



VIII Congreso Nacional de Ingeniería Física y
III International Applied Physics, Engineering and Innovation Conference

Popayán, Colombia — 20–24 de octubre de 2025

Comparación de Modelos de Clasificación a partir de datos Hiperespectrales usando aprendizaje Automático en Plantas de Banano

María Liliana Parra Osorno.

Marcela Ruiz Guzmán.

mlparrao@unal.edu.co

mruizg@unal.edu.co

Departamento de Estadística, Universidad Nacional de Colombia, sede Medellín, Colombia

I. RESUMEN

El **monitoreo del estado de salud en plantas** es esencial para el manejo agrícola sostenible y la **detección temprana de enfermedades**. En este trabajo exploramos la aplicación de técnicas de *aprendizaje automático* para clasificar el estado fisiológico de plantas de banano (*Musa spp.*) usando firmas espectrales en el rango de 350 a 2500 nm. El estudio se enfoca en distinguir entre dos condiciones: plantas sanas y afectadas por enfermedades (*Ralstonia*, *Fusarium*, entre otras) [1].

El análisis parte de una base de datos que contiene reflectancias espectrales por planta, información del tratamiento aplicado y la etiqueta de salud: sana (1), enferma (0) y estrés hídrico (-1). Mediante visualizaciones basadas en las curvas promedio por tramo espectral (VIS: 350–700 nm, NIR: 701–1300 nm, SWIR: 1301–2500 nm), en donde se evidenció una mayor similitud con las plantas sanas, se reclasificó el estrés hídrico como parte de la clase sana, simplificando la tarea a un problema de clasificación binaria.

Se aplicaron diferentes modelos supervisados como *Random Forest*, *XGBoost*, *LightGBM* y *CatBoost* [2]. Estos fueron integrados en pipelines con estandarización, entrenamiento y evaluación mediante validación cruzada. La métrica principal fue el **macro F1-score**, útil en contextos con desbalance de clases. Los mejores resultados fueron obtenidos con *CatBoost*, alcanzando una precisión del **98.3 %** y un macro F1-score de 0.980. La matriz de confusión evidenció una mínima confusión entre clases.

Adicionalmente, tras identificar si una planta está enferma, se plantea el entrenamiento de un segundo modelo exclusivamente con las muestras enfermas para predecir el tipo de enfermedad (*multi-clase*). Esta arquitectura permitiría diseñar un sistema jerárquico de clasificación [3], altamente útil para **agricultura de precisión**.

Este trabajo demuestra que la combinación de firmas hiperespectrales y algoritmos avanzados de clasificación ofrece una vía prometedora para la **detección temprana de condiciones fitopatológicas**.

Como ejercicio académico, los resultados abren puertas para futuras implementaciones en sistemas de **monitoreo automatizado en campo**, mejorando la toma de decisiones en el manejo fitosanitario del cultivo de banano.

Palabras clave: Aprendizaje automático; Imágenes hiperspectrales; Clasificación supervisada; Plantas de banano; CatBoost; Fusarium; Estrés hídrico.



(a) Planta sana



(b) Planta con enfermedad

Figura 1: Comparación visual entre una hoja de plátano sana [4] y una afectada por Fusarium raza 4 [5].

II. GRÁFICAS DEL RESUMEN

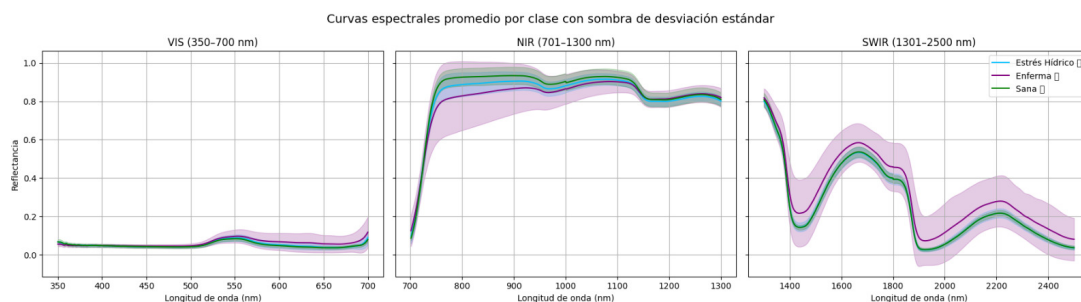


Figura 2: Curvas Espectrales Promedio por clase

- En el **VIS (350–700 nm)**: la reflectancia de estrés hídrico fue prácticamente igual a la de las sanas.
- En el **NIR (701–1300 nm)**: la estructura celular (mesófilo) tampoco mostró diferencias relevantes con las plantas sanas.
- En el **SWIR (1301–2500 nm)**: aunque se observó menor contenido de agua, esto no se relaciona con enfermedades infecciosas, sino con una condición reversible.

El tramo VIS (350–700 nm) está asociado con la captura de la luz por los pigmentos como la clorofila, graficamente se ve que las curvas de las tres clases (sana, enferma y estrés hídrico) están muy juntas y con baja reflectancia. No hay diferencias muy notorias en este rango, lo cual sugiere que la pigmentación foliar (clorofila) no cambia demasiado entre clases, o al menos, no lo suficiente para separarlas espectralmente.

El tramo NIR (701–1300 nm), este rango refleja cómo la luz interactúa con la estructura interna de las hojas, se aprecia como ondas espectrales de las plantas enfermas tienen una reflectancia un poco menor que las sanas y las plantas con estrés hídrico siguen una forma más parecida a las sanas. Aquí se empiezan a ver diferencias estructurales. Las enfermas probablemente tienen daño celular que altera cómo se refleja la luz, mientras que el estrés hídrico, al no destruir la célula, se comporta más como una sana, aunque con pequeñas variaciones.

El tramo SWIR (1301–2500 nm) aquí la clase de estrés hídrico se separa más notablemente de las sanas, especialmente en los valles de absorción de agua (como 1450, 1950 y 2500 nm). Las plantas enfermas también muestran diferencias, pero el estrés hídrico baja más en ciertos puntos, indicando menor contenido de agua [6].

Cuadro 1: Comparación de desempeño entre algoritmos de clasificación

Modelo	accuracy	precision_0	recall_0	f1_0	precision_1	recall_1	f1_1	macro_f1
CatBoost	0.983	0.978	0.998	0.988	0.995	0.949	0.972	0.980
XGBoost	0.980	0.974	0.998	0.986	0.995	0.940	0.967	0.976
LightGBM	0.979	0.972	0.998	0.985	0.995	0.935	0.964	0.975
Random Forest	0.901	0.893	0.974	0.932	0.925	0.737	0.821	0.876

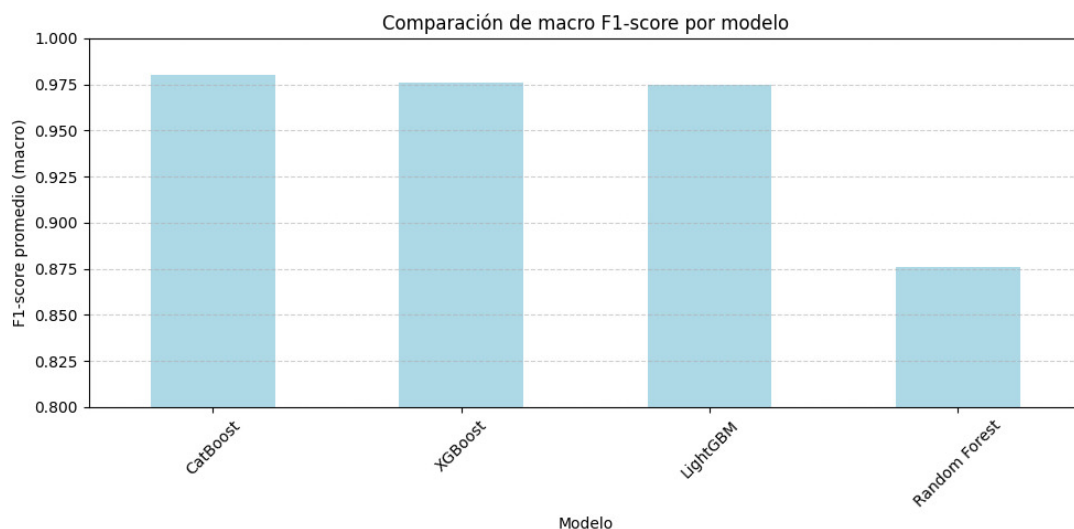


Figura 3: Comparación de F1 Score para modelos evaluados

La Figura 2 presenta una comparación del desempeño de cuatro algoritmos de clasificación supervisada aplicados al conjunto de datos hiperespectral de plantas de banano, utilizando el macro F1-score como métrica principal. Esta métrica es especialmente útil en contextos donde existe un cierto desequilibrio entre clases. Los tres primeros modelos (CatBoost, XGBoost y LightGBM), basados en técnicas de gradient boosting, obtuvieron un desempeño muy similar y superior, con macro F1-scores cercanos a 0.975, lo cual indica una excelente capacidad para clasificar correctamente tanto plantas sanas como enfermas. En contraste, el modelo Random Forest mostró un macro F1-score notablemente inferior, situándose alrededor de 0.875, lo que evidencia un menor rendimiento general comparado con los métodos de boosting.

Estos resultados resaltan la efectividad de los modelos de boosting en tareas de clasificación multiclase en contextos agrícolas, especialmente cuando se utilizan datos de alta dimensionalidad como los hiperespectrales.

REFERENCIAS

- [1] A. K. Mahlein, M. T. Kuska, J. Behmann, G. Polder, and A. Walter, “Hyperspectral imaging combined with machine learning for detecting plant diseases: A review,” *Biosystems Engineering*, vol. 174, pp. 203–215, 2018.
- [2] F. Gómez. (2023, junio) Genetic algorithms for feature selection in machine learning. Neural Designer. [Online]. Available: <https://www.neuraldesigner.com/blog/genetic-algorithms-for-feature-selection-in-machine-learning>
- [3] S. Ray. (2024, noviembre) Introduction to boosting algorithms in machine learning. Analytics Vidhya. [Online]. Available: <https://www.analyticsvidhya.com/blog/2024/11/introduction-to-boosting-algorithms-in-machine-learning>
- [4] jrperes, “Hoja sana de banano,” November 2014, fotografía disponible en Pixabay. De uso gratuito bajo la Licencia de contenido de Pixabay. [Online]. Available: <https://pixabay.com/es/photos/hoja-de-platano-banano-hoja-mojada-544208/>
- [5] ILGA Importadora. (2024, enero) Fusarium raza 4 en banano: qué es y cómo controlarlo. ILGA Importadora. Accedido el 13 de julio de 2025. [Online]. Available: <https://www.ilga.com.co/fusarium-raza-4-en-banano-que-es-y-como-controlarlo/>
- [6] A. Lowe, N. Harrison, and A. P. French, “Hyperspectral image analysis techniques for the detection and classification of the early onset of plant disease and stress,” *Plant Methods*, vol. 13, no. 80, 2017. [Online]. Available: <https://doi.org/10.1186/s13007-017-0233-z>