Informe de Avance: cambio relativo espectral en plantas

Ricardo Esteban Lopera Vasco

19 de enero de 2026

1. Análisis de Cambio Relativo Espectral

Este capítulo presenta la metodología y resultados del análisis de cambio relativo espectral, una técnica diseñada para cuantificar las variaciones en la reflectancia de plantas sometidas a diferentes tratamientos experimentales respecto a un grupo de referencia (control).

1.1. Selección de Longitudes de Onda de Referencia

1.1.1. Metodología de Selección

La selección de las longitudes de onda de referencia se fundamenta en el establecimiento de una línea base espectral representativa del estado fisiológico normal de las plantas. Para ello, se utilizó exclusivamente el conjunto de datos correspondiente al grupo control (`df0`), el cual contiene las mediciones espectrales de plantas no sometidas a ningún tratamiento experimental.

El proceso de selección se implementó mediante los siguientes pasos:

1. **Filtrado del grupo control:** Se extrajo un subconjunto del dataframe `df0` conteniendo únicamente las observaciones correspondientes al tratamiento "Control":

```
RefDf = df0[(df0['Tratamiento'] == 'Control')]
```

2. **Extracción de datos espectrales:** Se aislaron las columnas que contienen los valores de reflectancia para cada longitud de onda, excluyendo las columnas de metadatos ("Tratamiento" y "Planta"):

```
data_cols = RefDf.iloc[:, 2:]
```

3. **Cálculo del espectro de referencia:** Se computó la media aritmética de los valores de reflectancia para cada longitud de onda λ a través de todas las plantas control:

```
REF = data_cols.mean()
```

1.1.2. Justificación Estadística

La utilización de la media como estadístico de referencia se justifica por las siguientes razones:

- **Representatividad central:** La media proporciona el valor esperado de la distribución de reflectancias para cada longitud de onda, representando el comportamiento espectral típico de una planta sana bajo condiciones normales.

- **Sensibilidad a desviaciones:** Al emplear la media como referencia, cualquier desviación en las plantas tratadas se cuantifica directamente respecto al valor central de la distribución control, permitiendo detectar tanto aumentos como disminuciones en la reflectancia.
- **Robustez estadística:** Dado que se promedian múltiples observaciones del grupo control, el espectro de referencia `REF` mitiga el efecto de variaciones aleatorias individuales, proporcionando una estimación más estable del comportamiento espectral basal.

Matemáticamente, el valor de referencia para cada longitud de onda λ_i se define como:

$$\mathrm{REF}(\lambda_i) = \frac{1}{n} \sum_{j=1}^{n} R_j(\lambda_i) \tag{1}$$

donde $R_j(\lambda_i)$ representa la reflectancia de la planta j del grupo control en la longitud de onda λ_i, y n es el número total de plantas en el grupo control.

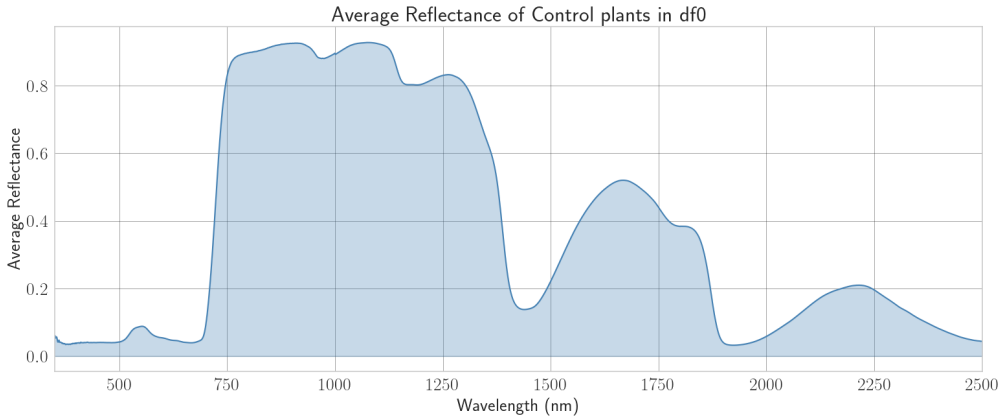

Figura 1: Reflectancia promedio del grupo control (df0) a lo largo del espectro electromagnético.

1.2. Metodología del Cambio Relativo

1.2.1. Formulación Matemática

El cambio relativo constituye una métrica normalizada que cuantifica la magnitud de la desviación espectral de una muestra respecto al espectro de referencia. La formulación implementada en el análisis se define como:

$$\Delta\lambda_{\mathrm{rel}} = \frac{|\lambda_{\mathrm{sample}} - \lambda_{\mathrm{Reference}}|}{\lambda_{\mathrm{Reference}}} \tag{2}$$

donde:

- $\Delta\lambda_{\mathrm{rel}}$ representa el cambio relativo adimensional
- $\lambda_{\mathrm{sample}}$ es el valor de reflectancia de la muestra en una longitud de onda específica

- $\lambda_{\text{Reference}}$ es el valor de referencia (media del control) para la misma longitud de onda

Para consolidar las múltiples mediciones de una planta en un único valor por longitud de onda, se calcula la magnitud euclidiana de los cambios relativos individuales:

$$M(\lambda_i) = \sqrt{\sum_{k=1}^{m} \left(\frac{|v_k - \text{REF}(\lambda_i)|}{\text{REF}(\lambda_i)} \right)^2} \tag{3}$$

donde v_k representa cada valor observado y m es el número de observaciones para la planta en cuestión.

1.2.2. Implementación Computacional

El cálculo se implementó iterando sobre cada tratamiento y planta, procesando cada columna espectral de manera independiente:

```
for col in data_cols.columns:
    values = data_cols[col].values
    REF_value = REF[col]

    if REF_value != 0:
        differences = np.abs(values - REF_value) / REF_value
        magnitude = np.sqrt(np.sum(differences**2))
    else:
        magnitude = np.nan

    magnitudes.append(magnitude)
```

1.2.3. Justificación de la Métrica

La elección del cambio relativo como métrica de análisis se fundamenta en las siguientes consideraciones:

1. **Normalización contra variabilidad basal:** Al dividir por el valor de referencia, la métrica compensa las diferencias inherentes en la magnitud absoluta de la reflectancia entre diferentes regiones del espectro. Esto permite comparar directamente cambios en longitudes de onda con valores base muy distintos.

2. **Detección de respuestas al estrés:** Las plantas sometidas a estrés biótico o abiótico exhiben alteraciones en su firma espectral, particularmente en regiones asociadas con pigmentos fotosintéticos, contenido de agua y estructura celular. El cambio relativo amplifica estas señales de estrés al expresarlas como proporción del estado basal.

3. **Interpretabilidad:** Un valor de $\Delta\lambda_{\text{rel}} = 0{,}1$ indica una desviación del 10 % respecto al control, facilitando la interpretación biológica de los resultados.

4. **Uso del valor absoluto:** La implementación utiliza $|\lambda_{\text{sample}} - \lambda_{\text{Reference}}|$, lo cual cuantifica la magnitud de la desviación independientemente de su dirección. Esto es apropiado cuando el objetivo es detectar cualquier perturbación respecto al estado normal, sin distinguir entre aumentos y disminuciones de reflectancia.

1.3. Análisis de Resultados: Cambio Relativo Promedio

1.3.1. Agregación por Tratamiento

Para obtener una visión consolidada del efecto de cada tratamiento, se calculó el cambio relativo promedio agrupando las observaciones por tipo de tratamiento:

```
df_means = df_all_magnitudes.groupby('Tratamiento').mean()
```

Esta agregación permite comparar directamente el perfil espectral de cambio relativo entre los diferentes grupos experimentales, suavizando la variabilidad inter-individual dentro de cada tratamiento.

1.3.2. Interpretación de Tendencias

El análisis del cambio relativo promedio revela patrones distintivos para cada tratamiento experimental:

- **Valores cercanos a cero:** Indican que el tratamiento no produce alteraciones significativas en la reflectancia espectral respecto al control. Esto sugiere que el tratamiento no afecta los procesos fisiológicos que determinan las propiedades ópticas de la planta.
- **Valores positivos elevados:** Señalan regiones espectrales donde el tratamiento induce cambios sustanciales. En el contexto biológico:
  - Cambios en la región visible (400-700 nm) pueden indicar alteraciones en el contenido de clorofila, carotenoides u otros pigmentos.
  - Cambios en el infrarrojo cercano (700-1300 nm) sugieren modificaciones en la estructura celular del mesófilo.
  - Cambios en el infrarrojo de onda corta (1300-2500 nm) están asociados con variaciones en el contenido de agua y compuestos bioquímicos.
- **Diferencias entre tratamientos:** La magnitud relativa del cambio entre tratamientos permite establecer un ranking de severidad del efecto, donde tratamientos con mayores valores de $\Delta\lambda_{\text{rel}}$ representan perturbaciones más intensas del estado fisiológico normal.

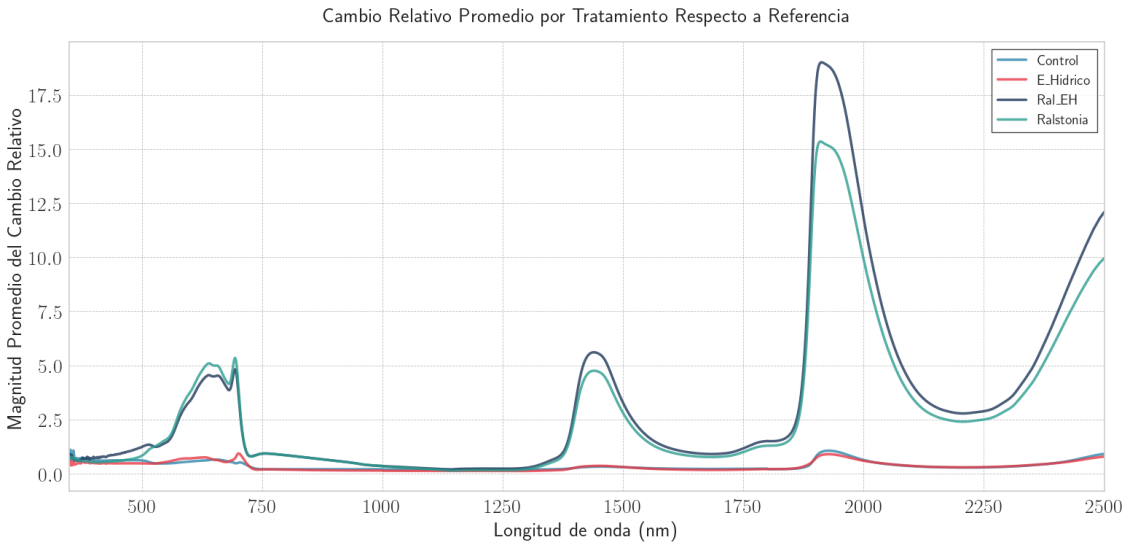

Figura 2: Cambio relativo promedio por tratamiento respecto a la referencia a lo largo del espectro completo.

1.3.3. Análisis por Rangos Espectrales Clave

Para un análisis más detallado, se segmentó el espectro en cuatro regiones de interés fisiológico:

1. **VIS-NIR (500-750 nm):** Región dominada por la absorción de pigmentos fotosintéticos. Cambios en esta zona reflejan alteraciones en el aparato fotosintético.

2. **SWIR-1 (1300-1600 nm):** Sensible al contenido de agua foliar y estructura celular. Variaciones indican estrés hídrico o cambios en la turgencia celular.

3. **SWIR-2 (1800-2200 nm):** Asociada con bandas de absorción de agua y compuestos orgánicos como celulosa, lignina y proteínas.

4. **SWIR-3 (2300-2500 nm):** Región sensible a compuestos bioquímicos específicos y contenido de agua residual.

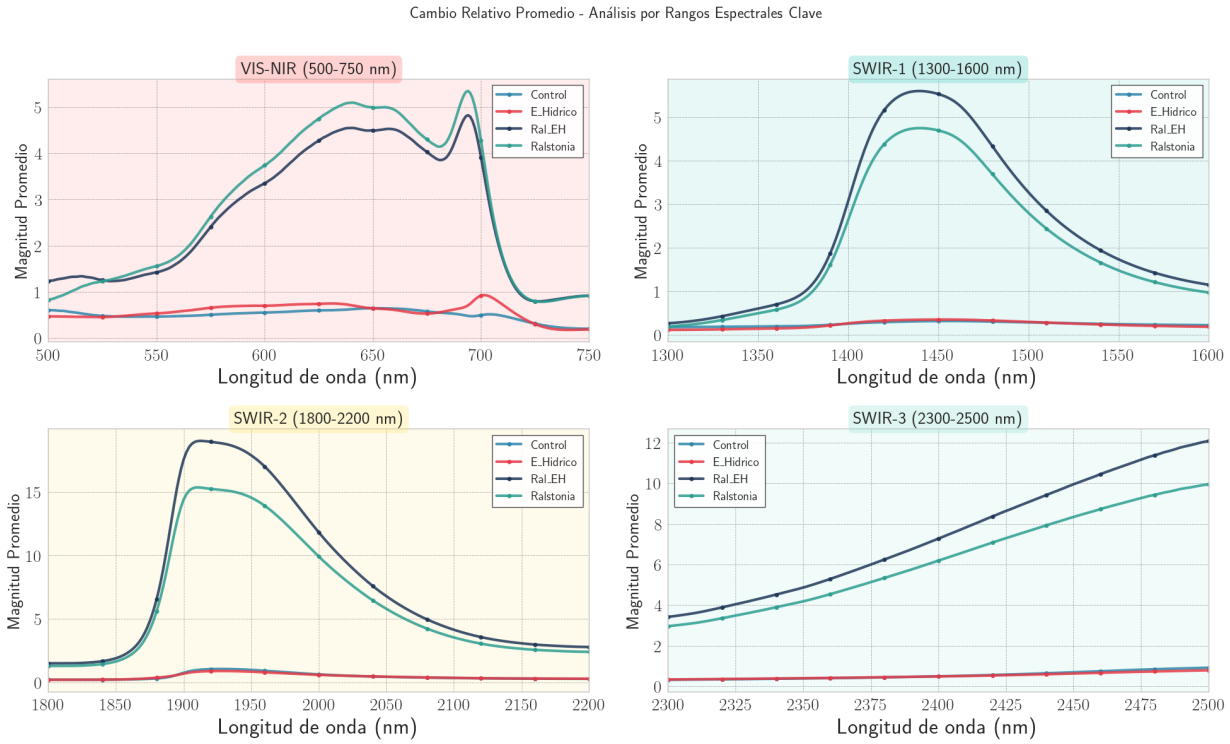

Figura 3: Análisis del cambio relativo promedio por rangos espectrales clave.

1.3.4. Conclusiones del Análisis

El análisis de cambio relativo permite establecer las siguientes observaciones:

- La metodología implementada proporciona una cuantificación objetiva y normalizada de las perturbaciones espectrales inducidas por los tratamientos experimentales.
- Las regiones espectrales con mayor discriminación entre tratamientos constituyen potenciales biomarcadores para la detección temprana de estrés vegetal.
- La agregación por tratamiento revela patrones consistentes que validan la sensibilidad de la técnica espectroscópica para caracterizar respuestas fisiológicas.
- El marco metodológico establecido sienta las bases para el desarrollo de modelos predictivos de clasificación basados en firmas espectrales de cambio relativo.

2. Resultados: Clasificación Biótico vs Abiótico

Esta sección presenta los resultados obtenidos de la evaluación comparativa de cinco algoritmos de aprendizaje automático para la clasificación binaria de estrés vegetal (biótico vs abiótico), utilizando las características de cambio relativo espectral como variables predictoras.

2.1. Configuración Experimental

Se implementó un pipeline de procesamiento que integra estandarización de datos, reducción de dimensionalidad mediante Análisis de Componentes Principales (PCA) y clasificación. Los hiperparámetros óptimos para cada modelo se determinaron mediante búsqueda exhaustiva en rejilla (GridSearchCV) con validación cruzada de 5 particiones, utilizando *balanced accuracy* como métrica de optimización.

Los modelos evaluados fueron:

1. Regresión Logística con regularización L1, L2 y ElasticNet
2. Máquina de Vectores de Soporte (SVM) con kernels lineal y RBF
3. Bosque Aleatorio (Random Forest)
4. Gradient Boosting
5. K-Vecinos más Cercanos (KNN)

2.2. Comparación de Rendimiento de Modelos

La Tabla 1 presenta las métricas de rendimiento obtenidas para cada clasificador. Se reportan tres métricas: el puntaje de validación cruzada (CV Score), la exactitud sobre el conjunto de prueba (Test Accuracy) y el puntaje F1 ponderado (Test F1 Score).

Tabla 1: Comparación de rendimiento de modelos para clasificación Biótico vs Abiótico basada en cambio relativo espectral.

Modelo	**CV Score**	**Test Accuracy**	**Test F1 Score**
SVM	0.7715	0.8644	0.8644
Random Forest	0.7962	0.8475	0.8475
Gradient Boosting	0.8071	0.8475	0.8472
Regresión Logística	0.7692	0.8136	0.8136
KNN	0.7364	0.7797	0.7795

2.2.1. Análisis de Resultados

Los resultados cuantitativos presentados en la Tabla 1 permiten establecer las siguientes observaciones:

- **Modelo de mejor rendimiento:** La Máquina de Vectores de Soporte (SVM) alcanzó el desempeño más alto en el conjunto de prueba, con una exactitud y puntaje F1 de 0.8644 (86.44 %). Este resultado posiciona a SVM como el clasificador más efectivo para la discriminación entre estrés biótico y abiótico utilizando características de cambio relativo espectral.

- **Discrepancia entre validación cruzada y prueba:** Se observa un patrón notable donde todos los modelos exhiben rendimientos superiores en el conjunto de prueba respecto a sus puntajes de validación cruzada. Gradient Boosting presenta el CV Score más alto (0.8071), sin embargo, su exactitud de prueba (0.8475) es inferior a la de SVM. Esta inversión sugiere que el CV Score más conservador de SVM (0.7715) refleja una estimación más robusta de su capacidad de generalización.

- **Rendimiento de métodos ensemble:** Random Forest y Gradient Boosting obtuvieron exactitudes de prueba idénticas (0.8475), aunque Gradient Boosting mostró una ligera ventaja en validación cruzada (0.8071 vs 0.7962). La diferencia marginal en sus puntajes F1 (0.8475 vs 0.8472) indica comportamientos predictivos prácticamente equivalentes.

- **Rendimiento inferior de KNN:** El clasificador K-Vecinos más Cercanos presentó el desempeño más bajo, con una exactitud de 0.7797 y F1 de 0.7795. La diferencia de 8.5 puntos porcentuales respecto a SVM sugiere que la estructura del espacio de características de cambio relativo favorece clasificadores basados en hiperplanos de separación sobre métodos basados en distancia local.

- **Regresión Logística como línea base:** Con una exactitud de 0.8136, la Regresión Logística proporciona un rendimiento intermedio que valida la separabilidad lineal parcial de las clases en el espacio reducido por PCA, aunque los métodos no lineales (SVM con kernel RBF, Random Forest) logran capturar patrones adicionales que mejoran la clasificación.

2.3. Visualización Comparativa

Para facilitar la interpretación de los resultados, se generaron representaciones gráficas del rendimiento de los clasificadores.

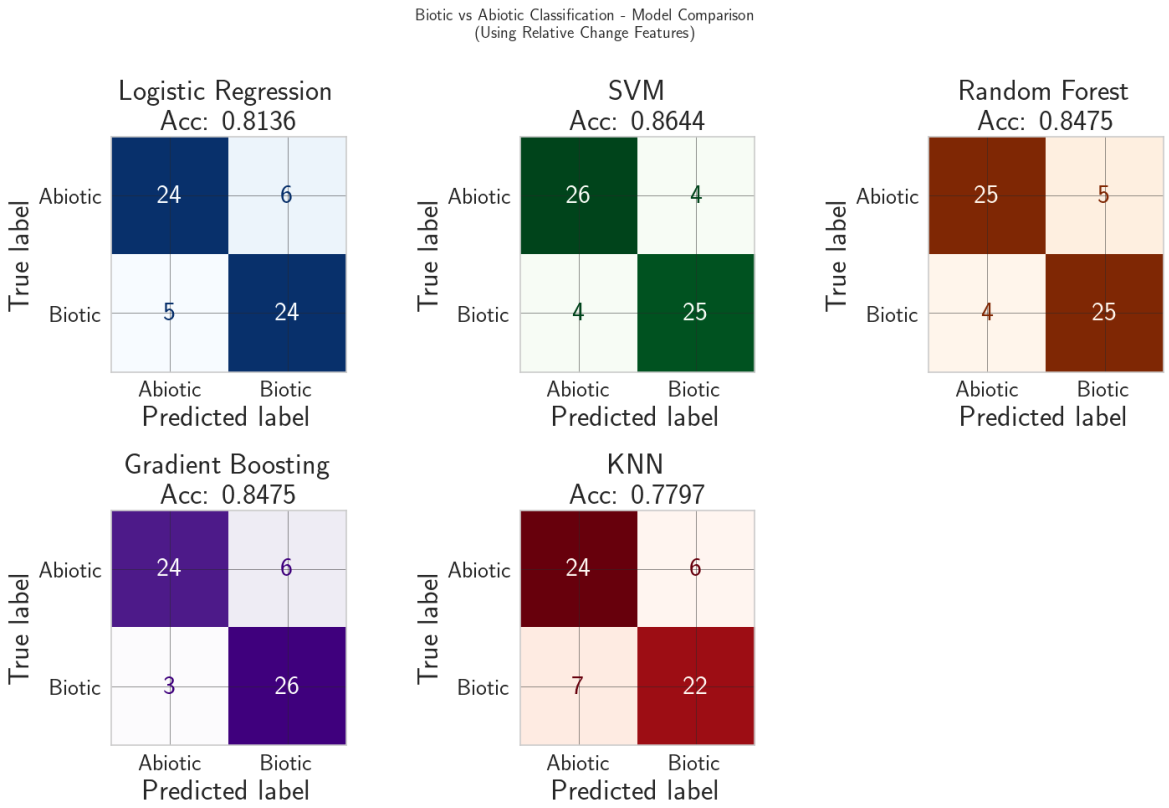

Figura 4: Matrices de confusión para los cinco modelos evaluados. Cada matriz muestra la distribución de predicciones correctas e incorrectas para las clases Abiótico y Biótico. Los valores en la diagonal principal representan clasificaciones correctas.

Las matrices de confusión (Figura 4) permiten evaluar el comportamiento de cada clasificador respecto a errores de tipo I (falsos positivos) y tipo II (falsos negativos). Un modelo óptimo exhibiría valores elevados en la diagonal principal y valores cercanos a cero en las celdas fuera de la diagonal.

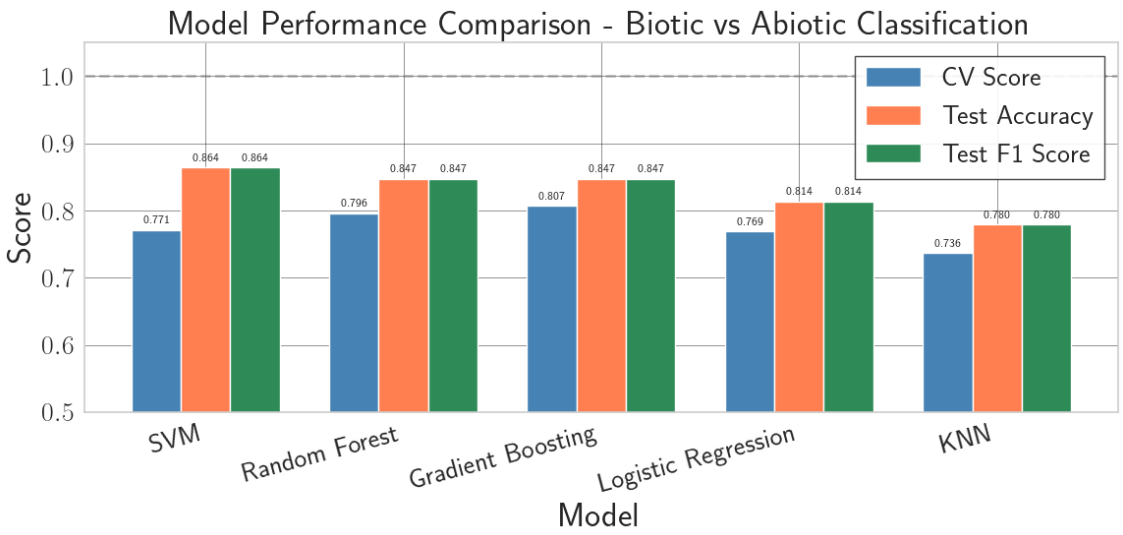

Figura 5: Comparación de métricas de rendimiento (CV Score, Test Accuracy y Test F1 Score) para cada modelo de clasificación. Las barras permiten una comparación visual directa entre algoritmos.

El gráfico de barras comparativo (Figura 5) sintetiza las tres métricas de evaluación, facilitando la identificación del modelo con mejor balance entre rendimiento en validación cruzada y generalización al conjunto de prueba.

2.4. Selección del Modelo Óptimo

Con base en el análisis comparativo presentado, se selecciona la **Máquina de Vectores de Soporte (SVM)** como el modelo óptimo para la clasificación de estrés biótico vs abiótico. Esta selección se fundamenta en los siguientes criterios:

1. **Máximo rendimiento en prueba:** SVM alcanzó la exactitud más alta (86.44 %) y el puntaje F1 más elevado (0.8644) sobre el conjunto de prueba independiente, superando al segundo mejor modelo (Random Forest) por 1.69 puntos porcentuales.

2. **Capacidad de generalización:** A pesar de presentar un CV Score moderado (0.7715), SVM demostró la mejor transferencia de aprendizaje al conjunto de prueba. La diferencia positiva de 9.29 puntos porcentuales entre Test Accuracy y CV Score indica que el modelo no presenta sobreajuste y generaliza efectivamente a datos no vistos.

3. **Robustez frente a métodos ensemble:** Aunque Gradient Boosting obtuvo el CV Score más alto (0.8071), su rendimiento de prueba (0.8475) fue inferior al de SVM. Esto sugiere que la complejidad adicional de los métodos ensemble no aporta beneficios para este problema específico, mientras que SVM logra una representación más eficiente del límite de decisión entre clases.

4. **Eficiencia computacional:** SVM con kernel RBF proporciona un balance óptimo entre complejidad del modelo y rendimiento predictivo, requiriendo menos parámetros

que Random Forest (número de árboles, profundidad) o Gradient Boosting (tasa de aprendizaje, número de estimadores).

2.4.1. Configuración del Modelo Seleccionado

Los hiperparámetros óptimos identificados mediante GridSearchCV para el modelo SVM fueron almacenados en el archivo `best_model_parameters.json`. El pipeline final integra:

- Estandarización de características mediante `StandardScaler`
- Reducción de dimensionalidad con PCA (número de componentes optimizado)
- Clasificador SVM con parámetros de regularización (C) y kernel optimizados