Informe de Avance: Procesamiento y Análisis de Datos

Ricardo Esteban Lopera Vasco

13 de noviembre de 2025

Índice

1. Introducción y Objetivos

El propósito de este informe es documentar la integridad y calidad de los datos recolectados para el proyecto, además de detallar los avances realizados en la producción de modelos y análisis preliminares.

Los objetivos específicos de este reporte son:

- Describir avances en los modelos y análisis realizados hasta la fecha.
- Presentar las métricas descriptivas preliminares obtenidas.
- Identificar y categorizar los errores sistemáticos en la captura de datos.

2. Descripción del Conjunto de Datos

La fuente de información consta de un conjunto de datos distribuido en 15 hojas, las cuales contienen entre 230 y 240 registros de plantas. El conjunto incluye 2150 variables correspondientes a longitudes de onda en el rango de 350 nm a 2500 nm. Para la realización del análisis se importaron las hojas como dataframes individuales.

Tabla 1: Diccionario de Variables Principales

Variable	**Tipo de Dato**	**Descripción**
Tratamiento	Cadena (str)	Nombre del estrés aplicado a la planta
$[350 - 2500$ nm]	Numérico (Float)	Indica la reflectancia en cada longitud de onda.
Planta	Numérico (Int[1:30])	Identificador único de cada planta.

Todas las muestras se dividen en 8 clases diferentes, las cuales representan distintos tipos de estrés aplicados a las plantas. El numero de muestras por clase son las siguientes:

Tabla 2: Distribución de Clases en el Conjunto de Datos

Clase	**Número de Muestras**
E_Hidrico	478
Control	461
Fusarium	448
Fus_EH_Ral	440
Ral_EH	436
Fus_EH	432
Ralstonia	428
Ral_Fus	412

3. Avances en el Procesamiento

Hasta la fecha, se han completado las fases de ingestión, procesamiento de datos y producción preliminar de modelos. A continuación, se detallan las observaciones clave:

3.1. PCA y Reducción de Dimensionalidad

Uno de los grandes problemas encontrados para el análisis y la producción de modelos fue la alta dimensionalidad del conjunto de datos. Se aplicó Análisis de Componentes Principales (PCA) para reducir la dimensionalidad y se encontró que las primeras 3 componentes explican en promedio el 96.68 % de la varianza total; sin embargo, hay dataframes donde el porcentaje de explicabilidad varía significativamente.

Tabla 3: Varianza total explicada por el PCA (3 componentes) para cada subconjunto de datos.

Modelo (DataFrame)	**Varianza Explicada (%)**
df0	95.03 %
df1	89.76 %
df2	97.47 %
df3	91.96 %
df4	90.82 %
df5	93.50 %
df6	95.76 %
df7	98.42 %
df8	99.28 %
df9	99.09 %
df10	99.34 %
df11	99.14 %
df12	99.40 %
df13	98.91 %
df14	99.52 %
df15	99.43 %

Lo anterior demuestra que, aun reduciendo significativamente el número de dimensiones, se preserva una alta proporción de la varianza original. Esto garantiza la representatividad de los datos reducidos, asegurando que el entrenamiento de los modelos subsiguientes sea robusto y consistente con la estructura subyacente de la información.

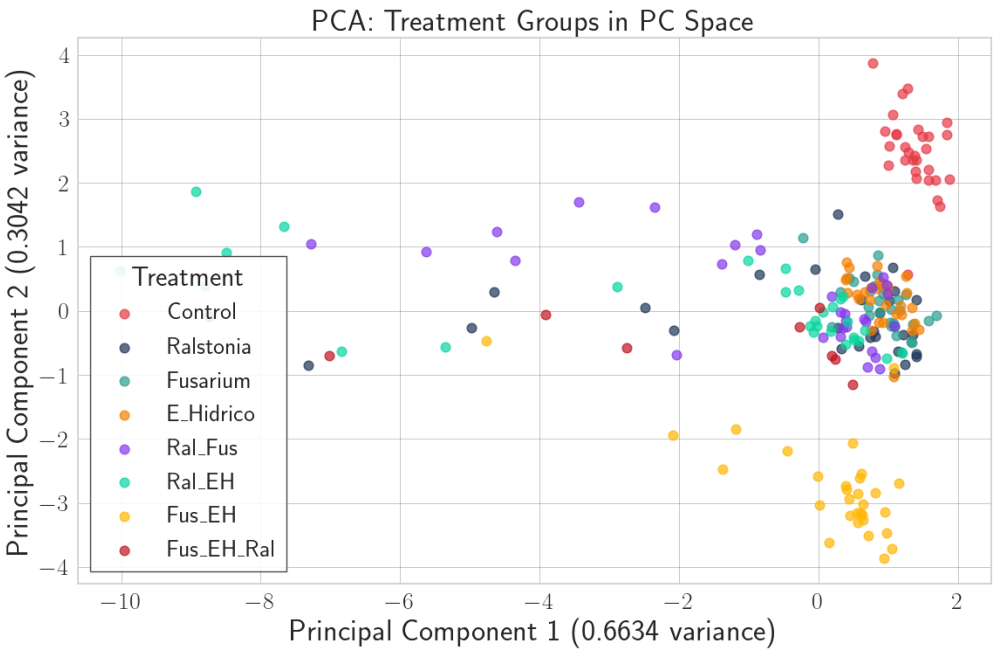

Figura 1: Resultados del PCA en el día 7

3.2. Balanceo algorítmico de datos vs Balanceo estadístico

El conjunto de datos presenta un desbalance de clases significativo, donde la clase minoritaria (plantas sanas) representa únicamente el 12 % del total de las muestras. Con el objetivo de mitigar este sesgo, se evaluaron estrategias de ponderación algorítmica (ajuste de pesos mediante el hiperparámetro *balanced*) y técnicas de generación de datos sintéticos (SMOTE y ADASYN).

Los resultados experimentales demostraron que las técnicas de sobremuestreo estadístico incrementaron el rendimiento de los modelos entre un 7 % y un 11 % en las métricas de exactitud (*accuracy*) y sensibilidad (*recall*). En contraste, la ponderación algorítmica no arrojó mejoras estadísticamente significativas. En consecuencia, se seleccionaron las técnicas de balanceo estadístico para las fases subsiguientes, debido a que proporcionaron una mayor estabilidad en la convergencia y optimizaron la identificación de la clase sana.

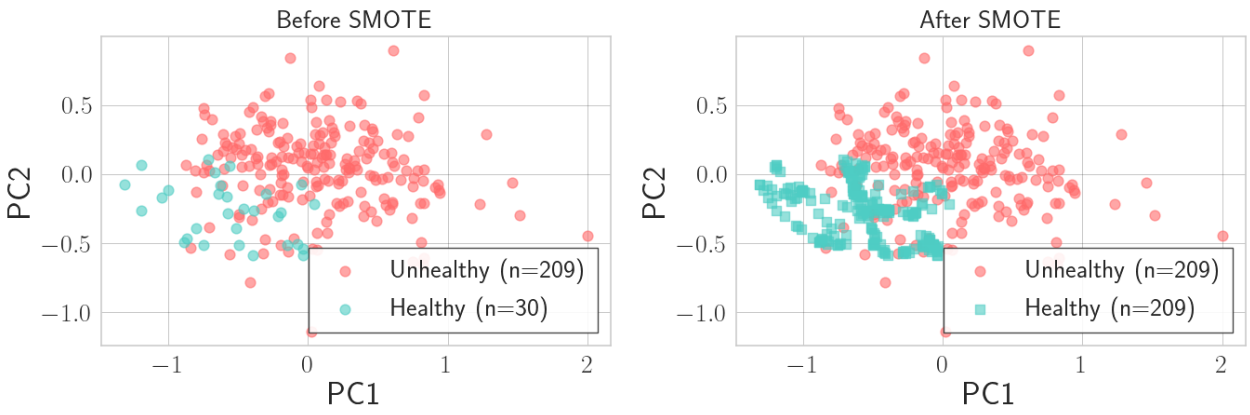

Figura 2: Reporte SMOTE para el día 2

3.3. Eleccion del orden de la Reducción y balanceo

Se evaluó el orden de aplicación de las técnicas de reducción de dimensionalidad (PCA) y balanceo estadístico (SMOTE) para determinar su impacto en el rendimiento del modelo. Dos enfoques fueron comparados: aplicar PCA antes de SMOTE y aplicar SMOTE antes de PCA. Los resultados indicaron que aplicar PCA antes de SMOTE condujo a una mejora significativa en las métricas de desempeño del modelo final, incluyendo un aumento del 6 % en la exactitud y un 8 % en la sensibilidad.

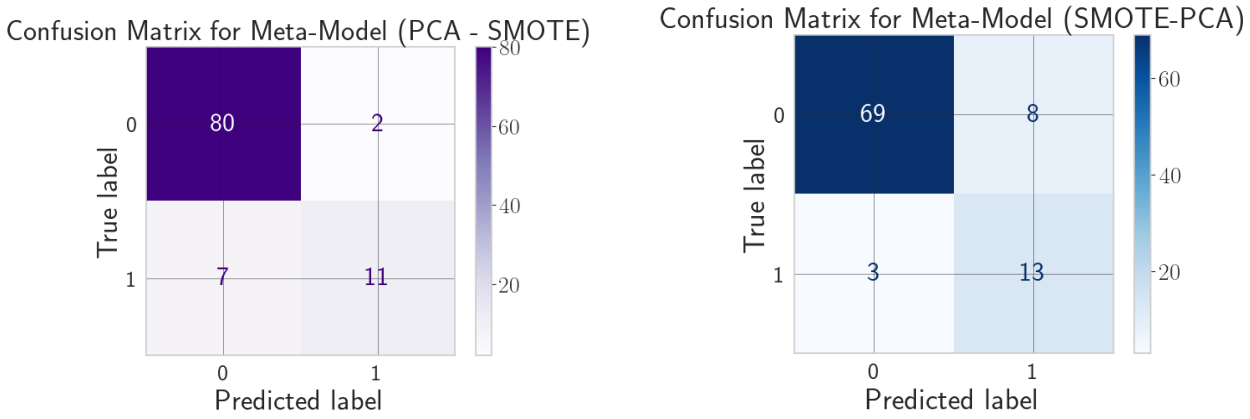

Figura 3: Rendimiento del modelo con PCA antes de SMOTE

Este enfoque permitió que SMOTE generara datos sintéticos en un espacio de menor dimensionalidad, lo que facilitó una mejor representación de la distribución de las clases y redujo el riesgo de sobreajuste. En contraste, aplicar SMOTE antes de PCA resultó en una menor calidad de los datos sintéticos generados, afectando negativamente el rendimiento del modelo. Por lo tanto, se concluyó que la secuencia óptima es aplicar PCA antes de SMOTE para maximizar la eficacia del balanceo estadístico en conjuntos de datos de alta dimensionalidad.

Una justificación adicional para esta elección es que al reducir la dimensionalidad primero, se eliminan características redundantes y ruido, lo que permite que SMOTE opere en un espacio más limpio y representativo. Esto mejora la calidad de los datos sintéticos generados, ya que SMOTE puede enfocarse en las características más relevantes para la clasificación, evitando la generación de muestras que no reflejen adecuadamente la distribución real de las clases. En resumen, aplicar PCA antes de SMOTE no solo optimiza el rendimiento del modelo, sino que también mejora la integridad y representatividad de los datos sintéticos generados.

A continuación se presentan los resultados comparativos entre ambos enfoques, para cada uno de los modelos entrenados por dia:

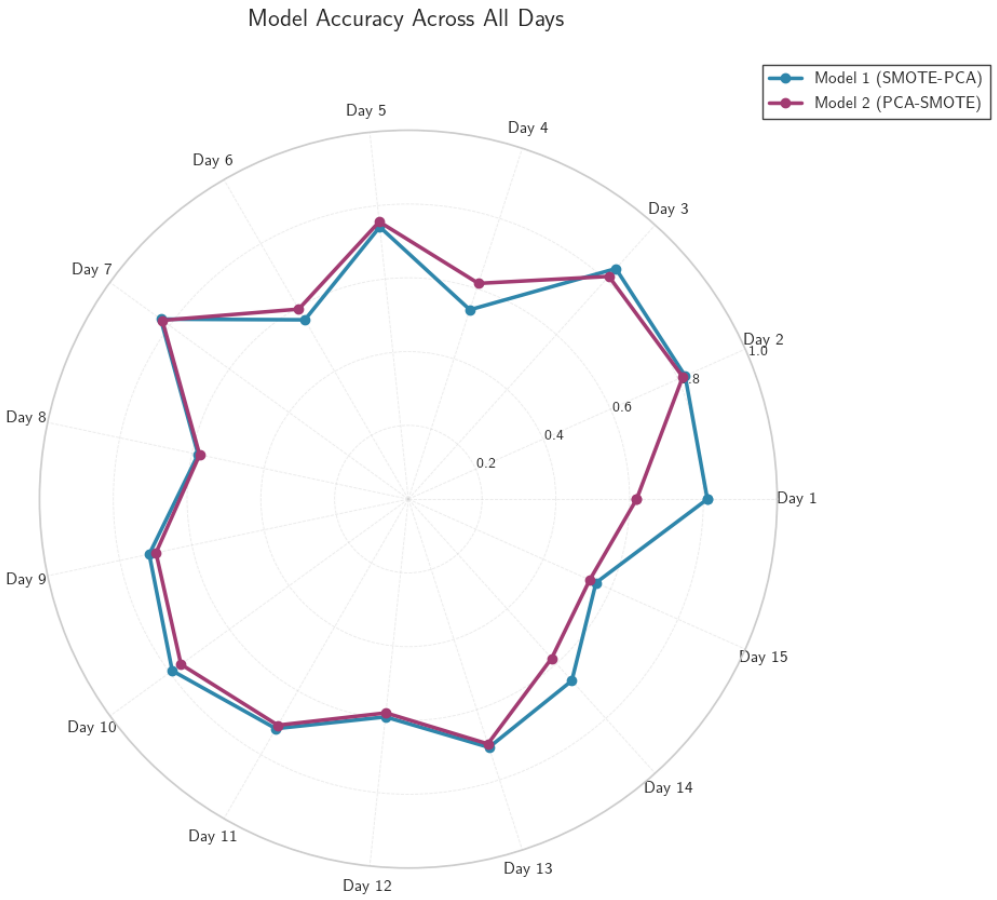

Figura 4: Comparación de modelos PCA-SMOTE vs SMOTE-PCA

4. Arquitectura del Modelo Principal

La arquitectura del modelo propuesto se estructura en dos niveles: un conjunto de modelos base (Nivel 1) y un meta-modelo de ensamblaje (Nivel 2).

El Nivel 1 consta de 15 clasificadores independientes, entrenados específicamente para los datos de cada día. Cada modelo base consiste en una Regresión Logística implementada mediante el siguiente pipeline:

- Reducción de dimensionalidad con PCA.
- Sobremuestreo sintético de la clase minoritaria con SMOTE.
- Ajuste de hiperparámetros de la Regresión Logística vía GridSearchCV.

Se optó por esta arquitectura debido a su equilibrio entre simplicidad y rendimiento. En pruebas preliminares, algoritmos de mayor complejidad (p. ej., Random Forest, SVM, XGBoost y Bagging) no ofrecieron ventajas significativas en las métricas de evaluación.

El Nivel 2 es un meta-modelo que integra las predicciones generadas por los 15 modelos base. Actualmente, se utiliza una Regresión Logística que pondera las salidas del Nivel 1, considerando el desempeño histórico y el día correspondiente a cada modelo. Se encuentra en fase de experimentación el uso de arquitecturas alternativas para este ensamblador (incluyendo redes neuronales), aunque hasta la fecha no se han observado mejoras sustanciales en el rendimiento.

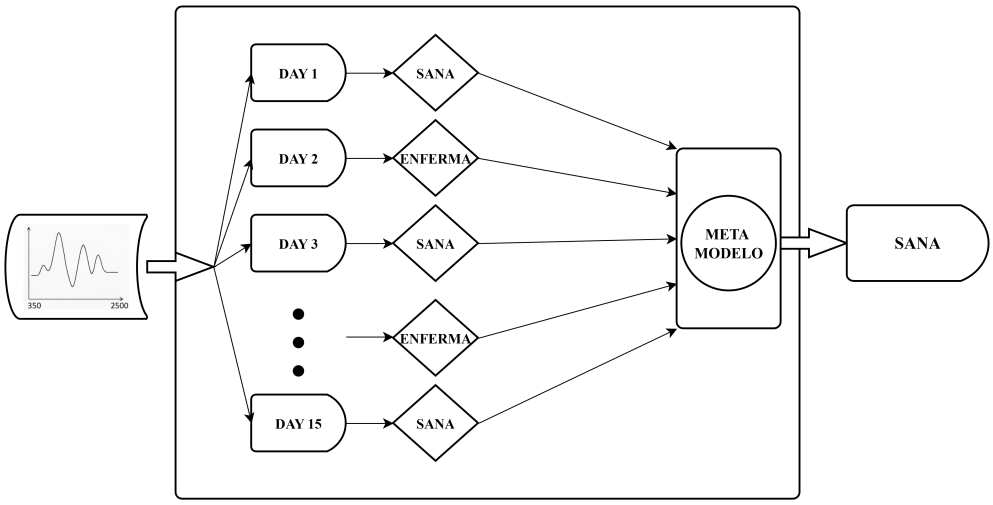

Figura 5: Arquitectura del Modelo de Ensamble

5. Errores en los datos

Los errores evidenciados en el conjunto de datos se clasifican en las siguientes categorías:

5.1. Datos Faltantes

Se identificaron valores faltantes en varias hojas, particulamente hay 2 casos críticos:

- **Hoja 0 (Dia 0):** En cada hoja deben haber 240 registros y 30 plantas por tratamiento, sin embargo, en la hoja 0 faltan las 30 plantas correspondientes a Fus_EH.
- **Hoja 15 (Dia 15):** En la hoja falta el 29.2 % de los datos. correspondientes a: 12 datos de Ralstonia, 23 datos de Ral_EH, 1 de Fus_EH y 4 de Fus_EH_Ral.

5.2. Datos no etiquetados

De las 15 hojas analizadas

1. **Imputación:** Utilizar la media/mediana para rellenar los vacíos en la Variable A, dado que el porcentaje es bajo (¡20 %).
2. **Filtrado:** Eliminar los registros con tiempos negativos y duplicados exactos.
3. **Validación:** Contactar al departamento de origen para verificar si los *outliers* son fenómenos reales o errores de sensor.

6. Conclusión

Aunque el conjunto de datos presenta desafíos en términos de completitud en variables secundarias, las variables críticas muestran una solidez suficiente para continuar con el análisis una vez aplicadas las técnicas de limpieza descritas.