PLS-PLR

El modelo predictivo PLS-PLR de BLSD se construyó utilizando dataset de entrenamiento 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.

Algoritmo PLS-PLR

Sea $\boldsymbol{Y}$ la variable respuesta binaria y $(\boldsymbol{X_1}, \dots, \boldsymbol{X_p})$ un conjunto de predictores. Para una muestra de tamaño n los datos pueden ser organizados en un vector respuesta $\boldsymbol{y} = (y_1, \dots, y_n)^T$ y una matriz de predictores $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_p) = (x_{ij})$ $(i = 1, \dots n; j = 1, \dots, p)$, 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 $\boldsymbol{X}$ se suponen centradas y escaladas.

La ecuación de regresión de la variable dependiente $\boldsymbol{y} = \boldsymbol{Tc} + \boldsymbol{F}$ en el algoritmo PLS que explica $\boldsymbol{y}$ desde las componentes $\boldsymbol{T}$ es adaptada para tomar en cuenta la respuesta binaria. Bastien et al. (2005) generaliza el método cuando la respuesta desde una familia exponencial usando $g(\hat{\boldsymbol{y}}) = \boldsymbol{Tc}$ y propone un algoritmo para estimar los parámetros. Nosotros incluimos una constante en el modelo porque la variable binaria no puede ser centrada.

$$g(\hat{\boldsymbol{y}}) = c_0 + \boldsymbol{Tc} \qquad (1)$$

El modelo Regresión Logística PLS es escrito como:

$$E(\boldsymbol{y}) = \hat{\boldsymbol{y}} = \frac{1}{\mathbf{1}+\boldsymbol{e}^{-\left(c_0\mathbf{1}+\sum_{h=1}^{m} c_h \boldsymbol{t_h}\right)}} \tag{2}$$

o

$$logit(\hat{\boldsymbol{y}}) = log\left(\frac{\hat{y}}{\mathbf{1}-\hat{y}}\right) = c_0\mathbf{1} + \sum_{h=1}^{m} c_h \boldsymbol{t}_h \tag{3}$$

O en forma matricial $logit(\hat{\boldsymbol{y}}) = c_0\mathbf{1} + \boldsymbol{Tc}$, donde $\hat{\boldsymbol{y}} = (\hat{y}_1, \dots, \hat{y}_n)^T$ es el vector de probabilidades estimadas de la presencia en cada individuo y $\boldsymbol{c} = (c_1, \dots, c_m)^T$ son los coeficientes de la regresión sobre las componentes.

En términos de las variables originales,

$$logit(\hat{\boldsymbol{y}}) = c_0\mathbf{1} + \boldsymbol{Tc} = c_0\mathbf{1} + \boldsymbol{XWc} = c_0\mathbf{1} + \boldsymbol{Xb} \tag{4}$$

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

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

1. Calculo de $\boldsymbol{t_1}$, 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 $\boldsymbol{y}$ sobre $\boldsymbol{x_j}$, para obtener $\boldsymbol{w_1} = (w_{11}, \dots, w_{1p})^T$
    b. Normalice el vector $\boldsymbol{w_1} := \boldsymbol{w_1}/\|\boldsymbol{w_1}\|$.
    c. Calcule las puntuaciones de la componente (scores) $\boldsymbol{t_1} = \boldsymbol{Xw_1}/\boldsymbol{w_1^T w_1}$
2. Calculo de $\boldsymbol{t_h}$, la h^{th} componente PLS. Las componentes $\boldsymbol{t_1}, \dots, \boldsymbol{t_{h-1}}$ han sido obtenidas.

a. Para cada predictor ($j = 1, \dots, p$), calcule el coeficiente de regresión w_{hj} de $\boldsymbol{x_j}$, en la regresión logística de $\boldsymbol{y}$ sobre $\boldsymbol{t_1}, \dots, \boldsymbol{t_{h-1}}$ y $\boldsymbol{x_j}$, para obtener $\boldsymbol{w_h} = (w_{h1}, \dots, w_{hp})^T$

b. Normalice el vector $\boldsymbol{w_h} := \boldsymbol{w_h}/\|\boldsymbol{w_h}\|$.

c. Calcule la matriz residual $\boldsymbol{X_{h-1}}$ de la regresión lineal de $\boldsymbol{X}$ sobre $\boldsymbol{t_1}, \dots, \boldsymbol{t_{h-1}}$.

d. Calcule las puntuaciones de la componente (scores) $\boldsymbol{t_h} = \boldsymbol{X w_h}/\boldsymbol{w_h^T w_h}$.

3. $\boldsymbol{X}$ es factorizado como $\boldsymbol{X} = \boldsymbol{TP}$

4. Regresión Logística de $\boldsymbol{y}$ sobre las componentes PLS retenidas

$$logit(\hat{y}) = c_0 1 + \sum_{h=1}^{m} c_h \boldsymbol{t_h}$$

5. Expresión del modelo en términos de las predictoras originales $b = W$.

$$logit(\hat{\boldsymbol{y}}) = c_0 \mathbf{1} + \boldsymbol{Tc} = c_0 \mathbf{1} + \boldsymbol{XWc} = c_0 \mathbf{1} + \boldsymbol{Xb}$$

Penalización Ridge

La función log-verosimilitud con penalización Ridge es,

$$L_{ridge}(\boldsymbol{\beta}) = L(\boldsymbol{\beta}) - \lambda \|\boldsymbol{\beta}\|^2 \tag{5}$$

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

(6)

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

Donde,

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

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

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.

Diferencia de Devianza

$$D_0 - D_M = \left(-2ln\frac{L_0(\boldsymbol{\beta})}{L_F(\boldsymbol{\beta})}\right) - \left(-2ln\frac{L_M(\boldsymbol{\beta})}{L_F(\boldsymbol{\beta})}\right) \quad (7)$$

$$D_0 - D_M = -2\left(ln\frac{L_0(\boldsymbol{\beta})}{L_F(\boldsymbol{\beta})} - ln\frac{L_M(\boldsymbol{\beta})}{L_F(\boldsymbol{\beta})}\right) \quad (8)$$

$$D_0 - D_M = -2ln\left(\frac{L_0(\boldsymbol{\beta})}{L_M(\boldsymbol{\beta})}\right) \quad (9)$$

$$DiffDeviance = -(2LL_0(\boldsymbol{\beta}) - 2LL_M(\boldsymbol{\beta})) \quad (10)$$

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 Chi^2 ($\chi^2_{s_M - s_0}$), 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

R^2McFadden

$$R^2{}_{McFadden} = 1 - \left(\frac{LL_M}{LL_0}\right) \tag{11}$$

Donde:

LL_M es log verosimilitud del modelo

LL_0 es log verosimilitud de modelo nulo (null model)

R^2Cox&Snell

$$R^2{}_{\boldsymbol{Cox\&Snell}} = 1 - \left(\frac{L_0}{L_M}\right)^{\frac{2}{n}} \tag{12}$$

Donde

L_M es el valor de verosimilitud del modelo.

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

R^2Nagelkerke

$$R^2{}_{\boldsymbol{Nagelkerke}} = \frac{R^2{}_{Cox\&Snell}}{1-(L_0)^{\frac{2}{n}}} \quad (13)$$

Donde,

L_M es el valor de verosimilitud del modelo.

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

Selección de parámetro de penalización Ridge (λ)

Tabla 1 Métricas de bondad de ajuste para modelos PLS-PLR con valores incrementales de penalización Ridge (λ)

λ	*Diff-Deviance*[a]	R^2*CoxSnell*[b]	R^2*Nagelkerke*[c]	R^2*MacFadden*[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^2Cox&Snell, [c] R^2Nagelkerke and [d] R^2MacFadden 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.

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

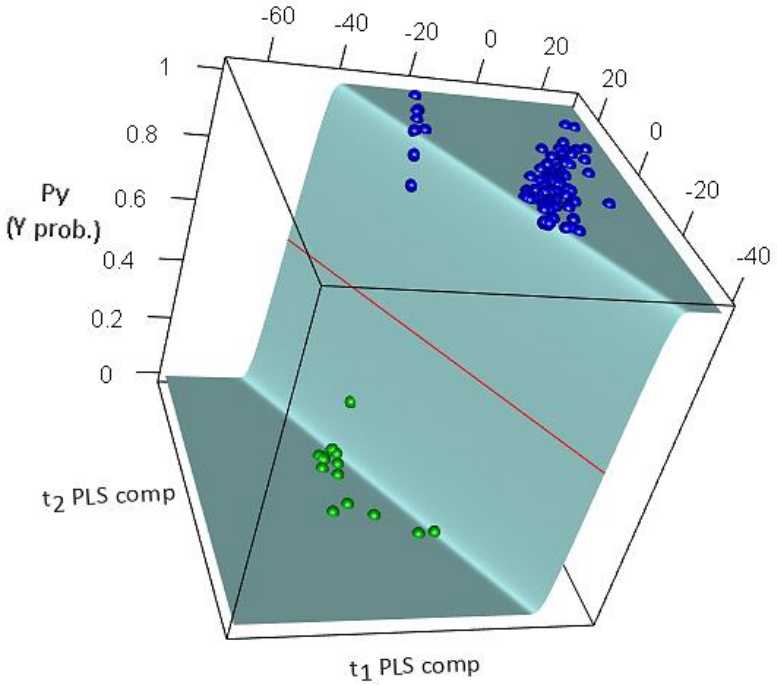

Figura 1 Gráfico 3D de la respuesta del modelo PLS-PLR.

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).

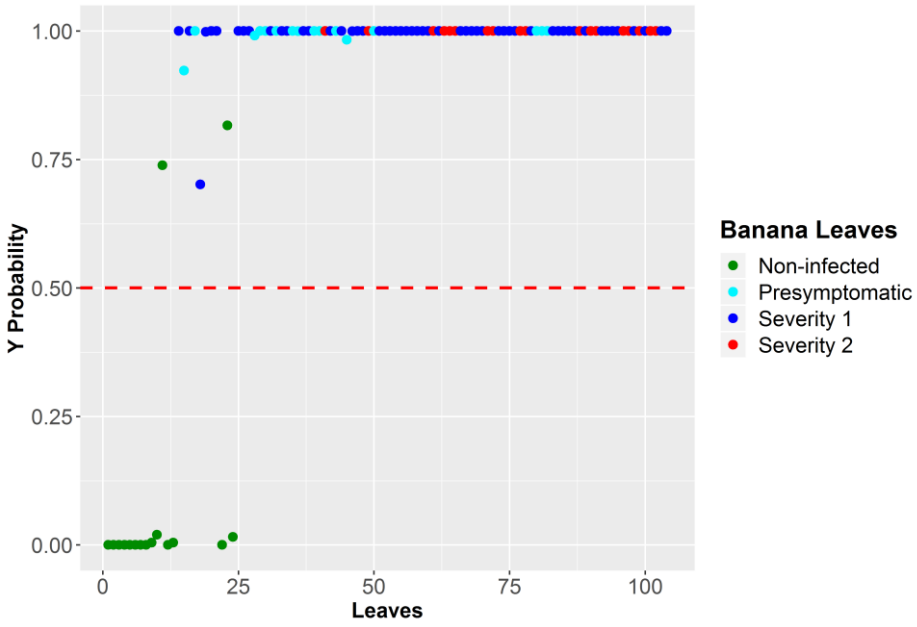

Figura 2 Probabilidad estimada por PLS-PLR con validación cruzada.

Matriz de confusión

Tabla 2 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

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.

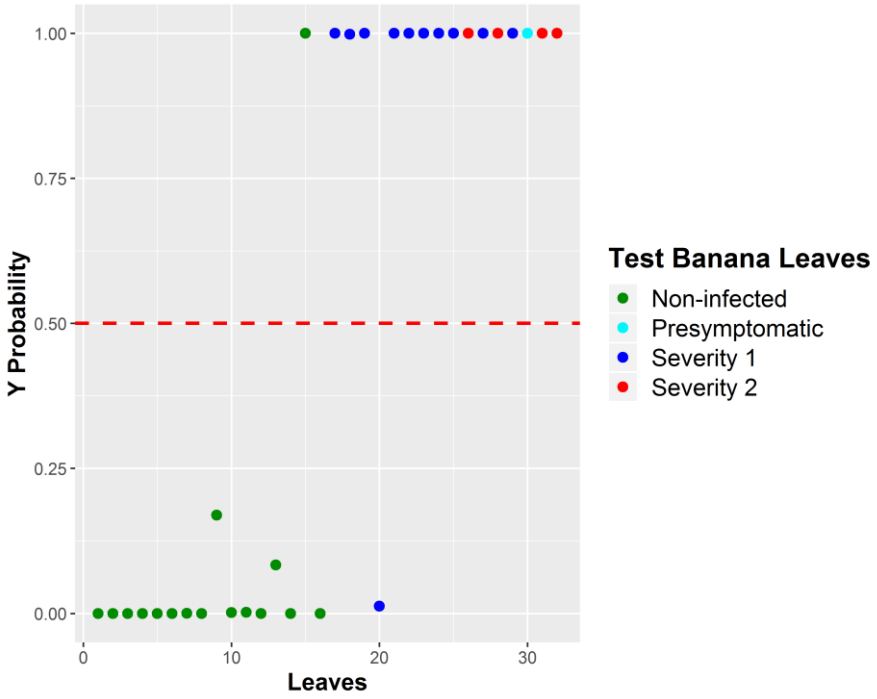

Figura 3 Probabilidad estimada por el modelo PLS-PLR en prueba de validación.

Tabla **¡Error! No hay texto con el estilo especificado en el documento.** 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

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.

HS-BIPLOT

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.

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

Las dos primeras componentes del modelo PLS-PLR fueron utilizadas para graficar el HS-Biplot. Las puntuaciones filas $\boldsymbol{T}$ (scores) y las puntuaciones columna $\boldsymbol{P}$ (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.

$$\boldsymbol{X} \cong \boldsymbol{TP} + \boldsymbol{E}$$

El porcentaje de variabilidad capturada por la aproximación es,

$$\rho^2 = \frac{tr(\hat{\boldsymbol{X}}^{\boldsymbol{T}}\hat{\boldsymbol{X}})}{tr(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})} \times 100 \qquad (14)$$

Donde $\widehat{\boldsymbol{X}}$ es la matriz de predictores aproximada y $\boldsymbol{X}$ es la matriz original.

Es posible identificar las variables relacionadas a las componentes PLS calculando bondad de ajuste por columna

$$\rho_j^2 = \frac{tr(\hat{x}_{[,j]}^T \hat{x}_{[,j]})}{tr(x_{[,j]}^T x_{[,j]})} \times 100 \ (j = 1, \dots, p) \qquad (15)$$

Donde $\widehat{\boldsymbol{x}}_{[,\boldsymbol{j}]}$ y $\boldsymbol{x}_{[,\boldsymbol{j}]}$ son el j^{th} columnas de la matriz ajustada y de la matriz original respectivamente.

La bondad de ajuste de cada fila es,

$$\rho_i^2 = \frac{tr(\hat{\boldsymbol{x}}_{[,\boldsymbol{i}]}\hat{\boldsymbol{x}}_{[,\boldsymbol{i}]}^{\boldsymbol{T}})}{tr(\boldsymbol{x}_{[,\boldsymbol{i}]}\boldsymbol{x}_{[,\boldsymbol{i}]}^{\boldsymbol{T}})} \times 100 \ (i = 1, \dots, n) \qquad (16)$$

Donde $\widehat{\boldsymbol{x}}_{[,\boldsymbol{i}]}$ y $\boldsymbol{x}_{[,\boldsymbol{i}]}$ 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.

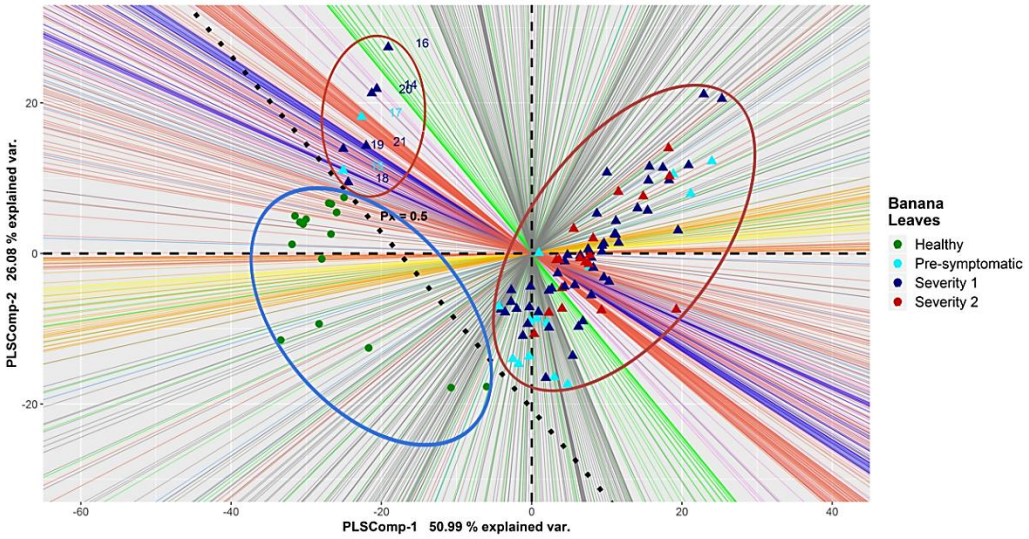

Figura 4 HS-Biplot del dataset de entrenamiento.

Las filas de la matriz reducida (hojas de banano) están representadas por puntos y las columnas (longitudes de onda) están representadas por líneas rectas.

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.

Se observó exactamente el mismo comportamiento en el conjunto de datos de validación externa.

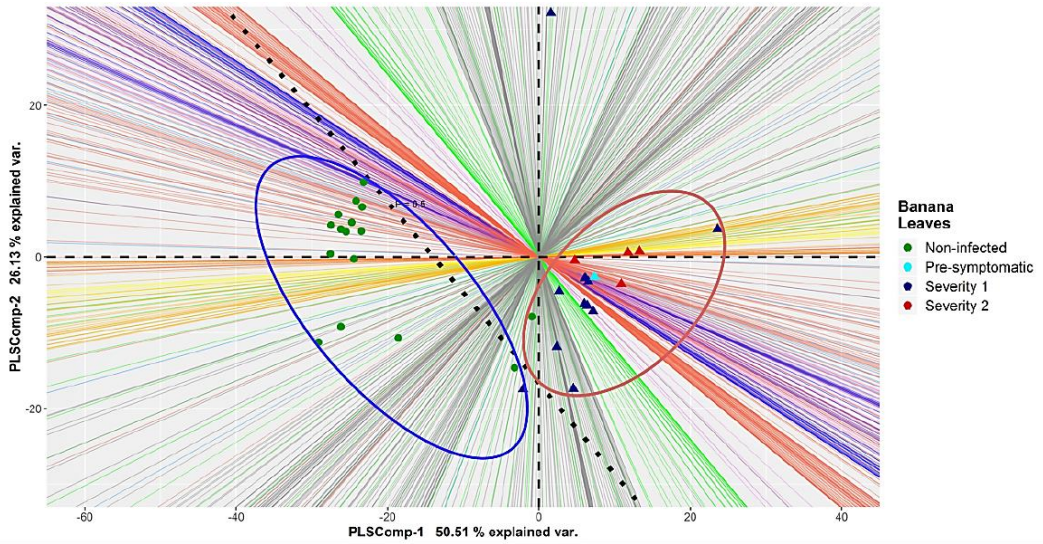

Figura 5 HS-Biplot del dataset de validación.

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

$$\rho^2 = \frac{tr(\widehat{\boldsymbol{X}}^{\boldsymbol{T}}\widehat{\boldsymbol{X}})}{tr(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})} \times 100 = 77.07$$

Donde

$\widehat{\boldsymbol{X}} = \boldsymbol{T}\boldsymbol{P}^{\boldsymbol{T}}$ es la matriz de datos aproximada.

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

La contribución de las componentes a cada variable

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

ρ_j^2 es la contribución de cada columna *j*

$\widehat{\boldsymbol{x}}_{[,j]}$ la columna *j* de la matriz aproximada $\widehat{\boldsymbol{X}}$

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

La contribución de las componentes a cada fila o individuo

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

ρ_i^2 es la contribución de cada columna *i*

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

$\boldsymbol{x}_{[i,]}$ es la fila *i* de la matriz original $\boldsymbol{X}$