early detection tool for rose rosette infection. *Planta*, 250(4):1247–1254.

[Fegan and Prior, 2006] Fegan, M. and Prior, P. (2006). Diverse members of the ralstonia solanacearum species complex cause bacterial wilts of banana. *Australasian Plant Pathology*, 35:93–101.

[García-Bastidas et al., 2020] García-Bastidas, F., Quintero-Vargas, J., Ayala-Vasquez, M., Schermer, T., Seidl, M., Santos-Paiva, M., Noguera, A., Aguilera-Galvez, C., Wittenberg, A., Hofstede, R., et al. (2020). First report of fusarium wilt tropical race 4 in cavendish bananas caused by fusarium odoratissimum in colombia. *Plant disease*, 104(3):994–994.

[Gazalba et al., 2017] Gazalba, I., Reza, N. G. I., et al. (2017). Comparative analysis of k-nearest neighbor and modified k-nearest neighbor algorithm for data classification. In *2017 2nd international conferences on information technology, information systems and electrical engineering (ICITISEE)*, pages 294–298. IEEE.

[Genin and Denny, 2012] Genin, S. and Denny, T. P. (2012). Pathogenomics of the ralstonia solanacearum species complex. *Annual review of phytopathology*, 50:67–89.

[Gold and Sollich, 2003] Gold, C. and Sollich, P. (2003). Model selection for support vector machine classification. *Neurocomputing*, 55(1-2):221–249.

[Gomez et al., 2008] Gomez, C., Rossel, R. A. V., and McBratney, A. B. (2008). Soil organic carbon prediction by hyperspectral remote sensing and field vis-nir spectroscopy: An australian case study. *Geoderma*, 146(3-4):403–411.

[Gull et al., 2019] Gull, A., Lone, A. A., and Wani, N. U. I. (2019). Biotic and abiotic stresses in plants. *Abiotic and biotic stress in plants*, pages 1–19.

[Guo et al., 2003] Guo, G., Wang, H., Bell, D., Bi, Y., and Greer, K. (2003). Knn model-based approach in classification. In *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings*, pages 986–996. Springer.

[Hanusz et al., 2018] Hanusz, Z., Enomoto, R., Seo, T., and Koizumi, K. (2018). A monte carlo comparison of jarque–bera type tests and henze–zirkler test of multivariate normality. *Communications in Statistics-Simulation and Computation*, 47(5):1439–1452.

[Hou et al., 2022] Hou, B., Hu, Y., Zhang, P., and Hou, L. (2022). Potato late blight severity and epidemic period prediction based on vis/nir spectroscopy. *Agriculture*, 12(7):897.

[(ICA), 2020] (ICA), I. C. A. (2020). Fusarium r4t. Fecha de acceso: 27/08/2023.

[Ignat et al., 2022] Ignat, T., Shavit, Y., Rachmilevitch, S., and Karnieli, A. (2022). Spectral monitoring of salinity stress in tomato plants. *Biosystems Engineering*, 217:26–40.

[Jie et al., 2009] Jie, L., Zifeng, W., Lixiang, C., Hongming, T., Patrik, I., Zide, J., and Shining, Z. (2009). Artificial inoculation of banana tissue culture plantlets with indigenous endophytes originally derived from native banana plants. *Biological control*, 51(3):427–434.

[Kaliramesh et al., 2013] Kaliramesh, S., Chelladurai, V., Jayas, D., Alagusundaram, K., White, N., and Fields, P. (2013). Detection of infestation by callosobruchus maculatus in mung bean using near-infrared hyperspectral imaging. *Journal of Stored Products Research*, 52:107–111.

[Khaled et al., 2018a] Khaled, A. Y., Abd Aziz, S., Bejo, S. K., Nawi, N. M., and Seman, I. A. (2018a). Spectral features selection and classification of oil palm leaves infected by basal stem rot (bsr) disease using dielectric spectroscopy. *Computers and Electronics in Agriculture*, 144:297–309.

[Khaled et al., 2018b] Khaled, A. Y., Abd Aziz, S., Bejo, S. K., Nawi, N. M., Seman, I. A., and Onwude, D. I. (2018b). Early detection of diseases in plant tissue using spectroscopy–applications and limitations. *Applied Spectroscopy Reviews*, 53(1):36–64.

[Kira and Rendell, 1992] Kira, K. and Rendell, L. A. (1992). The feature selection problem: Traditional methods and a new algorithm. In *Proceedings of the tenth national conference on Artificial intelligence*, pages 129–134.

[Klap et al., 2020] Klap, C., Luria, N., Smith, E., Bakelman, E., Belausov, E., Laskar, O., Lachman, O., Gal-On, A., and Dombrovsky, A. (2020). The potential risk of plant-virus disease initiation by infected tomatoes. *Plants*, 9(5):623.

[Koc et al., 2020] Koc, G., Fidan, H., Sari, N., and ÇALIŞ, O. (2020). A comparative study on apple chlorotic leafspot virus (aclsv) isolates from different hosts in the east mediterranean region of turkey. *APPLIED ECOLOGY AND ENVIRONMENTAL RESEARCH*, 18(1):141–157.

[Kononenko, 1994] Kononenko, I. (1994). Estimating attributes: Analysis and extensions of relief. In *European conference on machine learning*, pages 171–182. Springer.

[Learning, 1997] Learning, M. (1997). Tom mitchell. *Publisher: McGraw Hill*.

[LeCun et al., 1989] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551.

[Li et al., 2014] Li, M.-H., Xie, X.-L., Lin, X.-F., Shi, J.-X., Ding, Z.-J., Ling, J.-F., Xi, P.-G., Zhou, J.-N., Leng, Y., Zhong, S., et al. (2014). Functional characterization of the gene fooch1 encoding a putative α-1, 6-mannosyltransferase in fusarium oxysporum f. sp. cubense. *Fungal Genetics and Biology*, 65:1–13.

[Li et al., 2008] Li, R., Mock, R., Huang, Q., Abad, J., Hartung, J., and Kinard, G. (2008). A reliable and inexpensive method of nucleic acid extraction for the pcr-based detection of diverse plant pathogens. *Journal of Virological Methods*, 154(1-2):48–55.

[Lipton et al., 2014] Lipton, Z. C., Elkan, C., and Narayanaswamy, B. (2014). Thresholding classifiers to maximize f1 score. *arXiv preprint arXiv:1402.1892*.

[Luana et al., 2015] Luana, G., Fabiano, S., Fabio, G., and Paolo, G. (2015). Comparing visual inspection of trees and molecular analysis of internal wood tissues for the diagnosis of wood decay fungi. *Forestry: An International Journal of Forest Research*, 88(4):465–470.

[Macias-Echeverri et al., 2022] Macias-Echeverri, E., Hoyos-Carvajal, L. M., Botero-Fernández, V., Zapata-Henao, S., and Marín-Ortiz, J. C. (2022). Spectral behavior of banana with foc r1 infection: Analysis of williams and gros michel clones. *Agronomía Colombiana*, 40(3).

[Madihah et al., 2014] Madihah, A., Idris, A., and Rafidah, A. (2014). Polyclonal antibodies of ganoderma boninense isolated from malaysian oil palm for detection of basal stem rot disease. *African Journal of Biotechnology*, 13(34).

[Manzo-Sánchez et al., 2014] Manzo-Sánchez, G., Orozco-Santos, M., Martínez-Bolaños, L., Garrido-Ramírez, E., and Canto-Canche, B. (2014). Enfermedades de importancia cuarentenaria y económica del cultivo de banano (musa sp.) en méxico. *Revista mexicana de fitopatología*, 32(2):89–107.

[Marín-Ortiz et al., 2020] Marín-Ortiz, J. C., Gutierrez-Toro, N., Botero-Fernández, V., and Hoyos-Carvajal, L. M. (2020). Linking physiological parameters with visible/near-infrared leaf reflectance in the incubation period of vascular wilt disease. *Saudi Journal of Biological Sciences*, 27(1):88–99.

[Martens et al., 1983] Martens, H., Jensen, S., and Geladi, P. (1983). Multivariate linearity transformation for near-infrared reflectance spectrometry. In *Proceedings of the Nordic symposium on applied statistics*, pages 205–234. Stokkand Forlag Publishers Stavanger, Norway.

[Monroy and Rivera, 2012] Monroy, L. G. D. and Rivera, M. A. M. (2012). *Análisis estadístico de datos multivariados*. Universidad Nacional de Colombia.

[Montoya Rios et al., 2022] Montoya Rios, D. P., Molano Prieto, O. J., et al. (2022). Análisis de producción, rendimiento y exportación de banano en los principales países afectados por el hongo fusarium oxysporum f. sp. cubense (foc r4t) y recomendaciones para colombia.

[Morellos et al., 2020] Morellos, A., Tziotzios, G., Orfanidou, C., Pantazi, X. E., Sarantaris, C., Maliogka, V., Alexandridis, T. K., and Moshou, D. (2020). Non-destructive early

detection and quantitative severity stage classification of tomato chlorosis virus (tocv) infection in young tomato plants using vis–nir spectroscopy. *Remote Sensing*, 12(12):1920.

[Mosa et al., 2017] Mosa, K. A., Ismail, A., and Helmy, M. (2017). Introduction to plant stresses. In *Plant stress tolerance*, pages 1–19. Springer.

[Müller and Guido, 2016] Müller, A. C. and Guido, S. (2016). *Introduction to machine learning with Python: a guide for data scientists*. .ºReilly Media, Inc.".

[Muncan et al., 2022] Muncan, J., Jinendra, B. M. S., Kuroki, S., and Tsenkova, R. (2022). Aquaphotomics research of cold stress in soybean cultivars with different stress tolerance ability: Early detection of cold stress response. *Molecules*, 27(3):744.

[Murphy, 2012] Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

[Newey and Powell, 1987] Newey, W. K. and Powell, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica: Journal of the Econometric Society*, pages 819–847.

[Pal, 2005] Pal, M. (2005). Random forest classifier for remote sensing classification. *International journal of remote sensing*, 26(1):217–222.

[Pavia et al., 2014] Pavia, D. L., Lampman, G. M., Kriz, G. S., and Vyvyan, J. A. (2014). *Introduction to spectroscopy*. Cengage learning.

[Ploetz, 2006] Ploetz, R. C. (2006). Fusarium wilt of banana is caused by several pathogens referred to as fusarium oxysporum f. sp. cubense. *Phytopathology*, 96(6):653–656.

[Ploetz, 2015] Ploetz, R. C. (2015). Fusarium wilt of banana. *Phytopathology*, 105(12):1512–1521.

[R. Beghi and Guidetti, 2017] R. Beghi, V. Giovenzana, L. B. and Guidetti, R. (2017). Rapid evaluation of grape phytosanitary status directly at the check point station entering the winery by using visible/near infrared spectroscopy. *Journal of Food Engineering*, 204:46–54.

[Ramsay and Silverman, 2002] Ramsay, J. O. and Silverman, B. W. (2002). *Applied functional data analysis: methods and case studies*. Springer.

[Reyes-Matamoros et al., 2014] Reyes-Matamoros, J., Martínez-Moreno, D., Rueda-Luna, R., and Rodríguez-Ramírez, T. (2014). Efecto del estrés hídrico en plantas de frijol (phaseolus vulgaris l.) en condiciones de invernadero. *Revista Iberoamericana de Ciencias*, 1(2):191–203.

[Rinnan et al., 2009] Rinnan, Å., Van Den Berg, F., and Engelsen, S. B. (2009). Review of the most common pre-processing techniques for near-infrared spectra. *TrAC Trends in Analytical Chemistry*, 28(10):1201–1222.

[Roa Martínez and Loaiza Correa, 2011] Roa Martínez, S. M. and Loaiza Correa, H. (2011). Evaluation of techniques for relevance analysis of radiological images using filters. *Revista Ingeniería Biomédica*, 5(9):26–34.

[Rousseeuw et al., 1999] Rousseeuw, P. J., Ruts, I., and Tukey, J. W. (1999). The bagplot: a bivariate boxplot. *The American Statistician*, 53(4):382–387.

[Rumpf et al., 2010] Rumpf, T., Mahlein, A.-K., Steiner, U., Oerke, E.-C., Dehne, H.-W., and Plümer, L. (2010). Early detection and classification of plant diseases with support vector machines based on hyperspectral reflectance. *Computers and electronics in agriculture*, 74(1):91–99.

[Salazar et al., 2014] Salazar, E., Trujillo, I., Macías, M. P., Gutiérrez, M. A., Castro, L., Vallejo, E., and Torrealba, M. (2014). Respuesta fisiológica al estrés hídrico de plantas de banano cv.pineo gigante'(musa aaa) regeneradas a partir de yemas irradiadas. *Biotecnología Vegetal*, 14(3).

[Sankaran et al., 2010] Sankaran, S., Mishra, A., Ehsani, R., and Davis, C. (2010). A review of advanced techniques for detecting plant diseases. *Computers and electronics in agriculture*, 72(1):1–13.

[Savitzky and Golay, 1964] Savitzky, A. and Golay, M. J. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639.

[Schölkopf and Smola, 2002] Schölkopf, B. and Smola, A. J. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

[Shin et al., 2023] Shin, M.-Y., Viejo, C. G., Tongson, E., Wiechel, T., Taylor, P. W., and Fuentes, S. (2023). Early detection of verticillium wilt of potatoes using near-infrared spectroscopy and machine learning modeling. *Computers and Electronics in Agriculture*, 204:107567.

[Sun and Wu, 2008] Sun, Y. and Wu, D. (2008). A relief based feature extraction algorithm. In *Proceedings of the 2008 SIAM International Conference on Data Mining*, pages 188–195. SIAM.

[Svanberg, 2012] Svanberg, S. (2012). *Atomic and molecular spectroscopy: basic aspects and practical applications*, volume 6. Springer Science & Business Media.

[Tjandra Nugraha et al., 2021] Tjandra Nugraha, D., Zinia Zaukuu, J.-L., Aguinaga Bósquez, J. P., Bodor, Z., Vitalis, F., and Kovacs, Z. (2021). Near-infrared spectroscopy and aquaphotomics for monitoring mung bean (vigna radiata) sprout growth and validation of ascorbic acid content. *Sensors*, 21(2):611.

[Tu et al., 2022] Tu, Y.-K., Kuo, C.-E., Fang, S.-L., Chen, H.-W., Chi, M.-K., Yao, M.-H., and Kuo, B.-J. (2022). A 1d-sp-net to determine early drought stress status of tomato (solanum lycopersicum) with imbalanced vis/nir spectroscopy data. *Agriculture*, 12(2):259.

[Tunsagool et al., 2019] Tunsagool, P., Jutidamrongphan, W., Phaonakrop, N., Jaresitthikunchai, J., Roytrakul, S., and Leelasuphakul, W. (2019). Insights into stress responses in mandarins triggered by bacillus subtilis cyclic lipopeptides and exogenous plant hormones upon penicillium digitatum infection. *Plant cell reports*, 38(5):559–575.

[Urbanowicz et al., 2018] Urbanowicz, R. J., Meeker, M., La Cava, W., Olson, R. S., and Moore, J. H. (2018). Relief-based feature selection: Introduction and review. *Journal of biomedical informatics*, 85:189–203.

[Visa et al., 2011] Visa, S., Ramsay, B., Ralescu, A. L., and Van Der Knaap, E. (2011). Confusion matrix-based feature selection. *Maics*, 710(1):120–127.

[Walsh et al., 2020] Walsh, K. B., Blasco, J., Zude-Sasse, M., and Sun, X. (2020). Visible-nir 'point'spectroscopy in postharvest fruit and vegetable assessment: The science behind three decades of commercial use. *Postharvest Biology and Technology*, 168:111246.

[Wang et al., 2020] Wang, D., Peng, C., Zheng, X., Chang, L., Xu, B., and Tong, Z. (2020). Secretome analysis of the banana fusarium wilt fungi foc r1 and foc tr4 reveals a new effector oastl required for full pathogenicity of foc tr4 in banana. *Biomolecules*, 10(10):1430.

[Yu et al., 2021] Yu, K., Fang, S., and Zhao, Y. (2021). Heavy metal hg stress detection in tobacco plant using hyperspectral sensing and data-driven machine learning methods. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 245:118917.

[Zahir et al., 2022] Zahir, S. A. D. M., Omar, A. F., Jamlos, M. F., Azmi, M. A. M., and Muncan, J. (2022). A review of visible and near-infrared (vis-nir) spectroscopy application in plant stress detection. *Sensors and Actuators A: Physical*, page 113468.

[Zhang et al., 2019] Zhang, J., Huang, Y., Pu, R., Gonzalez-Moreno, P., Yuan, L., Wu, K., and Huang, W. (2019). Monitoring plant diseases and pests through remote sensing technology: A review. *Computers and Electronics in Agriculture*, 165:104943.

[Zhang et al., 2012] Zhang, J., Pu, R., Huang, W., Yuan, L., Luo, J., and Wang, J. (2012). Using in-situ hyperspectral data for detecting and discriminating yellow rust disease from nutrient stresses. *Field Crops Research*, 134:165–174.

[Zhang et al., 2021] Zhang, W., Zhang, W., Yang, Y., Hu, G., Ge, D., Liu, H., Cao, H., et al. (2021). A cloud computing-based approach using the visible near-infrared spectrum to classify greenhouse tomato plants under water stress. *Computers and Electronics in Agriculture*, 181:105966.