UNIVERSIDAD MIGUEL HERNÁNDEZ DE ELCHE

ESCUELA POLITÉCNICA SUPERIOR DE ELCHE

GRADO EN INGENIERÍA ELECTRÓNICA Y AUTOMÁTICA INDUSTRIAL


# "Caracterización de la degradación temporal de células solares orgánicas mediante técnicas de *machine learning*"

TRABAJO FIN DE GRADO

Junio - 2025

AUTOR: Manuel Vico Rodriguez

DIRECTOR/ES: David Valiente García

# Resumen

El presente trabajo tiene como objetivo optimizar la eficiencia y la durabilidad de células solares orgánicas (OSC) mediante el uso de técnicas de *machine learning*. Para ello, se llevaron a cabo experimentos con diferentes combinaciones de materiales, evaluando el impacto de factores externos como la temperatura y la humedad en la eficiencia de conversión de potencia (PCE). Los modelos de Random Forest y Gradient Boost se utilizaron para realizar predicciones sobre la eficiencia de las células, y la herramienta ROBERT se utilizó para realizar un análisis similar de manera automatizada y comparar los resultados. Los resultados muestran que las variables meteorológicas, como la temperatura y la humedad, tienen un efecto significativo en la PCE. Además, se identificaron áreas clave de mejora, tales como la necesidad de mejorar la calidad de los datos y reducir la redundancia entre variables. Se recomienda la ampliación del conjunto de datos y la exploración de modelos más complejos para mejorar la capacidad predictiva en futuros estudios.





# Abstract

This study aims to optimize the efficiency and durability of organic solar cells (OSCs) using machine learning techniques. Various material combinations were tested to evaluate the impact of external factors, such as temperature and humidity, on the power conversion efficiency (PCE). Random Forest and Gradient Boost models were employed to predict cell efficiency, while the ROBERT tool was used to perform a similar analysis in an automated way and compare the results. The results indicate that meteorological variables, such as temperature and humidity, significantly affect PCE. Key areas for improvement were identified, including enhancing data quality and reducing variable redundancy. It is recommended to expand the dataset and explore more complex models to improve predictive capacity in future research.

# ÍNDICE

# ÍNDICE DE FIGURAS

# ÍNDICE DE TABLAS

# 1. INTRODUCCIÓN

En un mundo cada vez más consciente del impacto ambiental del cambio climático y la dependencia de los combustibles fósiles, las energías renovables se han convertido en una alternativa crucial para un futuro sostenible. Estas fuentes de energía, como la solar, eólica, geotérmica e hidroeléctrica, ofrecen un suministro de energía limpio, abundante y renovable que puede reducir nuestra huella de carbono y mitigar los efectos del calentamiento global.

Si bien la adopción de energías renovables está viendo un crecimiento cada vez mayor a nivel global, aún queda un largo camino por recorrer para alcanzar un sistema energético totalmente limpio.

A pesar de los avances, solo unos pocos países se han acercado a este objetivo. Islandia, Noruega y Brasil son algunos de los países con mayor porcentaje de producción energética proveniente de fuentes renovables. En el mapa (Figura 1) se muestra este dato para un gran número de países.

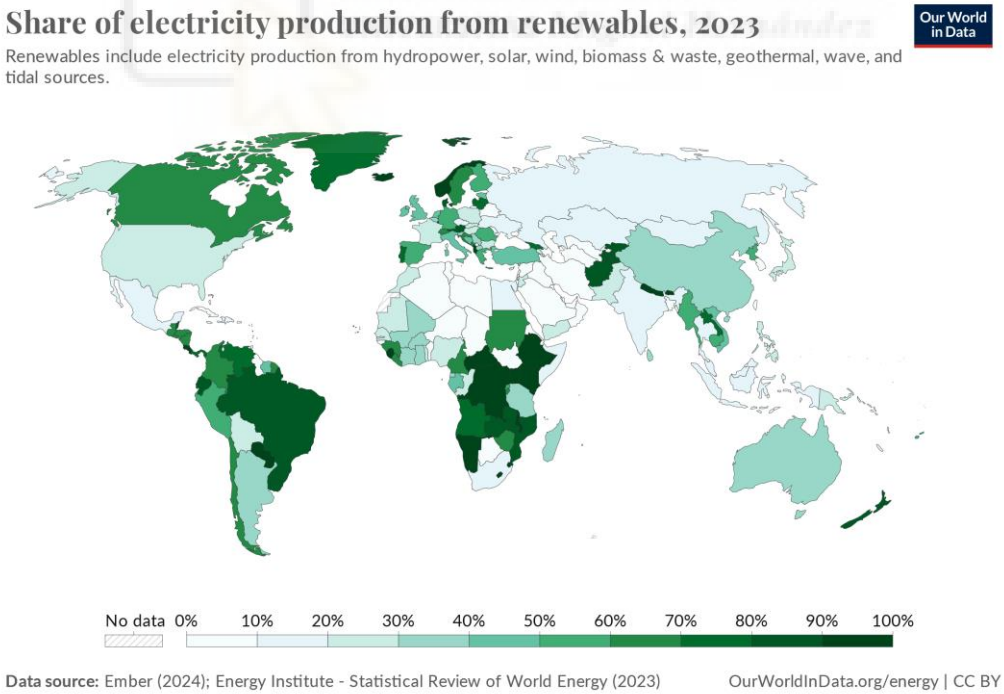



*Figura 1: Porcentaje de producción eléctrica proveniente de fuentes renovables [1].*

Entre las energías renovables, la energía solar destaca por su potencial ilimitado y su capacidad de generar electricidad a partir de la luz solar. Esta energía limpia y abundante puede ser aprovechada en diversas aplicaciones, desde hogares y comunidades hasta grandes instalaciones industriales.

En la última década, la inversión en energías renovables ha superado los 2.5 billones de dólares, con la energía solar a la vanguardia, demostrando ser tano una opción sostenible como económicamente viable. Esta inversión ha impulsado una capacidad instalada sin precedentes y ha jugado un papel clave en la reducción de las emisiones de $CO_2$, subrayando la importancia de continuar apoyando el crecimiento de la energía solar y otras renovables [2].

Para ilustrar este crecimiento, a continuación, se presenta una gráfica (Figura 2) que muestra la evolución de la capacidad de energía solar instalada a nivel mundial desde el año 2020.

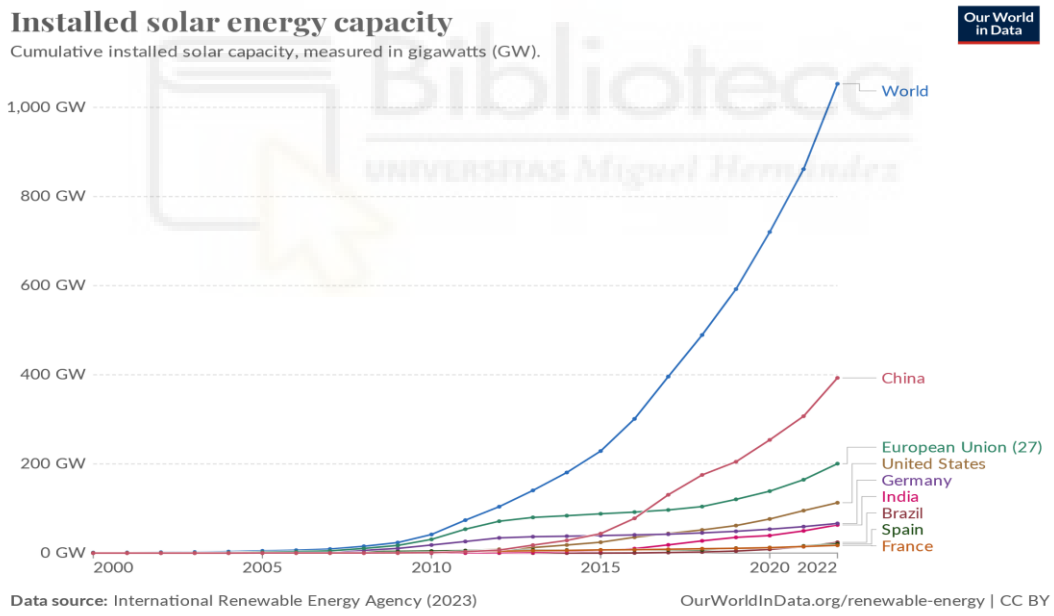



*Figura 2: Capacidad instalada de energía solar [3].*

Este crecimiento, se ha visto impulsado por la por la reducción de costos, el aumento de la eficiencia y el apoyo de políticas gubernamentales. Sin embargo, aún estamos muy lejos de los objetivos deseados, por ello es necesario seguir innovando en las tecnologías fotovoltaicas.

En este aspecto, se están produciendo avances significativos en el desarrollo de nuevas células fotovoltaicas con una mayor eficiencia, mayor durabilidad y costes reducidos.

Entre estas nuevas tecnologías se encuentran las células de perovskita, células tándem o células de película fina [4].

En el siguiente trabajo las células utilizadas son células solares orgánicas (OSC, por sus siglas en inglés) de tipo *bulk heterojunction*, las cuales se caracterizan por dos capas activas de materiales orgánicos diferentes, que se encuentran en contacto directo entre sí.

En la siguiente gráfica (Figura 3) se puede observar cómo la eficiencia de estas nuevas tecnologías ha ido en aumento, todo esto gracias a la innovación y al trabajo de distintos investigadores.

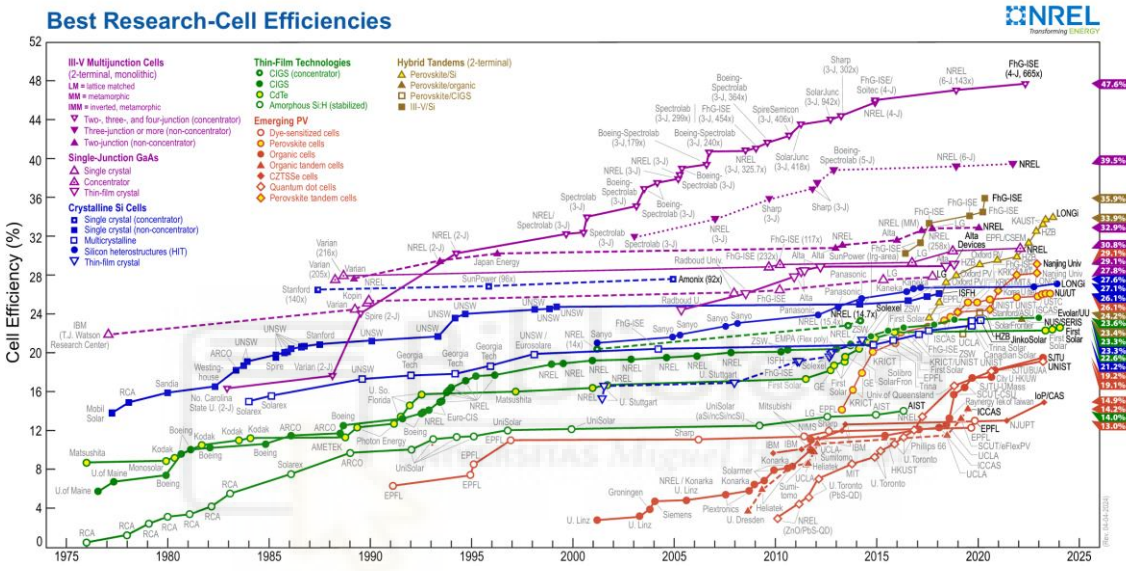



*Figura 3: Evolución de la eficiencia de conversión de potencia en distintas tecnologías de células solares [5].*

## 1.1 Objetivos

En este estudio, se propone identificar las combinaciones óptimas de materiales para la fabricación de OSC, mediante el modelado computacional, apoyado en técnicas de *machine learning* (ML). El objetivo se centra en optimizar la eficiencia y la durabilidad de las distintas células fabricadas en el laboratorio del área de electróncia de la UMH, buscando así potenciar su viabilidad. Para ello se han fabricado células con una variedad de materiales, cantidad y proporciones, lo que permite obtener un amplio conjunto de datos para su análisis.

A partir de los datos obtenidos se han planteado los siguientes objetivos específicos:

1. Estudiar el efecto del tiempo en la degradación de las OSC para identificar los factores que contribuyen a su estabilidad. Con ello, se buscará caracterizar la evolución de su eficiencia a lo largo del tiempo.

2. Evaluar el impacto de factores externos como la temperatura, humedad y presión en la eficiencia y degradación de las OSC.

3. Identificar combinaciones óptimas de materiales en las OSC utilizando algoritmos de aprendizaje automático, para maximizar la eficiencia conversión de potencia (PCE, del inglés "*Power conversion efficiency*") y durabilidad de estas a lo largo del tiempo.

Para alcanzar estos objetivos, se llevarán a cabo experimentos en el laboratorio con diversas configuraciones de OSC.

## 1.2 Justificación del Trabajo

La combinación de la tecnología de células solares orgánicas y las técnicas de *machine learning* ofrece un enfoque innovador y eficiente para mejorar el rendimiento y la sostenibilidad de las energías renovables. Este trabajo se justifica por la necesidad urgente de desarrollar fuentes de energía limpia y la capacidad de ML para transformar el proceso de investigación y desarrollo en este campo. A través de esta investigación, se pretende demostrar cómo el procesamiento de datos y los algoritmos de ML pueden ser utilizados para caracterizar automáticamente y optimizar las OSC, contribuyendo a avances significativos en la tecnología solar, como entre otros, maximizar la eficiencia del rendimiento energético o reducir los costes y residuos de fabricación.

La investigación además busca avanzar en el conocimiento teórico sobre las OSC y el aprendizaje automático, también proporcionar soluciones prácticas que puedan ser aplicadas en el desarrollo de tecnologías fotovoltaicas más eficientes y sostenibles. La optimización de materiales y el análisis de factores climatológicos tienen el potencial de mejorar significativamente el rendimiento y la durabilidad de las OSC, haciéndolas una opción más viable para una amplia gama de aplicaciones, desde pequeñas instalaciones residenciales hasta grandes parques solares.

Además, la capacidad de los algoritmos de aprendizaje automático para manejar y analizar grandes volúmenes de datos permite una caracterización más precisa y detallada de las OSC, lo que puede llevar a descubrimientos de nuevas combinaciones de materiales y configuraciones que de otro modo no serían posibles. En última instancia, este trabajo busca contribuir a la transición hacia un sistema energético más limpio y sostenible, alineándose con los objetivos globales de reducir las emisiones de carbono y combatir el cambio climático.

# 2. MARCO TEÓRICO

En esta sección se establecen las bases fundamentales y se proporciona el contexto necesario para comprender el desarrollo y los resultados del presente estudio. Esta sección explora los conceptos teóricos clave y las teorías subyacentes. Al examinar las bases científicas y tecnológicas de las células solares orgánicas, así como los avances en el campo de la inteligencia artificial aplicados a la optimización y análisis de dispositivos fotovoltaicos, este marco teórico no solo contextualiza el estudio dentro del campo de la ingeniería fotovoltaica, sino que también esboza las contribuciones teóricas que guían el análisis de datos y la interpretación de los resultados.

A lo largo de este apartado, se discutirán los principios operativos de las células solares orgánicas, incluyendo su composición, mecanismos de funcionamiento y los factores que afectan su eficiencia y estabilidad. Paralelamente, se revisarán las técnicas de aprendizaje automático que facilitan la extracción de patrones y relaciones complejas dentro de grandes conjuntos de datos, una capacidad esencial para la evaluación efectiva y eficiente de las tecnologías energéticas en desarrollo.

## 2.1 Fundamentos de las Células Solares

El funcionamiento de las células solares se basa en el efecto fotovoltaico, un proceso que permite la conversión directa de la energía lumínica en energía eléctrica. Este fenómeno comienza cuando un fotón de luz, con suficiente energía, impacta en el material semiconductor de la célula solar, como el silicio. Cuando esto ocurre, los fotones son absorbidos, y su energía se transfiere a los electrones en los átomos del semiconductor [6].

Al recibir esta energía, los electrones son "excitados", lo que significa que ganan suficiente energía para liberarse de sus posiciones originales en la estructura atómica. Esta excitación provoca la creación de un par electrón-hueco, donde el electrón es una partícula cargada negativamente, y el hueco es la ausencia del electrón, que actúa como una carga positiva móvil dentro del material [6], [7].

Para que se genere una corriente eléctrica utilizable, los electrones libres deben ser recolectados y dirigidos a través de un circuito externo. Las células solares logran esto mediante la creación de un campo eléctrico interno en la unión entre dos capas de material semiconductor con características eléctricas diferentes: una capa de tipo *n* (donde los electrones son las partículas mayoritarias) y una capa de tipo *p* (donde los huecos predominan). Este campo eléctrico empuja los electrones hacia el lado tipo *n* y los huecos hacia el lado tipo *p*, generando una diferencia de potencial o voltaje entre ambos lados. Cuando se conecta un circuito externo, los electrones fluyen a través de este, generando una corriente eléctrica [6], [7].

En resumen, el proceso físico clave que permite la generación de electricidad en una célula solar implica la absorción de luz, la generación de pares electrón-hueco, y la separación de estas cargas mediante un campo eléctrico, lo que da lugar al flujo de electrones en un circuito externo.

### 2.1.1 Tipo de Células Solares

Existen varios tipos de células solares, cada una con sus propias características, ventajas y limitaciones. A grandes rasgos, los tipos más comunes incluyen:

1. **Células de Silicio Cristalino**: estas son las más utilizadas comercialmente y se dividen en células monocristalinas y policristalinas. Las células monocristalinas

están hechas de un único cristal de silicio, lo que las hace más eficientes, pero también más caras de fabricar. Por otro lado, las células policristalinas están compuestas de múltiples cristales de silicio, lo que reduce su costo de producción, pero también su eficiencia [8].

2. **Células de Película Delgada**: estas células están compuestas por capas delgadas de materiales semiconductores, como teluro de cadmio (CdTe) o diseleniuro de cobre, indio y galio (CIGS). Las células de película delgada son menos eficientes que las de silicio cristalino, pero son más baratas de producir y pueden ser flexibles, lo que las hace adecuadas para aplicaciones como fachadas de edificios y dispositivos portátiles [4], [8].

3. **Células Solares Orgánicas**: este tipo de células, también conocidas como células solares orgánicas, son el foco principal de este trabajo debido a su potencial en aplicaciones específicas. Las células solares orgánicas están hechas de compuestos orgánicos (basados en carbono), que pueden absorber la luz solar y convertirla en electricidad. Su estructura suele estar compuesta por una capa activa, donde ocurre la excitación y la generación de pares electrón-hueco, situada entre dos electrodos. La eficiencia de estas células aún es inferior a la de las células de silicio, pero su bajo costo de producción, su flexibilidad, y su posibilidad de ser fabricadas mediante procesos como la impresión en rollo las hacen muy atractivas para aplicaciones futuras [4], [8].

4. **Células Solares de Perovskita**: estas células están ganando popularidad debido a su alta eficiencia y a las bajas temperaturas necesarias para su fabricación. Sin embargo, su estabilidad a largo plazo sigue siendo un desafío [4].

#### 2.1.1.1 Células Solares Orgánicas: Ventajas, Limitaciones y Aplicaciones

Las células solares orgánicas se distinguen por su estructura y propiedades únicas. Están formadas por materiales semiconductores orgánicos, que pueden ser polímeros o moléculas pequeñas. A diferencia de las células de silicio, los materiales orgánicos tienen la ventaja de ser más livianos, flexibles y fabricables mediante técnicas de bajo costo, como la impresión o el revestimiento en solución [9], [10].

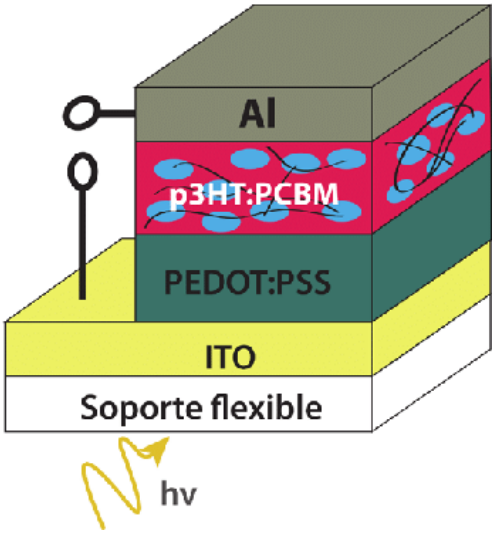



*Figura 4: Estructura de una célula solar orgánica de heterounión [11].*

Una de las ventajas más destacadas de las células solares orgánicas es su capacidad para ser producidas en grandes superficies a bajo costo, lo que las hace muy atractivas para aplicaciones como ventanas solares, ropa inteligente, o dispositivos portátiles. Sin embargo, presentan una eficiencia de conversión más baja en comparación con las células tradicionales de silicio, debido a factores como la recombinación de electrones y la inestabilidad de los materiales orgánicos frente a la exposición prolongada al ambiente [9], [10].

Además de la eficiencia, otro desafío es la vida útil de las células solares orgánicas, ya que su rendimiento tiende a degradarse con el tiempo, especialmente bajo la exposición a la luz ultravioleta y el oxígeno. Sin embargo, las investigaciones continúan para mejorar tanto su estabilidad como su eficiencia, lo que sugiere que podrían tener un papel crucial en aplicaciones específicas en el futuro cercano [12].

### 2.1.2 Cálculo del Rendimiento de las Células Fotovoltaicas

El rendimiento de una célula fotovoltaica se mide principalmente por su capacidad para convertir la energía solar en energía eléctrica utilizable. Este rendimiento se expresa a través de la Eficiencia de Conversión de Potencia o PCE (*Power Conversion Efficiency*), que es la relación entre la energía eléctrica generada y la energía solar incidente sobre la célula. Para calcular la PCE, es necesario analizar la curva I-V (corriente-voltaje), que es una representación gráfica fundamental en la caracterización de células solares [13].

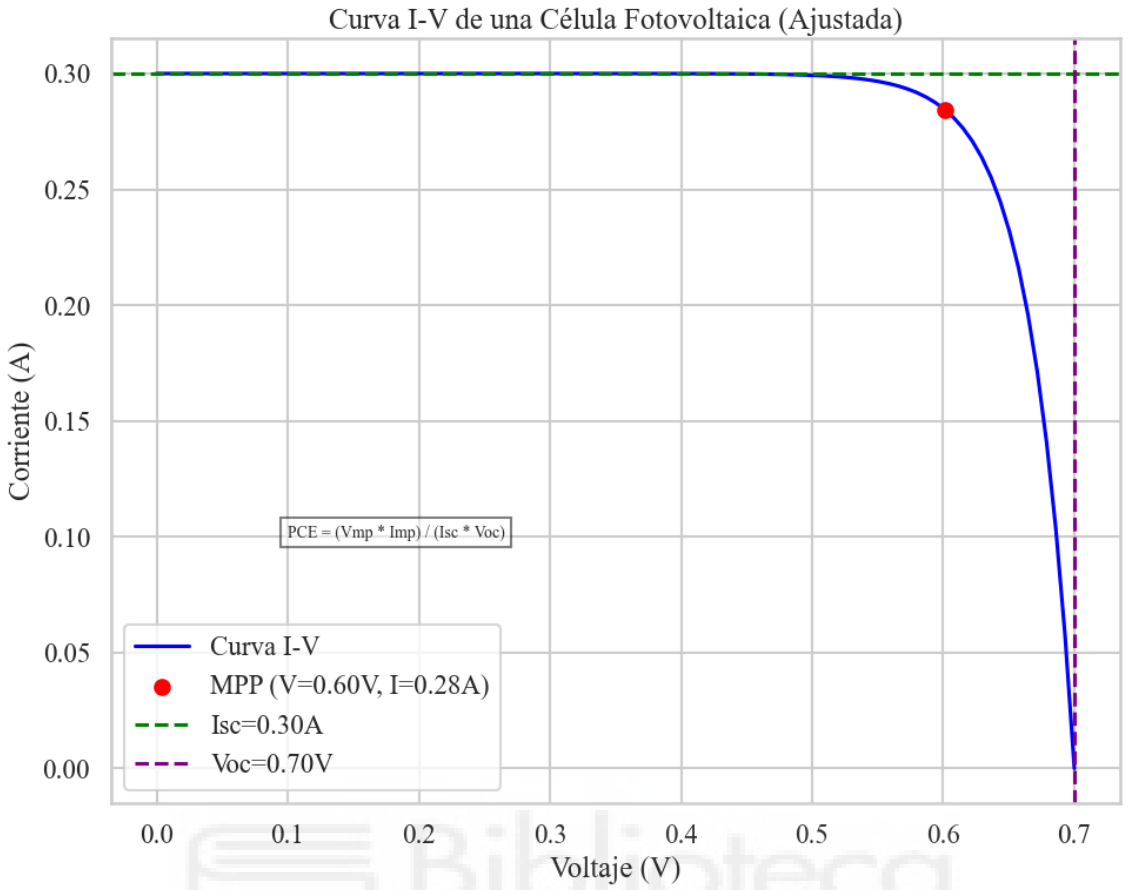



*Figura 5:Gráfica I-V típica de una célula fotovoltaica.*

### Curvas I-V y Principales Métricas

La curva I-V muestra la relación entre la corriente generada por la célula y el voltaje que se produce bajo condiciones de iluminación específicas. Para medir el rendimiento, se realiza una prueba estándar bajo la llamada condición de prueba estándar (STC), que incluye una irradiancia solar de 1000 W/m², una temperatura de la célula de 25°C y una masa de aire (AM) de 1.5 [14]. Al analizar la curva I-V, se obtienen varias métricas importantes:

1. **Corriente de Cortocircuito (Isc)**: es la corriente máxima que puede generar la célula cuando el voltaje es igual a cero, es decir, cuando los terminales de la célula están en cortocircuito. Esta métrica refleja la cantidad de luz solar absorbida por la célula y está relacionada con la eficiencia de absorción del material [7], [13].

2. **Voltaje de Circuito Abierto (Voc)**: representa el voltaje máximo que puede generar la célula cuando no hay flujo de corriente, es decir, cuando los terminales están desconectados. Este valor está relacionado con las propiedades del material

semiconductor y la diferencia de energía entre las bandas de conducción y valencia [7], [13].

3. **Punto de Potencia Máxima (MPP)**: en algún punto de la curva I-V, se alcanza la combinación óptima de corriente y voltaje, denominada punto de potencia máxima. Este punto es crucial para evaluar el rendimiento real de la célula, ya que es en este punto donde se genera la máxima potencia útil.

4. **Factor de Llenado (FF)**: el factor de llenado es la relación entre el producto de Isc y Voc, y el MPP. Un mayor factor de llenado implica que la célula es capaz de acercarse a su potencia máxima teórica. Se calcula usando la siguiente ecuación [7], [13]:

$$FF = \frac{I_{MP} \cdot V_{MP}}{I_{SC} \cdot V_{OC}}$$

5. **Eficiencia de Conversión de Potencia (PCE)**: finalmente, la PCE se calcula utilizando la siguiente fórmula [7], [13]:

$$PCE = \frac{I_{SC} \cdot V_{OC} \cdot FF}{P_{entrada}} \cdot 100\%$$



Donde $\mathbf{P_{entrada}}$ es la potencia de la radiación solar incidente. Esta métrica es esencial para comparar el rendimiento entre diferentes tipos de células solares, siendo las células de silicio cristalino las más eficientes actualmente, mientras que las células solares orgánicas presentan PCE menores, pero con potencial de mejora.

Las curvas I-V son una herramienta clave para comprender el rendimiento de las células fotovoltaicas, ya que permiten identificar pérdidas, evaluar la eficiencia bajo diferentes condiciones y ayudar en el diseño de mejoras en la estructura de las células.

## 2.2 Fundamentos del Aprendizaje Automático

El aprendizaje automático o *machine learning* es un campo de la inteligencia artificial que permite a los sistemas aprender de los datos y mejorar su desempeño sin necesidad de ser programados explícitamente. En el contexto de la caracterización de células solares, el *machine learning* puede ser una herramienta poderosa para analizar grandes volúmenes de datos generados por los experimentos y pruebas de rendimiento, permitiendo identificar patrones y realizar predicciones más precisas.

El proceso de caracterización de las células solares implica manejar un conjunto de datos multidimensionales, que incluye variables como la estructura de las células, el comportamiento bajo diferentes condiciones ambientales, y su rendimiento a lo largo del tiempo. Estos datos, en muchos casos, no presentan relaciones lineales simples, lo que dificulta el análisis mediante métodos tradicionales. Aquí es donde entra el *machine learning*, ya que ofrece técnicas avanzadas para modelar relaciones complejas entre variables, optimizar diseños y mejorar la eficiencia de las células solares [15].

Una ventaja clave del *machine learning* es su capacidad para generalizar a partir de los datos. Esto significa que, tras entrenar un modelo con un conjunto de datos de experimentos pasados, el sistema es capaz de predecir cómo se comportarán nuevas células solares bajo diferentes condiciones, permitiendo optimizar sus diseños sin necesidad de realizar pruebas exhaustivas para cada configuración posible.

En el ámbito de la caracterización de células solares, dos algoritmos de *machine learning* destacan por su capacidad de manejar grandes conjuntos de datos y producir predicciones precisas: Random Forest y Gradient Boost. A continuación, se ofrece una descripción detallada de cada uno y de la herramienta ROBERT que se utilizará también en el análisis.

### 2.2.1 Random Forest

Random Forest es un algoritmo de aprendizaje supervisado que se utiliza tanto para clasificación como para regresión. El algoritmo funciona construyendo múltiples árboles de decisión durante el entrenamiento, y la salida final es el promedio (en regresión) o la mayoría de los votos (en clasificación) de los resultados de todos los árboles. Este enfoque de combinar múltiples árboles mejora la precisión del modelo y reduce el riesgo de sobreajuste, ya que cada árbol individual opera de manera independiente en subconjuntos aleatorios de los datos [16].

En el contexto de la caracterización de células solares, Random Forest puede ser utilizado para identificar qué variables o características (como la estructura de la célula o las condiciones ambientales) tienen el mayor impacto en el rendimiento de las células. Al analizar grandes volúmenes de datos, este algoritmo puede detectar patrones complejos que son difíciles de identificar con técnicas tradicionales. Además, debido a su capacidad de manejar conjuntos de datos con muchas variables y posibles interacciones, Random Forest es una excelente opción cuando se trata de optimización de diseños [17].

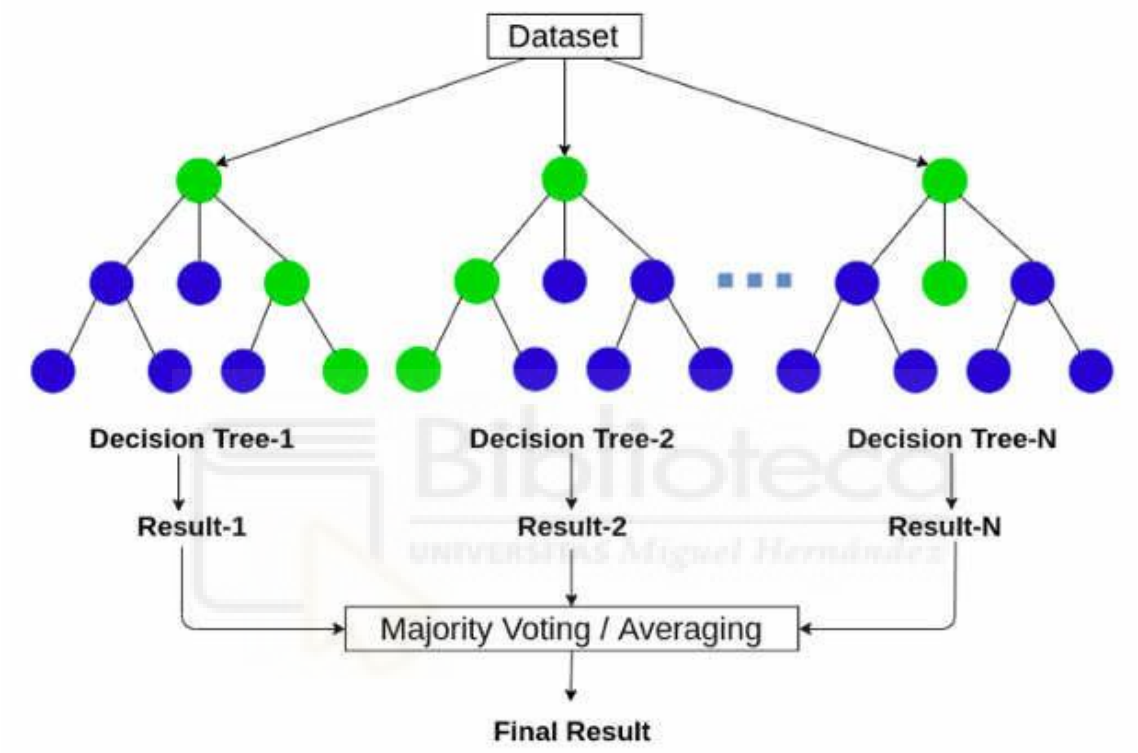



*Figura 6: Diagrama de funcionamiento del algoritmo Random Forest [18].*

## 2.2.2 Gradient Boost

Gradient Boost (o también conocido como Gradient Boosting) es otro algoritmo de aprendizaje supervisado, pero a diferencia de Random Forest, se enfoca en crear múltiples modelos "débiles" secuencialmente, donde cada modelo nuevo intenta corregir los errores cometidos por los anteriores. Este enfoque iterativo permite mejorar el desempeño del modelo de forma incremental, lo que lo convierte en una poderosa herramienta para problemas complejos donde se requiere precisión [19].

En el caso de la caracterización de células solares, Gradient Boosting puede ser particularmente útil cuando se desea predecir con precisión el impacto de diversas características de las células (como materiales o geometría) en su eficiencia bajo distintas

condiciones. Su capacidad para corregir errores sucesivos permite obtener un modelo altamente preciso, incluso en situaciones donde los datos son ruidosos o presentan muchas variaciones.

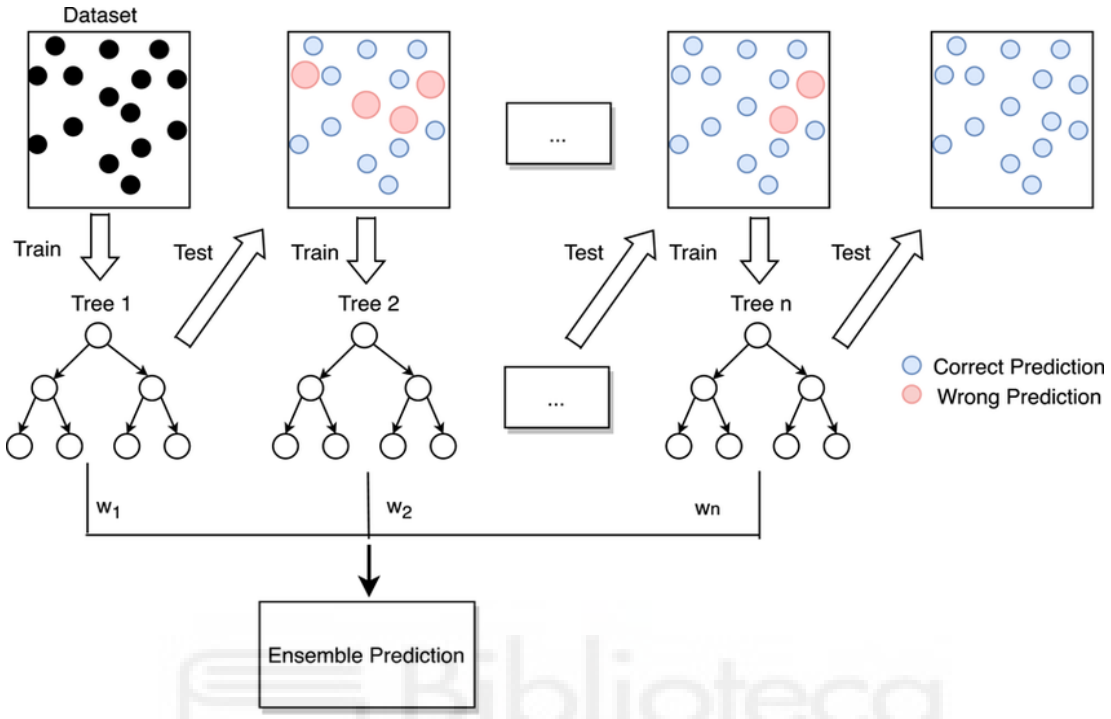



*Figura 7: Diagrama de funcionamiento del algoritmo Gradient Boost [20].*

## 2.2.3 Métricas de Ajuste y Error

En cualquier trabajo relacionado con modelos predictivos, resulta fundamental evaluar la calidad y precisión de las predicciones obtenidas. Esto implica el uso de métricas estadísticas que permitan cuantificar el grado de ajuste del modelo a los datos reales. En este contexto, las métricas más comúnmente utilizadas son el coeficiente de determinación ($R^2$), el error absoluto medio (MAE, "*Mean Absolute Error*") y la raíz del error cuadrático medio (RMSE, "*Root Mean Squared Error*"). A continuación, se describen brevemente estas métricas:

1. **Coeficiente de determinación ($R^2$):** Es una métrica que mide la proporción de la variabilidad total de los datos que es explicada por el modelo [21]. Sus valores varían entre 0 y 1, siendo más cercano a 1 cuando el modelo se ajusta mejor a los observados. Un valor de $R^2$ cercano a 0 indica que el modelo tiene una capacidad muy limitada para explicar la variabilidad de los datos [21].

2. **Error Absoluto Medio (MAE):** Es una métrica que mide la magnitud promedio de los errores en las predicciones, sin considerar la dirección. Se calcula como el promedio de las diferencias absolutas entre los valores reales y los predichos [21]. El MAE proporciona una interpretación clara, ya que indica directamente la magnitud promedio del error en las unidades originales de la variable que se estudia. Un MAE cercano a cero es indicativo de un modelo con predicciones precisas [21].
3. **Raíz del Error Cuadrático medio (RMSE):** Es una métrica que mide la magnitud promedio de los errores, asignando un peso mayor a errores más grandes debido al efecto del cuadrado en su cálculo [21]. Se obtiene tomando la raíz cuadrada del promedio de los errores al cuadrado entre los valores reales y predichos. Al igual que el MAE, valores cercanos a cero indican mayor precisión, aunque el RMSE es más sensible a la presencia de errores atípicamente grandes [21].

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

donde:

- $y_i$: valor observado.
- $\hat{y}_i$: valor predicho por el modelo.
- $n$: número total de observaciones.
- $\bar{y}$: promedio de los valores observados.

*Figura 8: Formulas para el cálculo de las distintas métricas de error [21].*

La elección de estas métricas permite una evaluación integral del desempeño predictivo, considerando tanto la precisión global del modelo como la sensibilidad ante errores significativos.

### 2.2.4 ROBERT

El módulo ROBERT es una herramienta integral que facilita el análisis, la generación y la validación de modelos de *machine learning*, especialmente en el contexto de la caracterización de datos complejos como células solares orgánicas [22], [23]. Su enfoque está en la simplificación del manejo de datos, la optimización de modelos predictivos, y la verificación rigurosa de resultados. ROBERT genera un PDF con los resultados obtenidos.

A continuación, se describe brevemente cada uno de sus submódulos clave:

1. **Curate**: el módulo "*Curate*" mejora la calidad de los datos mediante la eliminación de variables correlacionadas, ruido, y entradas duplicadas. También facilita la conversión de variables categóricas en numéricas o en codificación "*one-hot*". Esto reduce la complejidad de los predictores resultantes y garantiza una representación óptima de los datos.

2. **Generate**: "*Generate*" explora diversas combinaciones de algoritmos de *machine learning* y tamaños de partición de datos, utilizando modelos incorporados de *scikit-learn* y su acelerador *scikit-learn-intelex*. Mediante la optimización de hiperparámetros, este módulo genera modelos altamente precisos y realiza un análisis de la importancia de las características utilizando *Permutation Feature Importance* (PFI), permitiendo filtrar los descriptores menos influyentes.

3. **Predict**: el módulo "*Predict*" utiliza los modelos generados previamente para calcular métricas clave de evaluación como $R^2$, MAE y RMSE para tareas de regresión, o precisión y F1 score para clasificación. También realiza predicciones sobre datos externos y analiza la importancia de las características utilizando los métodos PFI y SHAP,, los cuales se explican más adelante. Además, permite la detección de valores atípicos mediante la medición de errores absolutos en las predicciones.

4. **Verify**: como muestra la imagen, el módulo "*Verify*" realiza una validación exhaustiva de los modelos mediante pruebas como la de la media de y ("*y-mean*"), el barajado de y ("*y-shuffle*"), y la codificación "*one-hot*". Estos métodos permiten evaluar la robustez del modelo y detectar posibles sobreajustes o subajustes. Los

resultados se codifican por colores: azul indica una prueba aprobada y rojo un fallo.

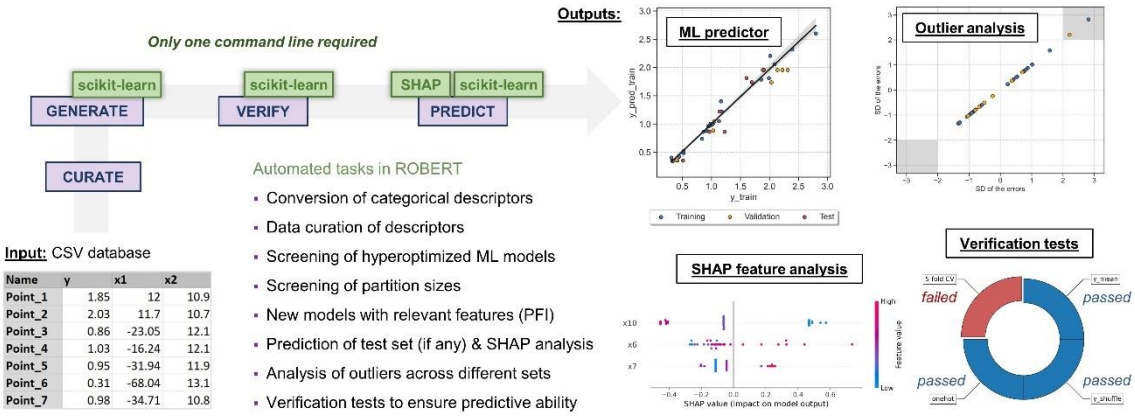



*Figura 9: Muestra el desarrollo completo que realiza ROBERT, obtenido de la documentación [22].*

### 2.2.4.1 SHAP (*Shapley Additive Explanations*)

SHAP es una herramienta avanzada utilizada para explicar el impacto de las características en los modelos predictivos. A través de la teoría de juegos, SHAP asigna un valor de "contribución" a cada característica del modelo, mostrando cómo las entradas individuales afectan a las predicciones. Esto proporciona una interpretación más detallada de cómo los modelos de *machine learning* toman decisiones, lo cual es crucial para asegurar la transparencia en la caracterización de células solares [24].

# 3. ESTADO DEL ARTE

En los últimos años, la investigación sobre células solares orgánicas ha experimentado un crecimiento significativo, impulsada por la necesidad de desarrollar tecnologías fotovoltaicas más eficientes, económicas y sostenibles. Este apartado explora los avances recientes en la caracterización de OSC, abordando tanto las innovaciones tecnológicas como las estrategias de análisis que han permitido un mayor entendimiento de sus propiedades y limitaciones. Se prestará especial atención a las nuevas técnicas que han mejorado la caracterización de las OSC, como la espectroscopía avanzada, las mediciones de transporte de carga, y las herramientas de simulación a nivel molecular [25], [26].

Además, este apartado analiza cómo la incorporación de técnicas de *machine learning* ha comenzado a transformar la forma en que se estudian y optimizan las OSC. El aprendizaje automático se ha convertido en una herramienta poderosa para analizar grandes volúmenes de datos experimentales, facilitando la identificación de patrones complejos y permitiendo la predicción del comportamiento celular bajo diferentes condiciones operativas. Estos enfoques están permitiendo tanto mejorar la eficiencia de las células, como optimizar su estabilidad y durabilidad [26], [27].

Finalmente, se discutirán las principales limitaciones y desafíos que persisten en el desarrollo de las OSC. A pesar de los avances, la estabilidad a largo plazo y la eficiencia en condiciones ambientales reales siguen siendo barreras clave para su adopción comercial. Este apartado abordará los problemas asociados con la degradación de las células, las dificultades para escalar su producción, y los retos que enfrentan los modelos de *machine learning* en términos de precisión y capacidad de generalización.

Este análisis integral proporcionará una visión clara del estado actual de la investigación y señalará los caminos futuros para superar las barreras tecnológicas que aún limitan el potencial de las OSC en el mercado energético global.

## 3.1 Avances Recientes en la Caracterización de Células Solares Orgánicas

En los últimos años, la caracterización de las células solares orgánicas ha evolucionado considerablemente, impulsada por el desarrollo de nuevas arquitecturas de dispositivos y la incorporación de materiales avanzados. Uno de los avances más destacados es el uso de aceptores no basados en fullerenos (*Non-Fullerene Acceptors*, NFAs), que ha permitido superar las limitaciones de los aceptores de fullerenos tradicionales, alcanzando eficiencias de conversión de energía superiores al 19%. Estas OSC de alta eficiencia se han beneficiado de arquitecturas en tándem y ternarias, las cuales optimizan la absorción de luz y reducen las pérdidas energéticas [28].

Los enfoques de caracterización se han centrado en técnicas como la espectroscopía avanzada y la microscopía de fuerza atómica, que han permitido un estudio más detallado de la morfología de las capas activas y las interfaces entre los materiales donadores y aceptores. El ajuste de la morfología a nano escala sigue siendo un factor crucial para mejorar la separación de cargas y reducir las recombinaciones [27].

Además, en los últimos años, la inteligencia artificial, y más específicamente el aprendizaje automático, ha comenzado a desempeñar un papel clave en la optimización de materiales y procesos. Los modelos de ML son utilizados para analizar grandes volúmenes de datos experimentales, lo que permite predecir propiedades clave de los materiales y seleccionar combinaciones de materiales prometedoras sin la necesidad de realizar pruebas exhaustivas. Esto acelera el proceso de innovación y mejora la eficiencia de las OSC al predecir comportamientos bajo diferentes condiciones operativas [27], [28].

Finalmente, la fabricación de dispositivos de gran escala ha progresado significativamente. Algunas OSC con áreas activas superiores a 1 $cm^2$ están logrando niveles de rendimiento mejorados gracias a nuevas técnicas de procesamiento que mantienen la homogeneidad en la fabricación, lo que permite una transición más fluida hacia la producción comercial [28].

## 3.2 Limitaciones y Desafíos Actuales

A pesar de los avances significativos en la investigación y desarrollo de células solares orgánicas, persisten varias limitaciones y desafíos que dificultan su comercialización a gran escala y su adopción generalizada. Entre los principales obstáculos se encuentran la estabilidad y durabilidad de las OSC, la variabilidad de los datos experimentales y las dificultades relacionadas con la implementación de técnicas de *machine learning*.

Uno de los problemas más notables en las OSC es la estabilidad a largo plazo. Las células solares orgánicas suelen degradarse más rápidamente que sus contrapartes inorgánicas, lo que reduce su vida útil bajo condiciones de operación reales. Las OSC son particularmente susceptibles a la degradación por exposición a la humedad, oxígeno y radiación ultravioleta. Aunque se han desarrollado estrategias para mejorar la estabilidad, como la encapsulación y la optimización de los materiales, la durabilidad sigue siendo un obstáculo importante para la comercialización [12], [28].

Otro desafío crítico es la eficiencia de conversión de energía en condiciones reales. Aunque en laboratorio se han alcanzado eficiencias de hasta el 19%, estos resultados no siempre se traducen en un rendimiento similar en el campo, donde factores como la temperatura y la variabilidad en la radiación solar afectan negativamente la eficiencia de las OSC. Las células solares deben demostrar su rendimiento en una variedad de condiciones ambientales para ser consideradas una alternativa viable a otras tecnologías fotovoltaicas [28].

Desde la perspectiva del *machine learning*, los desafíos incluyen la complejidad de los datos y la diversidad de las estructuras químicas de los materiales orgánicos. El *machine learnin*g se enfrenta a la dificultad de construir modelos que puedan generalizar correctamente para diferentes combinaciones de materiales y configuraciones de dispositivos. Además, la necesidad de grandes conjuntos de datos experimentales para entrenar los modelos limita su implementación eficiente en muchos laboratorios [27].

Finalmente, la escala de producción de OSC presenta desafíos técnicos y económicos. A pesar de los avances en la fabricación, el proceso sigue siendo costoso, y la producción de dispositivos grandes y homogéneos sigue siendo un reto. Además, el uso de disolventes y procesos que no siempre son respetuosos con el medio ambiente también plantea dificultades para la escalabilidad industrial [28].

Estos desafíos, aunque significativos, están siendo abordados por investigadores de todo el mundo, y las soluciones que surjan en los próximos años determinarán el futuro de las OSC como una alternativa viable en el sector de las energías renovables.

# 4. MATERIAL Y MÉTODOS

En esta sección se describen los procedimientos y técnicas empleadas para la caracterización automática de OSC mediante inteligencia artificial. Para llevar a cabo estos experimentos, se emplearon células solares orgánicas con diferentes estructuras, adaptadas para evaluar el impacto de las variaciones en los materiales y parámetros utilizados. Las estructuras base de las células se presentan en dos configuraciones principales: (ITO/PEDOT:PSS/P3HT:PCBM/Al) y (ITO/PEDOT:PSS/rGO /P3HT:PCBM/Al), según el tipo de experimento. Estas configuraciones permiten explorar cómo la cantidad de PEDOT:PSS, las proporciones de P3HT:PCBM, el tipo de dispersante para el rGO y la temperatura de dispersión afectan a la eficiencia y estabilidad de las células solares.

A lo largo de este apartado, se detallan los pasos seguidos desde la adquisición de los datos experimentales, la preparación y acondicionamiento de estos datos, hasta la implementación de algoritmos de *machine learning* para su análisis. Además, se incluye una descripción detallada de los experimentos realizados, en los cuales se variaron diversos parámetros de las células solares, como la cantidad de PEDOT:PSS, las proporciones de P3HT:PCBM, el tipo de dispersante utilizado para el rGO y la temperatura de dispersión, con el objetivo de observar cómo estas variaciones afectan a la eficiencia de las células.

Los experimentos se llevaron a cabo en diferentes fechas y bajo condiciones controladas para evaluar el impacto de varias modificaciones en la eficiencia de las células solares orgánicas. Asimismo, se repitieron las medidas periódicamente para estudiar la influencia con el paso del tiempo. Para llevar a cabo estos experimentos, se emplearon células s

A continuación, se describen los principales experimentos realizados:

1. **Variación de la cantidad de PEDOT:PSS (27/10/2021)**:
   **Estructura**: [ITO/PEDOT:PSS/P3HT:PCBM/Al], [P3HT:PCBM]: [1:1]
   Se prepararon células solares con diferentes cantidades de PEDOT:PSS**:** (0.25 mL, 0.5 mL, 0.75 mL y 1 mL) aplicadas mediante *spin-coating*. La eficiencia de conversión de energía de cada célula se midió periódicamente para determinar la cantidad óptima de PEDOT:PSS. Asimismo, se repitieron las medidas

periódicamente para estudiar la influencia con el paso del tiempo para poder identificar cada célula se especifica las características de cada una:

- Célula 1: 0.25 mL
- Célula 2: 0.50 mL
- Célula 3: 0.75 mL
- Célula 4: 1.00 mL

2. **Proporciones de P3HT:PCBM (04/11/2021)**:

   **Estructura**: [ITO/PEDOT:PSS/P3HT:PCBM/Al]

   Se evaluaron diversas proporciones de P3HT:PCBM ([1.2:1], [1.1:1], [1:1], [1:0.9], [1:0.8]) para identificar la proporción que maximiza la eficiencia de las células solares. Para poder identificar cada célula se especifica las características de cada una:

   - Célula 1: P3HT:PCBM [1.2:1]
   - Célula 2: P3HT:PCBM [1.1:1]
   - Célula 3: P3HT:PCBM [1:1]
   - Célula 4: P3HT:PCBM [1:0.9]
   - Célula 5: P3HT:PCBM [1:0.8]

3. **Dispersión del rGO (11/11/2021)**:

   **Estructura**: [ITO/PEDOT:PSS/rGO/P3HT:PCBM/Al], [P3HT:PCBM] : [1:0.8], [PEDOT:PSS/rGO] : [1:1]

   Se investigó el efecto de diferentes dispersantes (metanol, dimetilformamida, agua) en la dispersión del rGO y su influencia en la eficiencia de las células solares. Para poder identificar cada célula se especifica las características de cada una:

   - Célula 1: Célula de referencia sin rGO.
   - Célula 2: rGO disperso en metanol.
   - Célula 3: rGO disperso en dimetilformamida.
   - Célula 4: rGO disperso en agua.

4. **Temperatura de dispersión del rGO (26/11/2021)**:
   **Estructura**: [ITO/PEDOT:PSS/rGO/P3HT:PCBM/Al], [P3HT:PCBM] : [1:0.8], [PEDOT:PSS/rGO] : [1:1]
   Se estudió cómo la temperatura de la solución de rGO y agua (25 ºC, 50 ºC, 75 ºC) afecta la dispersión del rGO y, consecuentemente, la eficiencia de las células solares. Para poder identificar cada célula se especifica las características de cada una:
   - Célula 1: Célula de referencia sin rGO.
   - Célula 2: Temperatura de solución 25 ºC
   - Célula 3: Temperatura de solución 50 ºC
   - Célula 4: Temperatura de solución 75 ºC

5. **Temperatura de dispersión de rGO + PEDOT (10/12/2021)**:
   **Estructura**: [ITO/PEDOT:PSS/rGO/P3HT:PCBM/Al], [P3HT:PCBM] : [1:0.8], [PEDOT:PSS/rGO] : [1:1]
   Se analizó el efecto de calentar la mezcla de rGO y PEDOT a diferentes temperaturas (25 ºC, 50 ºC, 75 ºC) antes de aplicarla a las células solares para mejorar la dispersión del rGO y su eficiencia. Para poder identificar cada célula se especifica las características de cada una:
   - Célula 1: Célula de referencia sin rGO.
   - Célula 2: Temperatura de solución 25 ºC
   - Célula 3: Temperatura de solución 50 ºC
   - Célula 4: Temperatura de solución 75 ºC

Con estos experimentos se pretende identificar las condiciones óptimas para la fabricación de células solares orgánicas más eficientes, proporcionando datos cruciales para su caracterización mediante técnicas de inteligencia artificial. Para ello se han realizado medidas periódicas que permiten analizar la evolución de su rendimiento a lo largo del tiempo.

## 4.1 Adquisición de Datos

En nuestra investigación, nos enfocamos en la eficiencia de conversión de energía de las células solares orgánicas, ya que es un parámetro clave para evaluar su rendimiento. Para calcular la PCE, es esencial obtener las curvas de corriente-voltaje (I-V) de las células solares. Estas curvas nos permiten determinar parámetros cruciales como la corriente de cortocircuito (Isc), el voltaje de circuito abierto (Voc), y el punto de máxima potencia (MPP), entre otros. A partir de estos valores, y mediante los cálculos explicados previamente en el MARCO TEÓRICO, podemos derivar el PCE de cada célula solar estudiada.

En el laboratorio, utilizamos un sistema automatizado para obtener las curvas de corriente-voltaje (I-V) de las células solares. El equipo consta de un ordenador con el software Keithley, específicamente una versión de LabVIEW 18, conectado a un KEITHLEY 2400 SourceMeter, una lámpara de arco de xenón y un filtro AM 1.5G. Este conjunto simula las condiciones de exposición solar terrestre (100 mW/cm², 25 ºC y AM 1.5G).

Para llevar a cabo las mediciones, configuramos la fuente controlada Keithley con 100 puntos de medida en intervalos de -0,1 V a 1 V. Inicialmente, se realiza una medición en oscuridad para establecer una línea base. Posteriormente, se ejecutan las mediciones bajo iluminación. Este proceso permite medir la respuesta de la célula solar al barrido de tensión, con una retroalimentación constante de mediciones que focaliza el rango de tensión de interés, esencial para observar el comportamiento característico de un diodo.

Los datos obtenidos de cada muestra se almacenan en archivos “.txt” tal y como aparecen en la imagen (Figura 10).

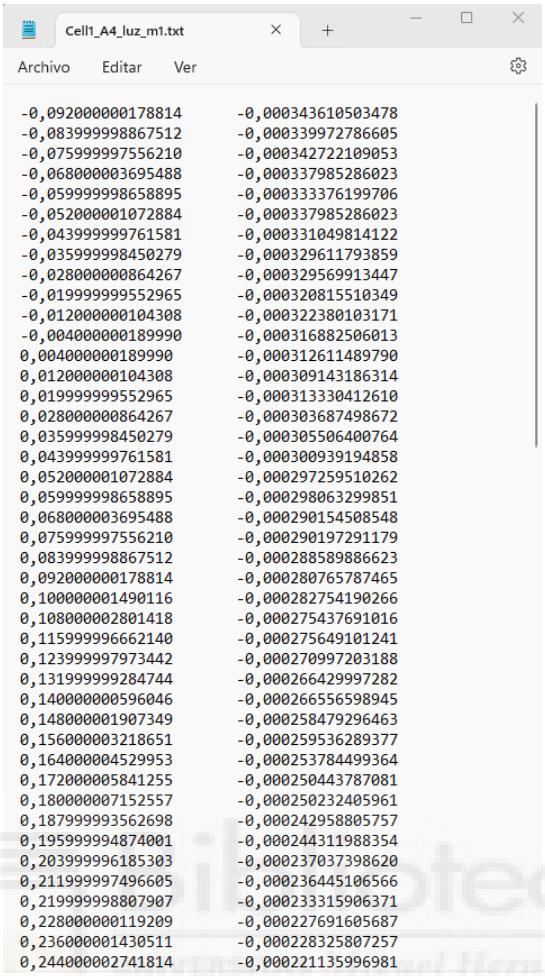



*Figura 10: Muestra de cómo se guardan los datos obtenidos por las mediciones.*

## 4.2 Preparación de Datos para Análisis

El preprocesamiento de datos es un paso crucial en el trabajo, ya que garantiza que los datos utilizados en los análisis sean precisos y consistentes. Para lograr esto, empleamos distintos *scripts* en MATLAB, se incluye un ejemplo en los anexos, que nos permite separar todas las células de los experimentos y calcular sus respectivas eficiencias de conversión de energía (PCEs). Este *script* procesa las curvas de corriente-voltaje (I-V) obtenidas y calcula los parámetros eléctricos clave de cada célula, almacenando los resultados en archivos ".mat" para un acceso y análisis posterior.

El *script* de MATLAB realiza las siguientes tareas:

1. **Separación de datos**: divide los datos experimentales en subconjuntos correspondientes a cada célula solar.
2. **Cálculo de PCE**: utiliza las curvas I-V para calcular la eficiencia de conversión de energía de cada célula, basándose en los métodos detallados en el Marco Teórico.
3. **Almacenamiento de datos**: guarda los datos procesados, incluidos los parámetros calculados, en archivos ".mat", facilitando su manejo y análisis posterior.

Una vez obtenidos los archivos ".mat", se procedió al preprocesamiento de datos utilizando Python. Este proceso incluyó varios pasos clave que se detallan a continuación.

### 4.2.1 Carga y Filtrado de Datos

El primer paso en el preprocesamiento de datos consistió en la visualización inicial de los datos para identificar y eliminar mediciones incorrectas. Se observaron mediciones con valores cero o muy dispares en comparación con los obtenidos en días cercanos, las cuales fueron eliminadas para asegurar la integridad del conjunto de datos.

A continuación, los archivos en formato ".mat" fueron transformados a archivos CSV para facilitar su posterior visualización y almacenamiento. Esta transformación se llevó a cabo leyendo los valores en Python con la función "scipy.io.loadmat", almacenándolos en un "DataFrame" de pandas y guardándolos en formato CSV con la fecha en la que se tomaron las muestras de cada experimento.

Una vez obtenidos los archivos CSV, se observó que los valores de tiempo estaban guardados como el tiempo transcurrido entre muestras en días. Para mejorar la

visualización y hacer los datos más intuitivos, esta columna se transformó a la fecha exacta de cada día en que se midieron las muestras mediante un *script* de Python.

En la siguiente tabla (Tabla 1) se presenta un fragmento de uno de los archivos CSV después de la limpieza y transformación realizadas.

| Fecha datos | célula | PCE (%) | rGO | cantidad DS HTL | Cantidad PEDOT:PSS | ratio P3HT:PCBM | Ta PEDOT:PSS | Ta rGO | disolución rGO |
|---|---|---|---|---|---|---|---|---|---|
| 04/11/2021 | 1 | 0.51142 | 0 | 0.5 | 0.5 | 1.2 | 25 | 0 | 0 |
| 04/11/2021 | 2 | 0.61682 | 0 | 0.5 | 0.5 | 1.1 | 25 | 0 | 0 |
| 04/11/2021 | 3 | 0.68237 | 0 | 0.5 | 0.5 | 1 | 25 | 0 | 0 |
| 04/11/2021 | 4 | 0.53104 | 0 | 0.5 | 0.5 | 1.11 | 25 | 0 | 0 |
| 04/11/2021 | 5 | 1.28911 | 0 | 0.5 | 0.5 | 1.25 | 25 | 0 | 0 |
| 11/11/2021 | 1 | 0.45504 | 0 | 0.5 | 0.5 | 1.2 | 25 | 0 | 0 |
| 11/11/2021 | 2 | 0.53163 | 0 | 0.5 | 0.5 | 1.1 | 25 | 0 | 0 |
| 11/11/2021 | 3 | 0.64917 | 0 | 0.5 | 0.5 | 1 | 25 | 0 | 0 |
| 11/11/2021 | 4 | 0.47007 | 0 | 0.5 | 0.5 | 1.11 | 25 | 0 | 0 |
| 11/11/2021 | 5 | 1.22812 | 0 | 0.5 | 0.5 | 1.25 | 25 | 0 | 0 |
| 17/11/2021 | 1 | 0.42438 | 0 | 0.5 | 0.5 | 1.2 | 25 | 0 | 0 |
| 17/11/2021 | 2 | 0.49906 | 0 | 0.5 | 0.5 | 1.1 | 25 | 0 | 0 |
| 17/11/2021 | 3 | 0.62497 | 0 | 0.5 | 0.5 | 1 | 25 | 0 | 0 |
| 17/11/2021 | 4 | 0.44081 | 0 | 0.5 | 0.5 | 1.11 | 25 | 0 | 0 |
| 17/11/2021 | 5 | 1.18218 | 0 | 0.5 | 0.5 | 1.25 | 25 | 0 | 0 |
| 24/11/2021 | 1 | 0.40174 | 0 | 0.5 | 0.5 | 1.2 | 25 | 0 | 0 |
| 24/11/2021 | 2 | 0.48215 | 0 | 0.5 | 0.5 | 1.1 | 25 | 0 | 0 |
| 24/11/2021 | 4 | 0.42168 | 0 | 0.5 | 0.5 | 1.11 | 25 | 0 | 0 |
| 24/11/2021 | 5 | 1.13557 | 0 | 0.5 | 0.5 | 1.25 | 25 | 0 | 0 |
| 25/11/2021 | 1 | 0.39934 | 0 | 0.5 | 0.5 | 1.2 | 25 | 0 | 0 |
| 25/11/2021 | 2 | 0.48076 | 0 | 0.5 | 0.5 | 1.1 | 25 | 0 | 0 |
| 25/11/2021 | 3 | 0.59842 | 0 | 0.5 | 0.5 | 1 | 25 | 0 | 0 |
| 25/11/2021 | 4 | 0.4198 | 0 | 0.5 | 0.5 | 1.11 | 25 | 0 | 0 |
| 25/11/2021 | 5 | 1.1295 | 0 | 0.5 | 0.5 | 1.25 | 25 | 0 | 0 |

*Tabla 1: Fragmento del CSV de uno de los experimentos.*

En los distintos experimentos, hay valores que se mantendrán constantes, como es el caso del rGO (grafeno), que en la tabla anterior siempre será 0 (1 si la célula contiene grafeno) dado que en este experimento no se añadió grafeno a las muestras. Sin embargo, se mantienen estos valores para poder comparar las medidas entre distintos experimentos.

### 4.2.2 Remuestreo e Interpolación de Datos

En esta fase del preprocesamiento, se abordó la irregularidad en la toma de datos. Los datos originales se recogieron en fechas no uniformes, lo que podría afectar la calidad del análisis. Para solucionar este problema, se decidió realizar un remuestreo de las fechas y una interpolación de las PCE para obtener curvas más adecuadas y continuas.

Primero, se ajustaron las fechas de medición a intervalos regulares. Para ello se utilizó el método "pandas.DataFrame.resample", con el cual se obtuvieron valores con una frecuencia de siete días entre ellos, a partir de la fecha inicial y final. Este proceso de remuestreo permitió uniformizar los datos, facilitando un análisis más coherente y preciso.

Después del remuestreo, se aplicó una técnica de interpolación a PCEs para rellenar los valores faltantes y suavizar las variaciones. Para ello se utilizó el método "pandas.DataFrame.interpolate", el cual permite a través de la columna PCE obtener una curva más suave y con valores equidistantes a lo largo del conjunto de datos, facilitando así el análisis.

El resto de las variables, que se mantienen constantes para cada célula dentro de cada experimento, se añadieron nuevamente después del proceso de remuestreo e interpolación. Este enfoque aseguró que cada registro en el conjunto de datos final mantuviera todas las características relevantes, sin perder la integridad de la información experimental. Permitiendo la generación de curvas más precisas y útiles para la caracterización de las células solares orgánicas.

En la siguiente figura (Figura 11) se puede observar todos los cambios realizados a las mediciones tomadas, la interpolación de datos faltantes y el suavizado de las curvas. A la izquierda se observan los datos sin procesar, mientras que a la derecha se puede ver cómo gracias a estos procesos contamos con unas curvas más suaves.

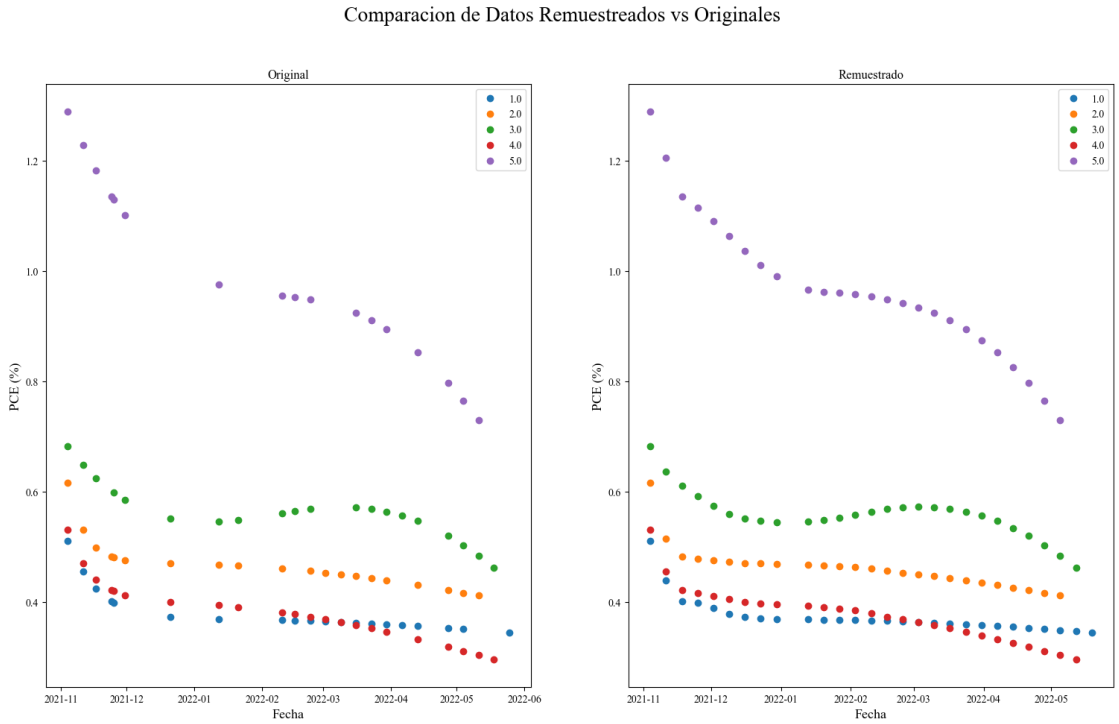



*Figura 11: Gráfica comparativa PCE-Fecha de uno de los experimentos*

### 4.2.3 Integración de Datos Climáticos

Para mejorar el análisis y comprender mejor los factores que influyen en el deterioro de las OSC y la consecuente disminución de su PCE, se decidió incorporar datos climáticos al conjunto de datos experimental. Las variables climáticas consideradas incluyen temperatura, presión, humedad y punto de rocío.

Los datos climáticos fueron obtenidos de la web Wunderground [29], que proporciona información detallada y precisa sobre diversas condiciones meteorológicas. Estos datos fueron recopilados para las mismas fechas en que se realizaron las mediciones de las OSC.

Una vez obtenidos, los datos climáticos se integraron con el conjunto de datos de las células solares mediante un proceso de combinación basado en las fechas de medición. Este proceso asegura que cada registro en el conjunto de datos final contenga tanto las mediciones de las OSC como las condiciones climáticas correspondientes.

La inclusión de estos datos adicionales permite analizar si existe alguna relación entre las variables climáticas y el deterioro de las OSC, así como la disminución de su PCE. Esta

integración de datos proporcionará una visión más completa y permitirá aplicar distintos algoritmos de inteligencia artificial para identificar posibles patrones o correlaciones.

Al incorporar datos climáticos, se espera obtener información valiosa sobre cómo factores externos pueden influir en la degradación de las OSC. Este análisis ayudará a mejorar el diseño y la durabilidad de las células solares orgánicas, promoviendo un mayor rendimiento y estabilidad a largo plazo.

En la tabla (Tabla 2) se puede observar un ejemplo de los datos añadidos al conjunto de datos para una de las células de uno de los experimentos.

| Fecha datos | Temperatura | Punto de rocío | Humedad | Presión |
|---|---|---|---|---|
| 04/11/2021 | 17.22 | 2.78 | 39 | 1007.79 |
| 11/11/2021 | 20 | 3.89 | 35 | 1011.85 |
| 17/11/2021 | 17.78 | 6.11 | 45 | 1014.9 |
| 24/11/2021 | 12.22 | 2.78 | 54 | 1007.79 |
| 25/11/2021 | 15 | 3.89 | 48 | 1002.03 |
| 21/12/2021 | 16.11 | 8.89 | 63 | 1015.92 |
| 12/01/2022 | 12.78 | 7.78 | 72 | 1023.71 |
| 10/02/2022 | 12.22 | 10 | 88 | 1023.71 |
| 16/02/2022 | 17.22 | 1.11 | 34 | 1020.66 |
| 23/02/2022 | 16.11 | 2.78 | 42 | 1021.67 |
| 02/03/2022 | 16.11 | 8.89 | 63 | 1017.95 |
| 09/03/2022 | 15 | 11.11 | 77 | 1017.95 |
| 16/03/2022 | 17.22 | 10 | 63 | 1007.79 |
| 23/03/2022 | 12.78 | 11.11 | 88 | 1014.9 |
| 30/03/2022 | 15 | 12.22 | 82 | 997.97 |
| 06/04/2022 | 12.78 | 6.11 | 63 | 1009.82 |
| 13/04/2022 | 17.22 | 12.78 | 77 | 1004.74 |
| 27/04/2022 | 18.89 | 10 | 56 | 1012.87 |
| 04/05/2022 | 15 | 13.89 | 94 | 1008.81 |
| 25/05/2022 | 22.78 | 6.11 | 33 | 1011.85 |

*Tabla 2: Ejemplo de los datos meteorológicos introducidos.*

## 4.2.4 Normalización de los Datos

La normalización de los datos es un paso crucial en el preprocesamiento para asegurar que todas las variables contribuyan equitativamente al análisis y modelos subsecuentes. Esto se debe a que las variables pueden estar en diferentes escalas y rangos, lo que puede afectar negativamente el rendimiento de ciertos algoritmos de aprendizaje automático y técnicas de análisis.

En este proyecto, utilizaremos herramientas como “StandardScaler” de la biblioteca “scikit-learn” de Python para normalizar nuestros datos. “StandardScaler” estandariza las características eliminando la media y escalando a la varianza unitaria, asegurando que los datos tengan una media de 0 y una desviación estándar de 1 [23]. Esta técnica es especialmente útil para algoritmos que suponen que los datos están centrados en cero.

Por otro lado, cuando se realicen las pruebas con ROBERT, este tiene un módulo integrado que procesa las variables automáticamente. Este módulo llamado CURATE nos filtra los datos y simplifica el trabajo ya que nos da una preparación óptima de los datos [22].

## 4.3 Selección de Características Relevantes

La selección de características relevantes es una etapa crítica en nuestro proyecto de caracterización de OSCs. Dado que nuestro objetivo principal es comprender qué componentes influyen más significativamente en la PCE y su deterioro a largo plazo, identificar las características más influyentes es fundamental. Esto nos permite optimizar las proporciones de ciertos componentes químicos involucrados en la fabricación y evaluar si las células con grafeno, por ejemplo, ofrecen un rendimiento superior o inferior en comparación con otras variantes o bien determinar variables climáticas relevantes.

### 4.3.1 Importancia de la Selección de Características

La selección adecuada de características mejora la precisión de los modelos predictivos, además de reducir la complejidad computacional y evita el sobreajuste. Al identificar qué características son más influyentes en la PCE, podemos orientar la investigación y el desarrollo hacia las áreas que más impactan el rendimiento de las células solares. Además, entender cómo estas características afectan la durabilidad y eficiencia de las células solares puede guiar la innovación en el diseño y fabricación de futuras tecnologías fotovoltaicas.

En nuestro estudio, empleamos técnicas avanzadas de *machine learning* para identificar las características más influyentes en la eficiencia de conversión de energía (PCE) de las células solares. Estas técnicas nos ayudan a evaluar cómo las variaciones en diferentes características afectan a la precisión de los modelos predictivos, permitiéndonos identificar aquellas que son esenciales para mejorar el rendimiento y la durabilidad de las células.

- **Importancia basada en permutación ("*permutation feature importance*", PFI):** integrada directamente en los algoritmos de aprendizaje automático, esta técnica mide el impacto en el rendimiento del modelo cuando los valores de una característica se alteran aleatoriamente. Este método proporciona una medida clara y directa de la importancia relativa de cada característica, destacando aquellas que son críticas para la precisión del modelo [16].
- **Selección Secuencial de Características ("*Sequential Feature Selection*", SFS):** esta técnica selecciona características de manera iterativa, comenzando con un conjunto vacío (o un conjunto completo), y agregando (o eliminando)

características una a una según su capacidad para mejorar el rendimiento del modelo. En cada paso, la característica que contribuye más a la precisión del modelo se incluye (o excluye). Este enfoque permite identificar un subconjunto óptimo de características que maximiza el rendimiento del modelo de manera eficiente, evitando la inclusión de características redundantes o irrelevantes [23].

La implementación de PFI y SFS nos ayuda a identificar las características más relevantes y a comprender cómo interactúan dentro del modelo para influir en la PCE. Al utilizar estas técnicas, aseguramos que nuestra investigación se enfoque en los componentes y las condiciones que maximizan la eficiencia de las células solares, basándonos en una evaluación robusta y relevante proporcionada por los algoritmos. Esta capacidad de desglosar la importancia de las características de manera precisa mejora significativamente nuestra habilidad para optimizar las células solares orgánicas y guiar futuras innovaciones en el diseño y fabricación de tecnologías fotovoltaicas.

### 4.3.2 Selección Automática de Características con ROBERT

Durante el desarrollo del proyecto, encontramos la herramienta ROBERT, un software que utiliza librerías de Python, que facilita considerablemente las tareas de curado de datos, entrenamiento y validación de modelos de manera automática. ROBERT permite automatizar muchos de los procesos involucrados en el análisis de datos, haciendo que el flujo de trabajo sea más eficiente y menos propenso a errores humanos [22].

ROBERT opera mediante una serie de módulos especializados que gestionan diferentes aspectos del proceso de análisis de datos. Más adelante se hablará sobre los distintos módulos, ahora es interesante mencionar el módulo GENERATE.

Este módulo realiza un ajustado de los descriptores que tienen un bajo PFI generando dos modelos distintos: uno que considera estos descriptores y otro que los elimina. De esta forma, se puede comparar los resultados obtenidos en ambos casos para determinar cuál es más preciso.

Además, el módulo GENERATE crea mapas de color que muestran las correlaciones entre las distintas variables para ambos modelos. Estos mapas de color ayudan a identificar el grado de correlación entre nuestras variables, lo cual es crucial, ya que no es beneficioso añadir descriptores altamente correlacionados al modelo, dado que pueden introducir redundancia y aumentar el error.

Este módulo facilita significativamente el análisis de los parámetros al identificar cuáles son los más relevantes para mantener en el modelo y asegurar que la inclusión de información adicional no introduzca errores. Al automatizar este proceso, ROBERT optimiza la selección de características y mejora la calidad del modelo predictivo final.

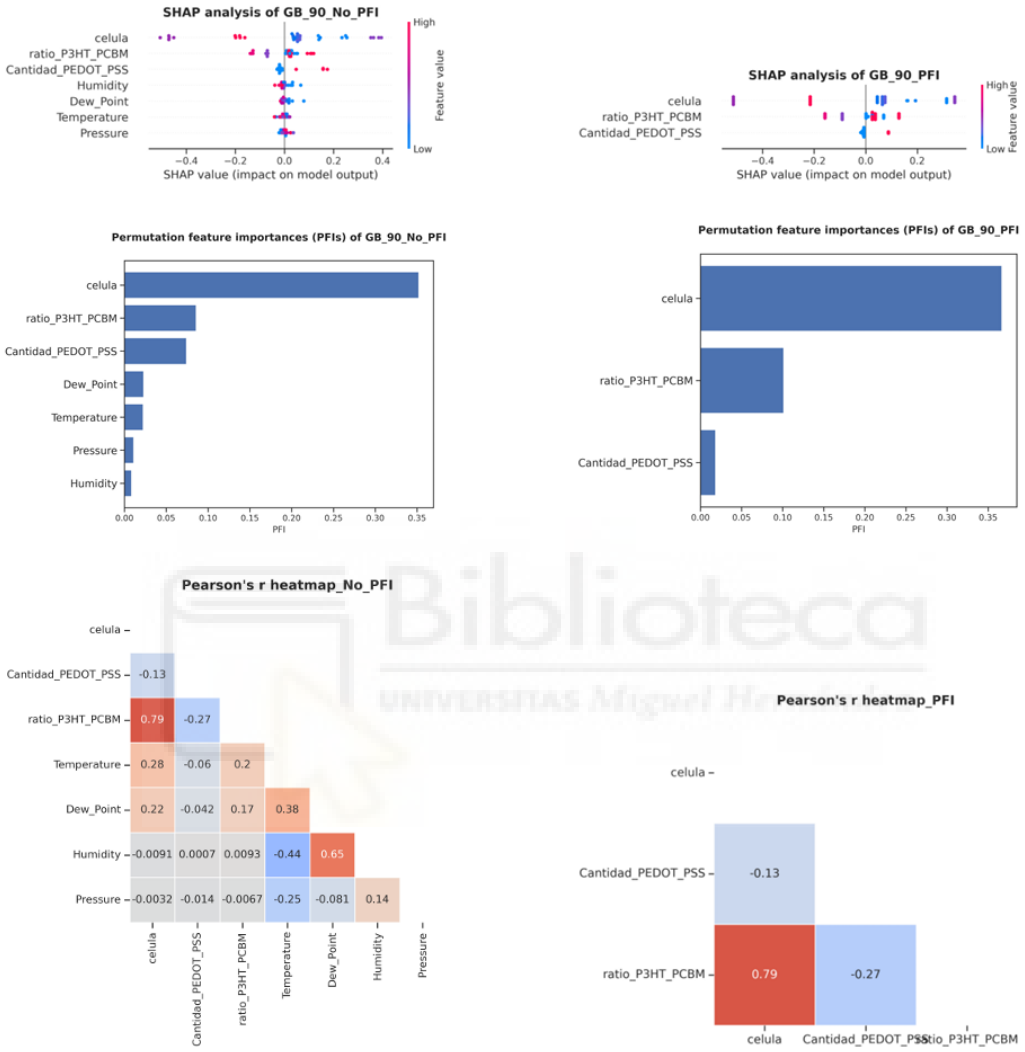



*Figura 12: Ejemplo de cómo muestra la importancia de las características el módulo ROBERT [30].*

## 4.4 Validación y Evaluación de los Datos Procesados

La validación y evaluación de los datos procesados es una etapa crucial en cualquier proyecto de análisis de datos, ya que garantiza la precisión y relevancia de los resultados obtenidos. Esta sección es fundamental para asegurar que los modelos desarrollados no solo sean precisos, sino que también sean generalizables a nuevos datos. Una adecuada validación y evaluación permiten identificar posibles sesgos, errores y áreas de mejora, lo que contribuye significativamente a la robustez y fiabilidad del modelo final.

Para evaluar el rendimiento de los modelos y verificar la precisión de las predicciones, se emplean métricas estándar de evaluación, incluyendo:

- **$R^2$ (Coeficiente de Determinación):** Esta métrica mide la calidad del ajuste de un modelo a sus datos, indicando qué proporción de la variación en la variable dependiente puede explicarse a través de las variables independientes.
- **MAE (Error Absoluto Medio):** MAE cuantifica la diferencia media entre los valores observados y los predichos por los modelos de ML obtenidos, ofreciendo una perspectiva del error real en términos de unidades de interés.
- **MSE (Error Cuadrático Medio):** MSE proporciona una medida de la magnitud del error, penalizando más los errores grandes al elevar al cuadrado las diferencias antes de promediarlas.

Estas métricas son cruciales para evaluar la efectividad de los modelos. En caso de obtener resultados subóptimos, es posible ajustar el conjunto de datos, los hiperparámetros del modelo, o la proporción de datos utilizados para las fases de prueba, entrenamiento y validación. Aquí, la herramienta ROBERT juega un papel interesante, ya que automatiza la prueba de diferentes hiperparámetros y configuraciones de datos para minimizar el error promedio.

Además de separar los datos en conjuntos de entrenamiento, test y validación —lo cual permite retener una porción de los datos desconocida para el modelo y verificar así su capacidad de generalización—, se utilizan técnicas adicionales para una validación más profunda:

- **Validación Cruzada**: Esta técnica implica dividir los datos en subconjuntos (*folds*), entrenando modelos en cada uno y evaluando su rendimiento en el subconjunto restante. Esto ayuda a garantizar la precisión y fiabilidad de los modelos [31]. La siguiente ilustración (Figura 13) muestra un esquema de cómo funciona la validación cruzada.

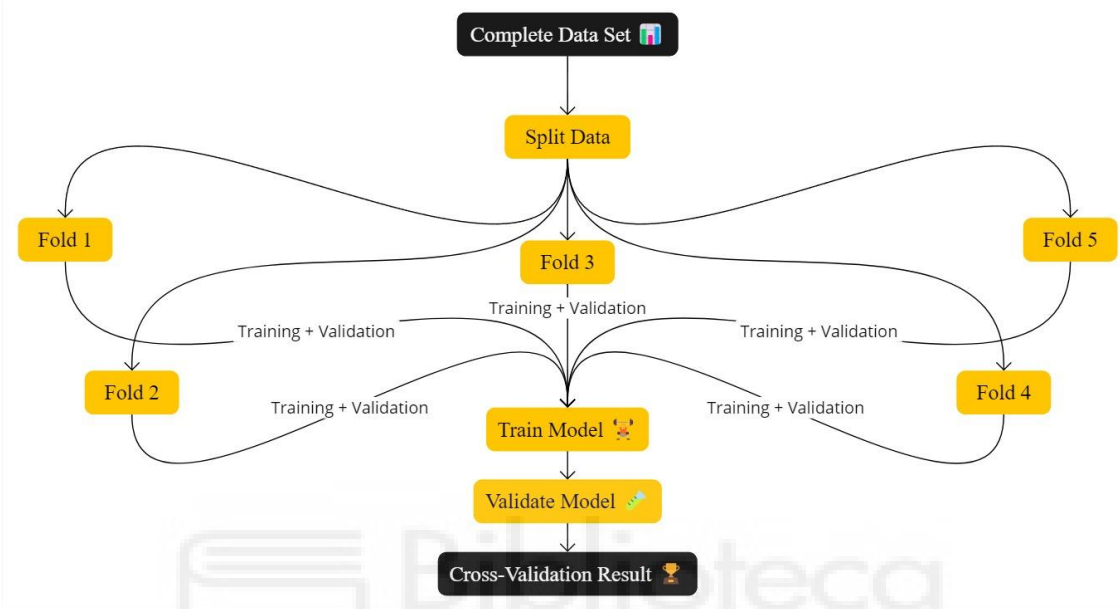



*Figura 13: Diagrama del funcionamiento de la validación cruzada.(Creado con Mermaid y Miro) [32], [33].*

ROBERT aplica la validación cruzada, incorporando a su vez tres métodos de validación innovadores:

1. **Precisión de referencia:** se evalúa la precisión del modelo asumiendo que siempre predice la media de los datos (en regresión) o la clase mayoritaria (en clasificación).
2. **Precisión con codificación one-hot:** se utiliza la codificación one-hot para las variables categóricas y se convierten todos los valores X distintos de cero a uno (incluyendo NaN a cero), y luego se evalúa la precisión del modelo.
3. **Precisión con permutación aleatoria de y:** los valores de la variable objetivo (y) se barajan aleatoriamente en el conjunto de validación para evaluar la precisión del modelo bajo condiciones de incertidumbre

Estas técnicas avanzadas proporcionan una comprensión más profunda de la capacidad del modelo para hacer predicciones precisas y fiables bajo diferentes condiciones,

aumentando la confianza en los resultados obtenidos y en la utilidad del modelo en aplicaciones prácticas.

## 4.5 Metodología de Procesamiento de Datos

En esta sección se describe el proceso llevado a cabo con los datos recopilados durante los experimentos para analizar y modelar la PCE) de las células solares orgánicas. Dado que los datos adquiridos provienen de diversos experimentos con condiciones variables, fue necesario aplicar técnicas de análisis exploratorio y selección de características para comprender la naturaleza de estos datos y determinar las variables que influyen significativamente en la PCE. Además, se emplearon modelos predictivos y metodologías de validación para garantizar la fiabilidad de los resultados obtenidos y optimizar el rendimiento de las células solares.

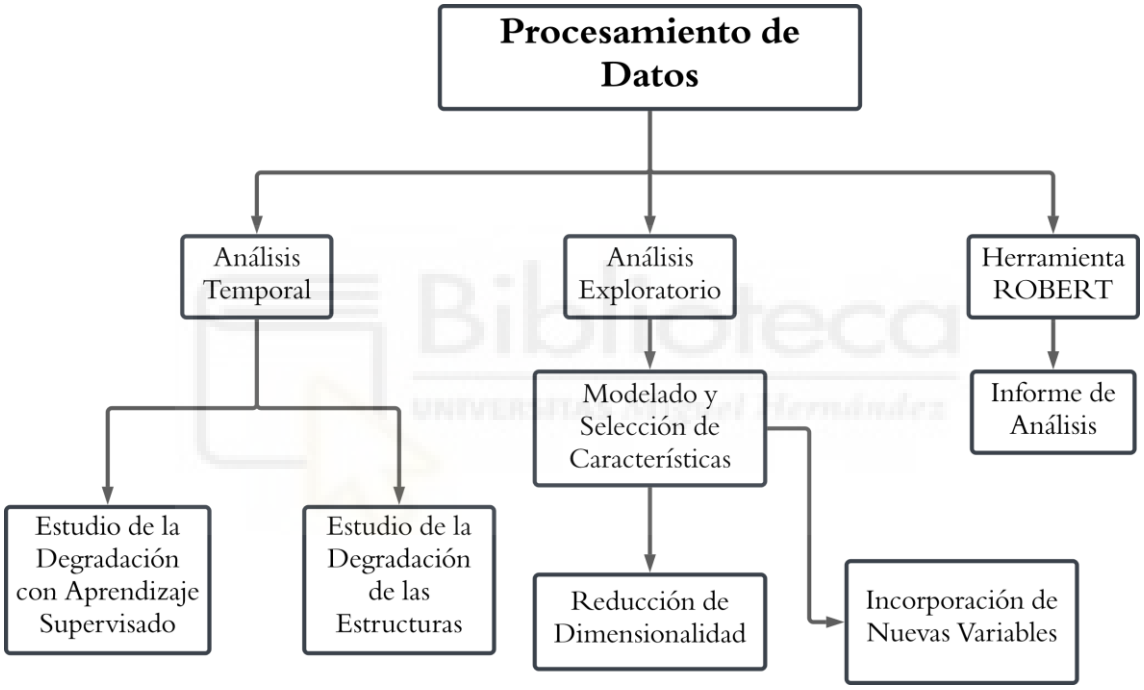



*Figura 14: Diagrama de los procesamientos realizados.*

En las siguientes subsecciones, se detalla el proceso de análisis exploratorio, selección de características, modelado predictivo, ajuste de hiperparámetros y reducción de dimensionalidad de los datos.

### 4.5.1 Análisis de Exploratorio de Datos

El análisis exploratorio de datos (EDA, "*Exploratory Data Analysis*") fue el primer paso para comprender la distribución y las relaciones entre las variables del conjunto de datos. Se utilizaron técnicas estadísticas y visualizaciones para identificar patrones, tendencias y posibles anomalías que pudieran afectar el rendimiento de los modelos predictivos. Entre las técnicas empleadas se encuentran:

1. D**istribución de la PCE**: s analizó la distribución de la eficiencia de conversión de potencia utilizando histogramas para visualizar la variabilidad de los valores obtenidos.
2. C**orrelación entre características**: se calculó la matriz de correlación entre las distintas variables del conjunto de datos y se representó en un mapa de color.
3. **Gráficos de dispersión**: se realizaron gráficos de dispersión para analizar la relación entre la PCE y las diferentes variables del conjunto de datos.

El análisis exploratorio fue crucial para detectar las relaciones subyacentes entre las variables y definir las estrategias a seguir en las etapas posteriores de selección de características y modelado.

### 4.5.2 Modelado y Selección de Características Relevantes

Una vez completado el análisis exploratorio de datos, se procedió con la selección de características, cuyo objetivo fue identificar las variables más relevantes para predecir la eficiencia de conversión de potencia (PCE). Dado que el conjunto de datos contenía múltiples variables, algunas de las cuales podían ser redundantes o no aportar valor significativo al modelo, se utilizaron técnicas de selección de características para simplificar el análisis y mejorar la precisión del modelo.

- **Importancia de características con Random Forest y Gradient Boost**: se utilizaron los modelos Random Forest y Gradient Boost para evaluar la importancia de cada variable en la predicción de la PCE. Ambos modelos poseen la capacidad de calcular la relevancia de las características a través de métricas internas, lo que permitió identificar las variables más influyentes.

- **Eliminación de variables poco relevantes**: tras el análisis de importancia de características, se decidió eliminar aquellas variables que presentaban poca relevancia o que podían estar introduciendo ruido en el modelo, como algunas variables meteorológicas. Esta eliminación permitió simplificar el modelo y centrarse en las características con mayor impacto en la PCE.

- **Validación de la selección de características**: para validar la selección de características realizada, se utilizaron Permutation Feature Importance (PFI) y Sequential Feature Selector (SFS). Estas técnicas ayudaron a confirmar la relevancia de las principales características identificadas y aseguraron que las variables seleccionadas eran efectivamente las más influyentes en el rendimiento del modelo.

La selección de características fue un paso fundamental para reducir la complejidad del modelo y garantizar que las variables utilizadas en el proceso de modelado fueran las que más contribuían a explicar la variabilidad en la PCE.

### 4.5.3 Ajuste de Hiperparámetro

Para mejorar el rendimiento de los modelos y evitar problemas de sobreajuste o infrajuste, se procedió al ajuste de hiperparámetros. El ajuste de hiperparámetros es un proceso crítico que permite optimizar los modelos para lograr un mejor equilibrio entre sesgo y varianza, garantizando así una mayor precisión y capacidad de generalización.

- **Búsqueda de hiperparámetros con Hyperopt**: se utilizó la librería **Hyperopt** para llevar a cabo una búsqueda eficiente de hiperparámetros, optimizando el rendimiento de los modelos Random Forest y Gradient Boost. En el caso de Random Forest, se ajustaron parámetros como el número de estimadores, que determina la cantidad de árboles en el bosque y afecta a la estabilidad y la precisión del modelo, la profundidad máxima de los árboles y el criterio de división. Para Gradient Boost, se ajustaron parámetros como la tasa de aprendizaje, el número de estimadores y la profundidad de los árboles.

### 4.5.4 Reducción de Dimensionalidad

Después del ajuste de hiperparámetros, se procedió a la reducción de dimensionalidad del conjunto de datos. Este proceso fue crucial para simplificar los modelos, reducir el ruido en los datos y mejorar la capacidad de generalización, eliminando características que, aunque importantes en ciertos contextos, no aportaban suficiente valor predictivo al modelo final. De esta forma, se intentó analizar el modelo:

- **Eliminación de variables redundantes**: tras la evaluación de la importancia de características, se decidió eliminar las variables meteorológicas que habían mostrado una importancia negativa o mínima en el análisis de PFI y SFS.

### 4.5.5 Modelado Predictivo

Una vez identificadas las características más relevantes, se procedió al modelado predictivo con el objetivo de predecir la eficiencia de conversión de potencia (PCE) de las células solares orgánicas. Se utilizaron dos modelos de *machine learning*: Random Forest y Gradient Boosting. Estos se entrenaron con los distintos conjuntos de datos creados, para observar el comportamiento de ambos algoritmos ante distintas combinaciones de variables.

- **Entrenamiento de los modelos**: los modelos fueron entrenados utilizando una división de 80-20 entre los conjuntos de entrenamiento y prueba. Para asegurar la fiabilidad de los resultados, se empleó validación cruzada, que permitió evaluar el rendimiento de los modelos de forma más robusta. Los hiperparámetros de ambos modelos fueron inicialmente configurados con sus valores predeterminados, salvo el número de estimadores, que se ajustó a 100.

- **Evaluación de rendimiento**: durante la fase de entrenamiento, se evaluó el rendimiento de los modelos utilizando métricas como el Error Absoluto Medio (MAE), el Error Cuadrático Medio (MSE) y el Coeficiente de Determinación ($R^2$). También se usó la función *cross_validation_score* de *scikit-learn* para medir el sobreajuste de los modelos.

El modelado predictivo permitió establecer una base sólida para predecir la PCE de las células solares orgánicas y proporcionó una primera aproximación a la identificación de las combinaciones de características que favorecen una mayor eficiencia.

### 4.5.6 Análisis Temporal de la Eficiencia

Finalmente, se realizó un análisis temporal de la eficiencia de conversión de potencia (PCE) para identificar posibles tendencias en el deterioro de las células solares orgánicas a lo largo del tiempo. Este análisis fue fundamental para comprender cómo las características de las células influían en su estabilidad y degradación.

- **Visualización de la evolución temporal**: para analizar la evolución de la PCE a lo largo del tiempo, se generaron gráficos de líneas que mostraban el rendimiento de cada célula solar durante el período de estudio. Estos gráficos permitieron identificar patrones de degradación, observando que algunas células presentaban una disminución rápida de la eficiencia mientras que otras mostraban una mayor estabilidad a lo largo del tiempo.
- **Cálculo de la tasa de degradación**: se calculó la tasa de degradación de cada célula solar mediante la diferencia porcentual entre mediciones sucesivas de PCE. Este análisis permitió identificar qué células eran más susceptibles a la degradación y en qué condiciones se presentaban los mayores descensos en la eficiencia.
- **Análisis comparativo de estructuras**: se compararon las tasas de degradación entre células con diferentes estructuras y composiciones.

El análisis temporal proporcionó información clave sobre cómo evolucionaba la eficiencia de las células solares orgánicas en el tiempo, permitiendo identificar combinaciones de materiales que favorecían una mayor estabilidad y eficiencia en el largo plazo.

### 4.5.7 Análisis Automatizado con ROBERT

La herramienta ROBERT automatiza la evaluación del conjunto de datos, realizando todos los procesos de optimización, entrenamiento, validación y evaluación automáticamente. Además de utilizar algoritmos avanzados como Random Forest (RF) y Gradient Boosting (GB), ROBERT también emplea Redes Neuronales (NN) y Máquinas de Vectores de Soporte (MVL), lo que amplía la gama de técnicas analíticas disponibles

para abordar los datos. Esto permite que ROBERT ajuste los modelos y a su vez compare diferentes enfoques para identificar el que ofrezca la mayor precisión y eficacia. Elegimos qué datos introducir, y ROBERT se encarga del resto, proporcionando un informe detallado de las actividades y métricas relevantes. Estos resultados serán analizados en la sección de resultados para determinar su impacto e implicaciones en nuestras conclusiones.

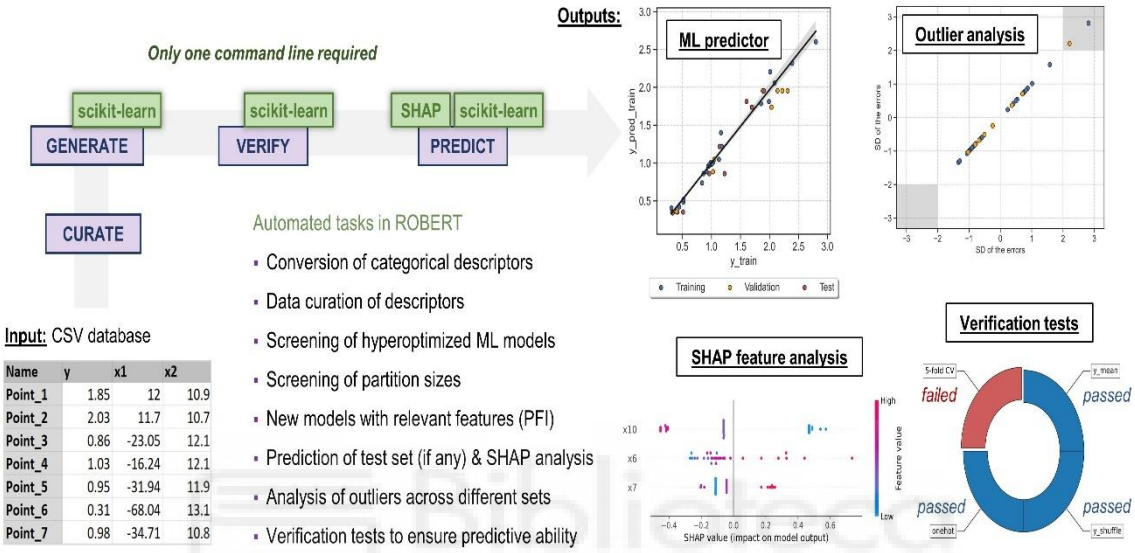



*Figura 15: Flujo de trabajo de trabajo de ROBERT . Imagen extraída de la documentación del software[30].*

# 5. RESULTADOS Y DISCUSIÓN

En esta sección, se presentan y analizan los resultados obtenidos en el estudio de la eficiencia de conversión de potencia (PCE) de las células solares orgánicas. Para ello, se ha dividido esta sección en dos apartados: primero, se describe en detalle el conjunto de datos utilizado, destacando sus características clave; posteriormente, se discuten los hallazgos principales derivados de los modelos predictivos, explorando la relevancia de las variables analizadas y la precisión de las predicciones realizadas.

## 5.1 Descripción del Conjunto de Datos Utilizado

En este apartado, se ofrece un análisis detallado de las características específicas del conjunto de datos empleado en esta investigación. Aunque en secciones anteriores ya se ha explicado el proceso de adquisición de los datos, aquí se profundizará en las variables que componen este conjunto y su relevancia para los objetivos del estudio. Este análisis es fundamental para comprender las relaciones subyacentes entre las diferentes variables y su impacto en la eficiencia de las células solares orgánicas.

A continuación, se muestra una lista con las variables que han sido estudiadas:

- Fecha datos: año, mes y día de la toma de la muestra.
- Célula: identificación de la célula a la que pertenece cada muestra en los diferentes experimentos.
- PCE (%): eficiencia de conversión de potencia (Power Conversion Efficiency), que es la variable objetivo.
- rGO: presencia de óxido de grafeno reducido (1 si está presente, 0 si no).
- Cantidad DS HTL: cantidad de material utilizado en la capa transportadora de huecos.
- Cantidad PEDOT:PSS: cantidad de PEDOT:PSS.
- Ratio P3HT:PCBM: proporción de P3HT a PCBM.
- Ta PEDOT:PSS: temperatura de la disolución de PEDOT:PSS.
- Ta rGO: temperatura de la disolución del grafeno (ºC).
- Disolución rGo: tipo de disolución utilizada de rGO.
- Temperatura (*Temperature*): temperatura medida en grados Celsius (°C).

- Presión (*Pressure*): presión atmosférica medida en hectopascales (hPa).
- Humedad (*Humidity*): humedad relativa medida en porcentaje (%).
- Punto de rocío (*Dew Point*): temperatura a la que el vapor de agua presente en el aire se condensa en rocío (ºC).

Variables como el tiempo y la célula se utilizan principalmente como identificadores, mientras que otras, como rGO, permiten diferenciar entre las células que contienen o no grafeno. Esto es útil para determinar si la adición de este material tiene un efecto positivo o negativo. Las variables relacionadas con el grafeno nos ayudarán a discernir cuáles son las condiciones óptimas en los experimentos que incluyen este material.

Por otro lado, variables como cantidad de DS HTL, cantidad de PEDOT:PSS y la ratio P3HT:PCBM son comunes a todas las muestras, ya que forman parte de la estructura de las células solares orgánicas. Estas variables son especialmente importantes, ya que afectan a todas las células y comprender qué proporciones resultan en un aumento de la PCE podría afectar al resultado de todos los experimentos. Finalmente, las variables meteorológicas se han añadido para observar cómo estas condiciones podrían repercutir a las células solares. Es posible que influyan en la PCE o que aceleren el deterioro de las células.

## 5.2 Análisis de los Resultados Obtenidos

En este apartado, se analizan los resultados obtenidos a lo largo de la investigación mediante diversas técnicas de análisis de datos, que incluyen tanto enfoques exploratorios como modelado predictivo avanzado. Se presentarán los hallazgos más significativos, discutiendo la relevancia de las variables estudiadas y la eficacia de los modelos predictivos aplicados. Este análisis permitirá identificar patrones y tendencias clave, ofreciendo una comprensión más profunda del comportamiento de las células solares bajo distintas condiciones experimentales.

### 5.2.1 Análisis Exploratorio de Datos

El análisis exploratorio de datos constituye un paso crucial en la comprensión de los patrones subyacentes en el conjunto de datos utilizado en esta investigación. Este análisis permite visualizar la distribución de las variables clave, como la eficiencia de conversión de potencia (PCE), además de identificar posibles relaciones entre las distintas características y su influencia en el rendimiento de las células solares orgánicas.

#### 5.2.1.1 Distribución de la PCE

Para analizar de la distribución de la PCE se ha utilizado un histograma (Figura 16), que nos ayuda a identificar los valores máximos y mínimos y proporciona una visión general de la variabilidad en los datos.

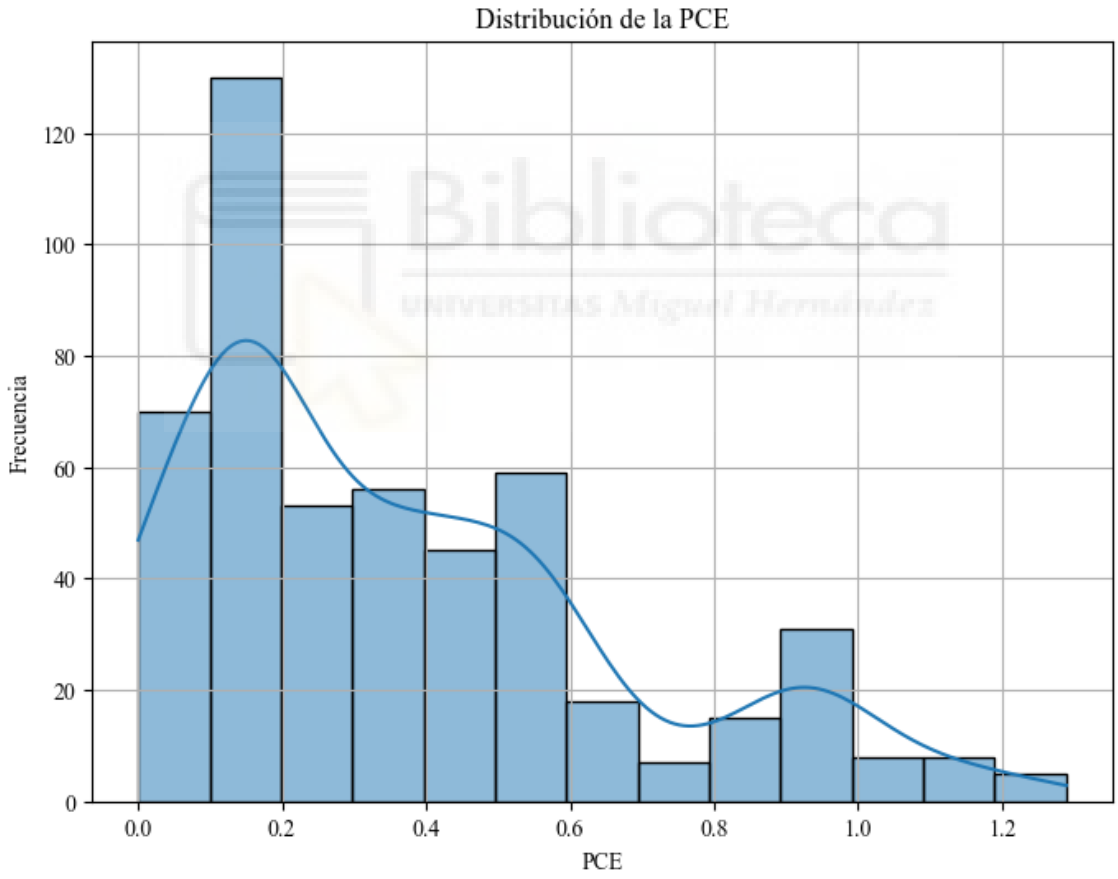



*Figura 16: Histograma de la distribución de los valores de PCE en el conjunto de datos.*

A partir de esta gráfica podemos observar cómo la PCE presenta una distribución amplia, con valores que oscilan entre aproximadamente 0% y 1.3%. La mayoría de las células tienen una PCE que cae por debajo de 0.6%, con un pequeño número alcanzando valores más altos. Esto sugiere que, en general, las OSC en el conjunto de datos tienen eficiencias

de conversión relativamente bajas, aunque existen casos excepcionales con mejores rendimientos.

### 5.2.1.2 Correlación entre Características

Se ha estudiado la correlación entre variables mediante una matriz de correlación representada en un mapa de color (Figura 17). Esto nos ayuda a obtener información sobre los componentes del conjunto de datos y permite identificar cuáles influyen más en la PCE.

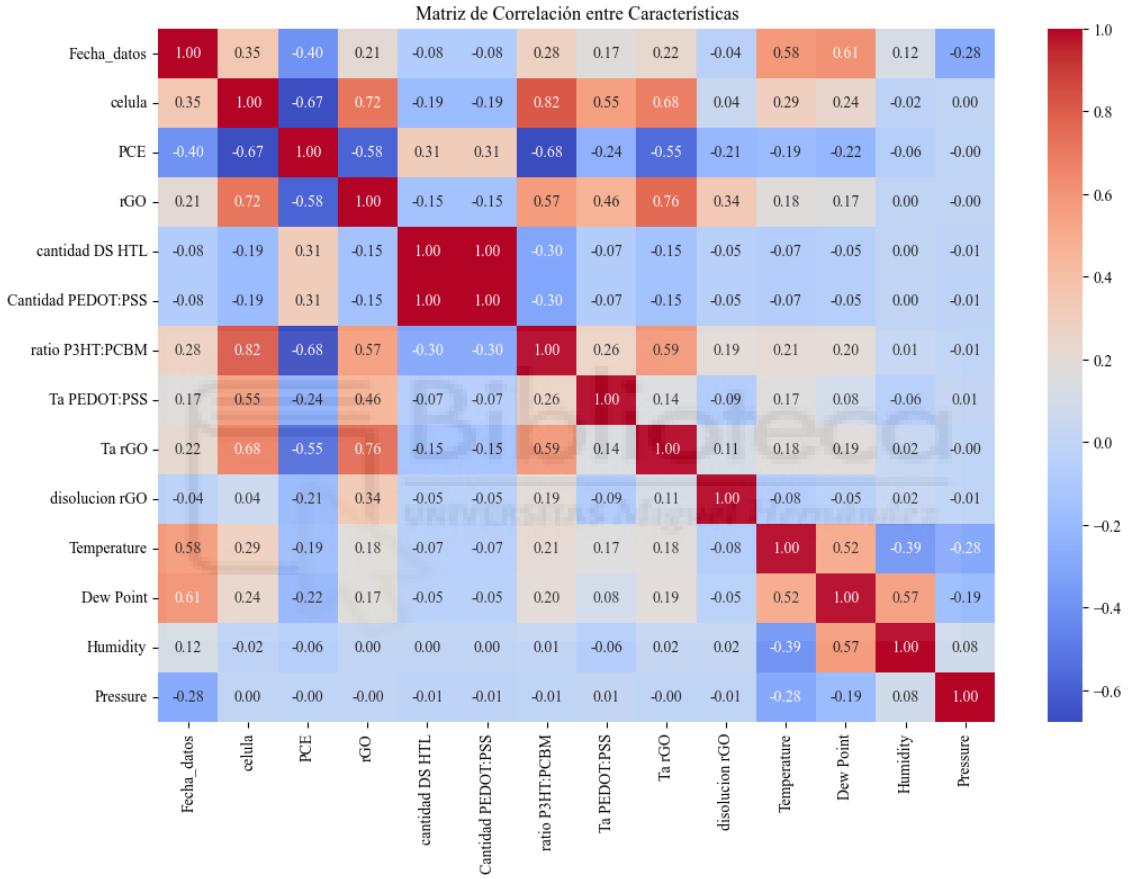



*Figura 17: Mapa de color de la correlación entre las variables del conjunto de datos.*

La PCE muestra una correlación negativa fuerte con el ratio de PEDOT:PSS, lo que sugiere que valores altos de esta ratio podrían estar asociados a una disminución en la eficiencia de las células. Así mismo, se observa cómo la presencia de grafeno en las células puede tener un impacto negativo en la PCE

El gráfico evidencia una fuerte correlación entre la cantidad de DS HTL y PEDOT:PSS, lo que indica una relación lineal entre ambas. Este valor de correlación tan alto sugiere que estas dos columnas representan esencialmente la misma información en el conjunto

de datos. Al revisar el contexto del experimento, se confirma que DS HTL hace referencia a la capa de transporte de huecos, que en este caso corresponde específicamente al material PEDOT:PSS.

Por lo tanto, la duplicidad en las columnas no aporta información adicional al análisis y puede generar redundancia en el modelo. Para evitar este problema de multicolinealidad y simplificar el conjunto de datos, se va a eliminar la columna cantidad DS HTL, ya que su información está completamente contenida en la columna cantidad PEDOT:PSS. Este ajuste ayudará a mejorar la eficiencia de los análisis posteriores y la interpretación de los resultados.

### 5.2.1.3 Gráficos de Dispersión

Se ha analizado la relación entre la PCE y las distintas variables del conjunto de datos mediante gráficos de dispersión (Figura 18) para detectar posibles tendencias.

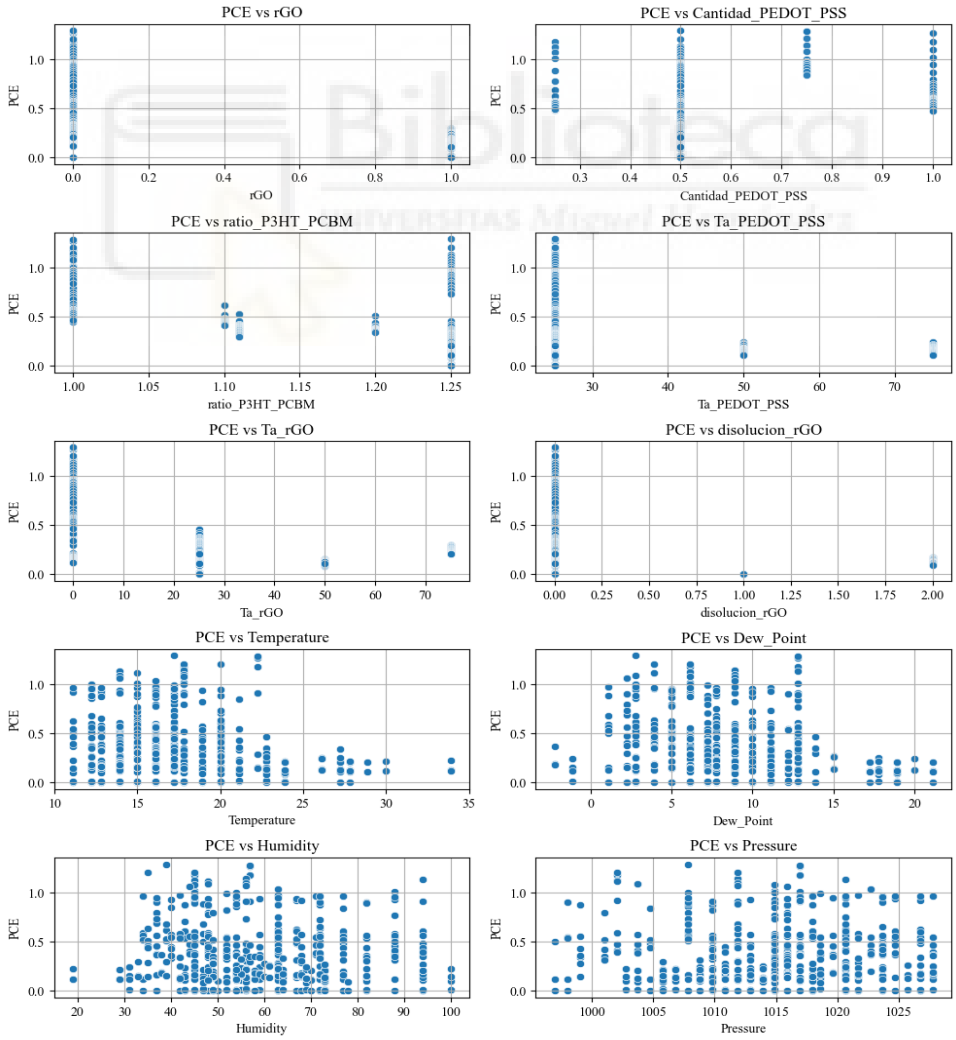



*Figura 18: Gráfico de dispersión de la PCE frente a las distintas variables.*

A partir de los distintos gráficos, se observa que la PCE no muestra una relación lineal con ninguna característica. Sin embargo, se nota que las muestras con rGO tienden a tener una PCE más baja, lo que sugiere que el uso de este componente podría estar perjudicando la eficiencia. Para el resto de las características, no se detectan patrones evidentes.

**Discusión del análisis Exploratorio**

La complejidad multivariable, junto con la falta de correlación fuerte y directa entre una sola característica y la PCE, sugiere que la eficiencia está determinada por una combinación de factores. Esto implica que el enfoque para mejorar la PCE debería considerar ajustes simultáneos en múltiples características.

Aunque el análisis no revela cambios claros en una única característica que puedan mejorar significativamente la PCE, podría ser útil investigar combinaciones específicas de valores para optimizar el rendimiento.

Finalmente, la observación de que la presencia de rGO podría estar relacionada con una disminución en la PCE, esto sugiere que eliminar este componente podría ser un camino que explorar para mejorar la eficiencia.

Con el objetivo de explorar diferentes perspectivas para caracterizar y comprender el rendimiento de las células solares orgánicas, se plantean dos enfoques complementarios en los siguientes apartados. En el primero de ellos, el apartado 5.2.2, se propone un análisis multivariable utilizando modelos de *machine learning* para identificar patrones y relaciones significativas entre las características estructurales de las células solares y su eficiencia de conversión de potencia (PCE). Este enfoque busca aprovechar la capacidad de estos modelos para analizar grandes volúmenes de datos y generar predicciones precisas, lo que permite optimizar las configuraciones de las células solares.

Por otro lado, el apartado 5.2.3 adopta una perspectiva temporal para abordar el problema, centrándose en el análisis de la tasa de degradación de las células solares a lo largo del tiempo. Este enfoque permite estudiar cómo las diferentes estructuras afectan la estabilidad de la PCE y qué características contribuyen a minimizar la degradación, manteniendo un buen rendimiento. Para ello, se introduce una variable identificadora para cada célula, lo que facilita el análisis de las tendencias individuales y la comparación entre diferentes configuraciones.

Ambos enfoques, aunque distintos en su naturaleza, son complementarios y ofrecen una visión integral del problema. Mientras que el análisis multivariable se centra en las características estructurales y su impacto inmediato en el rendimiento, el análisis temporal aporta una perspectiva dinámica que considera la evolución de la eficiencia a lo largo del tiempo, proporcionando información valiosa sobre la estabilidad y la durabilidad de las células solares orgánicas.

## 5.2.2 Selección de Características y Modelado Predictivo

Los resultados obtenidos en el análisis exploratorio sugieren que los datos presentan una complejidad considerable, lo que requiere un análisis más profundo. Para abordar esta complejidad, se han empleado técnicas de *machine learning* con el objetivo de identificar patrones complejos y formular recomendaciones precisas para mejorar la eficiencia de conversión de potencia (PCE) en las células solares orgánicas.

En esta sección, se describen los métodos utilizados para seleccionar las características más relevantes y los modelos predictivos entrenados para analizar la influencia de estas características en la PCE.

### 5.2.2.1 Análisis de la Importancia de Características

En esta primera etapa del análisis, se emplearon los modelos de Random Forest y Gradient Boost para identificar las características más influyentes en la eficiencia de conversión de potencia (PCE) de las células solares orgánicas. Estos algoritmos fueron seleccionados debido a su capacidad para manejar grandes cantidades de variables y capturar relaciones no lineales complejas entre las características y la PCE, lo que los hace ideales para este tipo de problemas multivariados.

El uso del atributo 'feature_importance_' en ambos modelos nos permite identificar la relevancia de cada característica en el cálculo de las predicciones. El resultado de este análisis se presenta en dos gráficos de barras (Figura 19), que muestran la importancia relativa de cada variable según ambos algoritmos. Como parte del proceso, los modelos fueron entrenados usando una división de 80-20 entre los conjuntos de entrenamiento y prueba, con los hiperparámetros configurados en sus valores predeterminados, salvo por el número de estimadores, que se ajustó a 100.

Los valores obtenidos durante el entrenamiento de los modelos para observar su rendimiento son los mostrados a continuación.

| Modelo | Random Forest | Gradient Boost |
|---|---|---|
| MAE | 0.0895 | 0.0959 |
| MSE | 0.0276 | 0.0213 |
| $R^2$ Score | 0.6661 | 0.7423 |

*Tabla 3: Métricas de error para el entrenamiento de los modelos.*

En términos de rendimiento, el modelo de Gradient Boost demostró ser más eficaz, obteniendo un error absoluto medio (MAE) de 0.0959 y un error cuadrático medio (MSE) de 0.0213, acompañado de un coeficiente de determinación ($R^2$) de 0.7425. Estos resultados indican que Gradient Boost captura de manera más precisa las relaciones no lineales entre las características y la PCE, en comparación con el modelo Random Forest, que obtuvo un $R^2$ de 0.6628. Este mejor desempeño sugiere que Gradient Boost es más adecuado para identificar y optimizar las variables clave que influyen en la eficiencia de las células solares.

Al analizarlos con validación cruzada obtenemos que el ajuste del modelo Random Forest obtiene un $R^2 = 0.7122$ +/- 0.0764, lo cual supone una mejoría que nos dice que es capaz de generalizar bien a datos no vistos mientras que en el caso de Gradient Boost $R^2 = 0.6948 \ +/- \ 0.0523$.

En cuanto a la importancia de las características, el análisis mostró que la variable más influyente en ambos modelos era la temperatura de la disolución de óxido de grafeno (Ta rGO). Sin embargo, análisis exploratorios previos habían revelado que las células solares que contenían rGO presentaban un rendimiento significativamente inferior. Para evitar que esta variable dominara el modelo y distorsionara la interpretación de otras características, se decidió eliminar tanto rGO como las células asociadas a él. Esta decisión permitió obtener una visión más equilibrada del conjunto de datos y mejorar la interpretación del impacto de las demás variables en la eficiencia de conversión de potencia.

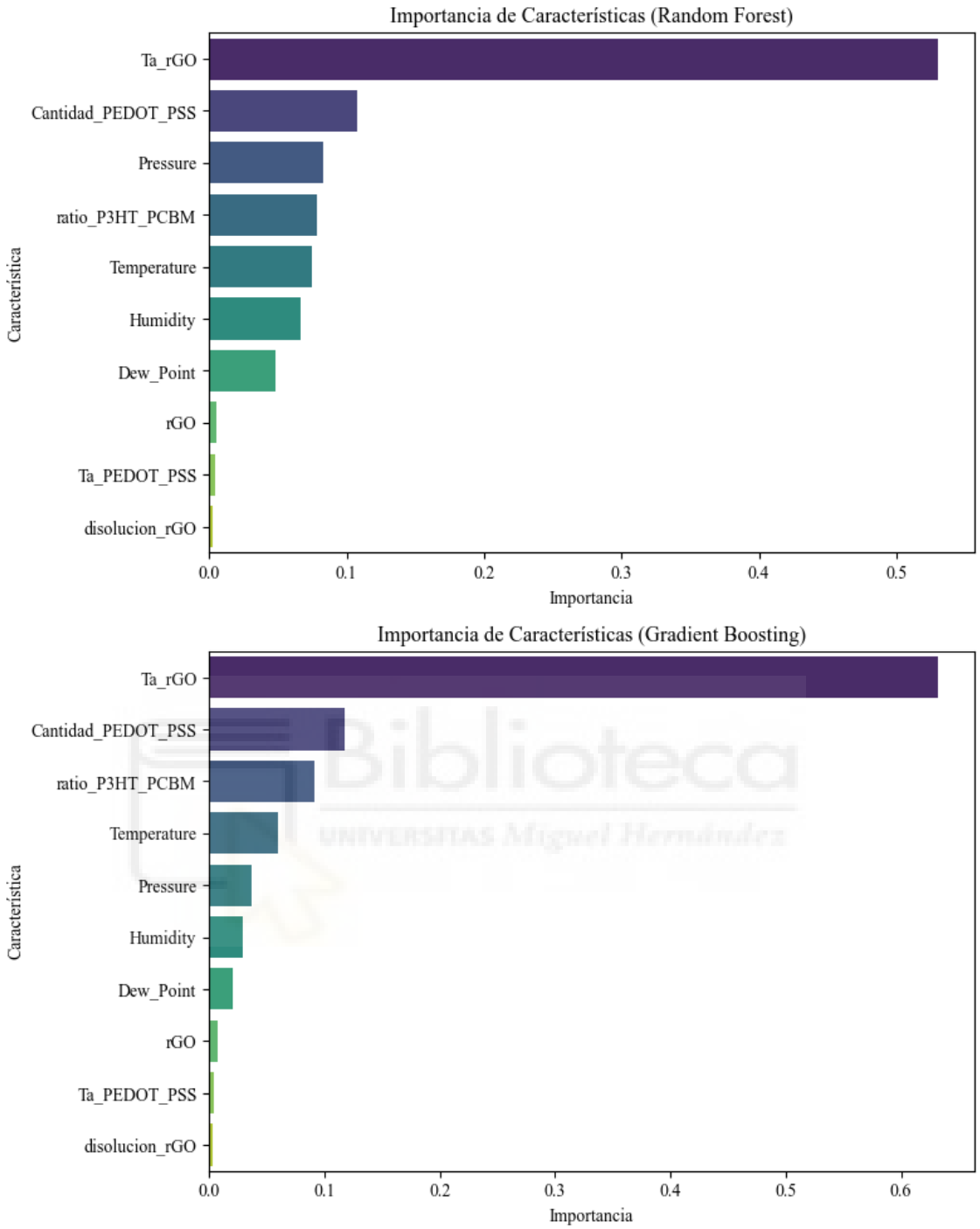



*Figura 19: Gráfico de barras, importancia de las distintas variables del conjunto de datos obtenida por Random Forest y Gradient Boost.*

Después de la eliminación de las células con rGO, se entrenaron nuevamente los modelos Random Forest y Gradient Boost, para evaluar cómo se comportan las variables en ausencia de los factores dominantes asociados al grafeno. Como era esperado, se observó una disminución en el desempeño de ambos modelos en cuanto a las métricas de error.

El modelo Random Forest obtuvo un $R^2$ de 0.1786 y el Gradient Boost un $R^2$ de 0.2921. Este resultado indica que la eliminación de las variables asociadas al rGO aumentó la dificultad para explicar la variabilidad en el conjunto de datos, dado que estas variables aportaban información significativa, aunque su impacto en el rendimiento de las células era negativo. Sin embargo, este análisis permite explorar patrones más relevantes en las células con mejor desempeño.

| Modelo | Random Forest | Gradient Boost |
|---|---|---|
| MAE | 0.1772 | 0.1650 |
| MSE | 0.0732 | 0.0631 |
| $R^2$ Score | 0.1786 | 0.2921 |

*Tabla 4: Métricas de error para el entrenamiento de los modelos una vez eliminado rGO del conjunto de datos.*

En cuanto a la importancia de las variables, esta cambió significativamente (Figura 20). Como se observa el ratio P3HT:PCBM se posiciona como la variable más relevante en ambos modelos. Además las variables meteorológicas también han aumentado su presencia. Este hallazgo es consistente con el objetivo de identificar factores clave para optimizar las células más eficientes.

La hipótesis radica en que, al eliminar la variable relacionada con el grafeno, los modelos pierden una fuente de variación que antes podía facilitar su capacidad de diferenciación. Además, el gran número de mediciones de eficiencia (PCE) provenientes de las células con grafeno, que presentan comportamientos muy similares, podría haber "inflado" la estadística de error, dando la impresión de un mejor desempeño de lo que realmente era. Para validar esta sospecha, más adelante se incorpora una nueva variable para cada estructura en los experimentos, lo que permitirá evaluar con mayor precisión el papel del grafeno y aislar adecuadamente su efecto en el modelo.

A pesar de la disminución en el desempeño de los modelos, este análisis refuerza la hipótesis de que las variables asociadas al grafeno, aunque dominantes, pueden enmascarar factores más relevantes para las células de mayor eficiencia. Este enfoque permite centrar los esfuerzos en optimizar las combinaciones estructurales y ambientales más prometedoras, lo cual será clave en futuras fases del proyecto. En los próximos pasos, se buscará refinar los modelos y validar estos hallazgos.

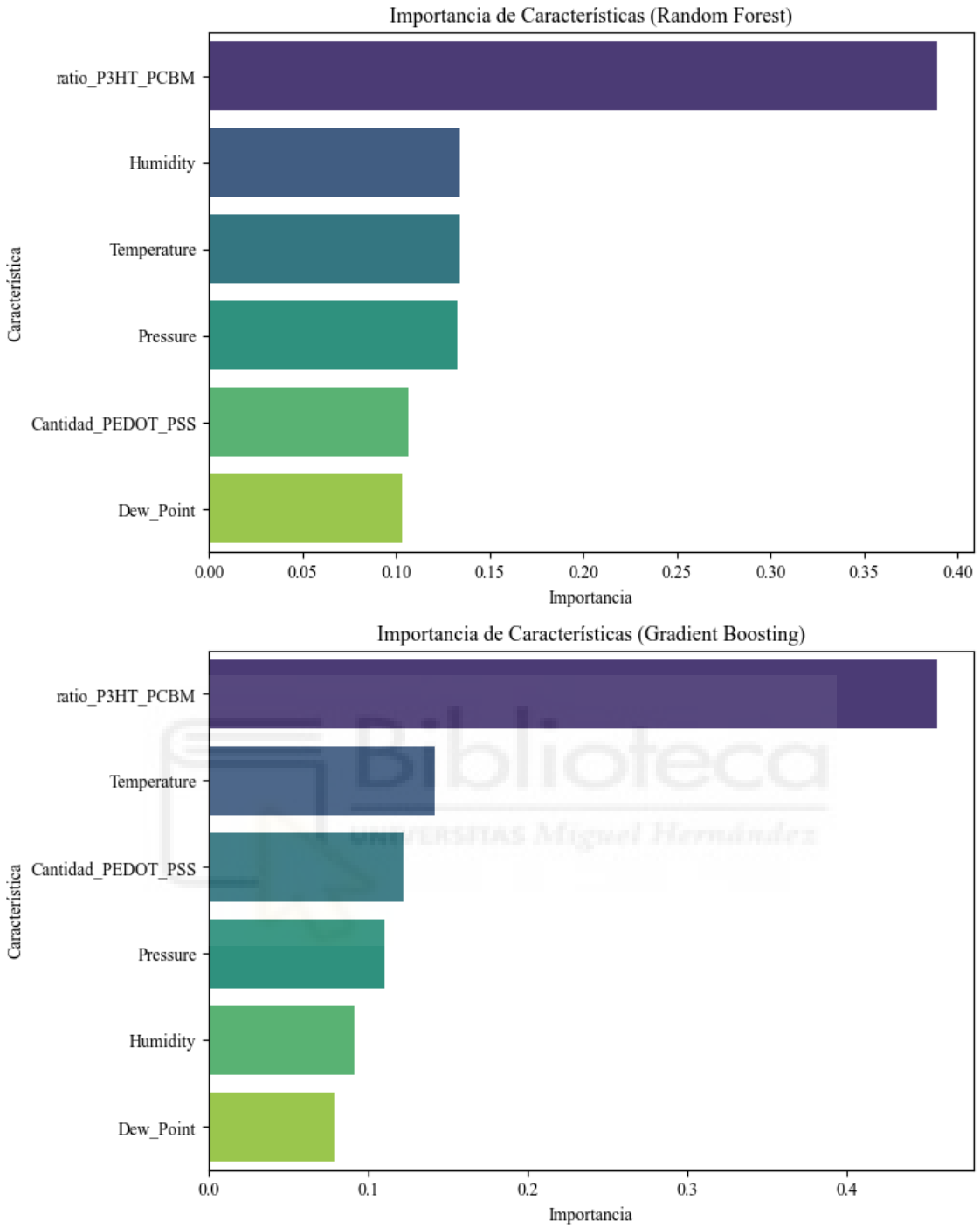



*Figura 20: Gráfico de barras, importancia de las distintas variables del conjunto de datos obtenida por Random Forest y Gradient Boost, una vez eliminado rGO del conjunto de datos.*

### 5.2.2.2 Ajuste de Hiperparámetros

Después de la etapa inicial de selección de características, se procedió al ajuste de los hiperparámetros de los modelos Random Forest y Gradient Boost con el objetivo de

mejorar su rendimiento predictivo. Este ajuste es esencial para optimizar el balance entre sesgo y varianza, y evitar problemas de sobreajuste o infraajuste en los modelos.

Para llevar a cabo el ajuste, se utilizó la librería **hyperopt**, que permite realizar una búsqueda eficiente de hiperparámetros, optimizando el rendimiento del modelo. En el caso de Random Forest, se ajustaron parámetros como el número de estimadores, la profundidad máxima de los árboles y el criterio de división. Al igual que en el caso anterior en el modelo Gradient Boost, se ajustaron parámetros como la tasa de aprendizaje, el número de estimadores y la profundidad de los árboles.

Los resultados obtenidos después del ajuste de hiperparámetros se presentan en la siguiente tabla (Tabla 5). A pesar de que los valores de $R^2$ mejoraron ligeramente, se mantienen bajos (0.3316 para Random Forest y 0.3327 para Gradient Boost), lo cual indica que los modelos aún tienen limitaciones en capturar las relaciones complejas en los datos. Esto podría deberse a una alta variabilidad en las características de las células solares o a mediciones inconsistentes entre las muestras. En este sentido, planteamos la necesidad de un análisis más exhaustivo en las siguientes etapas para abordar estas limitaciones y mejorar la capacidad de predicción de los modelos.

| Modelo | Random Forest | Gradient Boost |
|---|---|---|
| MAE | 0.1702 | 0.1669 |
| MSE | 0.0596 | 0.0595 |
| $R^2$ Score | 0.3316 | 0.3327 |

*Tabla 5: Métricas de error para el entrenamiento de los modelos, tras el ajuste de hiperparámetros, una vez eliminado rGO del conjunto de datos.*

Una vez ajustados los modelos, se volvió a analizar la importancia de las características seleccionadas (Figura 21). En el modelo Random Forest, se observó un cambio en la distribución de los pesos asignados a las variables, disminuyendo la relevancia de las variables meteorológicas. Esto sugiere que el ajuste de hiperparámetros permitió un enfoque más preciso en las variables más influyentes, mejorando la interpretación del modelo. En el modelo Gradient Boost, vemos pesos más distribuidos siendo las variables más relevantes ratio P3HT:PCBM y Cantidad de PEDOT:PSS.

En resumen, el ajuste de hiperparámetros mostró mejoras leves en los modelos, pero los bajos valores de $R^2$ indican la necesidad de una revisión más profunda de los datos. Esto podría incluir un análisis detallado de la variabilidad entre células solares, así como de

las condiciones de medición. Aunque estos ajustes proporcionan una base para futuras optimizaciones, los resultados destacan la importancia de abordar las limitaciones inherentes en los datos para mejorar la precisión de las predicciones de eficiencia de las células.

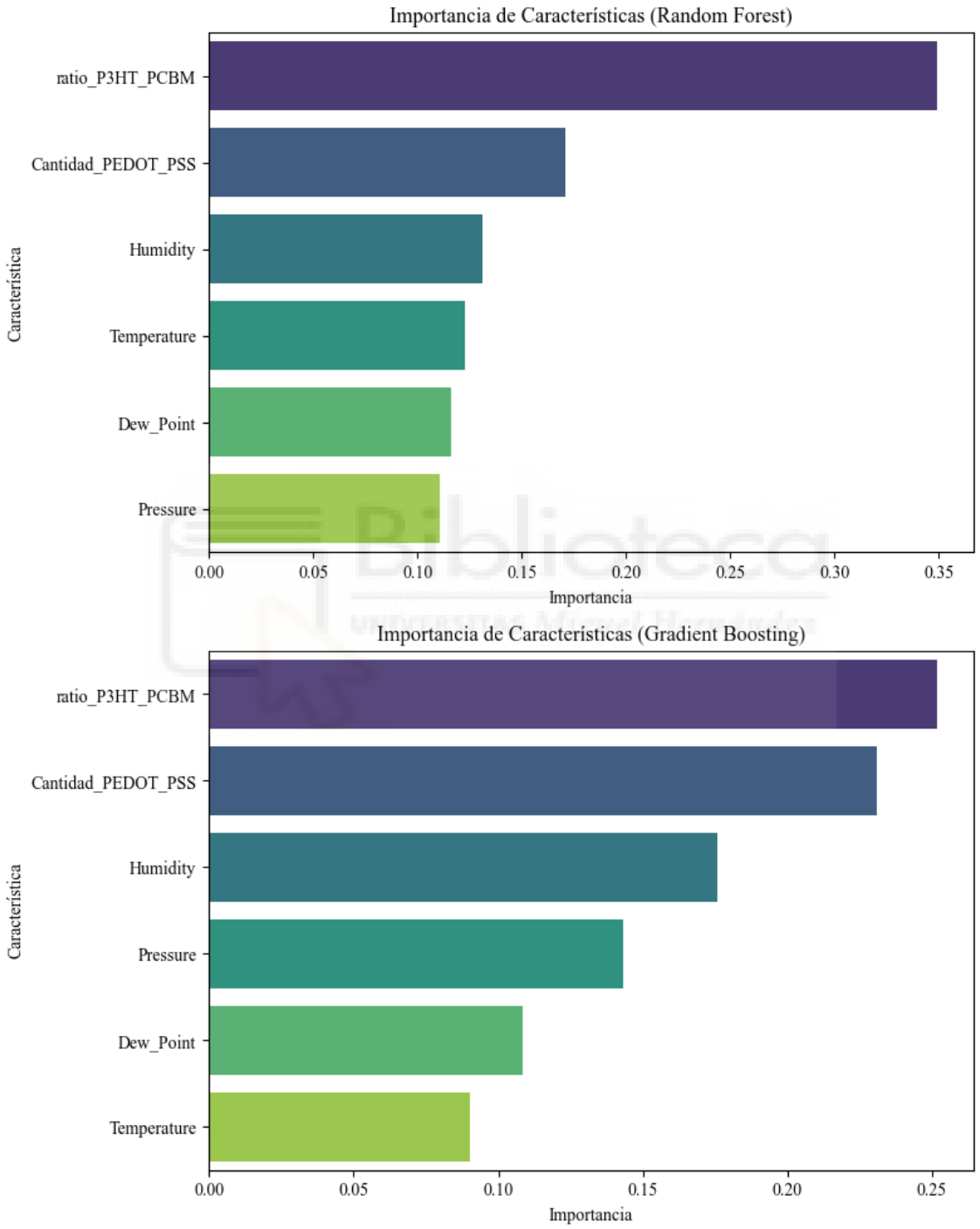



*Figura 21: Gráfico de barras, importancia de las distintas variables del conjunto de datos obtenida por Random Forest y Gradient Boost, una vez eliminado rGO del conjunto de datos y ajustados los hiperparámetros.*

### 5.2.2.3 Selección de Características Relevantes

Para complementar el análisis de selección de características y obtener una mejor comprensión de las variables clave, se aplicaron dos técnicas adicionales de evaluación de la importancia de las características: PFI y SFS Estas técnicas ofrecen una visión más robusta de cómo cada característica influye en el rendimiento de los modelos, ayudando a validar y mejorar los resultados obtenidos en la etapa anterior.

Para hacer esto hemos utilizado la función "*sklearn.inspection.permutation_importance*" y la clase "*sklearn.feature_selection.SequentialFeatureSelector*" las cuales aplican estos métodos utilizando nuestro modelo ya entrenado, de esta forma hemos obtenido dos listas con los valores más relevantes en cada caso.

En el gráfico de cajas (Figura 22), se ilustra la importancia de las características en ambos modelos, Random Forest y Gradient Boost. Los resultados del análisis muestran que la variable ratio P3HT:PCBM es consistentemente la característica más influyente en ambos modelos, indicando una relación significativa con la eficiencia de conversión de energía (PCE). Otras variables, como Cantidad PEDOT:PSS, temperatura y presión, también tienen un impacto considerable, aunque menor. Por otro lado, las variables meteorológicas como Humedad y Punto de Rocío mostraron valores negativos en el modelo Random Forest, lo que sugiere que podrían estar añadiendo ruido al modelo y no contribuyendo de manera positiva a las predicciones.

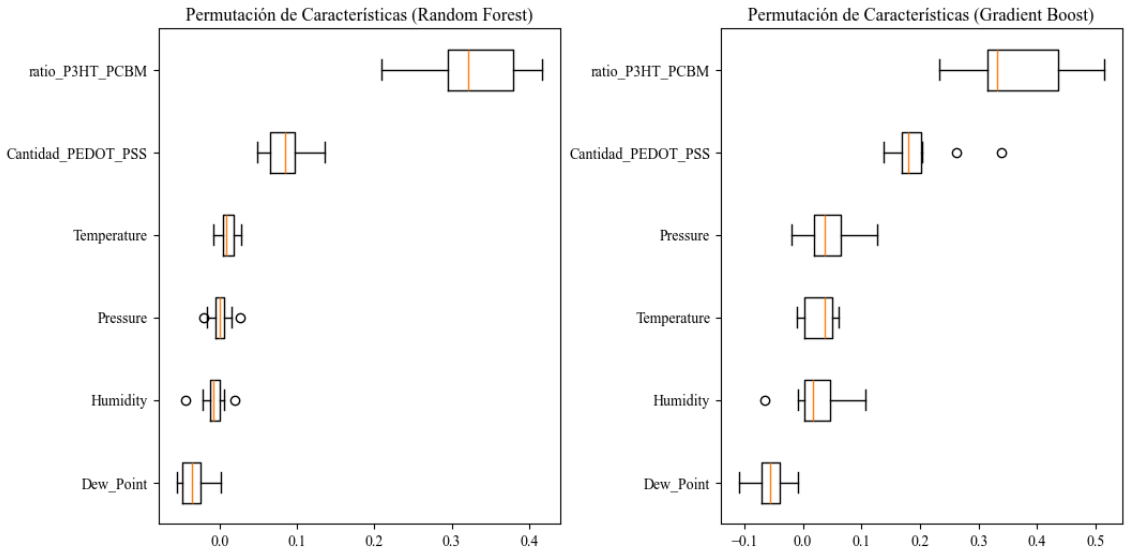



*Figura 22: Gráfico de cajas. Importancia de características por permutación en los modelos Random Forest (izquierda) y Gradient Boosting (derecha). Se observa el impacto en la precisión al permutar cada característica.*

Por otro lado, la técnica SFS fue aplicada para seleccionar iterativamente las características más relevantes. Esta técnica es útil para validar el conjunto óptimo de características y comparar los resultados con los obtenidos mediante PFI.

Los resultados obtenidos mediante SFS (Tabla 6), muestran una consistencia notable con los resultados del PFI. Tanto en el modelo Random Forest como en el modelo Gradient Boost, las características Cantidad PEDOT:PSS y ratio P3HT:PCBM aparecen como las más influyentes. Esto refuerza la confianza en la capacidad de estos modelos para identificar correctamente las variables clave que afectan la eficiencia de las células solares orgánicas.

| Modelo | Características Seleccionadas |
|---|---|
| Random Forest | Cantidad PEDOT:PSS, ratio P3HT:PCBM, Humidity y Pressure |
| Gradient Boost | Cantidad PEDOT:PSS, ratio P3HT:PCBM, Humidity y Pressure |

*Tabla 6: Resultados obtenidos de la SFS para ambos modelos de machine learning.*

Una observación importante es que en el modelo Random Forest, la característica Humidity fue seleccionada como relevante en el SFS, a pesar de mostrar una importancia negativa en el análisis de PFI. Esto podría indicar una compleja relación no lineal o una interacción con otras variables que puede hacer que su eliminación afecte de manera significativa el rendimiento del modelo. Esta discrepancia sugiere la necesidad de un análisis más profundo de las interacciones entre las características para identificar posibles fuentes de ruido o redundancia en los datos.

En resumen, la combinación de PFI y SFS permitió confirmar la relevancia de las principales características que afectan la PCE de las células solares orgánicas. Aunque ambos modelos coincidieron en las características clave, la variación en la importancia de las variables meteorológicas indica que estas podrían estar generando ruido en las predicciones. Estos resultados guiarán los próximos pasos en la reducción de dimensionalidad y en la mejora del modelo, con un enfoque en la optimización de las características que realmente aportan valor predictivo al modelo.

#### 5.2.2.4 Reducción de dimensionalidad

Tras los análisis de importancia de características y selección mediante los modelos Random Forest y Gradient Boost, se procedió a realizar una reducción de dimensionalidad del conjunto de datos. Esta etapa es crucial para simplificar el modelo, reducir el ruido y mejorar la capacidad de generalización, eliminando características que, aunque importantes en ciertos contextos, pueden no aportar valor predictivo suficiente en el modelo final.

Para llevar a cabo esta estrategia, se eliminaron las variables meteorológicas, ya que el Punto de Rocío y la Humedad en el análisis de *Permutation Feature Importance* (PFI) y su impacto en la precisión del modelo era mínimo o incluso perjudicial. Nos centramos, entonces, en las características más influyentes: Cantidad de PEDOT:PSS y ratio de P3HT:PCBM, que mostraron un peso considerable en ambos modelos de análisis. Estas variables fueron seleccionadas debido a su relevancia continua en los modelos, así como a su importancia práctica en la optimización de la eficiencia de conversión de potencia (PCE).

Una vez realizada la reducción, se entrenaron nuevamente los modelos Random Forest y Gradient Boost, y se evaluaron utilizando validación cruzada para comprobar la capacidad de generalización de ambos. Los resultados obtenidos tras la reducción de dimensionalidad se presentan en la Tabla 7. Ambos modelos muestran una leve mejora con respecto a experimentos anteriores, esto nos hace ver que seguimos teniendo dificultades para explicar la variabilidad del conjunto de datos.

| Modelo | Validación cruzada ($R^2$) | MAE | MSE | $R^2$ Score |
|---|---|---|---|---|
| Random Forest | 0.3575 +/- 0.0881 | 0.1634 | 0.0538 | 0.3960 |
| Gradient Boost | 0.2863 +/- 0.0986 | 0.1785 | 0.0573 | 0.3571 |

*Tabla 7: Métricas de error para el entrenamiento de los modelos, tras evaluar el modelo después de la reducción de dimensionalidad.*

En términos de error, el modelo Gradient Boost presentó peores resultados con un **MAE** de 0.1785 y un **MSE** de 0.0573, lo que indica que este modelo es menos preciso y tiene más error en comparación con Random Forest, cuyo MAE fue de 0.1634 y su MSE de 0.0538. Estos resultados destacan que, aunque ambos modelos mejoran muy levemente, esta estrategia no parece ser muy eficaz a la hora de mejorar los modelos.

Sin embargo, se observó que, a pesar de la leve mejora en la precisión, la variabilidad en el conjunto de datos sigue siendo un reto para ambos modelos. Esto es particularmente evidente en la distribución de las características seleccionadas. Como se puede observar en la Figura 23 la Cantidad de PEDOT:PSS y la Cantidad de DS HTL presentan una distribución muy desigual, con la mayoría de los valores concentrados en un único nivel, lo que dificulta la capacidad del modelo para hacer predicciones precisas en diferentes escenarios. Por otro lado, la ratio de P3HT:PCBM presenta una distribución algo más equilibrada, lo que explica en parte su importancia continua en las predicciones de ambos modelos.

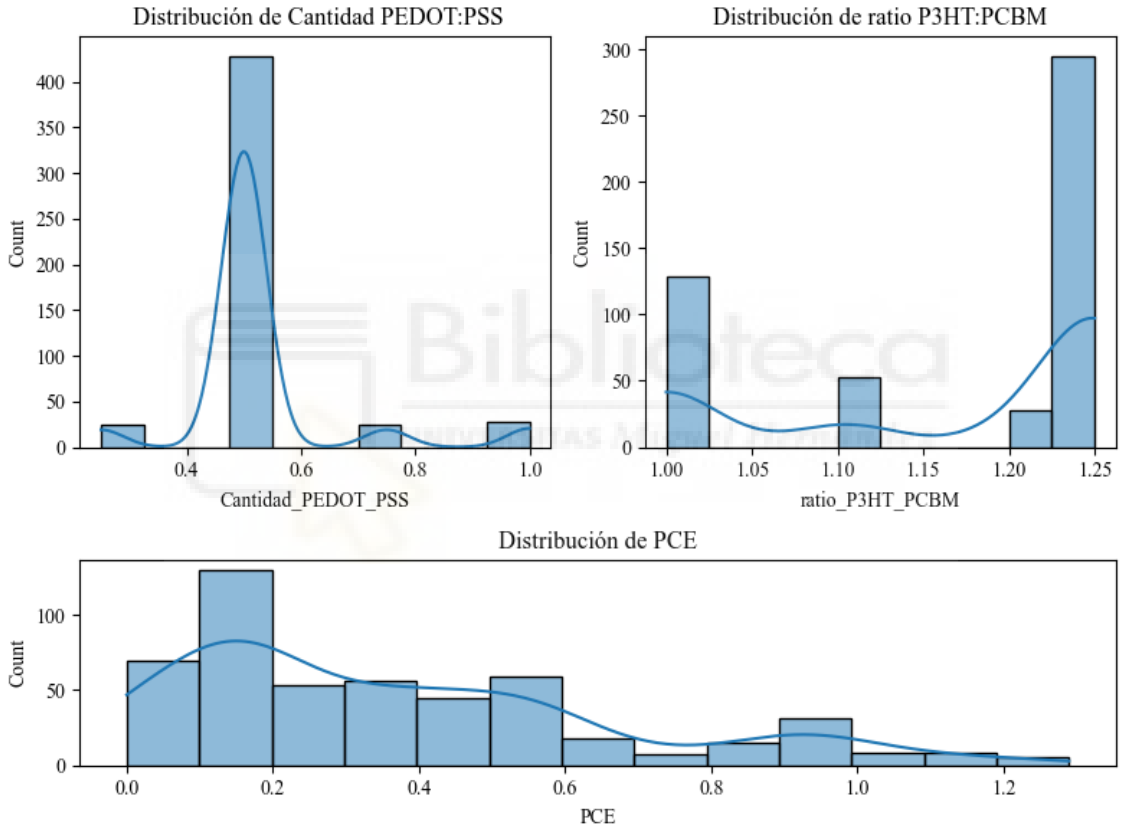



*Figura 23: Histogramas de la distribución de las distintas variables en el conjunto de datos tras la reducción de dimensionalidad.*

Estos resultados sugieren que, aunque la reducción de dimensionalidad ha permitido simplificar el modelo y mejorar su rendimiento en general, todavía existen limitaciones debidas a la naturaleza del conjunto de datos. Las distribuciones desiguales de algunas características pueden estar limitando la capacidad de los modelos para generalizar de manera efectiva. Por lo tanto, futuros esfuerzos deberían centrarse en mejorar el equilibrio

en la representación de las características clave y en la recolección de más datos que permitan una mayor diversidad en las variables.

En conclusión, la reducción de dimensionalidad ha sido efectiva para reducir el ruido en los modelos y mejorar su capacidad de predicción, especialmente en el caso del modelo Gradient Boost. No obstante, la naturaleza del conjunto de datos sigue presentando desafíos en términos de variabilidad, lo que sugiere la necesidad de continuar optimizando el proceso de recolección y selección de características. Esto permitirá mejorar aún más la precisión y robustez de los modelos, con el fin de maximizar la eficiencia de las células solares orgánicas.

#### 5.2.2.5 Refinamiento de la Selección de Datos: Introducción de Nuevas Variables

En base a los resultados obtenidos en los análisis anteriores, se decidió incorporar una nueva variable denominada Célula. El objetivo de esta variable es ayudar a los modelos a distinguir mejor entre los distintos experimentos, facilitando la identificación y análisis de las variaciones en la cantidad de PEDOT:PSS y ratio de P3HT:PCBM.

Para ello, se asignó un número único a cada célula de cada experimento, con el fin de diferenciarlas dentro del conjunto de datos. La tabla (Tabla 8) muestra cómo se ha realizado esta asignación, en la que cada célula de los diferentes experimentos recibió un valor numérico específico. La hipótesis detrás de esta estrategia es que, al añadir esta variable, los modelos podrán distinguir mejor las diferencias experimentales.

| Valor Asignado | Célula |
|---|---|
| 1 | 10/27 célula 1 |
| 2 | 10/27 celula 2 |
| 3 | 10/27 celula 3 |
| 4 | 10/27 celula 4 |
| 5 | 11/04 celula 1 |
| 6 | 11/04 celula 2 |
| 7 | 11/04 celula 3 |
| 8 | 11/04 celula 4 |
| 9 | 11/04 celula 5 |
| 10 | 11/11 celula 1 |
| 11 | 11/11 celula 2 |
| 12 | 11/11 celula 3 |
| 13 | 11/11 celula 4 |
| 14 | 11/26 celula 1 |
| 15 | 11/26 celula 2 |
| 16 | 11/26 celula 3 |
| 17 | 11/26 celula 4 |
| 18 | 12 /10 celula 1 |
| 19 | 12/10 celula 2 |
| 20 | 12/10 celula 3 |
| 21 | 12/10 celula 4 |

*Tabla 8: Asignación de valores a las células de cada experimento.*

Con esta nueva columna integrada en el conjunto de datos, se procedió a un nuevo análisis de la importancia de las características, excluyendo la variable **Punto de Rocío**, dado que esta había mostrado los peores resultados en el análisis de PFI. El modelo fue evaluado utilizando el mismo método que en los análisis previos.

Los resultados tras la incorporación de la nueva variable **Célula** se resumen en la siguiente tabla (Tabla 9). Ambos modelos, Random Forest y Gradient Boost, mostraron una mejora significativa en las métricas de rendimiento, especialmente el modelo Gradient Boost, que alcanzó un coeficiente de determinación ($R^2$) cercano a 1. Estas mejoras se reflejan también en una notable reducción en los errores absolutos medios (MAE) y los errores cuadráticos medios (MSE), lo que sugiere que la inclusión de esta nueva variable ha mejorado considerablemente la capacidad de los modelos para capturar relaciones complejas entre las variables.

| Modelo | Validación cruzada ($R^2$) | MAE | MSE | $R^2$ Score |
|---|---|---|---|---|
| Random Forest | 0.7029 ± 0.0336 | 0.1071 | 0.0175 | 0.8039 |
| Gradient Boost | 0.8747 ± 0.0256 | 0.0699 | 0.0072 | 0.9193 |

*Tabla 9: Métricas de error para el entrenamiento de los modelos, tras evaluar el modelo después de la adición de la variable Célula.*

En la gráfica (Figura 24), se observa cómo la importancia de las variables ha cambiado tras la inclusión de la nueva característica. En el caso de Random Forest, la distribución de los pesos de las demás características se mantuvo relativamente constante, mientras que la nueva variable Célula ganó una relevancia considerable. Por otro lado, en el modelo Gradient Boost, tanto la ratio P3HT:PCBM como la nueva variable Célula dominaron sobre el resto de las características, aunque las variables meteorológicas también adquirieron mayor peso en comparación con análisis anteriores.

En conclusión, la adición de la variable Célula ha tenido un impacto positivo en la capacidad predictiva de ambos modelos, con una mejora especialmente destacada en el modelo Gradient Boost. Estos resultados proporcionan una base sólida para futuras predicciones y análisis, ya que permiten identificar combinaciones de variables que optimicen la eficiencia de conversión de potencia (PCE) de las células solares orgánicas.

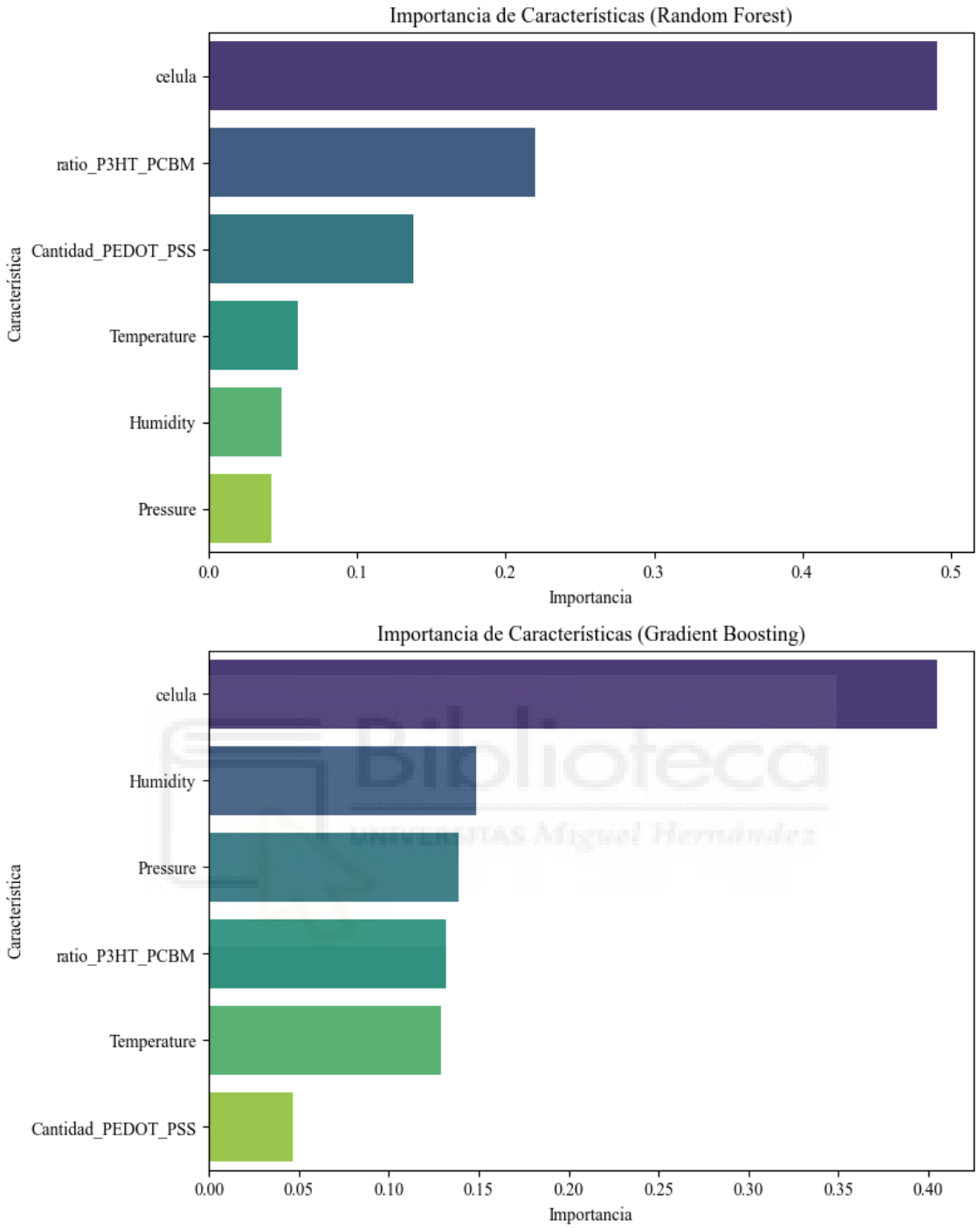



*Figura 24: Gráfico de barras, importancia de las distintas variables del conjunto de datos obtenida por Random Forest y Gradient Boost, una vez añadida la variable célula.*

#### 5.2.2.6 Predicción de Eficiencia de Conversión: Comparación de Modelos y Combinaciones

Una vez entrenados los modelos, se procedió a utilizarlos para predecir valores de eficiencia de conversión de potencia (PCE) con el fin de identificar las combinaciones óptimas de variables. Para realizar estas predicciones, se empleó la clase *itertools.product*, que permite generar un iterable con las diferentes combinaciones posibles de las variables sin necesidad de almacenar todos los datos en memoria, optimizando así el uso de recursos computacionales.

Para abarcar un amplio rango de posibilidades, se seleccionaron tres modelos previamente entrenados: el modelo sin la variable rGO tras el ajuste de hiperparámetros, el modelo con la nueva variable Célula, y el modelo reducido que solo incluye variables estructurales de las células orgánicas. La variable Célula se mantuvo con los mismos valores, mientras que el resto de los valores de las variables fueron ampliados ligeramente, estableciendo un rango de 0.1 a 1.2 para la cantidad de PEDOT:PSS, y de 0.75 a 1.5 para el ratio de P3HT:PCBM. Se generaron 10 puntos en esos rangos, lo que permitió explorar un rango más amplio de posibilidades.

En cuanto a las variables meteorológicas, se optó por utilizar sus valores medios para los modelos que las consideraban, ya que la incorporación de estas variables habría incrementado significativamente el número de combinaciones posibles, complicando la capacidad de cómputo sin aportar un control directo sobre las predicciones. Asimismo, por el análisis de correlación ya realizado se sabe que efecto tienen las variaciones de estos valores en la PCE.

**Predicciones del Modelo sin rGO tras el Ajuste de Hiperparámetros**

Los resultados de la primera prueba, realizados con el modelo sin la variable rGO tras el ajuste de hiperparámetros, se muestran en la tabla 10. Esta tabla presenta las diez mejores predicciones de PCE.

| Cantidad PEDOT:PSS | ratio P3HT:PCBM | Predicted PCE RF | Predicted PCE GB |
|---|---|---|---|
| 0.833333333 | 1 | 0.781434878 | 0.902928429 |
| 0.833333333 | 0.916666667 | 0.781434878 | 0.902928429 |
| 0.711111111 | 0.75 | 0.781434878 | 0.902928429 |
| 0.711111111 | 0.833333333 | 0.781434878 | 0.902928429 |
| 0.711111111 | 0.916666667 | 0.781434878 | 0.902928429 |
| 0.711111111 | 1 | 0.781434878 | 0.902928429 |
| 0.833333333 | 0.833333333 | 0.781434878 | 0.902928429 |
| 0.833333333 | 0.75 | 0.781434878 | 0.902928429 |
| 1.077777778 | 0.75 | 0.762968433 | 0.763472085 |
| 1.077777778 | 0.833333333 | 0.762968433 | 0.763472085 |

*Tabla 10: Mejores predicciones de ambos modelos con el conjunto de datos sin rGO tras el ajuste de hiperparámetros.*

Se observó que las mejores predicciones de PCE se obtenían con valores intermedios-altos en la cantidad de PEDOT:PSS junto con valores bajos de ratio P3HT:PCBM, lo que sugiere que estas combinaciones pueden ser óptimas para maximizar la eficiencia.

**Predicciones del Modelo Reducido**

En la segunda prueba, se utilizó el modelo reducido, que solo considera las variables estructurales de las células orgánicas. Los resultados obtenidos se resumen en la tabla (Tabla 11), y se observa una tendencia similar al modelo anterior en cuanto a las combinaciones de predicción. Sin embargo, en este caso, el modelo Random Forest predijo valores de PCE más altos que Gradient Boost.

| Cantidad PEDOT:PSS | ratio P3HT:PCBM | Predicted PCE RF | Predicted PCE GB |
|---|---|---|---|
| 0.833333333 | 1 | 0.970029623 | 0.687273366 |
| 0.833333333 | 0.916666667 | 0.970029623 | 0.687273366 |
| 0.711111111 | 0.75 | 0.970029623 | 0.687273366 |
| 0.711111111 | 0.833333333 | 0.970029623 | 0.687273366 |
| 0.711111111 | 0.916666667 | 0.970029623 | 0.687273366 |
| 0.711111111 | 1 | 0.970029623 | 0.687273366 |
| 0.833333333 | 0.833333333 | 0.970029623 | 0.687273366 |
| 0.833333333 | 0.75 | 0.970029623 | 0.687273366 |
| 0.711111111 | 1.083333333 | 0.723919578 | 0.469812983 |
| 0.833333333 | 1.083333333 | 0.723919578 | 0.469812983 |

*Tabla 11: Mejores predicciones de ambos modelos con el conjunto de datos reducido.*

### Predicciones del Modelo con la Variable Célula

Finalmente, la tercera prueba se realizó utilizando el modelo que incluía la nueva variable Célula, la cual había mejorado significativamente el rendimiento en etapas anteriores. Los valores máximos de PCE no coincidieron completamente entre Random Forest y Gradient Boost, pero los resultados (Tabla 12 - Tabla 13) fueron similares a los modelos previos, mostrando consistencia en las combinaciones óptimas de variables.

| Cantidad PEDOT:PSS | ratio P3HT:PCBM | Predicted PCE RF | Predicted PCE GB |
|---|---|---|---|
| 0.833333333 | 0.833333333 | 0.825956747 | 0.935894599 |
| 0.833333333 | 0.75 | 0.825956747 | 0.935894599 |
| 0.833333333 | 1 | 0.825956747 | 0.935894599 |
| 0.833333333 | 0.916666667 | 0.825956747 | 0.935894599 |
| 0.711111111 | 0.75 | 0.825956747 | 0.935894599 |
| 0.711111111 | 0.833333333 | 0.825956747 | 0.935894599 |
| 0.711111111 | 0.916666667 | 0.825956747 | 0.935894599 |
| 0.711111111 | 1 | 0.825956747 | 0.935894599 |
| 1.2 | 0.916666667 | 0.804792853 | 0.923837143 |
| 1.2 | 1 | 0.804792853 | 0.923837143 |

*Tabla 12: Mejores predicciones del modelo Gradient Boost con el conjunto de datos con la variable Célula.*

| Cantidad PEDOT:PSS | ratio P3HT:PCBM | Predicted PCE RF | Predicted PCE GB |
|---|---|---|---|
| 0.833333333 | 0.833333333 | 0.825956747 | 0.935894599 |
| 0.833333333 | 0.75 | 0.825956747 | 0.935894599 |
| 0.833333333 | 1 | 0.825956747 | 0.935894599 |
| 0.833333333 | 0.916666667 | 0.825956747 | 0.935894599 |
| 0.711111111 | 0.75 | 0.825956747 | 0.935894599 |
| 0.711111111 | 0.833333333 | 0.825956747 | 0.935894599 |
| 0.711111111 | 0.916666667 | 0.825956747 | 0.935894599 |
| 0.711111111 | 1 | 0.825956747 | 0.935894599 |
| 0.711111111 | 1 | 0.812604878 | 0.870039906 |
| 0.711111111 | 0.916666667 | 0.812604878 | 0.870039906 |

*Tabla 13: Mejores predicciones del modelo Random Forest con el conjunto de datos con la variable Célula.*

En general, los resultados obtenidos en las distintas pruebas revelan que combinaciones de valores intermedios-altos de cantidad de PEDOT:PSS, junto con valores bajos de la ratio P3HT:PCBM, tienden a maximizar la PCE en ambos modelos. El modelo Gradient Boost con la variable Célula fue el que mostró las mejores predicciones, sugiriendo que esta combinación de modelo y características es prometedora para futuras optimizaciones.

El comportamiento en las predicciones, donde algunos valores parecen mantenerse consistentes o con poca variación ante diferentes combinaciones de entradas, puede explicarse por varios factores. En primer lugar, la estructura del modelo Random Forest y Gradient Boost tiende a priorizar características con mayor impacto relativo en las predicciones, lo que puede llevar a resultados similares si ciertas variables dominan el conjunto de datos o si las combinaciones no aportan suficiente variabilidad informativa. Además, es posible que las relaciones entre las características seleccionadas y la variable objetivo (PCE) no sean lineales o estén limitadas por una alta redundancia entre las variables. Esto podría llevar a que los modelos encuentren patrones óptimos dentro de un rango específico de valores, explicando la similitud en las predicciones observadas. Finalmente, el rango estrecho de predicciones también podría estar influenciado por el tamaño del conjunto de datos y la naturaleza de las combinaciones generadas, que pueden limitar la capacidad de los modelos para discriminar entre configuraciones más diversas. Este análisis subraya la importancia de continuar explorando tanto la relevancia de las variables como la optimización de los modelos para maximizar el aprovechamiento de la información disponible.

### 5.2.3 Evaluación Temporal de la Eficiencia

El análisis del conjunto de datos como una serie temporal tiene como objetivo identificar patrones específicos que puedan acelerar el deterioro de las células solares, proporcionando información sobre las condiciones y características que influyen negativamente en la estabilidad de la eficiencia de conversión de potencia (PCE). Con este análisis, buscamos identificar qué características tienen un impacto negativo en la PCE a lo largo del tiempo y en qué células se observa un deterioro más rápido.

Se visualizaron todas las células para observar la tendencia de la PCE a lo largo del tiempo para los distintos experimentos.

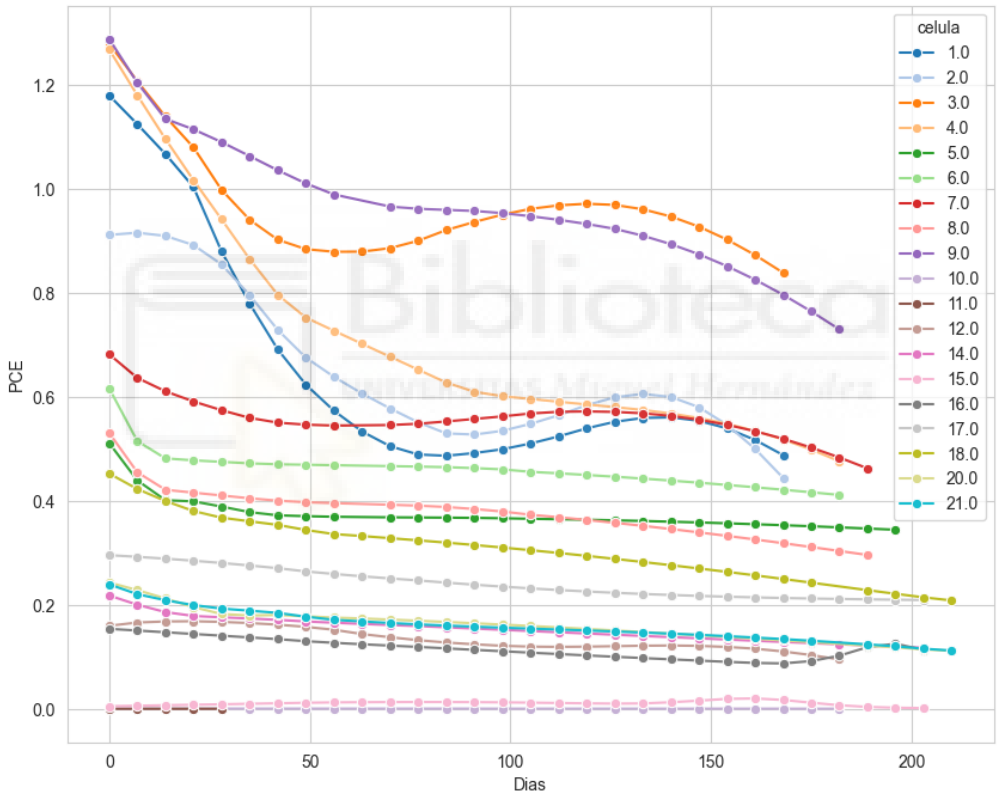



*Figura 25: Gráfico de líneas, visualiza los datos de PCE para los distintos experimentos a lo largo del tiempo.*

La gráfica (Figura 25) muestra la evolución de la eficiencia de conversión de potencia (PCE) a lo largo del tiempo para cada célula solar. Podemos observar variaciones significativas entre las diferentes células, indicando que algunas estructuras pueden tener un mejor desempeño en términos de estabilidad de PCE que otras.

### 5.2.3.1 Impacto de las estructuras de las células solares

Como primera prueba se calcula la tasa de degradación de las células. Para ello se determina la tasa de cambio de la PCE para cada célula a lo largo del tiempo. Esto se logra calculando la diferencia porcentual entre las mediciones sucesivas de PCE, proporcionando una métrica que mide la reducción de PCE. Con estos datos se comparan las tasas de degradación de diferentes células para identificar qué estructuras presentan una mayor o menor estabilidad.

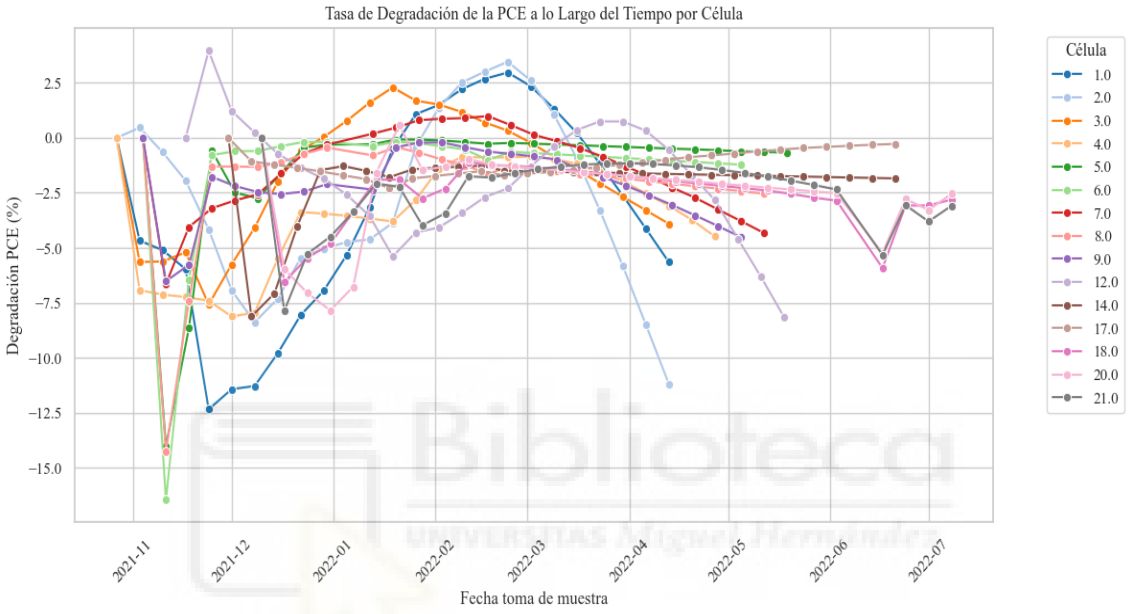



*Figura 26: Gráfico de líneas, muestra el porcentaje de disminución de la PCE entre toma de muestras.*

La gráfica anterior (Figura 26), muestra la tasa de degradación de la eficiencia de conversión de potencia de las células solares orgánicas a lo largo del tiempo. En los primeros meses del estudio (noviembre-diciembre de 2021), se observa una gran variabilidad en la degradación de las células. Algunas experimentando descensos abruptos en su eficiencia, mientras que otras mantienen cierta estabilidad. Esto sugiere una adaptación inicial a las condiciones ambientales, posiblemente influida por diferencias en la calidad de las células o en su fabricación.

Conforme avanza el tiempo, la mayoría de las células presentan una tendencia general hacia una degradación progresiva de la eficiencia, algo típico en materiales orgánicos expuestos a factores externos. Sin embargo, entre enero y abril de 2022, se aprecia una fase de estabilidad relativa, donde las tasas de degradación se estabilizan. Este comportamiento sugiere que, una vez superado el periodo inicial de deterioro, las células

alcanzan un estado más equilibrado en su rendimiento bajo condiciones ambientales controladas.

Las diferencias observadas en la degradación entre las distintas células también ponen de manifiesto la influencia de su estructura y composición. En particular, las células que incorporan materiales como rGO y PEDOT:PSS en configuraciones optimizadas parecen mostrar una mayor resistencia a la degradación. Es posible que algún fallo en la fabricación haga que estas apenas generen efecto fotovoltaico, y los valores obtenidos sean tan bajos que hagan aparentar que la tasa de degradación en buena.

Se observan anomalías, como una recuperación inesperada de la eficiencia en algunas células, visible en ciertos puntos (picos de degradación positiva). Estas recuperaciones podrían deberse a variaciones en las condiciones experimentales, errores de medición o factores externos.

A partir de estos datos se ha obtenido la media de la degradación de todas las células, los resultados muestran como las células 11, 4 y 1 tienen una degradación media mayor al resto. Estos datos pueden ser útiles para saber qué estructuras se han deteriorado más con el paso del tiempo.

| Célula | Degradación media PCE | Cantidad PEDOT:PSS | ratio P3HT:PCBM | rGO |
|---|---|---|---|---|
| 11 | -7.742019267 | 0.5 | 1.25 | 1 |
| 4 | -3.52969085 | 1 | 1 | 0 |
| 1 | -3.353928933 | 0.25 | 1 | 0 |
| 2 | -2.753461446 | 0.5 | 1 | 0 |
| 20 | -2.53988187 | 0.5 | 1.25 | 1 |
| 18 | -2.539563646 | 0.5 | 1.25 | 0 |
| 21 | -2.482789216 | 0.5 | 1.25 | 1 |
| 9 | -2.147260606 | 0.5 | 1.25 | 0 |
| 8 | -2.09849128 | 0.5 | 1.11 | 0 |
| 14 | -2.024656575 | 0.5 | 1.25 | 0 |
| 15 | -1.906645695 | 0.5 | 1.25 | 1 |
| 12 | -1.872071754 | 0.5 | 1.25 | 1 |
| 10 | -1.770094444 | 0.5 | 1.25 | 0 |
| 3 | -1.634962864 | 0.75 | 1 | 0 |
| 6 | -1.484578847 | 0.5 | 1.1 | 0 |
| 7 | -1.407650101 | 0.5 | 1 | 0 |
| 5 | -1.351494768 | 0.5 | 1.2 | 0 |
| 17 | -1.141494451 | 0.5 | 1.25 | 1 |
| 16 | -0.914503528 | 0.5 | 1.25 | 1 |

*Tabla 14: Tasa de degradación media de las distintas células, representada de más alta a más baja.*

Es importante tener en cuenta que, aunque algunas células presentan una tasa de degradación media más baja, también tienen una PCE promedio más reducida. Por ejemplo, la célula 1 es la tercera con mayor tasa de degradación, pero también es una de las que mayor PCE tienen durante las primeras semanas desde su fabricación. Por lo que tenemos que entender todos estos resultados como un conjunto antes de tomar conclusiones precipitadas.

### 5.2.3.2 Estudio de la Degradación con Aprendizaje Automático

Se estudió el conjunto de datos para identificar las estructuras con mayor resistencia al paso del tiempo. Para ello, se remuestrearon los valores a una frecuencia diaria, permitiendo una observación más detallada.

A partir del conjunto de datos se entrenaron dos modelos utilizando esta vez la variable temporal, la cual se obtuvo a partir de las fechas de muestreo. Se usaron los días desde la

primera fecha de toma de muestras para cada uno de los experimentos como variable temporal.

En este análisis también se descartaron las células con presencia de grafeno y se utilizó el valor medio de las variables meteorológicas para evitar el coste computacional. Esto fue posible porque, según el análisis de correlación ya realizado, se conoce el efecto de las variaciones de estos valores en la PCE.

Los modelos entrenados usando Random Forest y Gradient Boost obtuvieron los resultados mostrados en la siguiente tabla (Tabla 15). Estos muestran cómo el modelo en ambos casos solo puede explicar entorno al 50% de la variabilidad del conjunto de datos.

| Modelo | Validación cruzada ($R^2$) | MAE | MSE | $R^2$ Score |
|---|---|---|---|---|
| Random Forest | 0.4152 +/- 0.0127 | 0.1521 | 0.0475 | 0.4541 |
| Gradient Boost | 0.4308 +/- 0.0144 | 0.1393 | 0.0458 | 0.4735 |

*Tabla 15: Resultados de los modelos entrenados para observar la degradación de la PCE con el tiempo.*

Para este modelo, se realizaron predicciones cada 7 días en lugar de a diario, debido al gran número de combinaciones y la necesidad de observar cambios a lo largo del tiempo. Esto facilitó la identificación de estructuras con mejor resistencia al paso del tiempo. Cabe destacar que las predicciones no se realizaron a futuro, sino dentro del período en el que se muestrearon todas las células, permitiendo analizar la evolución de su comportamiento en ese intervalo de tiempo.

Visualizar estos datos puede ser complicado dado que tenemos una cantidad de combinaciones muy alta y queremos observar las estructuras que menos se deterioren pero que a su vez mejoren en la medida de lo posible la PCE máxima por lo que no podemos únicamente calcular la degradación media.

Teniendo todo esto en cuenta lo mencionado anteriormente se decidió dar una puntuación ponderada a las distintas predicciones siguiendo la siguiente expresión:

$$Puntuación = 0.7 \cdot (PCE\ máxima) - 0.3 \cdot (Tasa\ de\ deterioro)$$

siendo la PCE máxima la obtenida por el algoritmo Random Forest y la tasa de deterioro obtenida aproximando una recta a cada curva de PCE y obteniendo la pendiente de esta.

Una vez realizados los cálculos sobre las predicciones se obtuvo la siguiente tabla (Tabla 16). En ella se observa como las variables con cantidades de PEDOT:PSS más altas, parecen tener una mejor puntación. La PCE máxima en todos los casos para las predicciones ha sido en el primer día.

| Cantidad PEDOT:PSS | ratio P3HT:PCBM | PCE máxima predicha | Ratio de degradación | Puntuación de Rendimiento |
|---|---|---|---|---|
| 1.07777778 | 0.75 | 0.89992609 | -0.00048905 | 0.51143016 |
| 1.07777778 | 0.83333333 | 0.89992609 | -0.00048905 | 0.51143016 |
| 1.07777778 | 0.91666667 | 0.89992609 | -0.00048905 | 0.51143016 |
| 1.07777778 | 1 | 0.89992609 | -0.00048905 | 0.51143016 |
| 1.2 | 0.91666667 | 0.89992609 | -0.00048905 | 0.51143016 |
| 1.2 | 1 | 0.89992609 | -0.00048905 | 0.51143016 |
| 1.2 | 0.83333333 | 0.89992609 | -0.00048905 | 0.51143016 |
| 1.2 | 0.75 | 0.89992609 | -0.00048905 | 0.51143016 |
| 0.95555556 | 0.83333333 | 0.89992609 | -0.00048905 | 0.51143016 |
| 0.95555556 | 0.75 | 0.89992609 | -0.00048905 | 0.51143016 |

*Tabla 16: Mejores Resultados de las predicciones con variables temporales.*

## 5.2.4 Análisis del Conjunto de Datos con ROBERT

En este apartado usaremos el paquete ROBERT de Python para analizar el conjunto de datos y comparar los resultados con los obtenidos. Para ello vamos a utilizar dos de los subconjuntos creados: el que contiene la variable Célula y el utilizado en análisis temporal para analizar el deterioro en función de las estructuras.

### 5.2.4.1 Análisis con el Conjunto de Datos con la Variable 'Célula'

En este análisis, se utilizó el paquete ROBERT para generar dos informes que presentan los resultados tras realizar las pruebas pertinentes. Ambos informes se encuentran en Anexo II: Informes de ROBERT.

Los resultados del primer conjunto de datos fueron muy satisfactorios. En primer lugar, se observa que los modelos que obtuvieron los mejores resultados fueron Gradient Boosting (sin PFI) y Gradient Boosting (con PFI).

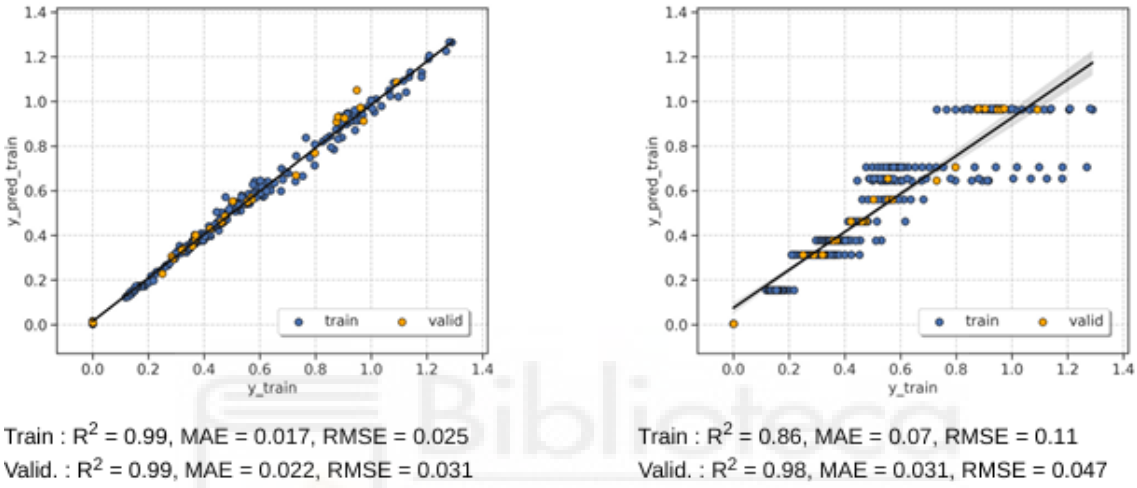



*Figura 27: Comparación del modelo Gradient Boosting sin (izq.) y con (der.) descriptores PFI. El uso de PFI no mejora el ajuste. Imagen procedente del reporte generado por ROBERT [30].*

El modelo Random Forest mostró un rendimiento muy alto, con un $R^2$ de 0.99 en el conjunto de entrenamiento y 0.99 en el de validación. Los errores medidos en términos de MAE y RMSE también fueron bajos, lo cual indica una excelente capacidad predictiva. En comparación, ROBERT parece haber encontrado un mejor ajuste de hiperparámetros o de la partición entre entrenamiento y prueba, lo que permitió mejorar nuestro resultado inicial de $R^2 = 0.8$ con Random Forest hasta alcanzar 0.99 en entrenamiento.

El modelo Gradient Boosting, que se entrenó usando solo los descriptores más importantes (PFI), tuvo un rendimiento algo menor durante el entrenamiento ($R^2$ de 0.86), pero excelente en validación ($R^2$ de 0.98), lo que sugiere que el modelo está bien ajustado y tiene buena capacidad de generalización. En nuestro caso, este modelo alcanzó un $R^2$ de 0.91 en entrenamiento, pero su desempeño en validación fue inferior ($R^2 = 0.87$). Esto indica que el ajuste realizado por ROBERT favorece una mejor generalización y ayuda a evitar el sobreajuste.

En cuanto a las advertencias, no se detectaron advertencias graves en el conjunto de datos. Sin embargo, se observó una distribución desigual de los datos, lo cual concuerda con lo deducido del análisis previo.

ROBERT también identificó algunos puntos del conjunto de datos como atípicos o *outliers*. Aunque no son muy numerosos (5.9% del conjunto de entrenamiento y solo tres *outliers* en validación), estos pueden impactar en la precisión del modelo. La presencia de valores atípicos sugiere la necesidad de revisar si estos datos son errores de medición o casos especiales que requieren un tratamiento específico. Estos datos ya fueron revisados y se eliminaron aquellos considerados errores de medición; sin embargo, los puntos detectados por ROBERT podrían deberse a que el umbral utilizado por el programa es más restrictivo.

Los resultados de los valores que maximizan la PCE según las predicciones se muestran en las siguientes tablas (Tabla 17 y Tabla 18).

| Cantidad PEDOT:PSS | ratio P3HT:PCBM |
|---|---|
| 0.83333333 | 1 |
| 0.83333333 | 0.91666667 |
| 0.83333333 | 0.83333333 |
| 0.83333333 | 0.75 |
| 0.71111111 | 1.5 |

*Tabla 17: Valores de las predicciones del modelo Random Forest.*

| Cantidad PEDOT:PSS | ratio P3HT:PCBM |
|---|---|
| 1.2 | 1 |
| 1.2 | 0.91666667 |
| 1.2 | 0.83333333 |
| 1.2 | 0.75 |
| 1.07777778 | 1 |

*Tabla 18: Valores de las predicciones del modelo Gradient Boost.*

El informe sugiere añadir más puntos de datos significativos para mejorar la fiabilidad del modelo. Además, podría considerarse la inclusión de más descriptores o la eliminación de aquellos menos útiles para aumentar la precisión predictiva.

En resumen, los resultados obtenidos son prometedores, con un buen rendimiento en validación para ambos modelos. Sin embargo, existen áreas que podrían mejorarse, como la distribución de los datos de salida y la reducción de la colinealidad entre los descriptores.

#### 5.2.4.2 Análisis con el Conjunto de Datos con la Variable 'Días'

En este análisis, se utilizó el conjunto de datos con la variable 'días' para evaluar el deterioro en el rendimiento de las celdas solares a lo largo del tiempo. Se entrenaron dos modelos: Random Forest con PFI y sin PFI para evaluar el rendimiento predictivo.

En ambos casos se obtuvieron resultados similares, siendo un poco mejor el modelo con PFI que ha obtenido un $R^2$ de 0.54 en el entrenamiento y un $R^2$ de 0.6 en la validación.

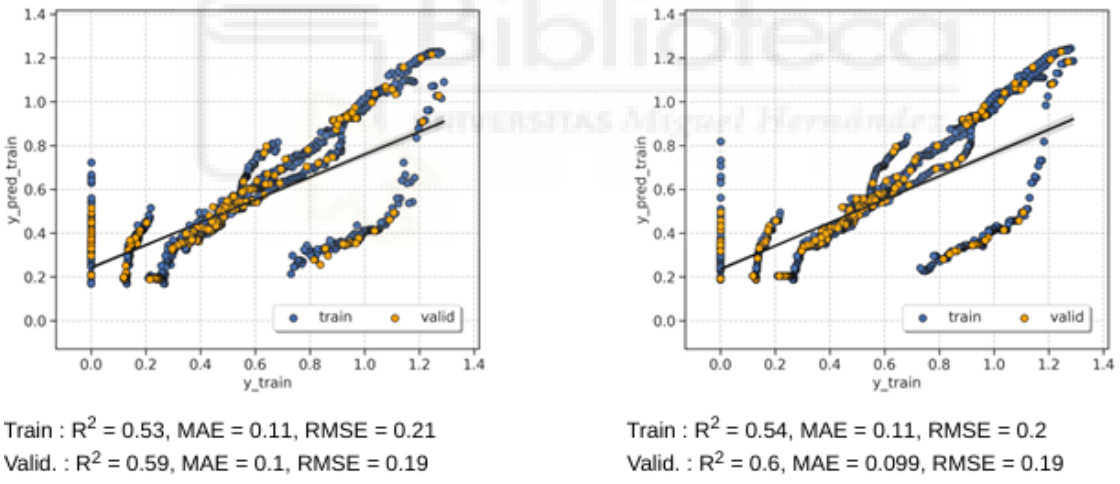



*Figura 28: Comparación del modelo Gradient Boosting sin (izq.) y con (der.) descriptores PFI.. Imagen procedente del reporte generado por ROBERT[30].*

En cuanto a las advertencias, ROBERT identificó varios aspectos que afectan la confiabilidad del modelo:

- **Distribución desigual de los valores de salida**: se observó una distribución no uniforme de los valores, lo cual puede afectar la capacidad del modelo para generalizar sus predicciones.

- **Predicciones imprecisas**: los modelos presentaron una alta variabilidad en las predicciones, como lo demuestra el análisis de la desviación estándar.

Además, se identificaron varios valores atípicos tanto en el conjunto de entrenamiento como en el de validación. En el entrenamiento, se identificaron 170 valores atípicos (8.5% del total), mientras que en la validación se encontraron 17 (7.6%). Estos valores atípicos podrían estar afectando negativamente el rendimiento de los modelos y es recomendable analizarlos en detalle para determinar si deben ser eliminados o tratados de manera especial.

La validación cruzada (5-fold CV) mostró un rendimiento bajo, con un $R^2$ promedio de 0.59 para el modelo sin PFI y 0.6 para el modelo con PFI, lo cual indica una baja capacidad de generalización y sugiere que estos modelos no son adecuados para este conjunto de datos.

Los resultados obtenidos de las predicciones se muestran en la siguiente tabla (Tabla 19), únicamente se muestran los valores del modelo sin PFI. Los valores mostrados son los que dan una mayor PCE.

| Cantidad PEDOT:PSS | ratio P3HT:PCBM |
|---|---|
| 0.83333333 | 1 |
| 0.83333333 | 0.91666667 |
| 0.83333333 | 0.83333333 |
| 0.83333333 | 0.75 |
| 0.71111111 | 1 |

*Tabla 19: Valores de las predicciones del modelo Random Forest.*

En resumen, los modelos utilizados en este análisis no mostraron un buen rendimiento predictivo, y existen varios factores que afectan su confiabilidad, como la distribución desigual de los datos, la alta correlación entre descriptores y la presencia de valores atípicos.

# 6. CONCLUSIONES

En este apartado se presentan las principales conclusiones derivadas del estudio sobre la PCE de las células solares orgánicas, utilizando modelos predictivos y técnicas de *machine learning*. El objetivo principal fue identificar patrones y características relevantes que permitan mejorar la eficiencia de estas celdas solares bajo diferentes condiciones experimentales.

## 6.1 Logros Alcanzados y Contribuciones del Trabajo

A lo largo de este trabajo se han logrado avances significativos en la caracterización y modelado de las células solares orgánicas. Los modelos de Random Forest y Gradient Boosting permitieron realizar predicciones sobre la PCE con un rendimiento prometedor, alcanzando valores de $R^2$ superiores a 0.9 en algunos casos, lo cual refleja una buena capacidad para capturar las relaciones entre las características del conjunto de datos y la eficiencia de las células solares.

El uso del paquete ROBERT también ha sido un punto clave en este análisis, facilitando la identificación de características relevantes y la evaluación de la calidad de los datos.

Los resultados obtenidos en este estudio permiten concluir que, con relación al primer objetivo, la degradación de las OSC presenta un comportamiento bifásico: una disminución inicial pronunciada de la PCE, seguida de una fase de estabilización con una degradación más gradual. Este fenómeno fue particularmente notable en las células 3 y 9, que mostraron una mayor retención de PCE a lo largo del tiempo, sugiriendo que su configuración material o estructural podría conferir una mayor resistencia a los procesos de degradación intrínsecos. Concretamente, la célula 9 se caracteriza por una cantidad de PEDOT:PSS de 0,5 y una ratio P3HT:PCBM de 1,25, mientras que la célula 3 presenta una cantidad de PEDOT:PSS de 0,75 y una ratio P3HT:PCBM de 1.

Respecto al segundo objetivo, el análisis exploratorio y predictivo permitió identificar las que las variables meteorológicas (como la temperatura y la humedad) afectaban de manera negativa a la eficiencia de las células solares. Valores elevados de temperatura resultaron en una disminución en la eficiencia de conversión, mientras que altos niveles de humedad contribuyeron a un deterioro más rápido de la estructura de las celdas, afectando negativamente la PCE.

Finalmente, el análisis mediante algoritmos de *machine learning* permitió identificar combinaciones óptimas de materiales para maximizar la PCE y la estabilidad. Se observó que altos niveles de PEDOT:PSS y bajos ratios de P3HT:PCBM favorecen una estructura más eficiente. Estos resultados se alinean con la eliminación experimental de componentes como el óxido de grafeno (rGO), cuyo impacto negativo en la eficiencia fue detectado durante el análisis exploratorio. En conjunto, los modelos computacionales presentan un marco para optimizar la selección de materiales y la resistencia ambiental, avanzando hacia su implementación práctica en condiciones realistas.

## 6.2 Limitaciones y Áreas de Mejora

El presente estudio, si bien ha aportado resultados valiosos para la optimización de células solares orgánicas, presenta ciertas limitaciones que abren oportunidades de mejora en futuras investigaciones. En primer lugar, la distribución desigual de los datos experimentales, derivada de variaciones en la cantidad de muestras por configuración material, podría introducir sesgos en los modelos predictivos, limitando la generalización de las conclusiones. Para mitigar este efecto, se recomienda diseñar experimentos con una distribución más equilibrada de variables clave, como proporciones de materiales, garantizando así un conjunto de datos más robusto y representativo. Asimismo, incrementar la frecuencia y el período de medición permitiría capturar con mayor precisión la evolución temporal de la degradación, particularmente en fases críticas como la estabilización post-degradación inicial.

En cuanto al modelado computacional, si bien los algoritmos de *machine learning* empleados (Random Forest y Gradient Boosting) demostraron alta capacidad predictiva ($R^2 > 0.9$), su dependencia de relaciones lineales o basadas en árboles podría no explotar plenamente patrones complejos en los datos. La implementación de técnicas de *Deep Learning*, como redes neuronales convolucionales o modelos de atención, podría mejorar la captura de interacciones no lineales entre materiales y factores ambientales, especialmente si se complementan con conjuntos de datos ampliados.

Finalmente, la eliminación de componentes como el óxido de grafeno, identificada como perjudicial, debería validarse en condiciones operativas más diversas para descartar efectos contextuales. En resumen, estas mejoras enriquecerían los modelos y también fortalecerían la transición desde el laboratorio hacia aplicaciones industriales, donde la escalabilidad y la estabilidad a largo plazo son críticas.

## 6.3 Perspectivas Futuras y Recomendaciones para Investigaciones Posteriores

Los hallazgos de este trabajo ofrecen un punto de partida sólido para futuras investigaciones en el diseño y optimización de células solares orgánicas. Una línea prioritaria es la mejora de los protocolos de muestreo y la expansión de los experimentos, tal como se señaló en las limitaciones anteriores.

Otra vía prometedora, dado que los experimentos presentaban similitudes en las variables analizadas, es la incorporación de variables adicionales que capturen con mayor precisión las condiciones experimentales. Por ejemplo, la variable 'Célula', utilizada para diferenciar configuraciones específicas, demostró ser crítica en la mejora de la capacidad predictiva de los modelos. Ampliar este enfoque incluyendo parámetros como métodos de fabricación, espesores de capas o técnicas de encapsulación podría refinar aún más la comprensión de los factores que influyen en la eficiencia y degradación de las OSC.

En el ámbito computacional, se recomienda explorar algoritmos avanzados de *machine learning* y *deep learning* para modelar interacciones no lineales y dinámicas temporales complejas, como la relación entre degradación y condiciones ambientales fluctuantes. Complementariamente, la optimización sistemática de hiperparámetros, respaldada por técnicas de validación cruzada, aumentaría la robustez y generalización de los modelos.

Aunque la eficiencia actual de las OSC es insuficiente para aplicaciones prácticas, los hallazgos orientan mejoras futuras. Se recomienda probar estrategias como encapsulación o exposición a estrés ambiental (humedad, ciclos térmicos, radiación UV) para evaluar su impacto en la estabilidad, incluso con bajas eficiencias. Validar combinaciones clave bajo estas condiciones podría optimizar su escalabilidad y resistencia, sentando bases para prototipos más viables.

En síntesis, la convergencia de enfoques experimentales rigurosos, modelos computacionales innovadores y una gestión inteligente de datos potenciaría la eficiencia de las OSC.

# 7. BIBLIOGRAFÍA


[1] «Share of electricity from renewables». [En línea]. Disponible en: https://ourworldindata.org/grapher/share-electricity-renewables

[2] UNEP, «Decade of renewable energy investment led by solar tops USD 25 trillion», *UNEP News Stories*, may 2023, [En línea]. Disponible en: https://www.unep.org/news-and-stories/press-release/decade-renewable-energy-investment-led-solar-tops-usd-25-trillion

[3] «Installed solar PV capacity». [En línea]. Disponible en: https://ourworldindata.org/grapher/installed-solar-pv-capacity?country=OWID_WRL~CHN~IND~ESP~BRA~OWID_EU27~FRA~DEU~USA

[4] Ossila, «Solar Cells: A Guide to Theory and Measurement». [En línea]. Disponible en: https://www.ossila.com/pages/solar-cells-theory

[5] N. R. E. Laboratory, «Best Research-Cell Efficiency Chart | Photovoltaic Research - NREL». [En línea]. Disponible en: https://www.nrel.gov/pv/cell-efficiency.html

[6] M. A. Green, «Photovoltaics: technology overview», *Energy Policy*, vol. 28, n.º 14, pp. 989-998, 2000, doi: https://doi.org/10.1016/S0301-4215(00)00086-0.

[7] A. S. Al-Ezzi y M. N. M. Ansari, «Photovoltaic Solar Cells: A Review», *Appl. Syst. Innov.*, vol. 5, n.º 4, 2022, doi: 10.3390/asi5040067.

[8] B. Parida, S. Iniyan, y R. Goic, «A review of solar photovoltaic technologies», *Renew. Sustain. Energy Rev.*, vol. 15, n.º 3, pp. 1625-1636, 2011, doi: https://doi.org/10.1016/j.rser.2010.11.032.

[9] B. Kippelen y J.-L. Brédas, «Organic photovoltaics», *Energy Env. Sci*, vol. 2, n.º 3, pp. 251-261, 2009, doi: 10.1039/B812502N.

[10] Y.-W. Su, S.-C. Lan, y K.-H. Wei, «Organic photovoltaics», *Mater. Today*, vol. 15, n.º 12, pp. 554-562, 2012, doi: https://doi.org/10.1016/S1369-7021(13)70013-0.

[11] W. Chamorro y S. Urrego Riveros, «Celdas solares orgánicas, una perspectiva hacia el futuro», *ELEMENTOS*, vol. 2, may 2013, doi: 10.15765/e.v2i2.181.

[12] E. K. Solak y E. Irmak, «Advances in organic photovoltaic cells: a comprehensive review of materials, technologies, and performance», *RSC Adv*, vol. 13, n.º 18, pp. 12244-12269, 2023, doi: 10.1039/D3RA01454A.

[13] T. Kita, Y. Harada, y S. Asahi, «The Conversion Efficiency of a Solar Cell as Determined by the Detailed Balance Model», en *Energy Conversion Efficiency of Solar Cells*, Singapore: Springer Singapore, 2019, pp. 55-79. doi: 10.1007/978-981-13-9089-0_5.

[14] T. Kita, Y. Harada, y S. Asahi, «Actual Calculation of Solar Cell Efficiencies», en *Energy Conversion Efficiency of Solar Cells*, Singapore: Springer Singapore, 2019, pp. 81-137. doi: 10.1007/978-981-13-9089-0_6.

[15] S. Shalev-Shwartz y S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[16] L. Breiman, «Random Forests», *Mach Learn*, vol. 45, n.º 1, pp. 5-32, oct. 2001, doi: 10.1023/A:1010933404324.

[17] A. Liaw y M. Wiener, «Classification and Regression by RandomForest», *Forest*, vol. 23, nov. 2001.

[18] V. Thanh Ha, «Experimental Study on Remaining Useful Life Prediction of Lithium-Ion Batteries Based on Three Regressions Models for Electric Vehicle Applications». junio de 2023. doi: 10.20944/preprints202306.0999.v1.

[19] J. Friedman, «Greedy Function Approximation: A Gradient Boosting Machine», *Ann. Stat.*, vol. 29, nov. 2000, doi: 10.1214/aos/1013203451.

[20] T. Zhang *et al.*, «Improving Convection Trigger Functions in Deep Convective Parameterization Schemes Using Machine Learning», *J. Adv. Model. Earth Syst.*, vol. 13, may 2021, doi: 10.1029/2020MS002365.

[21] D. Chicco, M. J. Warrens, y G. Jurman, «The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation», *PeerJ Comput. Sci.*, vol. 7, p. e623, 2021, doi: 10.7717/peerj-cs.623.

[22] D. Dalmau y J. V. Alegre Requena, «ROBERT: Bridging the Gap between Machine Learning and Chemistry», *ChemRxiv*, 2023, doi: 10.26434/chemrxiv-2023-k994h.

[23] F. Pedregosa *et al.*, «Scikit-learn Machine Learning in Python», *J. Mach. Learn. Res.*, vol. 12, pp. 2825-2830, 2011.

[24] S. M. Lundberg y S.-I. Lee, «A Unified Approach to Interpreting Model Predictions», en *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, y R. Garnett, Eds., Curran Associates, Inc., 2017, pp. 4765-4774. [En línea]. Disponible en: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[25] A. S. Mahdi, L. M. Shaker, y A. Alamiery, «Recent advances in organic solar cells: materials, design, and performance», *J. Opt.*, vol. 53, n.º 2, pp. 1403-1419, abr. 2024, doi: 10.1007/s12596-023-01262-2.

[26] L. Fu *et al.*, «Machine learning assisted prediction of charge transfer properties in organic solar cells by using morphology-related descriptors», *Nano Res.*, vol. 16, n.º 2, pp. 3588-3596, feb. 2023, doi: 10.1007/s12274-022-5000-4.

[27] A. Mahmood y J.-L. Wang, «Machine learning for high performance organic solar cells: current scenario and future prospects», *Energy Env. Sci*, vol. 14, n.º 1, pp. 90-105, 2021, doi: 10.1039/D0EE02838J.

[28] A. Jain *et al.*, «Advances in organic solar cells: Materials, progress, challenges and amelioration for sustainable future», *Sustain. Energy Technol. Assess.*, vol. 63, p. 103632, 2024, doi: https://doi.org/10.1016/j.seta.2024.103632.

[29] Weather Underground, «Weather Forecasts and Reports». 2024. [En línea]. Disponible en: {{url_wunderground}}

[30] D. Dalmau y J. V. Alegre Requena, «ROBERT: Bridging the Gap between Machine Learning and Chemistry», *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 2024. doi: 10.1002/WCMS.1733.

[31] D. Berrar y others, «Cross-validation.» 2019.

[32] K. Sveidqvist y Contributors to Mermaid, *Mermaid: Generate diagrams from markdown-like text*. (diciembre de 2014). [En línea]. Disponible en: https://github.com/mermaid-js/mermaid

[33] Miro, «Miro: The Visual Collaboration Platform for Any Team». 2024. [En línea]. Disponible en: https://miro.com

# 8. ANEXOS

## 8.1 Anexo I: Código para Ajuste y Entrenamiento de Modelos

```
from hyperopt import fmin, tpe, hp, Trials, STATUS_OK
from sklearn.ensemble import RandomForestRegressor,
GradientBoostingRegressor
from sklearn.metrics import r2_score
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
import numpy as np
import pandas as pd

# Cargar los datos
df = pd.read_csv('Datos.csv')
y = df['PCE']
X = df.drop(['PCE'], axis=1)

# Escalado de características
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# División en conjunto de entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y,
                                                    test_size=0.2,
                                                    random_state=42)

def optimize_and_train(X_train, X_test, y_train, y_test):
    # Espacio de búsqueda para Random Forest
    space_rf = {
        'random_state': hp.choice('random_state', range(1, 100)),
        'n_estimators': hp.choice('n_estimators', range(10, 301)),
        'max_features': hp.choice('max_features', ['sqrt', 'log2']),
        'max_depth': hp.choice('max_depth', range(10, 100)),
        'min_samples_split': hp.choice('min_samples_split',
                                       range(2, 11)),
        'min_samples_leaf': hp.choice('min_samples_leaf',
                                      range(1, 5)),
       'min_weight_fraction_leaf':
       hp.uniform('min_weight_fraction_leaf', 0.0, 0.5),
        'ccp_alpha': hp.uniform('ccp_alpha', 0.0, 0.5),
        'oob_score': hp.choice('oob_score', [True, False]),
        'max_samples': hp.choice('max_samples',
                                 np.linspace(0.1, 1.0, 10))
    }

    # Espacio de búsqueda para Gradient Boosting
    space_gb = {
        'random_state': hp.choice('random_state', range(1, 100)),
        'n_estimators': hp.choice('n_estimators', range(10, 301)),
        'learning_rate': hp.choice('learning_rate',
                                   [0.01, 0.05, 0.1]),
        'max_features': hp.choice('max_features', ['sqrt', 'log2']),
        'max_depth': hp.choice('max_depth', range(3, 6)),
```

```
        'min_samples_split': hp.choice('min_samples_split',
                                       range(2, 11)),
        'min_samples_leaf': hp.choice('min_samples_leaf',
                                      range(1, 5)),
        'min_weight_fraction_leaf':
        hp.uniform('min_weight_fraction_leaf', 0.0, 0.5),
        'subsample': hp.uniform('subsample', 0.1, 1.0),
        'validation_fraction': hp.uniform('validation_fraction',
                                          0.1, 0.3),
        'ccp_alpha': hp.uniform('ccp_alpha', 0.0, 0.5)
    }

    # Función objetivo para Random Forest
    def objective_rf(params):
        model = RandomForestRegressor(**params)
        model.fit(X_train, y_train)
        y_pred = model.predict(X_test)
        score = r2_score(y_test, y_pred)
        return {'loss': -score, 'status': STATUS_OK}

    # Función objetivo para Gradient Boosting
    def objective_gb(params):
        model = GradientBoostingRegressor(**params)
        model.fit(X_train, y_train)
        y_pred = model.predict(X_test)
        score = r2_score(y_test, y_pred)
        return {'loss': -score, 'status': STATUS_OK}

    # Optimización para Random Forest
    trials_rf = Trials()
    best_rf = fmin(fn=objective_rf, space=space_rf, algo=tpe.suggest,
                   max_evals=100, trials=trials_rf)

    # Optimización para Gradient Boosting
    trials_gb = Trials()
    best_gb = fmin(fn=objective_gb, space=space_gb, algo=tpe.suggest,
                   max_evals=100, trials=trials_gb)

    # Entrenar los modelos con los mejores hiperparámetros
    rf_model = RandomForestRegressor(**best_rf)
    gb_model = GradientBoostingRegressor(**best_gb)

    rf_model.fit(X_train, y_train)
    gb_model.fit(X_train, y_train)

    # Predicciones
    y_pred_rf = rf_model.predict(X_test)
    y_pred_gb = gb_model.predict(X_test)

    return rf_model, gb_model, y_pred_rf, y_pred_gb

modelo_rf, modelo_gb, predicciones_rf, predicciones_gb =
optimize_and_train(X_train, X_test, y_train, y_test)
```

## 8.2 Anexo II: Informes de ROBERT

En el siguiente apartado se presentan los informes generados mediante el uso del módulo ROBERT para el análisis automático de los datos. En primer lugar, se muestra el informe correspondiente al conjunto de datos con la variable célula añadida. A continuación, se presentan los resultados obtenidos al incluir la variable días.

ROBERT v 1.2.1 2025/02/19 00:28:29

**How to cite:** Dalmau, D.; Alegre Requena, J. V. WIREs Comput Mol Sci. 2024, DOI: 10.1002/WCMS.1733

## Section A. ROBERT Score

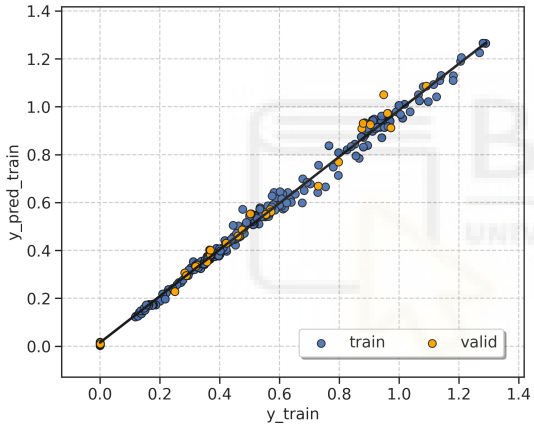
*This score is designed to evaluate the models using different metrics.*

**No PFI (standard descriptor filter):**

Model = GB · Train:Validation = 90:10

Points(train+valid.):descriptors = 323:7

Score = 9 / 10


Train : $R^2$ = 0.99, MAE = 0.017, RMSE = 0.025

Valid. : $R^2$ = 0.99, MAE = 0.022, RMSE = 0.031

**Severe warnings**

No severe warnings detected

**Moderate warnings**

Uneven y distribution (Section C)

**Overall assessment**

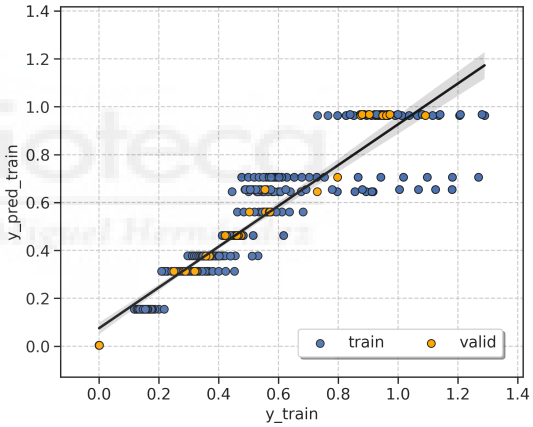
The model seems reliable

**PFI (only most important descriptors):**

Model = GB · Train:Validation = 90:10

Points(train+valid.):descriptors = 323:3

Score = 9 / 10


Train : $R^2$ = 0.86, MAE = 0.07, RMSE = 0.11

Valid. : $R^2$ = 0.98, MAE = 0.031, RMSE = 0.047

**Severe warnings**

No severe warnings detected

**Moderate warnings**

Uneven y distribution (Section C)

**Overall assessment**

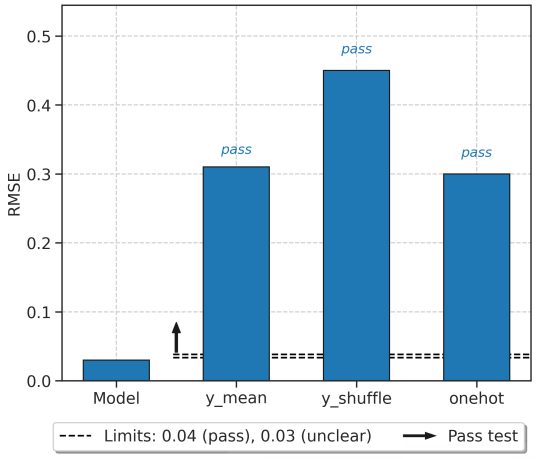
The model seems reliable

## Section B. Advanced Score Analysis

*This section explains each component that comprises the ROBERT score.*

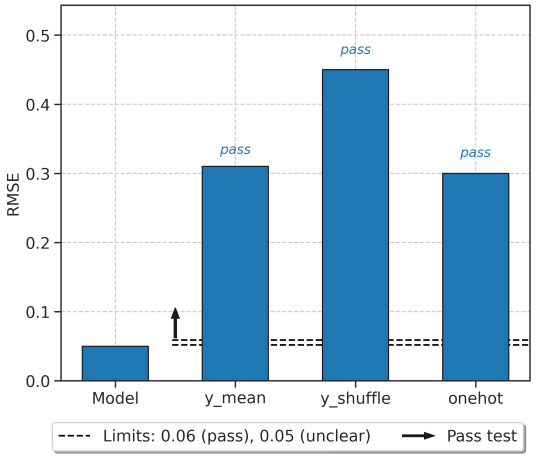
### 1. Model vs "flawed" models (3 / 3)

The model predicts right for the right reasons.
Pass: +1, Unclear: 0, Fail: -1. *Details here.*




### 1. Model vs "flawed" models (3 / 3)

The model predicts right for the right reasons.
Pass: +1, Unclear: 0, Fail: -1. *Details here.*




### 2. Predictive ability of the model (2 / 2)

Good predictive ability with $R^2$ (valid.) = 0.99.
$R^2$ 0.70-0.85: +1, $R^2$ >0.85: +2.

### 2. Predictive ability of the model (2 / 2)

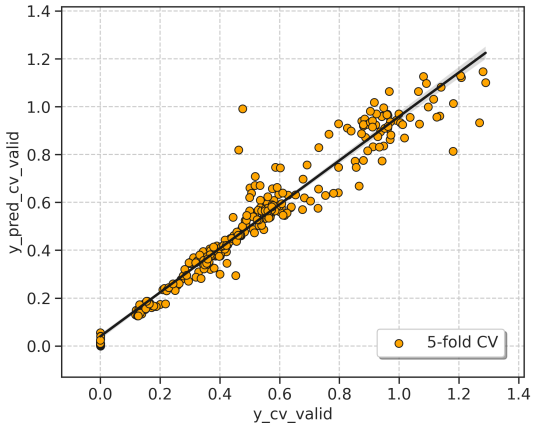
Good predictive ability with $R^2$ (valid.) = 0.98.
$R^2$ 0.70-0.85: +1, $R^2$ >0.85: +2.

### 3. Cross-validation (5-fold CV) of the model

Overfitting analysis on the model with 3a and 3b:

3a. CV predictions train + valid. (2 / 2)

Good predictive ability with $R^2$ (5-fold CV) = 0.94.
$R^2$ 0.70-0.85: +1, $R^2$ >0.85: +2.

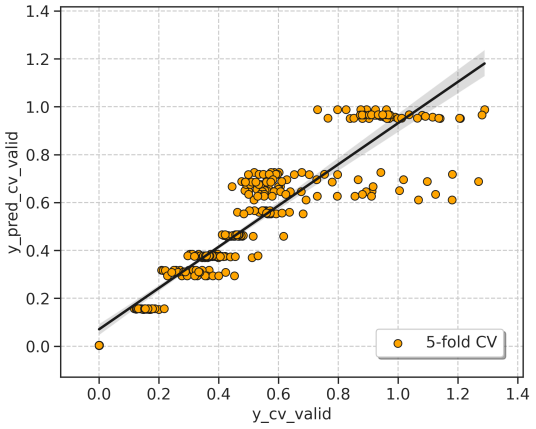



### 3. Cross-validation (5-fold CV) of the model

Overfitting analysis on the model with 3a and 3b:

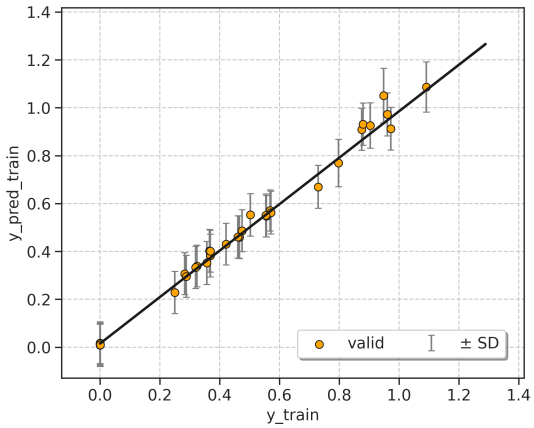
3a. CV predictions train + valid. (2 / 2)

Good predictive ability with $R^2$ (5-fold CV) = 0.86.
$R^2$ 0.70-0.85: +1, $R^2$ >0.85: +2.



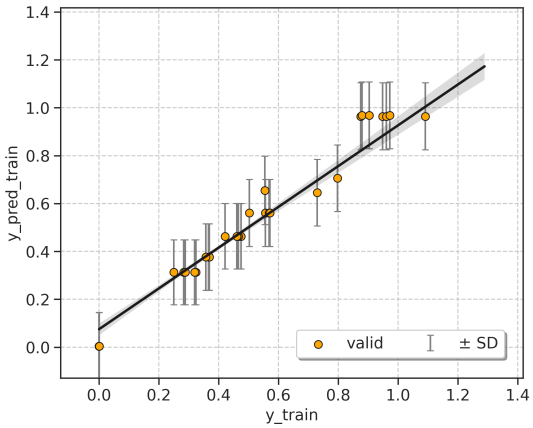
3b. Avg. standard deviation (SD) (1 / 2 )

Moderate variation, 4*SD (valid.) = 0.4 (28% y-range).
4*SD 25-50% y-range: +1, 4*SD < 25% y-range: +2.
*Details here.*




3b. Avg. standard deviation (SD) (1 / 2 )

Moderate variation, 4*SD (valid.) = 0.6 (43% y-range).
4*SD 25-50% y-range: +1, 4*SD < 25% y-range: +2.
*Details here.*




**4. Points(train+valid.):descriptors** (1 / 1 )

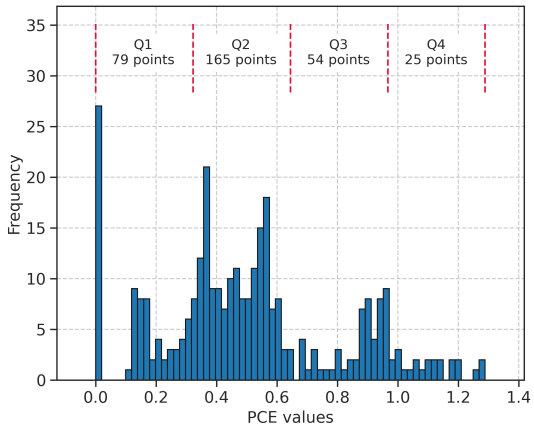
Decent number of descps. (ratio 323:7).
5 or more points per descriptor: +1.

**4. Points(train+valid.):descriptors** (1 / 1 )

Decent number of descps. (ratio 323:3).
5 or more points per descriptor: +1.

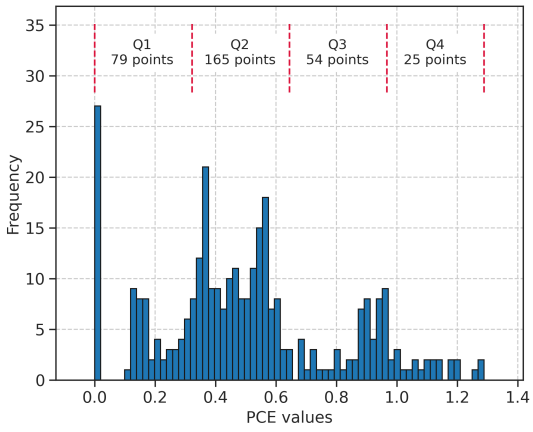
## Section C. Distribution of y Values

*This section shows the distribution of y values within the training and validation sets.*




**y distribution analysis**

x WARNING! Your data is not uniform (Q4 has 25 points while Q2 has 165)

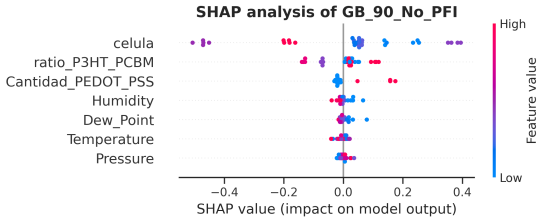


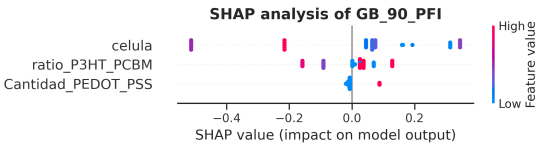

**y distribution analysis**

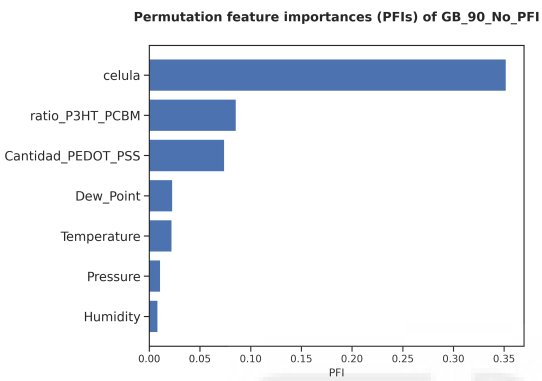
x WARNING! Your data is not uniform (Q4 has 25 points while Q2 has 165)

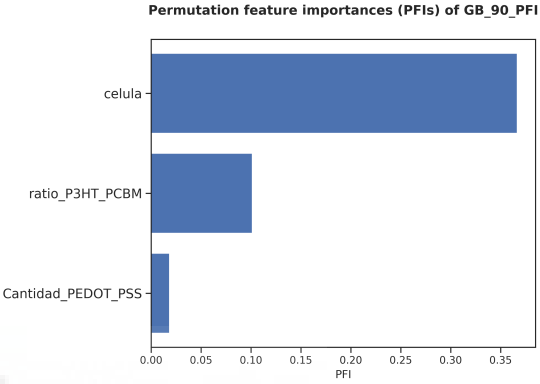
## Section D. Feature Importances

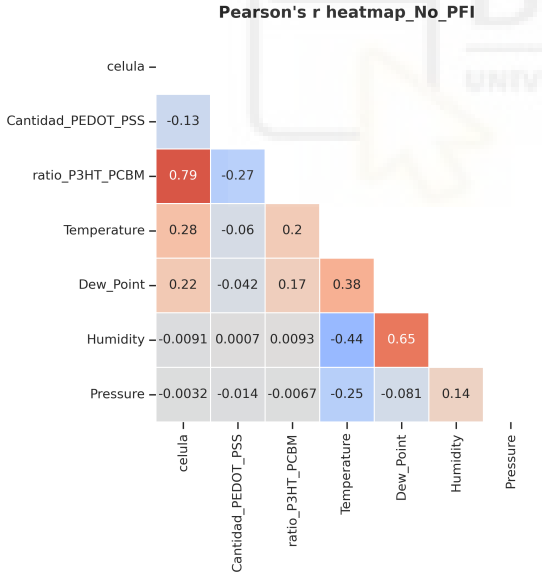
*This section presents feature importances measured using the validation set.*

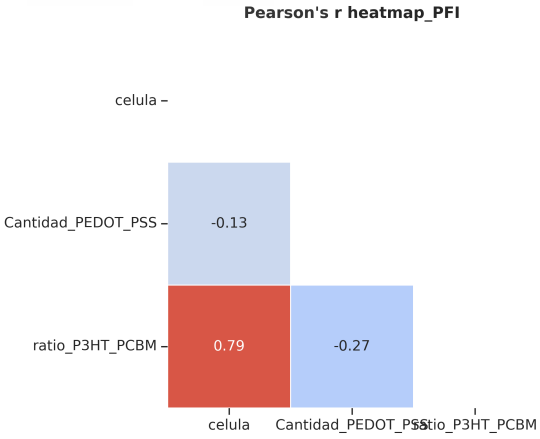



### Correlation analysis

o Correlations between variables are acceptable

### Correlation analysis

o Correlations between variables are acceptable

## Section E. Outlier Analysis

*This section detects outliers using the standard deviation (SD) of errors from the training set.*

**No PFI (standard descriptor filter):**

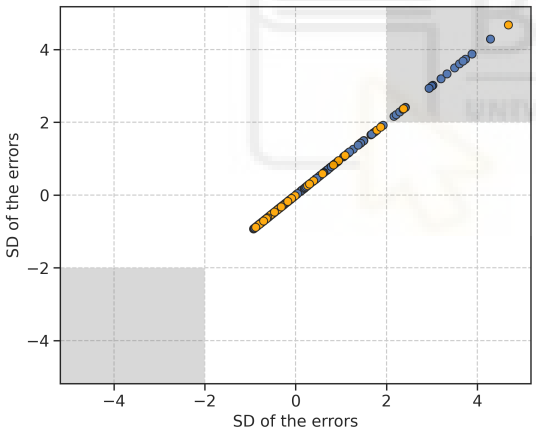
Outliers (max. 10 shown)
Train: 17 outliers out of 290 datapoints (5.9%)
- 88 (3.0 SDs)
- 98 (2.2 SDs)
- 100 (2.2 SDs)
- 112 (2.3 SDs)
- 222 (4.3 SDs)
- 224 (3.2 SDs)
- 225 (3.0 SDs)
- 226 (3.5 SDs)
- 231 (3.7 SDs)
- 233 (3.9 SDs)

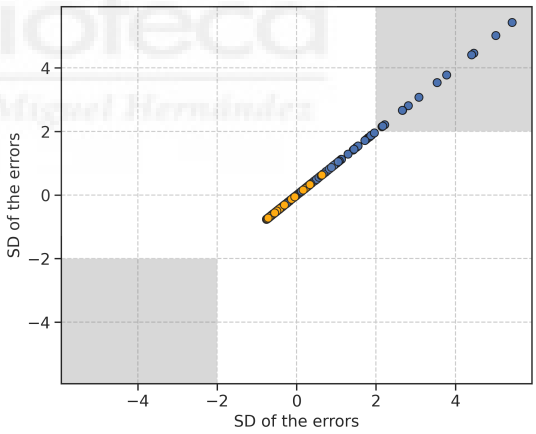
Validation: 3 outliers out of 33 datapoints (9.1%)
- 92 (4.7 SDs)
- 259 (2.4 SDs)
- 294 (2.4 SDs)

**PFI (only most important descriptors):**

Outliers (max. 10 shown)
Train: 13 outliers out of 290 datapoints (4.5%)
- 110 (2.8 SDs)
- 221 (2.7 SDs)
- 223 (4.5 SDs)
- 224 (3.5 SDs)
- 230 (5.4 SDs)
- 249 (2.7 SDs)
- 274 (2.2 SDs)
- 279 (2.2 SDs)
- 287 (2.2 SDs)
- 299 (3.1 SDs)

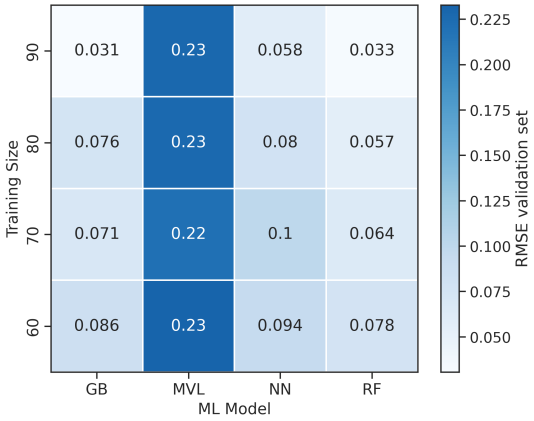
Validation: 0 outliers out of 33 datapoints (0.0%)

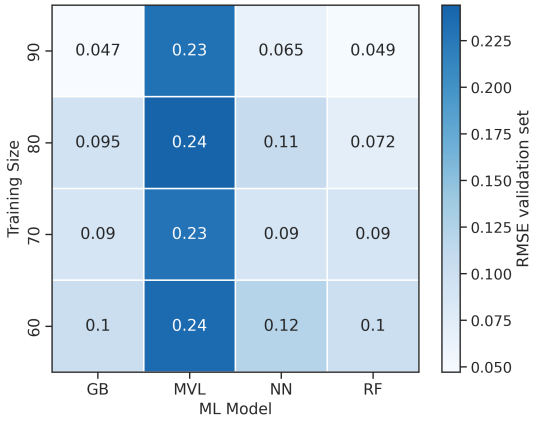


### Section F. Model Screening

*This section compares different combinations of hyperoptimized algorithms and partition sizes.*




### Section G. Reproducibility

*This section provides all the instructions to reproduce the results presented.*

**1. Download these files *(the authors should have uploaded the files as supporting information!)*:**

- CSV database (datos_final.csv)
- External test set (Pred_cel.csv)

**2. Install and adjust the versions of the following Python modules:**

- Install ROBERT and its dependencies: conda install -y -c conda-forge robert
- Adjust ROBERT version: pip install robert==1.2.1
- Install scikit-learn-intelex: pip install scikit-learn-intelex==2025.2.0

*(if scikit-learn-intelex is not installed, slightly different results might be obtained)*

**3. Run ROBERT using this command line in the folder with the CSV databases:**

python -m robert --ignore "Fecha_datos" --names "index" --y "PCE" --csv_name "datos_final.csv" --csv_test "Pred_cel.csv"

**4. Execution time, Python version and OS:**

Originally run in Python 3.11.11 using Linux #1 SMP PREEMPT_DYNAMIC Thu Jun 27 21:05:47 UTC 2024

Total execution time: 220.6 seconds *(the number of processors should be specified by the user)*

## Section H. Transparency

*This section contains important parameters used in scikit-learn models and ROBERT.*

**1. Parameters of the scikit-learn models (same keywords as used in scikit-learn):**

| No PFI (standard descriptor filter): | PFI (only most important descriptors): |
|---|---|
| sklearn model: GradientBoostingRegressor | sklearn model: GradientBoostingRegressor |
| random_state: 0 | random_state: 0 |
| names: index | names: index |
| n_estimators: 100 | n_estimators: 100 |
| max_depth: 5 | max_depth: 5 |
| max_features: 0.75 | max_features: 0.75 |
| min_samples_split: 2 | min_samples_split: 2 |
| min_samples_leaf: 1 | min_samples_leaf: 1 |
| min_weight_fraction_leaf: 0 | min_weight_fraction_leaf: 0 |
| ccp_alpha: 0 | ccp_alpha: 0 |
| learning_rate: 0.05 | learning_rate: 0.05 |
| subsample: 1.0 | subsample: 1.0 |
| validation_fraction: 0.2 | validation_fraction: 0.2 |

**2. ROBERT options for data split (KN or RND), predict type (REG or CLAS) and hyperopt error (RMSE, etc.):**

| No PFI (standard descriptor filter): | PFI (only most important descriptors): |
|---|---|
| split: RND | split: RND |
| type: reg | type: reg |
| error_type: rmse | error_type: rmse |

## Section I. Abbreviations

*Reference section for the abbreviations used.*

**ACC:** accuracy
**ADAB:** AdaBoost
**CSV:** comma separated values
**CLAS:** classification
**CV:** cross-validation
**F1 score:** balanced F-score
**GB:** gradient boosting
**GP:** gaussian process
**KN:** k-nearest neighbors
**MAE:** root-mean-square error
**MCC:** Matthew's correl. coefficient
**ML:** machine learning
**MVL:** multivariate lineal models
**NN:** neural network
**PFI:** permutation feature importance
**R2:** coefficient of determination
**REG:** Regression
**RF:** random forest
**RMSE:** root mean square error
**RND:** random
**SHAP:** Shapley additive explanations
**VR:** voting regressor

## Section J. New Predictions

*Predictions of the external test set added with the csv_test option.*

csv_test predictions (sorted, max. 20 shown)

From /PREDICT/csv_test/...No_PFI.csv

| index | PCE_pred ± sd |
|---|---|
| 2963 | 0.95 ± 0.09 |
| 2962 | 0.95 ± 0.09 |
| 2961 | 0.95 ± 0.09 |
| 2960 | 0.95 ± 0.09 |
| 2953 | 0.95 ± 0.09 |
| 2952 | 0.95 ± 0.09 |
| 2951 | 0.95 ± 0.09 |
| 2950 | 0.95 ± 0.09 |
| 2863 | 0.95 ± 0.09 |
| 2862 | 0.95 ± 0.09 |
| ... | ... |
| 9019 | 0.01 ± 0.09 |
| 9018 | 0.01 ± 0.09 |
| 9017 | 0.01 ± 0.09 |
| 9016 | 0.01 ± 0.09 |
| 9015 | 0.01 ± 0.09 |
| 9009 | 0.01 ± 0.09 |
| 9008 | 0.01 ± 0.09 |
| 9007 | 0.01 ± 0.09 |
| 9006 | 0.01 ± 0.09 |
| 9005 | 0.01 ± 0.09 |

csv_test predictions (sorted, max. 20 shown)

From /PREDICT/csv_test/..._PFI.csv

| index | PCE_pred ± sd |
|---|---|
| 2963 | 0.97 ± 0.14 |
| 2962 | 0.97 ± 0.14 |
| 2961 | 0.97 ± 0.14 |
| 2960 | 0.97 ± 0.14 |
| 2953 | 0.97 ± 0.14 |
| 2952 | 0.97 ± 0.14 |
| 2951 | 0.97 ± 0.14 |
| 2950 | 0.97 ± 0.14 |
| 2863 | 0.97 ± 0.14 |
| 2862 | 0.97 ± 0.14 |
| ... | ... |
| 9027 | 0.0 ± 0.14 |
| 9026 | 0.0 ± 0.14 |
| 9019 | 0.0 ± 0.14 |
| 9018 | 0.0 ± 0.14 |
| 9017 | 0.0 ± 0.14 |
| 9016 | 0.0 ± 0.14 |
| 9009 | 0.0 ± 0.14 |
| 9008 | 0.0 ± 0.14 |
| 9007 | 0.0 ± 0.14 |
| 9006 | 0.0 ± 0.14 |

**Miscellaneous**

*General tips to improve the models and instructions to predict new values.*

Some general tips to improve the score

1. Adding meaningful datapoints might help to improve the model. Also, using a uniform population of datapoints across the whole range of y values usually helps to obtain reliable predictions across the whole range. More information about the range of y values used is available in Section C.

2. Adding meaningful descriptors or replacing/deleting the least useful descriptors used might help. Feature importances are gathered in Section D.

How to predict new values with these models?

1. Create a CSV database with the new points, including the necessary descriptors.

2. Place the CSV file in the parent folder (i.e., where the module folders were created)

3. Run the PREDICT module as 'python -m robert --predict --csv_test FILENAME.csv'.

4. The predictions will be shown at the end of the resulting PDF report and will be stored in the last column of two CSV files called MODEL_SIZE_test(_No)_PFI.csv, which are in the PREDICT folder.

ROBERT v 1.2.1 2025/02/19 01:11:57

**How to cite:** Dalmau, D.; Alegre Requena, J. V. WIREs Comput Mol Sci. 2024, DOI: 10.1002/WCMS.1733

## Section A. ROBERT Score

*This score is designed to evaluate the models using different metrics.*

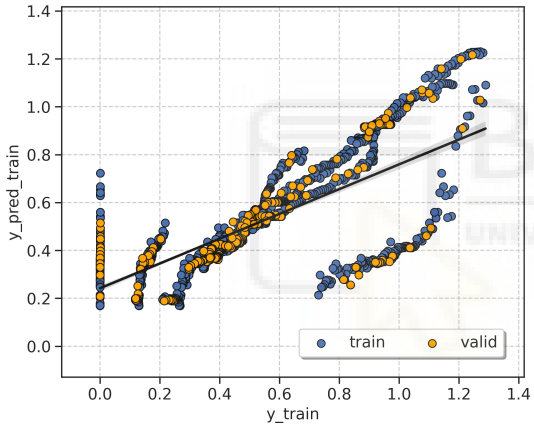
**No PFI (standard descriptor filter):**

Model = RF · Train:Validation = 90:10

Points(train+valid.):descriptors = 2232:6

Score = 4 / 10

WEAK




Train : $R^2$ = 0.53, MAE = 0.11, RMSE = 0.21

Valid. : $R^2$ = 0.59, MAE = 0.1, RMSE = 0.19

**Severe warnings**

No severe warnings detected

**Moderate warnings**

Imprecise predictions (Section B.3b)

Uneven y distribution (Section C)

**Overall assessment**

The model is unreliable

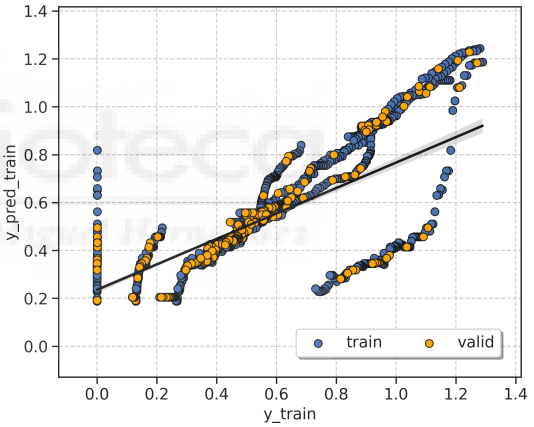
**PFI (only most important descriptors):**

Model = RF · Train:Validation = 90:10

Points(train+valid.):descriptors = 2232:3

Score = 4 / 10

WEAK




Train : $R^2$ = 0.54, MAE = 0.11, RMSE = 0.2

Valid. : $R^2$ = 0.6, MAE = 0.099, RMSE = 0.19

**Severe warnings**

No severe warnings detected

**Moderate warnings**

Imprecise predictions (Section B.3b)

Uneven y distribution (Section C)

**Overall assessment**

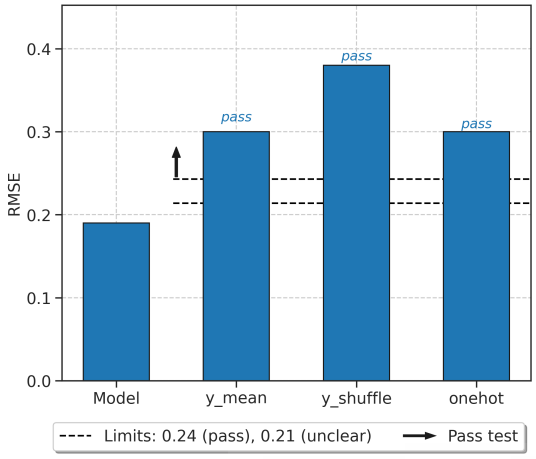
The model is unreliable

## Section B. Advanced Score Analysis

*This section explains each component that comprises the ROBERT score.*

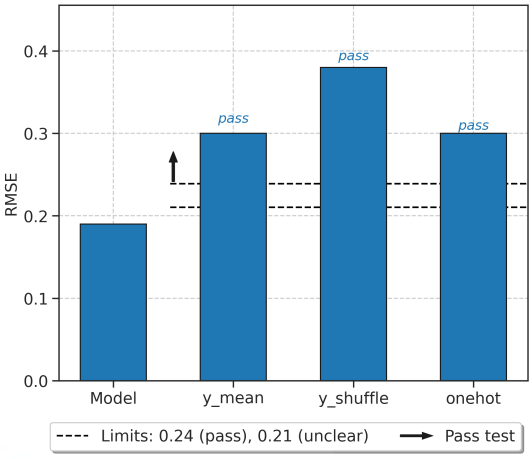
**1. Model vs "flawed" models** (3 / 3)

The model predicts right for the right reasons.
Pass: +1, Unclear: 0, Fail: -1. *Details here.*




**1. Model vs "flawed" models** (3 / 3)

The model predicts right for the right reasons.
Pass: +1, Unclear: 0, Fail: -1. *Details here.*




**2. Predictive ability of the model** (0 / 2)

Low predictive ability with $R^2$ (valid.) = 0.59.
$R^2$ 0.70-0.85: +1, $R^2$ >0.85: +2.

**2. Predictive ability of the model** (0 / 2)

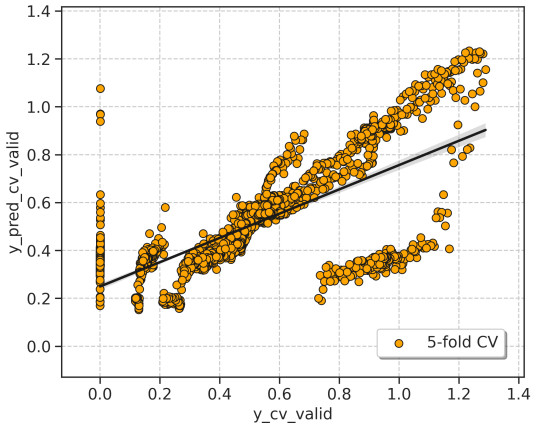
Low predictive ability with $R^2$ (valid.) = 0.6.
$R^2$ 0.70-0.85: +1, $R^2$ >0.85: +2.

**3. Cross-validation (5-fold CV) of the model**

Overfitting analysis on the model with 3a and 3b:

3a. CV predictions train + valid. (0 / 2)

Low predictive ability with $R^2$ (5-fold CV) = 0.51.
$R^2$ 0.70-0.85: +1, $R^2$ >0.85: +2.

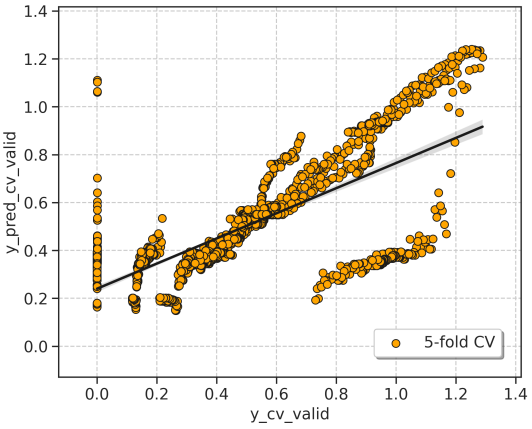



**3. Cross-validation (5-fold CV) of the model**

Overfitting analysis on the model with 3a and 3b:

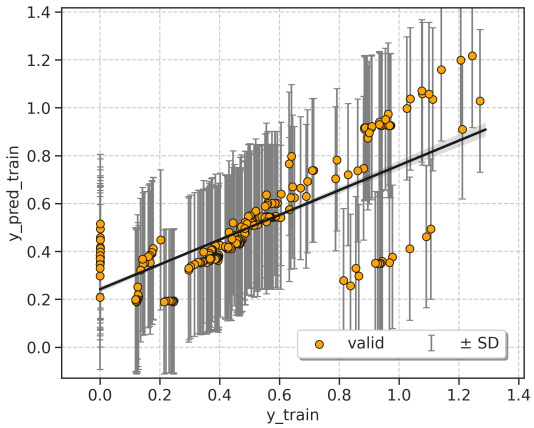
3a. CV predictions train + valid. (0 / 2)

Low predictive ability with $R^2$ (5-fold CV) = 0.51.
$R^2$ 0.70-0.85: +1, $R^2$ >0.85: +2.



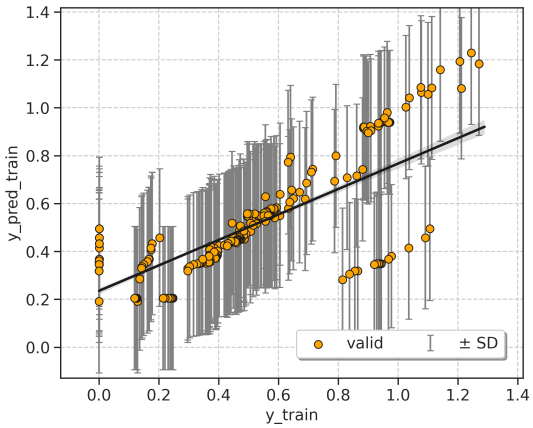
3b. Avg. standard deviation (SD) (0 / 2 )

High variation, 4*SD (valid.) = 1.2 (93% y-range).
4*SD 25-50% y-range: +1, 4*SD < 25% y-range: +2.
*Details here.*




3b. Avg. standard deviation (SD) (0 / 2 )

High variation, 4*SD (valid.) = 1.2 (93% y-range).
4*SD 25-50% y-range: +1, 4*SD < 25% y-range: +2.
*Details here.*




**4. Points(train+valid.):descriptors** (1 / 1 )

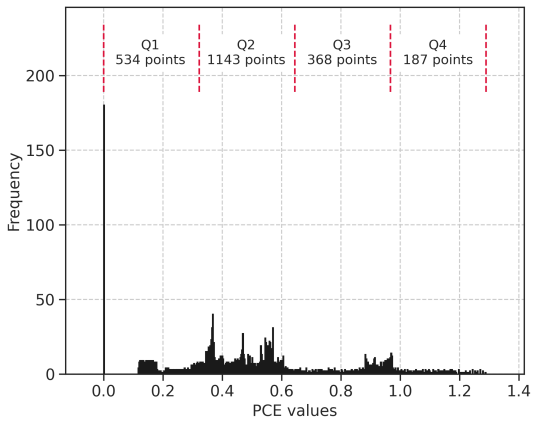
Decent number of descps. (ratio 2232:6).
5 or more points per descriptor: +1.

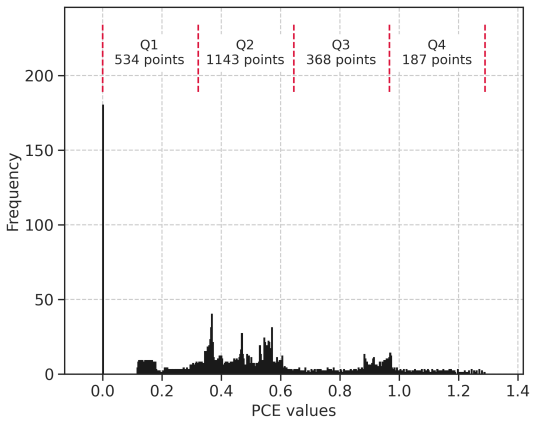
**4. Points(train+valid.):descriptors** (1 / 1 )

Decent number of descps. (ratio 2232:3).
5 or more points per descriptor: +1.

## Section C. Distribution of y Values

*This section shows the distribution of y values within the training and validation sets.*




**y distribution analysis**

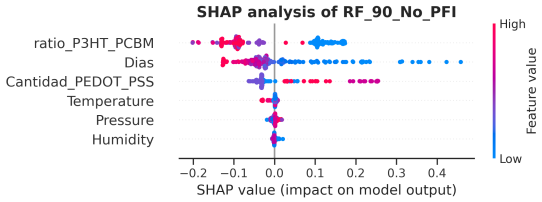
x WARNING! Your data is not uniform (Q4 has 187 points while Q2 has 1143)

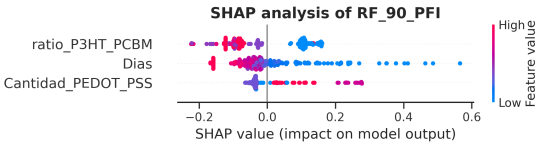
**y distribution analysis**

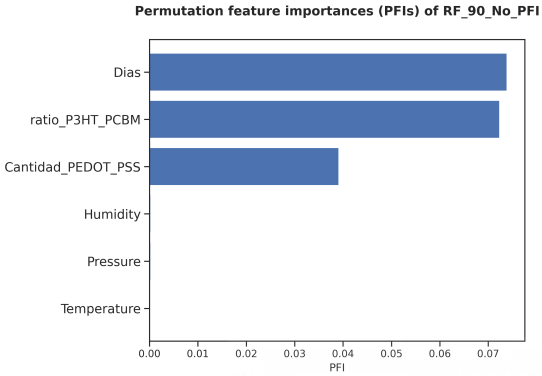
x WARNING! Your data is not uniform (Q4 has 187 points while Q2 has 1143)

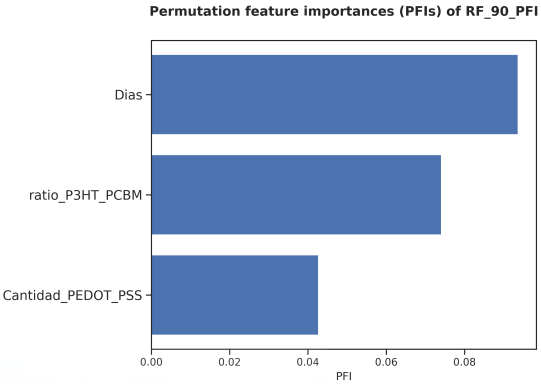
## Section D. Feature Importances

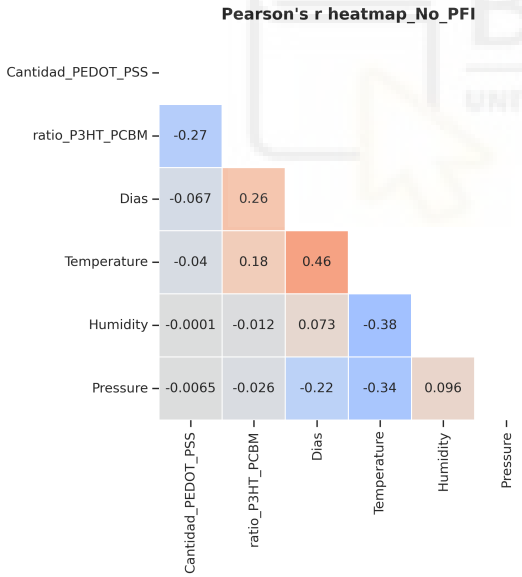
*This section presents feature importances measured using the validation set.*

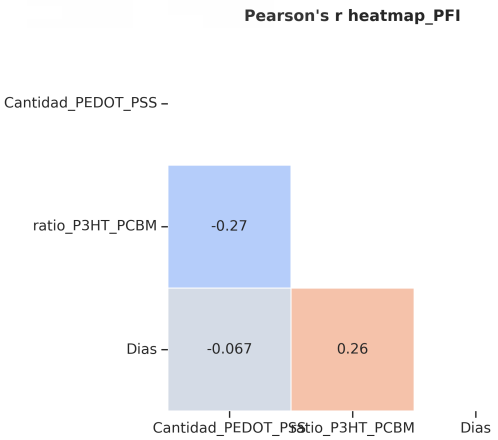



### Correlation analysis

o Correlations between variables are acceptable

### Correlation analysis

o Correlations between variables are acceptable

## Section E. Outlier Analysis

*This section detects outliers using the standard deviation (SD) of errors from the training set.*

**No PFI (standard descriptor filter):**

Outliers (max. 10 shown)

Train: 170 outliers out of 2008 datapoints (8.5%)

- 750 (2.4 SDs)
- 751 (3.0 SDs)
- 756 (2.5 SDs)
- 765 (3.0 SDs)
- 767 (2.4 SDs)
- 773 (2.2 SDs)
- 783 (2.2 SDs)
- 790 (2.7 SDs)
- 791 (2.8 SDs)
- 796 (2.6 SDs)

Validation: 17 outliers out of 224 datapoints (7.6%)

- 806 (2.9 SDs)
- 823 (3.0 SDs)
- 888 (3.0 SDs)
- 1009 (2.8 SDs)
- 1066 (2.9 SDs)
- 1256 (2.8 SDs)
- 1285 (2.8 SDs)
- 1298 (2.8 SDs)
- 1325 (2.7 SDs)
- 1347 (2.7 SDs)

**PFI (only most important descriptors):**

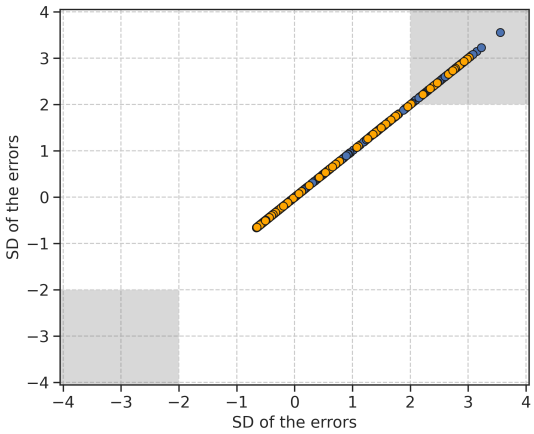
Outliers (max. 10 shown)

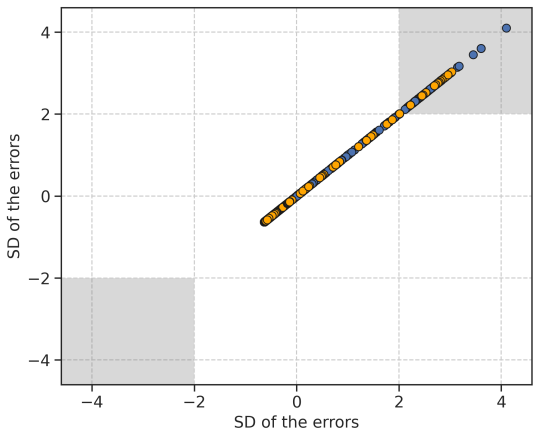
Train: 174 outliers out of 2008 datapoints (8.7%)

- 767 (2.3 SDs)
- 773 (2.2 SDs)
- 777 (2.3 SDs)
- 783 (2.3 SDs)
- 790 (2.6 SDs)
- 791 (2.9 SDs)
- 796 (2.9 SDs)
- 802 (3.0 SDs)
- 811 (2.9 SDs)
- 818 (3.0 SDs)

Validation: 18 outliers out of 224 datapoints (8.0%)

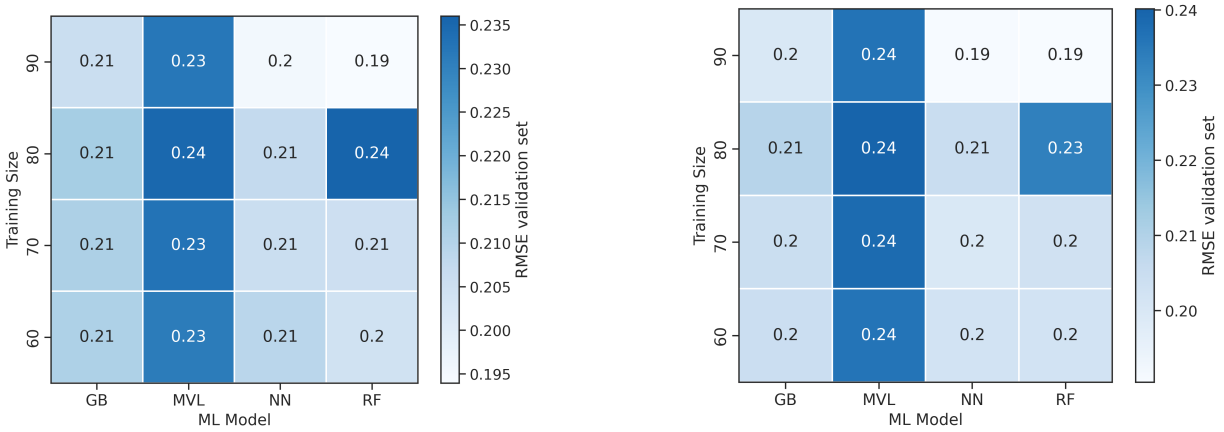
- 806 (2.9 SDs)
- 823 (3.0 SDs)
- 888 (3.0 SDs)
- 1009 (2.8 SDs)
- 1066 (2.8 SDs)
- 1256 (2.8 SDs)
- 1285 (2.8 SDs)
- 1298 (2.8 SDs)
- 1325 (2.7 SDs)
- 1347 (2.7 SDs)



### Section F. Model Screening

*This section compares different combinations of hyperoptimized algorithms and partition sizes.*




### Section G. Reproducibility

*This section provides all the instructions to reproduce the results presented.*

**1. Download these files *(the authors should have uploaded the files as supporting information!)*:**

- CSV database (Datos_final_1d_robert.csv)
- External test set (Pred_dias.csv)

**2. Install and adjust the versions of the following Python modules:**

- Install ROBERT and its dependencies: conda install -y -c conda-forge robert
- Adjust ROBERT version: pip install robert==1.2.1
- Install scikit-learn-intelex: pip install scikit-learn-intelex==2025.2.0

*(if scikit-learn-intelex is not installed, slightly different results might be obtained)*

**3. Run ROBERT using this command line in the folder with the CSV databases:**

python -m robert --csv_name "Datos_final_1d_robert.csv" --csv_test "Pred_dias.csv " names "index" --y "PCE" --y "PCE" --names "index"

**4. Execution time, Python version and OS:**

Originally run in Python 3.11.11 using Linux #1 SMP PREEMPT_DYNAMIC Thu Jun 27 21:05:47 UTC 2024

Total execution time: 678.17 seconds *(the number of processors should be specified by the user)*

## Section H. Transparency

*This section contains important parameters used in scikit-learn models and ROBERT.*

**1. Parameters of the scikit-learn models (same keywords as used in scikit-learn):**

| **No PFI (standard descriptor filter):** | **PFI (only most important descriptors):** |
|---|---|
| sklearn model: RandomForestRegressor | sklearn model: RandomForestRegressor |
| random_state: 0 | random_state: 0 |
| names: index | names: index |
| n_estimators: 20 | n_estimators: 20 |
| max_depth: 5 | max_depth: 5 |
| max_features: 0.75 | max_features: 0.75 |
| min_samples_split: 2 | min_samples_split: 2 |
| min_samples_leaf: 1 | min_samples_leaf: 1 |
| min_weight_fraction_leaf: 0 | min_weight_fraction_leaf: 0 |
| ccp_alpha: 0 | ccp_alpha: 0 |
| oob_score: False | oob_score: False |
| max_samples: 0.5 | max_samples: 0.5 |

**2. ROBERT options for data split (KN or RND), predict type (REG or CLAS) and hyperopt error (RMSE, etc.):**

| **No PFI (standard descriptor filter):** | **PFI (only most important descriptors):** |
|---|---|
| split: RND | split: RND |
| type: reg | type: reg |
| error_type: rmse | error_type: rmse |

## Section I. Abbreviations

*Reference section for the abbreviations used.*

**ACC:** accuracy
**ADAB:** AdaBoost
**CSV:** comma separated values
**CLAS:** classification
**CV:** cross-validation
**F1 score:** balanced F-score
**GB:** gradient boosting
**GP:** gaussian process
**KN:** k-nearest neighbors
**MAE:** root-mean-square error
**MCC:** Matthew's correl. coefficient
**ML:** machine learning
**MVL:** multivariate lineal models
**NN:** neural network
**PFI:** permutation feature importance
**R2:** coefficient of determination
**REG:** Regression
**RF:** random forest
**RMSE:** root mean square error
**RND:** random
**SHAP:** Shapley additive explanations
**VR:** voting regressor

## Section J. New Predictions

*Predictions of the external test set added with the csv_test option.*

csv_test predictions (sorted, max. 20 shown)

From /PREDICT/csv_test/...No_PFI.csv

| index | PCE_pred ± sd |
|---|---|
| 2394 | 1.22 ± 0.3 |
| 2356 | 1.22 ± 0.3 |
| 2318 | 1.22 ± 0.3 |
| 2280 | 1.22 ± 0.3 |
| 2014 | 1.22 ± 0.3 |
| 1976 | 1.22 ± 0.3 |
| 1938 | 1.22 ± 0.3 |
| 1900 | 1.22 ± 0.3 |
| 3534 | 1.21 ± 0.3 |
| 3496 | 1.21 ± 0.3 |
| ... | ... |
| 1055 | 0.24 ± 0.3 |
| 1017 | 0.24 ± 0.3 |
| 751 | 0.24 ± 0.3 |
| 713 | 0.24 ± 0.3 |
| 675 | 0.24 ± 0.3 |
| 637 | 0.24 ± 0.3 |
| 371 | 0.24 ± 0.3 |
| 333 | 0.24 ± 0.3 |
| 295 | 0.24 ± 0.3 |
| 257 | 0.24 ± 0.3 |

csv_test predictions (sorted, max. 20 shown)

From /PREDICT/csv_test/..._PFI.csv

| index | PCE_pred ± sd |
|---|---|
| 2622 | 1.26 ± 0.3 |
| 2584 | 1.26 ± 0.3 |
| 2546 | 1.26 ± 0.3 |
| 2508 | 1.26 ± 0.3 |
| 2242 | 1.26 ± 0.3 |
| 2204 | 1.26 ± 0.3 |
| 2166 | 1.26 ± 0.3 |
| 2128 | 1.26 ± 0.3 |
| 3762 | 1.25 ± 0.3 |
| 3724 | 1.25 ± 0.3 |
| ... | ... |
| 1055 | 0.19 ± 0.3 |
| 1017 | 0.19 ± 0.3 |
| 751 | 0.19 ± 0.3 |
| 713 | 0.19 ± 0.3 |
| 675 | 0.19 ± 0.3 |
| 637 | 0.19 ± 0.3 |
| 371 | 0.19 ± 0.3 |
| 333 | 0.19 ± 0.3 |
| 295 | 0.19 ± 0.3 |
| 257 | 0.19 ± 0.3 |

**Miscellaneous**

*General tips to improve the models and instructions to predict new values.*

Some general tips to improve the score

1. Adding meaningful datapoints might help to improve the model. Also, using a uniform population of datapoints across the whole range of y values usually helps to obtain reliable predictions across the whole range. More information about the range of y values used is available in Section C.

2. Adding meaningful descriptors or replacing/deleting the least useful descriptors used might help. Feature importances are gathered in Section D.

How to predict new values with these models?

1. Create a CSV database with the new points, including the necessary descriptors.

2. Place the CSV file in the parent folder (i.e., where the module folders were created)

3. Run the PREDICT module as 'python -m robert --predict --csv_test FILENAME.csv'.

4. The predictions will be shown at the end of the resulting PDF report and will be stored in the last column of two CSV files called MODEL_SIZE_test(_No)_PFI.csv, which are in the PREDICT folder.