

UNIVERSIDAD DE SALAMANCA

DEPARTAMENTO DE ESTADÍSTICA



TESIS DOCTORAL

**Detección temprana de la Sigatoka Negra en hojas de
banano mediante imágenes hiperespectrales: Un enfoque
aplicando los métodos PLS-PLR y HS-BIPLLOT.**

JORGE GUSTAVO UGARTE FAJARDO

Salamanca, 2020

**Detección temprana de la Sigatoka negra en hojas de
banano mediante imágenes hiperespectrales: Un enfoque
aplicando los métodos PLS-PLR y HS-BIPLLOT.**

Memoria para optar al título de Doctor en
Estadística Multivariante Aplicada
por la Universidad de Salamanca

Presenta:

JORGE GUSTAVO UGARTE FAJARDO

Salamanca, 2020

JOSE LUIS VICENTE VILLARDÓN

Profesor titular del Departamento de Estadística de la Universidad de Salamanca

CERTIFICA:

Que D. JORGE GUSTAVO UGARTE FAJARDO, ha realizado en el Departamento de Estadística de la Universidad de Salamanca, bajo su dirección, el trabajo para optar al Grado de Doctor en Estadística Multivariante Aplicada, que presenta con el título: “Detección temprana de la Sigatoka negra en hojas de banano usando imágenes hiperespectraless: Un enfoque aplicando los métodos PLS-PLR y HS-BIPLLOT”, autorizando expresamente su lectura y defensa.

Y para que conste, firma el presente certificado en Salamanca,
_____.

Fdo: José Luis Vicente Villardón

DANIEL ERICK OCHOA DONOSO

Profesor titular de la Facultad de Ingeniería en Electricidad y Computación de la
Escuela Superior Politécnica del Litoral.

CERTIFICA:

Que D. JORGE GUSTAVO UGARTE FAJARDO, ha realizado en el Departamento de Estadística de la Universidad de Salamanca, bajo su dirección, el trabajo para optar al Grado de Doctor en Estadística Multivariante Aplicada, que presenta con el título: “Detección temprana de la Sigatoka negra en hojas de banano usando imágenes hiperespectrales: Un enfoque aplicando los métodos PLS-PLR y HS-BIPLLOT”, autorizando expresamente su lectura y defensa.

Y para que conste, firma el presente certificado en Salamanca, 21 de octubre del 2020.

 Fdo: Daniel Erick Ochoa Donoso

*A la memoria de mi querida hermana
Sonia Piedad Ugarte F. fallecida el 20 de octubre del 2017
en la ciudad de Guayaquil, durante mi estancia en Salamanca.
No pudimos despedirnos, pero tu alegría y ganas
de vivir serán siempre una luz en nuestra vida.*

AGRADECIMENTOS

A mi esposa y mis hijas, quienes han sido mi fortaleza y me han brindado su amor, su apoyo y su confianza que ha sido la motivación para continuar.

A mis padres, quienes me han acompañado siempre en todos mis logros académicos brindándome su respaldo.

A la doctora Purificación Galindo quien siempre me brindó su apoyo para continuar en este programa doctoral y ha sabido guiarnos con dedicación por lo quedo eternamente agradecido.

A mis tutores de tesis los doctores José Luis Vicente-Villardón y Daniel Ochoa Donoso, quienes con su alto nivel de conocimientos y destacada experiencia como investigadores han encaminado este trabajo hacia una exitosa culminación.

A los amigos, compañeros y colaboradores que me brindaron su apoyo para realizar este trabajo.

ÍNDICE GENERAL

AGRADECIMENTOS.....	VI
ÍNDICE GENERAL.....	VII
ÍNDICE DE FIGURAS	X
ÍNDICE DE TABLAS	XIII
1 INTRODUCCIÓN.....	1
2 REVISIÓN BIBLIOGRÁFICA.....	10
2.1 FUNDAMENTOS DE LOS SENSORES REMOTOS.....	11
2.1.1 Las ondas electromagnéticas	11
2.1.2 Escáneres de teledetección	17
2.2 LAS IMÁGENES HIPERESPECTRALES Y SUS APLICACIONES	21
2.2.1 Sistema Hiperspectral	22
2.2.2 Firma Espectral.....	27
2.3 DETECCIÓN DE ENFERMEDADES EN PLANTAS UTILIZANDO HSI	29
2.3.1 Análisis de imágenes hiperspectrales.....	35
2.4 MÍNIMOS CUADRADOS PARCIALES (PLS).....	39
2.4.1 PLS (NIPALS)	41
2.4.2 Nway PLS-DA (NPLS-DA).....	43
2.5 BIPLLOT	47
2.5.1 Biplots clásicos.....	48
2.5.2 Biplot Logístico.....	51
2.6 MÁQUINA DE VECTORES SOPORTE (SVM)	56
2.6.1 Clasificador lineal SVM.....	57
2.6.2 El truco del Kernel.....	62
2.6.3 Funciones Kernel.....	62
2.7 REDES NEURONALES ARTIFICIALES.....	64
2.7.1 Perceptrón.....	65
2.7.2 Perceptrón multicapa (Multi-Layer Perceptron MLP).....	66
2.7.3 Funciones de activación.....	68
2.7.4 Dimensionalidad de las capas y conexiones	76
2.7.5 Función de pérdida	77
3 MATERIALES Y MÉTODOS.....	79

3.1	ORGANISMOS	80
3.1.1	Plantas	80
3.1.2	Patógeno	80
3.2	EQUIPOS.....	81
3.2.1	Sistema de adquisición de datos Hiperespectrales.....	81
3.2.2	Software.....	85
3.3	INOCULACIÓN DE PLANTAS.....	86
3.4	ADQUISICIÓN DE IMÁGENES.....	88
3.5	PRE-PROCESAMIENTO DE DATOS	90
3.6	MÉTODOS ESTADÍSTICOS	92
3.6.1	Análisis exploratorio	95
3.6.2	Modelo PLS-PLR	96
3.6.3	HS-Biplot.....	109
3.6.4	Análisis comparativo	112
3.6.5	Modelo NPLS-DA.....	114
3.6.6	Modelo de clasificación SVM	117
3.6.7	Redes Neuronales Artificiales MLP	121
4	RESULTADOS	129
4.1	ANÁLISIS EXPLORATORIO	132
4.1.1	Análisis de firmas espectrales.....	132
4.1.2	Prueba de normalidad	137
4.1.3	Análisis de multicolinealidad	138
4.2	PLS-PLR.....	144
4.2.1	Selección de parámetro de penalización Ridge (λ).....	144
4.2.2	Predicción y validación del modelo PLS-PLR	147
4.2.3	Validación Externa del modelo PLS-PLR	150
4.3	ANÁLISIS HS-BIPLLOT	155
4.4	NPLS-DA.....	161
4.4.1	Entrenamiento del modelo NPLS-DA	161
4.4.2	Validación del modelo NPLS-DA	166
4.5	SVM.....	170
4.5.1	SVM Lineal	171
4.5.2	SVM Polinomial	179
4.6	REDES NEURONALES	188
4.6.1	MLP con una capa oculta	188
4.6.2	MLP con dos capas ocultas	198
4.7	ANÁLISIS COMPARATIVO	209
5	DISCUSIÓN.....	217
6	CONCLUSIONES.....	225
7	BIBLIOGRAFÍA	229
	APÉNDICE A	251

APÉNDICE B	256
APÉNDICE C	259
APÉNDICE D	279

ÍNDICE DE FIGURAS

<i>Figura 2.1 Características de la onda electromagnética.....</i>	<i>12</i>
<i>Figura 2.2 Diagrama esquemático de los tipos de reflexión.</i>	<i>14</i>
<i>Figura 2.3 Espectro electromagnético.....</i>	<i>16</i>
<i>Figura 2.4 (A) Bandas multiespectrales y (B) Bandas hiperespectrales.....</i>	<i>18</i>
<i>Figura 2.5 Plataformas de teledetección según el alcance</i>	<i>20</i>
<i>Figura 2.6 Diagrama esquemático de un sistema HSI.....</i>	<i>23</i>
<i>Figura 2.7 Métodos de adquisición de imágenes hiperespectrales.....</i>	<i>25</i>
<i>Figura 2.8 Dimensiones de una imagen hiperespectral.....</i>	<i>26</i>
<i>Figura 2.9 Firmas espectrales de algunos materiales</i>	<i>27</i>
<i>Figura 2.10 Firma espectral de vegetación saludable.....</i>	<i>31</i>
<i>Figura 2.11 Estructura de la hoja de una planta.....</i>	<i>33</i>
<i>Figura 2.12 Simulación de espectros de hojas.....</i>	<i>34</i>
<i>Figura 2.13 Descomposición de matrices con el método PLS.</i>	<i>41</i>
<i>Figura 2.14 Modelo PLS1 de 3 vías.....</i>	<i>43</i>
<i>Figura 2.15 Desdoblamiento de una matriz de 3 vías en primer modo.</i>	<i>44</i>
<i>Figura 2.16 Interpretación general del biplot.</i>	<i>50</i>
<i>Figura 2.17 Escala de proyecciones en biplot logístico</i>	<i>54</i>
<i>Figura 2.18 Hiperplano generado por SVM en 2 y 3 dimensiones.....</i>	<i>57</i>
<i>Figura 2.19 Clasificación con SVM.....</i>	<i>61</i>
<i>Figura 2.20 Neurona biológica.....</i>	<i>64</i>
<i>Figura 2.21 Arquitectura del perceptrón.....</i>	<i>66</i>
<i>Figura 2.22 Perceptrón multicapa (MLP).</i>	<i>67</i>
<i>Figura 2.23 Proceso de aprendizaje del perceptrón multicapa (MLP).</i>	<i>68</i>
<i>Figura 2.24 Función de activación lineal.</i>	<i>69</i>
<i>Figura 2.25 Función de activación sign().</i>	<i>70</i>
<i>Figura 2.26 Función de activación sigmoid().</i>	<i>71</i>
<i>Figura 2.27 Función de activación tanh().</i>	<i>72</i>
<i>Figura 2.28 Funciones de activación ReLu() y softplus().</i>	<i>73</i>

<i>Figura 2.29 Función de activación Hard tanh()</i>	74
<i>Figura 2.30 Función de Activación Leaky ReLu</i>	75
<i>Figura 3.1 Sistema de Imágenes Hiperespectrales</i>	81
<i>Figura 3.2 Etapas de la enfermedad Sigatoka negra</i>	87
<i>Figura 3.3 Imagen de una hoja escaneada en el sistema HSI</i>	89
<i>Figura 3.4 Reducción de cubos hiperespectrales</i>	91
<i>Figura 3.5 Patrones de reflectancia SNV de regiones sanas e infectadas</i>	91
<i>Figura 3.6 Curva ROC para un modelo hipotético</i>	113
<i>Figura 3.7 Estructura de tensor de tercer orden en NPLS-DA</i>	116
<i>Figura 3.8 Tensor desplegado primer modo</i>	117
<i>Figura 3.9 Aplicación de hiperparámetro C en SVM de margen blando</i>	119
<i>Figura 3.10 Diagrama de MLP con una capa oculta</i>	124
<i>Figura 3.11 Diagrama de MLP con dos capas ocultas</i>	127
<i>Figura 4.1 Firmas espectrales de regiones sanas y enfermas</i>	133
<i>Figura 4.2 Variación de la reflectancia con BLSD</i>	135
<i>Figura 4.3 Firmas espectrales de hojas sanas e infectadas</i>	136
<i>Figura 4.4 Prueba de Kolmogorov Smirnov (Lilliefors)</i>	138
<i>Figura 4.5 Gráfico de calor de la matriz de correlación de variables predictoras</i>	143
<i>Figura 4.6 Gráfico 3D de la respuesta del modelo PLS-PLR</i>	146
<i>Figura 4.7 Probabilidad estimada por PLS-PLR con validación cruzada</i>	148
<i>Figura 4.8 Probabilidad estimada por el modelo PLS-PLR en prueba de validación</i>	151
<i>Figura 4.9 HS-Biplot del dataset de entrenamiento</i>	156
<i>Figura 4.10 HS-Biplot del dataset de validación</i>	158
<i>Figura 4.11 Tensor de características de cubos hiperespectrales</i>	161
<i>Figura 4.12 Predicción NPLS-DA con 1 a 6 componentes</i>	162
<i>Figura 4.13 Predicción del modelo NPLS_DA con datos de entrenamiento</i>	163
<i>Figura 4.14 Predicción del modelo NPLS_DA en prueba de validación</i>	166
<i>Figura 4.15 Probabilidad estimada con el modelo SVM lineal con datos de entrenamiento</i>	172
<i>Figura 4.16 Matriz de confusión de entrenamiento del modelo SVM lineal</i>	173
<i>Figura 4.17 Probabilidad estimada por el modelo SVM lineal en prueba de validación</i>	176
<i>Figura 4.18 Matriz de confusión de validación del modelo SVM lineal</i>	177
<i>Figura 4.19 Probabilidad estimada por el modelo SVM polinomial con datos de entrenamiento</i>	180
<i>Figura 4.20 Matriz de confusión del modelo SVM polinomial con datos de entrenamiento</i>	181
<i>Figura 4.21 Probabilidad estimada para dataset de prueba con el modelo SVM polinomial</i>	184

<i>Figura 4.22 Matriz de confusión del modelo SVM con Kernel polinomial con el dataset de prueba.....</i>	<i>184</i>
<i>Figura 4.23 Curva de exactitud en entrenamiento de la MLP con una capa oculta.....</i>	<i>189</i>
<i>Figura 4.24 Curvas de aprendizaje de la MLP con una capa oculta.....</i>	<i>190</i>
<i>Figura 4.25 Probabilidad estimada por la MLP con una capa oculta en fase de entrenamiento.....</i>	<i>191</i>
<i>Figura 4.26 Matriz de confusión de predicción de datos de entrenamiento por la MLP con una capa oculta.....</i>	<i>192</i>
<i>Figura 4.27 Probabilidad estimada por la MLP con una capa oculta en validación externa.....</i>	<i>195</i>
<i>Figura 4.28 Matriz de confusión de red neuronal de una capa oculta en validación externa.....</i>	<i>196</i>
<i>Figura 4.29 Curva de exactitud en entrenamiento de la MLP con dos capas ocultas.....</i>	<i>200</i>
<i>Figura 4.30 Curvas de aprendizaje de la MLP con dos capas ocultas.....</i>	<i>200</i>
<i>Figura 4.31 Probabilidad estimada por la MLP con dos capas ocultas en fase de entrenamiento.....</i>	<i>201</i>
<i>Figura 4.32 Matriz de confusión de predicción de datos de entrenamiento por la MLP con dos capas ocultas.....</i>	<i>202</i>
<i>Figura 4.33 Probabilidad estimada por la MLP con dos capas ocultas en validación externa.....</i>	<i>205</i>
<i>Figura 4.34 Matriz de confusión de MLP con dos capas ocultas en validación externa.....</i>	<i>206</i>
<i>Figura 4.35 Espectro de hoja 13.....</i>	<i>213</i>
<i>Figura 4.36 Espectro de hoja 15.....</i>	<i>214</i>
<i>Figura 4.37 Espectro de hoja 20.....</i>	<i>215</i>
<i>Figura 4.38 HS-Biplot de dataset de validación con hojas numeradas.....</i>	<i>216</i>

ÍNDICE DE TABLAS

<i>Tabla 1-1 Escala de severidad de la Sigatoka negra (BLSD).....</i>	<i>2</i>
<i>Tabla 2-1 Clasificación del Espectro electromagnético</i>	<i>17</i>
<i>Tabla 3-1 Longitudes de onda de las regiones visible y cercana al infrarrojo.....</i>	<i>109</i>
<i>Tabla 3-2 Número de épocas para entrenamiento de redes neuronales</i>	<i>125</i>
<i>Tabla 4-1 Intervalos del espectro en los cuales se producen cambios de reflectancia por efecto de la BLSD</i>	<i>135</i>
<i>Tabla 4-2 Métricas de bondad de ajuste para modelos PLS-PLR con valores incrementales de penalización Ridge (λ)</i>	<i>144</i>
<i>Tabla 4-3 Matriz de confusión del modelo PLS-PLR con validación cruzada.....</i>	<i>148</i>
<i>Tabla 4-4 Matriz de confusión del modelo PLS-PLR en prueba de validación</i>	<i>152</i>
<i>Tabla 4-5 Detalle de las hojas con bajo grado de infección.....</i>	<i>157</i>
<i>Tabla 4-6 Matriz de confusión del modelo NPLS-DA con datos de entrenamiento.....</i>	<i>164</i>
<i>Tabla 4-7 Matriz de confusión para evaluación del modelo NPLS-DA.....</i>	<i>167</i>
<i>Tabla 4-8 Métricas de predicción del modelo SVM lineal con datos de entrenamiento.....</i>	<i>173</i>
<i>Tabla 4-9 Evaluación de métricas de predicción de datos de prueba usando el modelo SVM lineal.....</i>	<i>177</i>
<i>Tabla 4-10 Métricas de predicción de datos de entrenamiento con el modelo SVM polinomial</i>	<i>181</i>
<i>Tabla 4-11 Evaluación de métricas de predicción de datos de prueba usando el modelo SVM polinomial.....</i>	<i>185</i>
<i>Tabla 4-12 Exactitud en entrenamiento y validación del modelo generado por la MLP con una capa oculta</i>	<i>189</i>
<i>Tabla 4-13 Métricas de predicción de datos de prueba usando la MLP con una capa oculta</i>	<i>192</i>
<i>Tabla 4-14 Métricas de predicción de MLP con una capa oculta con datos de prueba</i>	<i>196</i>
<i>Tabla 4-15 Exactitud en entrenamiento y validación del modelo generado por la MLP con dos capas ocultas.....</i>	<i>199</i>
<i>Tabla 4-16 Evaluación de métricas de predicción de datos de prueba usando la MPL con 2 capas ocultas</i>	<i>202</i>
<i>Tabla 4-17 Métricas de predicción de datos de prueba con MLP con dos capas ocultas en validación externa</i>	<i>206</i>
<i>Tabla 4-18 Tabla comparativa de métricas de predicción en fase de entrenamiento.....</i>	<i>210</i>

<i>Tabla 4-19 Tabla comparativa de métricas de predicción en fase de validación</i>	<i>210</i>
<i>Tabla 4-20 Errores de clasificación en prueba de validación externa</i>	<i>212</i>

1 INTRODUCCIÓN

El banano (*musa spp*) es uno de los productos agrícolas más cultivados en el mundo y es el principal producto agrícola en muchos países. Por sus beneficios nutricionales, esta fruta tropical es considerada un producto básico y contribuye a la seguridad alimentaria en gran parte de los países en desarrollo. Sus principales centros de producción están ubicados en Asia, América Central, Sudamérica y África, siendo los principales países exportadores de la fruta Ecuador, Filipinas, Guatemala, Costa Rica, Colombia y Honduras, en ese orden (Yeturu et al. 2016; FAO 2017).

Las plantaciones de banano son afectadas por una serie de problemas fitosanitarios entre los cuales se destaca la Sigatoka negra (BLSD, por sus siglas en ingles Black Sigatoka Disease), enfermedad foliar considerada la principal amenaza de la producción bananera por su impacto devastador que causa pérdidas de hasta 80% de los rendimientos. BLSD es originada por el hongo patógeno *Pseudocercospora fijiensis* (Perera, Kelaniyangoda, & Salgadoe, 2013) y su desarrollo ocasiona necrosis de la planta en seis estados sintomáticos (Bakache et al., 2019) (tabla 1-1).

Tabla 1-1 Escala de severidad de la Sigatoka negra (BLSD)

ESTADO	SÍNTOMAS
1	Manchas amarillentas de menos de 1 mm en el envés de la hoja.
2	Rayas cloróticas de color rojo o marrón de 1 a 5 mm.
3	Similar a la etapa 2 pero las rayas son mayores de 5 mm hasta 2cm.
4	Rayas elípticas marrones en el envés y rayas negras en el haz.
5	Manchas totalmente negras que se han extendido al envés de la hoja. Las manchas están rodeadas por un halo amarillo.
6	El centro de la macha es gris claro rodeado por un anillo negro y un halo amarillo.

La enfermedad afecta los tejidos de fotosíntesis de las hojas y causa degeneración de la producción de clorofila (Chaerle et al., 2007), lo que produce cambios en la estructura foliar. La Sigatoka negra se caracteriza por una fase biotrófica asintomática seguida de una fase necrotrófica con síntomas visibles. En los primeros estadios de la enfermedad se presentan pequeñas lesiones o manchas oscuras en la parte inferior de hoja que se desarrollan en forma de líneas finas de color marrón con una longitud entre 2 y 3 mm, en las cuales se establecen estructuras llamadas conidiosporas en donde se producen las esporas asexuales o conidios, la cuales se transiten por agua a cortas distancias, mientras que durante la fase sexual se producen gran cantidad de esporas ascosporas, que pueden ser dispersadas a largas distancias por las corrientes de aire y son las responsables de la diseminación de la enfermedad (Hidalgo et al., 2006). Como resultado del avance de la enfermedad las pequeñas líneas se unen y oscurecen hasta que se ennegrecen (Stover, 1980) presentando los primeros signos de necrosis. Luego, las zonas muertas se secan, causando la defoliación y la madurez temprana de la fruta. Una vez que aparecen los síntomas, el rendimiento de planta ya está comprometido (Marín et al., 2003; Hidalgo et al., 2006). Al inicio de la infección, la fase biotrófica pre-sintomática puede durar algunas semanas, después los síntomas se hacen visibles y la enfermedad se disemina afectando a la planta en forma irreversible (Marín et al., 2003) y causando pérdidas de la producción hasta del 85% (Luna-Moreno et al., 2019).

La Sigatoka negra deteriora el área foliar de la planta, retrasa tanto la floración como la cosecha y reduce el llenado del racimo (Hidalgo et al., 2006). En la actualidad, los cultivos afectados son tratados con la aplicación de fungicidas químicos los cuales afectan a la calidad de la fruta y generan resistencia del patógeno incrementando los

costos de control de la enfermedad y los daños al medio ambiente. Ante estas condiciones, los consumidores han reaccionado con mayor exigencia por una fruta libre del uso de plaguicidas. (Patiño L. F., Bustamante E., & Salazar L. M., 2007)

La investigación de métodos y técnicas que permitan detectar la Sigatoka negra ha sido, durante muchos años, un tema de alto interés para los investigadores con el objetivo de controlar la expansión de la enfermedad. Actualmente, la detección de BLSD se realiza por evaluación visual de los síntomas o mediante análisis destructivos como pruebas de DNA o inmunológicas (Luna-Moreno et al., 2019).

Entre los avances tecnológicos utilizados en Agricultura de Precisión se destacan los sistemas de imágenes hiperespectrales (HSI) y multiespectrales, los cuales, tradicionalmente han sido aplicados, con éxito, en estudios de largo y medio alcance de geografía, geología, ecología, meteorología e hidrología entre otros (Levin, 1999) y en la última década, han sido utilizados en aplicaciones de corto alcance para la agricultura, tales como el monitoreo de la calidad de frutas, detección de manifestaciones de insectos o hierba en cultivos, la detección y medición de la severidad de enfermedades de las plantas (Bock et al., 2010). En la Teledetección de largo alcance, el sensor se ubica a cientos a miles de kilómetros del objetivo, mientras que la de medio y corto alcance a distancias hasta cientos de kilómetros. Los sensores hiperespectrales registran la energía reflejada en materiales en un amplio rango espectral (Lowe, Harrison, & French, 2017).

Las propiedades ópticas de las hojas están relacionadas con los procesos fotosintéticos y de generación de energía que al ser afectados por una enfermedad producen cambios fisiológicos en las hojas que pueden ser detectados estudiando la variación de reflectancia (luz reflejada) en diferentes longitudes de onda.

Los cambios estructurales y químicos que ocurren durante una patogénesis han permitido la detección de enfermedades en otras plantas utilizando el análisis de imágenes hiperespectrales (Siche et al., 2016). Trabajos previos con remolacha (Mahlein, 2011), trigo (Ashourloo et al., 2014), tomate and lechuga (Lara et al., 2013) y otros detallados en el trabajo de Mishra et al. (2017) han provisto mayor conocimiento acerca de la relación entre las infecciones y las variaciones espectrales en las hojas. En plantas de banano, los trabajos de investigación utilizando imágenes hiperespectrales han sido enfocados en mediciones del tamaño de la fruta (Hu et al., 2015), control de madurez de la fruta (Intaravanne et al., 2012) y diferenciación entre Sigatoka negra y Amarilla (Bendini et al., 2015). Según nuestro mejor conocimiento, los métodos no destructivos basados en HSI para la detección temprana de BLSD no han sido evaluados todavía.

Los cambios internos y externos de las hojas producidos por las enfermedades producen cambios en la reflectancia en diferentes regiones del espectro electromagnético. Los cambios en la estructura de las hojas producen cambios en la reflectancia (luz reflejada) en longitudes de onda de la región cercana al infrarrojo (NIR, 780-1350 nm) producto de variaciones en la fotosíntesis y los cambios en los pigmentos fotosintéticos son evidentes en la región visible (VIS, 380-780 nm). Con el uso de cámaras hiperespectrales se puede obtener imágenes en los rangos espectrales en los que se han detectado evidencias de cambios fisiológicos en los cultivos (Mahlein, 2011) y la variación en la reflectancia medida en una imagen hiperespectral puede ser utilizada como indicador potencial de la presencia de la enfermedad (Bock et al., 2010).

Las imágenes hiperespectrales de las hojas conforman un set de datos tridimensional con altas dimensiones espaciales y espectrales que tienen características

distintivas como son: (1) alta colinealidad en las bandas adyacentes, (2) variabilidad de firmas hiperespectrales y (3) alta dimensionalidad debido a la elevada resolución espacial y espectral de los sensores hiperespectrales (Lu & Fei, 2014). Consecuentemente, es necesario aplicar metodologías de procesamiento multivariante capaces de correlacionar los perfiles hiperespectrales y las infecciones de las plantas con el fin de detectar e identificar en forma precisa las enfermedades en los cultivos para la aplicación oportuna de estrategias de control de plagas y prevención de enfermedades (Han, Haleem, & Taylor, 2015).

Varias investigaciones han asociado cambios metabólicos con signos pre-sintomáticos de las enfermedades y los efectos bioquímicos o estrés ambiental en las plantas (Dunn & Ellis, 2005). Específicamente, en relación al banano, el *P. fijiensis* modifica el patrón de transporte de los fotoasimilados (Hidalgo et al., 2006) destruyendo el tejido fotosintético, lo que induce un incremento en la fluorescencia, la emisión del calor y la reducción en la producción de clorofila debido a las lesiones necróticas y cloróticas, dando como resultado variaciones de reflectancia en regiones visibles (VIS) e infrarrojo cercano (NIR) del espectro electromagnético (Cevallos-Cevallos et al., 2018) que confirman los estudios realizados en otras plantas.

En estas condiciones, el pronóstico temprano de la BLSD tiene gran importancia tanto para productores como para consumidores de la fruta porque permite la racionalización de los fungicidas aplicados, logrando la reducción de los costos de producción y la mejora en el estado sanitario de los cultivos. El uso de imágenes hiperespectrales en la detección de la BLSD aventajan a los métodos tradicionales por ser un método no destructivo que permite la detección de la enfermedad cuando los síntomas

aún no son visibles por el ser humano. En los estados iniciales de BLSD, tales como, el pre-sintomático (hojas infectadas sin síntomas), primer y segundo (ver tabla 1-1), los cambios físicos en la planta son mínimos, haciendo difícil la identificación visual de daños en las hojas. Además, las esporas asexuales y sexuales se desarrollan a partir del segundo estado de la enfermedad en adelante y son responsables de la diseminación de la enfermedad. Por lo tanto, la detección de BLSD en las etapas iniciales es crítica para controlar la enfermedad en tiempos de tratamiento más cortos y reducir el impacto en la calidad de la fruta, el medio ambiente y los costos de producción.

La caracterización de enfermedades de plantas utilizando datos hiperespectrales han sido el objetivo de numerosos trabajos de investigación en los cuales se ha utilizado avanzados métodos de aprendizaje automático (ML) que han sido recopilados por Lowe, Harrison & French (2017) y Liakos et al. (2018), entre los cuales destacan: las máquinas de vector de soporte (Support Vector Machine SVM), redes neuronales artificiales (Artificial Neural Network ANN), Mínimos cuadrados parciales – análisis discriminante (Partial Least Squares-Discriminant Analysis PLS-DA), Análisis discriminante lineal (Linear Discriminant Analysis LDA), bosques aleatorios (random-forest RF) (Zhu et al., 2017).

Los métodos de aprendizaje automático buscan patrones en un alto número de variables con gran precisión pero a medida que mejora su capacidad predictiva se incrementa su complejidad y esto provoca una pérdida sensible de la capacidad de interpretabilidad de los resultados del modelo, esto enfrenta al investigador a la siguiente dicotomía: precisión o interpretabilidad? esto es, seleccionar un modelo simple,

posiblemente menos preciso, pero con mayor interpretabilidad o un modelo más complejo que reporte mayor eficiencia predictiva pero menos entendible para los investigadores.

Muchos de los problemas en los que se busca una excelente predicción también requieren del descubrimiento de puntos claves de entendimiento para la generación de nuevo conocimiento acerca de la causalidad o sobre el origen de los errores o predicciones no ajustadas al modelo porque que el modelo seleccionado requiere que se cumplan. En la mayoría de las tareas del mundo real, solo el resultado obtenido en predicción es una información incompleta. La interpretabilidad hace posible extraer este conocimiento adicional capturado por el modelo (Molnar, 2019).

En este trabajo nosotros presentamos dos técnicas de análisis multivariante, PLS (Partial Least Squares) con regresión logística penalizada (PLS-PLR) y el HS-Biplot (HyperSpectral Biplot), que han sido implementadas para el tratamiento de las condiciones inherentes a los datos hiperespectrales y que combinan el alto poder de predicción de PLS junto con la alta interpretabilidad de la estructura de los datos que ofrece el biplot.

PLS-PLR ha sido implementado aplicando mejoras en el algoritmo PLS que lo convierte en un método robusto para resolver los principales problemas de los datos hiperespectrales como lo es eliminar la multicolinealidad, reducir el sesgo (bias), controlar el sobre-ajuste (overfitting) y eliminar los efectos de la separación de datos. Por otro lado, el HS-Biplot permite la representación gráfica de la estructura de las tablas, es decir, la conformación de grupos de individuos (similaridad) y la relación con longitudes de onda espectrales. La complementación de las dos herramientas permite obtener

resultados con alta precisión en la predicción y mejorar la interpretación de las estructuras de los datos.

El objetivo principal de esta investigación es desarrollar un método no destructivo de detección temprana de la Sigatoka negra mediante la aplicación de las técnicas de análisis multivariante PLS con regresión logística penalizada (PLS-PLR) y HS-Biplot (HyperSpectral Biplot) en cubos hiperespectrales de hojas de banano obtenidos mediante un sistema de teledección de corto alcance.

Los objetivos específicos son los siguientes:

- Obtener imágenes hiperespectrales de las hojas de plantas de banano sanas e inoculadas con el hongo *Pseudocercospora fijiensis* de la Sigatoka negra.
- Realizar el pre-procesamiento de las imágenes que contienen la medición de la reflectancia de las hojas de banano para la aplicación de las técnicas propuestas.
- Desarrollar un modelo predictivo multivariante con dimensión reducida aplicando PLS-PLR en una muestra de hojas no-infectadas e infectadas y aplicarlo para evaluar la presencia de enfermedad Sigatoka negra en nuevas hojas de banano.
- Aplicar HS-Biplot para representar la estructura de la tabla de datos y obtener evidencia visual de las relaciones entre grupos de individuos y variables.
- Realizar un análisis comparativo con las técnicas NPLS-DA, SVM y redes neuronales en base a los resultados de la evaluación del poder predictivo y la capacidad explicativa de las técnicas.

2 REVISIÓN BIBLIOGRÁFICA

2.1 FUNDAMENTOS DE LOS SENSORES REMOTOS.

La teledetección, traducción del término en inglés Remote Sensing, también llamada percepción remota, sirve para designar las actividades de observación, adquisición e interpretación de información de alguna propiedad de un objetos o fenómenos sin que exista contacto material con el objeto o fenómeno en estudio (Levin, 1999). Los sistemas de teledetección recopilan datos mediante la detección de la radiación electromagnética que fluye desde los objetos observados hacia un sensor (cámara, escáner o un radar) situado en una plataforma (satélite, avión, dron, etc), luego son analizados y procesados para obtener información detallada acerca de las propiedades físicas de los objetos en observación.

Los sensores reciben la energía radiante reflejada en los objetos, la cual será registrada en una imagen que presenta la disposición espacial de los espectros de reflectancia. Si la energía reflejada proviene del Sol, los sensores utilizados se denominan pasivos y si proviene de la misma plataforma en la cual está ubicado el sensor, estos toman el nombre de sensores activos (Dyring, 1973).

2.1.1 Las ondas electromagnéticas

Las ondas electromagnéticas son combinaciones de campos eléctricos ($\vec{E}$) y magnéticos ($\vec{B}$) que se trasladan en el espacio a la velocidad de luz.

$$c = \lambda \nu \quad (2.1.1)$$

Donde,

c : es la velocidad de la luz

λ : es la longitud de onda

ν : es la frecuencia

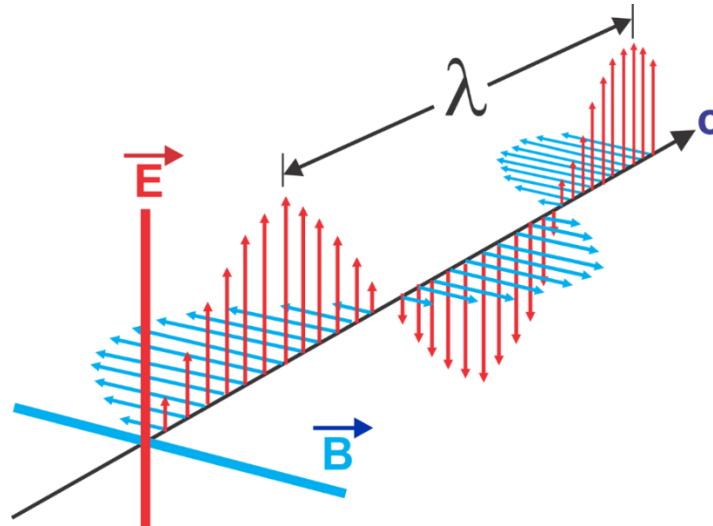


Figura 2.1 Características de la onda electromagnética.

(Recuperado de <https://www.needpix.com/photo/679154/electromagnetic-waves-wave-length-electric-field-oscillations-magnetic-field-oscillations>)

La amplitud es la altura máxima de la onda. La distancia de un ciclo completo o un periodo espacial es la longitud de onda (λ) y es medida en metros (m), nanómetros ($\text{nm} = 10^{-9} \text{ m}$) o micrómetros ($\mu\text{m} = 10^{-6} \text{ m}$). El número de ciclos o perturbaciones en un periodo de tiempo es la frecuencia (ν) y se mide en hercios o herz (Hz) que equivale al número de ciclos por segundo (figura 2.1). La velocidad de la luz c es un valor constante ($3 \times 10^8 \text{ m/s}$) (Chuvieco, 1991).

Cuando la energía electromagnética encuentra materia en estado sólido, líquido o gaseoso se producen cambios en la radiación incidente debido a la oscilación de las partículas que la componen. Estos cambios son principalmente de amplitud, dirección,

longitud de onda, polarización y fase. La composición y textura de la superficie de destino determina el tipo de interacción con la energía recibida, esto es, absorción, refracción o reflexión.

La absorción se produce cuando la frecuencia de la energía electromagnética incidente coincide o es próxima a la frecuencia de oscilación de los electrones del material, lo que ocasiona que el fotón transfiera toda su energía al electrón obligándolo a dar un salto desde un estado energético basal o fundamental a un estado de mayor energía (estado excitado), la misma que se transforma en energía térmica. En este caso, el material es opaco para la energía electromagnética retenida. La absorbancia o índice de absorción depende de la estructura atómica y de las condiciones del medio (pH, temperatura, fuerza iónica, constante dieléctrica) (Díaz et al., 2010).

La refracción se produce si la radiación incidente es re-emitida de un medio a otro y cambia de dirección y su velocidad. En este caso el material es transparente a esa radiación.

La reflexión se produce cuando la radiación regresa al medio donde se originó, pero en un ángulo diferente al de incidencia. Existen 2 tipos de reflexión: la reflexión especular y la reflexión difusa (figura 2.2). En el mundo real, se encuentra una combinación de las dos.

La reflexión especular, o reflexión en forma de espejo, se produce cuando toda la energía o la mayor parte de ella, se dirige fuera de la superficie de incidencia en una sola dirección.

La reflexión difusa ocurre cuando la luz incide en superficies rugosas por lo que se refleja en casi todas las direcciones. Depende del nivel de rugosidad o irregularidad de

la superficie si la energía se refleja en forma especular o difusa, o en algún punto intermedio (Kerle, Jansen, & Huurnerman, 2004).

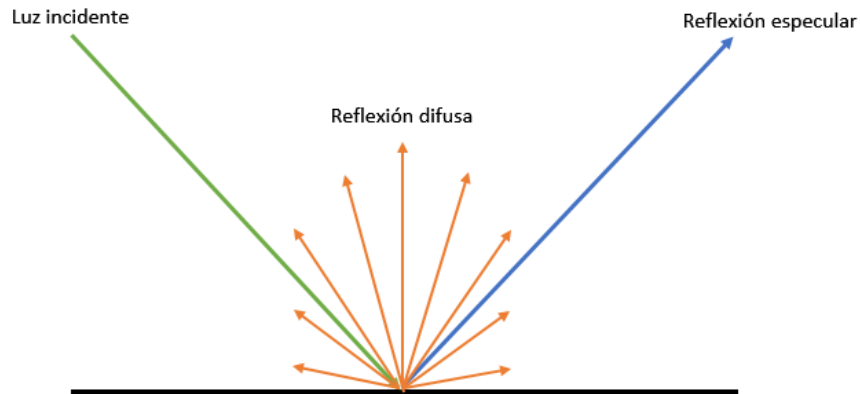


Figura 2.2 Diagrama esquemático de los tipos de reflexión.

2.1.1.1 Magnitudes Radiométricas

La energía absorbida o emitida se mide en fotones utilizando la ecuación de Planck:

$$Q = h\nu = h \frac{c}{\lambda} \quad (2.1.2)$$

Donde

Q es la energía de luz absorbida o emitida medida en *Joules (J)*,

ν es la frecuencia de la energía electromagnética, en *herz (Hz)* y

h es la constante de Planck, $6.626 \times 10^{-34} \text{ J.s.}$

Radiación. - es la medida de energía irradiada por un objeto en forma de ondas electromagnéticas.

Transmitancia. – se produce cuando la radiación atraviesa la sustancia. La relación entre la cantidad de luz transmitida que llega al detector después de que ha atravesado la muestra es llamada transmitancia.

Absorbancia.- Nos indica la cantidad de radiación que penetra una superficie, es absorbida e incorporada a la estructura molecular del objeto (Díaz et al., 2010).

Reflectancia. - La radiación es reflejada en forma especular o difusa. Se calcula mediante el cociente entre la energía reflejada y la energía incidente.

La energía incidente es la suma de la energía absorbida, transmitida y reflejada.

$$E_I = E_A + E_T + E_R \quad (2.1.3)$$

Donde,

E_I : Radiación que incide en un objeto

E_A : Radiación absorbida

E_T : Radiación transmitida

E_R : Radiación reflejada

La reflectancia es medida como el porcentaje de energía que se refleja en la superficie del material. Cuando un objeto es irradiado por el haz de luz se produce una reflectancia característica que depende de la composición de la muestra.

La reflectancia en una longitud de onda determinada, depende de las características física y químicas del objeto, por lo que es posible identificar cambios en las características del material comparando la reflectancia en cada longitud de onda, que en conjunto conforman firma espectral de cada material, antes y después del cambio. Las firmas espectrales son utilizadas en el análisis de imágenes hiperespectrales para obtener información sobre la estructura físico-química de los materiales observados. La

reflectancia toma valores entre 0 y 1 o se expresa en porcentaje entre 0 y 100 % (Levin, 1999).

$$\text{Reflectancia} = E_R / E_I \quad (2.1.4)$$

2.1.1.2 Espectro Electromagnético

El rango total de longitudes de onda es llamado espectro electromagnético (figura 2.3) La teledetección opera en varias regiones del espectro electromagnético, dependiendo de la capacidad del sensor utilizado.

Las ondas electromagnéticas han sido organizadas en grupos contiguos de acuerdo con las longitudes de onda que conforman el espectro electromagnético.

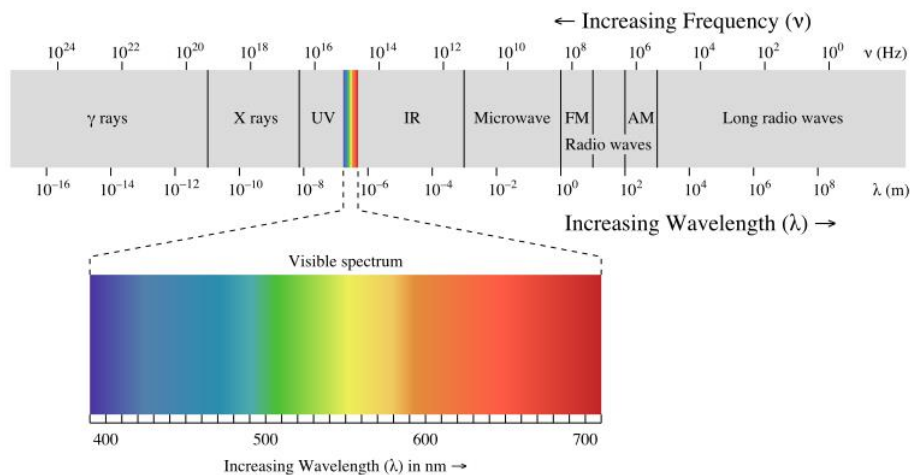


Figura 2.3 Espectro electromagnético.

(Recuperada de <http://chemgroups.ucdavis.edu/~larsen/ChemWiki.htm>).

La clasificación de acuerdo con la longitud de onda es la siguiente:

Cada uno de los tipos de onda corresponden a intervalos de longitud de onda o frecuencia. Así:

Tabla 2-1 Clasificación del Espectro electromagnético

ESPECTRO ELECTROMAGNÉTICO			
Tipo		Long. De Onda	Frecuencia
Radio	Muy Baja Frecuencia	> 10 Km	< 30 KHz
	Onda Larga	< 10 Km	> 30 KHz
	Onda Media	< 650 m	> 650 KHz
	Onda Corta	< 180 m	> 1.7 MHz
	Muy Alta Frecuencia	< 10 m	> 30 MHz
	Ultra Alta Frecuencia	< 1 m	> 300 MHz
Microondas		< 30 cm	> 1.0 GHz
Infrarrojo	Lejano	< 1 m	> 300 GHZ
	Medio	< 50 μ m	> 6 THz
	Cercano	< 2.5 μ m	> 120 THz
Luz Visible		< 780 nm	> 384 THz
Ultravioleta	Cercano	< 380 nm	> 789 THz
	Extremo	< 200 nm	> 1.5 PHz
Rayos X		< 10 nm	> 30 PHz
Rayos Gamma		< 10 pm	> 30 EhZ

2.1.2 Escáneres de teledetección

Los escáneres o sensores utilizados en teledetección son dispositivos que reciben las ondas electromagnéticas y las convierten en una imagen que está dividida en bandas del espectro electromagnético (Roman-Gonzalez & Vargas-Cuentas, 2013). Los sensores son instalados en plataformas estáticas o móviles dependiendo del tipo de información que sea requerida. A mayor altura, el sensor cubre mayor área, pero el detalle que puede ser observado se reduce.

Existen diferentes tipos de sensores para diferentes tipos de aplicaciones. Los sensores multiespectrales han sido optimizados a través del tiempo y son clasificados por el ancho de banda que pueden capturar, la resolución espectral y espacial. Los primeros sensores aparecieron en los años 60 con menos de 20 bandas espectrales en las regiones visible, infrarrojo reflectivo (IR) e infrarrojo termal. En primer escáner multiespectral se lanzó en 1972 y fue instalado en el satélite Landsat 1. El desarrollo tecnológico de los sensores da sus frutos en el año 2000 en el que se presentan sensores con elevado número de bandas y con un número mayor de píxeles como es el caso del Hyperion, sistema sensor con 220 bandas, 30 m por pixel y datos de 10 bits. Este tipo de sensor con muchas bandas espectrales son llamados hiperespectrales.(Landgrebe, 2002).

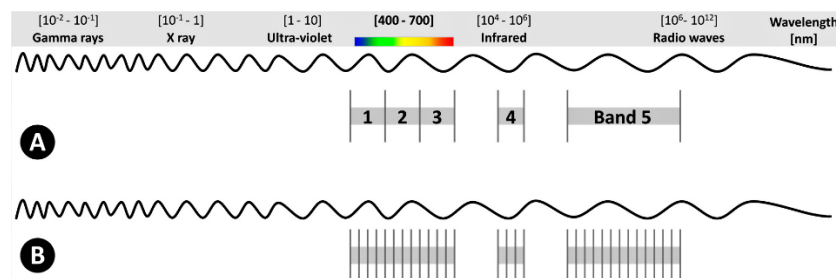


Figura 2.4 (A) Bandas multiespectrales y (B) Bandas hiperespectrales. (Adão, T. et al., 2017)

2.1.2.1 Tipos de sensores de teledetección

Cámara aérea digital. – la cámara digital, compuesto de lentes y CCD (Charge Couple Device) es el más antiguo de los sensores pasivos. Registra una imagen completa con una sola exposición en longitudes de onda en el rango entre 400 nm y 900 nm dividida en 3 bandas correspondientes a los 3 colores primarios azul, verde y rojo. El infrarrojo es

incorporado utilizando un color falso. Se las encuentra en aviones para realizar fotografías aéreas. Son utilizadas principalmente en mapeo topográfico y catastral.

Escáner multiespectral. – este tipo de sensor obtiene observaciones punto por punto y línea por línea mientras que una cámara aérea realiza registra una imagen completa con una sola exposición y la divide en decenas de bandas espectrales. El sensor escanea sistemáticamente la superficie a observar y mide la energía reflejada en algunas decenas de bandas de longitudes de onda.

Sensores hiperespectrales o Espectrógrafo. - Tiene el mismo funcionamiento que los sensores multiespectrales, pero poseen mejor resolución de longitud de onda (5 nm y 10 nm) y pueden funcionar en un amplio intervalo del espectro electromagnético, que incluyen luz visible, ultravioleta e infrarrojo, dividido en muchas bandas espectrales (50 a 500). Los sensores hiperespectrales pueden obtener miles de los espectros que identifican los componentes a nivel de pixel y al mismo tiempo su alta resolución espacial permite obtener la distribución espacial a distancias desde micrómetros hasta decenas de centímetros (Maldonado, Fuentes, & Contreras, 2018).

Otros sensores son el escáner térmico, radiómetro microonda, escáner laser, radar de imágenes, radar altímetro, sonar de barrido lateral, etc. (Kerle, Jansen, & Huurnerman, 2004).

Los sensores son instalados en distintas plataformas para estudiar objetivos a distintas distancias, lo que conlleva a una clasificación de la teledetección de acuerdo con el alcance en el que se realiza el estudio: de largo alcance (long range) cuando el sensor se ubica de cientos a miles de kilómetros del objetivo en una plataforma que, por lo general, es un satélite; de medio alcance (medium range), en este caso, el sensor se ubica a

distancias entre cientos de kilómetros en plataformas aéreas como aviones, helicópteros y vehículos no tripulados (UAVs Unmanned Aerial Vehicles) grandes; finalmente, la teledetección de corto alcance o rango corto, que cubren distancias menores a centenas de metros hasta centímetros o milímetros del objetivo. Entre las plataformas utilizadas en teledetección de corto alcance están pequeños UAVs, vehículos terrestres, dispositivos portátiles e instalaciones estacionarias (Frey, 2019).

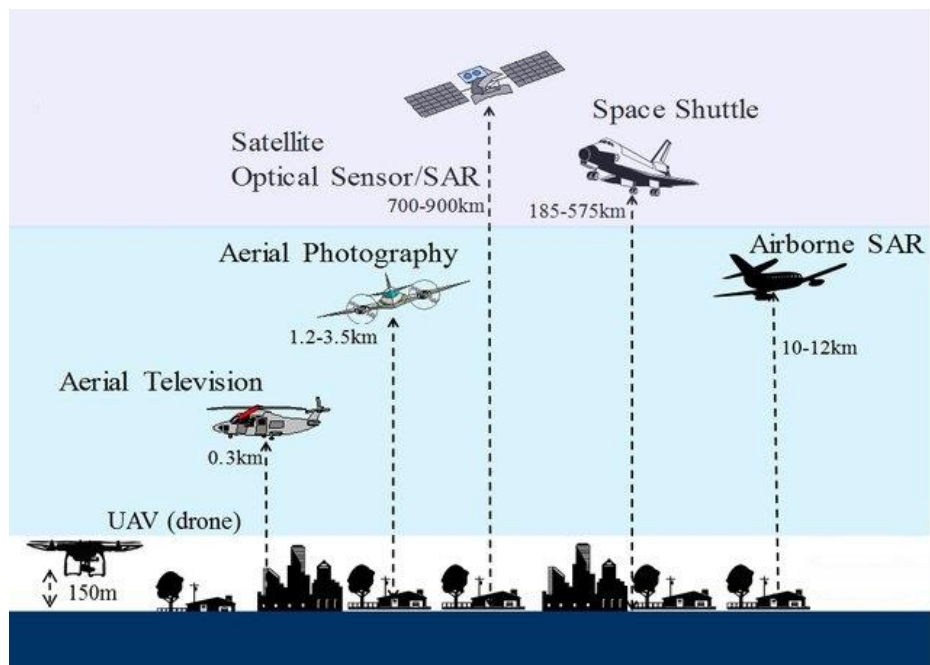


Figura 2.5 Plataformas de teledetección según el alcance (Yamazaki and Wen, 2016).

2.2 LAS IMÁGENES HIPERESPECTRALES Y SUS APLICACIONES

La teledetección hiperespectral, también conocida como espectropía de imágenes es una técnica que incluye los procesos de medición, análisis e interpretación de la radiación electromagnética registrada por un sensor hiperespectral (Plaza et al., 2006). Además de la gran capacidad para detectar las características físicas del material inspeccionado, la imagen hiperespectral tiene otras ventajas sobre los métodos tradicionales como lo son: procesos de preparación de la muestra reducidos, es una técnica no destructiva con tiempos de adquisición rápidos y visualización de la distribución espacial de las variaciones químicas del material (Elmasry & Sun, 2010). Por otro lado, entre sus desventajas están el costo superior frente a otros tipos de sensores ópticos y la elevada carga computacional que se requiere para procesar este tipo de datos con elevada dimensión espectral. Para atenuar esta última desventaja, una práctica muy utilizada es aplicar técnicas de reducción de la dimensionalidad de las imágenes, logrando una representación mínima que contiene las características necesarias para realizar el estudio deseado (Chang, 2016).

La diversidad y versatilidad de los sensores hiperespectrales compactos desarrollados durante la última década junto con los avances de las unidades de procesamiento gráficas (GPU's) han generado un gran impulso y flexibilidad en las aplicaciones de teledetección de corto alcance. Actualmente existen una gama muy alta de dispositivos para aplicaciones hiperespectrales que permite seleccionar sensores y plataformas para cada aplicación, tales como el uso de trípodes, UVAs, aviones ligeros o sistemas de laboratorio (Kurz & Buckley, 2016).

Los sensores hiperespectrales son utilizados en aplicaciones de Geología, Geografía y Agricultura para detectar materiales y especies en la superficie, como mapas de vegetación, análisis de suelos, identificación de minerales en la superficie y relieve superficies geográficas (Slaton, Raymond Hunt Jr., & Smith, 2001). Además, las imágenes hiperespectrales pueden utilizarse para obtener información más detallada de la calidad de alimentos y productos de fabricación, detección de residuos mineros y mapeo de presencia de microorganismos y contaminación en masas de agua (Aalderink, Klein, Padoan, Bruin, & Steemers, 2010; Tan, 2017). En los últimos treinta años se han logrado avances prometedores con el procesamiento de imágenes hiperespectrales (HSI Hyperspectral Imaging) de corto alcance aplicado a la agricultura (Mahlein et al., 2012), tecnología relativamente nueva que está siendo aplicada en la agricultura con excelentes resultados para monitoreo de la calidad de la fruta, detección de residuos fecales en alimentos, detección de insectos y maleza en cultivos y para la detección de enfermedades y fenotipado de las plantas (Bock et al. 2010; IB, Antonio, & Almorox 1999).

2.2.1 Sistema Hiperespectral

Los sistemas hiperespectrales están compuestos de cuatro elementos principales: fuente de luz, lentes, espectrógrafo y detector de área (figura 2.6).

Las fuentes de luz utilizadas en configuraciones HSI de laboratorio son las lámparas halógenas y las LED. Las fuentes halógenas cubren longitudes de onda en los rangos VIS y NIR de manera uniforme sin pico agudos y consumen baja energía. Sin embargo, los cambios de temperatura y voltaje producen picos en la respuesta espectral. Las fuentes LED (Diodos emisores de luz) tienen larga vida útil, bajo consumo y emisión de calor,

producen luz en longitudes de onda de banda corta en regiones ultravioleta, visible e infrarroja y de banda ancha, en diferentes disposiciones de luces.

Los lentes objetivos son elementos claves porque de ellos depende el enfoque de la luz entrante desde un área en un elemento detector para formar un pixel de la imagen.

El espectrógrafo es un componente fundamental del sistema HSI que sirve para separar la luz en las distintas longitudes de onda por medio de una rejilla de difracción que separa la luz policromática en las longitudes de onda que la constituyen. Cuando la luz incide en la rejilla de difracción, cada longitud de onda se refleja en diferentes ángulos. La rejilla está formada por una capa reflectante que contiene ranuras paralelas. El ángulo de la ranura permite que sea reflejada una longitud de onda determinada.

El detector de área convierte los fotones en energía eléctrica que luego es digitalizada. Para ello, el dispositivo de carga acoplada (Charge Coupled Device CCD) es el más utilizado. El CCD está formado por fotodiodos configurados como un arreglo 1-D para detectar por líneas o 2-D para detectar por áreas. Otra alternativa con mayor velocidad y menor costo es el sensor CMOS (Mishra et al., 2017).

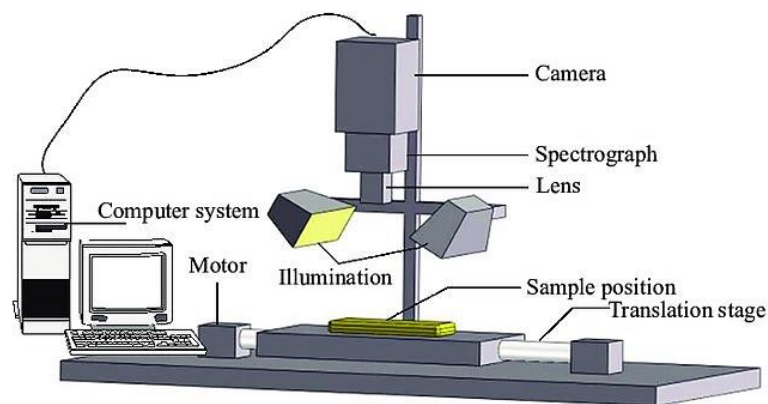


Figura 2.6 Diagrama esquemático de un sistema HSI (Yijie, W. A. N. G., & CHENG, J., 2018).

La operación de adquisición de las imágenes hiperespectrales en los sistemas presentados en la sección anterior puede realizarse aplicando una de las cuatro técnicas siguientes: exploración por punto (whisk-broom), barrido por línea (push-broom), barrido por área (spectral scanning) y captura instantánea (non-scanning) (figura 2.7).

En la exploración por punto el proceso de registro se realiza pixel por pixel. El detector o la muestra se mueve en las dos dimensiones espaciales. Aunque es muy útil para casos en los que se requiere alta resolución espectral, este proceso es lento.

En el barrido por línea el escáner o la muestra se mueven en una sola dirección realizando la captura línea por línea. Este método es el más utilizado.

En el barrido por área se realiza la captura de toda la imagen en una longitud de onda a la vez hasta completar todo el rango de longitudes de onda. Se puede utilizar para escanear un rango de longitudes de onda seleccionado.

En el método de captura instantánea se captura todas las longitudes de onda en toda la imagen, obteniendo en una sola exposición toda la información espectral y espacial.

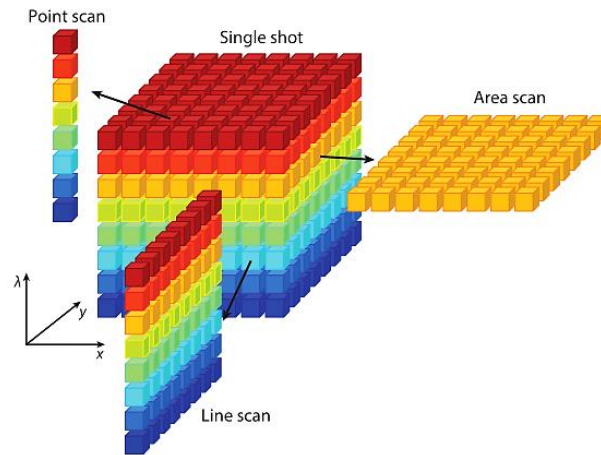


Figura 2.7 Métodos de adquisición de imágenes hiperespectrales (Ma et al., 2019).

Las imágenes captadas por los sensores hiperespectrales o espectrómetros de imagen conforman un conjunto de datos tridimensional con dos dimensiones espaciales y una dimensión espectral de longitud igual al número de bandas espectrales que recibe el sensor, que se conoce como el cubo HS o hipercubo (figura 2.8) (Vo-Dihn, 2004). Los valores capturados para cada pixel corresponden a la reflectación en cada banda a lo largo del espectro electromagnético y definen la firma espectral que permite distinguir diferentes componentes y los cambios que se producen en ellos (Richards, 2013).

La resolución espacial del sensor representa la capacidad de detectar una característica más pequeña posible, mientras que los pixeles son las unidades más pequeñas de una imagen. Cuando una imagen es presentada en resolución completa, cada pixel representa el área especificada por la resolución espacial. Los arreglos de pixeles presentan la distribución espacial de una imagen.

La resolución radiométrica describe la habilidad para discriminar mínimas cantidades de energía. Mientras menor es la resolución radiométrica, el sensor es más sensible para detectar pequeñas diferencias en energía reflejada.

La resolución digital es el número de bits que son utilizados para representar un valor de reflectancia en la imagen. Mientras mayor sea el número bits, mayor será el número de niveles de brillantez que puede ser registrado. La resolución radiométrica está directamente relacionada con la resolución digital.

La resolución espectral corresponde al número de bandas espectrales que el dispositivo puede medir. La elevada resolución espectral de los sensores hiperspectrales permiten detectar la energía reflejada en cientos de bandas espectrales muy finas lo que facilita la discriminación entre diferentes objetivos según su respuesta espectral (Levin, 1999).

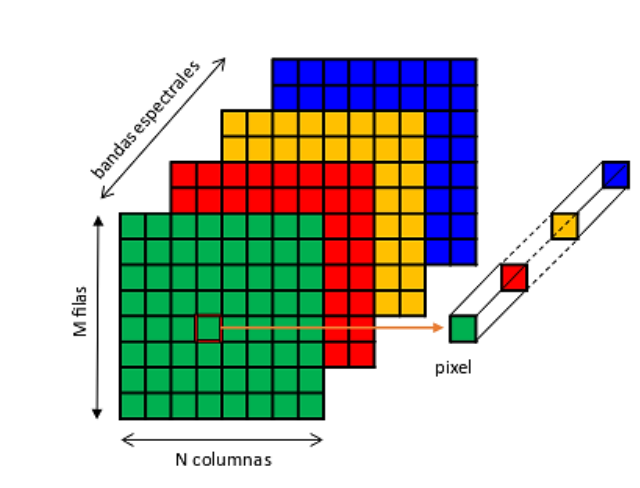


Figura 2.8 Dimensiones de una imagen hiperspectral.

2.2.2 Firma Espectral

La firma espectral o espectro en un punto de un determinado material es el conjunto de valores de radiancia o reflectancia en los diferentes canales espectrales del sensor. Si el número de bandas espectrales del sensor es muy grande y las bandas son muy estrechas, la firma espectral puede ser considerada como un espectro casi continuo (Muriel, 2009).

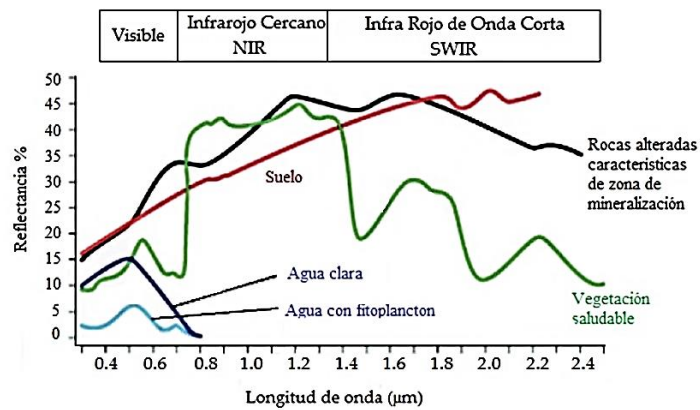


Figura 2.9 Firmas espectrales de algunos materiales (Fajardo Reina, 2019).

La firma espectral determina patrones del comportamiento de la reflectancia de cualquier material expuesto a una fuente de luz. En la figura 2.9 se puede observar las curvas de reflectancia características de algunos materiales en la superficie de la Tierra. Los cambios en los valores de reflectancia en distintas longitudes de onda permiten identificar los cambios en los materiales, por ejemplo, el espectro de agua clara muestra niveles de reflectancia mayores en el rango visible, los cuales pueden alterados por el contenido de fitoplancton en el agua y como resultado la reflectancia se reduce. La espectroscopia de reflectancia en el rango infrarrojo cercano (NIR) se ha convertido en la

herramienta más importante para el análisis de componentes sólidos y líquidos (Marten, Shenk, & Barton II, 1989).

2.3 DETECCIÓN DE ENFERMEDADES EN PLANTAS

UTILIZANDO HSI

En los últimos años se han producido importantes avances en la aplicación de técnicas de automatización y control de los procesos agrícolas, lo cual se ha enmarcado en el concepto de “Agricultura de Precisión”, mediante la aplicación de nuevas tecnologías para el control de la producción agrícola, entre las que se encuentra la teledetección hiperspectral (Gebbers & Adamchuk, 2010). La tecnología de sensores hiperspectrales ha tenido un desarrollo significativo logrando convertirse en una herramienta de análisis no destructiva prometedora para la evaluación de plantas que ofrece velocidad, precisión y confiabilidad por medio de la caracterización tanto la estructura como la concentración de los pigmentos de las hojas (Bousset et al., 2016), lo que ha dado lugar que las imágenes hiperspectrales se conviertan en una importante herramienta para evaluar las enfermedades en las plantas a para la detección de sus síntomas (Kuska, M. et al., 2015) y cambios físico-químicos causados por la patogénesis (Behmann, Steinrücken, & Plümer 2014).

Los sensores remotos capturan la energía reflejada por las plantas expuestas a una fuente de luz. Durante la exposición se desarrollan procesos simultáneos de emisión, absorción, reflexión y transmisión de energía radiante en las superficies irradiadas que generan un patrón característico de reflectancia de la estructura y de las condiciones fisiológica y químicas de las hojas. El patrón o firma espectral evidencia cualquier cambio que se produce en la hoja (IB, Antonio & Almorox, 1999).

Las propiedades ópticas de las hojas pueden ser caracterizadas por (i) la transmisión de luz a través de la hoja, (ii) la luz absorbida por los componentes químicos de la hoja

(tales como: pigmentos, agua, azúcares, lignina y aminoácidos) y (iii) la luz reflejada por la superficie o la estructura interna.

En la figura 2.10 se muestra el espectro de la vegetación saludable. En ella se puede observar alta absorción en la longitud de onda de 450 nm (color azul) y en 650 nm (color rojo) debido a la presencia de pigmentos como la clorofila (verde), xantofilas (amarillo), carotenos (naranja) y antocianinas (rojizo o púrpura) que se presentan en diferentes cantidades durante en el proceso de fotosíntesis. La clorofila predomina sobre los otros pigmentos, lo que causa una apariencia verdosa en las plantas saludables debido a la fuerte absorción de las longitudes de onda roja y azul y la reflexión de las longitudes de onda del verde, que produce un pico leve alrededor de la longitud de onda de 530 nm (Behmann, Steinrücken, , & Plümer 2014) . En la región cercana al infrarrojo (NIR 780 – 1300 nm) la reflectancia se incrementa a más del 50% por presencia del líquido intracelular y los espacios intracelulares del mesófilo, alrededor del 5% es absorbido y el resto se transmite (Govender, Chetty, & Bulcock, 2007). Este incremento masivo de la reflectancia, producido por la clorofila, se lo denomina el borde rojo (red edge) y se presenta en la longitud de onda de los 700 nm, justo en el borde entre el rojo y el infrarrojo (Bock et al., 2010).

Las concentraciones de los pigmentos cambian ante cualquier cambio en el entorno como, por ejemplo, contaminación del aire, acumulación de metales pesados, ataques virales, ataques de insectos o estrés por falta de agua (Mishra et al., 2017). El contenido de agua determina la reflectancia en la región medio-infrarroja (1350 -2500nm) debido a que el agua absorbe fuertemente la radiación electromagnética en la región MIR, particularmente entre 1.55 to 1.75 μm lo que ocasiona una reducción en la reflectancia.

El contenido de agua influye en forma importante en las propiedades espectrales de las hojas, la reflectancia MIR aumenta con la disminución del contenido de agua (Hunt & Rock, 1989).

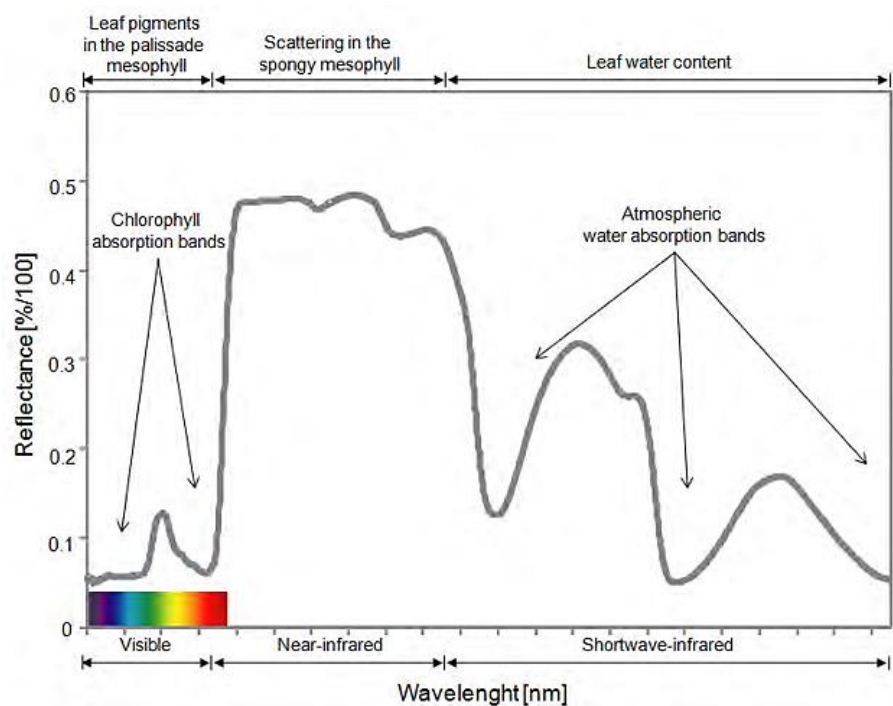


Figura 2.10 Firma espectral de vegetación saludable (Mahlein, 2011).

La reflectancia de las hojas depende de las propiedades ópticas de los materiales que las componen como, por ejemplo, proteínas, lignina, celulosa, azúcar, almidón. Las hojas están compuestas principalmente de cuatro elementos: hidrógeno, carbono, oxígeno y nitrógeno. La luz interactúa con los enlaces C-O, O-H, C-H, and N-H produciendo cambios en el movimiento rotacional de los electrones que generan vibraciones, armónicos y combinaciones de las vibraciones (Kokaly & Clark, 1999).

En resumen, la estructura foliar afecta la reflectancia en la región cercano al infrarrojo (NIR 750 – 1350 nm), mientras que los pigmentos fotosintéticos cambian la reflectancia

en la región visible (380 – 780 nm) y el contenido de agua determina la reflectancia en la región medio-infrarroja (1350 -2500 nm) (Hunt & Rock, 1989). La reflectancia en las bandas espectrales correspondientes al visible (VIS, 380-780 nm), Infrarrojo Cercano (NIR, 780-1350 nm) e Infrarrojo de onda corta (SWIR, 135-2500 nm) permite detectar cambios en la vitalidad de las plantas especialmente si estas se ven afectadas por algún patógeno que produce cambios fisiológicos en el metabolismo de la planta (Slaton, Raymond Hunt Jr., & Smith, 2001). La reflectancia de las hojas es el resultado de las interacciones entre la irradiación de la fuente de luz, la estructura de la hoja y sus características bioquímicas de la planta (Mahlein et al., 2012).

En un estudio realizado por Slaton, Raymond Hunt Jr. & Smith (2001) se investigó la relación entre la reflectancia cercana al infrarrojo en 800 nm (NIR) y las características de la estructura de la hoja utilizando 48 tipos de especies de Angiospermas alpinas. Las pruebas permitieron establecer que existe una fuerte correlación entre la reflectancia cercana al infrarrojo (NIR) y varias características estructurales de la hoja como la razón de área de superficie de células mesofílicas (IAS) por unidad de área de la superficie de la hoja, bicoloración de la hoja y la presencia de una cutícula gruesa. Mientras que la correlación de las longitudes de onda del infrarrojo cercano (NIR) con la densidad de tricomas foliares, el espesor de las hojas, la relación del tejido mesófilo en empalizada (PM) con el tejido mesófilo esponjoso (razón PM/SM) y la proporción del mesófilo por espacios intracelulares de aire (% IAS) fueron relativamente débiles. La figura 2.11 muestra un esquema de los cambios en la luz cuando incide en una hoja.

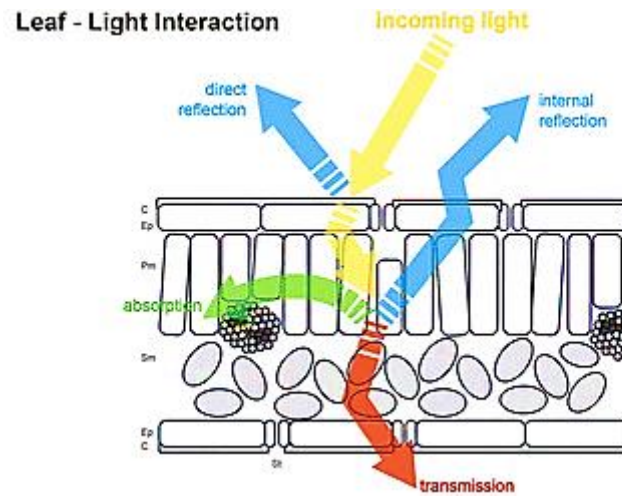


Figura 2.11 Estructura de la hoja de una planta (Mahelin, 2011).

Típicamente, las plantas saludables interactúan con la radiación electromagnética en forma diferente que las plantas infectadas. La figura 2.12 muestra los espectros de reflectancia y transmitancia simulados de hojas obtenidos con el simulador OPTICLEAF (<http://opticleaf.ipgp.fr/>). Las curvas de transmitancia y reflectancia han sido generadas utilizando diferentes parámetros ópticos para evidenciar los efectos de los cambios físico-químicos de las hojas en los espectros de las hojas. La primera corresponde a una hoja saludable con los siguientes parámetros definidos por el simulador: estructura de la hoja = 2.5, contenido de clorofila = $80 \mu\text{g}/\text{cm}^2$, contenido de carotenoides = $10 \mu\text{g}/\text{cm}^2$, pigmentos marrones = 0.1, espesor de agua equivalente = 0.015 cm y masa de la hoja por unidad de área = $0.009 \text{ g}/\text{cm}^2$. La segunda simulación se realizó aplicando cambios en los parámetros de estructura y de los pigmentos para evidenciar la variación del espectro causada por dichos cambios. Los parámetros son los siguientes: estructura de la hoja = 2.5, contenido de clorofila = $80 \mu\text{g}/\text{cm}^2$, contenido de carotenoides = $10 \mu\text{g}/\text{cm}^2$,

pigmentos marrones = 0.1, espesor de agua equivalente = 0.015 cm y masa de la hoja por unidad de área = 0.009 g/cm².

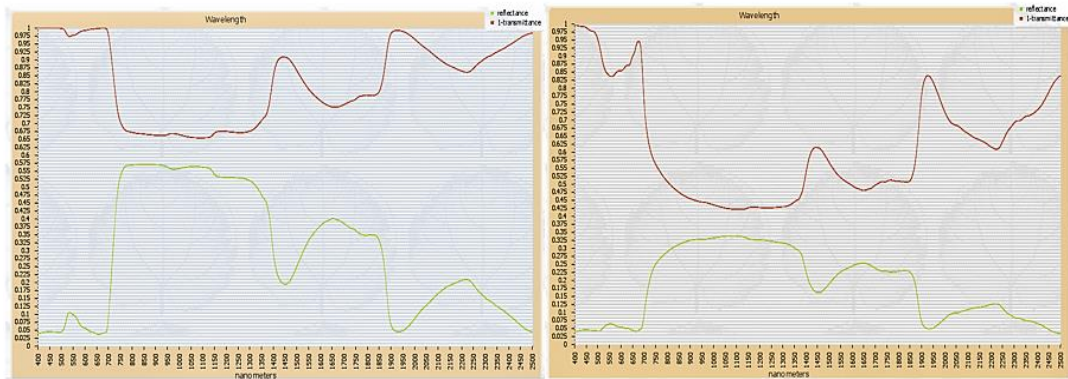


Figura 2.12 Simulación de espectros de hojas.

En plantas de banano, *p. fijiensis* destruye el tejido fotosintético y ocasiona cambios en los procesos metabólicos (Cevallos-Cevallos et al., 2018) que inducen un incremento en la fluorescencia y la emisión de calor en la hoja. Durante su desarrollo, el patógeno modifica el patrón de transporte de fotoasimilados, lo que ocasiona la destrucción del tejido fotosintético. Las esporas de *P. fijiensis* penetran las estomas de las hojas y colonizan el mesófilo foliar provocando la reducción de la conductividad estomática de la que depende la fotosíntesis (Hidalgo et al., 2006), afectando la producción de clorofila. Estos cambios fisiológicos se manifiestan como vaciones de la reflectancia en las regiones VIS y NIR del espectro.

2.3.1 Análisis de imágenes hiperespectrales

Los sistemas HSI proporcionan una gran cantidad de datos de alta calidad, pero no miden directamente los parámetros fisiológicos de la planta y requieren del análisis y procesamiento de imágenes para la detección e identificación de cambios en la salud de las plantas. Una gran variedad de métodos estadísticos y de aprendizaje automático que asimilan datos tan heterogéneos y proporciona información precisa y confiable han sido utilizados junto con métodos de procesamiento de imágenes en diferentes aplicaciones para brindar una efectiva y oportuna respuesta de la presencia de patógenos en las plantas. Los avances en inteligencia artificial, de manera especial, en aprendizaje automático (machine learning ML) han sido un factor preponderante en el desarrollo de las aplicaciones de HSI en plantas.

La caracterización de enfermedades de plantas utilizando datos hiperespectrales ha sido llevada a cabo utilizando avanzados métodos de aprendizaje automático (ML) principalmente supervisados, tales como, las máquinas de vector de soporte (Support Vector Machine SVM), redes neuronales artificiales (Artificial Neural Network ANN), Mínimos cuadrados parciales – análisis discriminante (Partial Least Squares-Discriminant Analysis PLS-DA), Análisis discriminante lineal (Linear Discriminant Analysis LDA), bosques aleatorios (random-forest RF).

Varios autores han realizado contribuciones en la aplicación de métodos de aprendizaje automático para la detección de enfermedades en plantas. Entre las contribuciones, se encuentra la de Huang et al. (2007) con el trabajo denominado “Identification of yellow rust in wheat by modeling the relationship between the Disease index (DI) and the Photochemical Reflectance Index (PRI)” en el cual, utilizando

regresión lineal simple logra una varianza explicada del 91% y un coeficiente de determinación (R^2) igual a 0.97.

Rumpf et al. (2010) utilizó SVM e índices de vegetación para la detección temprana de enfermedades en remolacha alcanzando una precisión en clasificación del 65 % con áreas de la hoja enferma del 1-2%, 95% con 6-9% y 100% cuando tiene más del 10% del área de la hoja infectada. En comparación con los resultados obtenidos con árboles de decisión y redes neuronales artificiales, el error de clasificación de SVM fue menor.

En el 2011, Anne-Katrin Mahlein realizó un estudio para identificar y clasificar tres enfermedades en las hojas de remolacha utilizando métodos no invasivos, basados en imágenes hiperespectrales. Las enfermedades consideradas en este estudio fueron afectaciones foliares de la remolacha como la mancha de la hoja por *Cercospora* (CLS), el mildiu pulverulento (PM) y la roya de la remolacha (SBR) causados por los hongos *Cercospora beticola* (Sacc.), *Erysiphe betae* (Vanha) Weltzien y *Uromyces betae* (Persoon) respectivamente. Se calcularon los índices de vegetación utilizando los valores de reflectancia medidos. Luego se evaluó las combinaciones de los índices de vegetación más adecuados para la discriminación entre las enfermedades mediante los niveles de correlación. SVM fue usado para la clasificación entre hojas saludables (no inoculadas) y hojas inoculadas con cada uno de los patógenos. La precisión fue entre el 93% y el 97%. La cuantificación fue realizada utilizando el Método SAM logrando coeficientes de clasificación Kappa de 0.98 con 17 días de inoculación con *Cercospora*, 0.95 con 17 días de inoculación con mildiu pulverulento y 0.56 con 20 días de roya de la remolacha (Mahlein, 2011).

Zhu et al. (2017) desarrollaron un estudio para detectar el virus del mosaico (Tobacco mosaic virus) en plantas de tabaco utilizando HSI, en el que combinó un método de selección de variable con modelos de aprendizaje automático. El número de longitudes de onda se redujo en un 98% después de la selección de longitudes de onda utilizando SPA (Successive Projections Algorithm) y se aplicó técnicas de aprendizaje automático para realizar la clasificación. Support Vector Machine (SVM), Back Propagation Neural Network (BPNN), Extreme Learning Machine (ELM), Least Squares Support Vector Machine (LS-SVM), Partial Least Squares-Discriminant Analysis (PLS-DA), Linear Discriminant Analysis (LDA) and Random Forest (RF) fueron usados para detectar la enfermedad con una precisión sobre el 85 %. La mayor precisión fue alcanzada con modelos BPNN y ELM (95%).

En la literatura (Lowe, Harrison & French 2017; Liakos et al. 2018) se detallan los avances en recientes trabajos de investigación para la detectar enfermedades en plantas que utilizan métodos de aprendizaje automático con imágenes hiperespectrales.

En esta investigación, nosotros presentamos dos métodos multivariantes, PLS-PLR (Partial Least squares – Penalized Logistic Regression) y HS-Biplot para la detección temprana de BLSD. PLS-PLR es un método con alto poder predictivo cuyo algoritmo hemos fortalecido para brindar solución a algunos problemas que se presentan en datos hiperespectrales, entre los que podemos anotar: La reducción de dimensionalidad de PLS suprime la multicolinealidad; PLS toma en cuenta la variable respuesta por lo que reduce el sesgo y por lo tanto se evita el sub-ajuste (underfitting); la penalización Ridge en la regresión logística limita el crecimiento de los coeficientes de la regresión, con lo cual se

reduce la varianza evitando el sobre-ajuste (overfitting) y los efectos de la separación de datos y la multicolinealidad.

Por otro lado, la alta dimensionalidad de los datos hiperespectrales, dificulta la interpretación de los resultados de la mayoría de los métodos de aprendizaje automático, orientándolos únicamente a cumplir objetivos de predicción. Una solución a esta limitación es el biplot, una herramienta muy útil y poderosa que le da transparencia a la información en una tabla de datos, revelando las estructuras principales de los datos, tales como patrones de correlaciones entre variables o similitudes entre los individuos (Greenacre, 2010). HS-Biplot (HyperSpectral Biplot) se basa en la conceptualización del biplot logístico presentado por Vicente-Villardón et al. (2006) al que se han agregado mejoras para la representación de datos hiperespectrales. En este estudio, el HS-Biplot realiza una representación gráfica simultánea de las hojas de banano (individuos) y las longitudes de onda (variables) en el espacio generado por las dos primeras componentes PLS-PLR y permite realizar un análisis visual de las relaciones entre individuos y variables, aportando un alto poder explicativo.

En las dos siguientes secciones realizamos una presentación teórica de los métodos PLS, Biplot, SVM y redes neuronales artificiales. Los métodos SVM y redes neuronales se encuentran entre los métodos más utilizados en aplicaciones de detección de enfermedades en plantas y en esta investigación serán utilizados para realizar un análisis comparativo con los resultados obtenidos con PLS-PLR y HS-Biplot.

2.4 MÍNIMOS CUADRADOS PARCIALES (PLS)

La regresión de mínimos cuadrados parciales (PLS, por sus siglas en inglés Partial Least Squares) es una técnica estadística introducida por Herman Wold en los 1960's con orientación hacia las ciencias sociales y económicas, que fue extendida posteriormente a otras ciencias tales como Quimiometría (Wold, Ruhe, & Wold, 1984; Hellberg et al. 1987), Administración de Empresas (Hulland, 1999), Marketing (Wu, 2010; Anderson & Swaminathan 2011), Genética (Pérez-Enciso & Tenenhaus 2003; Nguyen & Rocke 2002), Biología (Rodríguez et al., 2014), Agricultura (Diezma et al., 2011), Psicología (Sampson et al., 1989) y Ciencias Sociales (Martínez Ávila & Fierro Moreno, 2018), entre otras.

PLS relaciona dos conjuntos de variables y extrae variables latentes recogiendo la mayor variabilidad que permita modelar las variables respuesta de la mejor manera posible. Cuando existen muchas variables con problemas de multicolinealidad, la regresión múltiple (MCR) no es eficaz. A pesar de que el modelo MCR tenga un buen ajuste a los datos originales, falla en la predicción de nuevos datos. En estos casos, PLS es una técnica especialmente adecuada, principalmente si el número de predictores es mayor que el número de observaciones y a diferencia de MCR la matriz de datos no tiene que ser de rango completo (Takane & Loisel, 2014). Otra alternativa es utilizar PCA, pero no garantiza que las componentes principales sean adecuadas para explicar las respuestas. Mientras en PCA, el objetivo es calcular cada variable latente que explica mejor la varianza en las variables explicativas (X), en PLS se busca la proyección de las variables explicativas y variables respuesta en un espacio de variables latentes que explican mejor tanto a las variables explicativas (X) como a las variables respuesta (Y) y además estas

variables latentes tienen la relación más fuerte posible entre X y Y . Es decir, PLS encuentra hiperplanos de máxima varianza entre la variable de respuesta Y y las variables independientes X resultando variables latentes comprometidas tanto con el ajuste de X como con la predicción de Y . Estudios realizados comparando la aplicación de las dos técnicas han mostrado que PLS es superior al solucionar los problemas de multicolinealidad y dimensionalidad de mejor manera (Vega Vilca & Guzmán, 2011).

La idea central de la regresión PLS es aproximar X por un número (R) reducido de componentes que son los componentes de regresión de mínimos cuadrados parciales a las que denominaremos T y obtener Y mediante la regresión sobre las R componentes. Por lo tanto, PLS intenta modelar X y Y utilizando las componentes comunes T (Smilde, Bro, & Geladi, 2005). PLS descompone la matriz X en puntuaciones T (scores) y cargas P (loadings) de tal manera que maximicen la covarianza entre X y Y (Korkmazoglu, & Kemalbay, 2012).

$$X = TP' + E \quad (2.4.1)$$

$$Y = UQ' + F \quad (2.4.2)$$

PLS calcula los scores T y U y define un modelo de regresión entre scores en lugar de los datos originales. En la figura 2.13 se muestra la matriz X que se descompone en una matriz T (scores) y una matriz P (loadings) más el error E ; la matriz Y se descompone en la matriz U (scores) y la matriz Q (loadings) más el error F . PLS busca minimizar la norma de F manteniendo la correlación entre X y Y dada por la relación $U = BT$.

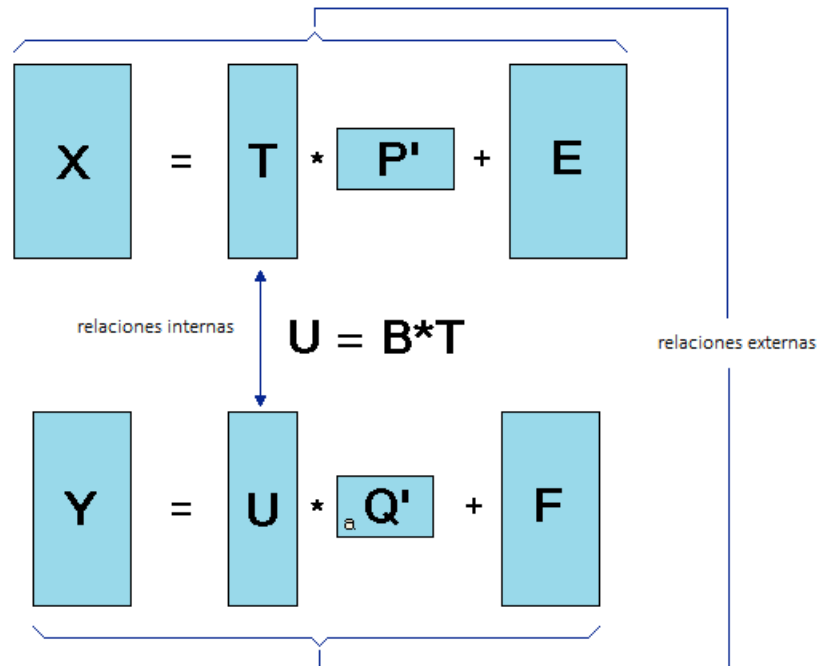


Figura 2.13 Descomposición de matrices con el método PLS.

2.4.1 PLS (NIPALS)

La forma clásica de PLS está basado en el algoritmo NIPALS (H. Wold, 1975; S. Wold, Sjöström, & Eriksson, 2001).

Se requiere estimar $y = \beta_1 x_1 + \dots + \beta_p x_p$, expresado con componentes ortogonales t del espacio de las predictoras correlacionadas con y , como $y = c_1 t_1 + \dots + c_h t_h$.

2.4.1.1 Algoritmo PLS NIPALS

Paso 1. Con X y y centrados por columna, inicializamos u con valores aleatorios o $u = y$

Paso 2. Repetir los siguientes pasos para h componentes ($h \leq \text{rango}(X)$):

2.1 Calcular w

$$w_i = \frac{x' u}{(u' u)}$$

2.2 Normalizar w

$$\mathbf{w} = \frac{\mathbf{w}}{\|\mathbf{w}\|}$$

2.3 Calcular scores $\mathbf{t}$

$$\mathbf{t} = \frac{\mathbf{X}\mathbf{w}}{\mathbf{w}'\mathbf{w}}$$

2.4 Calcular loadings $\mathbf{p}$

$$\mathbf{p} = \frac{\mathbf{X}'\mathbf{t}}{(\mathbf{t}'\mathbf{t})}$$

2.5 Calcular $\mathbf{c}$ ($\mathbf{Y}$ weights) coeficiente de $\mathbf{y}$ sobre $\mathbf{t}$

$$\mathbf{c} = \frac{\mathbf{y}'\mathbf{t}}{(\mathbf{t}'\mathbf{t})}$$

2.6 Calcular vector de scores $\mathbf{u}$ de $\mathbf{y}$

$$\mathbf{u} = \frac{\mathbf{y}'\mathbf{c}}{(\mathbf{c}'\mathbf{c})}$$

2.7 Calcular residuales de $\mathbf{X}$ y $\mathbf{y}$

$$\mathbf{X} = \mathbf{X} - \mathbf{t}\mathbf{p}'$$

$$\mathbf{y} = \mathbf{y} - \mathbf{t}\mathbf{c}'$$

2.8 Calcular las siguientes componentes de forma iterativa.

$$\mathbf{T} = \mathbf{T} + \mathbf{t} \quad \mathbf{X} \text{ Scores}$$

$$\mathbf{W} = \mathbf{W} + \mathbf{w} \quad \mathbf{X} \text{ Weights}$$

$$\mathbf{P} = \mathbf{P} + \mathbf{p} \quad \mathbf{X} \text{ Loadings}$$

Paso 3. Calcular coeficientes de regresión $\mathbf{b}$ sobre $\mathbf{X}$ para $\mathbf{S}$ componentes.

$$\hat{\mathbf{y}} = \mathbf{T}\mathbf{c}'$$

En función de $\mathbf{X}$

$$\text{Si } \mathbf{B} = \mathbf{W}\mathbf{c}'$$

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{W}\mathbf{c}' = \mathbf{X}\mathbf{B}$$

2.4.2 Nway PLS-DA (NPLS-DA)

Conociendo el modelo PLS de 2 vías, lo vamos a extender a estructuras de 3 vías (3-way). PLS1 es un modelo n-way con una variable dependiente (Andersson, 2009) . El algoritmo de NPLS-DA se basa en el algoritmo PLS1 por tener una variable respuesta, pero la respuesta es binaria (Ouertani et al., 2014).

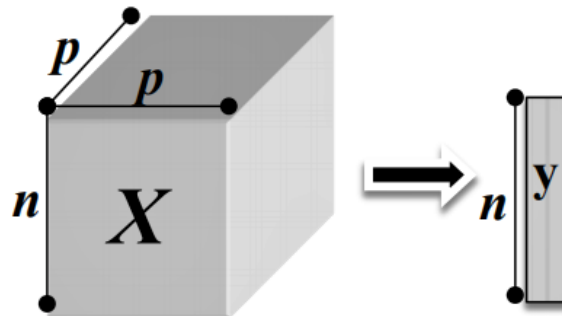


Figura 2.14 Modelo PLS1 de 3 vías.

La construcción de un modelo de calibración con datos de segundo orden (cada observación es una matriz) se puede realizar realizando un desdoblamiento de los datos (unfolding) de tal manera que para cada elemento de la muestra se tiene un tensor de primer orden (figura 2.15). En este caso se pueden utilizar los métodos de tres vías (three way data analysis) (Porcel, 2001).

En el caso, el arreglo de 3 vías de variables independiente ($\underline{X}$) de dimensiones $I \times J \times K$ es convertida en una matriz desdoblada $\mathbf{X}$ con dimensiones $I \times JK$. La matriz $\mathbf{X}$ es descompuesta en modelo triada similar al del modelo PARAFAC (Bro, 1998), pero el modelo se ajusta siguiendo la filosofía del PLS. La triada está formada por el vector $\mathbf{t}$

(scores) y 2 vectores $\mathbf{w}$ (weights) uno del segundo modo $\mathbf{W}^J$ ($J \times 1$) y uno del tercer modo $\mathbf{W}^K$ ($K \times 1$) con la restricción de que su longitud es 1 (Bro, Smilde, & de Jong, 2001).

Todos los modelos multilineales son modelos N-PLS y en general pueden escribirse:

$$\mathbf{X} = \mathbf{T}(\mathbf{W}^K \otimes \mathbf{W}^J)^T + \mathbf{E}_x \quad (2.4.2.1)$$

Donde

$\mathbf{X}$ tensor de tercer orden desdoblado en primer modo,

$\mathbf{T}$ es la matriz de scores,

$\mathbf{W}$ son las matrices de weights.

El problema consiste en encontrar los valores w^j y w^k :

$$x_{ijk} = t_i w_j^J w_k^K \quad (2.4.2.2)$$

De tal manera que los vectores $\mathbf{W}^J$ y $\mathbf{W}^K$ satisfagan

$$\max[cov(t, y) / \min(\sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K (x_{ijk} - t_i w_j^J w_k^K)^2)] \quad (2.4.2.3)$$

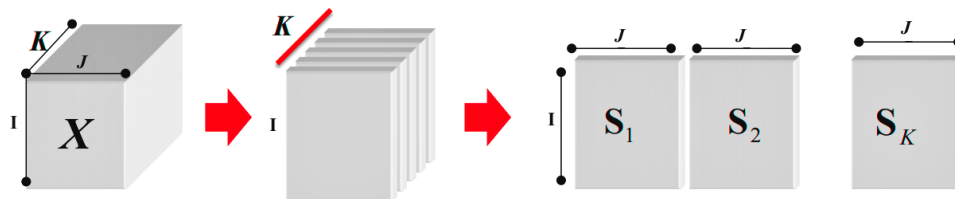


Figura 2.15 Desdoblamiento de una matriz de 3 vías en primer modo.

2.4.2.1 Algoritmo NPLS

Utilizando Descomposición en valores singulares SVD, se obtiene las componentes principales a partir del producto vectorial, obteniendo los espacios latentes $\mathbf{W}^k$ y $\mathbf{W}^j$ que optimizan la varianza entre $\mathbf{X}$ y $\mathbf{y}$. De la siguiente manera (Bro, 1996; Bro, 1998):

1.- Cálculo de $\mathbf{Z}$

$$\mathbf{Z} = \mathbf{X}^T \mathbf{y}$$

Donde,

$\mathbf{X}$ es la matriz desplegada en el primer modo ($I \times JK$)

I es el número de muestras

J es el número de características calculadas

K es el número de bandas espectrales capturadas por la cámara

$\mathbf{y}$: Vector respuesta de I filas.

$\mathbf{Z}$: matriz de covarianza

2.- Cálculo de los vectores singulares izquierdos y derechos. SVD de matriz $\mathbf{Z}$. Se obtiene el primer vector singular izquierdo y el primer vector singular derecho ($\mathbf{W}^k, \mathbf{W}^j$).

$$\mathbf{Z} = \mathbf{U} \Sigma \mathbf{V}'$$

3.- Cálculo del vector $\mathbf{t}$ de la matriz $\mathbf{T}$ (Scores) modelo de mínimos cuadrados

$$\mathbf{t} = \mathbf{X}(\mathbf{W}^k \otimes \mathbf{W}^j)$$

4.- Calcula la $\mathbf{b}$ de regresión. $\mathbf{T}$ incluye todos los scores calculados hasta el momento.

$$\mathbf{b} = (\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\mathbf{y}$$

5.- Cálculo de Residuos

$$\mathbf{y} = \mathbf{y}_0 - \mathbf{T}\mathbf{b}$$

$$\mathbf{X} = \mathbf{X} - \mathbf{T}\mathbf{W}^j(\mathbf{W}^k)^T$$

6.- Reemplaza y continua hasta el apropiado $\mathbf{y}_0$.

Una versión mejorada incluye un arreglo núcleo $\mathbf{G}$ de una descomposición Tucker3 para mejorar la aproximación del arreglo de 3 vías $\underline{\mathbf{X}}$ (Folch-Fortuny et al., 2016).

$$\mathbf{X} = \mathbf{T}\mathbf{G}(\mathbf{W}^k \otimes \mathbf{W}^j)^T \quad (2.4.2.4)$$

Más detalles el cálculo de la matriz $\mathbf{G}$ se puede encontrar en (Bro, Smilde, & de Jong 2001; Smilde, Bro & Geladi 2005).

2.5 BILOT

Un BILOT es una representación gráfica de datos multivariantes que permite visualizar la estructura de grandes matrices de datos (Gabriel, 1971). Las filas de una matriz de datos generalmente son unidades de muestreo observadas, como individuos, países, grupos demográficos, ubicaciones, casos, que en general llamaremos individuos y las columnas son variables que describen las filas, como las respuestas en un cuestionario, indicadores económicos, productos comprados, parámetros ambientales, marcadores genéticos (Greenacre, 2008). Los métodos biplot realizan una representación gráfica de filas (individuos) y columnas (variables) de una matriz de datos en un espacio de dimensión reducida, generalmente dos o tres dimensiones, en la cual, los individuos son representados por marcadores filas (puntos) y las variables por marcadores columna (vectores). Aunque esta situación causa una ligera pérdida de información, proporciona una interpretación más fácil de las relaciones. A través del biplot se proporciona una interpretación geométrica de estructuras de datos multivariadas (Alkan, Atakan, & Akdi, 2015).

Desde el punto de vista algebraico, el método se basa en la reducción de la dimensionalidad mediante la descomposición espectral de la matriz, de la misma forma que en otras técnicas factoriales de reducción de dimensionalidad, con la diferencia que en el biplot se busca la reproducción del dato y se realiza una representación conjunta de filas y columnas.

En un biplot se aproxima una matriz rectangular X de orden $(n \times p)$ y rango r , por otra de rango q ($q < r$), por medio de la factorización en 2 matrices A y B que contienen los

marcadores filas a_i y columna b_j , escogidos de tal manera que el producto interno $a_i^T b_j$ representa el elemento x_{ij} de la matrix X .

$$\hat{X} = AB^T \quad (2.5.1)$$

Si suponemos que las variables son centradas en la media, la proyección de los marcadores fila A en la dirección dada por los marcadores de las columnas B_j predicen el valor de las J variables para cada individuo. En otras palabras, el producto escalar de cada vector que representa una fila por cada vector que representa una columna es una aproximación al valor correspondiente a esa fila-columna en la matriz de datos original.

2.5.1 Biplots clásicos

En la práctica esta factorización se la realiza por Descomposición en Valores Singulares (SVD) (Cárdenas, Galindo-Villardón, & Vicente-Villardón, 2007).

$$X \cong U\Sigma V' \quad (2.5.2)$$

Donde,

U y V son matrices de vectores singulares ortonormales.

Σ es una matriz diagonal que contiene los mayores valores singulares.

De la expresión anterior se puede factorizar y dependiendo del valor de s se obtienen los biplot clásicos.

$$X \cong (U\Sigma^s)(\Sigma^{1-s}V')$$

Si $s = 0$, se obtiene el GH-Biplot en el que los **marcadores columna** tienen calidad de representación óptima.

Si $s = 1$, se obtiene el JK-Biplot. Los **marcadores fila** tienen calidad de representación óptima.

Si $s = \frac{1}{2}$ se obtiene el SQRT biplot.

La interpretación de un biplot se basa en conceptos geométricos (figura 2.16):

La similitud entre los puntos (individuos) es función inversa de la distancia entre los mismos.

Las longitudes de los vectores variables se interpretan como su variabilidad.

Los ángulos entre los vectores se interpretan como su covariabilidad de las variables.

Las relaciones entre individuos y variables se interpretan en términos de las proyecciones de los individuos sobre las variables (producto escalar).

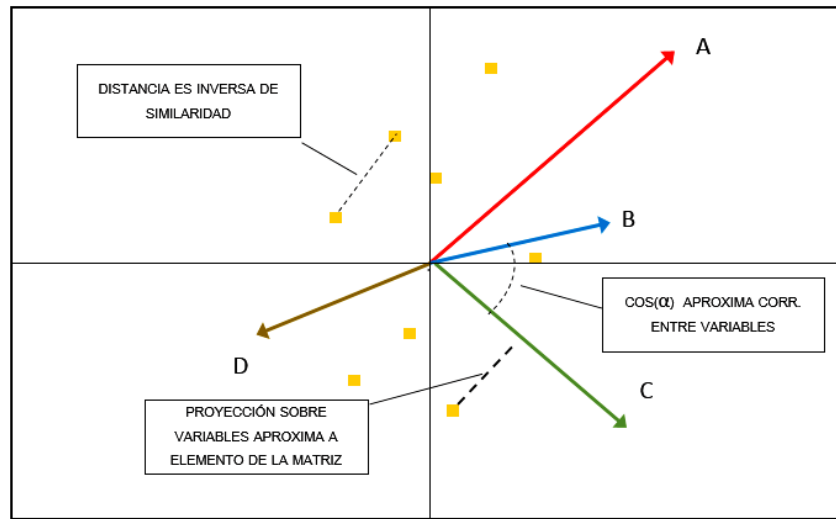


Figura 2.16 Interpretación general del biplot.

2.5.1.1 Bondad de ajuste

Si llamamos $\hat{\mathbf{X}}$ a la mejor aproximación obtenida mediante la reducción de dimensionalidad de $\mathbf{X}$, la bondad de ajuste total o calidad de representación global es la cantidad de variabilidad explicada por la predicción:

$$\rho^2 = \text{tr}(\hat{\mathbf{X}}' \hat{\mathbf{X}}) / (\mathbf{X}' \mathbf{X}) \quad (2.5.1.1.1)$$

La bondad de ajuste para cada columna también llamada calidad de representación o predictividad de la columna es:

$$\rho_j^2 = \text{diag}(\hat{\mathbf{X}}' \hat{\mathbf{X}}) \div \text{diag}(\mathbf{X}' \mathbf{X}) \quad (2.5.1.1.2)$$

donde $\div$ significa la operación elemento por elemento. ρ_j^2 es como el R^2 de la regresión de cada columna de $\mathbf{X}$ sobre $\mathbf{A}$.

Un buen ajuste general no implica que todos los individuos tengan la misma calidad de representación y que la interpretación de las posiciones de todos los puntos en el diagrama sean igualmente confiables. De acuerdo con Demey et al. (2008), un individuo está bien representado cuando la mayor parte de su información (medida a través de la variabilidad) se tiene en cuenta en la dimensión reducida. La bondad de ajuste para cada fila es:

$$\rho^2_i = \text{diag}(\hat{\mathbf{X}}\hat{\mathbf{X}}') \div \text{diag}(\mathbf{X}\mathbf{X}') \quad (2.5.1.1.3)$$

Estas medidas también se llaman calidad de la representación o predictividad. Las medidas separadas para cada dimensión también se denominan Contribuciones del Factor al Elemento (fila o columna) o Cosenos Cuadrados. Las medidas se utilizan para identificar qué dimensiones son útiles para diferenciar al individuo del resto. Las personas con calidad de representación bajas generalmente se colocan alrededor del origen.

2.5.2 Biplot Logístico

Los biplots clásicos son adecuados cuando la respuesta a lo largo de las dimensiones es lineal. Cuando los datos son binarios se debe considerar un biplot logístico. El biplot logístico para datos binarios fue propuesto por Vicente-Villardón et al. (2006) y luego extendido por Demey et al. (2008).

Un biplot logístico es un biplot lineal para datos binarios en el que la respuesta a lo largo de las dimensiones es logística. Cada individuo se representa como un punto y cada variable binaria, como una dirección a través del origen. La proyección de un punto individual en la dirección de un variable binaria predice la probabilidad de presencia de

esa variable. El método está relacionado con la regresión logística de la misma manera que el análisis biplot clásico está relacionado con la regresión lineal (Vicente-Villardón et al., 2006).

Sea $\mathbf{X}$ ($I \times J$) una matriz de datos que las filas corresponden a I individuos y las columnas a J variables binarias. Sea $\pi_{ij} = E(x_{ij})$ la probabilidad esperada de que la característica j esté presente en el individuo i , y x_{ij} la probabilidad observada, ya sea 0 o 1, resultando en una matriz de datos binarios. El biplot logístico S-dimensional en la escala logit se formula como:

$$\text{logit}(\pi_{ij}) = \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = b_{j0} + \sum_{s=1}^S b_{js}a_{is} = b_{j0} + a'_i b_j \quad (5.2.2.1)$$

Donde,

b_{js} y a_{is} son los marcadores fila y columna respectivamente.

El modelo es un modelo lineal (b_i) generalizado que tiene el logit como una función de enlace. En términos de probabilidades en lugar de logits (Hernández-Sanchez & Vicente-Villardón, 2017)

$$\pi_{ij} = \frac{e^{b_{j0} + \sum_k b_{jk}a_{ik}}}{1 + e^{b_{j0} + \sum_k b_{jk}a_{ik}}} \quad (2.5.2.2)$$

En forma matricial,

$$\text{logit}(\Pi) = \mathbf{1}_I \mathbf{b}'_0 + \mathbf{A} \mathbf{B}' \quad (2.5.2.3)$$

Donde,

Π es la matriz de probabilidades esperadas

1_I es un vector de unos y

b_0 es el vector que contiene los interceptos que deben ser agregado porque los datos no se pueden centrar.

Aunque el biplot en la escala logit puede ser útil, sería más interpretable en una escala de probabilidad.

Las predicciones en el biplot logístico se hacen de la misma manera que en los biplots lineales, i. e., proyectando un marcador de fila $a_i = (a_{i1}, a_{i2})$ en un marcador de columna $b_j = (b_{j1}, b_{j2})$.

Considerando que el modelo en escala logit de la ec. (2.5.2.3) los ejes son variables latentes que explican la asociación entre variables observadas, podemos asumir la independencia de los individuos respecto a las variables y que las variables son independientes para los valores dados de los rasgos latentes. Con estos supuestos, la función de probabilidad es

$$Prob(x_{ij} | (\mathbf{b}_0, \mathbf{A}, \mathbf{B})) = \prod_{i=1}^I \prod_{j=1}^J \pi_{ij}^{x_{ij}} (1 - \pi_{ij})^{1-x_{ij}} \quad (2.5.2.4)$$

Si obtenemos el logaritmo

$$L = \log Prob(x_{ij} | (\mathbf{b}_0, \mathbf{A}, \mathbf{B})) \quad (2.5.2.5)$$

$$= \sum_{i=1}^I \sum_{j=1}^J [x_{ij} \log(\pi_{ij}) + (1 - x_{ij}) \log(1 - \pi_{ij})] \quad (2.5.2.6)$$

Para $\mathbf{A}$ fijo, la expresión anterior puede ser separada en una parte para cada una de las J variables.

$$L = \sum_{j=1}^J L_j = \sum_{j=1}^J (\sum_{i=1}^I [x_{ij} \log(\pi_{ij}) + (1 - x_{ij}) \log(1 - \pi_{ij})]) \quad (2.5.2.7)$$

En el caso de una sola variable

$$L_1 = \sum_{i=1}^I [x_{ij} \log(\pi_{ij}) + (1 - x_{ij}) \log(1 - \pi_{ij})] \quad (2.5.2.7)$$

Maximizando L es equivalente a desarrollar una regresión logística.

Entonces, tenemos un biplot en escala logit, excepto por el vector de constantes y la geometría es la misma del caso lineal para predecir los logits y luego, para predecir las probabilidades. El intercepto debe incluirse porque no es posible centrar previamente los datos binarios. Los cálculos para obtener los marcadores de escala son similares al lineal, pero manteniendo el intercepto.

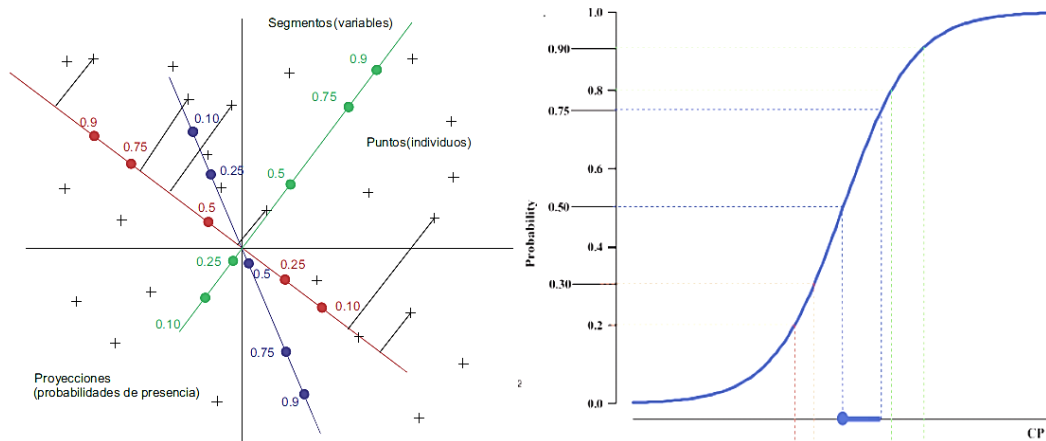


Figura 2.17 Escala de proyecciones en biplot logístico (Hernández-Sánchez, 2016).

La representación final es un biplot lineal interpretado por proyección, a pesar de que la superficie de respuesta no es lineal. El resultado es un biplot es lineal en escala logit.

La representación directa en la escala logit es difícil de interpretar; la escala de probabilidad es más simple y fácil de entender (figura. 2.17).

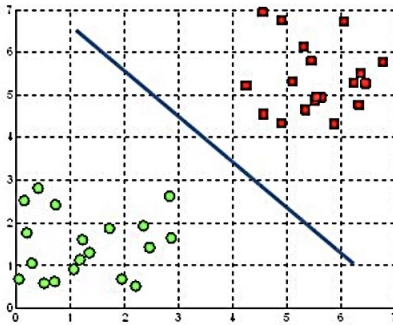
2.6 MÁQUINA DE VECTORES SOPORTE (SVM)

Las máquinas de vectores soporte con siglas SVM de su nombre en inglés (Support Vector Machine) fueron introducidas por Vapnik con fundamentación teórica en sus trabajos sobre aprendizaje estadístico (Vapnik & Vapnik, 1998). SVM es reconocido como un método de clasificación, pero también puede ser utilizado en problemas de clasificación binaria no lineal, clasificación múltiple, regresión y agrupamiento, aunque en este trabajo se le dará énfasis a las SVM aplicadas a problemas de clasificación.

SVM de clasificación es un método supervisado que crea un mapa en el que se diferencian distintas regiones de interés mediante un hiperplano multidimensional que separa las clases (figura 2.18). Es muy utilizado en clasificación de imágenes hiperespectrales por su fortaleza como clasificador lineal ya sea con datos originales (espacio original) si son separables o con datos transformados dentro del espacio de las características para el caso en que los datos originales no sean separables (Lu & Fei, 2014). Para realizar esta transformación de los datos utiliza un proceso llamado en “truco del kernel” (kernel trick) que toma los datos de entrada de baja dimensión y los transforma a una dimensión mayor mediante funciones kernel, logrando la separación de clases.

A diferencia de otros métodos de aprendizaje que se basan en minimizar el error generado por el modelo a partir de los datos de entrenamiento, SVM se centra en minimizar el riesgo estructural puesto que se busca seleccionar el hiperplano marginal con el mayor margen a los puntos más cercanos de cada clase a los que se denomina vectores de soporte (Anggraeni & Lin, 2011).

A hyperplane in $\mathbb{R}^2$ is a line



A hyperplane in $\mathbb{R}^3$ is a plane

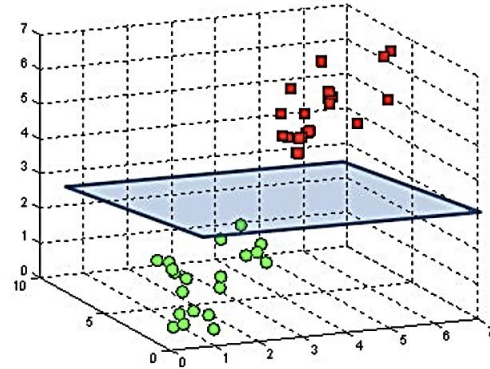


Figura 2.18 Hiperplano generado por SVM en 2 y 3 dimensiones (Gandhi, R., 2018).

2.6.1 Clasificador lineal SVM

Dado un conjunto de datos separables etiquetados $S = \{(x_1, y_1), \dots, (x_n, y_n)\}$ donde $x_i \in \mathbb{R}^d$ e $y_i \in \{-1, +1\}$

La función lineal de separación está dada por:

$$D(x) = (w_1x_1 + \dots + w_dx_d) + b = \langle w, x_i \rangle + b \quad (2.6.1.1)$$

Donde:

$$\mathbf{w} \in \mathbb{R}^d$$

$$b \in \mathbb{R}$$

$\langle w, x_i \rangle$ es el producto escalar entre $\mathbf{w}$ y x_i

Los puntos de x que se sitúan en el hiperplano satisfacen la ecuación,

$$\mathbf{w} \cdot \mathbf{x}_i + b = 0 \quad (2.6.1.2)$$

donde $\mathbf{w}$ es normal al hiperplano, por lo tanto $\frac{|b|}{\|\mathbf{w}\|_2}$ es la distancia perpendicular al origen y $\|\mathbf{w}\|_2$ es la norma euclídea de $\mathbf{w}$.

Si consideramos que los puntos más cercanos al hiperplano están a una distancia d_+ y d_- al hiperplano, se define el “margen” de separación como la distancia $d_+ + d_-$. Con el máximo margen, los datos de entrenamiento cumplen con las siguientes condiciones:

$$\mathbf{w} \cdot \mathbf{x}_i + b \geq +1 \text{ para } y_i = +1 \quad (2.6.1.3)$$

$$\mathbf{w} \cdot \mathbf{x}_i + b \leq -1 \text{ para } y_i = -1 \quad (2.6.1.4)$$

Lo que es equivalente a:

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1 \geq 0 \text{ para todo } i \quad (2.6.1.5)$$

Considerando 2 hiperplanos H_1 y H_2 paralelos equidistantes a una distancia d desde el hiperplano central y que en estos hiperplanos están situados los vectores soporte. La distancia entre H_1 y H_2 es $2/\|\mathbf{w}\|_2$ por lo tanto la máxima distancia se la obtiene minimizando:

$$\frac{1}{2} \|\mathbf{w}\|_2^2 \text{ sujeto a las restricciones } y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1 \geq 0 \text{ para todo } i \quad (2.6.1.6)$$

La solución de este problema se realiza utilizando técnicas de programación cuadrática.

A partir del problema de optimización definido en (2.6.1.6) se obtiene la función Lagrangiana (Boyd, & Vandenberghe, 2018) es:

$$L_p = \frac{1}{2} \|\mathbf{w}\|_2^2 - \sum_1^l \alpha_i y_i(\mathbf{w} \cdot \mathbf{x}_i + b) + \sum_1^l \alpha_i \quad (2.6.1.9)$$

Donde α_i son los multiplicadores de Lagrange aplicados a las ecuaciones de restricciones.

De acuerdo con la teoría de la optimización (Deisenroth, Faisal, & Ong, 2019), un problema de optimización tiene su dual si la función a optimizar y sus restricciones son convexas. Considerando que se puede demostrar que (2.6.1.6) y (2.6.1.7) son convexas y

por lo tanto el problema se puede resolver utilizando el problema dual (Wolfe de dual): maximizar L_p sujeto a las restricciones que el gradiente de L_p con respecto a $\mathbf{w}$ y b sea igual a cero y además con la restricción que el $\alpha_i \geq 0$ (Burges, 1998).

El gradiente de L_p respecto a $\mathbf{w}$ y b y aplicando las condiciones de Karush-Kuhn-Tucker (KKT) (Boyd, & Vandenberghe, 2018):

$$\frac{\partial L(\mathbf{w}, b, \alpha)}{\partial (\mathbf{w})} = \mathbf{w} - \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i = 0 \quad (2.6.1.10)$$

$$\frac{\partial L(\mathbf{w}, b, \alpha)}{\partial (b)} = \sum_{i=1}^n \alpha_i y_i = 0 \quad (2.6.1.11)$$

$$\alpha_i [1 - y_i (\mathbf{w} \cdot \mathbf{x}_i) + b] = 0 \quad (2.6.1.12)$$

por lo tanto:

$$\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i \quad (2.6.1.13)$$

$$\sum_i \alpha_i y_i = 0 \quad (2.6.1.14)$$

$$\alpha_i [1 - y_i (\mathbf{w} \cdot \mathbf{x}_i + b)] = 0, i = 1, \dots, n \quad (2.6.1.15)$$

Reemplazando estas condiciones en L_p :

$$L_D = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) \quad (2.6.1.16)$$

De acuerdo con Karush Kuhn Tucker (KKT) la gran mayoría de los coeficientes de Lagrange son cero y sólo pueden ser distintos de cero para los vectores de soporte, puntos que se encuentran exactamente a la distancia marcada por el margen. La maximización del margen, se transforma en un problema de minimización de una función cuadrática convexa sujeta a restricciones lineales en el modelo dual.

La solución del problema dual consiste en maximizar L_D con respecto a α_i sujeto a las restricciones en (2.6.1.9), (2.6.1.10) y $\alpha_i > 0$.

Hay un α_i para cada punto de entrenamiento. En la solución los puntos para los cuales $\alpha_i > 0$ son llamados vectores soporte y yacen en uno de los hiperplanos H1 o H2. El resto de puntos tienen $\alpha_i = 0$ y satisfacen a la ecuación (2.6.1.5). Por lo tanto, se puede afirmar que el hiperplano, se construirá solo como una combinación lineal de los vectores de soporte.

La solución del problema dual se reemplaza en la ecuación (2.6.1.1) y para encontrar el valor de b se hace uso de la ecuación (2.6.1.11) en la cual, si consideramos $\alpha_i > 0$, el factor de la derecha tendrá que ser necesariamente 0, por lo tanto:

$$[1 - y_i(\mathbf{w} \cdot \mathbf{x}_i + b)] = 0 \quad (2.6.1.17)$$

Despejando b ,

$$b = y_{vs} - (\mathbf{w} \cdot \mathbf{x}_{vs}) \quad (2.6.1.18)$$

Donde

$(\mathbf{x}_{vs}, y_{vs})$ representa la tupla del vector soporte ($\alpha_i > 0$), junto con su valor de clase. Para que el valor de b sea más robusto, se lo obtiene del promedio de todos los vectores soporte.

$$b^* = \frac{1}{N_{Vs}} \sum_{j \in Vs} [y_j - (\mathbf{w}^* \cdot \mathbf{x}_j)] \quad (2.6.1.19)$$

Donde Vs es el conjunto de los vectores soporte y N_{Vs} es la cardinalidad del conjunto Vs .

Para tomar en cuenta puntos potencialmente mal etiquetados, casos donde existe datos de entrada erróneos, ruido o alto solapamiento de las clases en los datos de entrenamiento, reescribimos las restricciones para obtener el mejor hiperplano clasificador que pueda tolerar el ruido en los datos de entrenamiento, al que se denomina SVM con margen blando. Esto se logra introduciendo una variable de holgura no negativa ξ_i .

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1 \geq 1 - \xi_i \quad (2.6.1.7)$$

ξ_i mide que tan lejos está un punto x_i desde el correcto margen por lo tanto si $0 \leq \xi_i < 1$, el punto se encuentra en el lado correcto del límite de decisión y si $\xi_i > 1$ el punto es asignado a la clase opuesta.

En este caso, el problema matemático queda definido como la minimización de:

$$\frac{1}{2} \|\mathbf{w}\|_2 + C \sum_{i=1}^l \xi_i \quad (2.6.1.8)$$

Sujeto a las restricciones: $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1 \geq 1 - \xi_i$ y $\xi_i \geq 0$

En el que se incluye un término de regularización $C \sum_{i=1}^l \xi_i$ el cual, depende de la variable de holgura y del parámetro C, que determina la holgura del margen blando. C es determinado a priori y su elección influye en el desempeño de la SVM.

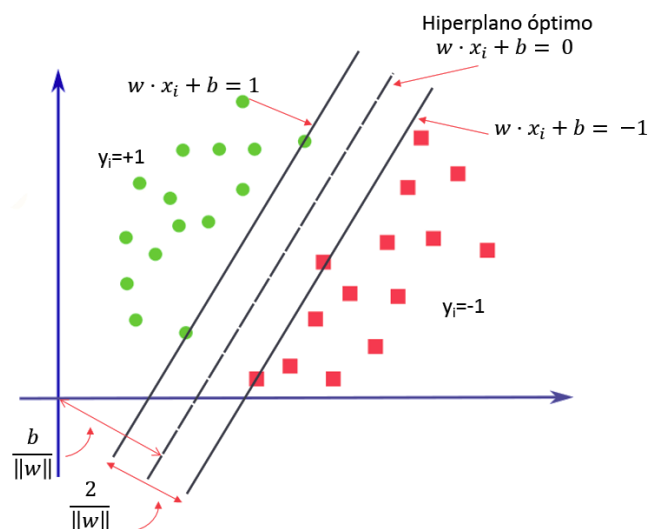


Figura 2.19 Clasificación con SVM.

2.6.2 El truco del Kernel

La implementación de SVM se realiza utilizando una técnica llamada el “truco del kernel”. La función Kernel transforma un set de entrada de un espacio de dimensión baja a en un espacio de alta dimensión, de esta forma convierte un problema no separable en un problema separable agregándole más dimensiones, principalmente cuando la separación no es lineal. Para esto, suponemos que los datos son llevados a un nuevo espacio euclideo H (posiblemente de dimensión infinita denominado espacio de *Hilbert*), mediante una transformación Φ y una función K , que calcula el producto punto de los puntos de entrada tal que $K(x_i \cdot x_j) = \Phi(x_i) \cdot \Phi(x_j)$. Al sustituir en la función del hiperplano (Deisenroth, Faisal, & Ong 2019):

$$D(x) = \sum_{i=1}^n \alpha_i y_i K(x_i \cdot x_i) \quad (2.6.2.1)$$

La solución corresponde a encontrar los valores de α_i que optimiza el problema dual y con ellos calcular el hiperplano que separa las clases.

2.6.3 Funciones Kernel

Kernel Lineal

$$K(x_i \cdot x_j) = (x_i \cdot x_j) \quad (2.6.2.2)$$

Kernel Polinomial

$$K(x_i \cdot x_j) = [x_i \cdot x_j + 1]^d \quad (2.6.2.3)$$

Kernel de base radial gaussiano (RBF)

$$K(x_i \cdot x_j) = e^{-\|x_i - x_j\|^2 / 2\sigma^2} \quad (2.6.2.4)$$

Kernel simoidal

$$K(x_i \cdot x_j) = \tanh(\gamma(x_i \cdot x_j) - \delta) \quad (2.6.2.5)$$

2.7 REDES NEURONALES ARTIFICIALES

Las redes neuronales artificiales (ANN Artificial Neural Networks) son sistemas que basan su funcionamiento en el comportamiento de las neuronas humanas y modelan la forma en que el cerebro humano resuelve los problemas. Las neuronas humanas se conectan entre sí por los impulsos nerviosos generados por axones y dendritas que convierten sustancias químicas en impulsos eléctricos que se transmiten a través de regiones llamadas sinapsis (figura 2.20). El cerebro humano es un sistema muy complejo, no lineal y paralelo lo que le da la capacidad de realizar muchas operaciones simultáneamente. Para lograr esto, muchas neuronas se interconectan, cada una con capacidad de realizar el procesamiento de la información en forma paralela y distribuida que en conjunto toma el nombre de red neuronal (Dony & Haykin, 1995).

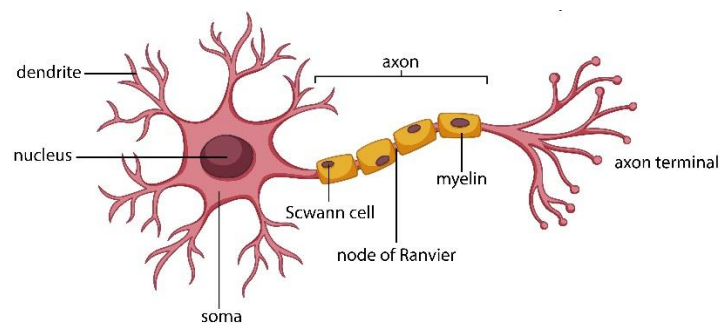


Figura 2.20 Neurona biológica.

El rendimiento de algunos sistemas naturales ha sido modelado con éxito por sistemas artificiales. Para emular las tareas que realiza el cerebro, las ANN se comportan de acuerdo con la forma que están conectadas las neuronas que lo componen y la fuerza de

las conexiones está dada por los pesos (w). Los pesos son ajustados automáticamente durante el entrenamiento hasta que la tarea sea ejecutada correctamente.

Las neuronas se agrupan de acuerdo a su comportamiento similar formando capas neurales que pueden resolver problemas lineales o no lineales. Las redes de propagación hacia atrás (retropropagación) son las más utilizadas en clasificación de imágenes. En este tipo de redes el patrón de entrada se propaga hacia delante y los errores se transmiten desde la salida de la red hacia capa de entrada (Lotfi et al., 2009). ANN puede ser utilizada para mapear diferentes tipos de vegetación en ambientes complejos aunque debe considerarse la elevada demanda de sistemas de computación cuando los datos son extensos (Adam, Mutanga, & Rugege, 2010). Su aplicación en la detección de enfermedades en plantas es muy extendida.

2.7.1 Perceptrón

El perceptrón, como se denomina a una neurona artificial, es un clasificador binario de modelo lineal que tiene una sola capa de entrada y un nodo de salida. Fue inventada en 1957 por Frank Rosenblatt, quien inicialmente la llamó Unidad de Umbral Lineal (Linear Threshold Unit LTU). El perceptrón es la estructura básica de procesamiento de las redes neuronales (figura 2.21). Consiste en un algoritmo que recibe un vector x de m valores ($x_1, x_2, \dots, x_n$) que son llamados características y genera una salida de clasificación con 2 valores 1 o 0. Para hacer esto, suma el producto de cada entrada (x_i) con su peso asociado (w_i) y la envía a una función de activación de escalón (Heaviside function) que permite la clasificación basada en un valor de umbral.

Este comportamiento se expresa matemáticamente con la siguiente función:

$$y = \begin{cases} 1 & \text{si } z \geq 0 \\ 0 & \text{si } z < 0 \end{cases} \quad z = \mathbf{w} \cdot \mathbf{x} + b \quad (2.5.1.1)$$

Donde,

$\mathbf{w}$ es el vector de pesos, $\mathbf{x}$ es el vector de entradas

$\mathbf{w} \cdot \mathbf{x}$ es el producto punto entre $\mathbf{w}$ y $\mathbf{x}$

b es el sesgo.

Si $\mathbf{x}$ se ubica sobre la línea recta, el resultado z es positivo y si se sitúa debajo de la recta, el resultado z es negativo.

y es la variable de clasificación, si z es positivo, tendrá valor 1 y si z es negativo tendrá el valor 0.

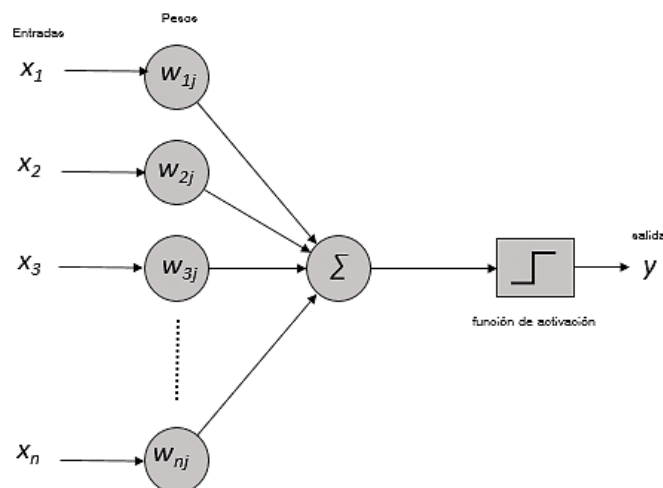


Figura 2.21 Arquitectura del perceptrón.

2.7.2 Perceptrón multicapa (Multi-Layer Perceptron MLP)

El perceptrón es una red de una capa simple (simple layer perceptron SLP). Una red formada por varias capas es llamada perceptrón multicapa (MLP) o red neuronal

multicapa, en las cuales las capas de entrada y salida están separadas por un grupo de capas ocultas (hidden layers).

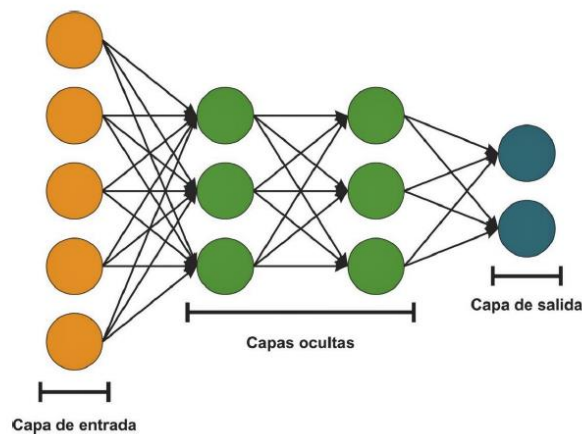


Figura 2.22 Perceptrón multicapa (MLP).

Esta arquitectura en capas de redes neuronales es conocida como redes de propagación hacia delante (feed forward network) porque cada capa alimenta a la capa siguiente desde la entrada hasta la salida (forward propagation) (figura 2.22). De la misma manera que el perceptrón simple, en las MLP hay un algoritmo de aprendizaje que cambia los pesos y el sesgo para cada neurona artificial. Luego se calcula el coste y se propaga hacia atrás, este proceso toma el nombre de retropropagación (backpropagation). Para optimizar los pesos y sesgos se busca que la función de pérdida sea cero utilizando la técnica descenso de gradiente (gradient descent) (figura 2.23).

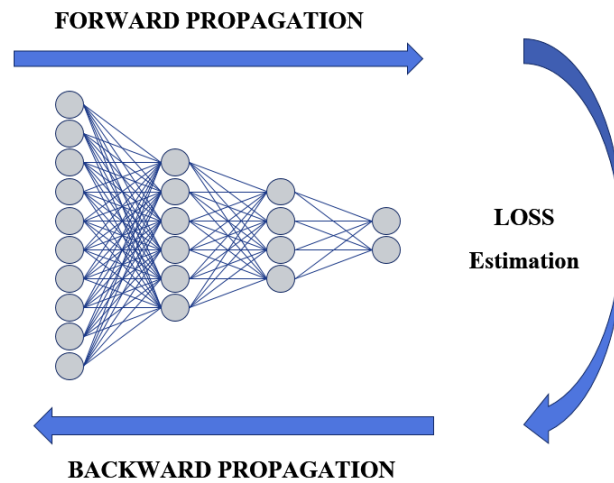


Figura 2.23 Proceso de aprendizaje del perceptrón multicapa (MLP).

El uso del principio inductivo ERM (Empirical Risk Minimization) en el que se basan las MLP, conduce a una buena generalización si elige los parámetros de la función aproximadora de tal manera que aseguren el número mínimo de errores sobre el conjunto de entrenamiento. Si la selección no es adecuada se puede provocar sobreentrenamiento o sobreajuste (overffiting), el mismo que puede ser solucionado mediante la aplicación de técnicas como criterios de parada temprana en el entrenamiento (early stopping rules), el decaimiento de pesos (weigth decay) o la regularización.

2.7.3 Funciones de activación

En una neurona, la suma ponderada de las entradas, al que llamamos valor pre-activación, es transformada por una función de activación, dando como resultado un valor post-activación. Para la capa de entrada la función de activación es lineal. En las capas ocultas, las entradas son el resultado de la activación de las capas anteriores. Si

consideramos que g es la función de activación, la salida de una neurona, es decir, el valor post-activación, está dado por la expresión (Patterson & Gibson, 2017):

$$z(x) = g(wx + b) \quad (2.7.3.1)$$

Entre las funciones de activación más utilizadas en redes neuronales podemos citar a la función de activación lineal, sigmoide (*sigmoid*), tangente hiperbólica (*tanh*), *ReLU* y tangente hiperbólica fuerte (*htanh*) (Aggarwal, 2018).

Función de activación lineal. - es la función más básica y también es conocida como función identidad, donde la variable dependiente tiene una relación directa y proporcional con la variable independiente. Es aplicada al nodo de salida cuando se desea obtener un valor real.

$$z(x) = x \quad (2.7.3.2)$$

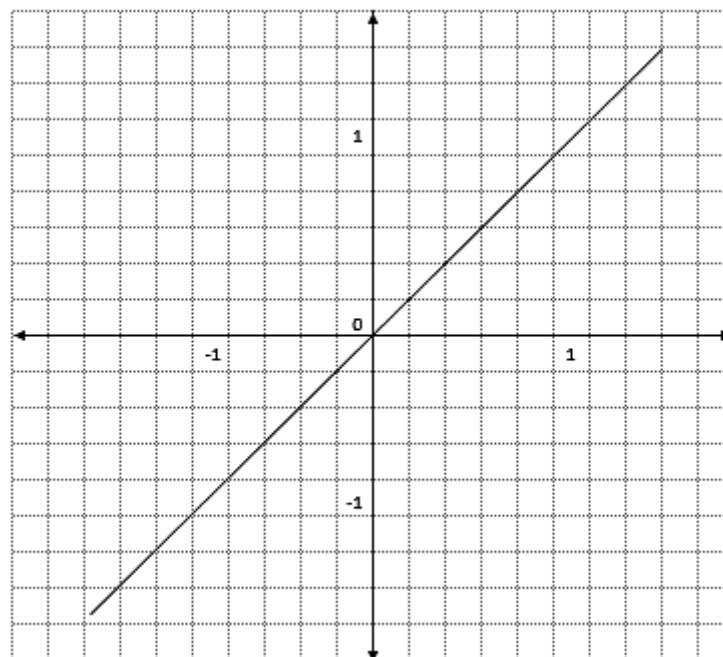


Figura 2.24 Función de activación lineal.

Función *signo*. - Esta función puede ser utilizada para la asignación de valores binarios a las salidas en la fase de predicción. Devuelve 1 si el argumento número es positivo y -1 si el argumento número es negativo.

$$z(x) = \text{sign}(x) \quad (2.7.3.3)$$

$$z(x) = \begin{cases} 1 & \text{para } x > 0 \\ -1 & \text{para } x < 0 \end{cases} \quad (2.7.3.4)$$

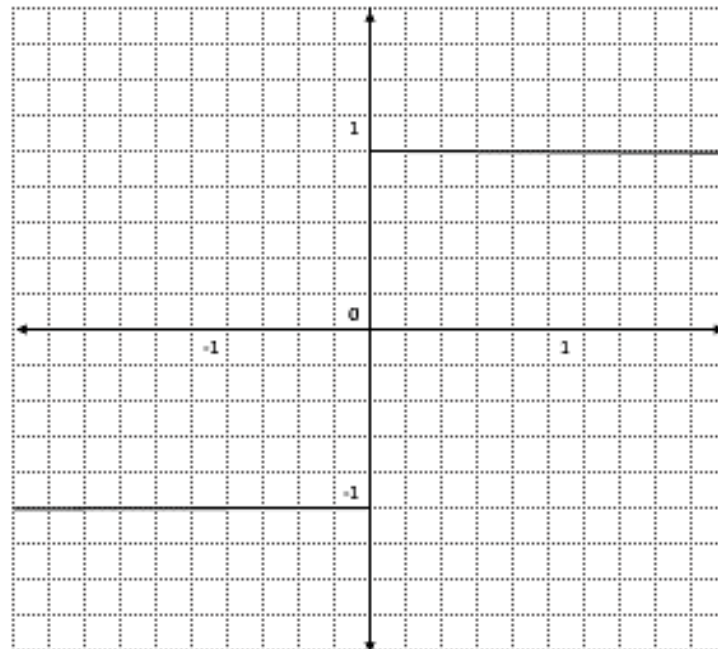


Figura 2.25 Función de activación *sign()*.

Función *sigmoide*. – La salida de la función *sigmoid()* está en rango entre 0 y 1 y puede ser aplicada cuando se requiera una salida de probabilidad y funciones de pérdida a partir de modelos de máxima verosimilitud.

$$z(x) = \frac{1}{1+e^{-x}} \quad (2.7.3.5)$$

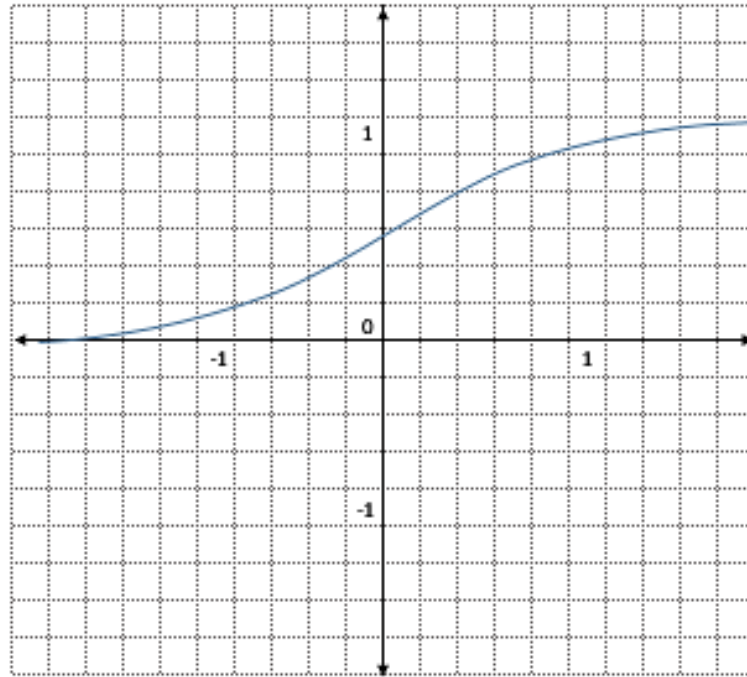


Figura 2.26 Función de activación sigmoid().

Función tangente hiperbólica (*tanh*). – tiene la forma similar a la *sigmoide* pero los valores de salida están entre -1 y 1. La función $\tanh(x)$ representa la razón entre el $\sinh(x)$ y $\cosh(x)$. Su fortaleza respecto a la función *sigmoide* es que puede manejar más fácilmente los números negativos.

$$z(x) = \frac{e^{2x}-1}{e^{2x}+1} \quad (2.7.3.6)$$

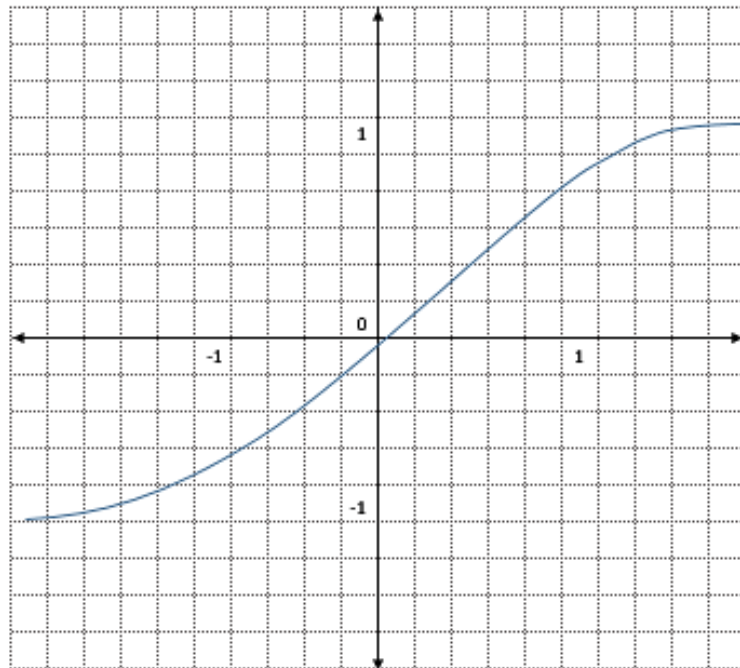


Figura 2.27 Función de activación $\tanh()$.

Se relaciona con la función *sigmoide* por medio de la siguiente ecuación:

$$\tanh(x) = 2\text{sigmoid}(2x) - 1 \quad (2.7.3.7)$$

Función *ReLu* (Rectified Linear Unit). – Esta función activa el nodo si la entrada es positiva, dando una salida relacionada linealmente con la entrada y en el caso en que la entrada es negativa, la salida es cero. La función *ReLu* tiene mejor funcionamiento en el entrenamiento que la función *sigmoide* porque su gradiente es cero o una constante.

$$z(x) = \max\{x, 0\} \quad (2.7.3.8)$$

Función *Softplus*. – Es otra variante de *ReLU* llamada la versión suave de *ReLU*.

$$z(x) = \ln[1 + e^x] \quad (2.7.3.9)$$

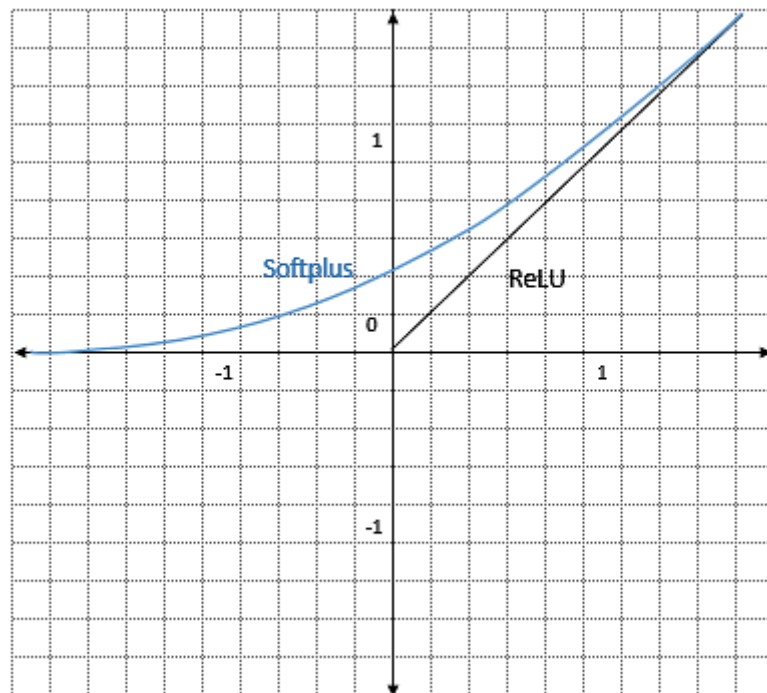


Figura 2.28 Funciones de activación ReLu() y softplus().

Función *Hard tanh*. - La ventaja de la función *Hard tanh* sobre la original *tanh* es que representa una carga computacional menor.

$$z(x) = \max\{\min[x, 1], 1\} \quad (2.7.3.10)$$

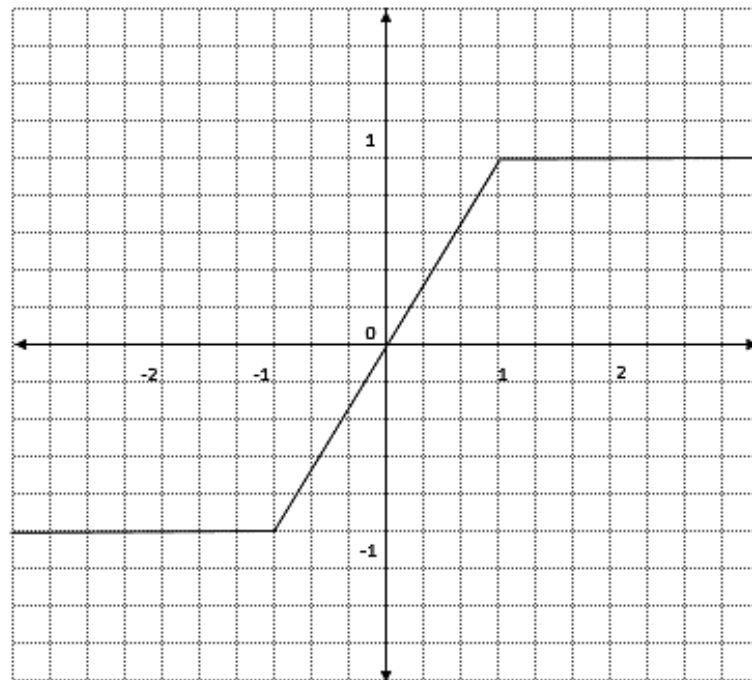


Figura 2.29 Función de activación Hard tanh().

Función *Leaky ReLu*. – Es una variante de *ReLu* en que incluye una pequeña pendiente en los valores negativos.

$$z(x) = \begin{cases} x, & x > 0 \\ 0.01x, & \text{otro caso} \end{cases} \quad (2.7.3.11)$$

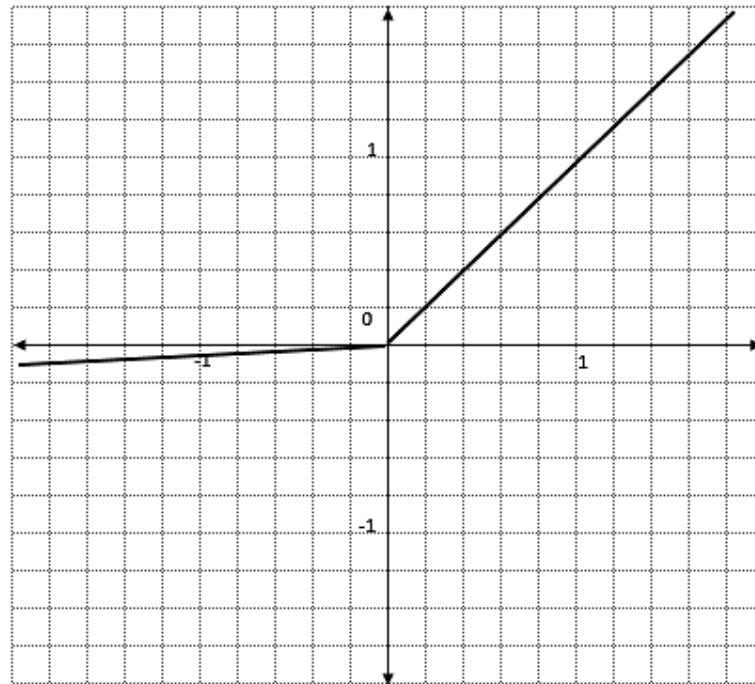


Figura 2.30 Función de Activación *Leaky ReLu*.

Función softmax. – Es una generalización de regresión logística con datos continuos, puede contener múltiples límites de decisión y maneja etiquetas de tipo multinomial.

Esta función es utilizada generalmente, en la capa de salida de un clasificador. Si se tiene un alto número de etiquetas (miles) se utiliza se utiliza una variante de *softmax* llamada *hierarchical softmax*. *Hierarchical softmax* crea una estructura jerárquica para las etiquetas y el clasificador *softmax* es entrenado en cada nodo del árbol para realizar la clasificación (Patterson & Gibson, 2017). También se la conoce con el nombre de función exponencial normalizada.

$$z(x) = p(y = j|x) = \frac{e^{(w_j^T x + b_j)}}{\sum_{k \in K} e^{(w_k^T x + b_k)}} \quad (2.7.3.12)$$

Donde,

j es j -ésima categoría de un total de K categorías.

2.7.4 Dimensionalidad de las capas y conexiones

En la arquitectura feed-forward network todos los nodos de una capa son conectados a la siguiente capa. La arquitectura queda casi definida cuando se establece el número de capas y el número y tipo de nodo en cada capa, quedando por definir únicamente la función de pérdida (loss function). El número de neuronas en cada capa es la dimensionalidad de esa capa, por lo tanto, si una red neuronal tiene k capas ocultas y cada capa oculta tiene $p_1, p_2, \dots, p_k$ neuronas en cada una de sus capas, entonces los vectores salida de cada capa oculta son referidos como $h_1, h_2, \dots, h_k$ con dimensionalidad $p_1, p_2, \dots, p_k$.

El vector de entrada x con dimensión d genera una matriz W_1 de pesos en las conexiones entre la capa de entrada y la primera capa oculta. La matriz W_1 tendrá un tamaño $d \times p_1$ y la matriz de pesos W_r entre las capas ocultas r y $(r+1)$ tendrá un tamaño $p_r \times p_{(r+1)}$. Finalmente, la capa de salida con o nodos recibirá W_k conexiones con un tamaño de $p_k \times o$. Las entradas a cada capa corresponden a las siguientes ecuaciones:

$$h_1 = z(W_1^T x + b_1) \quad (2.7.3.13)$$

$$h_{p+1} = z(W_{p+1}^T h_p + b_{p+1}) \quad (2.7.3.14)$$

$$O = z(W_{k+1}^T h_k + b_{k+1}) \quad (2.7.3.15)$$

2.7.5 Función de pérdida

El proceso de optimización requiere maximizar o minimizar la función objetivo. En redes neuronales se busca minimizar el error por lo que la función de pérdida es denominada función de coste o función de pérdida. El cálculo del error del modelo en el proceso de optimización es fundamental para lograr resultados satisfactorios, por lo que es necesaria una adecuada selección de la función de pérdida. Esta función debe ofrecer un conjunto de soluciones que puedan ser mapeadas con los resultados del algoritmo de optimización mediante actualizaciones iterativas de los pesos del modelo. Entre las funciones de pérdida más utilizadas se encuentran MSE (Mean Squared Error), binary cross-entropy y categorical cross-entropy (Gulli & Pal, 2017).

Error cuadrático medio (Mean Squared Error MSE)

MSE es el error cuadrático medio entre las predicciones y los valores verdaderos observados. El objetivo de esta función es obtener el promedio de los errores de todas las predicciones.

$$MSE = \frac{1}{n} \sum_{i=1}^n (\hat{Y} - Y)^2 \quad (2.7.3.15)$$

Donde $\hat{Y}$ es un vector de n predicciones y Y es un vector de n valores observados.

Entropía cruzada binaria (binary cross-entropy)

Binary cross-entropy es una función de pérdida logarítmica binaria adecuada para predicciones de etiquetas binarias. La entropía cruzada entre la distribución empírica

definida por el conjunto de datos de entrenamiento y el modelo es matemáticamente equivalente a la logverosimilitud negativa.

Está definida por:

$$L(W, b) = -\frac{1}{N} \sum_{i=1}^N y_i \log(\hat{y}_i) - (1 - y_i) \log(1 - \hat{y}_i) \quad (2.7.3.16)$$

Donde

W son los pesos y b es el sesgo

y_i clasificación objetivo

$\hat{y}_i$ son las predicciones

Entropía cruzada categórica (categorical cross-entropy)

Categorical cross-entropy es una función de pérdida logarítmica categórica que se utiliza para predicciones de etiquetas de categorías múltiples, se define por la siguiente función:

$$L(W, b) = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^M y_{i,j} \log(\hat{y}_{i,j}) \quad (2.7.3.17)$$

Donde

W son los pesos y b es el sesgo

M es el número de clases

y_i categorías objetivo

$\hat{y}_i$ son las predicciones

3 MATERIALES Y MÉTODOS

3.1 ORGANISMOS

3.1.1 Plantas

Banano Giant Cavendish del cultivar “Williams”.

100 plantas en la fase de establecimiento (3-4 meses en esta fase), provenientes de invernaderos de locales de propagación de banano ubicados en Guayaquil (Ecuador), fueron transportadas al invernadero del Centro de Investigaciones Biotecnológicas del ECUADOR (CIBE), en donde se las mantuvo a temperatura de 28 ° C, 70% HR (humedad) con luz natural (12 horas) y regadas cada 48 horas.

3.1.2 Patógeno

Hongo patógeno *Pseudocercospora fijiensis* del género *Ascomycete Mycosphaerella fijiensis* Morelet.

3.2 EQUIPOS

3.2.1 Sistema de adquisición de datos Hiperespectrales.

Para realizar la adquisición de las imágenes se utilizó un sistema de imágenes hiperespectrales diseñado y construido en el Centro de Visión por Computador y Robótica de la ESPOL (CVR-ESPOL).

El sistema está compuesto por un espectrómetro ImSpector V10E (Spectral Imaging Ltd.) conectado a una cámara 1500M-GE (Thorlabs). Estos dispositivos se montaron sobre un deslizador (WS70 Excitron) controlado por computador como se muestra en la figura 3.1.

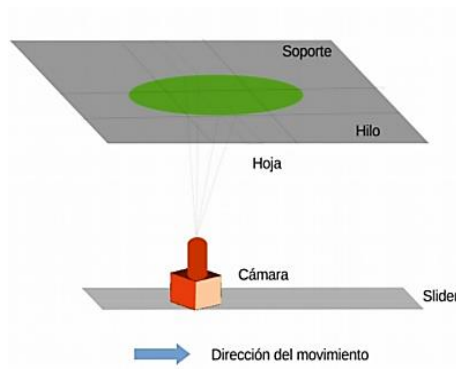


Figura 3.1 Sistema de Imágenes Hiperespectrales.

El ImSpector V10E es un escáner de línea (push-broom) que tiene una resolución espectral de 4.55 nm y opera en un rango espectral entre 386 y 1019 nm. La resolución espacial es de 0.5 mm (alto) x 176.6 mm (ancho), lo que implica que la luz de una sección de 0.5 x 176.6 mm del objeto es descompuesta por el espectrómetro. Para un pixel determinado, la luz que llega corresponde a una superficie de 0.125 x 0.125 mm. Dado que el sensor CCD de la cámara tiene resolución de 1040 (filas) x 1392 (columnas) píxeles y el espectro se proyecta de forma continua sobre el sensor, por cada imagen se obtienen 1040 muestras del espectro. Cada pixel es representado por 12 bits.

Durante el proceso de escaneo, la planta es colocada en la parte superior mientras la cámara se desplaza logrando un barrido completo de la hoja seleccionada. Dos lámparas halógenas de 50 W. iluminan el objetivo. El Sistema está controlado por un computador con una capacidad de almacenamiento de 1 TB, un procesador Intel Core i5 3.1 GHZ y 16 GB de RAM.

El sistema fue configurado para generar un cubo HS de dimensiones espaciales de 205 filas (M), 198 columnas (N) y una dimensión espectral de 520 longitudes de onda (J) para cada hoja. Para obtener el ancho de un cubo HS (dimensión N), se estableció un pixel binning-x (agrupamiento en x) de 7 píxeles, lo que resulta en una reducción de 1392 píxeles a 198 píxeles. Para calcular el número de longitudes de onda (J) de un cubo HS, configuramos un pixel binning-y (agrupamiento en y) de 2 píxeles, lo que resulta en una reducción de 1040 píxeles a 520 píxeles. Finalmente, la altura de un cubo HS (M) se estableció de acuerdo con la velocidad de fotogramas de adquisición de la cámara para escanear todo el soporte de la hoja, lo que resultó en 205 píxeles.

La calibración del sistema es crucial para la obtención de imágenes con la mejor calidad. En esta aplicación se utilizaron métodos estándar de calibración de imagen.

La calibración espectral se llevó a cabo utilizando fuentes de luz de espectros conocidos como mercurio (Hg), argón (Ar), helio (He) e hidrógeno (H). Los espectros generados se utilizaron para estimar la longitud de onda correspondiente a cada línea del dispositivo de acoplamiento de carga (CCD) relacionando los picos de los espectros de los gases con las posiciones de las bandas espectrales en la imagen. La relación fue establecida mediante la siguiente ecuación:

$$\lambda(row_i) = 0,000022 * row_i^2 + 0,586019 * row_i + 386,829(nm) \quad (3.2.1.1)$$

Donde,

row_i es la posición de la banda espectral

$\lambda_{(row_i)}$ es la longitud de onda en la posición row_i

En segundo lugar, se realizó la calibración radiométrica para reducir la influencia de las variaciones de intensidad de la luz y el ruido del sensor CCD. Para este propósito, se tomaron imágenes de referencias blancas y oscuras antes de cada sesión de exploración y se normalizó la imagen espectral en bruto. Para la normalización de la radiación medida se utilizó la siguiente ecuación:

$$R_\lambda = \frac{I_\lambda - D_\lambda}{W_\lambda - D_\lambda} \quad (3.2.1.2)$$

Donde,

R_λ es la reflectancia en la longitud de onda λ .

I_{λ} es la intensidad de luz medida en la longitud de onda λ .

W_{λ} es la intensidad de referencia (blanco) medida en la longitud de onda λ .

D_{λ} es la intensidad obtenida por el sensor cuando no recibe luz (negro).

A continuación, se realizó la calibración espacial que consiste en regular el movimiento del slider para que la secuencia de frames capturados (líneas de barrido) no se traslapen o queden espaciados. Si el espectrómetro se mueve muy rápido, se obtiene información incompleta del objeto. Por otro lado, si el espectrómetro se mueve muy lento, se registra información redundante produciéndose el traslape de las regiones. El primer caso se denomina submuestreo y el segundo sobremuestreo. Para eliminar este problema se calcula la velocidad de traslación del slider que permita muestrear espacialmente un objeto de forma correcta.

Finalmente, para obtener una imagen que contenga únicamente los de onda de 700 nm. y ponemos a cero el soporte de la hoja y los píxeles de fondo. Utilizamos píxeles que corresponden a la hoja, generamos una máscara segmentando la imagen a una longitud está máscara para eliminar automáticamente los elementos de la imagen que no corresponden a la hoja.

Más detalles de la construcción, calibración y funcionamiento del sistema se pueden obtener en el trabajo de Ochoa et al. (2016).

El sistema fue instalado en el Centro de Investigaciones Biotecnológicas del ECUADOR (CIBE).

3.2.2 Software

Los programas fueron desarrollados utilizando las siguientes herramientas:

Python versión 3.7, Anaconda, Jupiter Notes, Google Colab.

RStudio 1.1, R versión 3.5,

3.3 INOCULACIÓN DE PLANTAS

Con el fin de producir la enfermedad en las plantas se aplicó un protocolo de inoculación del patógeno manteniendo las condiciones necesarias para que el inóculo colonice la planta y se desarrolle la enfermedad.

Dieciséis plantas fueron seleccionadas al azar del invernadero, 10 fueron inoculadas con *P. fijiensis* y en 6 plantas de control se realizó una inoculación simulada (mock). La inoculación de la planta se realizó según el proceso planteado por Gbongue et al. (2019). Brevemente, hongos *P. fijiensis* aislados fueron inoculados sobre agar de papa y dextrosa (PDA) y se incubaron durante 2 semanas a 30 ° C. Luego, el micelio fue sumergido en 10 ml de agua estéril y se filtró para separar el micelio de los conidios. A continuación, la suspensión conidial se concentró por centrifugación a 3000 x g durante 10 minutos a 4 ° C.

Las hojas de banano se inocularon con la suspensión concentrada de conidios usando un atomizador de aerógrafo y los síntomas de la enfermedad se monitorearon usando la escala de severidad que se muestra en la tabla 1-1 (Fouré, 1986). Los síntomas visuales en cada etapa de la enfermedad se muestran en la figura 3.2. Las plantas de control se inocularon de forma simulada con agua destilada en autoclave.

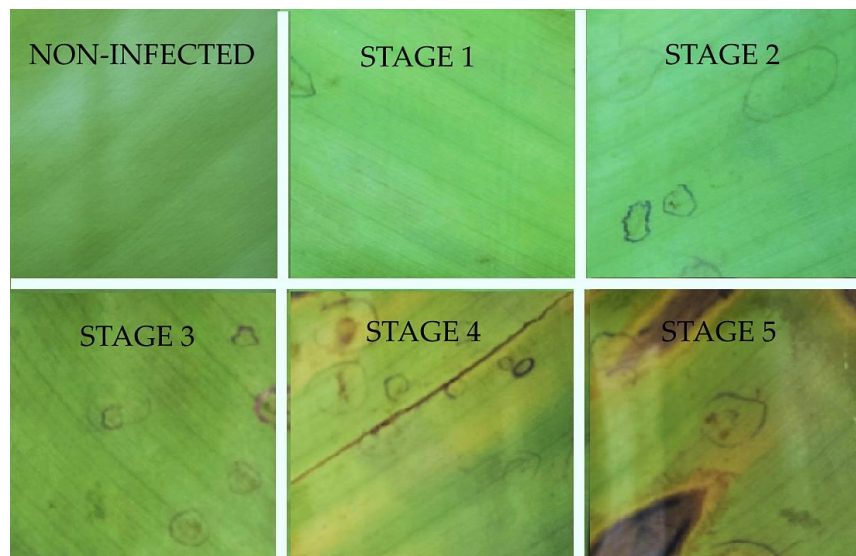


Figura 3.2 Etapas de la enfermedad Sigatoka negra.

Las marcas circulares fueron realizadas por los biólogos para resaltar las áreas afectadas.

3.4 ADQUISICIÓN DE IMÁGENES

En cada una de plantas seleccionadas en el apartado anterior, se identificaron las hojas que serían escaneadas y etiquetadas. Se seleccionaron tres hojas de las plantas de control (6), dos se dañaron debido a la manipulación lo que dio como resultado un subtotal de 16 imágenes de hojas no infectadas. De las plantas inoculadas (10), se seleccionaron dos hojas de cada planta y fueron escaneadas cada 3 días durante 3 meses. En cada sesión de escaneo, un experto evaluó los síntomas visualmente utilizando la escala de síntomas detallada en la tabla 1-1 y fueron etiquetadas. Durante este período, la progresión de la enfermedad en las hojas fue desigual. Los síntomas del nivel de severidad 1 aparecieron entre 7 y 31 días y aumentaron de manera irregular alcanzando niveles de gravedad más altos en diferentes períodos de tiempo. En algunas hojas, la enfermedad alcanzó el nivel de gravedad 5. Debido a la manipulación, varias hojas se dañaron y, por lo tanto, fueron descartadas durante el experimento.

De las imágenes escaneadas y etiquetadas por los expertos, seleccionamos las que pertenecen a las hojas infectadas en las etapas pre-sintomática, severidad 1 y severidad 2. Las imágenes pre-sintomáticas (16) fueron aquellas obtenidas seis días antes de que la hoja presentara síntomas de severidad 1. Luego, se tomaron las siguientes imágenes en intervalos de 6 días durante la progresión del nivel de severidad 1 (54) y el nivel de severidad 2 (18).

El conjunto de datos final consistió en 104 imágenes (16 no infectadas, 16 pre-sintomáticas, 54 niveles de gravedad 1 y 18 niveles de gravedad 2). El nivel de severidad reportado correspondió a la etapa de enfermedad más alta encontrada en la hoja.

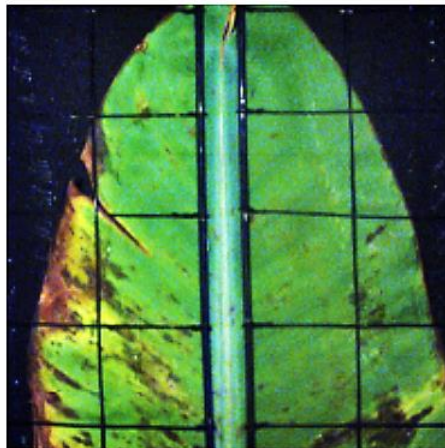


Figura 3.3 Imagen de una hoja escaneada en el sistema HSI.

3.5 PRE-PROCESAMIENTO DE DATOS

Para corregir diferencias de escala en las mediciones de reflectancia producidas por efectos de longitud de trayectoria y variaciones de fuente o detector u otros efectos relacionados a la sensibilidad instrumental, principalmente para compensar las variaciones de reflectancia debido a la orientación relativa de la superficie de la hoja y el sensor, cada cubo HS fue normalizado utilizando la técnica de varianza normal estándar (SNV).

$$x_{inorm} = \frac{x_i - \bar{x}_i}{\sigma_i} \quad (3.5.1)$$

Donde,

x_{inorm} es la reflectancia normalizada para una posición i .

x_i es la reflectancia sin normalizar en una posición i .

$\bar{x}_i$ es la media de la reflectancia en todas las longitudes de onda para una posición i .

σ_i es la desviación estándar de la reflectancia en una posición i .

Finalmente, se realizó una reducción de la dimensionalidad del cubo HS, calculando el promedio de los valores de reflectancia medidos en cada longitud de onda, dando como resultado una matriz de $I = 104$ filas (una por imagen) y $J = 520$ columnas (figura 3.4).

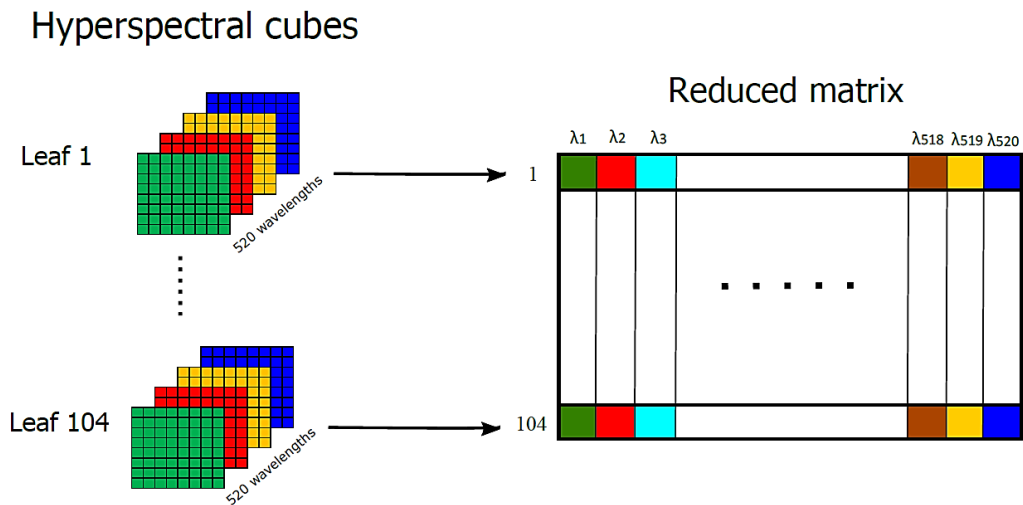


Figura 3.4 Reducción de cubos hiperespectrales

Un análisis preliminar de las regiones infectadas mostró diferencias en los patrones espectrales de las etapas de la enfermedad. La figura 3.5 muestra los patrones de reflectancia SNV para el nivel de gravedad 2 y las regiones no infectadas.

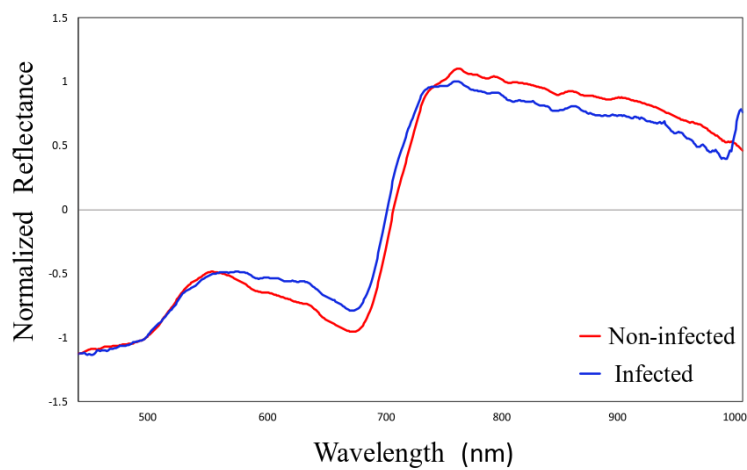


Figura 3.5 Patrones de reflectancia SNV de regiones sanas e infectadas.

3.6 MÉTODOS ESTADÍSTICOS

Los datos descritos en el apartado 3.5 fueron organizados en un vector y que corresponde a la variable respuesta binaria (infectado y no infectado) y una matriz X ($x_1, \dots, x_p$) que corresponde a la matriz reducida de dimensiones 104×520 formada por un set de predictores que cuantifican intensidad de la reflectancia normalizada en cada longitud de onda medida en 104 imágenes de hojas de banano (104 hojas $\times$ 520 longitudes de onda).

El modelo PLS_PLS fue entrenado utilizando la matriz X . La variabilidad explicada y su capacidad de predicción fue evaluada a partir de medidas de bondad de ajuste como diferencia de devianza, Pseudo *Cox&Snell* R^2 , *Nagelkerke* R^2 y *MacFadden* R^2 y con el método de validación cruzada LOOCV (Leave-One-Out-Cross-Validation).

La representación gráfica de la estructura de la matriz de entrada fue realizada utilizando el HS-Biplot y se realizó un análisis visual de los grupos de hojas y su relación con las longitudes de onda del espectro visual e infrarrojo cercano. La calidad de representación del HS-Biplot fue evaluado mediante la variabilidad capturada por las componentes PLS-PLS.

En la validación externa se utilizó una matriz con datos obtenidos a partir de nuevas hojas seleccionadas aleatoriamente. El nuevo set de datos incluyó 32 imágenes nuevas, formado por 16 sanas y 16 infectadas. El modelo PLS-PLR construido fue evaluado mediante la predicción de los nuevos datos.

Otros modelos de aprendizaje supervisado fueron construidos para realizar una evaluación comparativa con los resultados obtenidos en PLS_PLR y HS-Biplot. Los

métodos de clasificación seleccionados son utilizados ampliamente en distintas ramas de la ciencia. Ellos son: NPLS-DA, SVM y MLP (redes neuronales artificiales).

La aplicación de cada uno de los modelos se realizó en dos fases:

- Fase de entrenamiento o de calibración en la que se diseña y se construye el modelo o regla para la clasificación.
- Fase de predicción o prueba en la que se clasifican los objetos de los que se desconoce su clase de pertenencia. También se la conoce como validación externa.

En cada fase se evaluaron las siguientes métricas de predicción:

TP verdaderos positivos. Número de hojas infectadas que fueron clasificadas correctamente.

FP Falsos Positivos. Número de hojas sanas que fueron clasificadas como infectadas.

FN Falsos Negativos. Número de hojas infectadas que fueron clasificadas como sanas.

TN Verdaderos Negativos. Número de hojas sanas que fueron clasificadas correctamente.

Exactitud. - es el total de aciertos del modelo.

$$Exactitud = \frac{TP+TN}{TP+FP+TN+FN} \quad (3.6.1)$$

Error de clasificación. - es la tasa de errores de predicción del modelo.

$$Error\ de\ clasificación = \frac{FP+FN}{TP+FP+TN+FN} \quad (3.6.2)$$

Sensibilidad. - también llamada recall o tasa de verdaderos positivos en indica que el porporción de postivos que fueron clasificados correctamente.

$$Sensibilidad = \frac{TP}{TP+FN} \quad (3.6.3)$$

Especificidad. - es la tasa de verdaderos negativos en el resultado de la clasificación.

$$Especificidad = \frac{TN}{TN+FP} \quad (3.6.4)$$

Precisión. – es el valor de predicción positiva y corresponde a la tasa de verdaderos positivos en el resultado de la clasificación.

$$Precisión = \frac{TP}{TP+FP} \quad (3.6.5)$$

Valor de predicción negativa. – es la tasa de verdaderos negativos en el total de negativos clasificados.

$$Valor\ de\ predicción\ negativa = \frac{TN}{TN+FN} \quad (3.6.6)$$

Prevalencia. - es la tasa de positivos en el total de la muestra.

$$Prevalencia = \frac{TP+FN}{TP+FP+TN+FN} \quad (3.6.7)$$

F_1 . - se interpreta como el promedio ponderado de la precisión y la sensibilidad.

$$F_1 = 2 \frac{\text{Precisión} * \text{sensibilidad}}{\text{Precisión} + \text{sensibilidad}} \quad (3.6.8)$$

3.6.1 Análisis exploratorio

Un análisis preliminar de los datos se realizó mediante la observación de las curvas espectrales de las hojas infectadas y no infectadas. Inicialmente se consideró los píxeles de regiones etiquetadas por especialistas y se generaron los espectros promediando los valores de reflectancia medidos en los píxeles etiquetados para cada nivel de severidad. Posteriormente, el análisis se realizó considerando toda la hoja, lo que permitió evidenciar la potencialidad de la aplicación de técnicas clasificación para detectar la enfermedad en las imágenes hiperespectrales de las hojas. Utilizando un gráfico de diferencia de los espectros de hojas sanas e infectadas se observó la conformación de rangos espectrales en los que la reflectancia varía producto de los cambios producidos por la enfermedad. Este análisis permitió determinar regiones del espectro electromagnético que podrían ser utilizadas para realizar una preselección de variables en otras investigaciones en las cuales resulte prioritario y analizar los cambios en las hojas a través de las variaciones en la reflectancia.

A continuación, se realizó una prueba estándar en el análisis exploratorio, la prueba de normalidad. El test de Kolmogorov Smirnov modificado por Lilliefors fue utilizado para contrastar la hipótesis de normalidad. El objetivo es evaluar como hipótesis que los datos proceden de una distribución normal y como hipótesis alternativa que no lo hacen. Para ello la prueba estima el valor p que nos indica la probabilidad de encontrar un valor del

estadístico de contraste más alejado que lo observado en la muestra actual. Los resultados fueron presentados en un gráfico.

Para detectar la multicolinealidad, que consiste en la existencia de relaciones lineales entre dos o más variables explicativas se realizó la observación de la matriz de correlación presentada en un gráfico de calor y se evaluó el factor de inflación de la varianza (VIF). La multicolinealidad es un problema común en datos hiperespectrales que puede afectar la construcción de algunos modelos de aprendizaje automático por cuanto se incumple la premisa de que la matriz de datos debe tener rango completo. Los modelos PLS son una solución para la multicolinealidad como efecto de la reducción de la dimensionalidad para obtener los factores latentes ortogonales y específicamente PLS-PLR, el método propuesto en esta investigación incluye la aplicación de regularización ridge en la regresión logística para eliminar los efectos de la multicolinealidad.

El factor de inflación de la varianza se define como:

$$VIF = \frac{1}{(1 - R_j^2)}, j = 1, 2, 3, \dots, k \quad (3.6.1.1)$$

El VIF se evalúa para cada variable realizando una regresión sobre el resto de las variables independientes. R_j^2 es el coeficiente de determinación para cada regresión calculada.

3.6.2 Modelo PLS-PLR

PLS fue originalmente desarrollada para respuestas continuas. En el caso de respuestas binarias, la regresión lineal no garantiza que la respuesta se ajuste a valores entre 0 y 1.

En estos casos, una versión PLS muy cercana a análisis discriminante sigue siendo muy utilizada y es llamada PLS-DA (PLS Discriminant Analysis) (Barker & Rayens, 2003). Nosotros consideramos que un modelo lineal no es adecuado cuando la respuesta es binaria, por lo tanto, una transformación logit es necesaria para ajustar la regresión. Nosotros usamos regresión logística en lugar de regresión lineal para relacionar la respuesta a las componentes PLS y de esta forma garantizar el ajuste a la respuesta binaria. Este método lo denominamos Regresión logística penalizada PLS (PLS-PLR, PLS Penalized Logistic Regresión).

PLS-PLR fue seleccionada e implementada para resolver algunos problemas de los datos y que afecta su capacidad de predicción como: multicolinealidad, separación en los datos, sobre-ajuste y sub-ajuste.

La alta colinealidad entre los predictores indica que las variables que la producen comparten cantidades sustanciales de información. En estas condiciones, los coeficientes de regresión tienen alta varianza lo que ocasiona que cambios pequeños en los datos produzcan cambios fuertes en los resultados haciendo que el modelo sea inestable y por lo tanto la evaluación de la importancia relativa de las variables resulte difícil. Una alternativa, para superar este problema es la reducción de dimensionalidad utilizando PLS que genera nuevas variables latentes no correlacionadas y ortogonales a partir de las variables predictoras con la ayuda de la variable respuesta. El número de variables latentes que también son llamadas componentes PLS es mucho menor a las variables originales (Vega Vilca & Guzmán, 2011). La información relevante es resumida en las primeras variables latentes, mientras que el “ruido” es modelado por las últimas, por lo tanto, en los casos en que exista colinealidad (redundancia) entre las variables, el ruido

aleatorio se reduce. Por consiguiente, es posible minimizar el riesgo de cometer un error estadístico de Tipo II (Alciaturi, Escobar, De La Cruz, & Rincón, 2003).

Por otro lado, la intención de PLS es encontrar nuevas variables que son combinaciones lineales de los predictores y las variables de respuesta de tal manera que las nuevas variables, además de explicar la varianza observada, predicen la respuesta lo mejor posible. Este método hace uso de la variable de salida para obtener las variables latentes, lo que reduce el sesgo evitando así el sub-ajuste (underfitting). El sesgo es la diferencia entre la predicción promedio de nuestro modelo y el valor a predecir. El modelo con alto sesgo no se ajusta a los datos de entrenamiento y simplifica demasiado el modelo generando a un alto error en la predicción de los datos de entrenamiento y prueba.

Finalmente, para prevenir el problema de separación de datos y multicolinealidad (Albert & Anderson 1984; Santner & Duffy 1986), el algoritmo PLR-PLR incluye regularización Ridge (Le Cessie & Van Houwelingen, 1992) en la regresión logística para limitar el crecimiento de los coeficientes de regresión, lo que reduce la varianza, evita el sobreajuste y controla los efectos de separación en los datos. La penalización Ridge es calculada por la suma de los cuadrados de los coeficientes ($L2$ norm) multiplicada por el parámetro de penalización λ . El parámetro λ puede tener un valor entre 0 y 1 (Godínez-Jaimes et al., 2012). Con el propósito de encontrar el modelo que mejor describe los datos, nosotros probamos valores de λ en el rango [0.1 - 0.9] con pasos incrementales de 0.1 y se calculó para cada valor de λ , las siguientes medidas de bondad de ajuste: *DiffDeviance*, *Cox&Snell R^2* , *Nagelkerke R^2* y *MacFadden R^2* (Allison, 2014).

El poder predictivo del modelo es relevante cuando queremos usar el modelo para estimar el comportamiento de nuevos datos. En esta investigación, utilizamos Leave-One-Out-Cross-Validation (LOOCV) para evaluar el modelo obtenido que mejor se ajustó a los datos. LOOCV es un método que hace un uso intensivo de los datos mediante "remuestreo" o "reutilización simple", este procedimiento utiliza una observación como conjunto de validación y las observaciones restantes como conjunto de entrenamiento. Esto se repite para todas las observaciones. El método se basa en calcular la tasa de error al predecir la pertenencia o ausencia de la enfermedad para cada una de las observaciones utilizando un modelo generado con las $(n - 1)$ observaciones restantes (Cool, Winer, & Rados, 1987).

3.6.2.1 Algoritmo PLS-PLR

Sea $\mathbf{Y}$ la variable respuesta binaria y $(\mathbf{X}_1, \dots, \mathbf{X}_p)$ un conjunto de predictores. Para una muestra de tamaño n los datos pueden ser organizados en un vector respuesta $\mathbf{y} = (y_1, \dots, y_n)^T$ y una matriz de predictores $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_p) = (x_{ij})$ ($i = 1, \dots, n; j = 1, \dots, p$), donde y_i es 0 o 1 para presencia o ausencia de la característica principal y x_{ij} es el valor del i^{th} individuo en el j^{th} predictor. Las columnas de $\mathbf{X}$ se suponen centradas y escaladas.

La ecuación de regresión de la variable dependiente $\mathbf{y} = \mathbf{T}\mathbf{c} + \mathbf{F}$ en el algoritmo PLS que explica $\mathbf{y}$ desde las componentes $\mathbf{T}$ es adaptada para tomar en cuenta la respuesta binaria. Bastien et al. (2005) generaliza el método cuando la respuesta desde una familia exponencial usando $g(\hat{\mathbf{y}}) = \mathbf{T}\mathbf{c}$ y propone un algoritmo para estimar los parámetros.

Nosotros incluimos una constante en el modelo porque la variable binaria no puede ser centrada.

$$g(\hat{y}) = c_0 + Tc \quad (3.6.1.1.1)$$

Incluyendo una constante mejora el ajuste significativamente y tendrá mejores propiedades para el biplot.

Existe otro problema al ajustar un modelo logístico conocido como el problema de separación: cuando en el espacio generado por las X hay un hiperplano que separa las presencias y las ausencias, las estimaciones de máxima probabilidad no existen (Albert & Anderson, 1984), entendiéndose existencia como finitud. Incluso cuando la separación no es perfecta (cuasi separación) las estimaciones son muy inestables. La solución habitual es usar una versión penalizada de máxima verosimilitud (Heinze & Schemper 2002). Aquí usaremos una penalización Ridge (Le Cessie & Van Houwelingen, 1992) en el modelo que se muestra en la ecuación (3.6.1.1.2) y en el algoritmo que describimos a continuación.

Nosotros buscamos componentes que son combinaciones lineales de los predictores y que explican de la mejor manera la respuesta (en forma de regresión logística). Sea el $\mathbf{t}_h$ el vector que contiene las puntuaciones (scores) de cada individuo en una de esas combinaciones de componentes, entonces $\mathbf{t}_h = \sum_{j=1}^p w_{hj} \mathbf{x}_j = \mathbf{X} \mathbf{w}_h$ siendo $\mathbf{w}_h = (w_{h1}, \dots, w_{hp})^T$ el vector de coeficientes. Normalmente usamos m de esas componentes que son mutuamente ortogonales.

El modelo Regresión Logística PLS es escrito como:

$$E(\mathbf{y}) = \hat{\mathbf{y}} = \frac{1}{1 + e^{-(c_0 \mathbf{1} + \sum_{h=1}^m c_h \mathbf{t}_h)}} \quad (3.6.1.1.2)$$

o

$$\text{logit}(\hat{\mathbf{y}}) = \log\left(\frac{\hat{\mathbf{y}}}{1 - \hat{\mathbf{y}}}\right) = c_0 \mathbf{1} + \sum_{h=1}^m c_h \mathbf{t}_h \quad (3.6.1.1.3)$$

O en forma matricial $\text{logit}(\hat{\mathbf{y}}) = c_0 \mathbf{1} + \mathbf{T}\mathbf{c}$, donde $\hat{\mathbf{y}} = (\hat{y}_1, \dots, \hat{y}_n)^T$ es el vector de probabilidades estimadas de la presencia en cada individuo y $\mathbf{c} = (c_1, \dots, c_m)^T$ son los coeficientes de la regresión sobre las componentes. El modelo es una regresión logística estándar sobre las componentes PLS. La constante c_0 debe mantenerse porque no se puede centrar la variable binaria.

En términos de las variables originales,

$$\text{logit}(\hat{\mathbf{y}}) = c_0 \mathbf{1} + \mathbf{T}\mathbf{c} = c_0 \mathbf{1} + \mathbf{XW}\mathbf{c} = c_0 \mathbf{1} + \mathbf{Xb} \quad (3.6.1.1.4)$$

Donde $\mathbf{W} = (\mathbf{w}_1, \dots, \mathbf{w}_m)$ and $\mathbf{b} = (b_1, \dots, b_p)^T$ son los coeficientes sobre las variables observadas.

Para estimar $\mathbf{T}$, $\mathbf{W}$, $\mathbf{c}$, c_0 and $\mathbf{b}$ nosotros usamos el algoritmo desarrollado por Bastien et al., (2005) con las modificaciones que se detallaron en los primeros párrafos de este apartado.

1. Cálculo de $\mathbf{t}_1$, la primera componente PLS.
 - a. Para cada predictor ($j = 1, \dots, p$), calcule el coeficiente de regresión w_{1j} de x_j , en la regresión logística de $\mathbf{y}$ sobre x_j , para obtener $\mathbf{w}_1 = (w_{11}, \dots, w_{1p})^T$
 - b. Normalice el vector $\mathbf{w}_1: = \mathbf{w}_1 / \|\mathbf{w}_1\|$.

- c. Calcule las puntuaciones de la componente (scores) $\mathbf{t}_1 = \mathbf{X}\mathbf{w}_1/\mathbf{w}_1^T\mathbf{w}_1$
2. Cálculo de $\mathbf{t}_h$, la h^{th} componente PLS. Las componentes $\mathbf{t}_1, \dots, \mathbf{t}_{h-1}$ han sido obtenidas.
 - a. Para cada predictor ($j = 1, \dots, p$), calcule el coeficiente de regresión w_{hj} de $\mathbf{x}_j$, en la regresión logística de $\mathbf{y}$ sobre $\mathbf{t}_1, \dots, \mathbf{t}_{h-1}$ y $\mathbf{x}_j$, para obtener $\mathbf{w}_h = (w_{h1}, \dots, w_{hp})^T$
 - b. Normalice el vector $\mathbf{w}_h: = \mathbf{w}_h/\|\mathbf{w}_h\|$.
 - c. Calcule la matriz residual $\mathbf{X}_{h-1}$ de la regresión lineal de $\mathbf{X}$ sobre $\mathbf{t}_1, \dots, \mathbf{t}_{h-1}$.
 - d. Calcule las puntuaciones de la componente (scores) $\mathbf{t}_h = \mathbf{X}\mathbf{w}_h/\mathbf{w}_h^T\mathbf{w}_h$.
3. $\mathbf{X}$ es factorizado como $\mathbf{X} = \mathbf{TP}$
4. Regresión Logística de $\mathbf{y}$ sobre las componentes PLS retenidas

$$\text{logit}(\hat{y}) = c_0 \mathbf{1} + \sum_{h=1}^m c_h \mathbf{t}_h$$

5. Expresión del modelo en términos de las predictoras originales $\mathbf{b} = \mathbf{W}$.

$$\text{logit}(\hat{y}) = c_0 \mathbf{1} + \mathbf{T}\mathbf{c} = c_0 \mathbf{1} + \mathbf{X}\mathbf{W}\mathbf{c} = c_0 \mathbf{1} + \mathbf{X}\mathbf{b}$$

3.6.2.2 Penalización Ridge

La regresión logística es un modelo muy utilizado cuando la respuesta es binaria, pero presenta problemas cuando existe separación en los datos y multicolinealidad. En estos casos el método de máxima verosimilitud para regresión logística no converge y los estimadores tienden a infinito.

La separación en los datos existe cuando una variable o una combinación lineal de variables predice de manera perfecta la variable respuesta, es decir, que las dos clases pueden ser separadas por un hiperplano.

Albert & Anderson, (1984) demostraron que cuando hay separación completa o cuasi-separación en los datos la solución de máxima verosimilitud no existe, es decir no tiene un máximo finito. Mientras que, si los datos se traslapan la solución de máxima verosimilitud existe y es única.

Una solución para este problema es la aplicación de una penalización en la regresión logística para reducir el tamaño de los coeficientes y la varianza. Estudios comparativos con otros estimadores (Firth, Rousseeuw & Christmann y Shen & Gao) dieron la ventaja a la penalización Ridge en términos de error cuadrático y sesgo (Godínez-Jaimes et al., 2012). Adicionalmente, cuando hay muchas variables que influyen en la respuesta, como es el caso de los datos hiperespectrales, la penalización Ridge (Le Cessie & Van Houwelingen, 1992) ofrece mejores resultados que la penalización Lasso que es más eficiente cuando existen pocas variables que influyen en el resultado. En general, la regularización suele ser útil en situaciones en la que existe gran cantidad de variables predictoras o la relación del número de observaciones con respecto al número de variables es muy baja.

En el modelo de regresión logística, las variables independientes y la variable respuesta se relacionan por,

$$\pi_i = P(Y_i = 1|x_i^T) = \frac{e^{x_i^T \beta}}{(1 + e^{x_i^T \beta})} \quad (3.6.1.2.1)$$

Donde,

$x_i^T = (1, x_{i1}, x_{i2}, \dots, x_{ip})$ es la i -ésima observación de X_p variables independientes.

$\beta = (\beta_0, \beta_1, \beta_2, \dots, \beta_p)^T$ es el vector de parámetros de la regresión.

La función log-verosimilitud es,

$$L(\beta) = \sum_{i=1}^n Y_i \log(\pi_i) + \sum_{i=1}^n (1 - Y_i) \log(1 - \pi_i) \quad (3.6.1.2.2)$$

Para maximizar la función $L(\beta)$ se utiliza Newton-Raphson. La solución es el estimador MV $\hat{\beta}$ para .

$$\beta^{(s)} = \beta^{(s)} + I^{-1}(\beta^{(s)}) U(\beta^{(s)}) \quad (3.6.1.2.3)$$

Donde,

$U(\beta^{(s)}) = X^T(y - \pi(\beta))$ es el vector de primeras derivadas parciales de $L(\beta)$

$I(\beta) = X^T \hat{V} X$ es la matriz de información estimada con

$$\hat{V} = \text{diag}\{\hat{\pi}_1(1 - \hat{\pi}_1), \dots, \hat{\pi}_n(1 - \hat{\pi}_n)\}.$$

La función log-verosimilitud con penalización Ridge es,

$$L_{\text{ridge}}(\beta) = L(\beta) - \lambda \|\beta\|^2 \quad (3.6.1.2.4)$$

El estimador Ridge iterativo logístico $\hat{\beta}_{\text{Ridge}}$ se lo obtiene usando el método de Newton-Raphson.

$$(3.6.1.2.5)$$

$$\hat{\beta}_{Ridge} = \hat{\beta}_{Ridge} + \{X^T \hat{V}(\hat{\beta}_{Ridge})X + 2\lambda I\}^{-1} \{U(\hat{\beta}_{Ridge}) - 2\lambda \hat{\beta}_{Ridge}\}$$

Donde,

$U(\hat{\beta}_{Ridge}) = X^T [y - \pi(\hat{\beta}_{Ridge})]$ es el vector de primeras derivadas parciales de $L(\beta)$

y

$$\hat{V} = diag\{\hat{\pi}_1(1 - \hat{\pi}_1), \dots, \hat{\pi}_n(1 - \hat{\pi}_n)\}.$$

3.6.2.3 Bondad de ajuste del modelo

Para evaluar la bondad de ajuste del modelo se calculó la diferencia de devianza y los Pseudo R^2 de McFadden, Cox&Snell, Nagelkerke. A pesar de que el modelo fue utilizado para predecir la presencia o ausencia de la enfermedad BLSD en las hojas de banano y su desempeño será evaluado con otras métricas de predicción, hemos incluido estas medidas como indicadores descriptivos.

Diferencia de Devianza

La devianza se utiliza para medir la falta de ajuste en modelos logísticos y corresponde a una medida para contrastar la hipótesis de ajuste correcto del modelo, análoga a la suma de cuadrados de los residuales en la regresión simple. Es una comparación entre un modelo a evaluar con el modelo saturado que es el que ajusta de forma perfecta. La devianza sigue asintóticamente una distribución χ^2_{n-p-1} . Un menor valor de la devianza indica un mejor ajuste del modelo. Un valor D alto o un valor_p muy bajo indican que existe un porcentaje alto de varianza no explicada mediante el modelo.

$$D = -2\ln \frac{L_M(\boldsymbol{\beta})}{L_F(\boldsymbol{\beta})} \quad (3.6.1.3.1)$$

Donde

$L_M(\boldsymbol{\beta})$ Verosimilitud del modelo

$L_F(\boldsymbol{\beta})$ Verosimilitud del modelo saturado (Full)

La diferencia de devianza es la diferencia entre la devianza del modelo nulo (solo con el término independiente) y la devianza del modelo ajustado.

$$D_0 - D_M = \left(-2\ln \frac{L_0(\boldsymbol{\beta})}{L_F(\boldsymbol{\beta})}\right) - \left(-2\ln \frac{L_M(\boldsymbol{\beta})}{L_F(\boldsymbol{\beta})}\right) \quad (3.6.1.3.2)$$

$$D_0 - D_M = -2 \left(\ln \frac{L_0(\boldsymbol{\beta})}{L_F(\boldsymbol{\beta})} - \ln \frac{L_M(\boldsymbol{\beta})}{L_F(\boldsymbol{\beta})} \right) \quad (3.6.1.3.3)$$

$$D_0 - D_M = -2\ln \left(\frac{L_0(\boldsymbol{\beta})}{L_M(\boldsymbol{\beta})} \right) \quad (3.6.1.3.4)$$

$$DiffDeviance = -(2LL_0(\boldsymbol{\beta}) - 2LL_M(\boldsymbol{\beta})) \quad (3.6.1.3.5)$$

Donde:

$LL_M(\boldsymbol{\beta})$ Log verosimilitud del modelo.

$LL_0(\boldsymbol{\beta})$ Log verosimilitud del modelo nulo (null model)

La diferencia de devianza (*DiffDeviance*) es interpretada como una medida de la variación de los datos explicada por el modelo con predictores, pero sin constantes (modelo nulo). Este estadístico tiene una distribución Chi² ($\chi^2_{s_M-s_0}$), con grados de libertad igual a la diferencia entre los números de parámetros de los modelos. De esta manera, la hipótesis nula será rechazada para el nivel de significancia α cuando

$DiffDeviance > \chi^2$, lo que es equivalente a que el valor_p de contraste sea menor que el nivel de α fijado (Hosmer et al., 1998).

Pseudo R^2

El Pseudo R^2 indica que tan bien el modelo explica/predice la variable dependiente basado en las variables independientes también se las llamar medidas de poder predictivo o de calidad de ajuste. Normalmente suelen estar entre 0 y 1, siendo el valor igual a 1 el indicador de una predicción perfecta. En cuanto al umbral que determine si el modelo es aceptable o no aún no hay un consenso.

$R^2McFadden$ es definida como uno menos la relación entre el logaritmo de la verosimilitud del modelo con respecto al logaritmo de la verosimilitud del modelo solo con los interceptos (Modelo nulo), con un rango teórico de valores de $0 \leq R^2McFadden \leq 1$. Generalmente se considera Una buena calidad de ajuste cuando $0.2 \leq R^2McFadden \leq 0.4$ y excelente para valores más altos.

$$R^2_{McFadden} = 1 - \left(\frac{LL_M}{LL_0} \right) \quad (3.6.1.3.2)$$

Donde:

LL_M es log verosimilitud del modelo

LL_0 es log verosimilitud de modelo nulo (null model)

$R^2_{Cox\&Snell}$ es una medida de bondad de ajuste que generaliza el R^2 de la regresión lineal. Se basa en la comparación del valor de verosimilitud del modelo (L_M) con el valor de verosimilitud del modelo nulo (L_0). Su rango de valores es entre 0 and $(1 - (L_0)^{2/n})$

$$R^2_{Cox\&Snell} = 1 - \left(\frac{L_0}{L_M}\right)^{\frac{2}{n}} \quad (3.6.1.3.3)$$

Donde

L_M es el valor de verosimilitud del modelo.

L_0 es el valor de verosimilitud del modelo nulo (null model).

$R^2_{Nagelkerke}$ es el valor de $R^2_{Cox\&Snell}$ estandarizado sobre el valor máximo que podría tomar. Por lo tanto, el máximo valor de este pseudo R^2 es 1 (Allison, 2014; Walker & Smith, 2016).

$$R^2_{Nagelkerke} = \frac{R^2_{Cox\&Snell}}{1 - (L_0)^{\frac{2}{n}}} \quad (3.6.1.3.4)$$

Donde,

L_M es el valor de verosimilitud del modelo.

L_0 es el valor de verosimilitud del modelo nulo (null model).

3.6.3 HS-Biplot

Las dos primeras componentes del modelo PLS-PLR fueron utilizadas para graficar el HS-Biplot. Las puntuaciones filas T (scores) y las puntuaciones columna P (loadings) fueron proyectadas en un plano que tiene como ejes las componentes del modelo PLS-PLR. Las puntuaciones filas se muestran como puntos y representan las hojas de banano. Las puntuaciones columnas proveen la dirección de las líneas que representan las longitudes de onda.

Hyperspectral Biplot (HS-Biplot) es una representación gráfica de las hojas, las longitudes de onda y las regiones de predicción y permite una inspección visual de las relaciones entre ellos. Las longitudes de onda fueron representadas por líneas coloreadas de acuerdo con la banda espectral a la que pertenecen.

Tabla 3-1 Longitudes de onda de las regiones visible y cercana al infrarrojo

Espectro visible	
Color	Long. De onda
Violet	380 - 427 nm
Blue	427 - 476 nm
Cyan	476 - 497 nm
Green	497 - 570 nm
Yellow	570 - 581 nm
Orange	581 - 618 nm
Red	618 - 780 nm
Infrarrojo cercano	
Color	Long. De onda
Gray	780 - 1350 nm

Técnicamente un biplot es una descomposición de una matriz X en un producto de 2 matrices de bajo rango (usualmente $S = 2$ o 3) y una matriz de error.

$$\mathbf{X} \cong \mathbf{TP} + \mathbf{E} \quad (3.6.2.1)$$

De esta forma, las filas y columnas pueden ser representadas al mismo tiempo en un diagrama de dispersión usando $\mathbf{T}$ y $\mathbf{P}$ como marcadores filas ($t_1, t_2 \dots t_s$) y marcadores columnas ($p_1, p_2, \dots p_s$) respectivamente, por lo tanto el producto interno $t_i^T p_i$ representa el x_{ij} , $(i,j)^{th}$ elemento de la matriz $\mathbf{X}$, esto es $\hat{\mathbf{X}} = \mathbf{TP}^T$. En este caso, la construcción del HS-Biplot se realiza con dos componentes, por lo tanto, las matrices $\mathbf{T}$ y $\mathbf{P}$ tienen 2 columnas ($S=2$) (Oyedele & Lubbe, 2015).

El biplot muestra las direcciones del espacio generado por las columnas de $\mathbf{X}$ que mejor separa las presencias de las ausencias para la variable dependiente. Así, nosotros tenemos una aproximación de bajo rango $\hat{\mathbf{X}} = \mathbf{TP}$ de la matriz $\mathbf{X}$ que captura la parte que mejor explica la respuesta.

El porcentaje de variabilidad capturada por la aproximación es,

$$\rho^2 = \frac{tr(\hat{\mathbf{X}}^T \hat{\mathbf{X}})}{tr(\mathbf{X}^T \mathbf{X})} \times 100 \quad (3.6.2.2)$$

Donde $\hat{\mathbf{X}}$ es la matriz de predictores aproximada y $\mathbf{X}$ es la matriz original.

Es posible identificar las variables relacionadas a las componentes PLS calculando bondad de ajuste por columna, es decir, la cantidad de varianza de cada columna capturada por la aproximación, esto es,

$$\rho_j^2 = \frac{tr(\hat{\mathbf{x}}_{[j]}^T \hat{\mathbf{x}}_{[j]})}{tr(\mathbf{x}_{[j]}^T \mathbf{x}_{[j]})} \times 100 \quad (j = 1, \dots, p) \quad (3.6.2.3)$$

Donde $\hat{\mathbf{x}}_{[j]}$ y $\mathbf{x}_{[j]}$ son el j^{th} columnas de la matriz ajustada y de la matriz original respectivamente. Solamente las columnas con alto porcentajes son relacionadas a la

respuesta. Estas cantidades son llamadas contribuciones de las componentes a las variables predictoras. ρ_j^2 también es llamado calidad de representación de la variable j o predictividad de la columna.

La bondad de ajuste de cada fila es,

$$\rho_i^2 = \frac{tr(\hat{\mathbf{x}}_{[i]} \hat{\mathbf{x}}_{[i]}^T)}{tr(\mathbf{x}_{[i]} \mathbf{x}_{[i]}^T)} \times 100 \quad (i = 1, \dots, n) \quad (3.6.2.4)$$

Donde $\hat{\mathbf{x}}_{[i]}$ y $\mathbf{x}_{[i]}$ son el i^{th} filas de la matriz ajustada y de la matriz original respectivamente.

Estas medidas también se llaman calidad de la representación o predictividad. Las medidas separadas para cada dimensión también se llaman Contribuciones del Factor al Elemento (fila o columna). Las medidas se utilizan para identificar qué dimensiones son útiles para diferenciar al individuo del resto. Las personas con bajas cualidades generalmente se colocan alrededor del origen.

Los marcadores de las filas de $\mathbf{X}$ se usan en el paso 4 del algoritmo de PLS-PLR para predecir la respuesta binaria, entonces, la variable binaria también se puede proyectar en el biplot usando un Biplot logístico externo, propuesto por Demey et al. (2008) y Vicente-Villardón et al. (2006). La principal diferencia es que en la propuesta original los puntajes para los individuos se obtienen de las Coordenadas principales y aquí de la Regresión logística PLS. El vector $\mathbf{c}$ de coeficientes de regresión logística define la dirección en el espacio generado por las columnas de $\mathbf{T}$ que separa mejor las presencias y ausencias y las probabilidades esperadas de tener la presencia de la característica.

$$\text{logit}(\hat{y}_i) = \text{logit}(\hat{p}_i) = c_0 + \mathbf{t}_{[i]} \mathbf{c} \quad (3.6.2.4)$$

Donde $t_{[i]}$ es el i^{th} fila de $\mathbf{T}$. La probabilidad esperada es obtenida proyectando los puntos $t_{[i]}$ sobre el vector $\mathbf{c}$. El punto, en esa dirección, que predice una probabilidad esperada de 0.5 en el biplot tiene las coordenadas,

$$x = \frac{-c_0 c_1}{c_1^2 + c_2^2}; \quad y = \frac{-c_0 c_2}{c_1^2 + c_2^2} \quad (3.6.3.5)$$

Si nosotros predecimos presencia cuando la probabilidad esperada es mayor que 0.5, la dirección de $\mathbf{c}$ para ese valor de probabilidad, divide la representación en dos regiones, una que predice la presencia y otra que predice la ausencia. El límite de las dos regiones es una recta perpendicular a $\mathbf{c}$ y pasa por el punto (x, y) . Para más detalles ver en la sección 2.5.2 , Demey et al. (2008) o Vicente-Villardón et al. (2006).

3.6.4 Análisis comparativo

Tres modelos adicionales (NPLS-DA, SVM y MLP) fueron construidos con el objetivo de analizar el poder predictivo y explicativo de PLS-PLR y HS-Biplot. Los modelos fueron entrenados utilizando los mismos datos que fueron utilizados para ajustar el modelo PLS-PLR. Los resultados fueron evaluados utilizando matrices de confusión y métricas de predicción. La especificidad es la proporción de las hojas sanas clasificadas correctamente de todas las hojas sanas clasificadas. La sensibilidad es la proporción de hojas de banano infectadas correctamente clasificadas en relación con todas las hojas inoculadas clasificadas. La precisión nos permite conocer la tasa de verdaderos positivos que corresponde al porcentaje de hojas infectadas que fueron clasificadas correctamente. La exactitud es la proporción de hojas correctamente clasificadas en relación con el total

de hojas. Además, se calculó la métrica F1 que se puede definir como un promedio ponderado entre la precisión y la sensibilidad.

Otra forma de evaluar la idoneidad de los modelos son las curvas de la característica del receptor (ROC) y el área bajo la curva ROC (AUC). La curva ROC (Receiver Operating Characteristic) ilustra la "tasa de verdaderos positivos" (eje Y) versus la "tasa de falsos positivos" (eje X) para diferentes umbrales de decisión. Una línea diagonal en el gráfico describe lo que sería la curva ROC de un test diagnóstico incapaz de discriminar entre positivos y negativos, debido a que cada punto de corte que la compone determina la misma proporción de verdaderos positivos y de falsos positivos (Cerda & Cifuentes, 2012).

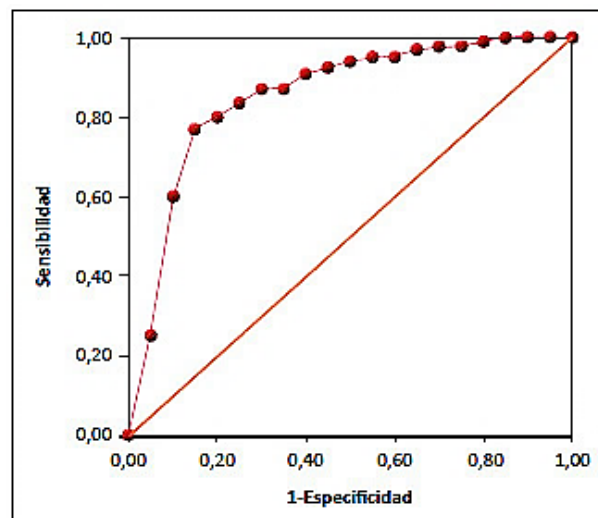


Figura 3.6 Curva ROC para un modelo hipotético (Cerde & Cifuentes, 2012).

El área bajo la curva AUC, es un indicador del rendimiento predictivo de un clasificador que permite comparar dos o más clasificadores en función de su capacidad discriminante. Este índice se puede interpretar como la probabilidad de que el modelo

pueda discriminar entre la clase positiva y la clase negativa. Un valor AUC alto significa una exactitud mayor (Jinzhu, Di, & Jiang, 2013).

3.6.5 Modelo NPLS-DA

El modelo NPLS-DA, basado en el algoritmo NPLS explicado en el apartado 2.4.2, fue generado utilizando un tensor de tercer orden construido a partir de los cubos hiperespectrales y fue transformado mediante un método de matrización. Para ello, siguiendo el procedimiento propuesto por Folch-Fortuny et al. (2016), se calcularon 5 características (media, desviación típica, simetría, curtosis, quinto momento) en cada longitud de onda obteniéndose una matriz reducida de 5 X 520 por cada cubo HS que inicialmente tenían 205 filas, 198 columnas y 520 longitudes de onda.

Número de píxeles en una imagen

$$n_p = \text{filas} \times \text{columnas} = 205 \times 198 = 40\,590$$

Media

$$\hat{x}_k = \frac{\sum x_{ik}}{n_p} \quad (3.6.5.1)$$

Donde

x_{ik} son los valores de reflectancia normalizada en el pixel i para una longitud de onda k .

n_p es número de píxeles en la imagen.

Desviación estándar

$$s_k = \sqrt{\frac{\sum (x_{ik} - \hat{x}_k)^2}{n_p - 1}} \quad (3.6.5.2)$$

Donde

x_{ik} son los valores de reflectancia normalizada en el pixel i para una longitud de onda k .

$\hat{x}_k$ es la media de la reflectancia en la longitud de onda k .

n_p es número de píxeles en la imagen.

Simetría (tercer momento)

$$\mu_{3k} = \frac{\sum (x_{ik} - \hat{x}_k)^3}{n_p} \quad (3.6.5.3)$$

Donde

x_{ik} son los valores de reflectancia normalizada en el pixel i para una longitud de onda k .

$\hat{x}_k$ es la media de la reflectancia en la longitud de onda k .

n_p es número de píxeles en la imagen.

Curtosis (cuarto momento)

$$\mu_{4k} = \frac{\sum (x_{ik} - \hat{x}_k)^4}{n_p} \quad (3.6.5.4)$$

x_{ik} son los valores de reflectancia normalizada en el pixel i para una longitud de onda k .

$\hat{x}_k$ es la media de la reflectancia en la longitud de onda k .

n_p es número de píxeles en la imagen.

Quinto momento

$$\mu_{5k} = \frac{\sum (x_{ik} - \hat{x}_k)^5}{n_p} \quad (3.6.5.5)$$

x_{ik} son los valores de reflectancia normalizada en el pixel i para una longitud de onda k .

$\hat{x}_k$ es la media de la reflectancia en la longitud de onda k .

n_p es número de píxeles en la imagen.

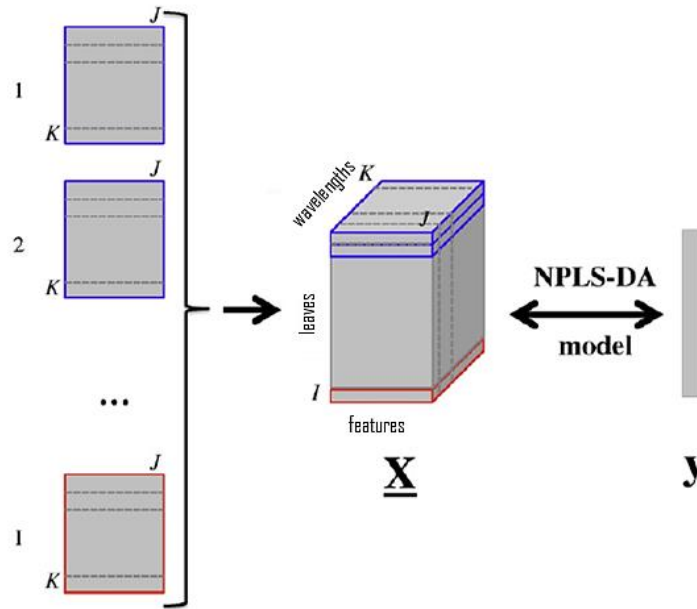


Figura 3.7 Estructura de tensor de tercer orden en NPLS-DA (Folch-Fortuny et al., 2016).

Una vez realizada la transformación de los cubos HS, un tensor de tercer orden con dimensiones $104 \times 5 \times 520$ se formó con las 104 matrices generadas (figura 3.7). Este nuevo cubo de datos fue desplegado en el primer modo y se obtuvo la matriz final de dimensiones 104 filas x 2600 columnas ($I \times JK$) (figura 3.8).

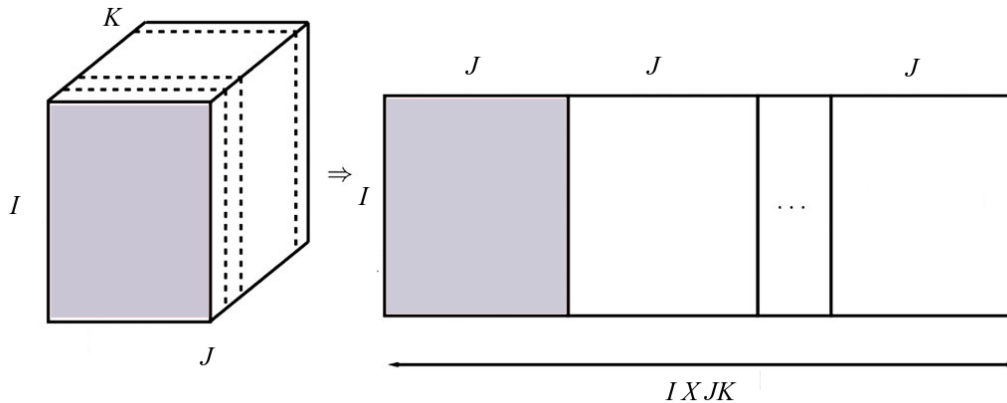


Figura 3.8 Tensor desplegado primer modo.

Durante la calibración del modelo se buscó el número mínimo de componentes para lograr la mejor predicción y luego, se realizó la prueba de validación externa utilizando la misma muestra utilizada en validación externa del modelo PLS-PLR.

El método NPLS-DA fue desarrollada en el lenguaje R.

3.6.6 Modelo de clasificación SVM

Las máquinas de vectores soporte (SVM) son clasificadores lineales que buscan separadores lineales llamados hiperplanos ya sea en el espacio de las entradas o en un espacio transformado (espacio de características) utilizando funciones kernel.

Dos modelos SVM de clasificación fueron entrenados aplicando a priori diferente criterio de separación de clases para seleccionar el kernel. En primer lugar, se consideró que las clases son linealmente separables para lo cual se seleccionó el kernel lineal.

Luego, se consideró la posibilidad de que los datos no sean linealmente separables y se creó un segundo modelo aplicando un kernel polinómico de segundo grado.

La optimización del hiperparámetro de regularización C permite el control de las violaciones de las observaciones sobre el margen del hiperplano, favoreciendo el equilibrio entre el sesgo (bias) y la varianza. El hiperparámetro de regularización C flexibiliza la clasificación proveyendo de una holgura para observaciones que se ubican dentro del margen máximo y establece un compromiso entre el error de entrenamiento y la complejidad del modelo. El hiperparámetro C limita el efecto de cualquier instancia de entrenamiento en la superficie de decisión.

Con un valor bajo de C se incrementa el margen generando un mayor bias, pero la varianza se reduce. El modelo es más simple a costa de un mayor error de entrenamiento. Con un valor alto de C , se aceptará un margen menor con un modelo más complejo y se ajustará bastante a los datos logrando un bias reducido, pero con una varianza alta, lo que aumenta el riesgo de sobreajuste.

La optimización del parámetro de regularización fue realizada mediante validación cruzada.

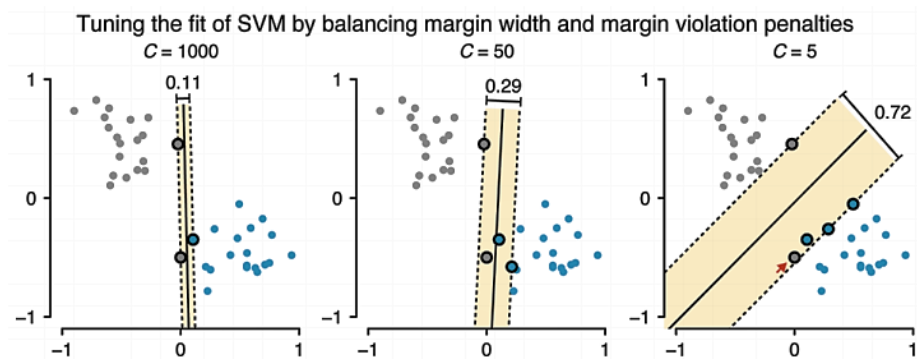


Figura 3.9 Aplicación de hiperparámetro C en SVM de margen blando (Bzdok, Krzywinski, Altman, 2018).

La aplicación del método SVM fue realizada utilizando el lenguaje PYTHON.

3.6.6.1 SVM Lineal

Tal como se indicó anteriormente, los hiperplanos de separación son buenos clasificadores cuando las clases son perfectamente separables o cuasi-perfectamente separables.

El modelo SVM lineal fue implementado mediante la función kernel lineal asumiendo que los datos son separables por un hiperplano. El hiperparámetro C, fue modificado entre 1 y 50.

Para cada modelo lineal se calculó la matriz de confusión y las métricas de predicción en entrenamiento y validación externa y se seleccionó el que presentó los mejores resultados.

El informe de clasificación incluye las siguientes métricas: precisión, sensibilidad, F1, exactitud, AUC.

3.6.6.2 SVM Polinomial

Cuando las clases no son separables linealmente el clasificador SVM (lineal) no es una solución práctica, en estos casos, se utiliza una transformación mediante una función kernel para aumentar las dimensiones del espacio de las predictoras.

Una posibilidad para tratar con límites no lineales entre clases consiste en utilizar el kernel polinomial. Los kernels son funciones que transforman un espacio de pocas dimensiones en un espacio de dimensiones mayores mediante transformaciones complejas de los datos. También puede definirse como una función que cuantifica la similitud entre dos observaciones en un nuevo espacio dimensional.

En esta investigación se avaluó varios modelos SVM con kernel polinomial modificando los valores del hiperparámetro C en un rango entre 1 y 100.

Utilizando los modelos generados construyó la matriz de confusión y el reporte de clasificación para predicción utilizando los datos de entrenamiento y validación externa. De la misma forma que el modelo SVM lineal, se seleccionó el modelo polinomial que presentó los mejores resultados de predicción.

El informe de clasificación incluye las siguientes métricas: precisión, sensibilidad, F1, exactitud, AUC.

3.6.7 Redes Neuronales Artificiales MLP

Se diseñó dos redes neuronales perceptrón multicapa (MLP) con retropropagación (back-propagation). En las MLP, cada neurona de la capa de entrada incorpora a la red el valor de una variable independiente que es recibido por las neuronas de las capas ocultas en donde se realiza la mayor parte del trabajo de modelo y las neuronas de la capa de salida calculan el valor de la variable respuesta sea esta cuantitativa o categórica.

Una MLP emplea la retropropagación como mecanismo de aprendizaje durante la etapa de entrenamiento en el cual se evalúa la función de pérdida para determinar la eficiencia del aprendizaje. El siguiente proceso, luego del entrenamiento, es la validación con información no etiquetada, esta etapa también es conocida como validación externa que utiliza un dataset de prueba (Garro, Sossa, & Vazquez, 2012).

3.6.7.1 Diseño de redes neuronales

El diseño de una MLP es determinante en el desempeño de una red neuronal. Los elementos que deben ser considerados para el diseño de una red neuronal son los siguientes:

Arquitectura. – es la topología de la red neuronal y está definida por el número de capas y el número de neuronas de cada capa. La capa de entrada contiene un número de neuronas igual al número de variables predictoras de la matriz de entrada. El número de neuronas de la capa de salida depende del tipo de problema a resolver (clasificación binaria, clasificación multicategorías, regresión). El número de capas ocultas está vinculado a la complejidad del modelo esperado.

Función de activación. - también es llamada función de transferencia. En el apartado 2.7.3 fueron descritas las principales funciones de activación entre las que se describieron las siguientes funciones: Lineal, Signo, Sigmoide, tangente hiperbólica, ReLu, Hard Tanh, Leaky Relu.

Función de pérdida. - La solución al problema de optimización para determinar los pesos sinápticos se obtiene mediante la determinación del error entre el valor estimado y el valor real utilizando una función de pérdida o de coste. Las funciones más son: Error cuadrático medio (MSE, Mean Squared Error), entropía cruzada binaria (binary cross-entropy) y entropía cruzada categórica (categorical cross-entropy).

Función de salida. – Corresponde a la función de la capa de salida. Las funciones de salida más utilizadas en problemas de clasificación son: la función sigmoide y la función softmax.

Optimizadores de gradiente. – son técnicas de aproximación para estimar el gradiente con la intención de alcanzar el mínimo de la función de pérdida en el menor tiempo posible. Entre los optimizadores más utilizados se encuentran: momentum, RMSProp (Root Mean Square Propagation), adam (Adaptive Moment Estimation), Adagrad y Adadelata. El más utilizado es adam por su rapidez en alcanzar el mínimo. Adam es un algoritmo de optimización de funciones estocásticas basada en gradiente de primer orden que utiliza estimaciones adaptativas de momentos de orden inferior propuesto por Kingma & Ba (2015). Este Optimizador es una extensión del descenso de gradiente estocástico que adapta las tasas de aprendizaje utilizando el primer momento (media) y el segundo momento (varianza).

La asignación de los parámetros del diseño se basó en el conocimiento de los datos hiperspectrales y en experiencias de otros autores en problemas similares utilizando un enfoque de arquitectura constructivo (Torres, 2018).

3.6.7.2 MLP con una capa oculta

La primera red fue diseñada con una capa oculta, la cual llamaremos “MLP 1”. El número de características o variables (longitudes de onda) corresponde al número de neuronas de la capa de entradas. La matriz de datos de entrada tiene 520 variables, por lo tanto, nuestra red neuronal fue diseñada con 520 neuronas en la capa de entrada.

Puesto que, el tipo de salida que requerimos es binaria (0,1), la capa de salida está formada por una neurona con una función de activación sigmoide.

En la capa oculta la función de activación es ReLu.

Para el cálculo del número de neuronas capas utilizamos la regla de la pirámide:

$$h = \sqrt{mn} \quad (3.6.7.2.1)$$

donde

h es el número de neuronas de la capa oculta

n es el número de neuronas de entrada

m es el número de neuronas de salida

Por lo tanto, utilizando la ecuación (3.6.7.2.1) el número de neuronas en la capa oculta es

$$h = \sqrt{1 * 520} = 22$$

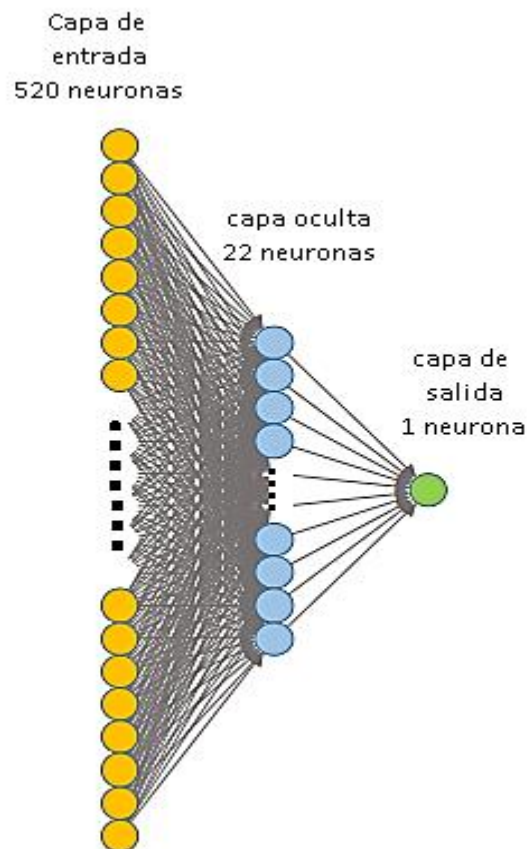


Figura 3.10 Diagrama de MLP con una capa oculta.

Se seleccionó el optimizador adam y la función de pérdida binary cross entropy.

El número de épocas de entrenamiento fue modificado según la tabla 3-2.

Tabla 3-2 Número de épocas para entrenamiento de redes neuronales

Número de épocas
300
500
700
1000
1200
1500
2000

Utilizando los datasets de entrenamiento y de validación se calculó la tabla de confusión y las métricas de exactitud, precisión, sensibilidad y F1. En la fase de validación se agregó el área bajo lo curva ROC, AUC.

3.6.7.3 MLP con dos capas ocultas

La segunda red fue diseñada con 2 capas ocultas (figura 3.11), será denominada “MLP 2”. La capa de entrada incluyó 520 neuronas que corresponden a las longitudes de onda en la matriz de datos.

De la misma manera que la primera MLP, el tipo de salida que requerimos es binaria (0,1) por lo tanto, la capa de salida está formada por una neurona con una función de activación sigmoide.

La función de activación en las dos capas ocultas es ReLu.

El número de neuronas de las capas ocultas se calculan mediante las ecuaciones (3.6.7.3.2) y (3.6.7.3.3).

$$r = \sqrt[3]{\frac{n}{m}} = \sqrt[3]{\frac{520}{1}} = 8 \quad (3.6.7.3.1)$$

Donde

n es el número de neuronas de la capa de entrada.

m es el número de neuronas de la capa de salida.

$$H1 = m * r^2 = 1 * 64 = 64 \quad (3.6.7.3.2)$$

Donde

H1 es el número de neuronas de la primera capa.

m es el número de neuronas de la capa de salida.

$$H2 = m * r = 1 * 8 = 8 \quad (3.6.7.3.3)$$

Donde

H2 es el número de neuronas de la segunda capa.

m es el número de neuronas de la capa de salida.

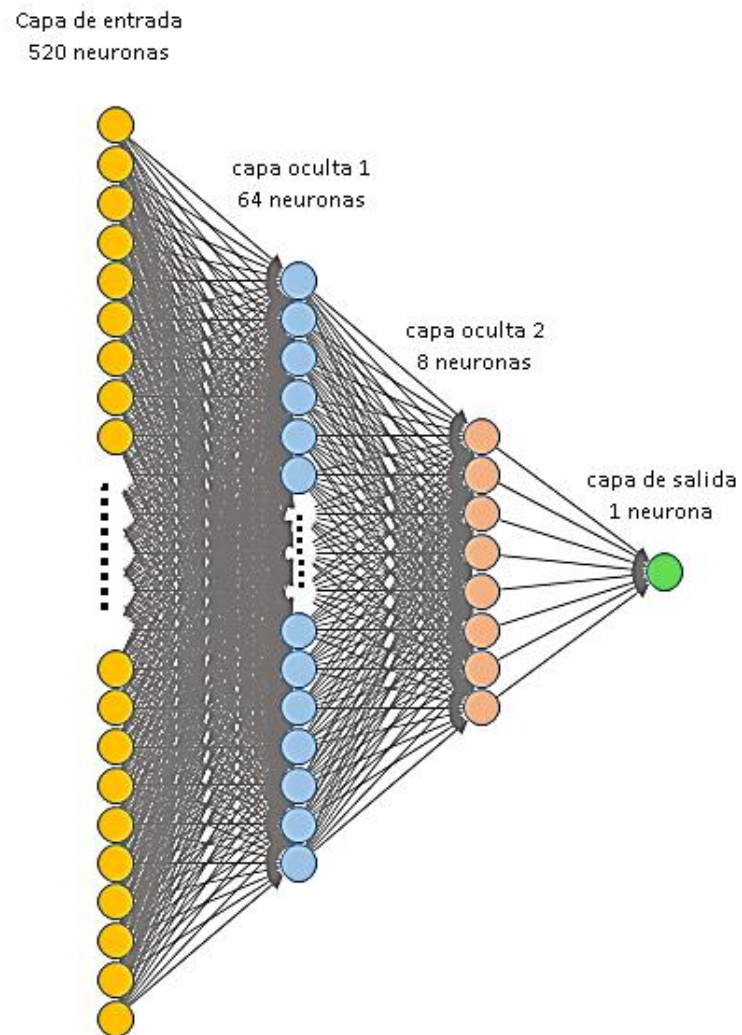


Figura 3.11 Diagrama de MLP con dos capas ocultas.

Se utilizó el optimizador de gradiente adam y la función de pérdida binay cross entropy.

Las épocas de entrenamiento se modificaron según la tabla 3-2.

Tanto en la fase de entrenamiento como en la de validación se calculó la tabla de confusión y las métricas de exactitud, precisión, sensibilidad, F1. En la fase de validación se incluyó el área bajo la curva ROC, AUC.

4 RESULTADOS

En enfoque principal del presente trabajo es el desarrollo de un modelo predictivo utilizando el método PLS-PLS con el propósito de detectar la Sigatoka negra en imágenes hiperespectrales de hojas de banano y realizar la visualización y análisis de la estructura de los datos utilizando HS-Biplot. La hipótesis que vamos a evaluar es que la Sigatoka negra influye en las propiedades ópticas de la planta de banano desde sus etapas iniciales y puede ser detectada en imágenes hiperespetrales de las hojas mediante la aplicación de las técnicas PLS-PLR y HS_Biplot.

El experimento fue realizado en condiciones controladas y se escanearon hojas en los primeros niveles de la infección. Considerando la detección temprana de la enfermedad como un factor clave para la aplicación de estrategias de control y prevención, se seleccionaron, dentro del grupo de estudio, hojas no infectadas y hojas infectadas en estado pre-sintomático, severidad 1 y 2.

El análisis preliminar de los espectros en diferentes niveles de severidad fue realizado utilizando las regiones infectadas etiquetadas por los especialistas. Las gráficas de los promedios de reflectancia de las regiones en diferentes niveles de severidad mostraron cambios en distintas regiones del espectro electromagnético producto de los cambios físico-químicos en la hoja causados por el hongo *p. fijiensis* a medida que se esparce la enfermedad. Las imágenes hiperespectrales fueron sometidas a un proceso de normalización para eliminar las distorsiones producidas por el dispositivo. Se aplicó la media de la reflectancia en cada longitud de onda lo que redujo la dimensión de los cubos HS. Una evaluación preliminar de las firmas espectrales de las hojas mostró diferencias entre los espectros de las hojas en diferentes estadios de la enfermedad respecto a las hojas no infectadas.

La inspección de la matriz de correlación y del factor de inflación de varianza (VIF) corroboraron la condición de alta multicolinealidad en datos hiperespectrales. PLS-PLR intenta solucionar el problema de multicolinealidad, pero además se incluye la penalización Ridge que garantiza el control de los efectos producidos por la separación de datos al aplicar la regresión logística sobre las variables latentes PLS. Los excelentes resultados de predicción fueron complementados con el HS-Biplot que utiliza las dos primeras componentes PLS para representar las hojas, las longitudes de onda y sus relaciones en un solo gráfico.

Tres métodos adicionales muy utilizados en la detección de enfermedades en plantas con HSI, tales como, NPLS-DA, SVM y redes neuronales artificiales (MLP) fueron aplicados y sus resultados fueron comparados con los resultados de PLS-PLR y HS-Biplot.

4.1 ANALISIS EXPLORATORIO

4.1.1 Análisis de firmas espectrales

Las hojas de banano seleccionadas y las las regiones de la hoja afectadas por el patógeno *p. fijiensis* fueron etiquetas por los especialistas en biotecnología. Luego se calculó la media de los valores de reflectancia medida en cada pixel generando una firma espectral por cada nivel de severidad. Los gráficos de los espectros de las regiones etiquetadas mostraron diferencias en los espectros en diferentes niveles de severidad de la enfermedad respecto al espectro de las hojas sanas.

En la figura 4.1 se muestra las curvas espectrales de las regiones sanas e infectadas. La curva color verde corresponde a las regiones sanas, los colores amarillo, café, marrón y gris son los espectros de las hojas con niveles de severidad 1 al 4. Los niveles 5 y 6 no se muestran debido a que los espectros presentan altos niveles de ruido producto de los daños físicos de la hoja.

A medida que la severidad de la enfermedad se incrementa, se observan cambios graduales en diferentes rangos espectrales. En el espectro visual, específicamente en el rango entre las longitudes de onda de 550 nm a 680 nm que corresponde al color amarillo (570 - 581 nm) y la color naranja (581 - 618 nm), se observa un incremento gradual del nivel de reflectancia debido al amarillamiento de la hoja por la presencia de zonas cloróticas y posterior necrosis debido a los cambios en los tejidos fotosintéticos que afectan la producción de la clorofila, lo que también produce un incremento en la absorción de la radiación en el rango del color verde (497 - 570 nm).

A partir de la longitud de onda de 700 nm aprox., en el borde rojo, se produce una disminución de la reflectancia que se extiende hasta el infrarrojo cercano. Esta disminución es causada por los cambios estructurales a nivel de mesófilo (Bendini et al., 2015).

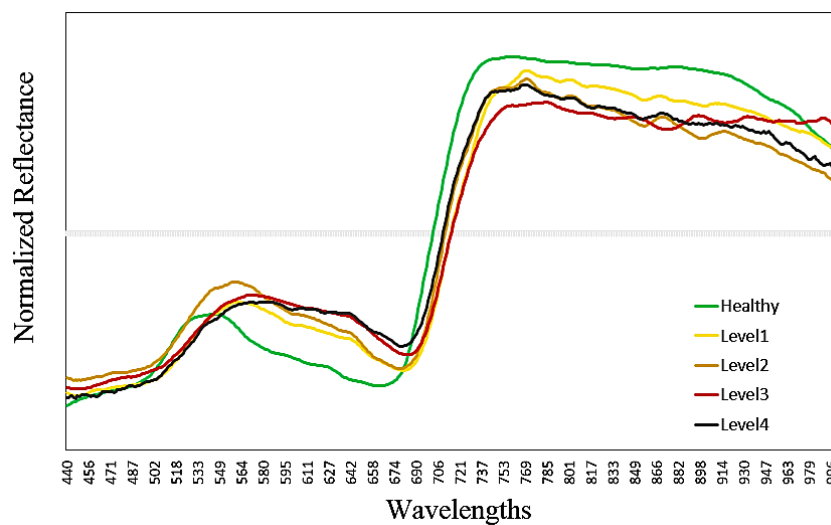


Figura 4.1 Firmas espectrales de regiones sanas y enfermas.

Las diferencias entre los espectros de regiones enfermas y sanas son un indicador potencial para la detección de la enfermedad en planas de banano utilizando análisis de los datos hiperespectrales.

En esta investigación, la detección temprana de BLSD se realiza en los estados presintomático, estadio 1 y estadio 2 por dos razones principales: la primera es la dificultad de detectar por observación directa los daños físicos mínimos producidos por la enfermedad en las etapas iniciales de la enfermedad y la segunda es que a partir del

nivel 2 de severidad se desarrollan las esporas conidios y ascosporas, las cuales son las causantes de la propagación de la enfermedad. Los conidios son esporas asexuales que se diseminan a distancias cortas por medio del agua, mientras las ascosporas son esporas sexuales que pueden ser transportadas a largas distancias por el viento (Marín et al., 2003). La detección de la enfermedad en estas etapas de la enfermedad es crucial para la aplicación de estrategias de control oportunas que permitan la racionalización de los fungicidas y la recuperación de la salud del cultivo en menor tiempo.

A partir de la matriz reducida (104 x 520) obtenida en la sección 3.5 se analizó los espectros de las hojas en diferentes estadios de la enfermedad. Cada fila de la matriz contiene la información de una hoja y genera un espectro específico de acuerdo a su nivel de infección. Los promedios de la reflectancia por el nivel de la enfermedad fueron calculados agrupando las hojas según el nivel de infección obteniendo un espectro característico por grupo.

El análisis de los espectros mostró que diferencia entre la media de la reflectancia de las hojas sanas y la media de la reflectancia de las hojas infectadas (figura 4.2) tiene seis puntos de inflexión en el cero, los cuales definen los rangos de longitudes de onda en los cuales la reflectancia de las hojas infectadas aumenta o disminuye respecto al patrón establecido por las hojas de control. En el gráfico, los valores negativos corresponden al incremento de la reflectancia en las hojas infectadas mientras que los valores positivos son el resultado de una disminución. Los puntos de inflexión en cero corresponden a las longitudes de onda en los cuales las reflectancias de las hojas sanas e infectadas son iguales por lo tanto la diferencia es cero. Los intervalos en los que se producen estos cambios de reflectancia están detallados en la tabla 4-1.

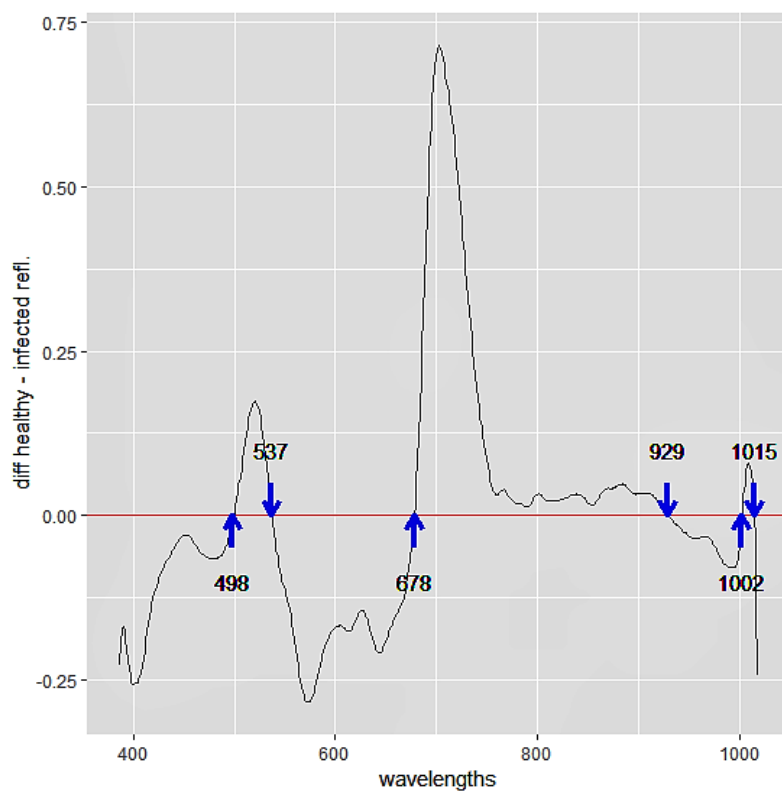


Figura 4.2 Variación de la reflectancia con BLSD.

Tabla 4-1 Intervalos del espectro en los cuales se producen cambios de reflectancia por efecto de la BLSD

RANGO ESPECTRAL		VARIACIÓN	COLOR
395	498	aumenta	violeta azul
499	537	disminuye	verde
538	678	aumenta	verde amarillo naranja rojo
679	929	disminuye	rojo infrarrojo
930	1002	aumenta	infrarrojo
1003	1015	disminuye	infrarrojo
1015	1018	aumenta	infrarrojo

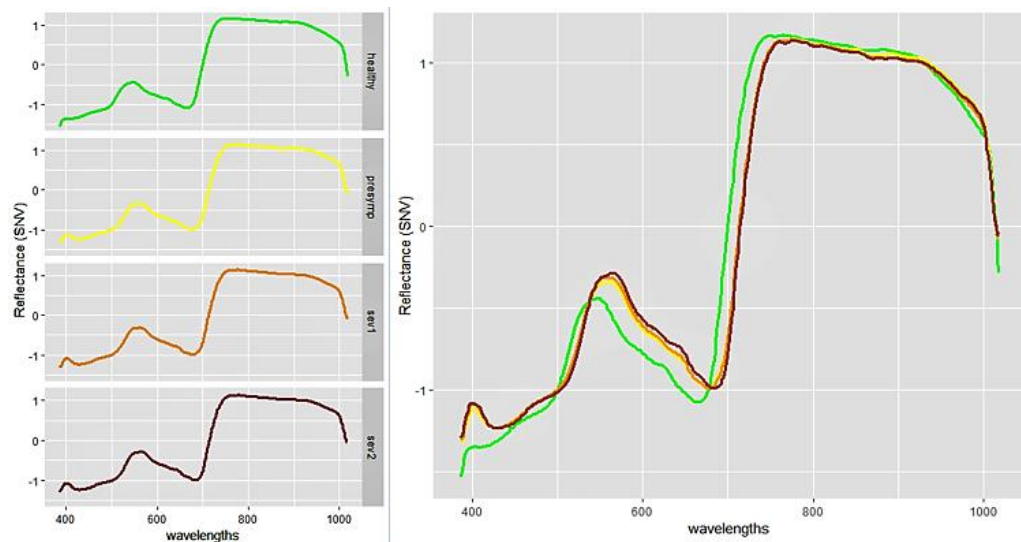


Figura 4.3 Firmas espectrales de hojas sanas e infectadas.

En la figura 4.3, se muestran los espectros de las hojas con diferentes niveles de incidencia de la enfermedad: el espectro de la hoja sana es presentado en color verde, el espectro de la hoja en estado pre-sintomático en color amarillo, el espectro de la hoja en nivel de severidad 1 en color naranja y el espectro de la hoja en nivel de severidad 2 en color marrón. Las firmas espectrales fueron obtenidas promediando los valores de reflectancia de cada grupo de hojas. El gráfico muestra que los espectros de las hojas enfermas presentan diferencias significativas respecto a los espectros de las hojas sanas.

En las primeras longitudes de onda localizadas en el rango del color verd (497 – 570 nm) se observa una reducción de la reflectancia producto de la disminución de la producción de la clorofila, mientras que en las longitudes de onda del color amarillo (570 - 581 nm) y del color naranja (581 - 618 nm) se produce un incremento, el cual se debe a la presencia de zonas cloróticas en la superficie de la hoja como consecuencia de alteraciones pigmentarias. En el borde rojo (618 - 780 nm) la reflectancia se reduce y sucede lo mismo, pero en menor escala, en en el infrarrojo cercano (780 – 1350 nm) debido a los cambios en estructuras intercelulares de la hoja. Es importante notar que estos cambios se producen inclusive en las hojas infectadas presintomáticas (espectro amarillo) lo que indica que los cambios fisiológicos en la hoja que pueden ser detectados por el sensor hiperespectral, aunque aún son imperceptibles por el ojo humano.

4.1.2 Prueba de normalidad

La prueba de normalidad de kolmogorov – Smirnov (corregida por Lilliefors) (Abdi & Molin, 2007) fue utilizada para contrastar la hipótesis de ajuste de los datos a una distribución normal. Los resultados mostraron que la mayor parte de las variables no cumple con la condición de normalidad tal como se muestra en el gráfico 4.4. La línea de color negro representa el nivel de significancia de 0.05. Los puntos corresponden a los valores_p obtenidos al evaluar cada longitud de onda. Los puntos que se ubican sobre la línea corresponden a los valores_p mayores al nivel de significancia, en los que la hipótesis no es rechazada y los que se encuentran debajo de la línea corresponden a los valores_p menores al nivel de significancia por lo que la hipótesis se rechaza. En 377 longitudes de onda el valor_p fue menor a 0.05 por lo que se rechazó la hipótesis de

normalidad. En las restantes 143 longitudes de onda no hubo suficiente evidencia estadística para rechazar la hipótesis de normalidad.

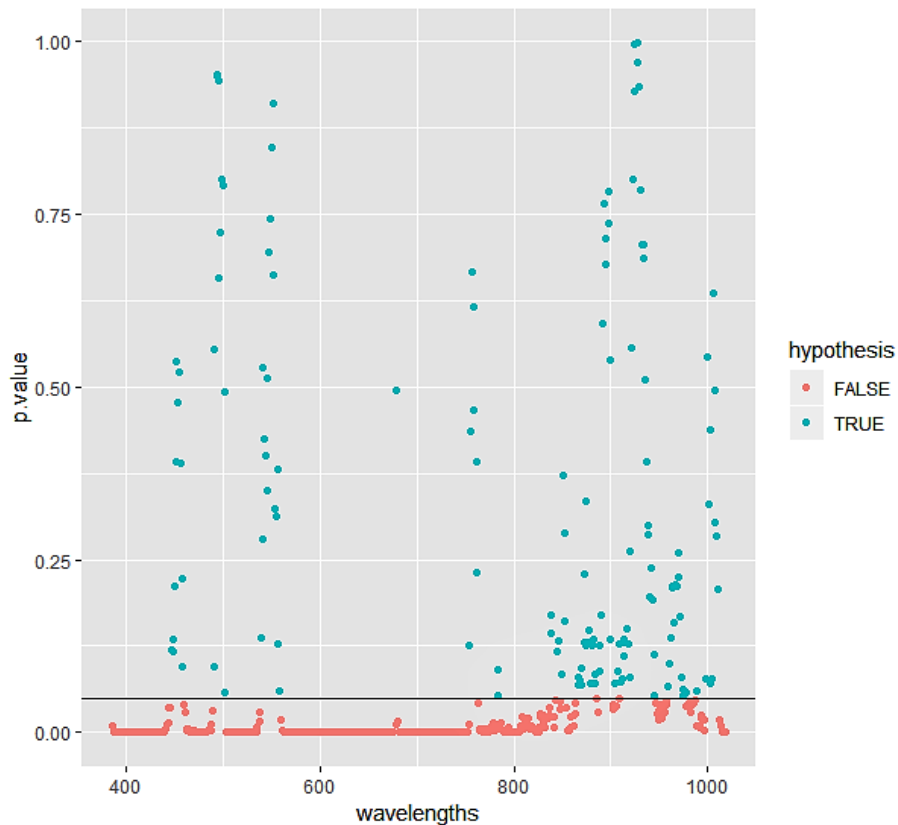


Figura 4.4 Prueba de Kolmogorov Smirnov (Lilliefors).

4.1.3 Análisis de multicolinealidad

La multicolinealidad es una condición de los datos en la que existe un alto grado de correlación entre las variables explicativas lo cual viola el supuesto de independencia lineal de estas variables establecido para regresión lineal múltiple y regresión logística.

La colinealidad sucede cuando alguno de los coeficientes de correlación simple o múltiple entre algunas de las variables independientes es 1.

Teóricamente hay 2 extremos de multicolinealidad, la multicolinealidad perfecta y no multicolinealidad. Si se cumple la segunda se dice que no hay no una relación lineal entre los regresores y por lo tanto son ortogonales. Cuando los predictores son ortogonales o no correlacionados, todos los valores propios de la matriz de diseño son iguales a uno y la matriz es de rango completo. En la práctica, cuando los regresores no son ortogonales, especialmente, si algún valor propio es igual a cero o cercano a cero, entonces existe multicolinealidad.

En aplicaciones de regresión múltiple surgen problemas de estimación de los coeficientes con la presencia de multicolinealidad porque la matriz $X'X$ es singular o casi singular por lo que los algoritmos de inversión de matrices deben dividir por un valor muy pequeño del determinante, siendo imprecisos e inestables con altos errores estándar producto de una matriz de varianzas elevadas. Además, la multicolinealidad dificulta la identificación de la importancia relativa de las variables correlacionadas incrementando la complejidad para la interpretación de los resultados por parte del investigador (Paul, 2006).

En el caso de la regresión logística, la multicolinealidad también afecta al supuesto de que la matriz de datos de entrada X es de rango completo (Hosmer Jr, Lemeshow, & Sturdivant, 2013). La solución de optimización de la función logverosimilitud utilizando el método de Newton-Raphson está dada por,

$$\beta = \beta + I^{-1}(\beta)U(\beta)$$

Donde,

$U(\beta) = X^T(y - \pi(\beta))$ vector de primeras derivadas parciales de la función logverosimilitud.

$I(\beta) = X^T \hat{V} X$ matriz de información estimada con $\hat{V} = \text{diag}\{\hat{\pi}_1(1 - \hat{\pi}_1), \dots, \hat{\pi}_n(1 - \hat{\pi}_n)\}$

Siendo X una matriz de rango incompleto es singular, lo que ocasiona que la matriz de información estimada no tenga inversa y por lo tanto no existe el estimador de máxima verosimilitud. Aún si la multicolinealidad no es perfecta, la matriz está cerca a la singularidad (Godínez-Jaimes et al., 2012)

4.1.3.1 Diagnóstico de multicolinealidad

Existen varias técnicas para detectar la presencia de multicolinealidad:

- El nivel de tolerancia
- El factor de inflación de la varianza VIF
- El test Farrar-Glauber
- Valores y vectores propios
- Examinación de la matriz de correlación

El nivel de tolerancia es estimado por $1 - R^2$ donde R^2 se calcula a partir de la regresión de la variable independiente de interés sobre las variables independientes restantes. Bajos niveles de tolerancia (menores que 0.4) indican alta multicolinealidad.

El VIF es el recíproco de la tolerancia, por lo tanto, altos valores del VIF (mayores que 2.5) indican relativamente altos niveles de multicolinealidad.

El test de Farrar-Glauber consiste en un procedimiento para detectar la multicolinealidad compuesto de tres test:

- Chi^2 indica la presencia de multicolinealidad
- F-test determina cuales son los regresores correlacionados.
- T-test determina la forma de la multicolinealidad.

Los valores propios de la matriz de correlación pequeños debido a las dependencias lineales entre las variables harán que el valor del número de condición de la matriz (CN) sea mayor. CN es la relación entre el mayor y el menor valor propio. Si CN es mayor que 100 a multicolinealidad es moderada y si es mayor que 1000 es severa (Adeboye, Fagoyinbo, & Olatayo, 2014).

Una forma simple evaluar la presencia de multicolinealidad es la examinación de los elementos diferentes a la diagonal de la matriz de correlación. Si dos regresores son linealmente dependientes la correlación será cercana a 1.

En este trabajo se utilizaron dos técnicas para detectar el grado de multicolinealidad: la inspección de la matriz de correlación y el factor de inflación de varianza (VIF) con la ayuda de funciones del lenguaje R.

4.1.3.2 Inspección de la Matriz de correlación

Utilizando las variables estandarizadas calculamos la matriz de correlación $X'X$ y hacemos una inspección de los valores altos de correlación.

La función *cor()* de R nos permite calcular la matriz de correlación. Debido al elevado número de variables el resultado se representó en el gráfico de calor de la figura 4.5. El color azul representa la correlación positiva perfecta (1) y el color rojo la correlación negativa perfecta (-1), mientras que el color blanco representa una correlación igual a cero, es decir que las variables son independientes. Los tonos entre los 2 colores principales disminuyen acorde con la disminución de la correlación. Se puede observar que la diagonal está coloreada con el color azul que representa una correlación igual a 1. El gráfico muestra alta correlación entre la mayoría de las variables. Alta multicolinealidad es una característica de los datos hiperespectrales que se evidencia por la presencia de áreas amplias con coloreadas con azul y rojo intensos.

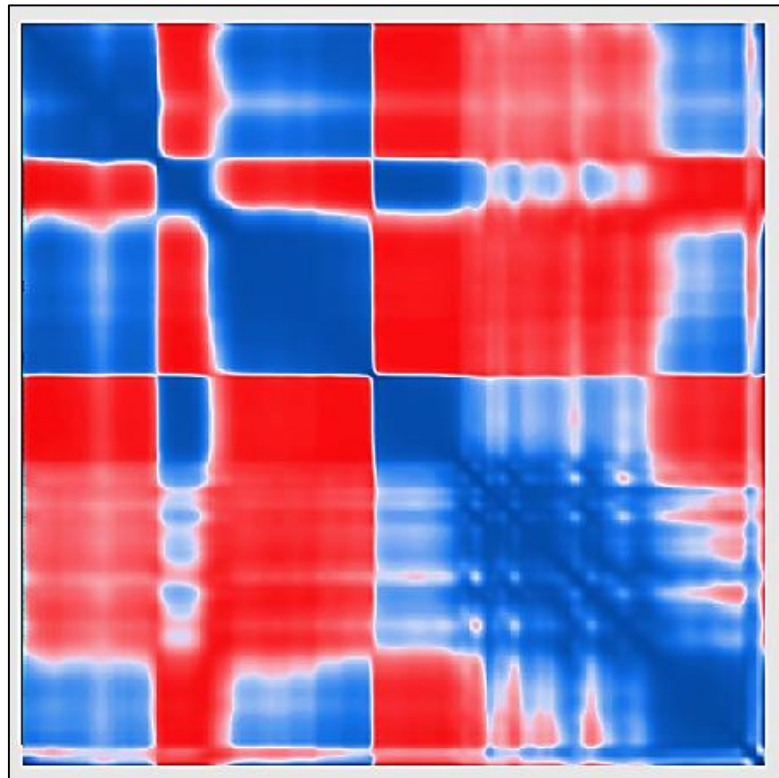


Figura 4.5 Gráfico de calor de la matriz de correlación de variables predictoras.

4.1.3.3 Factor de inflación de varianza (VIF)

Otra forma de evaluar la multicolinealidad es la inspección de los elementos de la diagonal de la inversa de la matriz $X'X$.

El valor VIF de una variable corresponde a la diagonal de la matriz $C=(X'X)^{-1}$, es decir,

$$VIF = C_{jj} = (1 - R_j^2)^{-1}$$

Los valores altos de VIF evidencian la presencia de multicolinealidad.

Utilizamos la función *vif()* del lenguaje R para calcular los VIF de variables explicativas.

Los resultados muestran un valor alto del VIF en todas las variables, lo que implica alta multicolinealidad acorde con el resultado observado en la matriz de correlación.

4.2 PLS-PLR

Utilizando la matriz de datos reducida se entrenó un modelo predictivo PLS-PLR de BLSD y se evaluó su desempeño utilizando un dataset de prueba. En primer lugar, se seleccionó el parámetro de penalización Ridge (λ), luego el modelo fue ajustado utilizando el parámetro de penalización que ofreció el mejor ajuste y finalmente fue validado usando validación cruzada y validación externa.

4.2.1 Selección de parámetro de penalización Ridge (λ)

El algoritmo PLS-PLR fue ejecutado utilizando diferentes valores del coeficiente de penalización λ y los modelos generados en cada iteración fueron evaluados utilizando métricas de bondad de ajuste cuyos resultados pueden ser observados en la tabla 4-2.

Tabla 4-2 Métricas de bondad de ajuste para modelos PLS-PLR con valores incrementales de penalización Ridge (λ)

λ	<i>Diff-Deviance</i> ^a	<i>R²CoxSnell</i> ^b	<i>R²Nagelkerke</i> ^c	<i>R²MacFadden</i> ^d
0.1	88.488	0.573	0.994	0.991
0.2	88.005	0.571	0.991	0.986
0.3	87.668	0.570	0.988	0.982
0.4	87.405	0.568	0.986	0.979
0.5	87.187	0.568	0.985	0.976
0.6	86.999	0.567	0.984	0.974
0.7	86.832	0.566	0.982	0.972
0.8	86.682	0.565	0.981	0.971
0.9	86.543	0.565	0.980	0.969

^a difference of Deviance (Hosmer et al.,1998).

^b R^2 Cox&Snell, ^c R^2 Nagelkerke and ^d R^2 MacFadden son índices pseudo R^2 para modelos de regresión logística binaria (Allison, 2014; Walker & Smith, 2016).

Con un valor de λ igual a 0.1, se obtuvieron las mejores medidas de bondad de ajuste. La penalización $\lambda = 0.1$ fue aplicada en cada iteración para el cálculo de los coeficientes, manteniéndolos estables mientras se controla el error durante el proceso de maximización de la función verosimilitud.

La diferencia de devianza (*DiffDeviance*) fue 88.488, mostrando que el modelo ajustado retuvo la mayor varianza. De acuerdo con lo expresado en el apartado 3.6.2.3, este estadístico tiene una distribución Chi-cuadrado, con grados de libertad igual a la diferencia entre los números de parámetros de los modelos. Por lo tanto, con un valor_p de $6,097 \times 10^{-20}$, se demuestra que existe una significativa asociación entre las variables latentes y la variable respuesta.

El resultado de los *pseudo* R^2 fue el siguiente: el *MacFadden's* R^2 igual a 0.991 muestra un alto poder explicativo del modelo. *Cox & Snell* (0.573) and *Nagelkerke* R^2 (0.994) indican también una alta calidad de ajuste.

El modelo generado con las dos primeras componentes PLS está representado en la siguiente función:

$$P_y = \frac{e^{(27.226+1.546t_1+1.318t_2)}}{1 + e^{(27.226+1.546t_1+1.318t_2)}}$$

Donde,

P_y es la probabilidad de la presencia de la enfermedad

t_1 es la primera componente PLS

t_2 es la segunda componente PLS

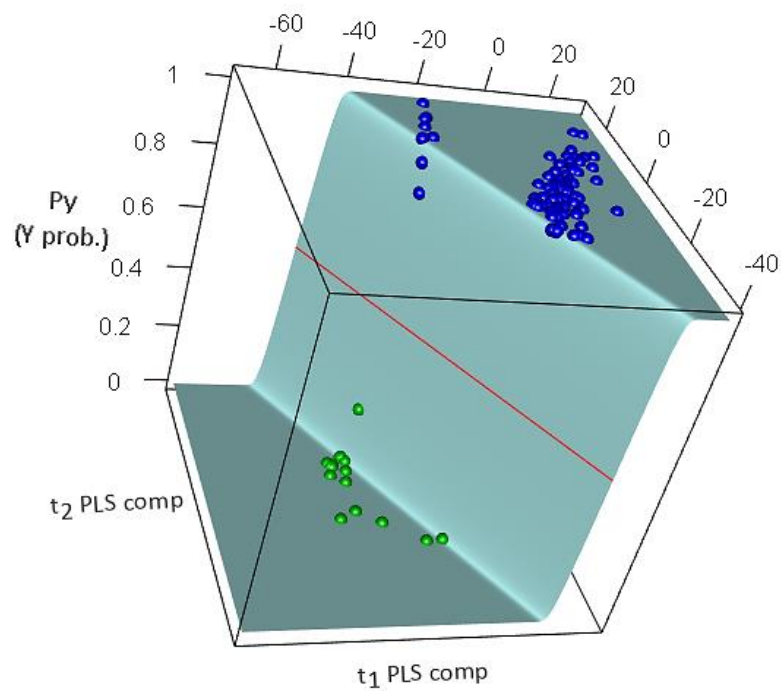


Figura 4.6 Gráfico 3D de la respuesta del modelo PLS-PLR.

La figura 4.6 muestra el modelo de respuesta logística ajustado en el espacio generado por las dos primeras componentes PLS. La línea roja corresponde al umbral de clasificación con probabilidad igual a 0.5 y separa los grupos con presencia y ausencia de la enfermedad. Las hojas de control se muestran como puntos verdes y las hojas infectadas se muestran como puntos azules.

4.2.2 Predicción y validación del modelo PLS-PLR

La capacidad predictiva del modelo fue valorada utilizando el método de validación cruzada LOOCV (Leave-One-Out-Cross-Validation). En esta prueba se evaluó cada una de las imágenes de las hojas (104 filas de la matriz de datos), separando una a una del set de datos y construyendo modelos PLS con las imágenes restantes ($n-1=103$ filas de la matriz). Como resultado del proceso de validación cruzada, 102 hojas fueron clasificadas correctamente, lo que representa una precisión en la clasificación del 98% (ver tabla 4-3). Dos hojas no infectadas fueron clasificadas erróneamente como infectadas con una probabilidad de 0.738 y 0.816, respectivamente. Todas las hojas infectadas fueron correctamente clasificadas.

La probabilidad estimada por el modelo como resultado del proceso de validación cruzada LOOCV es representada en la figura 4.7. Las hojas de control se muestran como puntos verdes y las hojas infectadas se presentan como puntos turquesa si son pre-sintomáticas, azules si pertenecen al nivel 1 de severidad y rojos si son de nivel 2 de severidad de la enfermedad.

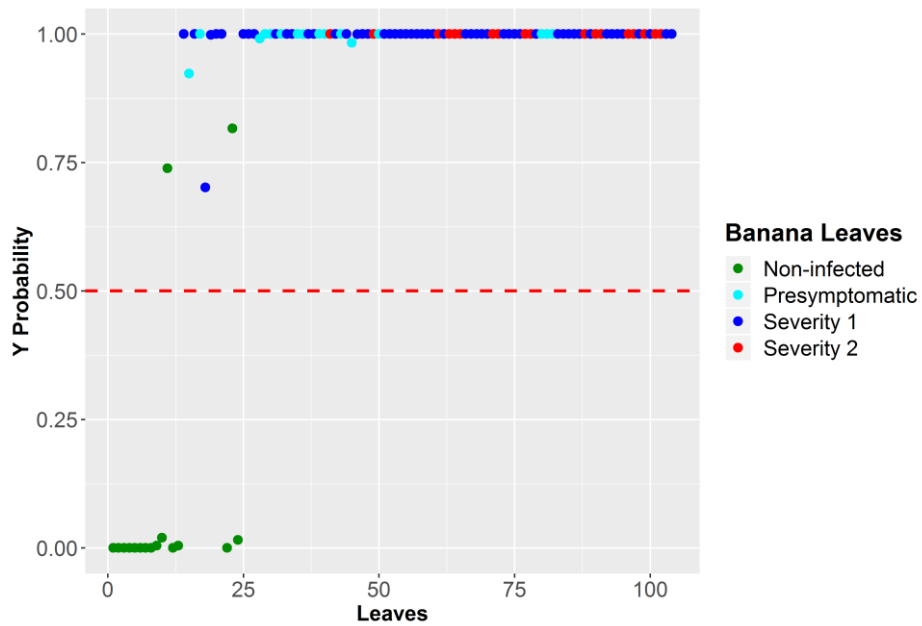


Figura 4.7 Probabilidad estimada por PLS-PLR con validación cruzada.

4.2.2.1 Matriz de confusión

Una forma de evaluar el desempeño de un algoritmo de clasificación es la matriz de confusión, que se muestra en la tabla 4-3. Las filas contienen el número de predicciones para cada clase realizadas por el modelo y las columnas corresponden a la clasificación real en la muestra.

Tabla 4-3 Matriz de confusión del modelo PLS-PLR con validación cruzada.

	Hojas Infectadas	Hojas no-infectadas	
Resultado Test	TP	FP	Precisión
	88	2	0.98
	FN	TN	Valor Pred. Negativo
	0	14	1
	Sensibilidad	Especificidad	Exactitud
	1	0.88	0.98

Exactitud.

$$Exactitud = \frac{TP + TN}{TP + FP + TN + FN} = \frac{88 + 14}{88 + 2 + 14 + 0} = 0.98$$

Error de clasificación.

$$Error\ de\ clasificación = \frac{FP + FN}{TP + FP + TN + FN} = \frac{2 + 0}{88 + 2 + 14 + 0} = 0.02$$

Sensibilidad.

$$Sensibilidad = \frac{TP}{TP + FN} = \frac{88}{88 + 0} = 1$$

Especificidad.

$$Especificidad = \frac{TN}{TN + FP} = \frac{14}{16} = 0.88$$

Precisión.

$$Precisión = \frac{TP}{TP + FP} = \frac{88}{88 + 2} = 0.98$$

Valor de predicción negativa.

$$Valor\ de\ predicción\ negativa = \frac{TN}{TN + FN} = \frac{14}{14} = 1$$

Prevalencia.

$$Prevalencia = \frac{TP + FN}{TP + FP + TN + FN} = \frac{88 + 0}{88 + 2 + 14 + 0} = 0.85$$

F₁.

$$F_1 = 2 \frac{Precisión * sensibilidad}{Precisión + sensibilidad} = 2 \frac{0.98 * 1}{0.98 + 1} = 0.99$$

Resumen:

En una muestra de 104 hojas, conformada por 88 hojas infectadas (prevalencia = 0.85) y 16 hojas sanas, el modelo pudo predecir correctamente 102 imágenes (98 %) mientras que 2 (2 %) fueron clasificadas en forma incorrecta (exactitud = 0.98). El total de las hojas infectadas fueron clasificadas correctamente (sensibilidad = 1) mientras que 2 hojas sanas fueron clasificadas como infectadas (especificidad = 0.88). El 98 % de las hojas infectadas que fueron clasificadas correctamente (precisión 0.98). F₁ igual a 0.99 muestra que el modelo tiene alta confiabilidad en predicción y alta capacidad de discriminar los verdaderos positivos.

4.2.3 Validación Externa del modelo PLS-PLR

El modelo PLS-PLR ajustado al conjunto de datos de entrenamiento se usó para predecir la presencia de la enfermedad en nuevas hojas y evaluar la eficacia del modelo. En el nuevo dataset se incluyó imágenes de 16 hojas no infectadas y 16 infectadas con

La matriz de confusión obtenida como resultado de la prueba de validación externa se presenta en la tabla 4-4. Las filas contienen el resultado de la predicción para cada clase y las columnas son las etiquetas en la muestra.

Tabla 4-4 Matriz de confusión del modelo PLS-PLR en prueba de validación

	Hojas Infectadas	Hojas no-infectadas	
Resultado Test	TP	FP	Precisión
	15	1	0.94
	FN	TN	Valor Pred. Negativo
	1	15	0.94
	Sensibilidad	Especificidad	Exactitud
	0.94	0.94	0.94

Exactitud.

$$Exactitud = \frac{TP + TN}{TP + FP + TN + FN} = \frac{15 + 15}{15 + 1 + 15 + 1} = 0.94$$

Error de clasificación.

$$Error\ de\ clasificación = \frac{FP + FN}{TP + FP + TN + FN} = \frac{1 + 1}{15 + 1 + 15 + 1} = 0.0625$$

Sensibilidad.

$$Sensibilidad = \frac{TP}{TP + FN} = \frac{15}{15 + 1} = 0.94$$

Especificidad.

$$Especificidad = \frac{TN}{TN + FP} = \frac{15}{15 + 1} = 0.94$$

Precisión.

$$Precisión = \frac{TP}{TP + FP} = \frac{15}{15 + 1} = 0.94$$

Valor de predicción negativa.

$$Valor\ de\ predicción\ negativa = \frac{TN}{TN + FN} = \frac{15}{15 + 1} = 0.94$$

F₁.

$$F_1 = 2 \frac{Precisión * sensibilidad}{Precisión + sensibilidad} = 2 \frac{0.938 * 0.938}{0.938 + 0.938} = 0.94$$

Prevalencia: el total de positivos en el total de la muestra fue del 50 %.

$$Prevalencia = \frac{TP + FN}{TP + FP + TN + FN} = \frac{15 + 1}{15 + 1 + 15 + 1} = 0.5$$

Área bajo la curva ROC (AUC): para calcular el área bajo la curva ROC se utilizó la función *auc()* del lenguaje R, el resultado fue 0.94.

Resumen:

En una muestra de 104 hojas, conformada con el 50 % de hojas infectadas (prevalencia = 0.5), un 94 % de las hojas (30) fue predicha correctamente por el mientras que 2 fueron clasificadas en forma incorrecta (exactitud = 0.94). El 94% de las hojas infectadas fueron clasificadas correctamente (sensibilidad = 0.94) de la igual forma el 94 % hojas sanas fueron clasificadas como infectadas (especificidad = 0.94). El 94 % de las hojas clasificadas como infectadas estaban etiquetadas como tales en la muestra. La métrica F1 tiene un valor de 0.94 y AUC fue igual a 0.94.

4.3 ANÁLISIS HS-BIPLLOT

La construcción de un modelo con rasgo latente como PLS-PLR busca la interpretación de las relaciones de los individuos y variables responsables de la reducción de las dimensiones, en otras palabras, su propósito se centra en la explicación de la relación entre variables observadas en términos de las variables latentes y su relación con los individuos representados por las puntuaciones en dichas variables latentes. En este caso, el HS-Biplot es útil para entender la estructura de los datos y para explorar posibles patrones y las relaciones entre los individuos y las variables, por lo que resulta una herramienta complementaria al PLS-PLR que agrega el componente explicativo de la estructura de datos al poder predictivo de PLS.

La estructura latente PLS fue usada para presentar el Hyperspectral Biplot (HS-Biplot), una herramienta gráfica que tiene el propósito de explorar las relaciones entre los grupos de hojas y las longitudes de onda. El HS-Biplot provee evidencia visual del agrupamiento de los individuos de la muestra con características relacionadas a conocidas propiedades químicas y físicas que se manifiestan en las longitudes de onda de las regiones visible e infrarrojo cercano. El HS-biplot del dataset de entrenamiento se muestra en la figura 4.9.

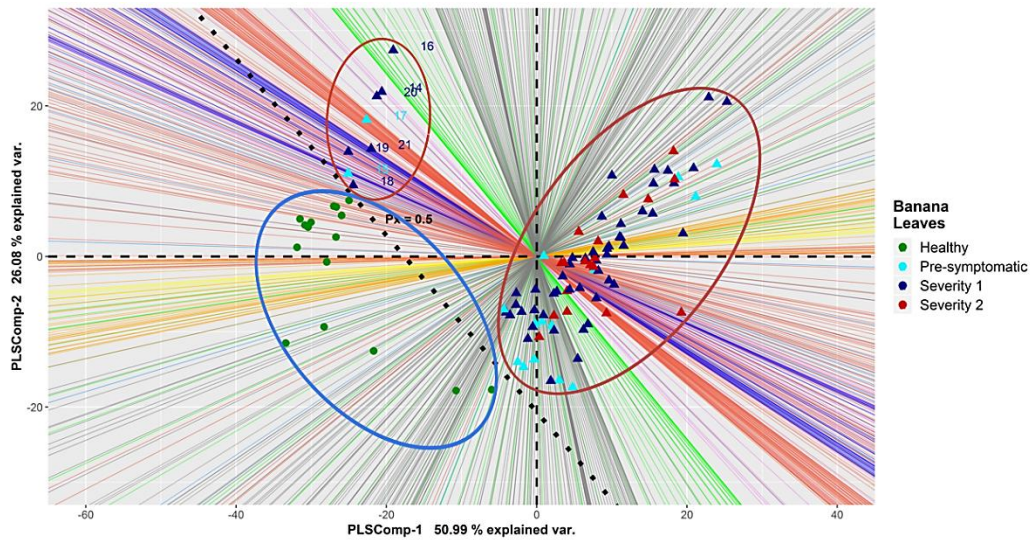


Figura 4.9 HS-Biplot del dataset de entrenamiento.

El plano cartesiano de la representación HS-Biplot está formado por las primera y segunda componente PLS como ejes de abscisas y ordenadas respectivamente. Las primeras dos componentes PLS contribuyeron con el 77% de variabilidad observada, la primera aportó con el 50.99% de la variabilidad y la segunda el 26.08%.

Las filas de la matriz reducida (hojas de banano) están representadas por puntos y las columnas (longitudes de onda) están representadas por líneas rectas. A partir de la factorización obtenida en el modelo PLS, las puntuaciones (scores) en la matriz T son coordenadas de las filas y las cargas (loadings) en la matriz P proporcionan la dirección de las variables de matriz original. Las variables originales han sido coloreadas de acuerdo con la banda espectral a la que pertenecen (ver tabla 3-1).

El espacio cartesiano está dividido por una línea puntuada oblicua formando dos regiones separadas que predicen la presencia o ausencia de enfermedad. La línea corta el

plano en el valor de predicción de 0.5 y corresponde al umbral de clasificación, lo que implica que los puntos ubicados sobre dicha línea son hojas clasificadas como infectadas y, por el contrario, si están abajo de la línea son hojas no infectadas.

El HS-biplot muestra tres agrupaciones principales: un grupo de hojas no infectadas (elipse azul) y dos grupos de hojas infectadas (elipses rojas). La agrupación de plantas no infectadas e infectadas se observó principalmente en el componente 1. Las longitudes de onda que más contribuyeron a la agrupación de las muestras no infectadas estuvieron en el rango de 577 nm a 651 nm (rango amarillo - rojo). En cuanto a las hojas infectadas, se observaron 2 grupos. El primer grupo (puntos numerados dentro de la elipse roja) se ubicó cerca del grupo de hojas no infectado, influenciado, principalmente, por una baja densidad de síntomas de enfermedad en las hojas, tal como se muestra en la tabla 4-5. El otro grupo de hojas infectadas (puntos no numerados dentro de la segunda elipse roja) estaba compuesto de hojas pre-sintomáticas y sintomáticas. Las longitudes de onda que más contribuyeron a la agrupación de las hojas pre-sintomáticas (turquesa), así como varias hojas en los niveles de severidad 1 (azul) y 2 (rojo) estaban en el rango NIR del espectro. Sin embargo, la agrupación de las otras hojas en los niveles sintomáticos 1 y 2 se produjo debido a su reflectancia en el rango de 577 nm a 651 nm (amarillo - rojo).

Tabla 4-5 Detalle de las hojas con bajo grado de infección

Número	Severidad	Observación
14	1	Presenta 2 pixeles severidad 1
15	0	Sin síntomas
16	1	Presnta 7 pixeles severidad 1
17	0	Sin síntomas
18	1	Sin síntomas

19	1	Presenta 10 pixeles severidad 1
20	1	Presenta 6 pixeles severidad 1
21	1	Presenta 19 pixeles severidad 1

Se observó exactamente el mismo comportamiento en el conjunto de datos de validación externa (fig. 4.10). Hs-biplot del conjunto de datos de prueba muestra la clasificación correcta de hojas sanas (elipse azul) e infectadas (elipse roja), excepto dos, una sana y una infectada.

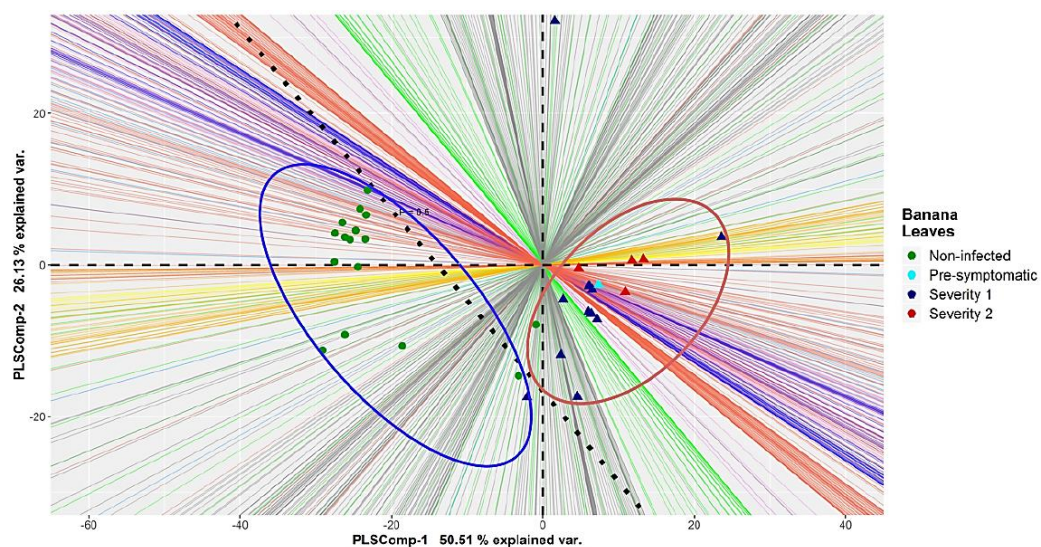


Figura 4.10 HS-Biplot del dataset de validación.

La bondad de ajuste global del HS-Biplot (el coeficiente de correlación al cuadrado entre los valores ajustados y observados) fue del 77.07%.

$$\rho^2 = \frac{\text{tr}(\hat{\mathbf{X}}^T \hat{\mathbf{X}})}{\text{tr}(\mathbf{X}^T \mathbf{X})} \times 100 = 77.07$$

Donde

$\hat{\mathbf{X}} = \mathbf{TP}^T$ es la matriz de datos aproximada.

$\mathbf{X}$ es la matriz de datos original

La contribución de las componentes a cada variable es la cantidad de varianza de cada columna capturada por la aproximación, también llamada calidad de representación o predictividad de columnas. La tabla de las contribuciones a las columnas se presenta en el APÉNDICE A.

$$\rho_j^2 = \frac{\text{tr}(\hat{\mathbf{x}}_{[j]}^T \hat{\mathbf{x}}_{[j]})}{\text{tr}(\mathbf{x}_{[j]}^T \mathbf{x}_{[j]})} \times 100 \quad (j = 1, \dots, p)$$

ρ_j^2 es la contribución de cada columna j

$\hat{\mathbf{x}}_{[j]}$ la columna j de la matriz aproximada $\hat{\mathbf{X}}$

$\mathbf{x}_{[j]}$ es la columna j de la matriz original $\mathbf{X}$

La contribución de las componentes a cada fila o individuo es la cantidad de varianza de cada fila capturada por la aproximación, también llamada calidad de representación o predictividad de filas. La tabla de las contribuciones a las filas se presenta en el APÉNDICE B.

$$\rho_i^2 = \frac{\text{tr}(\hat{\mathbf{x}}_{[i]} \hat{\mathbf{x}}_{[i]}^T)}{\text{tr}(\mathbf{x}_{[i]} \mathbf{x}_{[i]}^T)} \times 100 \quad (i = 1, \dots, n)$$

ρ_i^2 es la contribución de cada columna i

$\hat{x}_{[i,]}$ es la fila i de la matriz aproximada $\hat{X}$

$x_{[i,]}$ es la fila i de la matriz original X

4.4 NPLS-DA

4.4.1 Entrenamiento del modelo NPLS-DA

El método NPLS-DA, explicado en el apartado 3.6.4, fue utilizado para entrenar un modelo con las imágenes hiperespectrales de 104 hojas de banano. Los cubos HS fueron sometidos a una fase de preprocesamiento que consistió en la obtención de cinco características (media, desviación estándar, simetría, curtosis, quinto momento) en cada longitud de onda convirtiendo las estructuras tridimensionales en matrices de 5×520 . Las 104 matrices de 2 dimensiones fueron apiladas creando un tensor de tercer orden de dimensiones $I=104 \times J=5 \times K=520$ (figura 4.11).

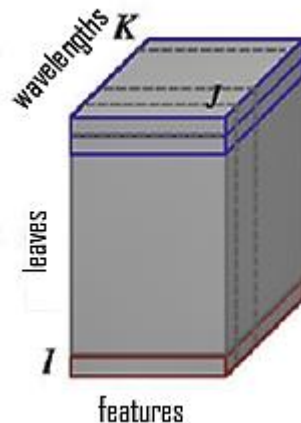


Figura 4.11 Tensor de características de cubos hiperespectrales.

Para lograr la mejor tasa de predicción se entrenó el modelo NPLS-DA incrementando el número de componentes (figura 4.12), logrando los mejores resultados con 7 componentes (figura 4.13). El modelo construido no logró la separación completa de los elementos de la muestra, aunque el rendimiento fue aceptable (tabla 4-6).

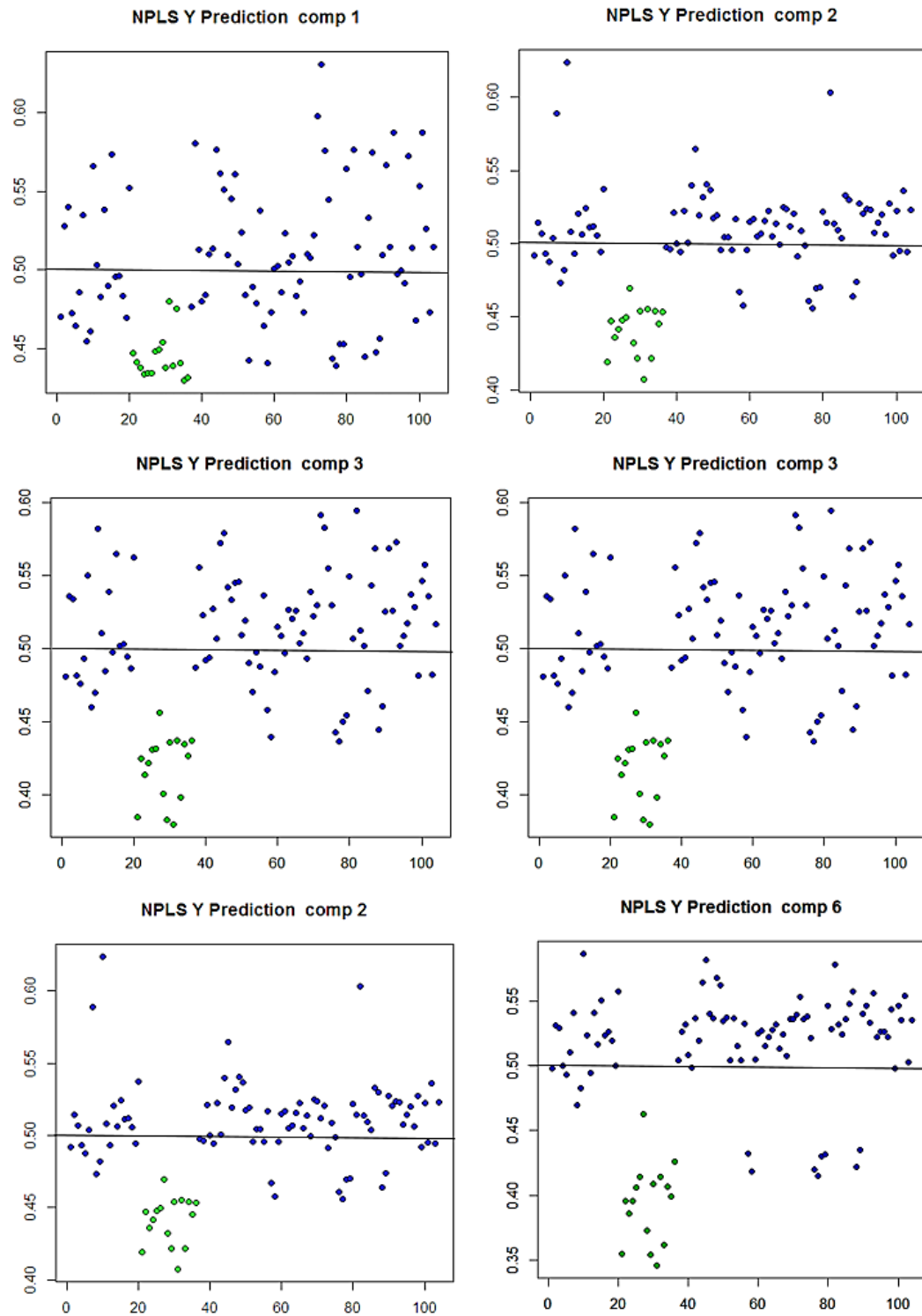


Figura 4.12 Predicción NPLS-DA con 1 a 6 componentes

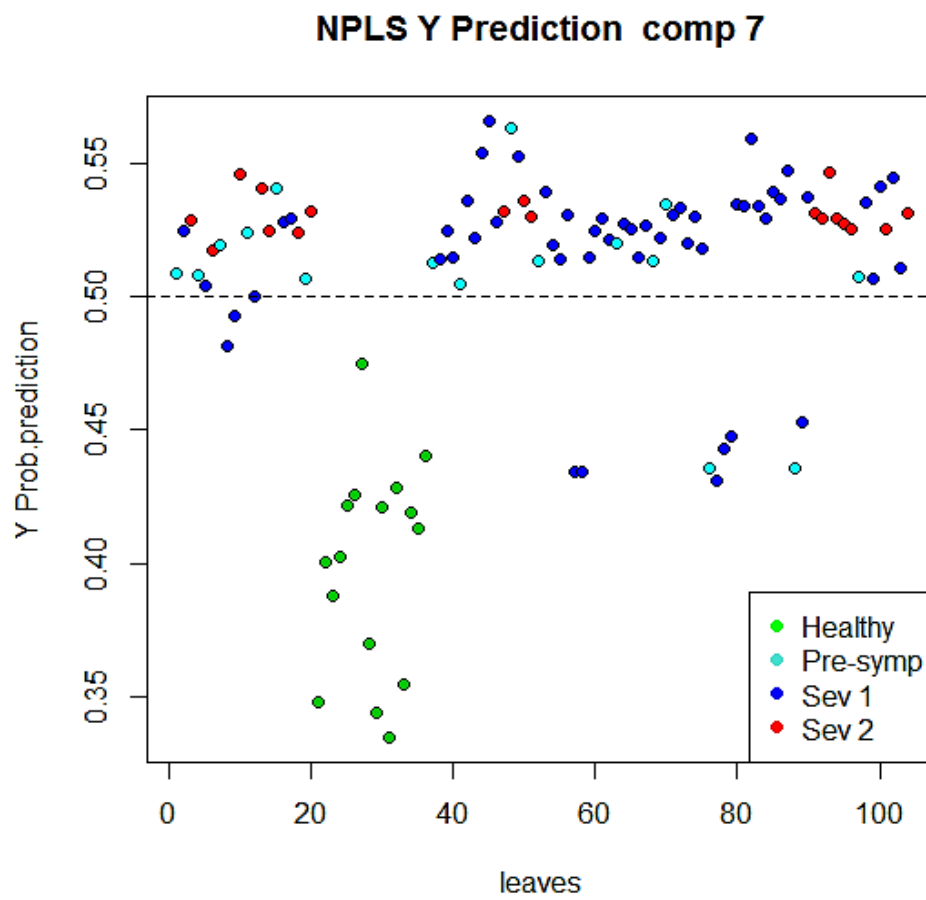


Figura 4.13 Predicción del modelo NPLS_{DA} con datos de entrenamiento.

Tabla 4-6 Matriz de confusión del modelo NPLS-DA con datos de entrenamiento

	Hojas Infectadas	Hojas no-infectadas	
Resultado Test	TP	FP	Precisión
	78	0	1
	FN	TN	Valor Pred. negativo
	10	16	0.62
	Sensibilidad	Especificidad	Exactitud
	0.89	1	0.90

Exactitud.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} = \frac{78 + 16}{78 + 0 + 16 + 10} = 0.904$$

Error de clasificación.

$$Error\ de\ clasificación = \frac{FP + FN}{TP + FP + TN + FN} = \frac{0 + 10}{88 + 2 + 14 + 0} = 0.096$$

Sensibilidad.

$$Sensibilidad = \frac{TP}{TP + FN} = \frac{76}{76 + 10} = 0.89$$

Especificidad.

$$\text{Especificidad} = \frac{TN}{TN + FP} = \frac{16}{16 + 0} = 1$$

Precisión.

$$\text{Precisión} = \frac{TP}{TP + FP} = \frac{78}{78 + 0} = 1$$

Valor de predicción negativa.

$$\text{Valor de predicción negativa} = \frac{TN}{TN + FN} = \frac{16}{16 + 10} = 0.62$$

Prevalencia.

$$\text{Prevalencia} = \frac{TP + FN}{TP + FP + TN + FN} = \frac{78 + 10}{78 + 0 + 16 + 10} = 0.85$$

F₁.

$$F_1 = 2 \frac{\text{Precisión} * \text{sensibilidad}}{\text{Precisión} + \text{sensibilidad}} = 2 \frac{1 * 0.88}{1 + 0.88} = 0.94$$

Resumen:

La muestra está conformada por 88 hojas infectadas que representan el 85% (prevalencia = 0.85) y 16 hojas sanas que representan el 15 % restante. El porcentaje de aciertos en la predicción fue del 90.4 % (exactitud = 0.904) que corresponde a 94 hojas

del total de 104, 10 fueron clasificadas en forma incorrecta. El 89 % de las hojas infectadas fueron clasificadas correctamente (sensibilidad = 0.89) y todas las hojas sanas fueron clasificadas correctamente (especificidad = 1). El resultado de F_1 fue 0.94.

4.4.2 Validación del modelo NPLS-DA

El modelo NPLS-DA ajustado al dataset de entrenamiento fue evaluado utilizando un nuevo dataset de prueba con 32 imágenes de hojas de banano. 16 imágenes de hojas sanas y 16 de hojas infectadas.

La figura 4.14 muestra los resultados de predicción.

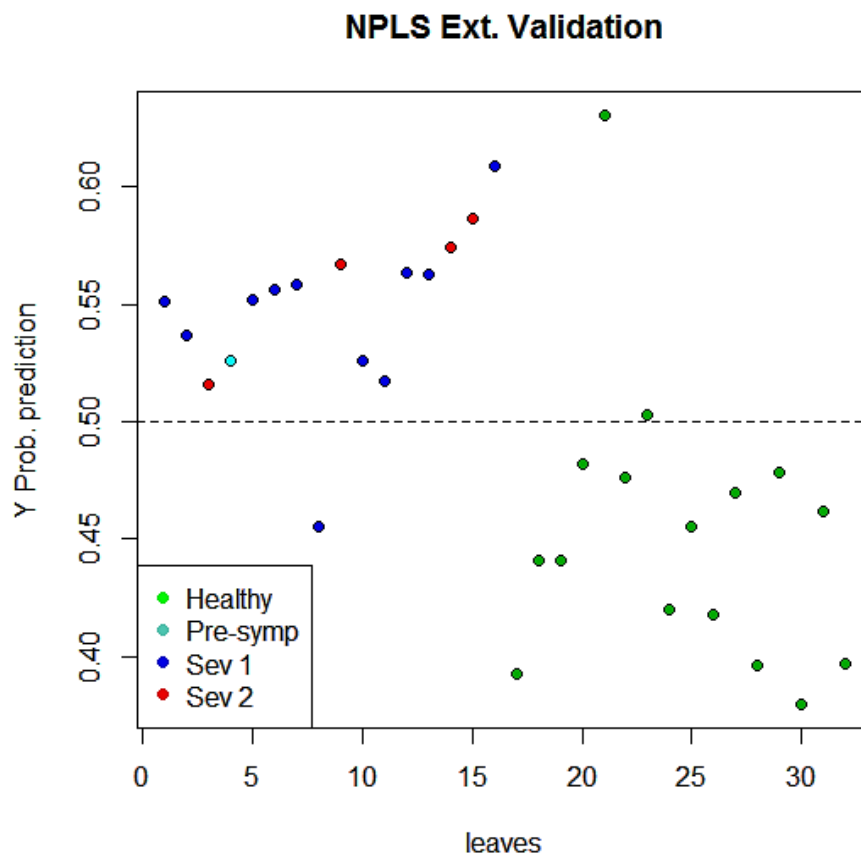


Figura 4.14 Predicción del modelo NPLS_DA en prueba de validación.

Tabla 4-7 Matriz de confusión para evaluación del modelo NPLS-DA

	Hojas Infectadas	Hojas no-infectadas	
Resultado Test	TP	FP	Precisión
	15	2	0.88
	FN	TN	Valor Pred. negativo
	1	14	0.93
	Sensibilidad	Especificidad	Exactitud
	0.94	0.88	0.91

Exactitud.

$$Exactitud = \frac{TP + TN}{TP + FP + TN + FN} = \frac{15 + 14}{15 + 2 + 14 + 1} = 0.91$$

Error de clasificación.

$$Error\ de\ clasificación = \frac{FP + FN}{TP + FP + TN + FN} = \frac{2 + 1}{15 + 2 + 14 + 1} = 0.09$$

Precisión.

$$Precisión = \frac{TP}{TP + FP} = \frac{15}{15 + 2} = 0.88$$

La sensibilidad.

$$\text{Sensibilidad} = \frac{TP}{TP + FN} = \frac{15}{15 + 1} = 0.94$$

Especificidad.

$$\text{Especificidad} = \frac{TN}{TN + FP} = \frac{14}{14 + 2} = 0.88$$

Valor de predicción negativa.

$$\text{Valor de predicción negativa} = \frac{TN}{TN + FN} = \frac{14}{14 + 1} = 0.93$$

F₁.

$$F_1 = 2 \frac{\text{Precisión} * \text{sensibilidad}}{\text{Precisión} + \text{sensibilidad}} = 2 \frac{0.88 * 0.94}{0.88 + 0.94} = 0.91$$

Prevalencia.

$$\text{Prevalencia} = \frac{TP + FN}{TP + FP + TN + FN} = \frac{15 + 1}{15 + 2 + 14 + 1} = 0.5$$

Área bajo la curva ROC (AUC). El valor AUC se obtuvo mediante la función *auc()* del lenguaje R, el resultado fue 0.91.

Resumen:

La muestra estuvo conformada por 16 hojas infectadas y 16 hojas sanas que corresponde a una prevalencia fue 0.5. El 91 % de las hojas fueron predichas correctamente, esto es 29 hojas mientras que 3 fueron clasificadas en forma incorrecta (exactitud = 0.91). El 94 % (15) de las hojas infectadas fueron clasificadas correctamente (sensibilidad = 0.94), mientras que 14 hojas sanas fueron clasificadas en forma correcta (especificidad = 0.88). F_1 fue 0.91 y el AUC fue 0.91.

4.5 SVM

La gran popularidad de las máquinas de soporte se debe a su capacidad para producir modelos con alto desempeño para múltiples tipos de aplicaciones. En principio se enmarcaron en resolver problemas de clasificación, pero con el tiempo ampliaron su aplicación a la clasificación multi-categorías, agrupamiento y regresión.

En nuestro caso, el modelo requerido es un clasificador binario para discriminar las hojas infectadas y no infectadas. Con este propósito se partió de un modelo de clasificación lineal y posteriormente se extendió su complejidad utilizando las funciones kernel polinómico y radial. El algoritmo SVM busca el hiperplano con menor riesgo estructural por medio de la maximización del margen de separación de las dos clases (*Maximal Margin Classifier*).

La función Kernel transforma los datos de entrada proyectándolos a un espacio de características de mayor dimensión (espacio Hilbert). En esta investigación fueron evaluados clasificadores SVM con kernel lineal, polinómico y radial, logrando los resultados de mayor precisión en la predicción con los dos primeros.

Si las clases son linealmente separables se puede encontrar un hiperplano de margen máximo que clasifica correctamente los elementos en este caso el margen es duro y no se permiten errores de entrenamiento. Si las clases no son linealmente separables entonces el margen debe ser permisivo a ciertos errores en la clasificación para encontrar un hiperplano de margen máximo, este tipo de margen se lo denomina blando.

En la función de coste, el parámetro C regula la permisibilidad de errores. Si el valor de C es muy grande, la anchura de margen óptimo será cada vez más estrecha y el modelo resultante se ajustará bastante a los datos con un bias reducido, pero una alta varianza y riesgo de sobreajuste. Por otro lado, si C es muy pequeño el margen óptimo será más amplio, permitiendo elementos de la muestra dentro del margen e incluso mal clasificados generando un mayor bias, pero reduciendo la varianza. La optimización del parámetro de regularización fue realizada mediante validación cruzada.

La implementación del clasificador SVM fue realizada en lenguaje PYTHON.

4.5.1 SVM Lineal

4.5.1.1 Entrenamiento del modelo SVM lineal

El modelo SVM inicial es un clasificador básico implementado mediante una función kernel lineal. En la fase de entrenamiento se utilizó la tabla reducida de dimensiones 104×520 que representa a 104 hojas de banano y 520 longitudes de onda.

La optimización del hiperparámetro de regularización C permite el control de las violaciones de las observaciones sobre el margen del hiperplano, favoreciendo al equilibrio entre el sesgo (bias) y la varianza. Utilizando un valor de regulación igual a 1, los resultados fueron los detallados a continuación.

La probabilidad de infección estimada por el modelo SVM lineal se muestra en la figura 4.15.

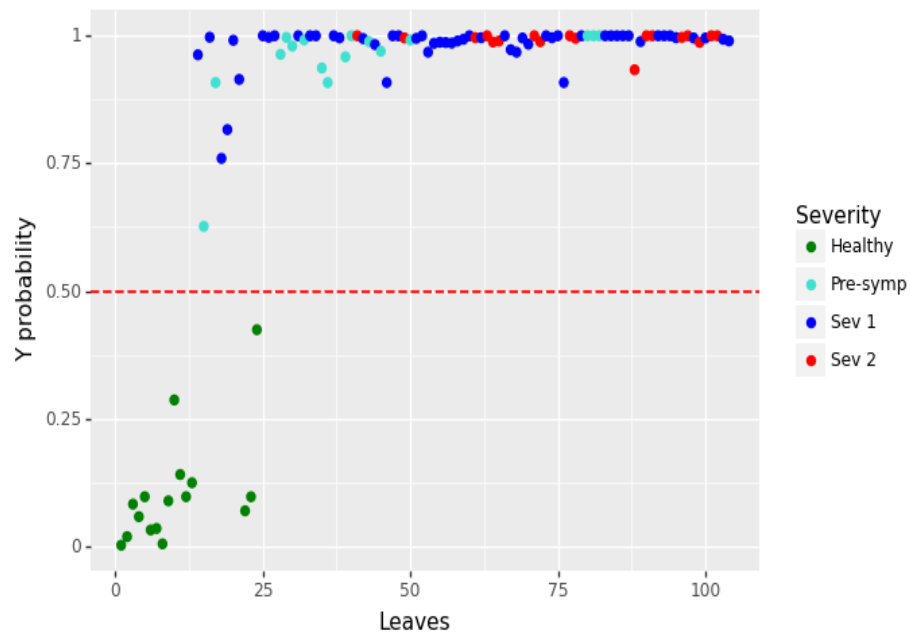


Figura 4.15 Probabilidad estimada con el modelo SVM lineal con datos de entrenamiento.

Las métricas de predicción utilizando el dataset de entrenamiento y la matriz de confusión (figura 4.16) se muestran a continuación.

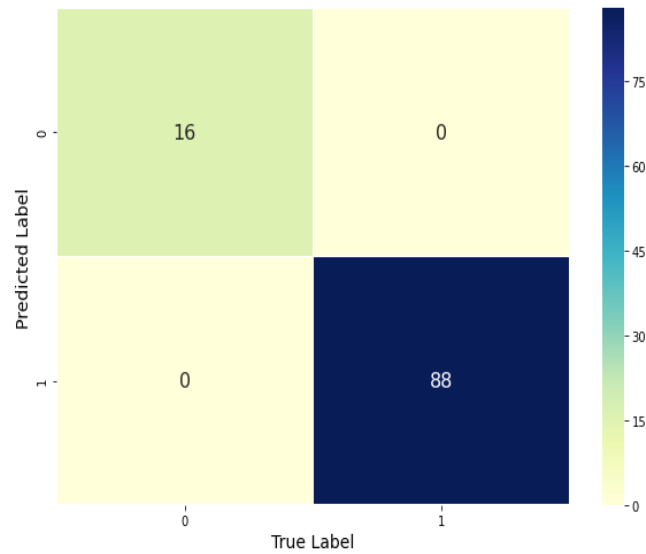


Figura 4.16 Matriz de confusión de entrenamiento del modelo SVM lineal.

Tabla 4-8 Métricas de predicción del modelo SVM lineal con datos de entrenamiento

	Hojas infectadas	Hojas no-infectadas	
Resultado Test	TP	FP	Precisión
	88	0	1
	FN	TN	Valor Pred. negativo
	0	16	1
	Sensibilidad	Especificidad	Exactitud
	1	1	1

Exactitud.

$$Exactitud = \frac{TP + TN}{TP + FP + TN + FN} = \frac{88 + 16}{88 + 0 + 16 + 0} = 1$$

Error de clasificación.

$$\text{Error de clasificación} = \frac{FP + FN}{TP + FP + TN + FN} = \frac{0 + 0}{88 + 0 + 16 + 0} = 0$$

Sensibilidad.

$$\text{Sensibilidad} = \frac{TP}{TP + FN} = \frac{88}{88 + 0} = 1$$

Especificidad.

$$\text{Especificidad} = \frac{TN}{TN + FP} = \frac{16}{16 + 0} = 1$$

Precisión.

$$\text{Precisión} = \frac{TP}{TP + FP} = \frac{88}{88 + 0} = 1$$

Valor de predicción negativa.

$$\text{Valor de predicción negativa} = \frac{TN}{TN + FN} = \frac{16}{16 + 0} = 1$$

F₁.

$$F_1 = 2 \frac{\text{Precisión} * \text{sensibilidad}}{\text{Precisión} + \text{sensibilidad}} = 2 \frac{1 * 1}{1 + 1} = 1$$

Prevalencia.

$$Prevalencia = \frac{TP + FN}{TP + FP + TN + FN} = \frac{88 + 0}{88 + 0 + 16 + 0} = 0.85$$

Resumen:

En una muestra conformada con un 85 % de hojas infectadas (prevalencia = 0.85), el 100 % de las hojas (88) fue predicha correctamente por lo tanto no hubo error en la predicción de ninguna hoja (exactitud = 1). El 100 % de las hojas infectadas fueron clasificadas correctamente (sensibilidad = 1) de la igual forma el 100 % hojas sanas fueron clasificadas como infectadas (especificidad = 1). El 100% de las hojas clasificadas como infectadas corresponden a hojas etiquetadas como infectadas (precisión = 1).

4.5.1.2 Validación del modelo SVM lineal

La validación externa se llevó a cabo utilizando el dataset de prueba con 32 imágenes, 16 sanas y 16 enfermas. La figura 4.17 muestra la predicción de probabilidad de infección del dataset de prueba con el modelo SVM lineal.

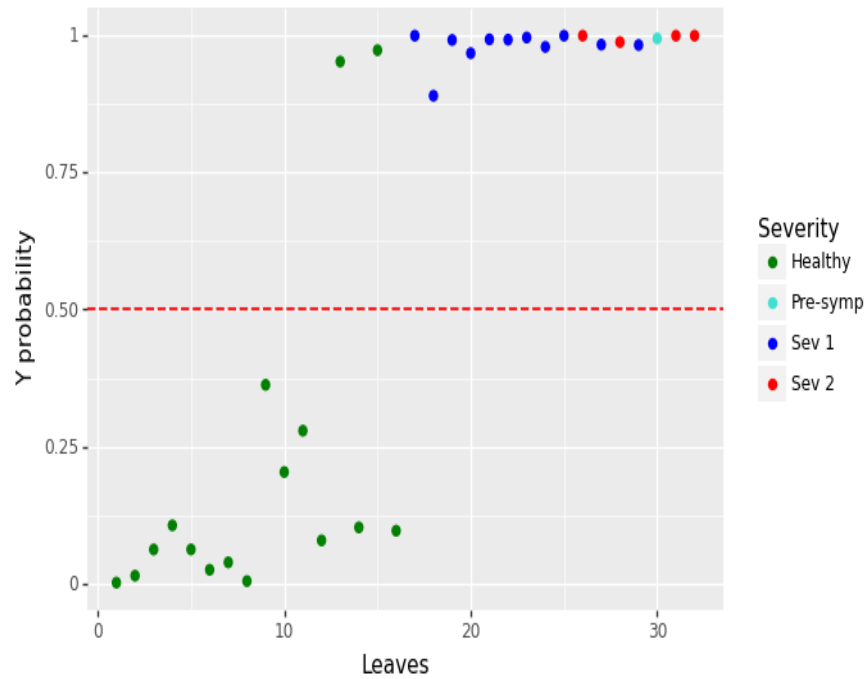


Figura 4.17 Probabilidad estimada por el modelo SVM lineal en prueba de validación.

La figura 4.18 muestra los resultados de la matriz de confusión.

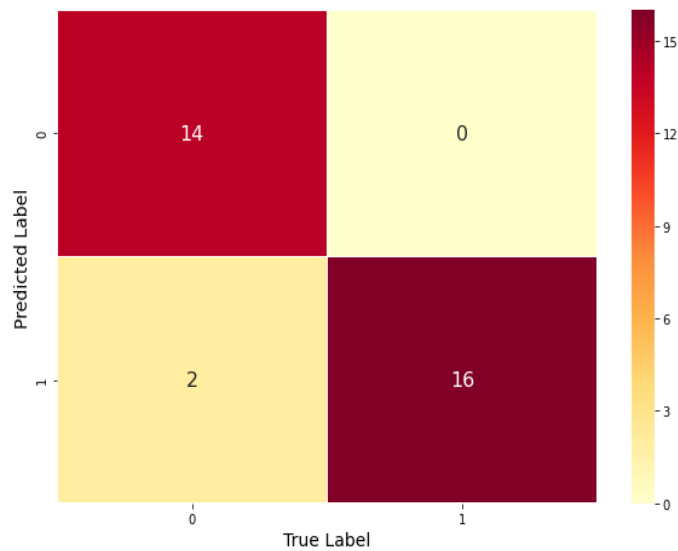


Figura 4.18 Matriz de confusión de validación del modelo SVM lineal.

Tabla 4-9 Evaluación de métricas de predicción de datos de prueba usando el modelo SVM lineal

	Hojas Infectadas	Hojas no-infectadas	
Resultado Test	TP	FP	Precisión
	16	2	0.89
	FN	TN	Valor Pred. negativo
	0	14	1
	Sensibilidad	Especificidad	Exactitud
	1	0.88	0.94

Exactitud.

$$Exactitud = \frac{TP + TN}{TP + FP + TN + FN} = \frac{16 + 14}{16 + 2 + 14 + 0} = 0.94$$

Error de clasificación.

$$\text{Error de clasificación} = \frac{FP + FN}{TP + FP + TN + FN} = \frac{2 + 0}{14 + 2 + 16 + 0} = 0.0625$$

Sensibilidad.

$$\text{Sensibilidad} = \frac{TP}{TP + FN} = \frac{16}{16 + 0} = 1$$

Especificidad.

$$\text{Especificidad} = \frac{TN}{TN + FP} = \frac{14}{14 + 2} = 0.88$$

Precisión.

$$\text{Precisión} = \frac{TP}{TP + FP} = \frac{16}{16 + 2} = 0.89$$

Valor de predicción negativa.

$$\text{Valor de predicción negativa} = \frac{TN}{TN + FN} = \frac{14}{14 + 0} = 1$$

Prevalencia.

$$\text{Prevalencia} = \frac{TP + FN}{TP + FP + TN + FN} = \frac{14 + 2}{16 + 2 + 14 + 0} = 0.5$$

F1.

$$F_1 = 2 \frac{\text{Precisión} * \text{sensibilidad}}{\text{Precisión} + \text{sensibilidad}} = 2 \frac{0.889 * 1}{0.889 + 1} = 0.94$$

Área bajo la curva ROC (AUC). El resultado se obtuvo utilizando las métricas de scikit learn de Python. El resultado fue 0.94.

Resumen:

En una muestra de 32 hojas, conformada con el 50 % de hojas infectadas (prevalencia = 0.5), un 94 % de las hojas (30) fue predicha correctamente por el modelo mientras que 2 fueron clasificadas en forma incorrecta (exactitud = 0.94). Todas las hojas clasificadas como infectadas fueron clasificadas correctamente (sensibilidad = 1) de la igual forma el 88 % hojas clasificadas como sanas fueron clasificadas correctamente (especificidad = 0.88). La métrica F1 fue igual a 0.94 y el área bajo la curva ROC (AUC) fue 0.94.

4.5.2 SVM Polinomial

En las secciones anteriores se ha explicado las bondades de los hiperplanos como clasificadores cuando las clases son linealmente separables, pero en el caso de clases que no son claramente separables linealmente el uso del kernel lineal no ofrece buenos resultados. En estos casos el aumento de la dimensionalidad mediante la aplicación de kernels no lineales son una buena alternativa. Los Kernels transforman un espacio de pocas dimensiones en un espacio de dimensiones mayores.

A continuación, presentamos los resultados de las pruebas de predicción del modelo SVM con kernel polinomial.

4.5.2.1 Entrenamiento del modelo SVM polinomial

El hiperparámetro grado (degree) corresponde al grado de la función kernel polinómico el mismo que fue establecido en 2 y 3. El hiperparámetro de regulación C se lo incrementó desde 1 hasta 100. Los mejores resultados se obtuvieron con el grado igual a 2 y C igual a 91. En la figura 4.19 muestra la probabilidad estimada para los datos de entrenamiento.

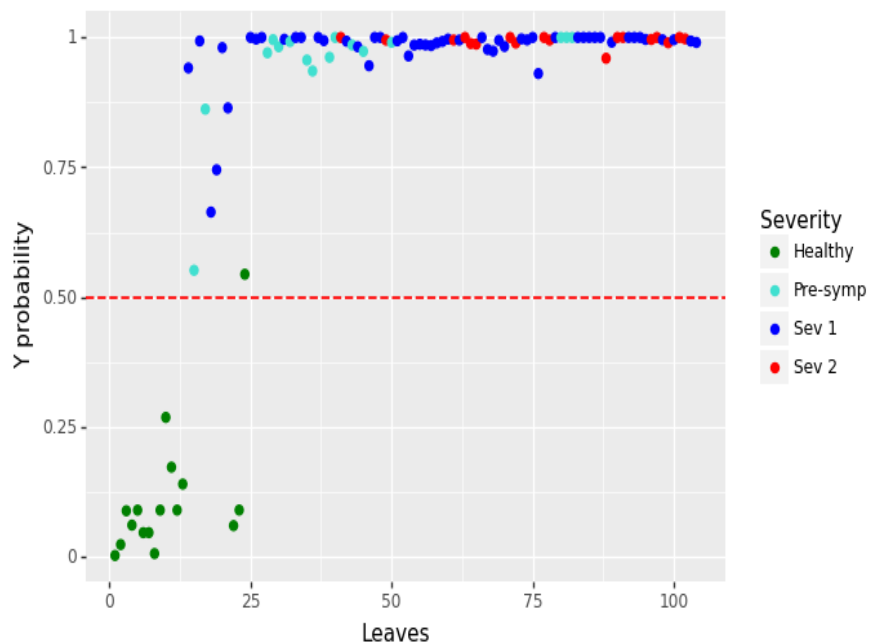


Figura 4.19 Probabilidad estimada por el modelo SVM polinomial con datos de entrenamiento.

La figura 4.20 se muestra la matriz de confusión para la predicción de datos de entrenamiento.

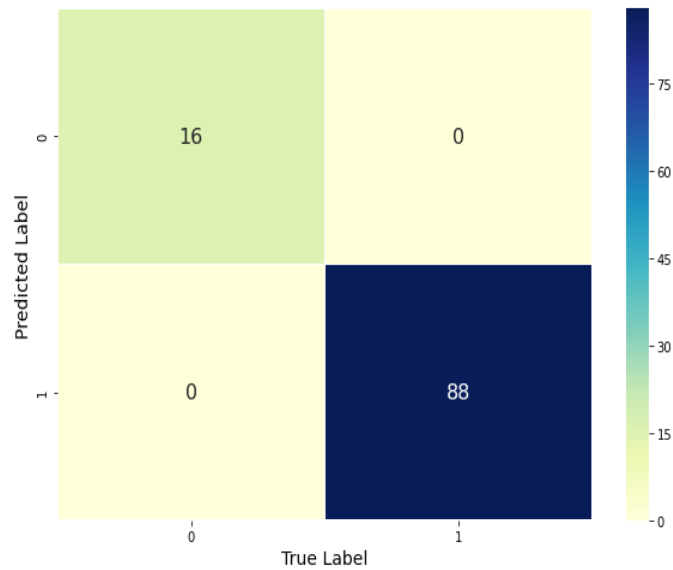


Figura 4.20 Matriz de confusión del modelo SVM polinomial con datos de entrenamiento.

Las métricas de predicción se describen a continuación.

Tabla 4-10 Métricas de predicción de datos de entrenamiento con el modelo SVM polinomial

	Hojas Infectadas	Hojas no-infectadas	
Resultado Test	TP	FP	Precisión
	88	0	1
	FN	TN	Valor Pred. negativo
	0	16	1
	Sensibilidad	Especificidad	Exactitud
	1	1	1

Exactitud.

$$Exactitud = \frac{TP + TN}{TP + FP + TN + FN} = \frac{88 + 16}{88 + 0 + 16 + 0} = 1$$

Error de clasificación.

$$\text{Error de clasificación} = \frac{FP + FN}{TP + FP + TN + FN} = \frac{0 + 0}{88 + 0 + 16 + 0} = 0$$

Sensibilidad.

$$\text{Sensibilidad} = \frac{TP}{TP + FN} = \frac{88}{88 + 0} = 1$$

Especificidad.

$$\text{Especificidad} = \frac{TN}{TN + FP} = \frac{16}{16} = 1$$

Precisión.

$$\text{Precisión} = \frac{TP}{TP + FP} = \frac{88}{88 + 0} = 1$$

Valor de predicción negativa.

$$\text{Valor de predicción negativa} = \frac{TN}{TN + FN} = \frac{16}{16 + 0} = 1$$

F₁.

$$F_1 = 2 \frac{\text{Precisión} * \text{sensibilidad}}{\text{Precisión} + \text{sensibilidad}} = 2 \frac{1 * 1}{1 + 1} = 1$$

Prevalencia.

$$Prevalencia = \frac{TP + FN}{TP + FP + TN + FN} = \frac{88 + 0}{88 + 0 + 16 + 0} = 0.85$$

Resumen:

En una muestra de 104 hojas con el 85 % de hojas infectadas (prevalencia = 0.85) y el 15% de hojas sanas, el 100 % fueron predicha correctamente (exactitud = 1). El 100 % de las hojas infectadas fueron clasificadas correctamente (sensibilidad = 1) de igual forma el 100 % hojas sanas fueron clasificadas como infectadas (especificidad = 1). F_1 fue 1.

4.5.2.2 Validación de modelo SVM polinomial

La validación externa se llevó a cabo utilizando el modelo SVM polinómico entrenado para predecir un dataset de prueba con información de 32 imágenes de hojas, 16 hojas sanas y 16 infectadas. La probabilidad estimada por el modelo se muestra en la figura 4.21.

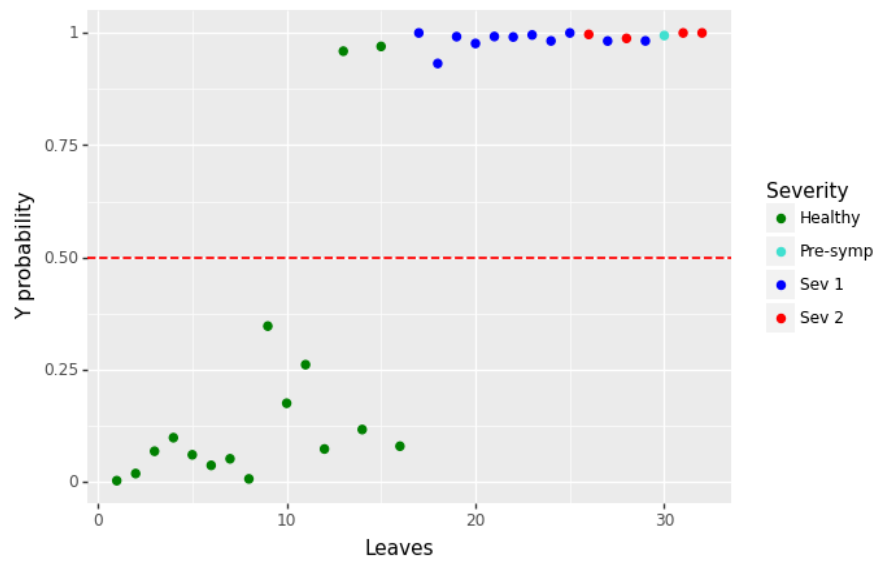


Figura 4.21 Probabilidad estimada para dataset de prueba con el modelo SVM polinomial.

Las métricas de predicción con datos de validación se describen a continuación.

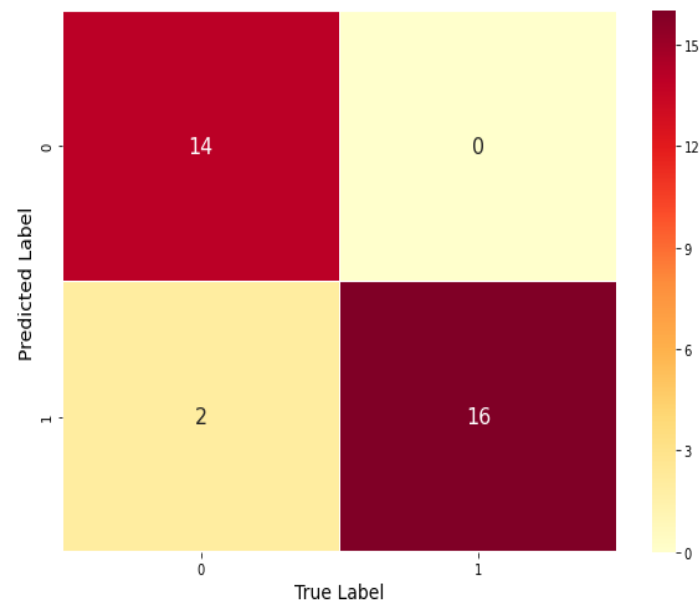


Figura 4.22 Matriz de confusión del modelo SVM con Kernel polinomial con el dataset de prueba.

Tabla 4-11 Evaluación de métricas de predicción de datos de prueba usando el modelo SVM polinomial

	Hojas Infectadas	Hojas no-infectadas	
Resultado Test	TP	FP	Precisión
	16	2	0.89
	FN	TN	Valor Pred. negativo
	0	14	1
	Sensibilidad	Especificidad	Exactitud
	1	0.88	0.94

Exactitud.

$$Exactitud = \frac{TP + TN}{TP + FP + TN + FN} = \frac{16 + 14}{16 + 2 + 14 + 0} = 0.94$$

Error de clasificación.

$$Error\ de\ clasificación = \frac{FP + FN}{TP + FP + TN + FN} = \frac{2 + 0}{14 + 2 + 16 + 0} = 0.0625$$

Sensibilidad.

$$Sensibilidad = \frac{TP}{TP + FN} = \frac{16}{16 + 0} = 1$$

Especificidad.

$$\text{Especificidad} = \frac{TN}{TN + FP} = \frac{14}{14 + 2} = 0.88$$

Precisión.

$$\text{Precisión} = \frac{TP}{TP + FP} = \frac{16}{16 + 2} = 0.89$$

Valor de predicción negativa.

$$\text{Valor de predicción negativa} = \frac{TN}{TN + FN} = \frac{14}{14 + 0} = 1$$

F₁ .

$$F_1 = 2 \frac{\text{Precisión} * \text{sensibilidad}}{\text{Precisión} + \text{sensibilidad}} = 2 \frac{0.889 * 1}{0.889 + 1} = 0.94$$

Prevalencia.

$$\text{Prevalencia} = \frac{TP + FN}{TP + FP + TN + FN} = \frac{16 + 0}{16 + 2 + 14 + 0} = 0.5$$

Área bajo la curva ROC (AUC). AUC fue igual a 0.94.

Resumen:

En una muestra de tamaño 32 conformada con un 50 % de hojas infectadas (prevalencia = 0.5) y el otro 50% de hojas sanas, el 94 % de las hojas (30) fue predicha correctamente por el modelo mientras que 2 fueron clasificadas en forma incorrecta (exactitud = 0.94). Todas las hojas clasificadas como infectadas fueron clasificadas correctamente (sensibilidad = 1) mientras que solo el 88 % hojas clasificadas como sanas fueron clasificadas correctamente (especificidad = 0.875). La precisión fue 0.89 y el AUC 0.94.

El modelo SVM lineal logró la separación completa de las hojas infectadas y no infectadas (exactitud = 1) en entrenamiento, mientras que la validación externa dio como resultado de predicción una exactitud de 0.94. En cuanto al modelo Polinomial, se obtuvo los mismos resultados, la exactitud con datos de entrenamiento fue del 1 y con datos de prueba el 0.94. Las evidencias obtenidas nos indican que el modelo lineal es el adecuado para la clasificación de las hojas sanas e infectadas puesto que fue configurado con un parámetro de regulación de 1. Adicionalmente, su simplicidad requiere menor carga computacional, por lo tanto, el modelo SVM lineal es el escogido para predecir el BLSD en plantas de banano.

4.6 REDES NEURONALES

En esta sección se exponen los resultados obtenidos en el proceso del entrenamiento y validación de los modelos de Perceptrón Multicapa (MLP) diseñados en la sección 3.6.7. La evaluación de la eficiencia predictiva de las redes neuronales se llevó a cabo utilizando matrices de confusión y las métricas de predicción. En cada prueba se realizó graficas de la probabilidad.

La matriz de datos de entrada de entrenamiento fue la matriz reducida con dimensiones 104 x 520. La salida es un vector que mantiene las etiquetas respecto a la presencia o ausencia de la enfermedad.

4.6.1 MLP con una capa oculta

La MLP 1 tiene una capa de entrada de 520 neuronas, una capa oculta con 22 neuronas con una función de activación ReLu y una capa de salida con una función sigmoide. El algoritmo de retropropagación utiliza la función de pérdida binary cross entropy. La estrategia de optimización seleccionada fue adam.

4.6.1.1 Entrenamiento de MLP con una capa oculta

El entrenamiento se ejecutó incrementando el número de épocas según la tabla 4-12. Los resultados de la exactitud tanto en entrenamiento como en validación externa determinan que la red logra su mayor eficiencia con 300 épocas de entrenamiento con una exactitud de entrenamiento igual a 1 y de validación externa de 0.94.

Tabla 4-12 Exactitud en entrenamiento y validación del modelo generado por la MLP con una capa oculta

Épocas	Validación cruzada	Validación externa
50	0.942	0.938
100	0.962	0.938
150	0.981	0.938
200	0.913	0.938
250	0.971	0.938
300	1	0.938
350	1	0.938
400	0.923	0.938

En las figuras 4.23 y 4.24 se muestra las curvas de la exactitud y de la función de pérdida.

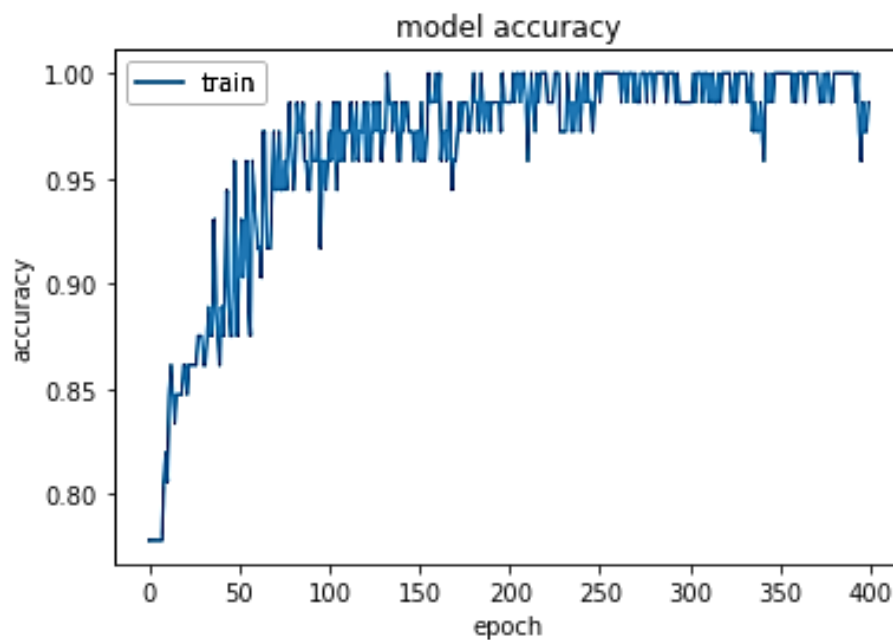


Figura 4.23 Curva de exactitud en entrenamiento de la MLP con una capa oculta.

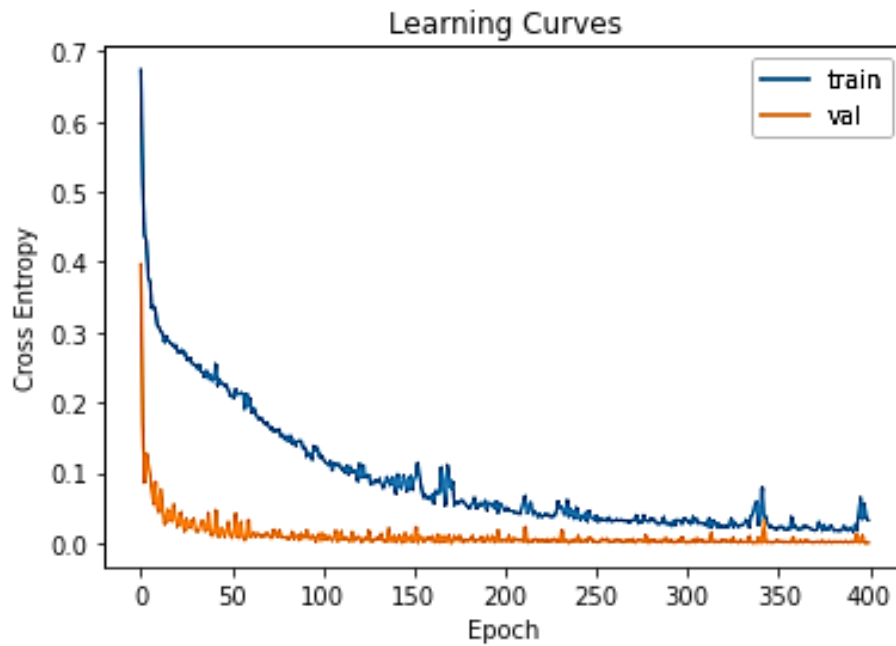


Figura 4.24 Curvas de aprendizaje de la MLP con una capa oculta.

El detalle de los resultados del modelo generado por esta red se presenta en las matrices de confusión y las tablas de métricas de predicción que se produjeron en las etapas de entrenamiento y validación. El gráfico de probabilidades estimadas a partir del dataset de entrenamiento se presenta en la figura 4.25.

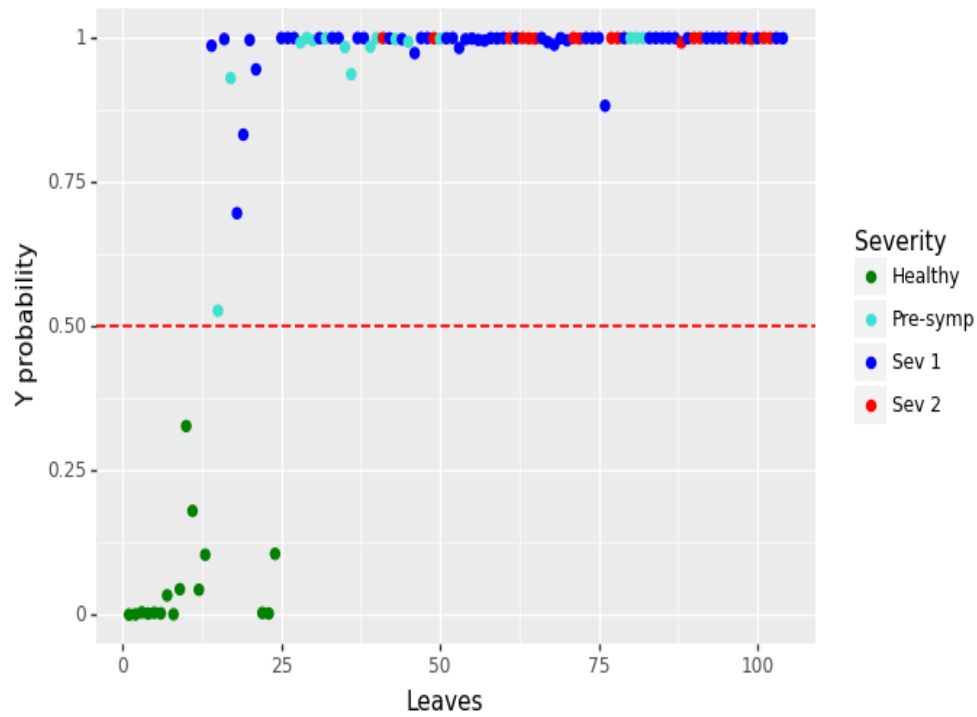


Figura 4.25 Probabilidad estimada por la MLP con una capa oculta en fase de entrenamiento.

Las métricas de predicción con la muestra de entrenamiento se muestran a continuación:

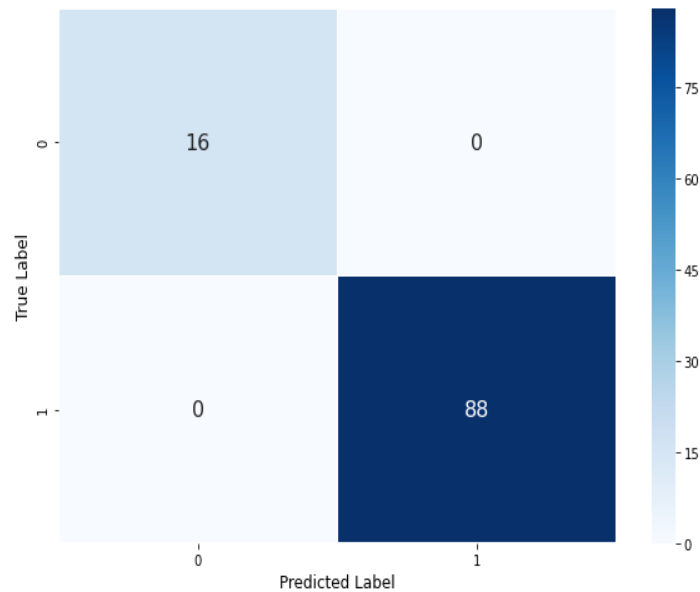


Figura 4.26 Matriz de confusión de predicción de datos de entrenamiento por la MLP con una capa oculta.

Tabla 4-13 Métricas de predicción de datos de prueba usando la MLP con una capa oculta

	Hojas Infectadas	Hojas no-infectadas	
Resultado Test	TP	FP	Precisión
	88	0	1
	FN	TN	Valor Pred. negativo
	0	16	1
	Sensibilidad	Especificidad	Exactitud
	1	1	1

Exactitud.

$$Exactitud = \frac{TP + TN}{TP + FP + TN + FN} = \frac{88 + 16}{88 + 0 + 16 + 0} = 1$$

Error de clasificación.

$$\text{Error de clasificación} = \frac{FP + FN}{TP + FP + TN + FN} = \frac{0 + 0}{88 + 0 + 16 + 0} = 0$$

Sensibilidad.

$$\text{Sensibilidad} = \frac{TP}{TP + FN} = \frac{88}{88 + 0} = 1$$

Especificidad.

$$\text{Especificidad} = \frac{TN}{TN + FP} = \frac{16}{16} = 1$$

Precisión.

$$\text{Precisión} = \frac{TP}{TP + FP} = \frac{88}{88 + 0} = 1$$

Valor de predicción negativa.

$$\text{Valor de predicción negativa} = \frac{TN}{TN + FN} = \frac{16}{16 + 0} = 1$$

Prevalencia.

$$\text{Prevalencia} = \frac{TP + FN}{TP + FP + TN + FN} = \frac{88 + 0}{88 + 0 + 16 + 0} = 0.85$$

F₁.

$$F_1 = 2 \frac{\textit{Precisi3n} * \textit{sensibilidad}}{\textit{Precisi3n} + \textit{sensibilidad}} = 2 \frac{1 * 1}{1 + 1} = 1$$

Resumen:

En una muestra conformada con un 85 % de hojas infectadas (prevalencia = 0.85), el 100 % de las hojas (88) fue predicha correctamente por lo tanto no hubo error en la predicci3n de ninguna hoja (exactitud = 1). El 100 % de las hojas infectadas fueron clasificadas correctamente (sensibilidad = 1) de la igual forma el 100 % hojas sanas fueron clasificadas como infectadas (especificidad = 1). La precisi3n fue 100%. El F₁ tambi3n fue 1.

4.6.1.2 Validaci3n de la MLP con una capa oculta

La validaci3n externa con la red de una capa oculta se llev3 a cabo utilizando el dataset de prueba con informaci3n de 32 im3genes de hojas, 16 hojas sanas y 16 infectadas. Las predicciones de probabilidad que estim3 el modelo se muestran en la figura 4.27.

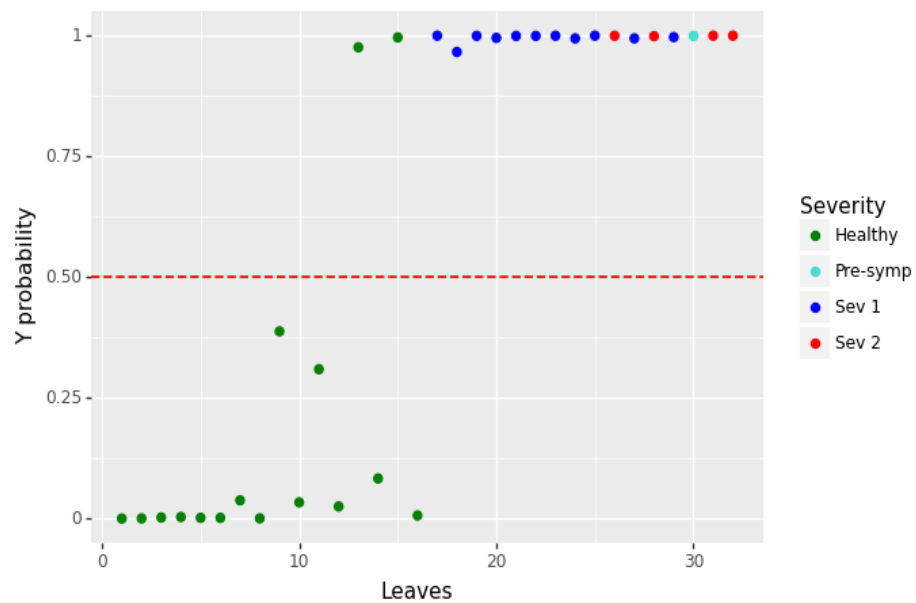


Figura 4.27 Probabilidad estimada por la MLP con una capa oculta en validación externa.

Los resultados de predicción se muestran en la figura 4.28 y la tabla 4-14.

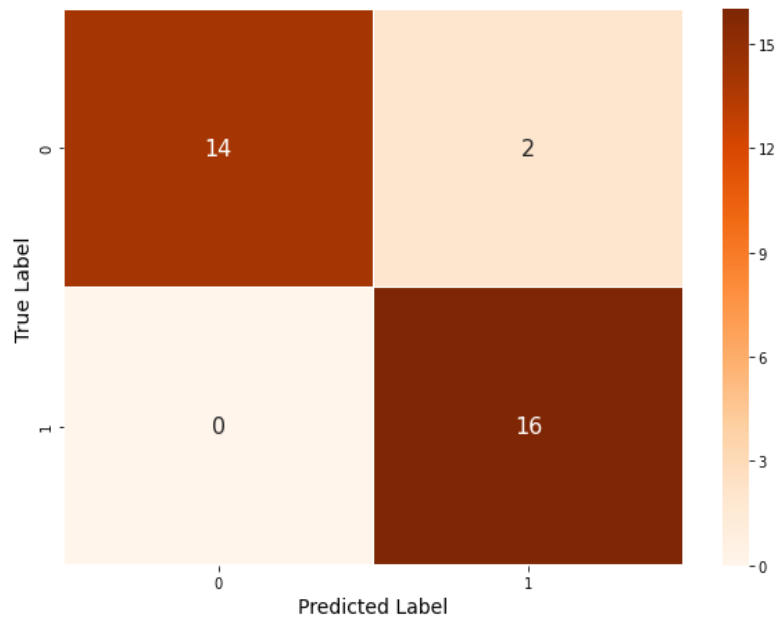


Figura 4.28 Matriz de confusión de red neuronal de una capa oculta en validación externa.

Tabla 4-14 Métricas de predicción de MLP con una capa oculta con datos de prueba

	Hojas Infectas	Hojas no-infectadas	
Resultado Test	TP	FP	Precisión
	16	2	0.89
	FN	TN	Valor Pred. negativo
	0	14	1
	Sensibilidad	Especificidad	Exactitud
	1	0.88	0.94

Exactitud.

$$Exactitud = \frac{TP + TN}{TP + FP + TN + FN} = \frac{16 + 14}{16 + 2 + 14 + 0} = 0.94$$

Error de clasificación.

$$\text{Error de clasificación} = \frac{FP + FN}{TP + FP + TN + FN} = \frac{2 + 0}{14 + 2 + 16 + 0} = 0.0625$$

Sensibilidad.

$$\text{Sensibilidad} = \frac{TP}{TP + FN} = \frac{16}{16 + 0} = 1$$

Especificidad.

$$\text{Especificidad} = \frac{TN}{TN + FP} = \frac{14}{14 + 2} = 0.88$$

Precisión.

$$\text{Precisión} = \frac{TP}{TP + FP} = \frac{16}{16 + 2} = 0.89$$

Valor de predicción negativa.

$$\text{Valor de predicción negativa} = \frac{TN}{TN + FN} = \frac{14}{14 + 0} = 1$$

Prevalencia.

$$\text{Prevalencia} = \frac{TP + FN}{TP + FP + TN + FN} = \frac{16 + 0}{16 + 2 + 14 + 0} = 0.5$$

F₁.

$$F_1 = 2 \frac{\text{Precisión} * \text{sensibilidad}}{\text{Precisión} + \text{sensibilidad}} = 2 \frac{0.889 * 1}{0.889 + 1} = 0.94$$

Área bajo la curva ROC. Su resultado fue 0.94.

Resumen:

En una muestra de 32 hojas con un 50 % de hojas infectadas (prevalencia = 0.5), el 94 % de las hojas (30) fue predicha correctamente por el modelo mientras que 2 hojas fueron clasificadas en forma incorrecta (exactitud = 0.937). Todas las hojas clasificadas como infectadas fueron clasificadas correctamente (sensibilidad = 1) de la igual forma el 88 % hojas clasificadas como sanas fueron clasificadas correctamente (especificidad = 0.875). La precisión fue 0.89. El F₁ y el AUC resultaron 0.94.

4.6.2 MLP con dos capas ocultas

La segunda red “MLP 2” tiene una capa de entrada de 520 neuronas, le siguen dos capas ocultas, la primera con 64 neuronas y la segunda con 8 neuronas, las dos con función de activación ReLu. La capa de salida con una neurona con función de activación sigmoide. El modelo se ajustó utilizando la función de pérdida binary cross entropy. La estrategia de optimización del gradiente que se aplicó fue adam.

4.6.2.1 Entrenamiento de MLP con dos capas ocultas

La MLP 2 fue entrenada en los mismos números de épocas que la red anterior alcanzando la máxima exactitud en validación con 350 épocas de entrenamiento tal como se muestra en la tabla 4-15.

Tabla 4-15 Exactitud en entrenamiento y validación del modelo generado por la MLP con dos capas ocultas

Épocas	Validación cruzada	Validación externa
50	0.904	0.938
100	0.904	0.938
150	0.962	0.938
200	0.981	0.938
250	0.981	0.938
300	0.981	0.938
350	1	0.938
400	1	0.938

De acuerdo con las tablas, los resultados son similares en las dos redes, la MLP 1 alcanza los mejores resultados con 300 épocas de entrenamiento y la MLP 2 lo hace en 350 épocas, por lo tanto, incrementar la complejidad de la red y la carga computacional incluyendo una capa oculta adicional no mejora los resultados.

En las figuras 4.29 y 4.30 se muestra las curvas de la exactitud y de la función de pérdida

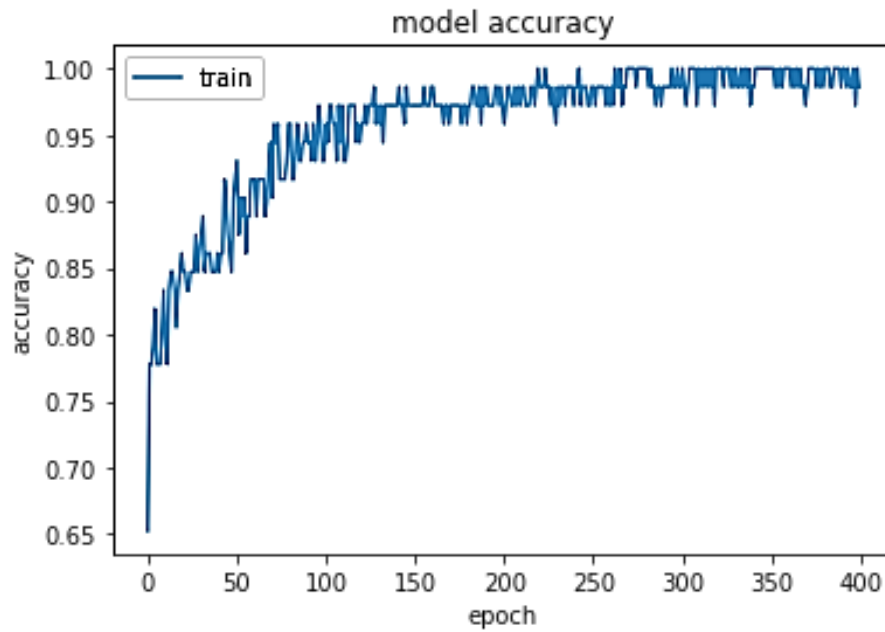


Figura 4.29 Curva de exactitud en entrenamiento de la MLP con dos capas ocultas.

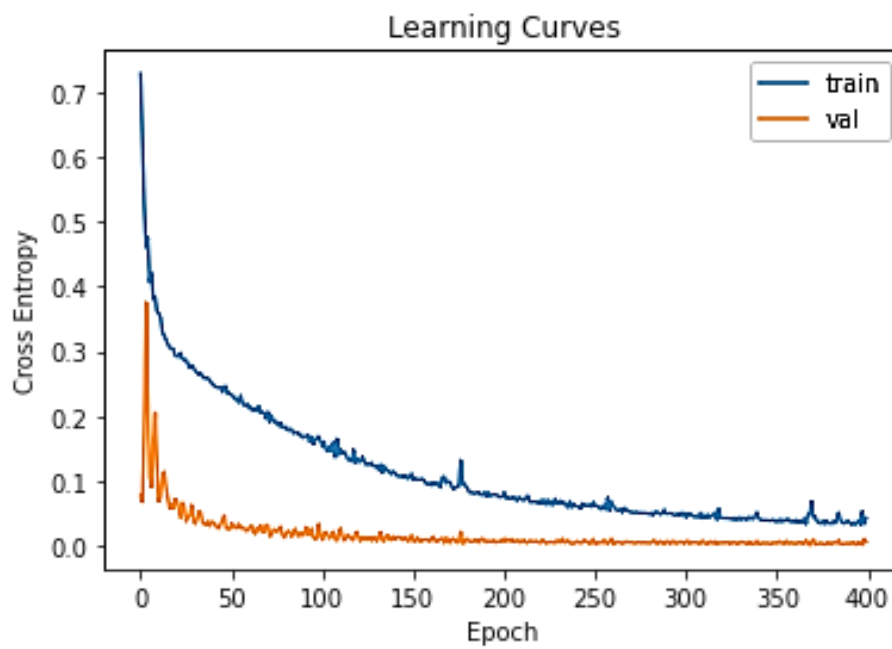


Figura 4.30 Curvas de aprendizaje de la MLP con dos capas ocultas.

El gráfico de probabilidades estimadas a partir del dataset de entrenamiento se presenta en la figura 4.31. El detalle de los resultados del modelo generado por esta red se presenta en las matrices de confusión y las tablas de métricas de predicción (figura 4.32 y tabla 4-16) que se produjeron en las etapas de entrenamiento y validación.

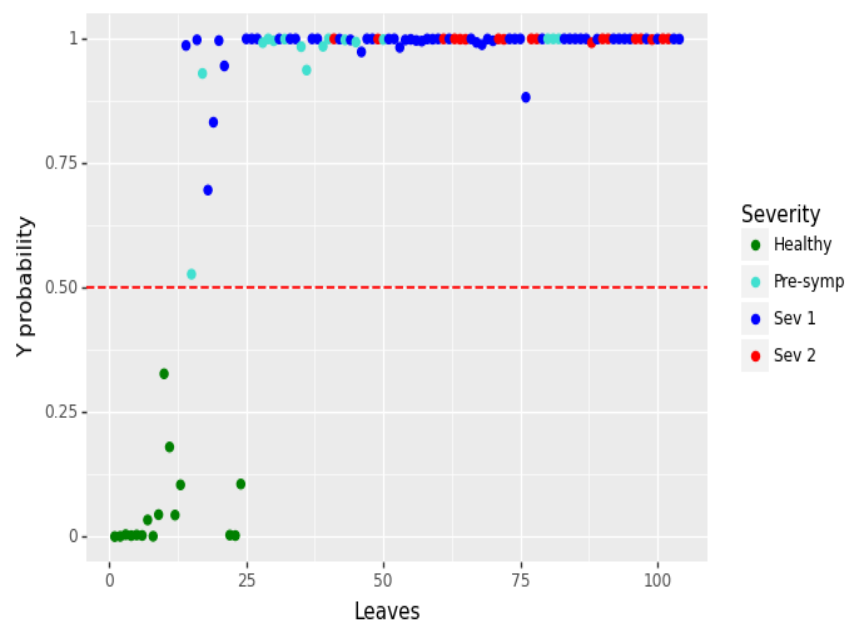


Figura 4.31 Probabilidad estimada por la MLP con dos capas ocultas en fase de entrenamiento.

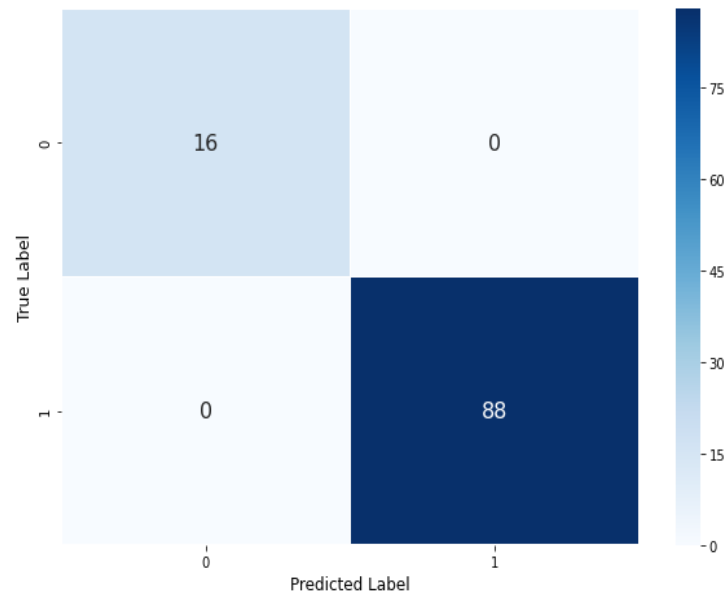


Figura 4.32 Matriz de confusión de predicción de datos de entrenamiento por la MLP con dos capas ocultas.

Las métricas de predicción se describen a continuación.

Tabla 4-16 Evaluación de métricas de predicción de datos de prueba usando la MPL con 2 capas ocultas

	Hojas Infectadas	Hojas no-infectadas	
Resultado Test	TP	FP	Precisión
	88	0	1
	FN	TN	Valor Pred. negativo
	0	16	1
	Sensibilidad	Especificidad	Exactitud
	1	1	1

Exactitud.

$$Exactitud = \frac{TP + TN}{TP + FP + TN + FN} = \frac{88 + 16}{88 + 0 + 16 + 0} = 1$$

Error de clasificación.

$$Error\ de\ clasificación = \frac{FP + FN}{TP + FP + TN + FN} = \frac{0 + 0}{88 + 0 + 16 + 0} = 0$$

Sensibilidad.

$$Sensibilidad = \frac{TP}{TP + FN} = \frac{88}{88 + 0} = 1$$

Especificidad.

$$Especificidad = \frac{TN}{TN + FP} = \frac{16}{16} = 1$$

Precisión.

$$Precisión = \frac{TP}{TP + FP} = \frac{88}{88 + 0} = 1$$

Valor de predicción negativa.

$$Valor\ de\ predicción\ negativa = \frac{TN}{TN + FN} = \frac{16}{16 + 0} = 1$$

F1.

$$F_1 = 2 \frac{\text{Precisión} * \text{sensibilidad}}{\text{Precisión} + \text{sensibilidad}} = 2 \frac{1 * 1}{1 + 1} = 1$$

Prevalencia.

$$\text{Prevalencia} = \frac{TP + FN}{TP + FP + TN + FN} = \frac{88 + 0}{88 + 0 + 16 + 0} = 0.85$$

Resumen:

En una muestra de 104 hojas, conformada con un 85 % de hojas infectadas (prevalencia = 0.85) y 15% de hojas sanas, el 100 % de las hojas fue predicha correctamente por lo tanto no hubo error en la predicción de ninguna hoja (exactitud = 1). El 100 % de las hojas infectadas fueron clasificadas correctamente (sensibilidad = 1) de la igual forma el 100 % hojas sanas fueron clasificadas como infectadas (especificidad = 1). La precisión fue 1.

4.6.2.2 Validación de MLP con dos capas ocultas

La validación externa con la MLP con dos capas ocultas se llevó a cabo con un dataset de prueba con información de 32 imágenes de hojas, 16 hojas sanas y 16 infectadas. Las predicciones de probabilidad que estimó el modelo se muestran en la figura 4.33.

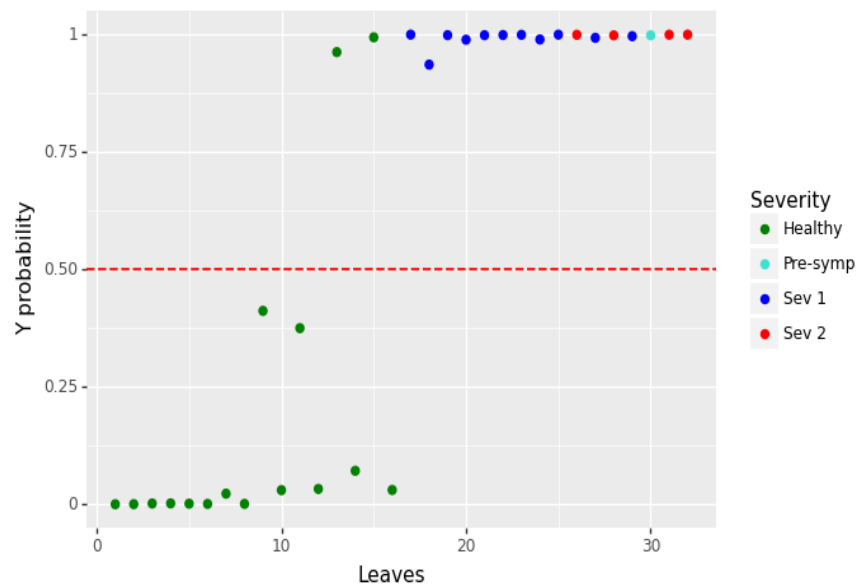


Figura 4.33 Probabilidad estimada por la MLP con dos capas ocultas en validación externa.

Los resultados de predicción se muestran en la figura 4.34 y la tabla 4-17.

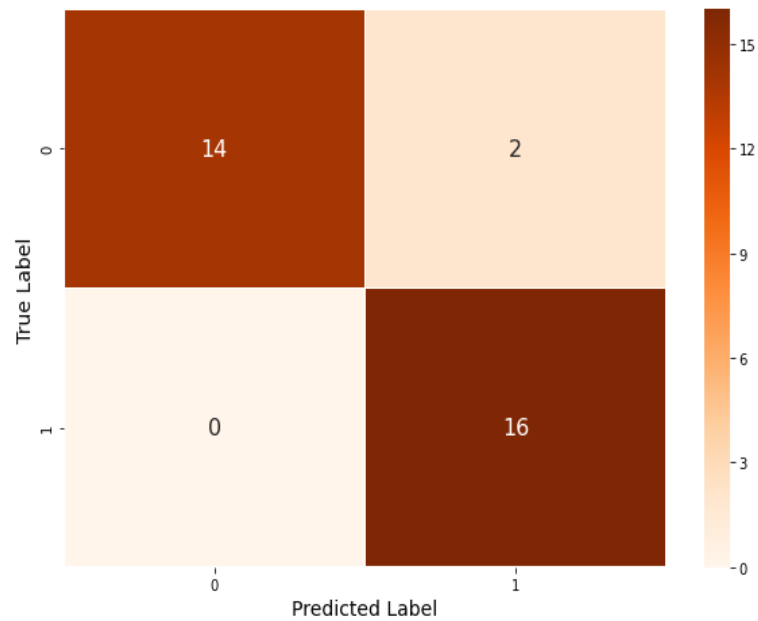


Figura 4.34 Matriz de confusión de MLP con dos capas ocultas en validación externa.

Tabla 4-17 Métricas de predicción de datos de prueba con MLP con dos capas ocultas en validación externa

	Hojas infectadas	Hojas no-infectadas	
Resultado Test	TP	FP	Precisión
	16	2	0.89
	FN	TN	Valor Pred. negativo
	0	14	1
	Sensibilidad	Especificidad	Exactitud
	1	0.88	0.94

Exactitud.

$$Exactitud = \frac{TP + TN}{TP + FP + TN + FN} = \frac{16 + 14}{16 + 2 + 14 + 0} = 0.94$$

Error de clasificación.

$$\text{Error de clasificación} = \frac{FP + FN}{TP + FP + TN + FN} = \frac{2 + 0}{14 + 2 + 16 + 0} = 0.0625$$

Sensibilidad.

$$\text{Sensibilidad} = \frac{TP}{TP + FN} = \frac{16}{16 + 0} = 1$$

Especificidad.

$$\text{Especificidad} = \frac{TN}{TN + FP} = \frac{14}{14 + 2} = 0.88$$

Precisión.

$$\text{Precisión} = \frac{TP}{TP + FP} = \frac{16}{16 + 2} = 0.89$$

Valor de predicción negativa.

$$\text{Valor de predicción negativa} = \frac{TN}{TN + FN} = \frac{14}{14 + 0} = 1$$

F₁.

$$F_1 = 2 \frac{\text{Precisión} * \text{sensibilidad}}{\text{Precisión} + \text{sensibilidad}} = 2 \frac{0.889 * 1}{0.889 + 1} = 0.94$$

Prevalencia.

$$Prevalencia = \frac{TP + FN}{TP + FP + TN + FN} = \frac{16 + 0}{16 + 2 + 14 + 0} = 0.5$$

Área bajo la curva ROC (AUC). El resultado de AUC fue 0.94 y se lo obtuvo mediante las métricas de scikit learn de Python.

Resumen:

En una muestra de 32 hojas conformada con un 50 % de hojas infectadas (prevalencia = 0.5) y el 50 % de hojas sanas, un 94 % de las hojas (30) fue predicha correctamente por el modelo mientras que 2 fueron clasificadas en forma incorrecta (exactitud = 0.94). Todas las hojas clasificadas como infectadas fueron clasificadas correctamente (sensibilidad = 1) de la igual forma el 88 % hojas clasificadas como sanas fueron clasificadas correctamente (especificidad = 0.88). La precisión fue 0.89, F_1 y el AUC 0.94.

4.7 ANÁLISIS COMPARATIVO

El enfoque propuesto en esta investigación es la aplicación de una metodología para la detección de la Sigatoka negra en plantas de banano basada en la construcción de un modelo PLS-PLR con alto poder predictivo en combinación con HS-Biplot para mejorar la interpretabilidad de los resultados y la estructura de los datos. En esta sección se realiza una comparación de los resultados del desempeño de los métodos propuestos junto con otros modelos que reportan alta capacidad de predicción y mantienen una elevada preferencia entre los métodos de aprendizaje automático supervisado para clasificación como son: NPLS-DA, SVM y MLP.

El análisis exploratorio de los datos mostró alta multicolinealidad entre las variables predictoras (longitudes de onda) para lo cual se utilizó la exploración de la matriz de correlación y el Factor de Inflación de la Varianza (VIF). Se realizó la prueba de normalidad mediante el método de Kolmogorov-Smirnov mejorado por Lilliefors que demostró que la mayoría no cumple con las condiciones de normalidad (377 de 520 longitudes de onda). Además, el análisis de los espectros promedio de las hojas infectadas y sanas, mostró que existen cambios positivos y negativos en diferentes rangos de longitudes de onda producidos por los cambios fisiológicos de la hoja durante la progresión de la enfermedad y son indicadores potenciales para la detección de BLSD.

Los resultados de predicción con los datos de entrenamiento se muestran en la tabla 4-18. La evaluación del rendimiento de los modelos de clasificación se realizó comparando las métricas de exactitud, precisión, sensibilidad, F_1 para cada modelo.

Tabla 4-18 Tabla comparativa de métricas de predicción en fase de entrenamiento

ENTRENAMIENTO				
Modelos	Exactitud	Precisión	Sensibilidad	F1
PLS-PLR *	0.98	0.98	1	0.99
NPLS-DA	0.9	1	0.88	0.94
SVM lineal	1	1	1	1
SVM polinómico	1	1	1	1
MLP 1 capa oculta	1	1	1	1
MLP 2 capas ocultas	1	1	1	1

* El modelo PLS-PLR fue evaluado utilizando el método de validación cruzada LOOCV.

Los modelos SVM y MLP clasificaron correctamente todos los datos de entrenamiento. Mientras que el modelo NPLS-DA no logró la separación de las clases. El modelo PLS-PLR, a diferencia de los otros modelos, fue evaluado con validación cruzada LOOCV con excelentes resultados. Es importante destacar que el modelo PLS-PLR también logró la separación completa de los datos de la misma forma que SVM y MLP, tal como se muestra en el HS-Biplot del dataset de entrenamiento (fig. 2.10).

Tabla 4-19 Tabla comparativa de métricas de predicción en fase de validación

VALIDACIÓN					
Modelos	Exactitud	Precisión	Sensibilidad	F1	AUC
PLS-PLR	0.94	0.94	0.94	0.94	0.94
NPLS-DA	0.91	0.88	0.94	0.91	0.91
SVM lineal	0.94	0.89	1	0.94	0.94
SVM polinómico	0.94	0.89	1	0.94	0.94
MLP 1 capa oculta	0.94	0.89	1	0.94	0.94
MLP 2 capas ocultas	0.94	0.89	1	0.94	0.94

Las pruebas de validación dieron resultados similares en todos los modelos, aunque en precisión tomó cierta ventaja el modelo PLS-PLR, la sensibilidad fue menor. El modelo NPLS-DA obtuvo las menores evaluaciones.

La métrica AUC (área bajo la curva ROC) es la más utilizada para comparar el rendimiento de modelos de clasificación utilizando la siguiente guía para interpretar los resultados:

Tabla 4-20 Capacidad discriminante según el valor AUC

Rango AUC	Capacidad discriminante
[0.5]	Sin capacidad discriminante
[0.5, 0.6)	Baja
[0.6, 0.75)	Regular
[0.75, 0.9)	Buena
[0.9, 0.97)	Muy buena
[0.97, 1)	Excelente

El AUC puede ser interpretado como la probabilidad de que un modelo clasifique un verdadero positivo aleatorio más alto que un falso positivo aleatorio. Los cuatro modelos se encuentran en el nivel de muy buena capacidad discriminante.

Un análisis de los individuos con errores nos permitirá evaluar la interpretabilidad del HS-Biplot. Para ello, se identificó las hojas que fueron erróneamente clasificadas en tres modelos que lograron los mayores indicadores de predicción, estos son: PLS-PLR, SVM

lineal, MLP 1 capa oculta. Los modelos SVM lineal y MLP 1 capa oculta fueron seleccionados por tener mayor simplicidad y la menor carga computacional (tabla 4-21).

Tabla 4-21 Errores de clasificación en prueba de validación externa

ERRORES				
Modelo	Número	Probabilidad	Predicción	Etiqueta
PLS-PLR	15	0.9999	Infectada	Sana
	20	0.006168	Sana	Infectada
SVM lineal	13	0.954854	Infectada	Sana
	15	0.975763	Infectada	Sana
MLP 1 capa o.	13	0.99319	Infectada	Sana
	15	0.99427	Infectada	Sana

Las hojas clasificadas erróneamente fueron la 13, 15 y 20. La tabla muestra coincidencia en las hojas mal clasificadas por los métodos SVM y MLP. El análisis de los espectros generados por los errores muestra que los 3 casos tienen características que los diferencian de los espectros promedios para cada nivel de la enfermedad. Se observa la coincidencia en las hojas erradas de los métodos SVM y MLP.

En el caso de la hoja 13 etiquetada como sana, fue clasificada como infectada por los modelos SVM y MLP con una probabilidad de 0.96 y 0.99 respectivamente. La figura 4.35 muestra el espectro de la hoja 13 en color negro, los espectros promedios de las hojas con niveles de infección presintomático, severidad 1 y 2 con colores amarillo, naranja y café respectivamente y el espectro de las hojas sanas con color verde. El espectro de la muestra 13 muestra que, en el rango de los colores amarillo, naranja y rojo (560 nm – 780 nm) tiene valores de reflectancia similares a los espectros de las hojas enfermas un poco

más bajos, mientras que en el rango del infrarrojo cercano ($> 780 \text{ nm}$) la reflectancia presenta valores similares al espectro de las hojas sanas.

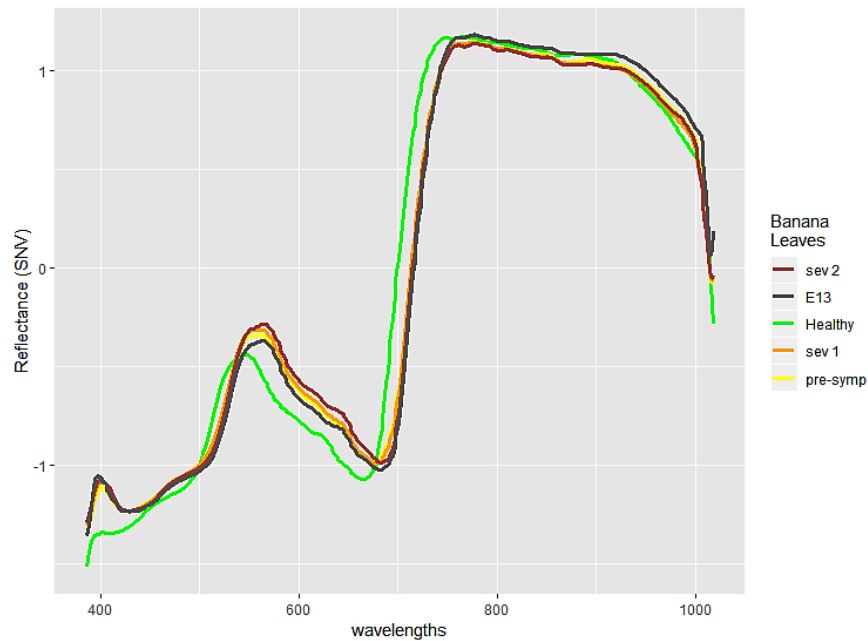


Figura 4.35 Espectro de hoja 13.

En el caso de la hoja 15 etiquetada como sana, fue clasificada como infectada por los 3 modelos, PLS-PLR, SVM y MLP, con una probabilidad de 0.99, 0.98, 0.99 respectivamente. En la figura 4.36 se muestra el espectro de la hoja 15 con color azul. Es muy similar al de la hoja 13 presentado en la figura anterior. En el rango de los colores amarillo, naranja y rojo ($560 \text{ nm} - 780 \text{ nm}$) tiene valores de reflectancia un poco más bajos pero similares a los espectros de las hojas enfermas, mientras que en el rango del infrarrojo cercano ($>780 \text{ nm}$) la reflectancia presenta valores similares al espectro de las hojas sanas.

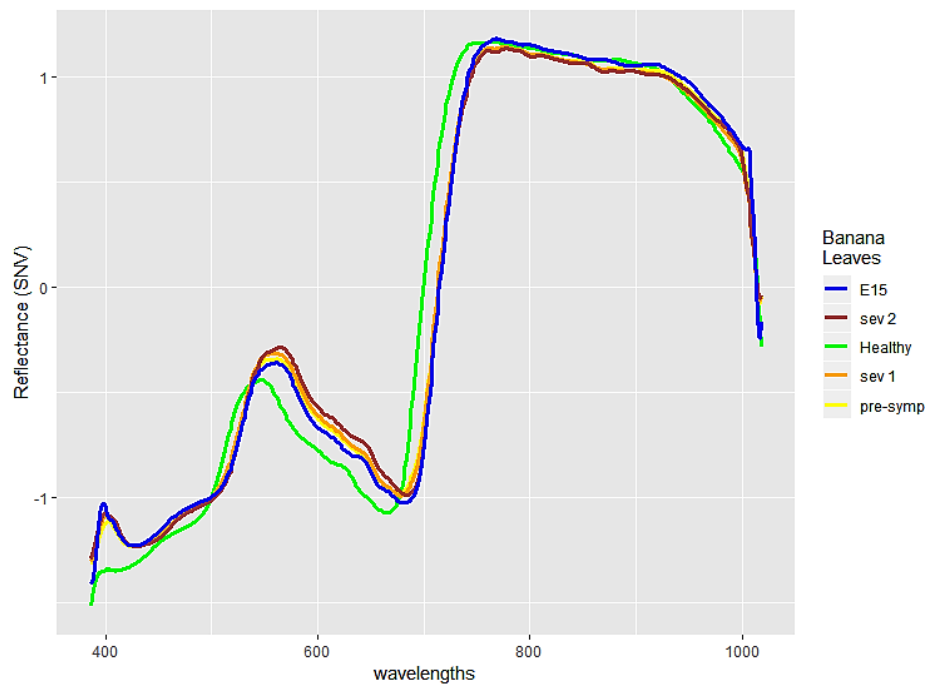


Figura 4.36 Espectro de hoja 15.

Finalmente, la hoja 20 que tenía etiqueta como infectada, fue clasificada como sana por el modelo PLS-PLR, con una probabilidad de 0.0062. El espectro de la hoja 20 se muestra en la figura 4.37 en color negro, aunque presenta algunas sutiles diferencias de los espectros de las hojas 13 y 15. En el rango de los colores amarillo, naranja y rojo (560 nm – 780 nm) tiene valores de reflectancia similares a los espectros de las hojas enfermas, pero más bajos que en los dos casos anteriores, es decir más cercanos al espectro de las hojas sanas, mientras que en el rango del infrarrojo cercano (> 780 nm) la reflectancia presenta valores inclusive superiores al espectro de las hojas sanas.

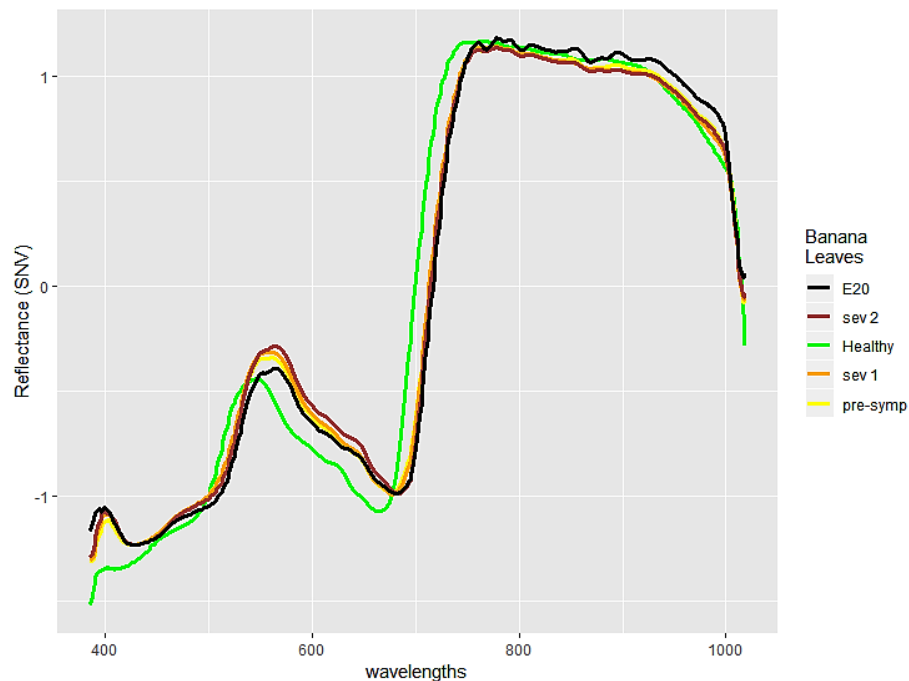


Figura 4.37 Espectro de hoja 20.

El HS-Biplot del dataset de validación nos muestra la posición de las hojas analizadas (figura 4.38). Las 3 hojas analizadas se encuentran en una posición cercana al grupo de hojas infectadas con una mayor influencia de las longitudes de onda del infrarrojo cercano (líneas color gris) que las ubica hacia la parte inferior del umbral de clasificación (0.5). De acuerdo con el análisis de los espectros realizado en los párrafos anteriores, las 3 hojas tenían un espectro similar al espectro de las hojas enfermas debido a esto, las hojas se muestran en el HS-Biplot cerca del grupo de las hojas infectadas. Por otro lado, también se pudo observar que el espectro en el rango infrarrojo cercano era similar al espectro de las hojas sanas, lo cual se muestra en el HS-Biplot como una relación más fuerte con las variables del rango del infrarrojo cercano que los acerca al umbral de clasificación.

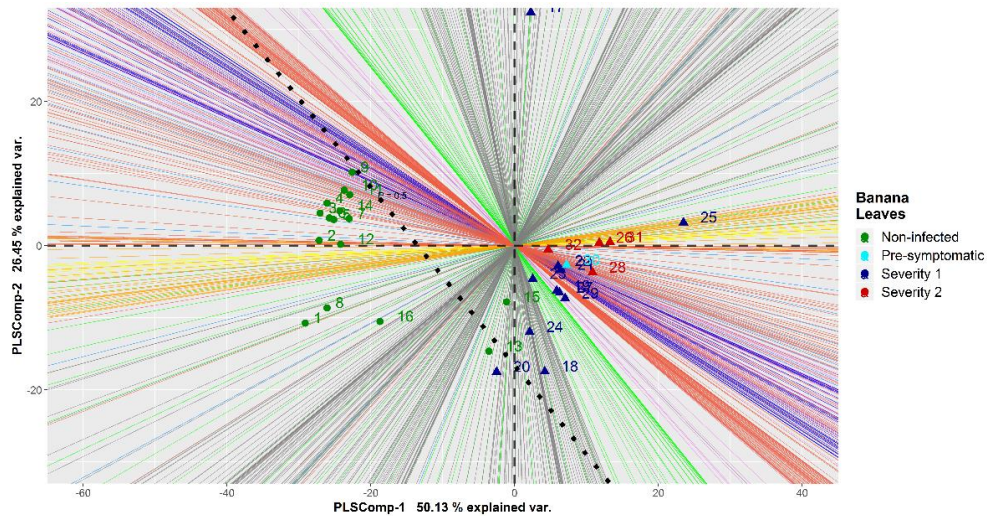


Figura 4.38 HS-Biplot de dataset de validación con hojas numeradas.

La hoja 20 presenta mayor influencia de las longitudes de ondas del rango infrarrojo y valores mas bajos que los obtenidos en las hojas 13 y 15 en los rangos del amarillo y naranja que lo acerca a las hojas sanas.

El análisis realizado con el HS-Biplot nos permitió evidenciar los cambios en los patrones de reflectancia que diferencian a las hojas con errores de los grupos de hojas sanas y enfermas.

En conclusión, los 4 métodos tienen altas capacidades de predicción y discriminación, aunque su principio inductivo es diferente. Los métodos NPLS-DA, SVM y MLP tienen un bajo poder interpretativo, mientras que el PLS-PLR es el único que, siendo complementado con el HS-Biplot, alcanza gran capacidad de interpretación de los resultados.

5 DISCUSIÓN

La agricultura es una actividad productiva que constantemente se ve sometida a ataques de plagas y enfermedades de las plantas que reducen la producción y la calidad. La escasez de herramientas de diagnóstico de última tecnología en los países subdesarrollados tiene un impacto devastador en su productividad y medio ambiente. Recientemente, la aplicación de nuevas estrategias y técnicas de agricultura para identificar enfermedades de las plantas se basan en métodos no destructivos que realizan la detección de la enfermedad en etapas iniciales permitiendo la gestión oportuna para evitar la propagación de la enfermedad y minimizar el efecto de los fungicidas en el medio ambiente. Con este propósito, en este trabajo presentamos una metodología de detección temprana de la Sigatoka negra en imágenes hiperespectrales basada en la aplicación de dos métodos multivariantes, PLS-PLR y HS-Biplot, que cumplen con los requerimientos intrínsecos de los datos hiperespectrales con desempeño confiable en predicción y alta capacidad interpretativa. La principal contribución es la implementación de PLS-PLR y HS-Biplot y su aplicación dentro de un marco metodológico para el análisis de imágenes hiperespectrales.

Las imágenes hiperespectrales de hojas de banano miden la energía reflejada de una fuente de luz y producen firmas espectrales únicas que pueden ser utilizadas para la detección de la Sigatoka negra y otras enfermedades. La calidad de las señales espectrales depende de varios factores como el sistema del sensor (lentes, espectrógrafo y detector de área), la escala de medición (hoja, planta o campo) y el análisis e interpretación de los datos.

La detección de enfermedades implica procesos como la adquisición de imágenes, el preprocesamiento de imágenes, la segmentación de imágenes, la extracción de

características y la clasificación. El sistema presentado combina tecnología hiperespectral con análisis avanzado de datos y métodos estadísticos para realizar la predicción de enfermedades de las plantas con elevada exactitud, a través de los cambios en la reflectancia que resultan de cambios fisiológicos característicos causados por una infección de patógenos. Actualmente está configurado como un sistema de imágenes estacionario y puede usarse en un laboratorio para el reconocimiento de enfermedades foliares en otras plantas y evaluaciones de calidad de los alimentos. El sistema se puede transformar, con pocos cambios, en un sistema aéreo de imágenes para escanear imágenes en campos de cultivo.

En este estudio, las imágenes hiperespectrales fueron utilizadas para clasificar las hojas de banano sanas e infectadas por BLSD debido a las marcadas diferencias en los espectros VIS y NIR de ambos grupos. Los resultados están de acuerdo con informes anteriores que muestran que los síntomas de la enfermedad pueden aumentar la reflectancia espectral en los rangos visibles (400-700 nm) y reducirla en infrarrojo cercano (700-1100 nm) (Ayala-Silva & Beyl, 2005). Los cambios en la reflectancia que ocurren durante las interacciones planta-patógenos se han asociado con alteraciones en la estructura de la hoja y la composición química del tejido durante la patogénesis, lo que se puede observar por la sucesión del tejido clorótico al necrótico (Mahlein, 2016).

El método propuesto PLS-PLR utiliza regresiones logísticas en lugar de lineales para correlacionar la variable de respuesta con los componentes PLS acorde con la respuesta binaria y es complementado con el HS-Biplot. Trabajos previos sobre detección de enfermedades de plantas han reportado una precisión y sensibilidad significativas de otros algoritmos de aprendizaje automático, pero la interpretabilidad de los modelos es baja y

en la mayoría de los estudios no incluyeron hojas en etapas pre-sintomáticas. En nuestro estudio, la aplicación de PLS-PLR mostró una exactitud del 98% en las etapas pre-sintomáticas y tempranas del BLSD utilizando validación cruzada (LOOCV).

HS-Biplot presenta las hojas como puntos y las longitudes de onda como líneas. En general, las hojas no infectadas se caracterizaron por la presencia prominente de longitudes de onda en el espectro visible, mientras que las hojas infectadas se asociaron a rangos visibles e infrarrojos cercanos, dependiendo de la etapa de la enfermedad o la proporción de tejido sintomático respecto al tejido sano. Los resultados están de acuerdo con estudios previos, en los que se ha reportado que los síntomas de la enfermedad en las plantas aumentan la reflectancia espectral tanto en el rango visible (400-780 nm) como en el infrarrojo cercano (780-1300 nm). Los cambios generales en la reflectancia que ocurren durante las interacciones entre plantas y patógenos se han asociado con alteraciones en la estructura de la hoja y cambios en la composición química del tejido durante la patogénesis, lo que se puede observar mediante la aparición de tejido clorótico y necrótico (Mahlein, 2016).

El agrupamiento de plantas no infectadas e infectadas se observó principalmente en la primera componente. Las longitudes de onda que más contribuyeron a la primera componente estuvieron en el rango de 577 nm a 651 nm (rango amarillo - rojo). Los cambios en el rango amarillo del espectro sugieren la detección de la clorosis de la planta que ocurre en las etapas iniciales de BLSD. La clorosis es causada por la insuficiencia de clorofila en las plantas, lo que lleva al amarillamiento de la hoja, generalmente medido con índices de amarillez (Adams, Philpot, & Norvell, 1999). De manera similar, los cambios en el rango naranja-rojo del espectro sugieren la presencia de tejido clorótico a

necrótico como se observa en las rayas rojas o marrones que aparecen en la etapa 2 del BLSD (Fouré, 1986). Curiosamente, las plantas pre-sintomáticas sin clorosis ni necrosis se agruparon aparte de las no infectadas, lo que sugiere cambios en la superficie de la hoja de las plantas infectadas. Las etapas pre-sintomáticas biotróficas generalmente no causan cambios observables en las hojas, pero algunos patógenos fúngicos pueden producir estructuras en la superficie de la hoja que pueden influir en las propiedades ópticas de la planta (Mahlein, 2016). Las plantas de banano no infectadas o infectadas formaron dos grupos en el HS-Biplot.

Como resultado de la prueba de validación, PLS-PLR y HS-Biplot demostraron su eficacia para realizar la clasificación y el análisis de los grupos de hojas con características similares. La conformación de los grupos fue similar a la presentada en la fase de entrenamiento mostrando que las longitudes de onda que mayor influencia tienen en la clasificación son las del rango visible y específicamente de los colores amarillo, naranja y rojo. Además, se obtuvo evidencia gráfica para explicar los errores presentados mostrando que las longitudes de onda del infrarrojo cercano (780-1300 nm) tuvieron mayor influencia en la clasificación errónea y que están vinculados a cambios en la estructura de la hoja. Este conocimiento puede ser utilizado por los investigadores para descubrir sus causas.

Estos resultados confirman la alta capacidad de predicción del modelo PLS-PLR y la eficiencia de HS-Biplot para representar las relaciones entre grupos de individuos (hojas de banano) y variables (longitudes de onda).

La comparación con los resultados obtenidos utilizando otros modelos mostró que el rendimiento de PLS-PLR está al nivel de otros métodos con alta capacidad predictiva como NPLS-DA, SVM y las redes neuronales (MLP) cuya evaluación se llevó a cabo calculando las métricas de predicción y el AUC (área bajo la curva ROC).

EL modelo NPLS-DA se basa en el algoritmo PLS1 y ha sido adaptado para la clasificación con datos de 3 vías (3 way) con respuesta binaria. Su implementación requiere de la construcción de un tensor que resume la información de cada imagen en cinco características estadísticas calculadas para cada longitud de onda. El tensor es desplegado en el primer modo y se aplicó un algoritmo recurrente para calcular una regresión lineal que estima la salida la cual, es transformada mediante una función logit que entrega la probabilidad de infección. Este método fue el de menor desempeño puesto que no logró separar todos los datos de entrenamiento, aunque los resultados de predicción en la prueba de validación fueron aceptables y similares a los otros métodos. Tiene un algoritmo muy rápido debido a los pocos parámetros que se deben calcular, lo que resulta en un bajo costo computacional. Uno de los inconvenientes que tiene este modelo es que está diseñado para respuestas continuas y su uso para predecir una respuesta binaria no incluye la estimación de una constante en la regresión, lo cual reduce su poder predictivo. El método ofrece buenos resultados para datos cuyas clases son linealmente separables.

La máquina de vectores soporte (SVM) es un método reconocido en el ámbito científico por su eficiencia en diferentes aplicaciones de clasificación y regresión que tiene algunas características que lo han puesto en ventaja respecto a otras técnicas. Su algoritmo se basa en el principio inductivo de minimización del riesgo estructural (SRM)

que busca un hiperplano, definido por los vectores soporte, que separa las clases y junto con el uso de una función kernel tiene una gran capacidad de generalización para casos con datasets de entrenamiento grandes o pequeños con clases separables o no. La solución del problema de programación cuadrática logra una solución única. Se entrenaron 2 modelos SVM, uno con kernel lineal y otro con kernel polinómico. Aplicando variaciones del parámetro de regularización, los dos modelos SVM lograron la separación completa de los datos de entrenamiento y en la prueba de validación presentaron 2 errores.

MLP utiliza un procedimiento de aprendizaje basado en la minimización de la función de pérdida mediante el procedimiento de retropropagación, lo que lo convierte en un aproximador universal, aunque no logra una solución única. Se desarrollaron dos modelos de Perceptrón multicapa (MLP), el primero con una capa oculta y el segundo con dos capas ocultas y fueron entrenados con el mismo set de datos que se utilizó para entrenar los modelos anteriores. El entrenamiento se inició utilizando un bajo número de etapas y se fue incrementando hasta alcanzar la máxima exactitud. Una vez entrenados los modelos fueron evaluados utilizando el dataset de prueba y como resultado se obtuvieron 2 errores.

Los errores de clasificación obtenidos en los modelos PLS-PLR, SVM y MLP fueron sometidos a un análisis de los espectros. El HS-Biplot mostró que las hojas con errores se ubicaron cerca del grupo de las hojas infectadas, pero con una fuerte influencia de las longitudes de onda del infrarrojo cercano (780 nm – 1300 nm), relacionada con cambios en la estructura de la hoja. Esto estuvo acorde con el resultado de la comparación de los espectros de los errores con los espectros promedio por nivel de enfermedad. Además, el

HS-Biplot mostró una relación directa entre los niveles de variación de la reflectancia con el acercamiento de las hojas mal clasificadas al grupo de hojas infectadas.

En la mayoría de los trabajos de investigación, el entendimiento y aprendizaje acerca del porqué se presentan ciertos resultados es crucial. La interpretabilidad permite aprovechar este conocimiento adicional capturado por el modelo. Utilizando métodos de clasificación interpretables se pueden explicar los resultados obtenidos principalmente cuando los modelos no tienen un impacto significativo en la toma de decisiones.

Las máquinas superan a los humanos en muchas tareas y tiene grandes ventajas en velocidad, reproducibilidad y escala. Una vez que se ha entrenado un modelo producirá los mismos resultados a partir de la misma entrada y se puede replicar con bajo costo. Una gran desventaja del uso del aprendizaje automático es que el problema de los datos y la forma como la maquina lo soluciona están ocultos dentro de modelos cada vez más complejos. Las redes neuronales y las máquinas de vectores soporte son un ejemplo de los modelos tipo caja negra que tiene alto desempeño en los que los modelos se convierten en fuente de conocimiento en lugar de los datos por lo tanto la interpretación es difícil o imposible de acuerdo con la complejidad. Esta desventaja tiene mayor impacto en aplicaciones de alta dimensionalidad como la tedetección. PLS-PLR y HS-Biplot permiten reducir o eliminar esta desventaja y ofrecen un alto poder predictivo y alta interpretabilidad para aplicaciones con datos hiperespectrales.

6 CONCLUSIONES

1. El vertiginoso ritmo de la innovación tecnológica a puesto a disposición de la humanidad nuevos métodos y tecnologías para enriquecer nuestro conocimiento acerca de la superficie del planeta. La búsqueda de la maximización del costo-beneficio dentro del campo ha motivado la incorporación de la teledetección a la agricultura de precisión. La utilización de sistemas de sensores de imágenes hiperespectrales permiten obtener información sobre la estructura fisiológica de las plantas sin tener contacto directo con ellas.
2. HSI proporciona un método no destructivo para analizar plantas infectadas con la Sigatoka negra y probablemente otros patógenos. El desarrollo de la tecnología ofrece una mayor capacidad de almacenamiento de datos, computadoras rápidas, detectores sensibles y diferentes técnicas analíticas para imágenes hiperespectrales que, junto con las técnicas estadísticas adecuadas, permiten la detección de enfermedades de las plantas incluso en las primeras etapas y demuestran que los cambios fisiológicos en las hojas se pueden capturar y modelar utilizando datos hiperespectrales.
3. La detección temprana de enfermedades infecciosas juega un papel crucial tanto en el tratamiento como en la aplicación de estrategias de prevención y control, por lo tanto, los métodos estadísticos con alto nivel de predicción e interpretabilidad utilizados para detectar la enfermedad en las primeras etapas, principalmente en la fase presintomática son herramientas promisorias para el desarrollo de la agricultura y el control del medio ambiente.

4. El proceso de preparación de las plantas requiere del seguimiento y control de especialistas en la rama biológica quienes deben realizar el monitoreo del proceso de inoculación y desarrollo de la enfermedad. El sistema HSI debe ser construido y calibrado tomando en cuenta las características específicas de la enfermedad y del cultivo para evitar deformaciones o presencia de ruido en las imágenes que afecten el estudio realizado a fin de obtener la mayor sensibilidad y especificidad.
5. La principal contribución de este trabajo es el desarrollo de una metodología para la detección temprana de la enfermedad del banano Sigatoka negra mediante la aplicación de PLS-PLR y HS-Biplot utilizando imágenes hiperespectrales. PLS-PLR y la representación gráfica HS-Biplot son técnicas prometedoras para analizar datos hiperespectrales, incluso teniendo en cuenta la alta reducción de la dimensionalidad después de la fase de pre-procesamiento de los datos.
6. Los metodos de aprendizaje automático comúnmente utilizados clasifican imágenes hiperespectrales sin incorporar información de la estructura de los datos. La eficiencia predictiva e interpretativa de los datos multidimensionales son factores de gran relevancia en los métodos para detectar y caracterizar enfermedades de las plantas. La aplicación del PLS-PLR proporciona un excelente poder predictivo que se complementa con el alto nivel de interpretación visual que ofrece el HS-Biplot, garantizando una adecuada clasificación y los insights para el entendimiento del comportamiento de los datos y sus relaciones con las variables.

7. Los resultados obtenidos de la aplicación de PLS-PLR confirman su alto nivel predictivo, similar al obtenido por técnicas de caja negra conocidas tales como SVM y las redes neuronales, pero toma ventaja en la interpretación tanto de los resultados como de la estructura de los datos cuando se utiliza el HS-Biplot. La reducción drástica de la dimensionalidad que ofrece el método PLS-PLR permite la reducción del tiempo y recursos de procesamiento y la visualización de los datos mediante el HS-Biplot.
8. La metodología aplicada para la detección de la Sigatoka negra en plantas de banano utilizando PLS-PLR y HS-Biplot puede ser transferida a otros cultivos con otras enfermedades para mejorar los sistemas de monitoreo de enfermedades de las plantas que ayude a los productores agrícolas a realizar una mejor evaluación y control de las enfermedades de las plantas. Esta tecnología contribuirá a optimizar el uso de los recursos naturales, mejorar la productividad y lograr altos estándares de calidad de los cultivos con el menor impacto en el medioambiente favoreciendo la sostenibilidad de la actividad agrícola.

7 BIBLIOGRAFÍA

- Aalderink, B. J., Klein, M. E., Padoan, R., De Bruin, G., & Steemers, T. (2010). Quantitative Hyperspectral Imaging Technique for Condition Assessment and Monitoring of Historical Documents. In *Poster presented at AIC's 38th Annual Meeting*.
- Abdi, H., & Molin, P. (2007). Lilliefors/Van Soest's test of normality. *Encyclopedia of measurement and statistics* (pp. 540–544).
- Adão, Telmo; Hruška, Jonáš; Pádua, Luís; Bessa, José; Peres, Emanuel; Morais, Raul; Sousa, Joaquim J. (2017). "Hyperspectral Imaging: A Review on UAV-Based Sensors, Data Processing and Applications for Agriculture and Forestry." *Remote Sens.* 9, no. 11: 1110.
- Adam, E., Mutanga, O., & Rugege, D. (2010). Multispectral and hyperspectral remote sensing for identification and mapping of wetland vegetation : a review. *Wetlands Ecol Manage*, 18(3), 281–296. <https://doi.org/10.1007/s11273-009-9169-z>
- Adams, M. L., Philpot, W. D., & Norvell, W. A. (1999). Yellowness index: An application of spectral second derivatives to estimate chlorosis of leaves in stressed vegetation. *International Journal of Remote Sensing*, 20(18), 3663–3675. <https://doi.org/10.1080/014311699211264>
- Adeboye, N. O., Fagoyinbo, I. S., & Olatayo, T. O. (2014). Estimation of the Effect of Multicollinearity on the Standard Error for Regression Coefficients. *Journal of Mathematics*, 10(4), 16–20. <https://doi.org/10.9790/5728-10411620>
- Aggarwal, C. C. (2018). *Neural Networks and Deep Learning*. Springer International Publishing AG. <https://doi.org/10.1007/978-3-319-94463-0>

- Albert, A., & Anderson, J. A. (1984). On the Existence of Maximum Likelihood Estimates in Logistic Regression Models. *Biometrika*, 71(1), 1–10.
<https://doi.org/https://doi.org/10.1093/biomet/73.3.755>
- Alciaturi, C. E., Escobar, M. E., De La Cruz, C., & Rincón, C. (2003). Partial least squares (PLS) regression and its application to coal analysis. *Revista Técnica de La Facultad de Ingeniería Universidad Del Zulia*, 26(3), 197–204.
- Alkan, B. B., Atakan, C., & Akdi, Y. (2015). Visual Analysis Using Biplot Techniques of Rainfall Changes over Turkey. *MAPAN-Journal of Metrology Society of India*, 30(1), 25–30. <https://doi.org/10.1007/s12647-014-0119-8>
- Allison, P. D. (2014). Measures of Fit for Logistic Regression. In *Proceedings of the SAS Global Forum 2014 Conference* (pp. 1–13).
- Anderson, R. E., & Swaminathan, S. (2011). Customer Satisfaction and Loyalty in E - Markets : A PLS Path Modeling Approach. *Journal of Marketing Theory and Practice*, 19(2), 221–234. <https://doi.org/10.2753/MTP1069-6679190207>
- Andersson, M. (2009). A comparison of nine PLS1 algorithms. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 23(April), 518–529.
<https://doi.org/10.1002/cem.1248>
- Anggraeni, A., & Lin, C. (2011). Application of SAM and SVM Techniques to Burned Area Detection for Landsat TM Images in Forests of South Sumatra. In *International Conference on Environmental Science and Technology*, 6 (January), (pp. V2160–V2164).

- Ashourloo, D., Mobasheri, M., and Huete, A. (2014). Evaluating the effect of different wheat rust disease symptoms on vegetation indices using hyperspectral measurements. *Remote Sensing*, 6(6):5107–5123.
- Ayala-Silva, T., & Beyl, C. A. (2005). Changes in spectral reflectance of wheat leaves in response to specific macronutrient deficiency. *Advances in Space Research*, 35(2), 305–317. <https://doi.org/10.1016/j.asr.2004.09.008>
- Bakache, A., Douzals, J.-P., Bonichelli, B., Cotteux, E., De Lapeyre de Bellaire, L., & Sinfort, C. (2019). Development of a rapid methodology for biological efficacy assessment in banana plantations: application to reduced dosages of contact fungicide for Black Leaf Streak Disease (BLSD) control. *Pest Management Science*, 75(4), 1081–1090. <https://doi.org/10.1002/ps.5219>
- Barker, M., & Rayens, W. (2003). Partial least squares for discrimination. *Journal of Chemometric: A Journal of the Chemometrics Society*, 17(3), 166–173. <https://doi.org/10.1002/cem.785>
- Bastien, P., Vinzi, V. E., & Tenenhaus, M. (2005). PLS generalised linear regression. *Computational Statistics and Data Analysis*, 48, 17–46. <https://doi.org/10.1016/j.csda.2004.02.005>
- Behmann, J., Steinrücken, J., & Plümer, L. (2014). Detection of early plant stress responses in hyperspectral images. *ISPRS JOURNAL OF PHOTOGRAMMETRY AND REMOTE SENSING*, 93, 98–111. <https://doi.org/10.1016/j.isprsjprs.2014.03.016>
- Bendini, Hugo & Jacon, Aline & Moreira Pessôa, Ana Carolina & Pompeu Pavanelli,

- João & Moraes, Wilson & Ponzoni, Flávio & Fonseca, Leila. (2015). Spectral characterization of banana leaves (*Musa* spp.) for detection and differentiation of black Sigatoka and yellow sigatoka. In *Anais XVII Simpósio Brasileiro de Sensoriamento Remoto* (pp. 2536–2543). João Pessoa - Brazil.
- Korkmazoglu, O. B., & Kemalbay, G. (2012). Econometrics application of partial least squares regression: an endogeneous growth model for Turkey. *Procedia-Social and Behavioral Sciences*, 62, 906-910. <https://doi.org/10.1016/j.sbspro.2012.09.153>
- Bock, C. H., Poole, G. H., Parker, P. E., & Gottwald, T. R. (2010). Plant disease severity estimated visually, by digital photography and image analysis, and by hyperspectral imaging. *Critical Reviews in Plant Sciences*, 29(2), 59–107. <https://doi.org/10.1080/07352681003617285>
- Bousset, L., Jumel, S., Picault, H., Domin, C., Lebreton, L., Ribulé, A., & Delourme, R. (2016). An easy, rapid and accurate method to quantify plant disease severity: application to phoma stem canker leaf spots. *European Journal of Plant Pathology*, 145(3), 697-709. <https://doi.org/10.1007/s10658-015-0739-z>
- Boyd, S., & Vandenberghe, L. (2018). *Introduction to Applied Linear Algebra Vectors , Matrices , and Least Squares*. Cambridge University Press. <https://doi.org/10.1017/9781108583664>
- Bro, R. (1996). Multiway calibration. multilinear pls. *Journal of Chemometrics*, 10(1), 47–61. [https://doi.org/https://doi.org/10.1002/\(SICI\)1099-128X\(199601\)10:1<47::AID-CEM400>3.0.CO;2-C](https://doi.org/10.1002/(SICI)1099-128X(199601)10:1<47::AID-CEM400>3.0.CO;2-C)
- Bro, R. (1998). Multi-way Analysis in the Food Industry: Models, Algorithms and

- Applications. *Multi-Way Analysis in the Food Industry: Models, Algorithms, and Applications*, (1998).
- Bro, R., Smilde, A. K., & de Jong, S. (2001). On the difference between low-rank and subspace approximation: improved model for multi-linear PLS regression. *Chemometrics and Intelligent Laboratory Systems*, 58(1), 3–13.
[https://doi.org/https://doi.org/10.1016/S0169-7439\(01\)00134-4](https://doi.org/10.1016/S0169-7439(01)00134-4)
- Burges, C. J. C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2(2), 121–167.
- Bzdok, D., Krzywinski, M., Altman, N., 2018. Machine learning: Supervised methods, SVM and kNN. Nature Publishing Group, págs.1-6. {Hal-01657491}
- Cárdenas, O., Galindo-Villardón, M. P. & Vicente-Villardón, J. L. (2007). Los métodos Biplot : Evolución y aplicaciones. *Revista Venezolana de Análisis de Coyuntura*, 13, 279–303. Retrieved from <http://redalyc.uaemex.mx>
- Cevallos-Cevallos, J. M., Jines, C., Maridueña-Zavala, M. G., Molina-Miranda, M. J., Ochoa, D. E., & Flores-Cedeno, J. A. (2018). GC-MS metabolite profiling for specific detection of dwarf somaclonal variation in banana plants. *Applications in Plant Sciences*, 6(11), e01194. <https://doi.org/10.1002/aps3.1194>
- Chaerle, L., Leinonen, I., Jones, H. G., & Van Der Straeten, D. (2007). Monitoring and screening plant populations with combined thermal and chlorophyll fluorescence imaging. *Journal of Experimental Botany*, 58(4), 773–784.
<https://doi.org/10.1093/jxb/erl257>

- Chang, C. I. (2016). *Real-Time Progressive Hyperspectral Image Processing*. Springer Berlin Heidelberg, New York, NY.
- Chuvieco, E. (1991). *FUNDAMENTOS DE TELEDETECCION ESPACIAL*. Estudios Geográficos, 52(203), 371.
- Cerda, J., & Cifuentes, L. (2012). Uso de curvas ROC en investigación clínica. Aspectos teórico-prácticos. *Revista Chilena de Infectología*, 29(2), 138–141. <https://doi.org/10.4067/S0716-10182012000200003>
- Cooil, B., Winer, R. S., & Rados, D. L. (1987). Cross-Validation for Prediction. *Journal of Marketing Research*, 24(3), 271–279. <https://doi.org/10.2307/3151637>
- Deisenroth, M. P., Faisal, A. A., & Ong, C. S. (2019). *MATHEMATICS FOR MACHINE LEARNING*. Cambridge University Press.
- Demey, J. R., Vicente-Villardón, J. L., Galindo-Villardón, M. P., & Zambrano, A. Y. (2008). Identifying molecular markers associated with classification of genotypes by External Logistic Biplots. *Bioinformatics*, 24(24), 2832–2838. <https://doi.org/10.1093/bioinformatics/btn552>
- Díaz, N. A., Ruiz, J. A. B., Reyes, E. F., Cejudo, A. G., Novo, J. J., Peinado, J. P., Meléndez-Valdés, F. T., Fiñana, I. T. (2010). Espectrofotometría : Espectros de absorción y cuantificación colorimétrica de biomoléculas. Universidad de Córdoba.
- Diezma, B., Lleó, L., Herrero, A., Lunadei, L., Roger, J. M., & Ruiz-Altisent, M. (2011). La imagen hiperespectral como herramienta de evaluación de la calidad de hortaliza de hoja mínimamente procesada. In *VI Congreso Ibérico de Agroindustria* (pp. 5–

7).

Dony, R. D. & Haykin, S. (1995). Neural Network Approaches to Image Compression.

In proceedings of the *IEEE*, 83(2). <https://doi.org/10.1109/5.364461>

Dunn, W. B., & Ellis, D. I. (2005). Metabolomics: Current analytical platforms and methodologies. *TrAC - Trends in Analytical Chemistry*, 24(4), 285–294.

<https://doi.org/10.1016/j.trac.2004.11.021>

Dyring, E. (1973). The Principles of Remote Sensing The Principles of Remote Sensing.

Jstor, 2(3), pp 57–69. Retrieved from <http://www.jstor.org/stable/25066564>

Elmasry, G., & Sun, D. (2010). Principles of Hyperspectral Imaging Technology. In *Hyperspectral Imaging for Food Quality Analysis and Control* (pp. 3–43). Academic Press.

<https://doi.org/10.1016/B978-0-12-374753-2.10001-2>

Fajardo Reina, Luis (2019). *Firmas Espectrales*.

FAO. (2017). FAOSTAT. Retrieved from <http://www.fao.org/faostat/en/#home>

Folch-Fortuny, A., Prats-Montalbán, J. M., Cubero, S., Blasco, J., & Ferrer, A. (2016).

VIS / NIR hyperspectral imaging and N-way PLS-DA models for detection of decay lesions in citrus fruits. *Chemometrics and Intelligent Laboratory Systems*, 156, 241–

248. <https://doi.org/10.1016/j.chemolab.2016.05.005>

Fouré, E. (1986). Varietal reactions of bananas and plantains to black leaf streak disease.

In G. Persley & E. Langhe De (Eds.), *Banana and Plantain Breeding Strategies: Proceedings of an International Workshop* (vol 21, pp. 110–113). Cairns (AUS): ACIAR.

- Frey, J. (2019). *Evaluating close range remote sensing techniques for the retention of biodiversity-related forest structures*. (Doctoral dissertation, Universität), Albert-Ludwigs-University. <https://doi.org/10.6094/UNIFR/151315>
- Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58(3), 453–467. <https://doi.org/10.1093/biomet/58.3.453>
- Gandhi, R. (2018). Towards Data Science: Support Vector Machine — Introduction to Machine Learning Algorithms. Retrieved from <https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47>
- Garro, B. A., Sossa, H., & Vazquez, R. A. (2012). Diseño automático de Redes Neuronales Artificiales mediante el uso del Algoritmo de Evolución Diferencial (ED). *Polibits*, (46), 13–27. <https://doi.org/10.17562/PB>
- Gbongue, L.-R., Lalaymia, I., Zeze, A., Delvaux, B., & Declerck, S. (2019). Increased Silicon Acquisition in Bananas Colonized by *Rhizophagus irregularis* MUCL 41833 Reduces the Incidence of *Pseudocercospora fijiensis*. *Frontiers in Plant Science*, 9, 1977. <https://doi.org/10.3389/fpls.2018.01977>
- Gebbers, R., & Adamchuk, V. I. (2010). Precision Agriculture and Food Security. *SCIENCE* 327, 5967, pp. 828(2010). <https://doi.org/10.1126/science.1183899>
- Godínez-Jaimes, F., Ramírez-Valverde, G., Reyes-Carretero, R., & Barrera-Rodríguez, E. (2012). Collinearity and separated data in the Logistic Regression Model. *Agrociencia (Montecillo)*, 46(4), 411–425.

- Govender, M., Chetty, K., & Bulcock, H. (2007). A review of hyperspectral remote sensing and its application in vegetation and water resource studies. *Water SA*, 33(2), 145–151. <https://doi.org/http://dx.doi.org/10.4314/wsa.v33i2.49049>
- Greenacre, M. (2008). La práctica del análisis de correspondencias. Fundacion BBVA.
- Greenacre, M. J. (2010). *Biplots in practics*. (R. Editorial, Ed.). Fundación BBVA.
- Gulli, A., & Pal, S. (2017). *Deep Learning with Keras*. Brimangham: PACKT PUBLISHING LTD.
- Han, L., Haleem, M. S., & Taylor, M. (2015). A Novel Computer Vision-based Approach to Automatic Detection and Severity Assessment of Crop Diseases. In *Science and Information Conference 2015 (SAI)* (pp. 638-644). IEEE. London. <https://doi.org/10.1109/SAI.2015.7237209>
- Heinze, G., & Schemper, M. (2002). A solution to the problem of separation in logistic regression, *Statistics in medicine*, 21(16), 2409–2419. <https://doi.org/10.1002/sim.1047>
- Hellberg, S., Sjoestroem, M., Skagerberg, B., & Wold, S. (1987). Peptide Quantitative Structure-Activity Relationships , a Multivariate Approach, 30(7), 1126–1135. <https://doi.org/10.1021/jm00390a003>
- Hernández-Sanchez, J. C. (2016). *BILOT LOGÍSTICO PARA DATOS NOMINALES Y ORDINALES*.
- Hernández-Sanchez, J. C., & Vicente-Villardón, J. L. (2017). logistic biplot for nominal data. *Advances in Data Analysis and Classification*, 11(2), 307–326. Retrieved from

<https://link.springer.com/article/10.1007/s11634-016-0249-7>

- Hidalgo, M., Tapia, A., Rodriguez, W., & Serrano, E. (2006). EFECTO DE LA SIGATOKA NEGRA (*Mycosphaerella fijiensis*) SOBRE LA FOTOSÍNTESIS Y TRANSPIRACIÓN FOLIAR DEL BANANO (*Musa* sp. AAA, cv. Valery). *Agronomía Costarricense*, 30(1), 35–41.
- Hosmer, D. W., Hosmer, T., Le Cessie, S., & Lemeshow, S. (1998). A comparison of goodness-of-fit tests for the logistic model. *Statistical Medicine*, 16(9), 965–980.
[https://doi.org/https://doi.org/10.1002/\(SICI\)1097-0258\(19970515\)16:9<965::AID-SIM509>3.0.CO;2-O](https://doi.org/https://doi.org/10.1002/(SICI)1097-0258(19970515)16:9<965::AID-SIM509>3.0.CO;2-O)
- Hosmer Jr, D., Lemeshow, S., & Sturdivant, R. (2013). *Applied Logistic Regression* (vol 398). (Thrid Edit). Jhon Wiley & Sons.
- Hu, M., Dong, Q., Malakar, P. K., Liu, B., & Jaganathan, G. K. (2015) Determining Banana Size Based on Computer Vision, *International Journal of Food Properties*, 18:3, 508-520, DOI: 10.1080/10942912.2013.833223
- Huang, W., Lamb, D. W., Niu, Z., Youngjiang, Z., Liu, L., & Wang, Æ. J. (2007). Identification of yellow rust in wheat using in-situ spectral reflectance measurements and airborne hyperspectral imaging. *Precision Agriculture*, (8), 187–197.
<https://doi.org/10.1007/s11119-007-9038-9>
- Hulland, J. (1999). USE OF PARTIAL LEAST SQUARES (PLS) IN STRATEGIC MANAGEMENT RESEARCH: A REVIEW OF FOUR RECENT STUDIES. *Strategic Management Journal*, 20(2), 195–204.

- Hunt, E. R., & Rock, B. N. (1989). Detection of Changes in Leaf Water Content Using Near and Middle-Infrared Reflectance. *Remote Sensing of Environment*, 30(1), 43–54. [https://doi.org/https://doi.org/10.1016/0034-4257\(89\)90046-1](https://doi.org/https://doi.org/10.1016/0034-4257(89)90046-1)
- IB, S., Antonio, R., & Almorox, J. A. (1999). Aplicación de sensores remotos en la detección y evaluación de plagas y enfermedades en la vegetación. *Teledetección. Avances y Aplicaciones.*, 64–67.
- Intaravanne, Y., Sumriddetchkajorn, S., and Nukeaw, J. (2012). Ripeness level indication of bananas with visible and fluorescent spectral images. In *2012 9th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology*, pages 1–4. IEEE.
- Jinzhu, L. U., Di, C. U. I., & Jiang, H. (2013). Discrimination of tomato yellow leaf curl disease using hyperspectral imaging. In *2013 Kansas City, Missouri, July 21-July 24, 2013* (p. 1). *American Society of Agricultural and Biological Engineers*. [https://doi.org/10.1016/s0889-8529\(03\)00006-9](https://doi.org/10.1016/s0889-8529(03)00006-9)
- Kerle, N., Jansen, L. L. F., & Huurnerman, G. C. (2004). *PRINCIPLES OF REMOTE SENSING* (Third edit). Enschede, The Netherlands: The International Institute for Geo-Information Science and Earth Observation (ITC).
- Kingma, D. P., & Ba, J. L. (2015). ADAM: A METHOD FOR STOCHASTIC OPTIMIZATION. In *ICLR 2015* (pp. 1–15). arXiv preprint.
- Kokaly, R. F., & Clark, R. N. (1999). Spectroscopic Determination of Leaf Biochemistry Using Band-Depth Analysis of Absorption Features and Stepwise Multiple Linear Regression. *Remote Sensing of Environment*, 67(98), 267–287.

[https://doi.org/https://doi.org/10.1016/S0034-4257\(98\)00084-4](https://doi.org/https://doi.org/10.1016/S0034-4257(98)00084-4)

- Kurz, T. H., & Buckley, S. J. (2016). A REVIEW OF HYPERSPECTRAL IMAGING IN CLOSE RANGE APPLICATIONS. In *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLI-B5, 2016 XXIII* (Vol. XLI, pp. 865–870). Prague, Czech Republic: XXIII ISPRS Congress. <https://doi.org/10.5194/isprsarchives-XLI-B5-865-2016>
- Kuska, M., Wahabzada, M., Leucker, M., Dehne, H., Kersting, K., Oerke, E. C., Steiner, U., & Mahlein, A. K. (2015). Hyperspectral phenotyping on the microscopic scale : towards automated characterization of plant-pathogen interactions. *Plant Methods*, 11(1), 28. <https://doi.org/10.1186/s13007-015-0073-7>
- Landgrebe, D. (2002). Hyperspectral image data analysis. *IEEE Signal Processing Magazine*, 19(1), 17–28. <https://doi.org/10.1109/79.974718>
- Lara, M., Diezma Iglesias, B., Lleó García, L., Roger, J.-M., Garrido, Y., Gil, M., and Ruiz-Altisent, M. (2013). Aplicación de imagen hiperespectral para observar el efecto de la salinidad en hojas de lechuga.
- Le Cessie, S., & Van Houwelingen, J. C. (1992). Ridge Estimators in Logistic Regression. *Journal of the Royal Statistical Society: Series C (applied Statistics)*, 41(1), 191–201. <https://doi.org/https://doi.org/10.2307/2347628>
- Levin, N. (1999). *Fundamentals of Remote Sensing. 1st Hydrographic Data Management Course, IMO—International Maritime Academy, Trieste*, 76.
- Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Botchis, D. (2018). Machine

- Learning in Agriculture : A Review. *Sensors*, 18(8), 2674.
<https://doi.org/10.3390/s18082674>
- Lotfi, M., Solimani, A., Dargazany, A., Afzal, H., & Bandarabadi, M. (2009). Combining Wavelet Transforms and Neural Networks for Image Classification. *IEEE Xplore Digital Library*, 44–48.
- Lowe, A., Harrison, N., & French, A. P. (2017). Hyperspectral image analysis techniques for the detection and classification of the early onset of plant disease and stress. *Plant Methods*, 13(1), 80. <https://doi.org/10.1186/s13007-017-0233-z>
- Lu, G., & Fei, B. (2014). Medical hyperspectral imaging: a review. *Journal of Biomedical Optics*, 19(1), 010901. <https://doi.org/10.1117/1.JBO.19.1.010901>
- Luna-Moreno, D., Sánchez-Álvarez, A., Islas-Flores, I., Canto-Canche, B., Carrillo-Pech, M., Villarreal-Chiu, J. F., & Rodríguez-Delgado, M. (2019). Early detection of the fungal banana black sigatoka pathogen *Pseudocercospora fijiensis* by an SPR immunosensor method. *Sensors*, 19(3), 465.
- Ma, J., Sun, D., Pu, H., Cheng, J. H., & Wei, Q. (2019). Advanced Techniques for Hyperspectral Imaging in the Food Industry : Principles and Recent Applications. *Annual Review of Food Science and Technology*, (10), 197–220.
<https://doi.org/https://doi.org/10.1146/annurev-food-032818-121155> Copyright
- Mahlein, A. K. (2011). *Detection, identification and quantification of fungal diseases of sugar beet leaves using imaging and non-imaging hyperspectral techniques*. Rheinischen Friedrich-Wilhelms-Universität Bonn. <https://doi.org/10.1.1.407.4184>

- Mahlein, A. K., Steiner, U., Hillnhütter, C., Dehne, H. W., & Oerke, E. C. (2012). Hyperspectral imaging for small-scale analysis of symptoms caused by different sugar beet diseases. *Plant Methods*, 8(1), 3. <https://doi.org/10.1186/1746-4811-8-3>
- Mahlein, A. K. (2016). Plant Disease Detection by Imaging Sensors – Parallels and Specific Demands for Precision Agriculture and Plant Phenotyping. *Plant Disease*, 100(2), 241–251. <https://doi.org/10.1094/PDIS-03-15-0340-FE>
- Maldonado, A. I. L., Fuentes, H. R., & Contreras, J. A. V. (2018). Introductory Chapter: Trends on Hyperspectral Imaging Development. In *Hyperspectral Imaging in Agriculture, Food and Environment*, 1, (pp. 2–7). <https://doi.org/10.5772/intechopen.70213>
- Marín, D. H., Romero, R. A., Guzmán, M., & Sutton, T. B. (2003). Black Sigatoka: An increasing threat to banana cultivation, 87(3), 208-222. <https://doi.org/https://doi.org/10.1094/PDIS.2003.87.3.208>
- Marten, G. C., Shenk, J. S., & Barton II, F. E. (1989). *Near Infrared Reflectance Spectroscopy (NIRS): Analysis of Forage* Agriculture handbook (Washington), 643. U.S. Department of Agriculture.
- Martínez Ávila, M., & Fierro Moreno, E. (2018). Aplicación de la técnica PLS-SEM en la gestión del conocimiento: un enfoque técnico práctico. *RIDE Revista Iberoamericana Para La Investigación y El Desarrollo Educativo*, 8(16), 130–164. <https://doi.org/10.23913/ride.v8i16.336>
- Mishra, P., Asaari, M. S. M., Shahrimie, M., Herrero-langreo, A., Lohumi, S., Diezma, B. & Scheunders, P. (2017). Close range hyperspectral imaging of plants : A review.

Biosystems Engineering, 164, 49–67.

<https://doi.org/10.1016/j.biosystemseng.2017.09.009>

Molnar, C. (2019). *Interpretable Machine Learning A Guide for Making Black Box Models Explainable*. Leanpub book. Retrieved from <https://www.lulu.com/>

Muriel, M. R. (2009). *Caracterización de imágenes hiperespectrales utilizando Support Vector Machines y técnicas de extracción de características. Proyecto fin de carrera*, UNIVERSIDAD DE EXTREMADURA.

Naji Al Najm, M. (2018). *The electromagnetic spectrum [Image]*. Retrieved from https://www.researchgate.net/publication/328702140_The_electromagnetic_spectrum

Nguyen, D. V, & Rocke, D. M. (2002). Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, 18(1), 39–50. <https://doi.org/https://doi.org/10.1093/bioinformatics/18.1.39>

Ochoa, D., Cevallos, J., Vargas, G., Criollo, R., Romero, D., Castro, R., & Bayona, O. (2016). Hyperspectral imaging system for disease scanning on banana plants. In *Sensing for Agriculture and Food Quality and Safety VIII*, 4(1), (Vol. 9864, p. 98640M). <https://doi.org/10.1117/12.2224242>

Ouertani, S. S., Mazerolles, G., Boccard, J., Rudaz, S., & Hanafi, M. (2014). Multi-way PLS for discrimination: Compact form equivalent to the tri-linear PLS2 procedure and its monotony convergence. *Chemometrics and Intelligent Laboratory Systems*, 133, 25–32. <https://doi.org/10.1016/j.chemolab.2014.01.015>

- Oyedele, O. F., & Lubbe, S. (2015). The Construction of a Partial Least Squares Biplot Opeoluwa. *Journal of Applied Statistics*, 42(11), 2449–2460. <https://doi.org/https://doi.org/10.1080/02664763.2015.1043858>
- Patiño, L. F., Bustamante, E., & Salazar, L. M. (2007). Efecto de Sustratos Foliares Sobre la Sigatoka Negra (*Mycosphaerella fijiensis* Morelet) en Banano (*Musa × paradisiaca*L.) y Plátano (*Musa acuminata* Colla). *Agricultura Técnica*, 67(4), 437–445. <https://doi.org/10.4067/S0365-28072007000400012>
- Patterson, J., & Gibson, A. (2017). *Deep Learning: A practitioner's approach*. "O'Reilly Media Inc".
- Paul, R. K. (2006). MULTICOLLINEARITY : CAUSES , EFFECTS AND REMEDIES. *IASRI, New Delhi*, 1(1), 58-65.
- Perera, N. A. T. T., Kelaniyangoda, D. B., & Salgadoe, A. S. A. (2013). Leaf Spot Diseases in Banana (*Musa ssp.*) and their control (in vitro). In *ISAE 2013. Proceedings of the International Symposium on Agriculture and Environment 2013, 28 November 2013, University of Ruhuna, Sri Lanka* (pp. 287-290). Faculty of Agriculture, University of Ruhuna.
- Pérez-Enciso, M., & Tenenhaus, M. (2003). Prediction of clinical outcome with microarray data: A partial least squares discriminant analysis (PLS-DA) approach. *Human Genetics*, 112(5-6), 581–592. <https://doi.org/10.1007/s00439-003-0921-9>
- Plaza, A, Benediktsson, J. A., Boardman, J., Brazile, J., & Bruzzone, L., Camps-Valls, G., Chanussot, J., Fauvel, M., Gamba, P., Gualtieri, A., Tilton, J. C., Trianni, G. (2006). Advanced Processing of Hyperspectral Images. In 2006 *IEEE International*

- Symposio* (pp. 1974–1978). IEEE. <https://doi.org/10.1109 / IGARSS.2006.511>
- Porcel, M. (2001). *TÉCNICAS QUIMIOMÉTRICAS PARA EL DESARROLLO DE NUEVOS MÉTODOS CINÉTICO-ESPECTROFOTOMÉTRICOS DE ANÁLISIS*. (Tesis Doctoral) Universidad Autónoma de Barcelona. España.
- Richards, J. A. (2013). *Remote sensing digital image analysis: An introduction*. (Vol. 9783642300). <https://doi.org/10.1007/978-3-642-30062-2>
- Rodríguez, Y. E. T., Ones, V. G., Sánchez García, J. E. S., & Velar, R. C. (2014). Utilización combinada de métodos exploratorios y confirmatorios para el análisis de la actividad antibacteriana de la cefalosporina (PARTE II). *Investigacion Operacional*, 33(2), 114–120.
- Roman-Gonzalez, A. & Vargas-Cuentas, N. I. (2013). Análisis de imágenes hiperespectrales. *Revista Ingenieria & Desarrollo*, 2013, Año 9(N 35), pp.14-17. Retrieved from <https://hal.archives-ouvertes.fr/hal-00935014>
- Rumpf, T., Mahlein, A. K., Steiner, U., Oerke, E. C., Dehne, H. W., & 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. <https://doi.org/10.1016/j.compag.2010.06.009>
- Sampson, P. D., Streissguth, A. P., Barr, H. M., & Bookstein, F. L. (1989). Neurobehavioral effects of prenatal alcohol: Part II. Partial Least Squares analysis. *Neurotoxicology and Teratology*, 11(5), 477–491. [https://doi.org/http://dx.doi.org/10.1016/0892-0362\(89\)90025-1](https://doi.org/http://dx.doi.org/10.1016/0892-0362(89)90025-1)

- Santner, T. J., & Duffy, D. E. (1986). A note on A . Albert and J . A . Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 73(3), 755–758.
<https://doi.org/https://doi.org/10.1093/biomet/73.3.755>
- Siche, R., Vejarano, R., Aredo, V., Velasquez, L., Saldaña, E., and Quevedo, R. (2016). Evaluation of food quality and safety with hyperspectral imaging (hsi). *Food Engineering Reviews*, 8(3):306–322.
- Slaton, M. R., Raymond Hunt Jr., E., & Smith, W. K. (2001). ESTIMATING NEAR - INFRARED LEAF REFLECTANCE FROM LEAF STRUCTURAL CHARACTERISTICS. *American Journal of Botany*, 88(2), 278–284.
- Smilde, A., Bro, R., & Geladi, P. (2005). *Multi-way Analysis: Applications in the Chemical Sciences*. John Wiley & Sons, Ltd.
- Stover, R. H. (1980). Sigatoka leaf spots of bananas and plantains. *Plant Disease*. La Lima Honduras: American Phytopathological Society.
- Takane, Y., & Loisel, S. (2014). On the PLS Algorithm for Multiple Regression (PLS1). In *The Multiple Facets of Partial Least Squares and Related Methods* (pp. 17–29).
- Tan, S. Y. (2017). Developments in Hyperspectral Sensing. In *Handbook of Satellite Applications* (pp. 1137–1157). https://doi.org/10.1007/978-3-319-23386-4_101
- Torres, J., (2018). *Deep Learning introducción práctica con Keras* (3rd ed.). Barcelona: Kindle direct publishing.
- Vapnik, V., & Vapnik, V. (1998). *Statistical learning theory*. John Wiley & Sons. New

York, 1, 624.

Vega Vilca, J. C., & Guzmán, J. (2011). Regresión pls y pca como solución al problema de multicolinealidad en regresión múltiple. *Revista de Matemática: Teoría y Aplicaciones*, 18(1), 2011, 9–20.

Vicente-Villardón, J. L., Galindo-Villardón, M. P., & Blázquez-Zaballos, A. (2006). Logistic Biplots. In *Multiple Correspondence Analysis and Related Methods*. London. Chapman & Hall, 503 - 521. <https://doi.org/10.1201/9781420011319.ch23>

Vo-Dihn, T. (2004). A Hyperspectral Imaging System for In Vivo Optical Diagnostics. In *IEEE Engineering in Medicine and Biology Magazine*, 23(5), 40–49. <https://doi.org/10.1109/MEMB.2004.1360407>

Walker, D. A., & Smith, T. J. (2016). Nine Pseudo R2 Indices for Binary Logistic Regression Models. *Journal of Modern Applied Statistical Methods*, 15(1), 848–854. <https://doi.org/10.22237/jmasm/1462078200>

Wold, H. (1975). Soft Modelling by Latent Variables : The Non-Linear Iterative Partial Least Squares (NIPALS) Approach. *Journal of Applied Probability*, 12(S1), 117–142.

Wold, S., Ruhe, H., & Wold, H. (1984). III, WD, The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses. *Journal of Scientific and Statistical Computations*, 5(3), 735–743. <https://doi.org/10.1137/0905052>

Wold, S., Sjöström, M., & Eriksson L. (2001). PLS-regression : a basic tool of

- chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58, 109–130.
[https://doi.org/10.1016/S0169-7439\(01\)00155-1](https://doi.org/10.1016/S0169-7439(01)00155-1)
- Wu, W. W. (2010). Linking Bayesian networks and PLS path modeling for causal analysis. *Expert Systems With Applications*, 37(1), 134–139.
<https://doi.org/10.1016/j.eswa.2009.05.021>
- Yamazaki, F. and Wen, L., (2016). *REMOTE SENSING TECHNOLOGIES FOR POST-EARTHQUAKE DAMAGE ASSESSMENT: A CASE STUDY ON THE 2016 KUMAMOTO EARTHQUAKE*.
- Yeturu, S., Mendez, K., Garrido, P., Serrano, S., & Garrido, A. (2016). Serological and molecular identification of cucumber mosaic virus (CMV) infecting banana crops in Ecuador. *ECUADOR ES CALIDAD - Revista Científica Ecuatoriana*, 3, 17–22.
- Yijie, W. A. N. G., & CHENG, J. (2018). Rapid and Non-destructive Prediction of Protein Content in Peanut Varieties Using Near-infrared Hyperspectral Imaging Method. *Grain & Oil Science and Technology*, 1(1), 40-43.
- Zhu, H., Chu, B., Zhang, C., Liu, F., Jiang, L., & He, Y. (2017). Hyperspectral Imaging for Presymptomatic Detection of Tobacco Disease with Successive Projections Algorithm and Machine-learning Classifiers. *Sci Rep* 7, 4125.
<https://doi.org/10.1038/s41598-017-04501-2>

APÉNDICE A

CONTRIBUCIONES RELATIVAS DE LOS FACTORES PLS-PLR A LAS VARIABLES (LONGITUDES DE ONDA).



Long onda	t1	t2	Long onda	t1	t2	Long onda	t1	t2	Long onda	t1	t2
386	49.73	20.13	450	23.03	3.66	490	55.51	0.36	543	6.55	72.1
388	45.29	21.59	451	22.4	3.64	491	51.4	0.02	544	10.79	67.74
389	41.85	20.99	452	22.53	3.71	493	45.88	0.16	545	15.42	62.97
390	41.96	19.89	453	23.17	3.79	494	38.35	1.15	546	20.33	57.89
391	45.05	18.6	455	24.42	3.84	495	28.53	3.6	547	25.38	52.75
392	49.52	16.97	456	26.5	3.98	496	16.76	8.49	549	30.4	47.77
393	54.34	15.21	457	29.02	4.03	497	5.68	16.53	550	35.27	43.13
395	57.74	13.97	458	31.83	4.05	498	0.11	26.46	551	39.95	38.89
396	60.11	12.91	459	34.83	4.21	500	2.83	35.17	552	44.46	34.95
397	63.63	11.13	461	37.9	4.47	501	11.25	40.41	553	48.92	31.15
398	65.85	10.16	462	41.04	4.75	502	21.05	42.38	555	53.5	27.27
399	67.23	9.03	463	44.16	4.97	503	29.87	42.31	556	58.3	23.2
400	68.29	8	464	47.03	5.01	504	37.04	41.21	557	63.31	19.01
402	68.59	7.31	465	49.25	4.91	506	42.76	39.71	558	68.35	14.87
403	68.88	6.6	466	51.1	4.74	507	47.26	38.17	559	73.09	11.12
404	69.71	5.8	468	52.75	4.55	508	50.75	36.87	561	77.24	7.96
405	70.53	5.54	469	53.94	4.34	509	53.4	35.89	562	80.68	5.49
406	70.98	5.42	470	55.15	4.24	510	55.43	35.2	563	83.39	3.68
407	71.25	5.18	471	56.19	4.32	512	56.94	34.81	564	85.47	2.41
409	71.96	4.99	472	57.21	4.48	513	57.99	34.66	565	87.03	1.58
410	72.35	4.81	474	58.33	4.61	514	58.72	34.63	567	88.18	1.05
411	72.22	4.45	475	59.45	4.63	515	59.21	34.63	568	89.03	0.72
412	72.35	4.24	476	60.21	4.54	516	59.45	34.7	569	89.67	0.53
413	71.72	4.55	477	60.62	4.33	518	59.5	34.79	570	90.18	0.42
415	70.94	4.64	478	60.89	4.02	519	59.34	34.98	571	90.61	0.36
416	69.79	4.9	479	61.05	3.69	520	58.96	35.33	573	91	0.31
417	68.58	5.36	481	61.19	3.42	521	58.33	35.9	574	91.38	0.27
418	67.34	5.92	482	61.38	3.18	522	57.36	36.83	575	91.74	0.23
419	66.12	6.46	483	61.61	3	523	56.02	38.17	576	92.06	0.2
420	65.19	6.9	484	61.88	2.83	525	54.24	39.99	577	92.32	0.18
422	64.36	7	485	61.97	2.56	526	51.98	42.3	579	92.5	0.18
423	63.3	7.19	487	61.59	2.12	527	49.14	45.1	580	92.61	0.21
424	62.75	7.63	488	60.5	1.56	528	45.66	48.38	581	92.66	0.26
425	62.13	7.87	489	58.51	0.92	529	41.48	52.09	582	92.66	0.35
426	61.45	7.92	438	46.54	7.96	531	36.52	56.24	583	92.62	0.46
427	60.43	8.04	439	43.86	7.57	532	30.76	60.79	585	92.54	0.57
429	59.05	8.04	440	40.89	7.07	533	24.34	65.58	586	92.43	0.67
430	57.73	8.03	442	37.94	6.44	534	17.53	70.4	587	92.32	0.77
431	56.39	8.26	443	34.92	5.94	535	10.93	74.78	588	92.19	0.86
432	54.82	8.29	444	31.99	5.46	537	5.35	78.18	589	92.07	0.97
433	53.3	8.37	445	29.62	5.01	538	1.58	80.06	591	91.97	1.09
435	51.78	8.55	446	27.46	4.57	539	0.04	80.18	592	91.87	1.2
436	50.29	8.6	448	25.68	4.15	540	0.67	78.65	593	91.78	1.29
437	48.64	8.29	449	24.08	3.81	541	3.03	75.82	594	91.71	1.32

Long onda	t1	t2	Long onda	t1	t2	Long onda	t1	t2	Long onda	t1	t2
595	91.67	1.28	648	92.87	4.06	702	76.43	16.84	756	61.05	2.11
597	91.68	1.19	650	92.68	4.28	703	76.26	17.18	757	56.49	1.42
598	91.72	1.07	651	92.32	4.62	704	76.02	17.56	758	53.78	0.92
599	91.84	0.98	652	91.78	5.12	706	75.72	17.97	759	52.75	0.62
600	92	0.94	653	91.08	5.74	707	75.39	18.39	761	52.48	0.49
601	92.25	0.93	654	90.28	6.45	708	75.04	18.81	762	51.91	0.51
603	92.54	0.94	656	89.43	7.19	709	74.66	19.23	763	50.03	0.65
604	92.85	0.91	657	88.61	7.9	710	74.3	19.62	764	46.97	0.85
605	93.16	0.81	658	87.85	8.54	712	73.95	19.97	765	43.78	1.03
606	93.47	0.65	659	87.19	9.06	713	73.64	20.25	767	41.15	1.09
607	93.78	0.45	661	86.59	9.46	714	73.41	20.43	768	39.63	0.94
609	94.06	0.26	662	86.04	9.76	715	73.26	20.51	769	39.3	0.6
610	94.28	0.13	663	85.53	9.97	717	73.18	20.51	770	40.02	0.17
611	94.46	0.06	664	85.02	10.11	718	73.18	20.44	772	41.6	0.02
612	94.59	0.03	665	84.46	10.23	719	73.24	20.32	773	43.3	0.78
613	94.69	0.01	667	83.81	10.35	720	73.36	20.17	774	44.44	3.24
615	94.77	0.01	668	82.96	10.49	721	73.54	19.99	775	44.16	7.65
616	94.82	0.01	669	81.82	10.71	723	73.78	19.8	777	42.38	13.15
617	94.82	0.01	670	80.29	10.99	724	74.05	19.6	778	40.13	17.9
618	94.75	0	671	78.16	11.3	725	74.35	19.4	779	38.59	20.27
619	94.61	0	673	75.14	11.58	726	74.71	19.18	780	37.98	20.29
621	94.4	0.01	674	70.63	11.62	728	75.11	18.97	781	38.25	18.56
622	94.16	0.01	675	63.38	11.07	729	75.58	18.79	783	38.51	16.04
623	93.97	0.02	676	50.54	9.24	730	76.07	18.66	784	38.18	13.43
624	93.91	0.03	678	27.8	5.1	731	76.57	18.59	785	36.92	11.39
625	94	0.03	679	2.39	0.3	732	77.06	18.57	786	35.08	10.12
627	94.21	0.02	680	9.15	2.47	734	77.5	18.57	788	32.92	9.57
628	94.48	0	681	34.31	8.31	735	77.85	18.56	789	31.12	9.41
629	94.7	0.04	682	51.26	12.09	736	78.15	18.47	790	30.46	9.17
630	94.79	0.21	684	60.49	14.06	737	78.41	18.24	791	31.49	8.68
632	94.72	0.57	685	65.73	15.08	739	78.78	17.79	793	34.07	7.94
633	94.52	1.05	686	68.96	15.61	740	79.26	17.14	794	37.83	6.89
634	94.25	1.59	687	71.09	15.91	741	79.99	16.17	795	41.46	5.62
635	93.99	2.08	688	72.59	16.07	742	80.99	14.93	796	44.07	4.26
636	93.76	2.49	690	73.69	16.15	743	82.12	13.53	797	45.26	3.07
638	93.54	2.8	691	74.52	16.18	745	83.31	12	799	45.3	2.27
639	93.33	3.04	692	75.16	16.17	746	84.35	10.47	800	44.99	1.87
640	93.12	3.23	693	75.65	16.15	747	85.15	9.01	801	44.8	1.88
641	92.95	3.41	695	76.02	16.11	748	85.46	7.66	802	44.88	2.35
642	92.86	3.56	696	76.29	16.1	750	84.92	6.51	804	45.44	3.41
644	92.83	3.69	697	76.47	16.12	751	83.04	5.5	805	46.07	5.39
645	92.86	3.78	698	76.58	16.2	752	79.11	4.63	806	46.23	8.49
646	92.9	3.86	699	76.6	16.35	753	73.59	3.76	807	45.57	12.5
647	92.93	3.94	701	76.55	16.56	754	67.14	2.91	809	44.11	16.55

Long onda	t1	t2	Long onda	t1	t2	Long onda	t1	t2	Long onda	t1	t2
810	42.73	19.5	862	47.65	19.55	914	28.97	62.71	967	8.42	83.58
811	41.93	21.15	863	51.28	16.43	915	25.25	66.62	968	8.85	82.63
812	41.94	21.85	864	53.65	13.7	917	21.29	70.8	969	9.38	81.58
813	42.39	22.02	866	54.67	11.74	918	17.28	74.93	970	10.13	80.29
815	42.91	21.85	867	55.08	10.71	919	13.68	78.69	972	11	78.88
816	43.65	21.16	868	55.34	10.62	920	10.52	82.03	973	11.96	77.42
817	44.15	20.19	869	55.96	11.36	922	7.61	84.9	974	12.91	76.09
818	44.62	18.85	870	56.78	13.06	923	5.28	87.29	975	13.92	74.84
820	45.25	17.45	872	57.88	15.89	924	3.48	89.13	977	14.95	73.66
821	45.98	16.3	873	58.82	19.76	925	2.18	90.4	978	15.92	72.64
822	46.97	15.45	874	59.04	24.35	927	1.28	91.27	979	16.82	71.6
823	48.02	15.36	875	58.64	28.81	928	0.66	91.81	981	17.65	70.33
825	49.06	15.96	877	57.5	32.55	929	0.27	91.95	982	18.24	69.07
826	50.04	16.97	878	55.65	35.47	930	0.06	91.83	983	18.7	67.88
827	50.84	17.92	879	53.66	37.38	932	0	91.71	984	19.14	66.87
828	51.7	18.29	880	51.71	38.68	933	0.03	91.58	986	19.64	66.47
830	52.67	18.14	882	49.81	39.64	934	0.17	91.43	987	20.51	66.49
831	53.59	17.75	883	48.05	40.58	935	0.38	91.31	988	21.46	66.73
832	54.39	17.73	884	46.36	41.84	937	0.7	91.23	989	22.47	66.13
833	54.86	18.52	885	44.27	43.8	938	1.12	91.23	991	23.12	64.7
835	54.89	20.32	887	41.86	46.06	939	1.61	91.23	992	23.32	62.91
836	54.38	23.3	888	38.93	48.27	940	2.25	90.96	993	23.34	61.36
837	53.39	27	889	35.5	50.28	942	2.95	90.52	994	23.8	60.76
838	51.57	30.95	890	32.15	51.56	943	3.76	89.97	996	24.59	60.78
839	49.27	34.7	892	29.25	52	944	4.52	89.43	997	24.93	61.44
841	46.87	37.89	893	27.17	51.86	945	5.31	88.73	998	22.55	62.1
842	44.86	40.66	894	25.81	51.71	947	6.09	88.02	999	16.09	62.72
843	42.87	43.5	895	25.51	51.28	948	6.83	87.4	1001	9.13	62.47
844	40.74	46.6	897	26.41	50.63	949	7.61	86.72	1002	3.09	59.77
846	38.14	49.84	898	27.91	50.13	950	8.38	86.07	1003	0.26	56.45
847	35.14	52.42	899	30.04	49.71	952	8.93	85.65	1004	0.14	52.73
848	31.72	53.74	900	32.69	49.37	953	9.34	85.32	1006	1.01	49.43
849	28.38	53.16	902	35.31	49.16	954	9.61	85.09	1007	1.51	51.79
851	25.42	50.86	903	37.8	48.86	955	9.75	84.91	1008	1.65	51.96
852	23.48	47.39	904	40.03	48.6	957	9.64	84.9	1009	0.74	51.35
853	22.74	43.29	905	41.42	48.63	958	9.37	84.91	1011	0.11	46.44
854	23.23	39.11	907	41.84	49.18	959	9.08	84.91	1012	0	39.24
856	24.88	35.34	908	41.58	50	960	8.78	84.88	1013	0.48	31.33
857	27.87	31.98	909	40.5	51.18	962	8.57	84.73	1015	2.71	27.16
858	32.09	28.93	910	38.37	53.19	963	8.42	84.6	1016	10.05	27.03
859	37.36	25.88	912	35.42	56.03	964	8.29	84.36	1017	30.51	29.13
861	42.87	22.78	913	32.43	59.12	965	8.24	84.09	1018	58.39	23.68

APÉNDICE B

CONTRIBUCIONES RELATIVAS DE LOS FACTORES PLS-PLR A LOS ELEMENTOS FILA.



HOJA	T1	T2
1	86.18	5.38
2	93.85	2.38
3	84.5	8.95
4	89.36	7.82
5	89.59	9.05
6	70.43	5.81
7	83.84	5.33
8	74.89	3.94
9	83.33	13.58
10	76.65	15.51
11	51.86	14.13
12	91.68	0.13
13	75.82	15.38
14	29.61	68.61
15	65.23	32.17
16	16.67	68.08
17	39.95	55.22
18	68.01	27.57
19	56.88	40.93
20	31.23	64.54
21	48.86	47.24
22	59.73	18.88
23	5.19	24.38
24	3.72	91.91
25	67.69	22.32
26	73.54	0
27	58.82	44.71
28	0.06	89.72
29	38.45	1.97
30	2.19	53.85
31	2.7	14.41
32	11.13	68.32
33	35.96	0
34	58.07	14.56
35	6.46	83.69
36	1.23	83.67
37	83.46	9.34
38	1.56	22.55
39	4.86	39.99
40	6.67	0.08
41	44.63	9.22
42	0.01	31.99
43	1.79	18.2
44	48.64	8.29

HOJA	T1	T2
45	12.63	77.09
46	0.05	82.4
47	4.27	90.13
48	67.88	13.68
49	29.17	0.61
50	23.11	16.37
51	4.31	47.62
52	5.74	45.31
53	50.29	45.74
54	1.14	36.3
55	55.03	23.79
56	33.07	19.2
57	57.63	2.85
58	11.26	16.56
59	37.56	0
60	0.98	39.34
61	93.87	0
62	42.48	0.49
63	25.67	6.66
64	88.9	1.67
65	12.22	47.65
66	36.41	0.47
67	82.55	1.08
68	28.22	51.2
69	31.04	42.31
70	39.83	15.4
71	0.5	76.01
72	77.4	0
73	3.24	73.19
74	66.3	0
75	68.17	0
76	61.03	17.84
77	14.42	60.41
78	49.52	1.38
79	22.42	40.78
80	60.61	14.17
81	70.59	13.13
82	61.94	14.27
83	69.16	5.51
84	78.94	5.9
85	57.77	24.87
86	53.07	30.42
87	76.82	0
88	62.66	0.61

HOJA	T1	T2
89	39.75	8.52
90	65.76	9.56
91	54.24	0.01
92	54.1	20.24
93	0.14	7.94
94	0.14	14.16
95	52.42	4.26
96	1.48	25.69
97	59.96	13.73
98	66.02	11.86
99	38.22	0
100	50.24	31.87
101	79.58	9.44
102	56.78	27.59
103	71.01	1.36
104	51.97	0
105	16.99	19.02

APÉNDICE C

Artículo publicado por la revista

APPLICATIONS IN PLANT SCIENCES

**INVITED SPECIAL ARTICLE for Special
Issue “Advances in Plant Phenomics: From
Data and Algorithms to Biological Insights.”**

Article DOI: 10.1002/aps3.11383

DECISION LETTER

Date: Jun 10 2020 11:53AM
To: "Jorge Ugarte Fajardo"
From: "APPS" apps@botany.org
Subject: Decision for Manuscript APPS-D-19-00179R2

Ref.: Ms. No. APPS-D-19-00179R2

Early detection of black Sigatoka in banana leaves using hyperspectral images

Dear Mr. Jorge Ugarte Fajardo,

Your manuscript is accepted for publication in Applications in Plant Sciences. We are pleased to publish your research and thank you for contributing to the journal.

Your manuscript will now move into the eGalley, or page proof, stage. Based on our current schedule, you should receive your eGalley within approximately four to six weeks. Your review of the e-Galleys is important, and we will need to receive your response within 48 hours after receipt of the e-Galleys. Consequently, please keep us informed as to times when you will be away from your office during the next weeks and whom we can contact during this period. This will ensure that final publication will not be delayed by our failure to locate you when necessary.

If you have any questions about this, please feel free to contact the editorial office.

Sincerely,
Theresa M. Culley, Ph.D.
Editor-in-Chief
Applications in Plant Sciences

Editorial Office:
Beth Parada
Managing Editor
Applications in Plant Sciences
314-577-9486 apps@botany.org

Please be aware that if you ask to have your user record removed, we will retain your name in the records concerning manuscripts for which you were an author, reviewer, or editor.

In compliance with data protection regulations, you may request that we remove your personal registration details at any time. (Use the following URL: <https://www.editorialmanager.com/apps/login.asp?a=r>). Please contact the publication office if you have any questions.



INVITED SPECIAL ARTICLE

For the Special Issue: *Advances in Plant Phenomics: From Data and Algorithms to Biological Insights*

Early detection of black Sigatoka in banana leaves using hyperspectral images

Jorge Ugarte Fajardo^{1,6} , Oswaldo Bayona Andrade² , Ronald Criollo Bonilla² , Juan Cevallos-Cevallos^{3,4} , María Mariduena-Zavala¹ , Daniel Ochoa Donoso² , and José Luis Vicente Villardón⁵

Manuscript received 20 December 2019; revision accepted 10 June 2020.

¹Facultad de Ciencias Naturales y Matemáticas (FCNM), Escuela Superior Politécnica del Litoral (ESPOL), Guayaquil, Ecuador

²Facultad de Ingeniería Eléctrica y Computación (FIEC), Escuela Superior Politécnica del Litoral (ESPOL), Guayaquil, Ecuador

³Centro de Investigaciones Biotecnológicas del Ecuador (CIBE), Escuela Superior Politécnica del Litoral (ESPOL), Guayaquil, Ecuador

⁴Facultad de Ciencias de la Vida (FCV), Escuela Superior Politécnica del Litoral (ESPOL), Guayaquil, Ecuador

⁵Department of Statistics, Salamanca University (USAL), Salamanca, Spain

⁶Author for correspondence: gugarte@espol.edu.ec

Citation: Ugarte Fajardo, J., O. Bayona Andrade, R. Criollo Bonilla, J. Cevallos-Cevallos, M. Mariduena-Zavala, D. Ochoa Donoso, and J. L. Vicente Villardón. 2020. Early detection of black Sigatoka in banana leaves using hyperspectral images. *Applications in Plant Sciences* 8(8): e11383.

doi:10.1002/aps3.11383

PREMISE: Black Sigatoka is one of the most severe banana (*Musa* spp.) diseases worldwide, but no methods for the rapid early detection of this disease have been reported. This paper assesses the use of hyperspectral images for the development of a partial-least-squares penalized-logistic-regression (PLS-PLR) model and a hyperspectral biplot (HS biplot) as a visual tool for detecting the early stages of black Sigatoka disease.

METHODS: Young (three-month-old) banana plants were inoculated with a conidia suspension of the black Sigatoka fungus (*Pseudocercospora fijiensis*). Selected infected and control plants were evaluated using a hyperspectral imaging system at wavelengths in the range of 386–1019 nm. PLS-PLR models were run on the hyperspectral data set. The prediction power was assessed using leave-one-out cross-validation as well as external validation.

RESULTS: The PLS-PLR model was able to predict the presence of the disease with a 98% accuracy. The wavelengths with the highest contribution to the classification ranged from 577 to 651 nm and from 700 to 1019 nm.

DISCUSSION: PLS-PLR and HS biplot effectively estimated the presence of black Sigatoka disease at the early stages and can be used to graphically represent the relationship between groups of leaves and both visible and near-infrared wavelengths.

KEY WORDS: banana; black Sigatoka; HS biplot; hyperspectral imaging; penalized logistic regression (PLS-PLR); plant disease.

Banana (*Musa* L. spp.) is one of the most commonly cultivated crops worldwide and is the top agricultural commodity in many countries (Yeturu et al., 2016; Food and Agriculture Organization of the United Nations, 2017). However, banana production is severely affected by black leaf streak disease (BLSD, also known as black Sigatoka disease), which is caused by the fungal pathogen *Pseudocercospora fijiensis*. BLSD is considered the most widespread and damaging disease affecting bananas worldwide, causing plant necrosis in six symptomatic stages (Bakache et al., 2019).

BLSD is characterized by a biotrophic phase followed by a necrotrophic phase with visible symptoms. The disease affects the photosynthetic tissues of banana leaves and decreases chlorophyll

production (Chaerle et al., 2007), resulting in changes to the structure of the leaves. The first symptoms of the disease are small dark spots on the underside of the leaf that develop to form fine brown lines 2–3 mm long, which are also visible on the adaxial surface of the infected leaves. As the disease progresses, the stripes join together and gradually turn black, showing the first signs of necrosis. The dead zones of the leaves then dry out, causing defoliation and the early maturation of the fruit. The presymptomatic biotrophic phase can last for several weeks, and by the time symptoms are visible the banana plants are irreversibly affected and the disease has already spread (Marin et al., 2003), potentially inducing production losses of up to 85% (Luna-Moreno et al., 2019). In the initial

Applications in Plant Sciences 2020 8(8): e11383: <http://www.wileyonlinelibrary.com/journal/AppsPlantSci> © 2020 Ugarte Fajardo et al. *Applications in Plant Sciences* published by Wiley Periodicals LLC on behalf of Botanical Society of America. This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

stages of BLS (i.e., presymptomatic infected leaves and those in stages 1 and 2; see Table 1), the physical changes in the plant are minimal, making the visual identification of leaf damage difficult. Furthermore, asexual and sexual spores develop from stage 2 of the disease onward. Conidial (asexual) spores are waterborne over short distances, whereas ascospores (sexual spores) can be carried over long distances and are responsible for the spread of the disease; therefore, the early detection of BLS and the timely application of fungicides is crucial for controlling a *P. fijiensis* infestation. Early treatments reduce production costs and improve the health of crops while using shorter treatment times.

Currently, the detection of BLS relies on the visual observation of symptoms or the use of destructive analyses, such as those based on DNA or immunological assays (Luna-Moreno et al., 2019). To the best of our knowledge, hyperspectral image-based non-destructive methods have not yet been evaluated for the early detection of BLS.

The optical properties of a leaf can be characterized by (1) the light transmission through the leaf, (2) the light absorbed by chemicals within the leaf (e.g., pigments, water, sugars, lignin, and amino acids), and (3) the light reflected by the leaf surface or internal structures. Light reflectance levels depend on complex biophysical and biochemical interactions within the leaves. Changes in the photosynthetic pigments are evident in the visible region (VIS, 400–750 nm wavelength), changes in leaf structure and the scattering process affect reflectance in the near-infrared region (NIR, 750–1350 nm), and the water content influences the reflectance in the mid-infrared region (MIR, 1350–2500 nm) (Hunt and Rock, 1989). Many reflectance-related changes have been observed in diseased plants (Mahlein et al., 2012). In banana plants, *P. fijiensis* destroys photosynthetic tissue, which induces an increase in fluorescence and heat emission in the leaf blade and modifies the transport pattern of the photoassimilates, affecting the production of chlorophyll (Hidalgo et al., 2006). The resulting necrotic and chlorotic lesions cause variations of reflectance in the VIS and NIR regions of the spectrum.

Structural and chemical changes occurring in leaves during pathogenesis have enabled disease detection using hyperspectral image analyses (Siche et al., 2016). Previous works on sugar beet (*Beta vulgaris* L. subsp. *vulgaris*) (Mahlein, 2011), wheat (*Triticum aestivum* L.) (Ashourloo et al., 2014), tomato (*Solanum lycopersicum* L.), and lettuce (*Lactuca sativa* L.) (Lara et al., 2013) have provided some insights into the relationship between pathogen infections and spectral variations in leaves. In banana plants, hyperspectral image research has mostly focused on fruit measurements (Hu et al., 2015), quality control (Intaravanne et al., 2012), and the differentiation of black Sigatoka from yellow Sigatoka disease

(Bendini et al., 2015); however, the use of hyperspectral images for early detection of BLS has not been evaluated.

Data from hyperspectral leaf images usually present both spatial and spectral dimensions showing (1) high collinearity in the adjacent bands, (2) variability of hyperspectral signatures, and (3) high dimensionality due to the increased sensitivity and resolution of hyperspectral sensors (Lu and Fei, 2014). It is therefore necessary to apply multivariate data processing methods to be able to correlate hyperspectral fingerprints to plant infections.

Partial least squares (PLS) is a dimension reduction technique that is useful when the number of variables is greater than the number of observations. This tool maximizes the covariance between dependent and independent (predictors) variables, extracting from the predictors a set of orthogonal latent factors that are linear combinations of the original variables with the best predictive power (Abdi, 2010). When the response variable is categorical, PLS discriminant analysis (PLS-DA) has been used by other researchers (Brereton and Lloyd, 2014). PLS-DA is a linear classifier and its objective is to find a straight line that separates the regions. In this paper, we use an alternative method based on the algorithm proposed by Bastien et al. (2005), which takes into account the binary nature of the response using logistic regression rather than linear regression. This method is complemented with a hyperspectral biplot (HS biplot) that generalizes the proposal of Oyedele and Lubbe (2015), whose method was based on linear PLS components, whereas our study uses logistic PLS components. Additionally, penalized logistic regressions (PLR) have been introduced into the main algorithm to avoid the separation problems that occur when positive groups are presented separately to the negative groups, which prevents finding the maximum likelihood estimators (Albert and Anderson, 1984). This PLS-PLR model has not previously been used for the early detection of BLS from hyperspectral images.

The objective of this study was to assess the use of hyperspectral imaging for the early detection of BLS. The PLS-PLR technique was used to classify banana leaves, and was complemented by an HS biplot representation that facilitates the observation of the influence of the disease-associated changes in the reflectance of artificial light irradiated on a banana leaf.

METHODS

Plant inoculation

Banana plants (*Musa acuminata*, AAA Group, Cavendish subgroup, cultivar 'Williams') were obtained from commercial banana propagation facilities (Sociedad Ecuatoriana de Biotecnología C.A. [SEBIOCA], Guayaquil, Ecuador). A total of 100 plants that had been established for 3–4 months in a greenhouse were transported to our greenhouse located at the Centro de Investigaciones Biotecnológicas del Ecuador (Escuela Superior Politécnica del Litoral [ESPOL]) and grown at 28°C in 70% relative humidity and 12 hours of natural light, and were watered every 48 hours. Sixteen plants were randomly selected from this population, of which 10 were inoculated with *P. fijiensis* and six were mock-inoculated for use as a control. The inoculation was carried out as reported by Gbongue et al. (2019). Briefly, isolates of *P. fijiensis* were inoculated onto potato dextrose agar and incubated for two weeks at 30°C. The mycelium was then ground in 10 mL of sterile water and filtered to separate the mycelium from the conidia. The conidial suspension

TABLE 1. Severity scale of black leaf streak disease (BLS).

Stage	Symptoms
1	Yellowish spots <1 mm in diameter on the abaxial leaf surface
2	Red or brown streaks from 1 to 5 mm
3	Similar to stage 2, but streaks are >5 mm
4	Brown elliptical streaks on the abaxial leaf surface, black streaks on the adaxial leaf surface
5	The streak is totally black and has spread to the abaxial leaf surface. The streak is surrounded by a yellow halo.
6	The center of the streak is light gray surrounded by a black ring and a yellow halo.

was concentrated by centrifugation at $3000 \times g$ for 10 min at 4°C . Banana leaves were spray-inoculated with the concentrated conidia suspension using an aerograph atomizer (Gerensa, Guayaquil, Ecuador), and disease symptoms were monitored using the severity scale shown in Table 1, as suggested by Fouré (1986). The visual symptoms at each disease stage are shown in Fig. 1. The control plants were mock inoculated with autoclaved distilled water.

Imaging

The imaging system used in our experiments (Appendix S1) was exactly as described by Ochoa et al. (2016) and included a spectrometer (ImSpector V10E, Specim, Oulu, Finland) connected to a 12 bits per pixel camera (1500M-GE; Thorlabs Inc., Newton, New Jersey, USA) with high sensitivity in the NIR region. The camera was mounted on a slider in a push-broom configuration and controlled by a computer with a storage capacity of 1 TB, an Intel Core i5 3.1-GHz processor, and 16 GB of RAM. The system operated in a spectral range between 386–1019 nm, with a spectral resolution of 4.55 nm and a spatial resolution of 1040 (rows) by 1392 (columns).

The system allowed the nondestructive scanning of individual leaves from intact plants. Each hyperspectral image was composed of three dimensions, including one spectral and two spatial dimensions, resulting in a hyperspectral cube. The spatial dimensions consisted of the position of each pixel on the image ($M \times N$),

whereas the spectral dimension was made of the wavelength of reflected light (J). The system generated a hyperspectral cube for each leaf, with a resolution of 205 rows (M), 198 columns (N), and 520 wavelengths (J). To obtain the width of a hyperspectral cube (N), we set a pixel binning- x camera of 7, resulting in a reduction of 1392 pixels to 198 pixels. To calculate the number of wavelengths in a hyperspectral cube (J), we set a pixel binning- y camera of 2, resulting in a reduction of 1040 pixels to 520 pixels. Finally, the height of a hyperspectral cube (M) was determined according to the acquisition frame rate of the camera to scan the entire leaf holder, resulting in 205 pixels.

The plants were placed in the imaging system and the selected leaves were tagged and scanned. Three leaves were selected from each of the six control plants, but two were damaged due to manipulation, resulting in 16 images of non-infected leaves. Two leaves were selected from each of the 10 inoculated plants. The leaves were scanned every three days for three months. At every scan, the infection symptoms were visually assessed by an expert using the symptoms scale, as detailed in Table 1. During this period, the disease progression in the leaves was unequal. The level 1 symptoms appeared between seven and 31 days after the inoculation and increased irregularly, reaching higher severity levels in different time periods. In some leaves, the disease reached a severity level of 5. Due to manipulation, several leaves were damaged and were therefore discarded during the experiment.

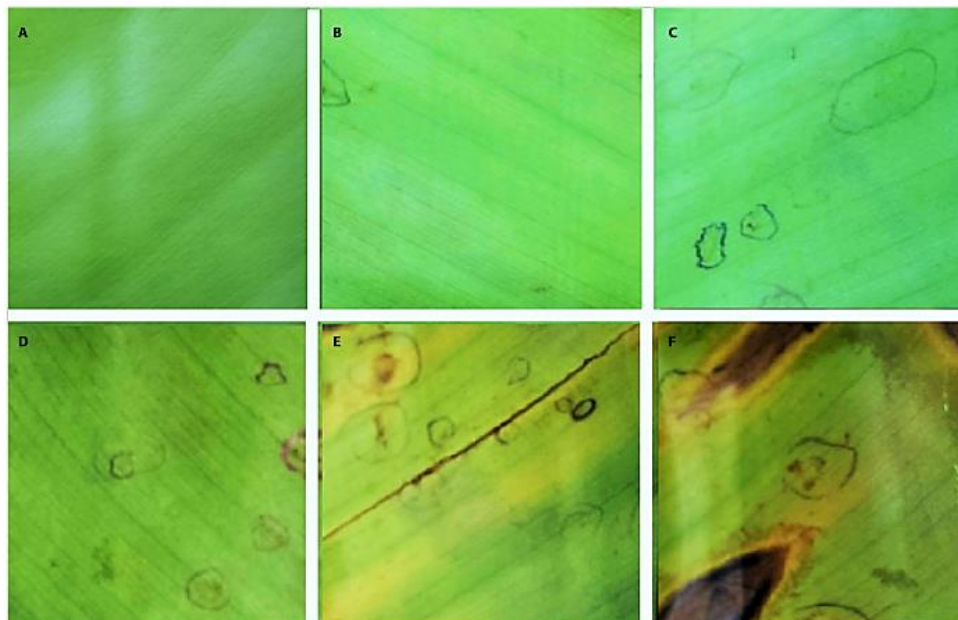


FIGURE 1. Stages of black leaf streak disease (BLSD): (A) non-infected, (B) stage 1, (C) stage 2, (D) stage 3, (E) stage 4, and (F) stage 5. Pen marks highlight the affected areas.

From the images scanned and tagged by experts, we selected those that belonged to infected leaves at the presymptomatic stage and stages 1 and 2 of BLSD. The presymptomatic images (16) were obtained six days before the leaf presented symptoms at severity level 1. The following images were taken in intervals of six days, during the progression of severity level 1 (54) and severity level 2 (18).

The final data set consisted of 104 images (16 non-infected, 16 presymptomatic, 54 severity level 1, and 18 severity level 2). The severity level reported corresponded to the highest disease stage found in the leaf.

Data preprocessing

Standard image calibration methods were applied. First, spectral calibration was applied using mercury (Hg), argon (Ar), helium (He), and hydrogen (H) light sources to estimate the wavelength corresponding to each charge-coupled device (CCD) line. Second, radiometric calibration was applied to reduce the influence of light intensity variation and the noise of the CCD sensor. For this purpose, white and dark references were imaged before each scanning session and the raw spectral image was normalized. In addition, the spatial calibration of the slider was performed to avoid overlap between leaf regions when the images were acquired.

To retain only the leaf pixels of each hyperspectral cube, we generated a mask using the image at a wavelength of 700 nm, and setting the leaf holder and background pixels to zero. Next, each hyperspectral cube was normalized using the standard normal variate technique (Barnes et al., 1989). This step is necessary to compensate for reflectance variation caused by the relative orientation of the leaf surface and the sensor. Finally, the dimensionality reduction was achieved by calculating the average of the reflectance values measured at each wavelength of the hyperspectral cube, resulting in a matrix of $I = 104$ rows (one per image) and $J = 520$ columns (Appendix S2). A preliminary analysis of infected regions showed differences in the spectral

patterns of the disease stages. Figure 2 shows the reflectance patterns for severity level 2 and non-infected regions.

Statistical methods

PLS–PLR model and HS biplot—The data were arranged into one vector and one matrix. The Y vector consisted of the binary (infected and non-infected) response variable, whereas the X matrix ($X_1, \dots, X_J$) consisted of a set of predictors that correspond to the reflectance intensity at each wavelength.

The general logistic PLS model and its associated biplot were included in the MultibiotR package (Vicente-Villardón, 2017). The package implements the algorithm that estimates PLS components and applies a logistic regression on the PLS components and the response vector Y . To prevent the separation problem reported in previous studies (Albert and Anderson, 1984; Santner and Duffy, 1986) and detected in our initial tests, the package applies a ridge penalty (Le Cessie and Van Houwelingen, 1992), which is calculated by the sum of the squares of the coefficients (L2 norm) multiplied by a penalty parameter λ . The parameter λ can have a value between 0 and 1, and can be adjusted by cross-validation. For the purpose of finding the model that best describes the data, we tested values of λ in the [0.1–0.9] range using incremental steps of 0.1 and calculated, for each value of λ , the following goodness-of-fit measures: difference of deviance (DiffDeviance), Cox and Snell's R^2 , Nagelkerke's R^2 , and McFadden's R^2 (Allison, 2014; Walker and Smith, 2016).

The biplot that represents the labeled leaves, the wavelengths, and the boundary of the prediction regions (i.e., the HS biplot) allows for the visual inspection of the PLS–PLR model. The wavelengths were represented by lines colored according to the spectral band to which they belong. The statistical procedures applied in PLS–PLR and HS biplot are detailed in Appendix 1.

The PLS–PLR model yielded estimated values between 0 and 1. Leaves with model values above or below 0.5 were considered to be predicted-infected or predicted-healthy, respectively. The goodness-of-fit of the logistic regression was calculated using the pseudo R^2 measures described above.

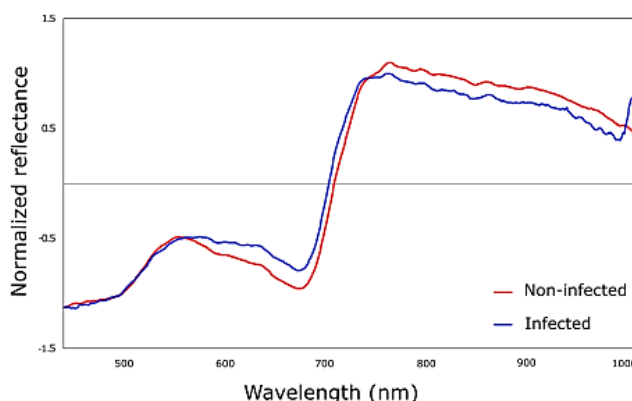


FIGURE 2. Normalized reflectance curves of labeled regions in non-infected and infected banana (*Musa acuminata*) leaves. The curves were normalized using the standard normal variate technique. Infected leaves correspond to severity level 2.

RESULTS

Penalty parameter (λ) selection and model goodness-of-fit

First, the penalty parameter of the PLS–PLR model was selected from different λ coefficients. The goodness-of-fit values obtained after each interaction are shown in Table 2.

We obtained the best goodness-of-fit measures (highest DiffDeviance value of 88.488) using a λ value of 0.1, showing that the fitted model maintained the greatest variance. Furthermore, the P value of 6097×10^{-20} shows a significant association between the latent variables and the response variable. The pseudo R^2 measures were also estimated; McFadden's R^2 (0.991) indicated a high explicative power of the model, while the Cox and

TABLE 2. Goodness-of-fit measures for the PLS–PLR model using incremental penalty values (λ).

λ	DiffDeviance ^a	Cox and Snell's R^{2b}	Nagelkerke's R^{2b}	McFadden's R^{2b}
0.1	88.488	0.573	0.994	0.991
0.2	88.005	0.571	0.991	0.986
0.3	87.668	0.570	0.988	0.982
0.4	87.405	0.568	0.986	0.979
0.5	87.187	0.568	0.985	0.976
0.6	86.999	0.567	0.984	0.974
0.7	86.832	0.566	0.982	0.972
0.8	86.682	0.565	0.981	0.971
0.9	86.543	0.565	0.980	0.969

^aDifference of deviance (Hosmer et al., 1997).

^bPseudo R^2 indices for binary logistic regression models (Allison, 2014; Walker and Smith, 2016).

Snell's R^2 (0.573) and Nagelkerke's R^2 (0.994) also indicated a high goodness-of-fit.

The penalty $\lambda = 0.1$ was applied in each iteration for calculating the coefficients, providing a stable value that maximized the likelihood function while controlling the error. Figure 3 shows the logistic response surface fitted on the space spanned by the first two PLS components. The model is:

$$P_y = \frac{e^{(27.226+1.546t_1+1.318t_2)}}{1 + e^{(27.226+1.546t_1+1.318t_2)}}$$

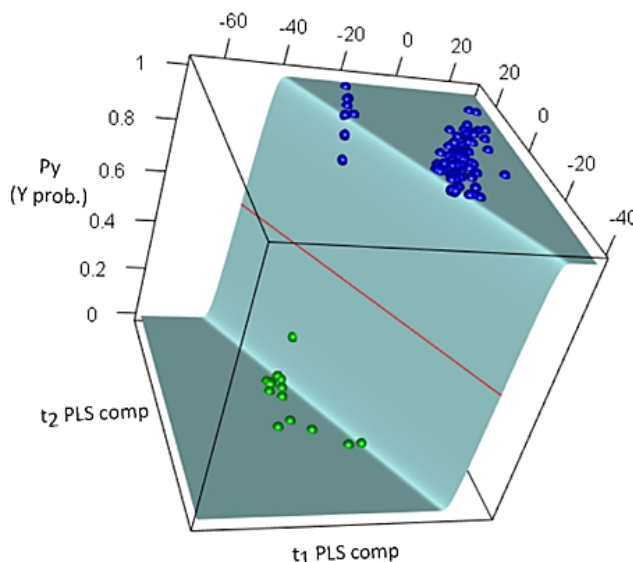


FIGURE 3. Response surface for the PLS–PLR model in BLSD detection. The red line corresponds to the probability equal to 0.5. The control leaves are shown as green points and infected leaves are shown as blue points.

where P_y is the infected leaf probability, t_1 is the first PLS–PLR component, and t_2 is the second PLS–PLR component.

HS biplot

The HS biplot of the training data set is shown in Fig. 4. The first two PLS components contributed 77% of the observed variability.

The HS biplot indicates three main groupings: a group of non-infected leaves (blue ellipse) and two other groups of infected leaves (red ellipses). The clustering of non-infected and infected plants was mostly observed in component 1. The wavelengths that contributed the most to the grouping of the non-infected samples ranged from 577 to 651 nm (yellow to red range). The first of the two groups of infected leaves was located near the non-infected group (numbered points inside the red ellipse in Fig. 4) in the HS biplot, mainly influenced by a low density of disease symptoms in the leaves (Table 3). The second group of infected leaves was composed of both pre-symptomatic and symptomatic leaves. The wavelengths that contributed the most to the grouping of the pre-symptomatic leaves (turquoise), as well as several leaves at severity levels 1 (blue) and 2 (red), were in the NIR range of the spectrum. However, the grouping of the other leaves at symptomatic levels 1 and 2 was determined by their reflectance in the 577 to 651 nm (yellow to red) range. This was also observed in the external validation data set (Fig. 5). The HS biplot of the test data set showed that the healthy (blue ellipse) and infected (red ellipse) leaves had all been correctly classified, with the exception of two samples, one healthy and one infected.

Model prediction and validation

The model was validated using the leave-one-out cross-validation (LOOCV) method, as shown in Appendix S3. During the cross-validation process, 102 leaves were properly classified as non-infected or infected, which represented an overall classification accuracy of 98% (Table 4). Only two non-infected samples were classified as being infected, with a probability of 0.738 and 0.816, respectively. All infected leaves at the different disease stages were correctly classified. The positive predictive value was 98%, while the sensitivity or recall value was 100% (Table 4), indicating that all infected leaves were correctly identified. The estimated global HS biplot goodness-of-fit (the squared correlation coefficient between the adjusted and observed values) was 77.07%.

External validation

The PLS–PLR model fitted to the initial training data set was used to predict the presence of the disease in new leaves. A new data set with images of 16 non-infected and 21 infected leaves was used to evaluate the efficacy of the model. The prediction accuracy of the new samples was 95% (35 successful identifications and two errors). Appendix S4 indicates the predicted probability for the external validation test.

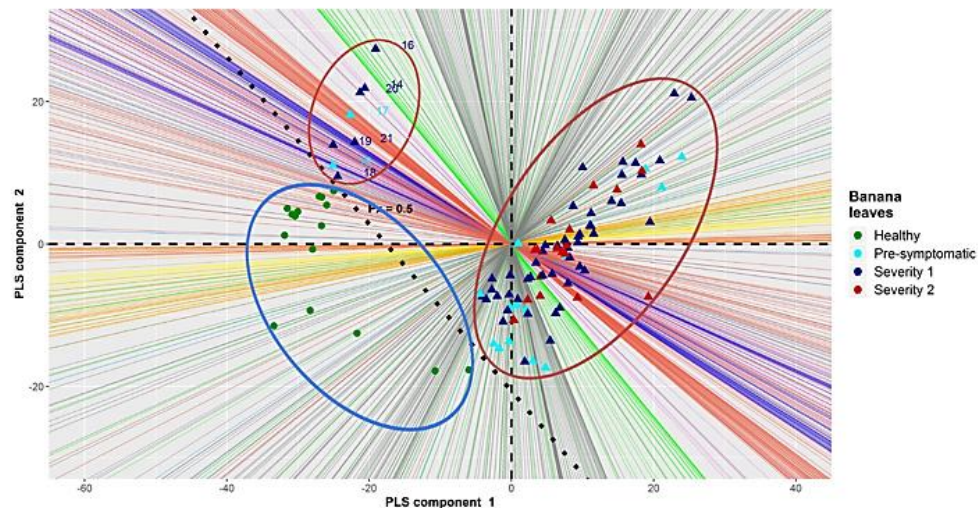


FIGURE 4. HS biplot of the training data set. Banana leaves are represented by points. Each wavelength is represented by straight lines colored according to the colors of the electromagnetic spectrum. The diagonal dotted line separates the healthy and infected leaves. The blue ellipse encloses healthy leaves and the red ellipses contain infected leaves. PLS component 1 explains 50.99% of variance, while PLS component 2 explains 26.08% of the variance. The numbered points correspond to banana leaves with a low-severity infection (Table 3).

TABLE 3. HS biplot description of banana leaves with a low-severity infection.

Leaf sample no. ^a	Severity ^b	Observations ^c
14	1	Presents two pixels (severity 1)
15	0	Without symptoms
16	1	Presents seven pixels (severity 1)
17	0	Without symptoms
18	1	No visible infected pixels
19	1	Presents 10 pixels (severity 1)
20	1	Presents six pixels (severity 1)
21	1	Presents 19 pixels (severity 1)

^aNumbers correspond to the numbers displayed on Fig. 4. On the HS biplot, these leaves were plotted near the healthy leaves due to their less severe symptoms.

^bSeverity levels correspond to the scale presented in Table 1.

^cNumber of pixels in the infected area.

DISCUSSION

Agricultural crops are constantly threatened by pests and plant diseases that reduce production and quality. New agricultural strategies and techniques to identify plant diseases are based on non-destructive methods that detect the disease in early stages, allowing timely management to prevent the spread of the disease and minimize the effect of fungicides in the environment. Remote sensing allows early detection of plant diseases using methods based on reflectance in the VIS and NIR regions. Disease detection involves steps such as image acquisition, image pre-processing, image segmentation, feature extraction, and classification.

In this study, hyperspectral images were used to classify BLSD-infected and healthy banana plants based on the marked differences

in the VIS and NIR spectra of both groups. Our results are consistent with previous reports showing that disease symptoms in plants can change spectral reflectance in the VIS (400–700 nm) and NIR (700–1100 nm) ranges (Ayala-Silva and Beyl, 2005). General changes in reflectance occurring during plant–pathogen interactions have been associated with impairments in the leaf structure and chemical composition of the tissue during pathogenesis, which can be observed by the succession from chlorotic to necrotic tissue (Mahlein, 2016).

PLS was originally developed for continuous response variables. In the case of binary responses, a linear regression cannot guarantee that response-fitted values fall at 0 or 1. In this study, we used logistic rather than linear regressions to correlate the response variable with the PLS components. Previous studies on plant disease detection have reported relatively high accuracies and sensitivities similar to PLS models, but the model interpretability is low, and most studies did not include leaves in presymptomatic stages (Rumpf et al., 2010; Mahlein, 2011; Zhu et al., 2017). In our study, the application of PLS–PLR showed a prediction accuracy of 98% in the presymptomatic and early stages of BLSD. The goodness-of-fit measures obtained for the PLS–PLR model were similar to those reported in the literature (Huang et al., 2007; Bock et al., 2010; Cevallos-Cevallos et al., 2018).

As indicated by the HS biplot, which shows leaves as points and wavelengths as lines, the non-infected leaves were characterized by the prominent presence of wavelengths in the VIS spectrum, whereas infected leaves were associated with wavelengths in both the VIS and NIR ranges, depending on the stage of the disease or the proportion of symptomatic to healthy tissue. These results are in agreement with previous studies, as disease

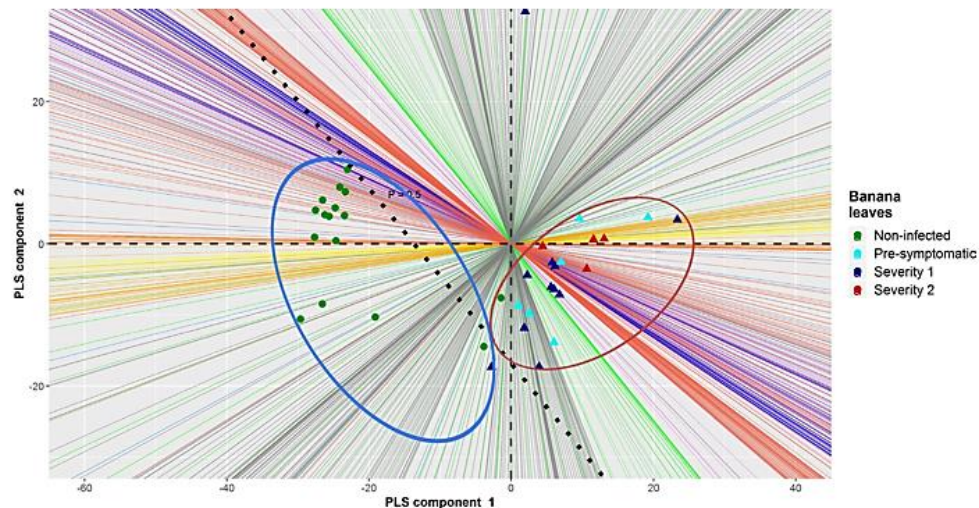


FIGURE 5. HS biplot of the external validation data set. The diagonal dotted line separates predicted healthy (blue ellipse) and infected leaves (red ellipse). Each wavelength is represented by straight lines colored according to the electromagnetic spectrum colors. PLS component 1 explains 50.68% of variance, while PLS component 2 explains 26.21% of the variance.

symptoms in plants have been reported to increase spectral reflectance in both the VIS (400–700 nm) and NIR (700–1100 nm) ranges (Ayala-Silva and Beyl, 2005). General changes in reflectance occurring during plant–pathogen interactions have been associated with impairments in the leaf structure and changes in the chemical composition of the tissue during pathogenesis, which can be observed by the succession of chlorotic and necrotic tissue (Mahlein, 2016).

The clustering of non-infected and infected plants was mostly observed within PLS component 1. The wavelengths that contributed the most to PLS component 1 ranged from 577 to 651 nm (yellow to red). Changes in the yellow range of the spectrum suggest that detection of plant chlorosis occurs in the initial stages of BLS. Chlorosis is caused by insufficient chlorophyll accumulated in the plants, leading to the yellowing of the leaf, which is usually measured with yellowness indexes (Adams et al., 1999). Similarly, changes in the orange–red range of the spectrum suggest the

succession of chlorotic to necrotic tissue, as observed in the red or brown streaks that appear in stage 2 of BLS (Fouré, 1986). Interestingly, presymptomatic plants showing neither chlorosis nor necrosis were clustered apart from the non-infected ones, suggesting changes in the leaf surface of these infected plants despite the lack of visible symptoms. Biotrophic presymptomatic stages do not usually cause observable changes in leaves, but some fungal pathogens can produce structures on the leaf surface that can influence the optical properties of the plant (Mahlein, 2016). Non-infected or infected banana plants were scattered in two groups in the HS biplot, which is in agreement with previous reports showing two groups of non-infected and dwarf banana plants due to the natural biological diversity observed in agricultural conditions (Cevallos-Cevallos et al., 2018). These results confirm the high prediction capacity of the PLS–PLR model and the efficiency of the HS biplot to represent the relationships between the variables and the groups of individuals with non-infected and infected leaves.

Hyperspectral imaging provides a non-destructive method for analyzing plants infected with BLS and possibly other pathogens. The development of technologies offering larger data storage capacities, faster computers, more sensitive detectors, and different analytical techniques for hyperspectral images, combined with suitable statistical techniques, makes it possible to detect plant diseases even at early stages, and enables the capture and modeling of physiological changes of leaves using close-range hyperspectral data. The early detection of infectious diseases plays a crucial role in both treatment and prevention strategies.

PLS–PLR and HS biplot visual representation are promising techniques for the analysis of hyperspectral data, even after considering the high reduction of the dimensionality after preprocessing

TABLE 4. Confusion matrix for the classification accuracy assessment of the PLS–PLR model.

	True infected leaf	False non-infected leaf	Prediction metrics
Test result	True positive 88	False positive 2	Positive predictive value 0.98
	False negative 0	True negative 14	Negative predictive value 1
Prediction metrics	Sensitivity 1	Specificity 0.88	Accuracy 0.98



of the raw data. The PLS–PLR model provides excellent predictive power, which is complemented by the high level of visual interpretation offered by the HS biplot.

The system presented here combines hyperspectral technology with advanced data analysis and statistical methods to accurately predict plant disease by measuring the reflectance differences resulting from the biophysical and biochemical characteristic changes following infection. It is currently configured as a stationary imaging system and could be used in a laboratory for the recognition of foliar diseases in other plants and for food quality evaluations. In the future, however, the system could be transformed, with few changes, into an airborne imaging system for scanning images of crop fields.

Future research should concentrate on improving the detection of different severity levels of the disease and on the more detailed analysis of the wavelengths that have greater influence, taking into account other factors that can cause spectral changes.

ACKNOWLEDGMENTS

This research was supported by VLIR - UOS grant “VLIR Network Ecuador.”

AUTHOR CONTRIBUTIONS

D.O.D. and J.C.C. formulated the research problem. D.O.D., J.C.C., J.L.V.V., and J.U.F. designed the approaches. D.O.D., R.C.B., and O.B.A. collected the data. J.L.V.V., D.O.D., J.C.C., M.M.Z., R.C.B., O.B.A., and J.U.F. developed the processing workflow. J.L.V.V. and J.U.F. performed the data analysis. All authors contributed to the writing and development of the manuscript. All authors read and approved the final manuscript.

DATA AVAILABILITY STATEMENT

The two data sets used in this study (i.e., Training data set [data_banana.txt] and Validation data set [data_banana_test.txt]) are available at: https://drive.google.com/drive/folders/1JtHKfCPedxof0TY_3Yhvb_LBHsBldRL?usp=sharing.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

APPENDIX S1. Hyperspectral imaging system used in this study: (A) camera, (B) spectrometer, (C) light source, (D) slider, (E) holder.

APPENDIX S2. Dimensionality reduction of the hyperspectral cubes. The mean of the reflectance matrix at each wavelength is calculated.

APPENDIX S3. Predicted probability using leave-one-out cross-validation (LOOCV) to assess the PLS–PLR model.

APPENDIX S4. Predicted probability for the external validation of the PLS–PLR model.

LITERATURE CITED

- Abdi, H. 2010. Partial least squares regression and projection on latent structure regression (PLS regression). *Wiley Interdisciplinary Reviews: Computational Statistics* 2(1): 97–106.
- Adams, M. L., W. D. Philpot, and W. A. Norvell. 1999. Yellowness index: An application of spectral second derivatives to estimate chlorosis of leaves in stressed vegetation. *International Journal of Remote Sensing* 20(18): 3663–3675.
- Albert, A., and J. A. Anderson. 1984. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* 71(1): 1–10.
- Allison, P. D. 2014. Measures of fit for logistic regression. In *Proceedings of the SAS Global Forum*, 1–13. Washington, D.C., USA.
- Ashourloo, D., M. Mobasheri, and A. Huete. 2014. Evaluating the effect of different wheat rust disease symptoms on vegetation indices using hyperspectral measurements. *Remote Sensing* 6(6): 5107–5123.
- Ayala-Silva, T., and C. A. Beyl. 2005. Changes in spectral reflectance of wheat leaves in response to specific macronutrient deficiency. *Advances in Space Research* 35(2): 305–317.
- Bakache, A., J.-P. Douzals, B. Bonicelli, E. Cotteux, L. de Lapeyre de Bellaire, and C. Sinfort. 2019. Development of a rapid methodology for biological efficacy assessment in banana plantations: Application to reduced dosages of contact fungicide for black leaf streak disease (BLSD) control. *Pest Management Science* 75(4): 1081–1090.
- Barnes, R. J., M. S. Dhanoa, and S. J. Lister. 1989. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied Spectroscopy* 43(5): 772–777.
- Bastien, P., V. E. Vinzi, and M. Tenenhaus. 2005. PLS generalised linear regression. *Computational Statistics and Data Analysis* 48: 17–46.
- Bendini, H., A. Jacon, A. C. Moreira Pessôa, J. A. Pompeu Pavanelli, W. da Silva Moraes, F. J. Ponzoni, and L. Fonseca. 2015. Spectral characterization of banana leaves (*Musa spp.*) for detection and differentiation of black sigatoka and yellow sigatoka. XVII Simpósio Brasileiro de Sensoriamento Remoto, João Pessoa, Brazil.
- Bock, C., G. Poole, P. Parker, and T. Gottwald. 2010. Plant disease severity estimated visually, by digital photography and image analysis, and by hyperspectral imaging. *Critical Reviews in Plant Sciences* 29(2): 59–107.
- Brereton, R. G., and G. R. Lloyd. 2014. Partial least squares discriminant analysis: Taking the magic away. *Journal of Chemometrics* 28: 213–225.
- Cevallos-Cevallos, J. M., C. Jines, M. G. Maridueña-Zavala, M. J. Molina-Miranda, D. E. Ochoa, and J. A. Flores-Cedeno. 2018. GC-MS metabolite profiling for specific detection of dwarf somaclonal variation in banana plants. *Applications in Plant Sciences* 6(11): e01194.
- Chaerle, L., I. Leinonen, H. G. Jones, and D. Van Der Straeten. 2007. Monitoring and screening plant populations with combined thermal and chlorophyll fluorescence imaging. *Journal of Experimental Botany* 58(4): 773–784.
- Demey, J., J. L. Vicente-Villardón, M. Galindo-Villardón, and A. Zambrano. 2008. Identifying molecular markers associated with classification of genotypes by external logistic biplots. *Bioinformatics* 24(24): 2832–2838.
- Food and Agriculture Organization of the United Nations. 2017. FAOSTAT statistical database. Website <http://www.fao.org/faostat/en/#home> [accessed 15 June 2018].
- Fouré, E. 1986. Varietal reactions of bananas and plantains to black leaf streak disease. *Banana and Plantain Breeding Strategies* 21: 110–113.
- Fu, J. W. 1998. Penalized regressions: The bridge versus the lasso. *Journal of Computational and Graphical Statistics* 7(3): 397–416.
- Gbongue, L.-R., I. Lalaymia, A. Zeze, B. Delvaux, and S. Declercq. 2019. Increased silicon acquisition in bananas colonized by *Rhizoglyphus irregularis* MUCL 41833 reduces the incidence of *Pseudocercospora fijiensis*. *Frontiers in Plant Science* 9: 1977.
- Hidalgo, M., A. Tapia, W. Rodriguez, and E. Serrano. 2006. Efecto de la Sigatoka negra (*Mycosphaerella fijiensis*) sobre la fotosíntesis y transpiración foliar del banano (*Musa sp. AAA*, cv. Valery). *Agronomía Costarricense* 30(1): 35–41.
- Hosmer, D. W., T. Hosmer, S. Le Cessie, and S. Lemeshow. 1997. A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine* 16(9): 965–980.
- Hu, M.-H., Q.-L. Dong, P. K. Malakar, B.-L. Liu, and G. K. Jaganathan. 2015. Determining banana size based on computer vision. *International Journal of Food Properties* 18(3): 508–520.



- Huang, W., D. W. Lamb, Z. Niu, Y. Zhang, L. Liu, and J. Wang. 2007. Identification of yellow rust in wheat using in-situ spectral reflectance measurements and airborne hyperspectral imaging. *Precision Agriculture* 8(11): 187–197.
- Hunt Jr., E. R., and B. N. Rock. 1989. Detection of changes in leaf water content using near- and middle-infrared reflectances. *Remote Sensing of Environment* 30(1): 43–54.
- Intaravanne, Y., S. Sumriddetchkajorn, and J. Nukeaw. 2012. Ripeness level indication of bananas with visible and fluorescent spectral images. In 9th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, 1–4. IEEE, New York, New York, USA.
- Lara, M., B. Diezma Iglesias, L. Lleó García, J.-M. Roger, Y. Garrido, M. Gil, and M. Ruiz-Altisent. 2013. Aplicación de imagen hiperespectral para observar el efecto de la salinidad en hojas de lechuga. In VII Congreso Ibérico de Agroingeniería y Ciencias Hortícolas, Madrid, Spain.
- Le Cessie, S., and J. C. Van Houwelingen. 1992. Ridge estimators in logistic regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 41(1): 191–201.
- Lu, G., and B. Fei. 2014. Medical hyperspectral imaging: A review. *Journal of Biomedical Optics* 19(1): 010901.
- Luna-Moreno, D., A. Sánchez-Álvarez, I. Islas-Flores, B. Canto-Canche, M. Carrillo-Pech, J. E. Villarreal-Chiu, and M. Rodríguez-Delgado. 2019. Early detection of the fungal banana black Sigatoka pathogen *Pseudocercospora fijiensis* by an SPR immunosensor method. *Sensors* 19(3): 465.
- Mahlein, A.-K. 2011. Detection, identification, and quantification of fungal diseases of sugar beet leaves using imaging and non-imaging hyperspectral techniques. Doctoral thesis, Rheinischen Friedrich-Wilhelms-Universität Bonn, Bonn, Germany.
- Mahlein, A.-K. 2016. Plant disease detection by imaging sensors—parallels and specific demands for precision agriculture and plant phenotyping. *Plant Disease* 100(2): 241–251.
- Mahlein, A., E. Oerke, U. Steiner, and H.-W. Dehne. 2012. Recent advances in sensing plant diseases for precision crop protection. *European Journal of Plant Pathology* 133: 197–209.
- Marin, D. H., R. A. Romero, M. Guzmán, and T. B. Sutton. 2003. Black Sigatoka: An increasing threat to banana cultivation. *Plant Disease* 87(3): 208–222.
- Ochoa, D., J. Cevallos, G. Vargas, R. Criollo, D. Romero, R. Castro, and O. Bayona. 2016. Hyperspectral imaging system for disease scanning on banana plants. In Sensing for Agriculture and Food Quality and Safety VIII, vol. 9864, 98640M. International Society for Optics and Photonics, Bellingham, Washington, USA.
- Oyedele, O. F., and S. Lubbe. 2015. The construction of a partial least-squares biplot. *Journal of Applied Statistics* 42(11): 2449–2460.
- Rumpf, T., A.-K. Mahlein, U. Steiner, E.-C. Oerke, H.-W. Dehne, and L. Plumer. 2010. Early detection and classification of plant diseases with support vector machines based on hyperspectral reflectance. *Computers and Electronics in Agriculture* 72(11): 91–99.
- Santner, T. J., and D. E. Duffy. 1986. A note on A. Albert and J. A. Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models. *Biometrika* 73(3): 755–758.
- Siche, R., R. Vejarano, V. Aredo, L. Velasquez, E. Saldaña, and R. Quevedo. 2016. Evaluation of food quality and safety with hyperspectral imaging (HSI). *Food Engineering Reviews* 8(3): 306–322.
- Vicente-Villardón, J. L. 2017. MultiBiplotR: Multivariate analysis using biplots. R package version 0.1.0. Website <http://biplot.usal.es/multibiplot/multibiplot-in-r/> [accessed 20 November 2019].
- Vicente-Villardón, J. L., M. P. Galindo-Villardón, and A. Blázquez-Zaballos. 2006. Logistic biplots. In M. Greenacre, and J. Blasius [eds.], Multiple correspondence analysis and related methods. Chapman and Hall/CRC Press, Boca Raton, Florida, USA.
- Walker, D. A., and T. J. Smith. 2016. Nine pseudo R^2 indices for binary logistic regression models. *Journal of Modern Applied Statistical Methods* 15(1): 848–854.
- Yeturu, S., K. Méndez, P. Garrido, S. Serrano, and A. Garrido. 2016. Serological and molecular identification of cucumber mosaic virus (CMV) infecting banana crops in Ecuador. *Ecuador Es Calidad: Revista Científica Ecuatoriana* 3: 17–22.
- Zhu, H., B. Chu, C. Zhang, F. Liu, L. Jiang, and Y. He. 2017. Hyperspectral imaging for presymptomatic detection of tobacco disease with successive projections algorithm and machine-learning classifiers. *Scientific Reports* 7: 4125.

APPENDIX 1. Partial-least-squares penalized-logistic-regression (PLS-PLR) method applied to classify BLSD-infected and healthy banana leaves.

Let Y be the binary response variable and $(X_1, \dots, X_p)$ a set of predictors. For a sample of size n , the data can be organized in a response vector $y = (y_1, \dots, y_n)^T$ and a matrix of predictors $X = (x_1, \dots, x_p) = (x_{ij})$ ($i = 1, \dots, n; j = 1, \dots, p$), where y_i is either 0 or 1 for the presence or absence of the main characteristic and x_{ij} is the value of the i th individual on the j th predictor. The columns of X are supposed to be centered and possibly standardized.

We are searching for components that are linear combinations of the predictors and that best explain the response (in a logistic regression manner). Let t_h the vector containing the scores of each individual on one of those combined components, then $t_h = \sum_{j=1}^p w_{hj} x_j = X w_h$ with $w_h = (w_{h1}, \dots, w_{hp})^T$ being the vector of coefficients. Normally we will use m of those components that are mutually orthogonal.

The logistic PLS regression model is written as

$$E(y) = \hat{y} = \frac{1}{1 + e^{-(c_0 + \sum_{h=1}^m c_h t_h)}}$$

or

$$\text{logit}(\hat{y}) = \log\left(\frac{\hat{y}}{1 - \hat{y}}\right) = c_0 + \sum_{h=1}^m c_h t_h$$

Or in matrix form $\text{logit}(\hat{y}) = c_0 + Tc$, where $\hat{y} = (\hat{y}_1, \dots, \hat{y}_n)^T$ is the vector of fitted probabilities of the presence of the disease in each individual, and $c = (c_1, \dots, c_m)^T$ are the coefficients of the regression on the components. The model is a standard logistic regression on the PLS components. The constant c_0 must be kept because the binary variable cannot be centered. In terms of the original variables,

$$\text{logit}(\hat{y}) = c_0 + Tc = c_0 + XWc = c_0 + Xb$$

where $W = (w_1, \dots, w_m)$ and $b = (b_1, \dots, b_p)^T$ are the coefficients on the observed variables. The problem now is to estimate T , W , c , c_0 , and b .

Estimation algorithm

Here, we use the algorithm developed by Bastien et al. (2005), with slight modifications.

- Calculation of t_1 , the first PLS component.
 - For each predictor ($j = 1, \dots, p$), compute the regression coefficient w_{1j} of x_j in the logistic regression of y on x_j to obtain $w_1 = (w_{11}, \dots, w_{1p})^T$.
 - Normalize the vector w_1 : $w_1 = w_1 / \|w_1\|$.
 - Compute the component scores $t_1 = X w_1 / w_1^T w_1$.
- Calculation of t_h , the h th PLS component. The components $t_1, \dots, t_{h-1}$ have been already obtained.
 - For each predictor ($j = 1, \dots, p$), compute the regression coefficient w_{hj} of x_j in the logistic regression of y on $t_1, \dots, t_{h-1}$ and x_j to obtain $w_h = (w_{h1}, \dots, w_{hp})^T$.

- b. Normalize the vector $w_h := w_h / \|w_h\|$.
 - c. Compute the residual matrix X_{h-1} of the linear regression of X on $t_1, \dots, t_{h-1}$.
 - d. Compute the component scores $t_h = Xw_h / w_h^T w_h$.
- X is factorized as $X = TP$.
 - Logistic regression of y on the retained PLS components

$$\text{logit}(\hat{y}) = c_0 + \sum_{h=1}^m c_h t_h$$

- Expression of the model in terms of the original predictors $b = Wc$.

Remarks

- Although the original algorithm is not clear, we think that all the logistic models in steps 1a, 2a, and 3 must include a constant.
- When the model is good, i.e., when it is able to discriminate almost perfectly among presences and absences, the maximum likelihood method for logistic regression does not converge and the estimators tend to infinity. This is known as the separation problem (Albert and Anderson, 1984; Santner and Duffy, 1986), and is easily solved using a penalty. Here, we use the ridge (Le Cessie and Van Houwelingen, 1992) because of its simplicity. Additionally, when there are many variables that influence the response, as is the case of analysis of this work, this penalty (ridge) offers better results, whereas the lasso penalty (Fu, 1998) is more efficient when fewer variables influence the result. The ridge penalty is equal to the sum of the squares of the coefficients (L_2 norm) by the penalty parameter λ . The parameter λ may have a value between 0 and 1 and can be adjusted by cross-validation.

PLS-PLR technique was selected and implemented to resolve the bias, overfitting, data separation, and multicollinearity problems. Dimensionality reduction by PLS eliminates perfect or quasi-perfect correlation between predictors. Furthermore, it makes use of the output variable, which reduces the bias, and therefore avoids underfitting. A ridge penalty in logistic regression limits the growth of the regression coefficients, which reduces variance, avoids overfitting, and controls the effects of data separation.

Goodness of fit of the model

The goodness-of-fit measures used are described below.

Deviance difference—The deviance difference (DiffDeviance) is interpreted as a measure of the variation of the data explained by the model with predictors and the model without predictors (only the constant). This statistic has a chi-square distribution with degrees of freedom equal to the difference in the numbers of the model parameters. Thus, the null hypothesis will be rejected for the significance level α when $\text{DiffDeviance} > \chi^2_\alpha$, which is equivalent to the P value of the contrast being less than the fixed level of α (Hosmer et al., 1997).

$$\text{DiffDeviance} = 2LL_M(\beta) - 2LL_0(\beta)$$

where $LL_M(\beta)$ is the log likelihood of the model and $LL_0(\beta)$ is the log likelihood of the null model.

Pseudo R^2 —Pseudo R^2 tells us how well the model can explain/predict the dependent variable based on the independent variables. Several different values were calculated as detailed below:

- McFadden's R^2 is defined as one minus the ratio between the logarithms of the likelihood for the model with respect to the log likelihood for the intercept only model (null model), with its theoretical range of values being $0 \leq \text{McFadden's } R^2 \leq 1$. It is usually considered a good quality of fit when $0.2 \leq \text{McFadden's } R^2 \leq 0.4$ and higher values show an excellent fit.

$$R^2_{\text{McFadden}} = 1 - \left(\frac{LL_M}{LL_0} \right)$$

where LL_M is the log likelihood of the model and LL_0 is the log likelihood of the null model.

- Cox and Snell's R^2 is a goodness-of-fit measure that generalizes the R^2 of the linear regression. It is based on the comparison of the likelihood of the model (L_M) with the likelihood of the null model (L_0). Its range of values is between 0 and $(1 - L_0)^{2/n}$.

$$R^2_{\text{Cox\&Snell}} = 1 - \left(\frac{L_0}{L_M} \right)^{\frac{1}{n}}$$

where L_M is the likelihood of the model and L_0 is the likelihood of the null model.

- Nagelkerke's R^2 is the value of Cox and Snell's R^2 that is standardized based on the maximum value it could take. The maximum value of this pseudo R^2 is therefore 1 (Allison, 2014; Walker and Smith, 2016).

$$R^2_{\text{Nagelkerke}} = \frac{R^2_{\text{Cox\&Snell}}}{1 - (L_0)^{\frac{1}{n}}}$$

where L_M is the likelihood of the model and L_0 is the likelihood of the null model.

Hyperspectral biplot (HS biplot)—Technically a biplot is a decomposition of a matrix X in the product of two low-rank (usually two or three) matrices and an error matrix.

$$X \cong TP + E$$

in such a way that rows and columns can be jointly represented on a scatter diagram using T and P as markers, respectively. In this case, we use the factorization obtained from the logistic PLS regression as a biplot representation. The biplot shows the directions of the space spanned by the columns of X that better separate the presences and absences for the dependent variable.

So, we have a low-rank approximation $\hat{X} = TP$ of the matrix X that captures the part that better explains the response. The percent of variability captured by the approximation is

$$\rho^2 = \frac{\text{tr}(\hat{X}^T \hat{X})}{\text{tr}(X^T X)} \times 100$$

It is possible to identify the variables related to the PLS components calculating the amount of variance of each column captured by the approximation as follows:



$$\rho_j^2 = \frac{\text{tr}(\hat{x}_{[j]}^T \hat{x}_{[j]})}{\text{tr}(x_{[j]}^T x_{[j]})} \times 100 \quad (j = 1, \dots, p)$$

where $\hat{x}_{[j]}$ and $x_{[j]}$ are the j th columns of the fitted and original matrix, respectively. Only the columns with high percentages are related to the response. These quantities are called contributions of the components to the variables or predictiveness.

The row scores are also used to predict the binary response in step 4 of the algorithm, then the binary variable can also be projected on the biplot using an external logistic biplot, proposed by Demey et al. (2008) based on the proposal of Vicente-Villardón et al. (2006). The main difference is that, in the original proposal, the scores for the individuals are obtained from the principal coordinates, whereas here they are obtained from the logistic PLS regression. The vector c of logistic regression coefficients defines the direction in the space spanned by the columns of T , which better separates the presences and absences and the expected probabilities of having the presence of the characteristic

$$\text{logit}(\hat{y}_i) = \text{logit}(\hat{p}_i) = c_0 + t_{[i]}c,$$

where $t_{[i]}$ is the i th row of T . The expected probability is obtained by projecting the point $t_{[i]}$ onto the vector c . The point on that direction that predicts an expected probability of 0.5 in a two-dimensional biplot has the coordinates

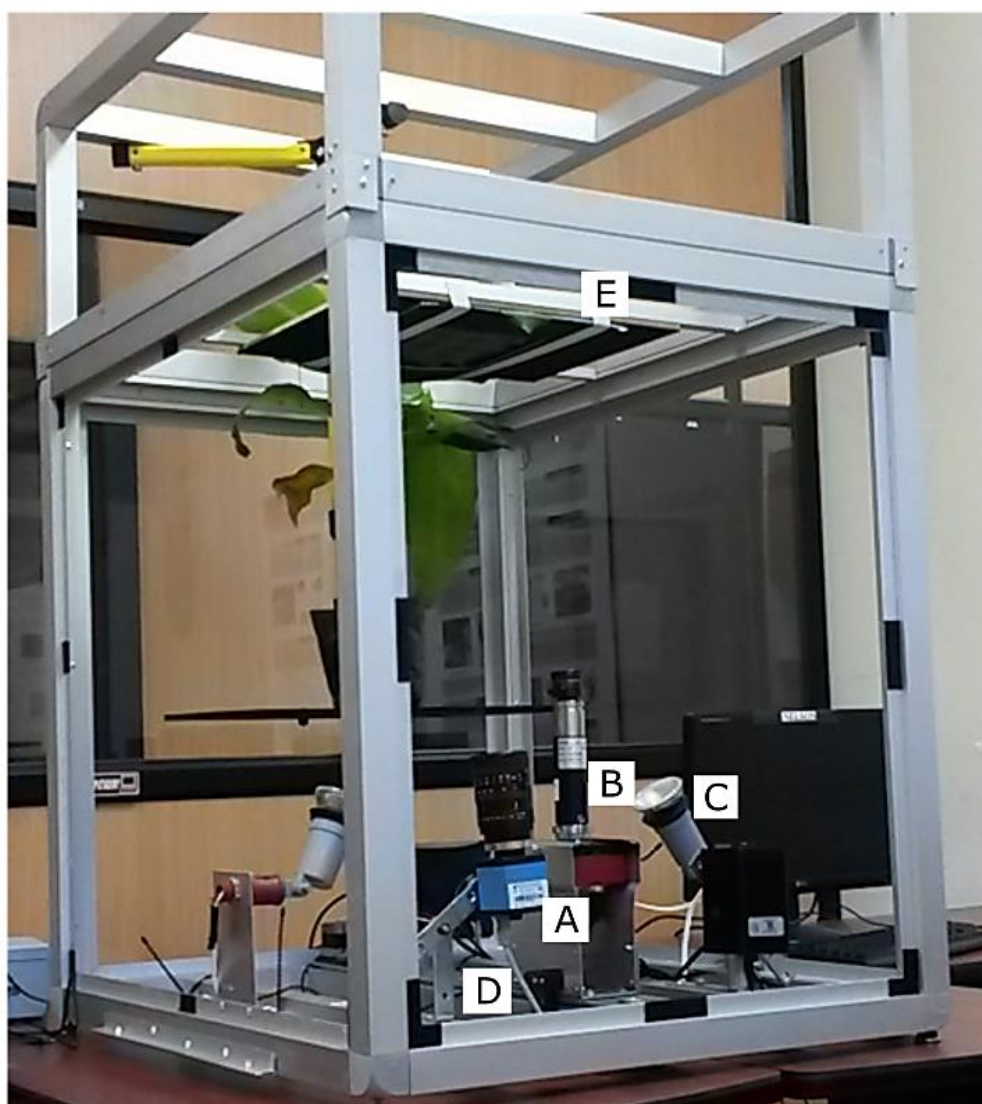
$$x = \frac{-c_0 c_1}{c_1^2 + c_2^2}; y = \frac{-c_0 c_2}{c_1^2 + c_2^2}$$

If we predict presence when the expected probability is greater than 0.5, the direction of c divides the representation into two regions, one predicting the presence and the other predicting absence. The boundary of the two regions is a straight line perpendicular to c and passing through the point (x, y) . For more details, see Demey et al. (2008) or Vicente-Villardón et al. (2006).

The goodness of fit of the logistic regression is measured using pseudo R^2 measures or the deviance. Although those measures are not completely adequate with the ridge penalizations, they can still be used as descriptive indicators equivalent to the contributions of the continuous biplot.

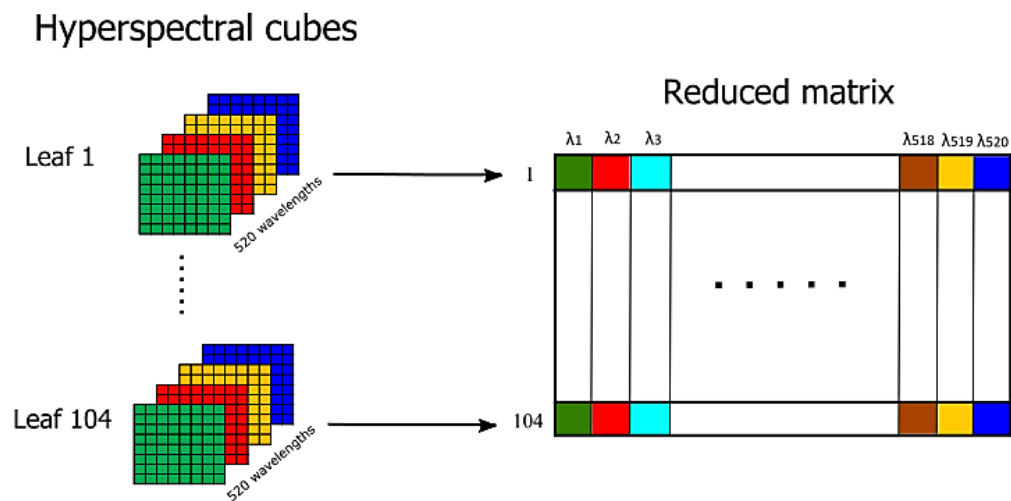
Ugarte Fajardo et al.—Applications in Plant Sciences 2020 8(8)—Data Supplement S1. Page 1 of 1.
DOI 10.1002/aps3.11383

APPENDIX S1. Hyperspectral imaging system used in this study: (A) camera, (B) spectrometer, (C) light source, (D) slider, (E) holder.



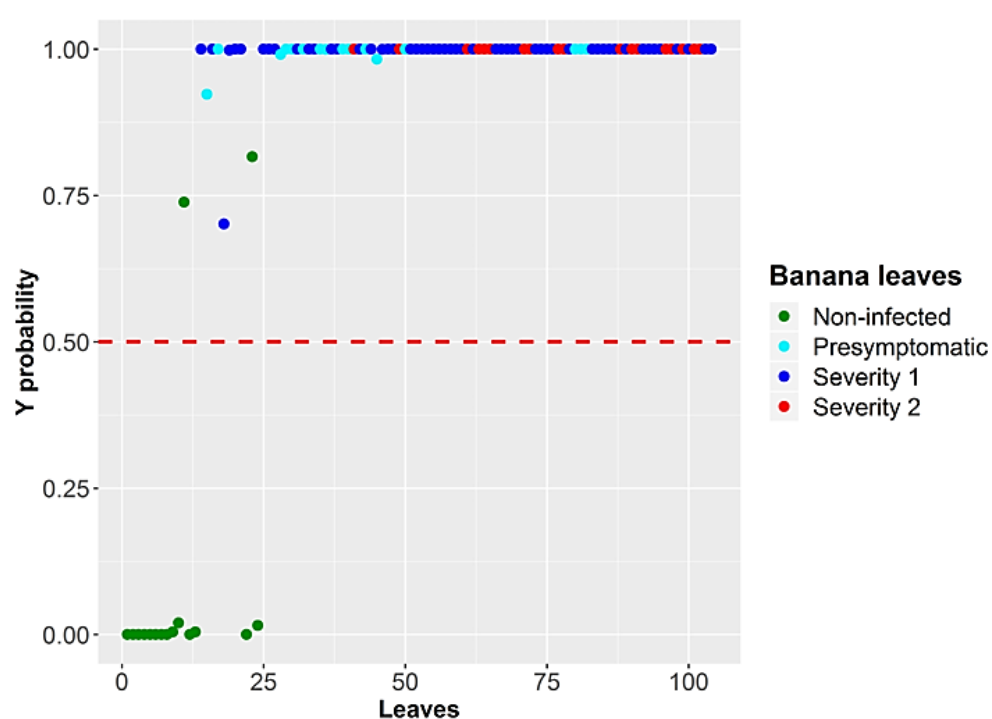
Ugarte Fajardo et al.—Applications in Plant Sciences 2020 8(8)—Data Supplement S2. Page 1 of 1.
DOI 10.1002/aps3.11383

APPENDIX S2. Dimensionality reduction of the hyperspectral cubes. The mean of the reflectance matrix at each wavelength is calculated.



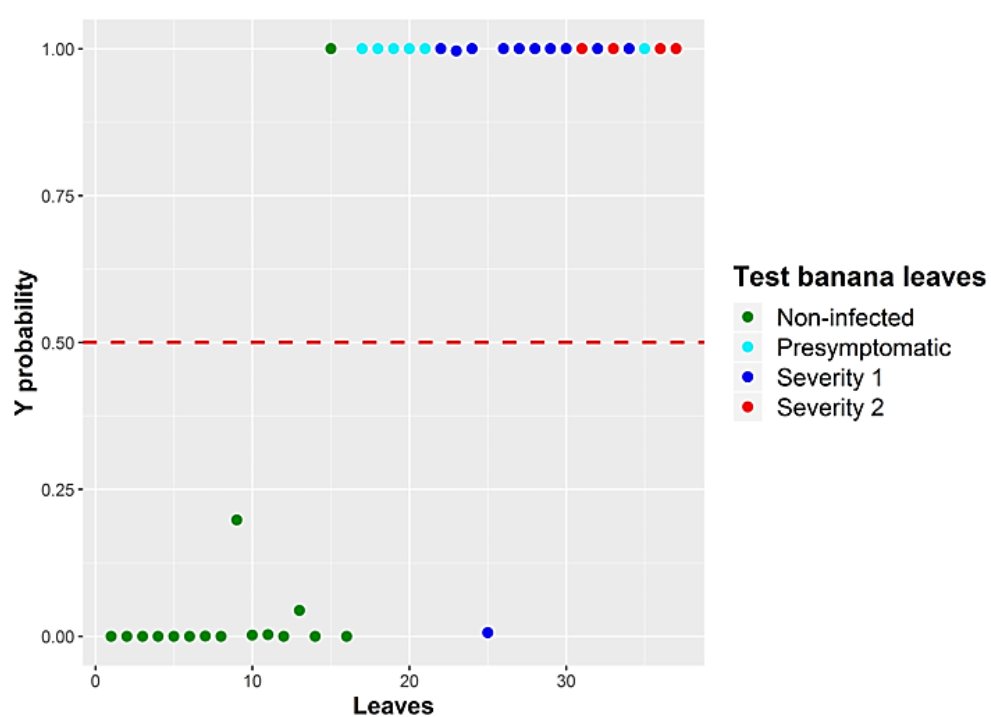
Ugarte Fajardo et al.—Applications in Plant Sciences 2020 8(8)—Data Supplement S3. Page 1 of 1.
DOI 10.1002/aps3.11383

APPENDIX S3. Predicted probability using leave-one-out cross-validation (LOOCV) to assess the PLS–PLR model.



Ugarte Fajardo et al.—Applications in Plant Sciences 2020 8(8)—Data Supplement S4. Page 1 of 1.
DOI 10.1002/aps3.11383

APPENDIX S4. Predicted probability for the external validation of the PLS–PLR model.



Applications in Plant Sciences

Published by Wiley on behalf of Botanical Society of America (the "Owner")

LICENSE AGREEMENT FOR PUBLISHING CC-BY-NC-ND

Date: August 10, 2020

Contributor name: Jorge Ugarte Fajardo

Contributor address:

Manuscript number: APPS-D-19-00179

Re: Manuscript entitled Early detection of black Sigatoka in banana leaves using hyperspectral images (the "Contribution")
for publication in Applications in Plant Sciences (the "Journal")
published by Wiley Periodicals LLC ("Wiley")

Dear Contributor(s):

Thank you for submitting your Contribution for publication. In order to expedite the editing and publishing process and enable Wiley to disseminate your Contribution to the fullest extent, we need to have this Agreement executed. If the Contribution is not accepted for publication, or if the Contribution is subsequently rejected, this Agreement will be null and void.

Publication cannot proceed without a signed copy of this Agreement and payment of the appropriate article publication charge.

A. TERMS OF USE

1. The Contribution will be made Open Access under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivatives License](#) which permits use, distribution and reproduction in any medium, provided that the Contribution is properly cited, the use is non-commercial and no modifications or adaptations are made.
2. For an understanding of what is meant by the terms of the Creative Commons License, please refer to [Wiley's Open Access Terms and Conditions](http://www.wileyauthors.com/OAA) (<http://www.wileyauthors.com/OAA>).
3. The Owner (and Wiley, where Wiley is not the Owner) reserves the right to require changes to the Contribution, including changes to the length of the Contribution, as a condition of acceptance. The Owner (and Wiley, where Wiley is not the Owner) reserves the right, notwithstanding acceptance, not to publish the Contribution if for any reason such publication would in the reasonable judgment of the Owner (and Wiley, where Wiley is not the Owner), result in legal liability or violation of journal ethical practices. If the Owner (or Wiley, where Wiley is not the Owner) decides not to publish the Contribution, no Article Processing Charge or any other fee shall be charged. The Contributor is free to submit the Contribution to any other journal from any other publisher.

B. RETAINED RIGHTS

The Contributor or, if applicable, the Contributor's Employer, retains all proprietary rights in addition to copyright, such as patent rights in any process, procedure or article of manufacture described in the Contribution.

APÉNDICE D

**Poster presentado en el IV INTERNATIONAL
WORKSHOP ON PROXIMITY DATA,
MULTIVARIATE ANALYSIS AND
CLASSIFICATION Organizado por el Group
of Multivariate Analysis and Classification of
the SEIO, 2019.**

PLS Regression for Detection of Black Sigatoka using Hyperspectral Images

Jorge Ugarte F.¹, Daniel Ochoa D.¹, Ronald Criollo B.¹, José Luis Vicente Villardón²

¹Escuela Superior Politécnica del Litoral, ESPOL, Ecuador

²Salamanca University, Spain

INTRODUCTION

Black Sigatoka is the disease of banana (Fig.1) most devastating in many country around the world. Currently, it's treated with chemical fungicides, which increase cost, detract the quality of the fruit and produce a negative impact on the environment.

Hyperspectral imaging (HSI) capture the spectral data for each pixel en the image, providing a "data cube" of entire image with a specific spectral signature (Fig 2). The changes in the leaf, produced by the disease, cause changes in the reflectance in different spectral regions.

The following issues are likely to be present when using HSI :

- (1)high collinearity in the adjacent bands,
- (2)variability of hyperspectral signatures, and
- (3)high dimensionality.

In this work, Logistic PLS is used in combination with HS-Biplot to classify healthy and infected leaves and to interpret the inter e intrarelations of individuals groups.



Fig. 1 Black Sigatoka stages

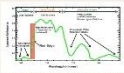


Fig. 2 Spectral response pattern of healthy vegetation

MATERIALS



Fig. 3 Hyperspectral imaging system CVR Laboratory - ESPOL - Ecuador

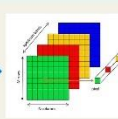


Fig. 4 Hyperspectral cube

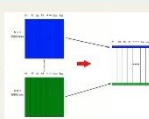


Fig. 5 Dimensionality reduction of data cube

- 1.- **HSI system**, includes a spectrometer ImSpector V10E connected to a camera 1500M-GE. The camera is mounted on a slider controlled by a computer.
- 2.- **Data**, 104 images of leaves of Cavendish banana: 16 of healthy leaves (control) and 88 of infected leaves in initial stages of the disease, 16 asymptomatic inoculate leaves, 54 severity 1 infected leaves and 18 severity 2 infected leaves.
- 3.- **Software**, Multiplot R program made by J.L. Vicente-Villardón, Ph.D., PLS prediction programs, HS-Biplot and complementary test programs made by J. Ugarte Fajardo.

RESULTS

PLS Good-of-fit

We fit the model and validate it using cross-validation. With penalty (λ) equal 0.1, we get the best goodness-of-fit measures.

Prediction procedure and model accuracy

Using Leave-One-Out-Cross-Validation (LOOCV), conditional Probability was estimated for each observation.

Observed	True	False	True	False
Healthy	16	2	18	0
Infected	0	12	0	1
Asymptomatic	16	0	16	0
Severity 1	54	0	54	0
Severity 2	18	0	18	0

Table 2 Confusion table.

The model accuracy was 0.98
Positive precision was 0.88
Sensitivity or recall value was 1

HS-Biplot

The HS-Biplot graphic (Fig. 7) shows:

- The threshold line separates the healthy and infected leaves.
- The infected leaves are into 2 groups.

The first sick leaves group is located near healthy leaves (Fig 9), indicating similar characteristics mainly influenced by wavelengths of the visible spectrum due to a low presence of disease symptoms and the presence of large healthy areas on the sheet.

In the second sick leaves group, asymptomatic leaves (turquoise) have a strong relationship with wavelengths near-infrared spectrum due as have very low or nothing presence of visible symptoms. Among infected leaves with severity level 1 (blue) and 2 (red) show that some of them are located in the near-infrared, which indicates a strong presence of no visible symptoms of the disease. While others are present on visible zone due to higher visible evidence of the infection.

- Global goodness-of-fit resulting is 77.07%.

A	DIF-Deviance	R ² CoxSnell	R ² Nagelkerke	R ² MacFadden
0.1	88.48775	0.5729456	0.5942404	0.6000133

Table 1 Logistic PLS good-of-fit measures

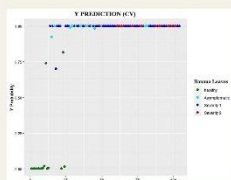


Fig. 6 Y Prediction (CV) Results of cross-validation process LOOCV

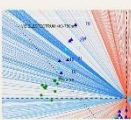


Fig. 9 HS-Biplot section show leaves low presence of disease

METHODS

Preprocessing and Data Transformation.

- 1.- Distortion reduction caused by the system: radiometric calibration, spectral calibration and spatial.
- 2.- Data cube standardization, using Standard Normal Variate (SNV).
- 3.- Dimensionality reduction of SNV cubes. Each matrix in the cube is transformed in a column and mean is calculated on each them obtaining one row for each cube.

PLS Logistic Regression.

We use here the algorithm developed by Bastien et al. (2015) with some improvements.

$$\logit(\hat{y}) = \log\left(\frac{\hat{y}}{1-\hat{y}}\right) = c_0 + \sum_{i=1}^m c_i t_i$$

$$\logit(\hat{y}) = C_0 + Tc = C_0 + XWc = C_0 + Xb$$

Where:

$\hat{y}$ is the vector of fitted probabilities of presence for each individual.
 T are the components mutually orthogonal.
 c the coefficients of the regression on the components.
 W and b are coefficients on X variables.

The graphic of the PLS model

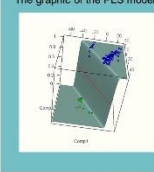


Fig. 8 Banana PLS 3D Model with 2 components

HS-Biplot (Hyperspectral Biplot)

We use the factorization obtained from the Logistic PLS Regression: $X = TP$

X-Scores (T matrix) are coordinates of the rows

X-Loadings (P matrix) provide the direction of variables of the original X matrix.

The original variables have been colored according to the spectral band to which it belongs.

Wavelengths of the Visible Spectrum (between 380x10-9m and 780x10-9m) are blue.

Wavelengths of the Near Infrared (> 780x10-9m) are red.

Two regions are separated by the classification threshold line $Px = 0.5$. If the value is above that threshold indicates "infected"; if it's below indicates "healthy".

The cosine of the angle formed by vectors representing the variables, estimate the correlation between them.
The distance between row markers represents the similarity or dissimilarity between individuals.

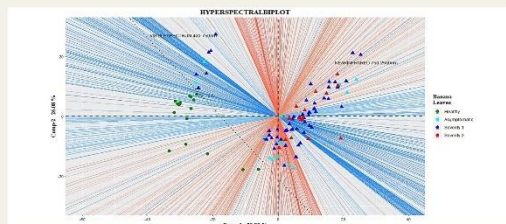


Fig. 7 HS-Biplot

CONCLUSIONS

- Hyperspectral Imaging provides non-destructive sampling that allows analysis of plants infected with the black Sigatoka, and probably other pathogens.
- Logistic PLS Regression together with the visual representation in a HS-Biplot are promising techniques to analyze hyperspectral data even considering the high reduction of the amount of data after preprocessing the raw observations.
- Considering that the sample includes leaves images in initial stages, the results show that the applied technique classifies correctly leaves in the early stages of infection black Sigatoka and confirms that during the initial stages changes in the structure of the leaves are detected by spectrometers HSI.

REFERENCES

- Mahlein, A., 2010. Detection, identification and quantification of fungal diseases of sugar beet leaves using imaging and non-imaging hyperspectral techniques. Rheinischen Friedrich Wilhelms Universität.
- Castro, R., Ochoa, D., Criollo, R., 2017. On the influence of spectral calibration in hyperspectral image classification of leaves. A. 2017 CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies, CHILECON 2017 - Proceedings, Vol. 2017 Janua, p. 1-6. DOI 10.1109/CHILECON.2017.8229687.
- Bastien, P., Espósito, V., Tenenhaus, M., 2005. PLS generalised linear regression. A. Computational Statistics and Data Analysis, Vol. 48, p. 17-46. DOI 10.1016/j.csda.2004.02.005
- Villardón, J.L.V. et al., 2006. Logistic biplots. A. Multiple correspondence analysis and related methods, London: Chapman & Hall [on line], núm. March, p. 503-521. DOI 10.1201/9781420011319.ch23. Disponible a: [http://biplot.usal.es/DOCTORADO/CI/CILO/BENIO-08-08/Logistic Biplots final.pdf](http://biplot.usal.es/DOCTORADO/CI/CILO/BENIO-08-08/Logistic%20Biplot%20.pdf) http://www.academia.edu/2135984/Logistic_biplots.
- Siehe, R. et al., 2016. Evaluation of Food Quality and Safety with Hyperspectral Imaging (HSI). A. Food Engineering Reviews, Vol. 8, núm. 3, p. 306-322. ISSN 18667929. DOI 10.1007/s12393-015-9137-8.