IMPLEMENTACIÓN DE UN ALGORITMO DE IMPUTACIÓN DE DATOS DE SÍNTESIS DE CELDAS SOLARES DE PEROVSKITA UTILIZANDO EL CRITERIO DE BAYESIAN LINEAR REGRESSION (BLR)

FRANCY JESSENIA CALDERÓN OSORIO
JOSEPH MATHEO CRUZ CALDERÓN

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FISICOMECÁNICAS
ESCUELA DE INGENIERÍAS
ELÉCTRICA, ELECTRÓNICA Y DE TELECOMUNICACIONES
BUCARAMANGA
2023

IMPLEMENTACIÓN DE UN ALGORITMO DE IMPUTACIÓN DE DATOS DE SÍNTESIS DE CELDAS SOLARES DE PEROVSKITA UTILIZANDO EL CRITERIO DE BAYESIAN LINEAR REGRESSION (BLR)

FRANCY JESSENIA CALDERÓN OSORIO
JOSEPH MATHEO CRUZ CALDERÓN

Trabajo de grado para optar al título de Ingeniero Electrónico

Director
Franklin Alexánder Sepúlveda Sepúlveda
Doctor en Ingeniería Electrónica
Codirectora
Mónica Andrea Botero Londoño
Dra. en Ciencia Física

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FISICOMECÁNICAS
ESCUELA DE INGENIERÍAS
ELÉCTRICA, ELECTRÓNICA Y DE TELECOMUNICACIONES
BUCARAMANGA
2023

AGRADECIMIENTOS

Primero que todo, agradecer a Dios y a toda mi querida familia, especialmente a mi madre por todo su amor, sin su ayuda, apoyo y comprensión no hubiese podido lograrlo, a mi hermana, mis nonos, tíos, mis padrinos, a todos aquellos con quien me crucé en este camino y han contribuido con un granito de arena en este logro, a Cris, mi soporte, mi amigo, mi compañero de vida, gracias; a mis amados compañeros de cuatro patas Kira, Eros, Pitt, Mimi, Venus, Kiko, Duncan, Moana, Scott aunque algunos no me acompañen más, siempre estarán presentes. A mi director y codirectora por la oportunidad brindada y por supuesto a mi compañero de proyecto. Finalmente pero no menos importante a la UIS, la gloriosa, por abrirme las puertas a esta gran oportunidad.
En memoria de mi abuelo, Luis Calderón.

Francy Jessenia Calderón Osorio

En este momento de culminación de mi proyecto de tesis, quiero expresar mi más sincero agradecimiento a todas las personas que contribuyeron y apoyaron de diversas formas para hacer posible la realización de este trabajo. Sus valiosas aportaciones, orientación y aliento han sido fundamentales en cada etapa de este arduo proceso.
En primer lugar, me gustaría expresar mi gratitud a mi director de tesis, Franklin Alexander Sepúlveda, y codirectora Mónica Andrea Botero, por su dedicación, paciencia y guía a lo largo de todo este trayecto. Sus conocimientos, experiencia y comentarios constructivos han sido invaluables para enriquecer mi investigación y asegurar la calidad de este trabajo. No puedo dejar de mencionar el apoyo incondicional de mi familia y amigos durante todo este proceso. Sus palabras de aliento, comprensión y paciencia fueron un gran respaldo emocional en los momentos más desafiantes. Gracias por creer en mí y por motivarme a seguir adelante.

Joseph Matheo Cruz Calderón

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

pág.

RESUMEN

TÍTULO: IMPLEMENTACIÓN DE UN ALGORITMO DE IMPUTACIÓN DE DATOS DE SÍNTESIS DE CELDAS SOLARES DE PEROVSKITA UTILIZANDO EL CRITERIO DE BAYESIAN LINEAR REGRESSION (BLR) [*]

AUTORES: FRANCY JESSENIA CALDERÓN OSORIO, JOSEPH MATHEO CRUZ CALDERÓN [**]

PALABRAS CLAVE: CELDAS SOLARES, PEROVSKITA, REGRESIÓN LINEAL BAYESIANA, ESTIMACIÓN DE PROBABILIDAD, APRENDIZAJE AUTOMÁTICO.

DESCRIPCIÓN:

El cambio climático ha impulsado el uso de energías renovables, como la solar, para reducir las emisiones de gases de efecto invernadero y la dependencia de los combustibles fósiles. Las celdas solares de perovskita tienen altas expectativas, debido a la mejora en pocos años de su eficiencia de conversión de potencia, PCE, que para 2009 fue de $3.8\,\%$, y para el presente año, de $26.1\,\%$. A pesar del progreso que han logrado, el reporte de la información incompleta respecto a los datos de síntesis de estas celdas ha resultado en carencias significativas en la información científica disponible. En este trabajo se desarrolla una solución a este problema, la metodología desarrollada implementa un algoritmo de imputación de datos basado en el modelo de regresión lineal Bayesiana, realizado en Python, aplicándolo en una base de datos depurada de 63 observaciones. Los resultados demuestran el potencial de combinar nuevas tecnologías como las celdas solares de perovskita con metodologías y estadísticas avanzadas y no convencionales como lo es la estadística bayesiana y generar conocimientos en áreas emergentes. Futuros trabajos podrán usar este algoritmo para imputar y estimar datos de otras fuentes y campos de investigación.

---

[*] Trabajo de grado

[**] Facultad de Ingenierías Físico-Mecánicas. Escuela de Ingenierías Eléctrica, Electrónica y Telecomunicaciones. Director: Franklin Alexander Sepúlveda Sepúlveda, Doctorado en Ingeniería Electrónica . Codirectora: Mónica Andrea Botero Londoño, Dra. en Ciencia Física

ABSTRACT

TITLE: IMPLEMENTATION OF A PEROVSKITE SOLAR CELL SYNTHESIS DATA IMPUTATION ALGORITHM USING THE BAYESIAN LINEAR REGRESSION (BLR) CRITERION [*]

AUTHORS: FRANCY JESSENIA CALDERÓN OSORIO, JOSEPH MATHEO CRUZ CALDERÓN [**]

KEYWORDS: SOLAR CELLS, PEROVSKITE, BAYESIAN LINEAR REGRESSION (BLR), PROBABILITY ESTIMATION, MACHINE LEARNING.

DESCRIPTION:

Climate change has driven the use of renewable energies, such as solar energy, to reduce greenhouse gas emissions and dependence on fossil fuels. Perovskite solar cells have high expectations, due to the improvement in a few years of their power conversion efficiency, PCE, which for 2009 was $3.8\,\%$, and for the current year, $26.1\,\%$. Despite the progress they have achieved, the incomplete reporting of information regarding the synthesis data of these cells has resulted in significant gaps in the available scientific information. In this work a solution to this problem is developed, the developed methodology implements a data imputation algorithm based on the Bayesian linear regression model, performed in Python, applying it on a cleaned database of 63 observations. The results demonstrate the potential to combine new technologies such as perovskite solar cells with advanced and unconventional methodologies and statistics such as Bayesian statistics and generate knowledge in emerging areas. Future work could use this algorithm to impute and estimate data from other sources and research fields.

---

[*] Bachelor Thesis

[**] Faculty of Engineering Physicomechanics. School of Electrical, Electronic and Telecommunications Engineering. Director: Franklin Alexander Sepúlveda Sepúlveda, PhD in Electronic Engineering . Co-director: Mónica Andrea Botero Londoño, PhD in Physical Sciences

INTRODUCCIÓN

La preocupación por el cambio climático ha aumentado en los últimos años. Como consecuencia de esto, se han estado buscando mejores fuentes de energía renovable; y en particular, mejores celdas solares fotovoltaicas. Por ello algunos autores resaltan: .[en] el corazón de ese interés hay una clase de materiales conocidos como perovskitas que encuentran su uso en células fotovoltaicas, celdas de combustible, dispositivos de memoria, entre otros". [1]

Las celdas solares de perovskita han estado en investigación exhaustiva los últimos años, centrándose en la eficiencia de estas. Estos resultados experimentales se han reportado en escritos científicos en artículos científicos presentados mediante revistas internacionales, los cuales contienen información muy valiosa desde el punto de vista del aprendizaje automático. Sin embargo, ésta información está dispersa en cientos y miles de artículos científicos. Al respecto, "se han utilizado técnicas de procesamiento del lenguaje natural para implementar algoritmos de extracción de información con el fin de avanzar en la investigación de materiales"[2], aunque para otros materiales diferentes a celdas de perovskita. Otra forma de obtener datos experimentales de artículos consiste en hacerlo de

---

[1] George Stephen Thoppil y Alankar Alankar. “Predicting the formation and stability of oxide perovskites by extracting underlying mechanisms using machine learning”. En: *Computational Materials Science* 211 (2022). DOI: https://doi.org/10.1016/j.commatsci.2022.111506.

[2] Elsa A. Olivetti et al. “Data-driven materials research enabled by natural language processing and information extraction”. En: *Applied Physics Reviews* 7 (2020). DOI: https://doi.org/10.1063/5.0021106.

forma manual, como muestran los siguientes autores, [3], [4] y [5] lo cual deriva en una labor muy costosa.

El aprendizaje automático ha sido una de las herramientas claves para procesar los datos de celdas solares de perovskita. En [1] se menciona el uso de clasificadores de aprendizaje automático para predecir nuevas composiciones estables de perovskita. Independientemente del método de obtención de los datos, se ha encontrado el problema de datos faltantes. Se ha observado una presencia significativa de datos faltantes en estos conjuntos de datos, lo que dificulta obtener modelos con mejores propiedades y desempeño.

Aunque los datos faltantes están presentes en muchas áreas de investigación, se han abordado varias métodos para realizar imputación de datos. Por ejemplo, los autores Rui et al, proponen ün marco para mejorar el popular método de imputación multivariante por ecuaciones encadenadas (MICE)".[6] Igualmente, Lee et al, "plantean una imputación de datos faltantes basada en la red generativa adversarial (GAN) mejorada para la industria de semiconductores llamada Semi-GAN".[7]

En este trabajo de grado se propone la utilización de un método probabilístico, la Regresión Lineal Bayesiana (BLR). Se opta por este método en particular debido a su capacidad

---

[3] Olga Kononova et al. “Text-mined dataset of inorganic materials synthesis recipes”. En: *Sci Data* 6 (2019), pág. 203. DOI: 10.1038/s41597-019-0224-1.

[4] Pranav Shetty y Rampi Ramprasad. “Automated knowledge extraction from polymer literature using natural language processing”. En: *iScience* 24.1 (2021), pág. 101922. DOI: 10.1016/j.isci.2020.101922.

[5] Juan David Arroyave Zapata y Sebastian Camilo Toscano Higuera. *Enriquecimiento automático de datos extraídos mediante el procesamiento de lenguaje natural en la minería de información de celdas solares de perovskita*. Trabajo de Grado para optar al título de Ingeniera Electrónica. Universidad Industrial de Santander. 2022.

[6] Rui Wu et al. “Data Imputation for Multivariate Time Series Sensor Data With Large Gaps of Missing Data”. En: *IEEE Sensors Journal* 22.11 (2022), págs. 671-683. DOI: 10.1109/JSEN.2022.3166643.

[7] Sun-Yong Lee et al. “Semi-GAN: An Improved GAN-Based Missing Data Imputation Method for the Semiconductor Industry”. En: *IEEE Access* 10 (2022), págs. 328-338. DOI: 10.1109/ACCESS.2022.3188871.

para trabajar eficazmente con conjuntos de datos, incluso aquellos de dimensiones reducidas. Es importante destacar que si bien BLR es óptimo para pequeños volúmenes de datos, también puede aplicarse en contextos con volúmenes más grandes, aunque no sea el enfoque principal de esta investigación. En [8], los autores resaltan una característica crucial para la utilización de la Regresión Lineal Bayesiana (BLR), la cual radica en el hecho de que cada configuración subóptima conlleva un alto costo computacional. Debido a esto, resulta inviable la obtención de un extenso conjunto de datos. Por tanto, es imperativo que el método de regresión empleado para modelar dichos datos sea capaz de prevenir los fenómenos de sobreajuste. En este contexto, la Regresión Lineal Bayesiana y el Proceso Gaussiano cumplen con éxito este requisito.

Para llevar a cabo el desarrollo del algoritmo de regresión lineal bayesiana, se parte de una base de datos inicial que contiene 221 registros. Antes de proceder al análisis de los datos, se realiza un proceso de preprocesamiento que implica la limpieza de registros con valores faltantes. Entre todas las características, el tamaño del grano es la que resulta más afectada por este proceso, ya que se encuentra insuficientemente documentada en la literatura científica, lo que conlleva a una mayor cantidad de valores faltantes en comparación con otras características, como se evidencia en la base de datos utilizada [9]. Tras este proceso, se dispone de un conjunto de 63 datos, los cuales se utilizan para la implementación del algoritmo BLR.

---

[8] Pablo Garcia-Aunon y Antonio Barrientos Cruz. “Control optimization of an aerial robotic swarm in a search task and its adaptation to different scenarios”. En: *Journal of Computational Science* 29 (2018), págs. 107-118. DOI: https://doi.org/10.1016/j.jocs.2018.10.004.

[9] Jeisson Velez et al. “Absorber layer thickness as a new feature in statistical learning tools of Perovskite solar cells”. En: *Journal of Applied Research and Technology* In Press (2023).

Estado del Arte

La imputación de datos ha crecido notablemente en los últimos años. Al realizar una búsqueda en ScienceDirect la base de datos de investigación, de la editorial científica Elsevier, muestra que en 2010 hay un total de 1189 artículos publicados en esta base de datos, relacionados con imputación de datos, para el año 2022 fueron 6916 artículos y hasta el momento en 2023 hay publicados 6595, además se muestra que incluso ya hay 21 artículos publicados para 2024 con 21 artículos publicados relacionados a este tema.[10] Este crecimiento se evidencia de forma similar en otras bases de datos como lo son, Web of Science, Scopus, Nature e IEEE.

Existen varios enfoques de imputación de datos. Dos que son tradicionales y fáciles de implementar son: *listwise deletion*, que descarta aquellas observaciones que tengan uno o más valores faltantes; y, *pairwise deletion*, que solo utiliza los datos disponibles para cada par de variables que se están analizando. Otros enfoques corresponden a imputación única e imputación múltiple. El término imputación única proviene del hecho de que estos métodos generan un único valor de sustitución para cada punto de datos que falta. En contraste, se tiene al enfoque de imputación múltiple, que busca generar conjuntos de valores plausibles para los valores imputar.

En cuanto a imputación simple, el método más básico consiste en realizar imputación por media aritmética, el cual adopta la aparentemente atractiva táctica de rellenar los valores que faltan con la media aritmética de los casos disponibles.

De otra parte están los métodos de imputación por regresión, que sustituyen los valores que faltan por predicciones obtenidas a partir de una función o modelo de regresión. La imputación por regresión estocástica es un método de imputación de datos que utiliza

---

[10] ScienceDirect. *Estadísticas de publicación de la base de datos, respecto a celdas solares de perovskita*. [Sitio web]. Consulta realizada el 2 de septiembre 2023. Disponible en. URL: https://www-sciencedirect-com.bibliotecavirtual.uis.edu.co.

ecuaciones de regresión para predecir las variables incompletas a partir de las variables completas. Sin embargo, a diferencia de la imputación por regresión estándar, la imputación por regresión estocástica agrega un término residual normalmente distribuido a cada puntuación predicha.

La imputación *hot-deck* es un "método de imputación de datos que consiste en reemplazar los valores faltantes con los valores de casos similares. Este método es simple y eficaz, pero es importante tener en cuenta que puede introducir sesgo en los datos, especialmente si la selección de casos similares no se realiza de forma cuidadosa".[11]

Dentro de los métodos de imputación se pueden mencionar "K-vecino más cercano (KNN) y otros métodos de regresión del área del aprendizaje automático; pero también se pueden utilizar modelos probabilísticos tales como modelos de mezclas gaussianas (GMM), redes Bayesianas (BN), cadenas de Markov de Monte Carlo (MCMC) y modelos normales multivariados (MN)"[12].

A pesar de las comparaciones con otros métodos, la imputación mediante el método de BLR no se ha utilizado ampliamente. Sin embargo, un estudio reciente presentó un algoritmo para la regresión lineal bayesiana robusta multivariante con datos faltantes. .[El] algoritmo iterativo calcula un posterior aproximado para los parámetros del modelo basado en el método variacional de Bayes (VB), que se basa en la distribución de probabilidad posterior del modelo".[13] Por lo tanto, el estudio presenta un algoritmo prometedor para la imputación de datos faltantes mediante el método de BLR, pero se necesita más investi-

---

[11] Craig K Enders. “Applied Missing Data Analysis”. En: ed. por Todd D Little. New York London: The Guilford Press, 2010, págs. 39-49.

[12] Yaohui Ding y Arun Ross. “A comparison of imputation methods for handling missing scores in biometric fusion”. En: *Pattern Recognition* 45.3 (2012), págs. 919-933. DOI: https://doi.org/10.1016/j.patcog.2011.08.002.

[13] Juha Ala-Luhtala y Robert Piché. “Gaussian Scale Mixture Models for Robust Linear Multivariate Regression with Missing Data”. En: *Communications in Statistics - Simulation and Computation* 45.3 (2016), págs. 791-813. DOI: 10.1080/03610918.2013.875565.

gación para evaluar su eficacia en una variedad de escenarios.

Este modelo en particular se ha aplicado en diversos campos y para diversas aplicaciones. Entre ellos, se ha utilizado para predecir el consumo de agua en el área metropolitana de Montreal, Canadá, durante las próximas tres décadas.[14] También se ha empleado en la predicción del contenido de fallas en módulos de software.[15] En este contexto, un estudio reciente propone ün nuevo modelo de aprendizaje automático basado en la regresión lineal bayesiana, diseñado específicamente para abordar la desafiante relación muestra-variable que suele encontrarse en los estudios de neuroimagen".[16]

Al tener como precedente este panorama en la implementación de la regresión lineal bayesiana, es propio mencionar que esta propuesta es novedosa. Por ello, se espera que la solución propuesta para la imputación de datos faltantes en los datos de síntesis de celdas solares de perovskita aporte un nuevo uso a este modelo estadístico, ampliando sus aplicaciones. Luego de conocer algunas de las aplicaciones del modelo BLR, se aborda el tema de la regresión y otros conceptos relacionados, con el objetivo de profundizar en la comprensión de su aplicación, estos se presentan en el marco teórico.

Regresión Lineal Bayesiana BLR

La regresión es un proceso que busca encontrar una relación o mapeo $g(\cdot)$ entre variables aleatorias, e.g. entradas $\mathbf{x}$ y salidas $\mathbf{y}$. Se considera un conjunto de datos de entrada $x_1, ..., x_N$ y un conjunto de datos de salida $y_1, ..., y_N$, reunidos un conjunto de datos $D = (x_1, y_1), ..., (x_N, y_N)$, donde N es la cantidad de

---

[14] N. Rasifaghihi, S.S. Li y F. Haghighat. “Forecast of urban water consumption under the impact of climate change”. En: *Sustainable Cities and Society* 52 (2020), pág. 101848. DOI: https://doi.org/10.1016/j.scs.2019.101848.

[15] Rohit Singh y Santosh Singh Rathore. “Linear and non-linear bayesian regression methods for software fault prediction”. En: *International Journal of System Assurance Engineering and Management* 13 (2022), págs. 1864-1884. DOI: https://doi-org.bibliotecavirtual.uis.edu.co/10.1007/s13198-021-01582-1.

[16] Albert Belenguer-Llorens et al. “A Novel Bayesian Linear Regression Model for the Analysis of Neuroimaging Data”. En: *Applied Sciences* 12.5 (2022), pág. 2571. DOI: 10.3390/app12052571.

observaciones disponibles. Se busca encontrar el mejor θ, donde θ son los parámetros que gobiernan la forma de $g(\cdot)$, tal que se minimice el error de entrenamiento

$$\theta = \arg\min_{\theta \in \mathbb{R}^d} \sum_{n=1}^{N} L(y_n, g_\theta(x_n)) \tag{1}$$

donde $L(.,.)$ es la pérdida entre un par de observaciones verdaderas y_n y la predicción $g(x_n)$. Si $y = g(\cdot) = g(\theta, x)$ se restringe a que sea lineal respecto a los parámetros θ, entonces se llega al caso particular de regresión lineal [17].

En BLR en lugar de tener un único conjunto de coeficientes de regresión, se emplea una distribución de probabilidad para modelar los coeficientes de regresión desconocidos. En contraste con la regresión lineal clásica, la BLR incorpora incertidumbre en los parámetros del modelo y permite la actualización de las creencias a medida que se recopila nueva información.

.[En] análisis bayesiano, en primer lugar los parámetros del modelo no son valores, y en su lugar, son variables a aleatorias y por tanto se acepta un grado de incertidumbre de los mismos que se cuantifica mediante la probabilidad. En principio de tiene o establece un conocimiento (al que llamamos información a priori) acerca de la variables que se estable en términos de probabilidad; luego, los datos observados se utilizan para actualizar esa información o creencias a priori a con el fin de convertirlas en información a posteriori".[18]

En un modelo básico se tiene un conjunto de parámetros de interés θ y una observación $\mathbf{x}$. Se considera la distribución de probabilidad conjunta $p(\theta, \mathbf{x})$, para luego aplicar la definición que relaciona la probabilidad condicional con la probabilidad conjunta con el fin de obtener la expresión Bayesiana dada por,

$$p(\theta \mid \mathbf{x}) = \frac{p(\theta) \cdot p(\mathbf{x} \mid \theta)}{p(\mathbf{x})} \tag{2}$$

---

[17] Stanley H. Chan. “Regression”. En: *Introduction to Probability for Data Science*. Michigan Publishing, 2021, págs. 389-390.

[18] Michael D. Lee y Eric-Jan Wagenmakers. “The Basics of Bayesian Analysis”. En: *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press, 2014, pág. 3.

a $p(\theta)$ se denomina distribución de probabilidad a priori de θ y $p(\mathbf{x} \mid \theta)$ se denomina función de verosimilitud. $p(\mathbf{x})$ corresponde a la distribución marginal que se obtiene al integrar la función de probabilidad conjunta respecto a θ. Esta ecuación suele expresarse como:

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{marginal}} \tag{3}$$

Para algún valor de x se tiene que la distribución a posteriori es proporcional a verosimilitud multiplicada por la distribución a priori debido a que el término correspondiente a la probabilidad marginal no está relacionada con el parámetro θ, por ello se puede expresar que [19]:

$$p(\theta \mid \mathbf{x}) \propto p(\mathbf{x} \mid \theta) \cdot p(\theta) \tag{4}$$

Descripción del modelo En BLR se considera el modelo,

$$p(\theta) = \mathcal{N}(\mathbf{m}_0, \mathbf{S}_0) \tag{5}$$

$$p(y|\boldsymbol{x}, \boldsymbol{\theta}) = \mathcal{N}(y|\phi^\top(x)\boldsymbol{\theta}, \sigma^2) \tag{6}$$

donde 5 representa el priori y 6 representa la verosimilitud. "Donde ahora se coloca una Gaussiana priori $p(\theta) = \mathcal{N}(\mathbf{m}_0, \mathbf{S}_0)$ en θ, lo que convierte al vector de parámetros en una variable aleatoria. El modelo probabilístico completo, es decir, la distribución conjunta de las variables observadas y no observadas, y y θ, respectivamente, es"[19]:

$$p(y, \boldsymbol{\theta}|\boldsymbol{x}) = p(y|\boldsymbol{x}, \boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{7}$$

Predicciones prior En el entorno bayesiano, cuando se realizan predicciones, se toma la distribución de los parámetros y la media de todas las configuraciones plausibles de los parámetros. [19] Más

---

[19] Marc Peter Deisenroth, A. Aldo Faisal y Cheng Soon Ong. *Mathematics for Machine Learning*. This version is free to view and download for personal use only. Not for re-distribution, re-sale, or use in derivative works. `https://mml-book.com`. Cambridge University Press, 2020, págs. 309-321.

concretamente, para hacer predicciones con una entrada x_*, se integra θ y se obtiene

$$p(y_*|\boldsymbol{x_*}) = \int p(y_*|\boldsymbol{x_*}, \boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} = \mathbb{E}_\theta \left[p(y_*|\boldsymbol{x_*}, \boldsymbol{\theta})\right] \tag{8}$$

Los cuales se interpretan como la predicción media de $y_*|x_*, \theta$ para todos los parámetros plausibles θ según la distribución a priori $p(\theta)$.

En 5 y 6 se eligió una prior conjugada Gaussiana sobre θ, por ello para que la distribución predictiva sea Gaussiana también, con la distribución a priori mostrada en 5, se obtiene la distribución predictiva como:

$$p(y_*|\boldsymbol{x_*}) = \mathcal{N}(\boldsymbol{\phi}^\top(\boldsymbol{x_*})\boldsymbol{m}_0, \boldsymbol{\phi}^\top(\boldsymbol{x_*})\boldsymbol{S}_0\boldsymbol{\phi}(\boldsymbol{x_*}) + \sigma^2) \tag{9}$$

Con esto, se puede destacar que (i) la predicción es Gaussiana debido a la conjugación y a la propiedad de marginalización de las Gaussianas, (ii) el ruido Gaussiano es independiente de modo que

$$\mathbb{V}[y_*] = \mathbb{V}_\theta[\boldsymbol{\phi}^\top(\boldsymbol{x_*})\boldsymbol{\theta}] + \mathbb{V}_\epsilon[\epsilon] \tag{10}$$

y (iii) y_* es una transformación lineal de θ y por ello se pueden aplicar las reglas para hallar la media y la covarianza.

En 9, el término $\boldsymbol{\phi}^\top(\boldsymbol{x_*})\boldsymbol{S}_0\boldsymbol{\phi}(\boldsymbol{x_*})$ de la varianza predictiva da cuenta explícitamente de la incertidumbre asociada a los parámetros θ, mientras que σ^2 es la contribución de incertidumbre debida al ruido de medición. Si se presenta el interés de predecir valores libres de ruido, $f(\boldsymbol{x_*}) = \boldsymbol{\phi}^\top(\boldsymbol{x_*})\theta$, en lugar de valores con ruido y_*, se obtiene:

$$p(f(\boldsymbol{x_*})) = \mathcal{N}(\boldsymbol{\phi}^\top(\boldsymbol{x_*})\boldsymbol{m}_0, \boldsymbol{\phi}^\top(\boldsymbol{x_*})\boldsymbol{S}_0\boldsymbol{\phi}(\boldsymbol{x_*}) \tag{11}$$

la cuál solo se diferencia de la ecuación 9, en eliminar el término de la varianza del ruido σ^2, en la varianza predictiva. [19]

Distribución posterior Dado un conjunto de datos de entrada x_n y observaciones correspondientes $y_n \in \mathbb{R}, n = 1, ..., N$, se calcula la posterior sobre los parámetros usando el teorema de Bayes como:

$$p(\boldsymbol{\theta}|\boldsymbol{X,Y}) = \frac{p(Y|X, \boldsymbol{\theta})p(\boldsymbol{\theta})}{P(Y|X)} \tag{12}$$

donde x es el conjunto de datos de entrada y y la colección de objetivos de entrenamiento correspondientes. Además, $p(y|x, \boldsymbol{\theta})$ es la verosimilitud, $p(\theta)$ la distribución prior de los parámetros, y

$$p(Y|X) = \int p(Y|X, \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} = \mathbb{E}_\theta[p(Y|X, \boldsymbol{\theta})] \tag{13}$$

la verosimilitud marginal evidencia, que es independiente de los parámetros θ y asegura que la posterior esté normalizada, es decir, que se integre a 1. Se puede pensar en la verosimilitud marginal como la verosimilitud promediada sobre todas las posibles configuraciones de parámetros en relación con la distribución priori $p(\theta)$.

En [19] se presenta un teorema llamado "parámetro posterior", donde se define:

$$p(\boldsymbol{\theta}|X, Y) = \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{m}_N, \boldsymbol{S}_N) \tag{14}$$

$$\boldsymbol{S}_N = (\boldsymbol{S}_0^{-1} + \sigma^{-2}\boldsymbol{\phi}^\top \boldsymbol{\phi})^{-1} \tag{15}$$

$$m_N = S_N(S_0^{-1} m_0 + \sigma^{-2}\boldsymbol{\phi}^\top \boldsymbol{y}) \tag{16}$$

donde N indica el tamaño del conjunto de entrenamiento.

Predicciones posterior En 8, calculamos la distribución predictiva de y_* en una entrada de prueba usando el parámetro prior $p(\theta)$. En principio, predecir con la distribución posterior de los parámetros $p(\theta|x, y)$ no es fundamentalmente diferente, dado que en nuestro modelo conjugado, la distribución prior y la posterior son ambas Gaussianas (con parámetros diferentes). Por lo tanto, siguiendo el mismo razonamiento empleado en las "*Predicciones prior*", obtenemos la distribución predictiva posterior:

$$\begin{aligned} p(y_*|X, Y, x_*) &= \int p(y_*|x_*, \boldsymbol{\theta}) p(\boldsymbol{\theta}|X, Y) d\boldsymbol{\theta} \\ &= \int \mathcal{N}(y_*|\boldsymbol{\phi}^\top(x_*)\theta, \sigma^2 \mathcal{N}(\boldsymbol{m}_N, \boldsymbol{S}_N) d\boldsymbol{\theta} \\ &= \mathcal{N}(y_*|\boldsymbol{\phi}^\top(x_*)\boldsymbol{m}_N, \boldsymbol{\phi}^\top(\boldsymbol{x_*})\boldsymbol{S}_N\boldsymbol{\phi}(\boldsymbol{x_*}) + \sigma^2) \end{aligned} \tag{17}$$

El término $\boldsymbol{\phi}^{\top}(\boldsymbol{x_*})\boldsymbol{S}_N\boldsymbol{\phi}(\boldsymbol{x_*})$ refleja la incertidumbre posterior asociada con los parámetros θ. Teniendo en cuenta que $\boldsymbol{S}_N$ depende de las entradas de entrenamiento a través de $\boldsymbol{\phi}$; como se evidencia en ecuación 15. La media predictiva $\boldsymbol{\phi}^{\top}(x_*)\boldsymbol{m}_N$ coincide con las predicciones realizadas con la estimación *Maximum a posteriori* (MAP) θ_{MAP} [19]

Celdas solares de Perovskita

Las células solares de Perovskita, también conocidas como células solares de Perovskita (PSC), son un tipo de célula solar de tercera generación que utiliza un compuesto con estructura tipo perovskita como capa absorbente. Estas células solares están hechas de materiales artificiales llamados perovskitas, que son compuestos químicos con una estructura similar a la del mineral original.

Al mencionar la perovskita, "no cabe duda de que los compuestos con estructura similar a la perovskita vuelven a estar en el punto de mira, sobre todo por sus posibles aplicaciones en células solares".[20] .El mineral de perovskita tiene la capacidad de absorber radiación solar en espesores de menos de $1\,[\mu m]$ de material para aprovechar la misma cantidad de luz solar que otras células solares".[21]

"Las celdas solares de perovskita fueron incorporadas por primera vez en 2009 por Miyasaka, la eficiencia de conversión (PCE) fue tan solo del $3.8\,\%$ con muy baja estabilidad debido a la presencia de un electrolito líquido".[21] En 2023, .aunque se ha logrado una impresionante eficiencia de conversión de potencia (PCE) del $26,1\,\%$, aún hay margen

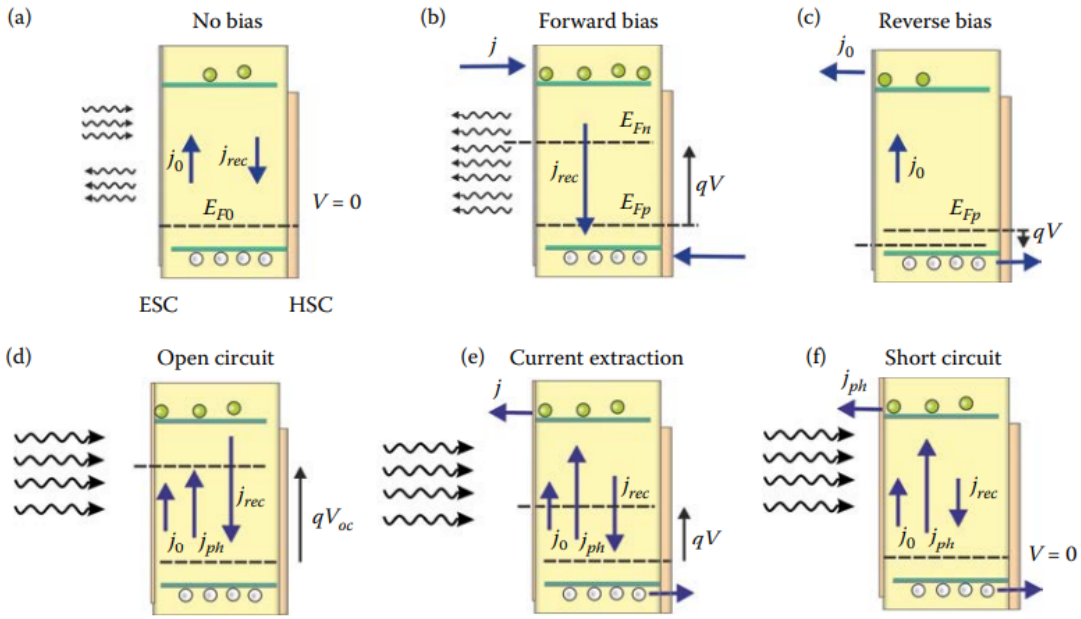
---

[20] Luis Ortega-San-Martin. “Introduction to Perovskites: A Historical Perspective”. En: *Revolution of Perovskite Synthesis, Properties and Applications*. Ed. por Narayanasamy Sabari Arul y Vellalapalayam Devaraj Nithya. Advanced Structured Materials. Springer, 2020. DOI: 10.1007/978-981-15-1267-4_12.

[21] Sarat Kumar Sahoo, Balamurugan Manoharan y Narendiran Sivakumar. “Introduction: Why Perovskite and Perovskite Solar Cells?” En: *PEROVSKITE PHOTOVOLTAICS: Basic to Advanced Concepts and Implementation*. Ed. por Sabu Thomas y Aparna Thankappan. Academic Press, Elsevier, 2018, págs. 1-22. DOI: https://doi.org/10.1016/B978-0-12-812915-9.00001-0.

para nuevas mejoras antes de que estas células alcancen su límite teórico".[22]

Figura 1. Esquema de material semiconductor tipo p.



Fuente: Juan Bisquert. “Basic Operation of Solar Cells”. En: *The Physics of Solar Cells Perovskites, Organics, and Photovoltaic Fundamentals*. Boca Raton, FL: CRC Press, 2018, págs. 139-143.

Parámetros de rendimiento

Los parámetros de rendimiento son indicadores clave que se utilizan para evaluar el rendimiento de las celdas solares de perovskita. Estos parámetros incluyen el coeficiente de rendimiento de la celda solar (PCE), el factor de llenado (FF), la corriente de cortocircuito (J_{sc}) y la tensión de circuito abierto (V_{oc}). El coeficiente de rendimiento de la celda solar (PCE) es la relación entre la potencia eléctrica generada por la celda solar y la potencia solar incidente. El factor de llenado (FF) es la relación entre la potencia máxima que puede generar la celda solar y la potencia generada en cortocircuito. La corriente de cortocircuito (J_{sc}) es la corriente que fluye a través de la celda solar cuando los terminales están conectados en cortocircuito.

---

[22] Zhifang Wu et al. “Passivation strategies for enhancing device performance of perovskite solar cells”. En: *Nano Energy* 115 (2023). DOI: https://doi.org/10.1016/j.nanoen.2023.108731.

La tensión de circuito abierto (V_{oc}) es la tensión que se produce a través de la celda solar cuando los terminales están abiertos.

Estos parámetros son importantes para evaluar el rendimiento de las celdas solares de perovskita. Un alto coeficiente de rendimiento de la celda solar (PCE) indica que la celda solar es capaz de generar una gran cantidad de potencia eléctrica. Un alto factor de llenado (FF) indica que la celda solar es capaz de generar la máxima potencia posible. Una alta corriente de cortocircuito (J_{sc}) indica que la celda solar es capaz de generar una gran cantidad de corriente. Una alta tensión de circuito abierto (V_{oc}) indica que la celda solar es capaz de generar una gran cantidad de tensión. La figura 1 muestra las características eléctricas de una celda solar de perovskita. La figura 1.d muestra la característica de circuito abierto, en la que no se extrae corriente. La figura 1.e muestra que la tensión del circuito es menor que la tensión de circuito abierto ($V < V_oc$). La figura 1.f muestra que los portadores no recombinados, por conservación, producen una densidad de fotocorriente.

Se tiene,

$$j = j_{ph} - j_{rec} + j_0 \tag{18}$$

El equilibrio entre los tres términos del lado derecho de la ecuación 18 da lugar a la corriente que fluye en el circuito externo en términos de electrones que abandonan el estado excitado a través de su contacto selectivo y entran en el estado de masa a través del contacto selectivo del agujero. Ahora, la densidad de corriente-voltaje característica de la celda solar es

$$j = j_{ph} - j_0(e^{\frac{qV}{k_B T}} - 1)(V_{forward} > 0) \tag{19}$$

La figura 2.a muestra la forma de la curva $j - V$, donde la tensión en circuito abierto V_{oc} está dada por,

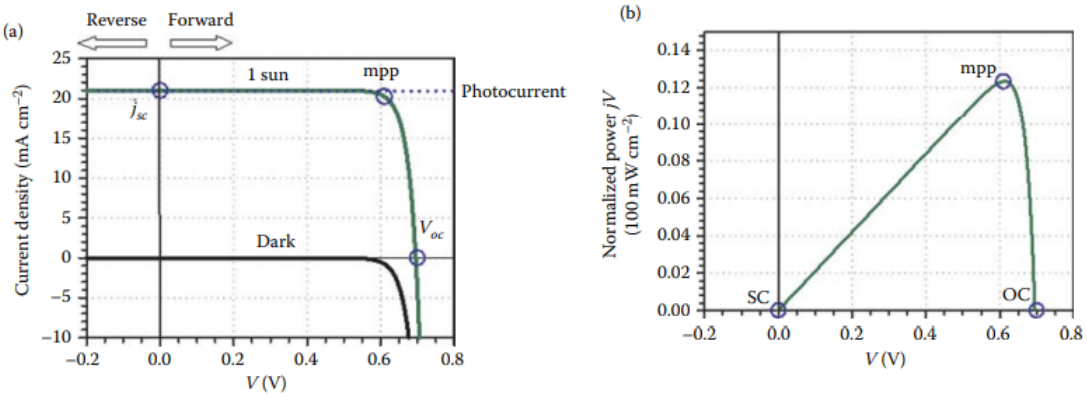
$$V_{oc} = \frac{mk_B T}{q} ln(\frac{j_{ph}}{j_0} + 1) \tag{20}$$

El segundo punto de interés es la condición de cortocircuito, mostrada en la figura 1.e, en este caso, se presenta $j_{rec} = j_0$ por lo tanto, todos los portadores generados en exceso contribuyen a la corriente externa, la fotocorriente de cortocircuito,

$$j_{sc} = j_{ph} \tag{21}$$

La principal aplicación de una celda solar es servir de fuente de alimentación que produce la electricidad a partir de la radiación solar. La característica central para evaluar el rendimiento de la celda es la poten-

Figura 2. Características corriente, voltaje.



Fuente: Juan Bisquert. “Basic Operation of Solar Cells”. En: *The Physics of Solar Cells Perovskites, Organics, and Photovoltaic Fundamentals*. Boca Raton, FL: CRC Press, 2018, pág. 89.

cia eléctrica que se puede extraer del nivel de radiación disponible. La potencia eléctrica en un punto de funcionamiento de tensión de la celda solar P_{el}, es

$$P_{el} = jV \tag{22}$$

En la figura 2.b se muestra la potencia suministrada por la celda solar en función de la tensión. La potencia es nula tanto en condiciones de circuito abierto como de cortocircuito. Entre los casos extremos de baja potencia se encuentra el punto de máxima potencia mpp a cuya tensión V_{mp} debe funcionar la célula solar para la producción de electricidad. La potencia máxima proporcionada por el dispositivo fotovoltaico es

$$P_{el,max} = j_{mp}V_{mp} \tag{23}$$

Para la caracterización de dispositivos convertidores de energía, la principal cifra de mérito es la máxima eficiencia de conversión de la celda solar, es decir, el PCE, que consiste en la potencia eléctrica suministrada en el mpp con respecto a la energía fotónica entrante, se tiene,

$$\eta PCE = \frac{P_{el}}{\Phi_{E,tot}} \tag{24}$$

$$\eta PCE = \frac{j_{mp}V_{mp}}{\Phi_{E,tot}^{source}} \tag{25}$$

El PCE de una celda solar suele indicarse bajo el espectro terrestre estándar simulado $AM1.5G$ que tiene una potencia integrada de $\Phi_E^{AM1.5G} = 1kWm - 2 = 100mWcm - 2$. Este tipo de iluminación suele denominarse "1 sol". Teniendo en cuenta los conceptos anteriores, si se extraen los electrones a alta energía útil (tensión) hay un precio en la disminución del número que se puede extraer. Si este punto estopa cerca de V_{oc}, entonces la tensión y la corriente operativas son mucho mayores que si V_{mpp} se produce a baja tensión cerca de $\frac{V_{oc}}{2}$. El parámetro que sigue esta propiedad es el factor de llenado FF definido como:[23]

$$FF = \frac{j_{mp}V_{mp}}{j_{ph}V_{oc}} \tag{26}$$

Teniendo en cuenta esto, se obtienen las expresiones convenientes,

$$P_{el,max} = j_{ph} \times FF \times V_{oc} \tag{27}$$

$$\eta PCE = \frac{^{ph} \times FF \times V_{oc}}{\Phi_{E,tot}^{source}} \tag{28}$$

En conclusión, los parámetros PCE, FF, J_{sc} y V_{oc} son los más importantes para evaluar la eficiencia de las celdas solares de perovskita. Los valores más altos de estos parámetros indican una mayor eficiencia de las celdas. Lograr valores altos de estos parámetros es esencial para mejorar la eficiencia de las celdas solares de perovskita, lo que las hace más competitivas con otras tecnologías de celdas solares.

Con el propósito de encaminar la implementación del algoritmo planteado, se establecen los siguientes objetivos, los cuales nos orientarán para el cumplimiento exitoso del presente trabajo de grado.

---

[23] Juan Bisquert. "Basic Operation of Solar Cells". En: *The Physics of Solar Cells Perovskites, Organics, and Photovoltaic Fundamentals*. Boca Raton, FL: CRC Press, 2018, págs. 139-143.

Objetivo general

- Desarrollar un algoritmo de imputación de datos de síntesis de materiales de celdas solares de perovskita, resultantes del proceso de extracción de información sobre artículos de investigación, mediante regresión lineal bayesiana.

Objetivos específicos

1. Construir funciones de densidad de probabilidad a priori, a partir de conocimiento de expertos, que permitan estimar las variables aleatorias de interés mediante el criterio bayesiano.

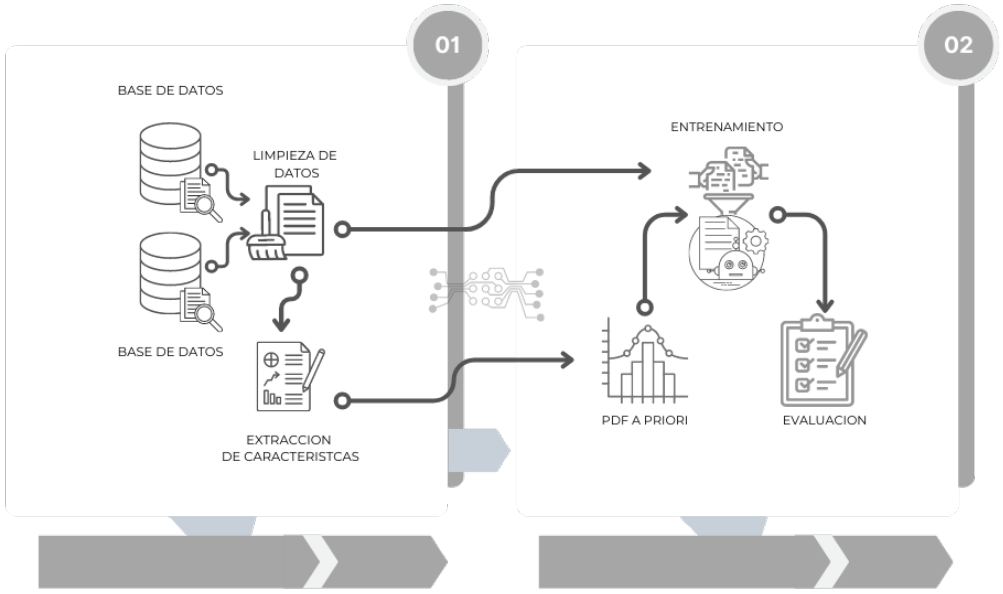
2. Desarrollar un algoritmo de regresión lineal bayesiana que permita la estimación del valor y conocer la incertidumbre asociada a dicha estimación, en datos de desempeño y variables físicas de celdas solares de perovskita.

3. Comparar el desempeño del algoritmo desarrollado con algoritmos basados en GMMs y Máximo Likelihood en escenarios de pequeños y medianos volúmenes de datos.

Luego de explorar los conceptos claves y la teoría relevante, es importante definir la metodología que dirigirá la implementación de este trabajo de investigación. Por ello en la siguiente sección se detalla como se lleva a cabo el proceso de implementación del algoritmo de imputación de datos.

1. METODOLOGÍA

Figura 3. Flujo general del proceso de imputación de datos faltantes de celdas solares de perovskite.



Representación gráfica del flujo general del tratamiento de datos, la construcción, entrenamiento y evaluacion de un modelo de imputación de datos faltantes usando Regresión Lineal Bayesiana.

El proceso de imputación propuesto en este trabajo de investigación consta de dos etapas generales. En la primera de estas se crea una rutina de preprocesamiento, y extracción de información previa de los datos. Mientras en la segunda etapa se lleva a cabo los pasos necesarios para construir, entrenar y evaluar el modelo generado. En la Figura 3 se muestra una vista general de la investigación y el flujo que se lleva a cabo para realizar

el proceso de imputación de forma automática.

Teniendo en cuenta las etapas que se deben llevar a cabo, y la cantidad de recursos que se emplean para lograrlas satisfactoriamente es necesario usar un equipo de software robusto, por ello se ha empleado el servicio en la nube de Google Colaboratory, una plataforma la cuál permite escribir, ejecutar y compartir código en el lenguaje Python, dado que en este entorno se pueden procesar grandes volúmenes de datos, y su procesamiento lleva menos tiempo, en el procesamiento se emplearon varias librerías tales como: PyMC, Pandas, NumPy, Matplotlib.pyplot, Seaborn, Pickle, Sklearn.metrics, Scipy entre otros.

1.1. Datos

Una de las grandes problemáticas a la hora de construir o generar modelos de predicción, es la construcción de un conjunto o base de datos que otorgue los mejores resultados posibles, además de permitir que el modelo construido a través de esta converja a una solución aceptable, para esto es importante una rutina de preparación y depuración de los datos. En este proyecto se cuentan con dos bases de datos, la primera donde se incluyen los valores de "Grain size, DeltaH y DeltaL", reportada por Vélez et.al [9]; y la segunda es la base de datos complementaria reportada en [24]. La primera cuenta con 221 datos de 16 variables ; y, la segunda cuenta con 42496 datos la cual contaba con una cantidad de 13 variables de interés para este proyecto. El porque de utilizar dos bases de datos, una bastante más grande que la otra, es por el enfoque del problema y el modelo de solución propuesto, ambos apuntando a la imputación de datos faltantes cuando se cuenta con pequeños volúmenes de datos. Por ende, la primera base de datos es aquella que se busca imputar denominada base de datos a imputar, y la segunda sirve

---

[24] T. Jesper Jacobsson et al. "An open-access database and analysis tool for perovskite solar cells based on the FAIR data principles". En: *Nature Energy* 7.1 (2022), págs. 107-115. DOI: 10.1038/s41560-021-00941-3.

como complemento informativo para la construcción de información previa denominada base de datos complementaria. La Tabla 1 muestra como se configuran las variables en entradas y salidas, en el proceso de síntesis de las celdas solares de perovskita.

Tabla 1. Parámetros de síntesis (Entradas) y rendimiento (Salidas) de las celdas solares de Perovskita.

Variable	Tipo
Ion$_A$	Entrada
Ion$_B$	Entrada
Ion$_X$	Entrada
Band Gap	Entrada
DeltaH	Entrada
DeltaL	Entrada
Grain Size	Entrada
Thcikness	Entrada
PCE	Salida
Voc	Salida
Jsc	Salida
FF	Salida

La tabla muestra las variables y su correspondencia en la síntesis de la celda solar de perovskite, las entradas que se representan como parámetros de síntesis y las salidas que representan como parámetros de rendimiento de la síntesis

Esta primera etapa que busca generar una rutina integral para el tratamiento y extracción de información previa de los datos se encuentra dividida en distintos procesos los cuales son complementarios entre si.

Estandarización Los artículos científicos funcionan como base fundamental en la extracción y construcción de bases de datos, dado que estos pueden ser presentados en formatos diferentes, lo cual puede generar inconsistencias en sus valores. Se genera una rutina de acondicionamiento, donde se busca estandarizar las unidades de cada variable como se muestra en la Tabla 2, a excepción de aquellas variables cuya unidad es

adimensional.

Tabla 2. Unidades de las variables.

Variable	Unidades
Ion$_A$	-
Ion$_B$	-
Ion$_X$	-
Band gap	eV
DeltaH	eV
DeltaL	eV
Grain size	mm
Thickness	mm
PCE	$\%$
Voc	V
Jsc	mA/cm^2
FF	$\%$

La tabla muestra como se estandarizan las unidades de los datos presentes por variable.

Dado la existencia de tantas variables dentro del modelo, con la intención de minimizar su complejidad matemática, se generan relaciones para los iones A, B y X, de esta manera reduciendo el tamaño de la base de datos de entrenamiento de 17 a 12 variables, utilizando las mismas relaciones se reduce la base de datos complementaria de 15 a 10 variables.

$$Ion_A = MA - FA - Cs \tag{29}$$

$$Ion_B = Pb - Sn \tag{30}$$

$$Ion_X = I + Cl - Br \tag{31}$$

La codificación original del conjunto de datos esta representado en los compuestos mas utilizados para cada uno iones que conforman el perovskita, es así que se observan características como Metylamonio (MA), Formaldehido (FA) y Cesio (Cs) para el ion A como se muestra en la ecuación 29, en 30 se muestran el Plomo (Pb) y Estaño (Sn) para

el ion B y como último el Bromo (Br), Cloro ((Cl) y Yodo(I) para el ion X en la ecuación 31. Esta codificación aumenta la complejidad final del modelo y por ende ser susceptible a los efectos del overfitting. Por otro lado, en cada ion que compone el perovskita puede estar formada por una combinación de cada uno de los elementos mencionados. Todo lo anterior da pie a proponer una codificación que permita reducir la complejidad del modelo y a su vez contenga la información de todos los compuestos presentes en cada ion.

Depuración A modo de etapa previa resulta importante identificar datos inconsistentes; es decir, datos que no guardan un sentido lógico para la tarea en cuestión. Debido a que el presente trabajo se enfoca en celdas de película simple, aquellas de película doble (que además poseen mayor eficiencia) son descartadas. Se establece como estrategia el limitar los valores a los rangos de las variables de la Tabla (3).

Tabla 3. Limitación datos atípicos.

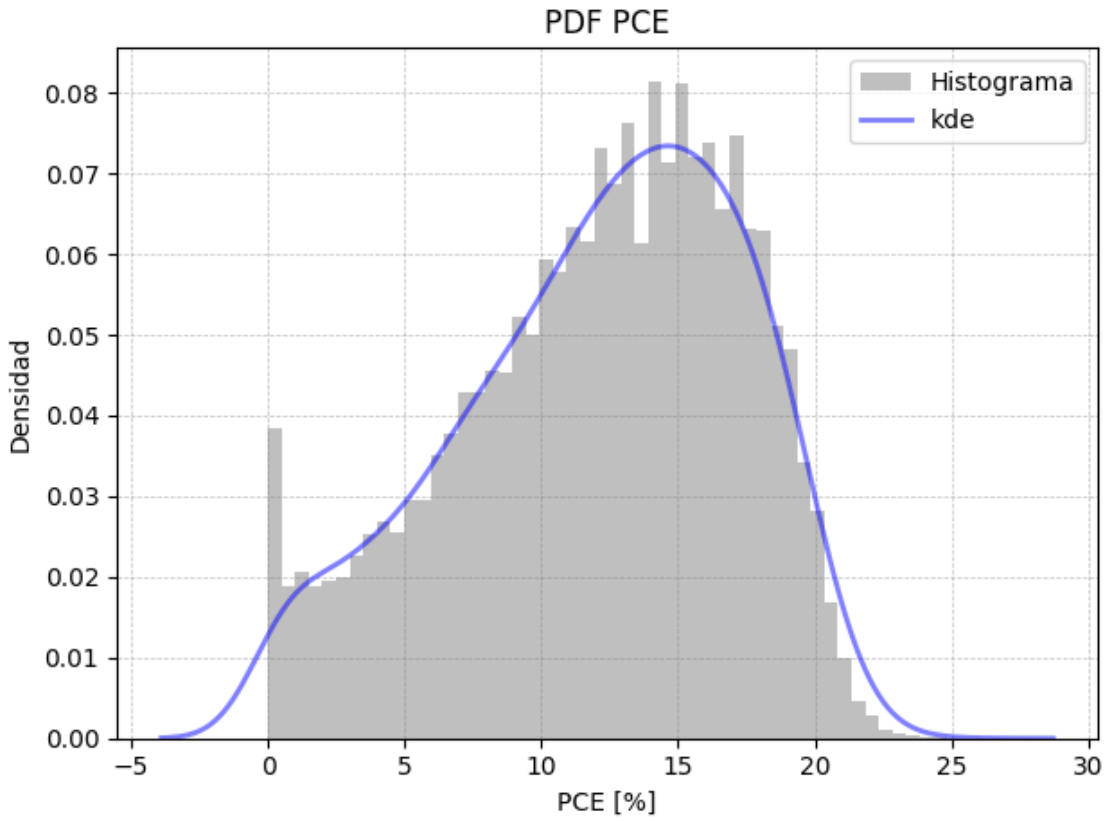
Variable	Limites
Ion_A	-
Ion_B	-
Ion_X	-
Band gap	1 - 2.5
DeltaH	-
DeltaL	-
Grain size	-
Thickness	0 - 1
PCE	0 - 25
Voc	0 - 2
Jsc	0.5 - 30
FF	0 - 100

Limitacion de datos correspondiente a la distribucion tipica de las variables.

Construcción de la distribución previa Este proceso busca la construcción de una función de densidad de probabilidad (PDF) a priori. Para este propósito se utiliza el método “displot” de la librería Seaborn para generar un gráfico de distribución de datos, generando un histograma asociado al conjunto de datos de interés y activando la estimación de densidad de kernel (KDE), como se muestra en la Figura 4.

En la generación del KDE, se utiliza una función de kernel Gaussiano para crear una

Figura 4. Función de probabilidad asociada al histograma de PCE.



Representacion grafica de la construccion de la PDF a priori extraida del histograma para la variable PCE.

función de probabilidad suave a partir de un conjunto de datos discretos, utilizando la función “get_lines” se obtienes los datos de la línea del KDE, y la función “get_data” para

obtener los valores de la grilla del KDE, de esta manera se extraen las coordenadas x e y, que sirven como características para la construcción de la PDF a priori. Este proceso se realiza en la base de datos complementaria, que dado la gran densidad de datos que contiene se puede generar PDF a priori confiable. En este proceso se excluyen las variables $Grainsize$, $DeltaH$ y $DeltaL$, dado que estos datos no se encuentran en dicha base de datos complementaria, por lo que se opta por construir una PDF a priori uniforme limitada por el rango de valores definido para cada variable.

Transformación de variables Se aplica transformación a las variables con el fin de mejorar el ajuste del modelo al fenómeno en cuestión. Dentro de las opciones planteadas (logarítmica, exponencial, cuadrática y raíz cuadrática) se utilizan aquellas que ofrecen mayor correlación respecto a la variables de salida del modelo. Utilizando la correlación de Pearson implementada mediante la librería Scipy, medimos la correlación lineal entre dos variables y generamos valores numéricos de correlación para todas la variables, incluyendo sus transformaciones. De esta manera, creando automáticamente un conjunto de datos relacionado con una variable de interés que tenga la correlación más alta posible. Este proceso implica encontrar la transformación óptima de las variables independientes (las entradas del modelo) en función de su correlación con la variable dependiente (la variable a imputar o de salida).

Establecemos un umbral mínimo de correlación absoluta de 0.55 entre las variables dependientes y la variable independiente. Aquellas transformaciones que, aunque tengan una correlación óptima, es decir, igual o por encima del umbral, que generen indeterminaciones, son descartadas. Para llevar acabo esta tarea, hemos desarrollado una función, que dado un conjunto de datos, una variable de interés y un valor umbral mínimo de correlación absoluta, transforma el conjunto de datos de manera automática con el objetivo de lograr que la correlación sea igual o superior al umbral especificado para cada una de las variables respecto a la variable de interés.

En la Tabla 4, presentamos las correlaciones entre la variable PCE y su entorno de va-

riables relacionadas, incluyendo las transformaciones de estas últimas. Estas transformaciones se obtienen a través de la función mencionada anteriormente. En la Tabla 5, mostramos específicamente aquellas transformaciones que superan el umbral mínimo de correlación establecido para la variable PCE. En función de no extender la explicación, este proceso es replicable para el resto de variables de interés.

Además de todo lo anteriormente mencionado es importante explicar que la variable PCE no sufre una transformación, esta se mantiene en su versión original, dado que se busca que la PDF prior anteriormente construida pueda ingresar información previa al modelo, lo cual no seria posible si esta variable sufre alguna transformación.

Tabla 4. Correlaciones del entorno de variables con la variable PCE.

Variable	None	Logaritmo	Exponencial	Cuadratica	Raiz Cuadrada
Ion$_A$	-0.154	NaN	-0.152	-0.199	NaN
Ion$_B$	0.007	NaN	0.007	0.007	0.007
Ion$_X$	0.228	NaN	0,073	-0.207	NaN
Band gap	-0.284	-0.252	-0.331	-0.313	-0.268
DeltaH	-0.464	-0.518	-0.454	-0.435	-0.416
DeltaL	-0.038	0.085	-0.086	-0.177	0.068
Grain size	0.219	0.362	0.165	0.159	0.281
Thickness	0.407	0.594	0.335	0.255	0.513
Voc	0.283	0.287	0.267	0.270	0.286
Jsc	0.842	0.844	0.045	0.795	0.849
FF	0.543	0.529	0.549	0.552	0.537

La tabla muestra la correlación de pearson del entorno de variables y sus transformaciones respecto a la variable de interés PCE.

Segmentación de los datos de entrenamiento y testeo Para realizar el proceso segmentación de los datos de entrenamiento y testeo, se tiene presente la baja densidad de datos con los que se cuenta par la construcción y evaluación del modelo, por lo que es importante implementar una validación cruzada (CV), para esta tarea se utiliza el método

Tabla 5. Transformaciones del entorno de variables con la variable PCE.

Variable	Transformaciones
Ion_A	None
Ion_B	None
Ion_X	None
Band gap	None
DeltaH	None
DeltaL	None
Grain size	None
Thickness	Log
Voc	None
Jsc	Sqrt
FF	None

La tabla muestra las transformación óptima para variable dado el valor de correlación obtenido respecto a PCE.

Kfold de la librería Scikit-learn, que permite dividir el conjunto de datos en K particiones de igual tamaño. Luego el modelo se entrena K veces, donde en cada iteración se utiliza cada una de las K particiones como conjunto de validación y las K-1 particiones restantes se utilizan para entrenamiento del modelo. Para este proyecto se escogió utilizar K = 5, lo que representa un 20 % de los datos, como datos de validación, mientras que el 80 % restante se empleó como datos de entrenamiento. Dado que contamos con un total de 63 datos completos, esto implica que 50 de ellos se utilizaron para el entrenamiento y 13 para las validaciones en cada iteración. Esto nos permitió obtener 5 modelos de imputación distintos.

Normalización Después de realizar las transformaciones y segmentación de los datos, se procede con el proceso de normalización. Este procedimiento se repite para cada uno de los conjuntos de datos generados mediante el método de validación cruzada (fold). Durante la normalización, se emplean la media y la desviación estándar calcula-

das únicamente a partir de los datos de entrenamiento. Esto se hace con el propósito de evitar la inclusión de información de los datos de prueba en los datos de entrenamiento, garantizando así que los resultados de prueba no estén influenciados por sesgos. La normalización se lleva a cabo utilizando la ecuación 32, donde X representa el conjunto de datos de la variable que se desea normalizar, X_{mean} la media del conjunto de datos de entrenamiento y X_{std} la desviación estándar de dicho conjunto.

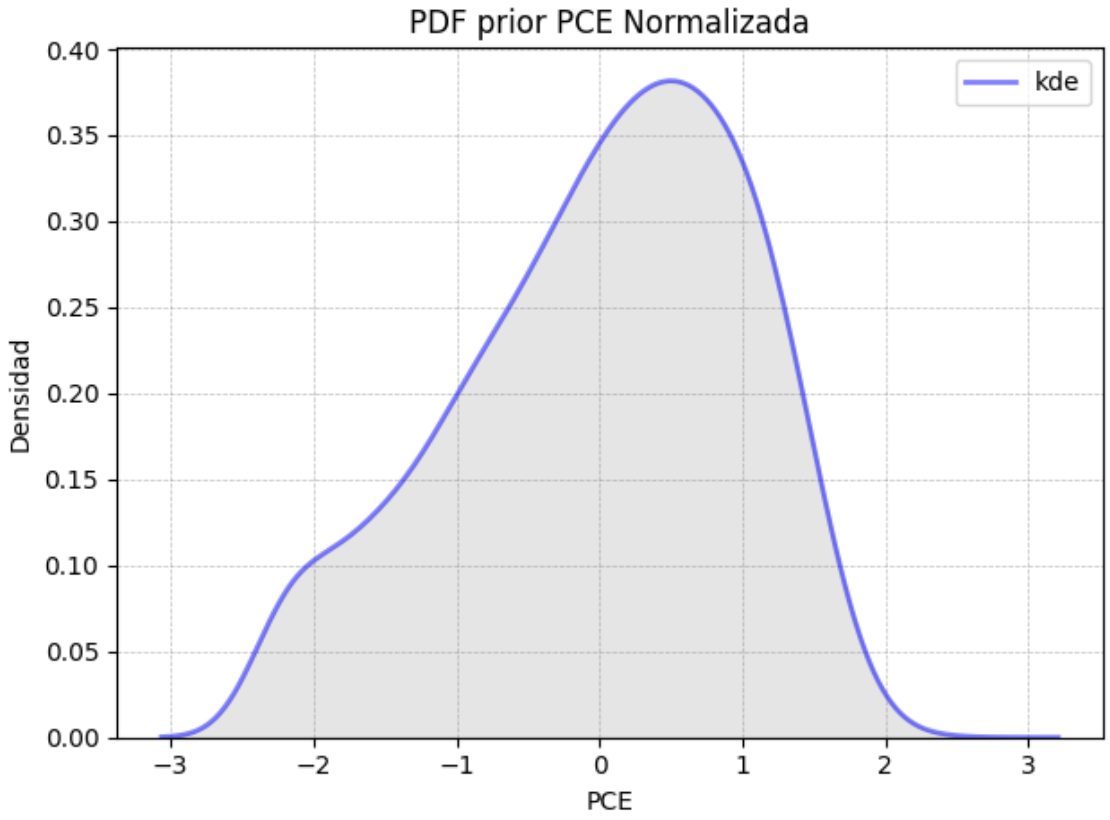
$$X_{normalizada} = \frac{X - X_{mean}}{X_{std}} \tag{32}$$

De manera similar al proceso anteriormente mencionado se recrea una normalización para la PDF priori, con la intención de conservar la información que esta otorga al modelo Figura 5.

1.2. Entrenamiento y evaluación

En esta sección, se explicará el procedimiento que se lleva a cabo, para la construcción del modelo bayesiano, el entrenamiento y la evaluación del modelo de imputación, y como adicional la imputación generada por la estimación donde solo se imputan los parámetros de rendimiento (salidas) de las celdas solares respecto a los parámetros de síntesis (entradas) de las mismas como se muestra en la Tabla 1, al cual llamaremos “imputación por estimación” la cual busca comprobar la existencia de una relación entre los parámetros de síntesis y los parámetros de rendimiento. El proceso de construcción, entrenamiento y evaluación es exactamente el mismo para las dos formas de imputación, solo varían los parámetros a imputar y los datos de entrada que se utilizan, por lo que la explicación para ambos métodos es la misma y en orden de no extender la explicación se hará de forma general y se presentara el proceso mencionado anteriormente para una variable en particular (PCE), el procedimiento es fácilmente extensible para las otras variables.

Figura 5. Función de densidad de probabilidad de PCE normalizada.



La figura muestra la PDF prior construida de PCE normalizada.

Construcción del modelo Para la construcción del modelo es importante definir las PDFs asociadas a las variables del modelo, la PDF asociada al coeficiente de la variable a imputar o de salida, denominada Alpha y las PDFs de los coeficientes de las variables de entrada del modelo denominadas Betas.

Dadas las necesidades presentes de este proyecto, es necesario la construcción de una PDF a priori informativa para Alpha, es decir una que permita al modelo incorporar información previa que se tenga de la variable a imputar, con el fin de mejorar la precisión del modelo. Esta variable es generada en la etapa de “construcción de distribución previa” de la sección anterior del capítulo. Esta información se consigna en el modelo a través

de la función “Interpolated” de la librería PyMC, la cual permite construir en el contexto del modelo la PDF a priori Alpha a partir de la grilla generada por el KDE generado en la etapa anteriormente mencionada Figura 4.

Por otra parte, para la construcción de los priores betas, se opta por utilizar PDFs asociadas a las variables de entrada del modelo no informativos, es decir PDFs que no ingresen información previa de las variables de entrada, se opta por este tipo de prior para las variables de entrada, aun teniendo información previa de estas, dado que se busca permitir que la creencia sobre las variables de entrada se ajuste en función de la información observada. Para esto se utiliza la función “Normal” de la librería PyMC, para generar una distribución normal con media 0 y varianza 10, para todas las variables de entrada.
Luego, se define la varianza del error del modelo utilizando la distribución half-normal, que es una distribución de probabilidad positiva continua que se usa comúnmente para modelar la varianza de un error. Para esto se utiliza la función “HalfNormal” de la librería PyMC que toma como parámetro de entrada una desviación estándar de 1.
La media del modelo se construye mediante la combinación lineal de las variables predictoras de entrada con las correspondientes betas asociados y el prior predictivo Alpha de la variable a imputar. Esta se representa como la estimación de la media del modelo de regresión lineal bayesiano.
Por otra parte, se representa como “likelihood” a la distribución condicional de la variable a imputar dada las variables de entrada. Esta se modela como una distribución normal con media igual a la estimación de la media del modelo. La idea detrás de esto es que, dada una hipótesis sobre los parámetros del modelo, esta variable mida la calidad de esta hipótesis para explicar los datos observados. En otras palabras, “likelihood” es quien permite cuantificar cuán probable es que los datos observados se hayan generado a partir de los parámetros del modelo que se están evaluando. Para esto se utiliza la función “Normal” de la librería PyMC, que toma como media la estimación construida con la re-

gresión lineal, y como desviación estándar la distribución de la varianza del error definida anteriormente.
Finalmente, se agregó una regularización L2 para controlar la complejidad del modelo. Este término se incluyó en el modelo con un “Potential” de PyMC con valor negativo.

Entrenamiento En el BLR, no hay una etapa de entrenamiento como tal, ya que se trata de un modelo probabilístico que se infiere mediante la estimación de la distribución posterior de los parámetros. En este sentido, se podría considerar como una etapa de inferencia o estimación de los parámetros del modelo, a partir de los datos de entrenamiento y los priores especificados. Para esta etapa se utiliza la función “sample” de PyMC para realizar el muestreo de la distribución posterior del modelo.
Esta función implementa un algoritmo llamado NUTS (No-U-Turn Sampler), basado en el método de Monte Carlo Markov Chain (MCMC), para realizar la inferencia. El NUTS es una herramienta eficiente que permite muestrear la distribución posterior de los parámetros del modelo y obtener estimaciones bayesianas precisas.
Su enfoque de exploración del espacio de los parámetros es eficiente, lo que permite trabajar con modelos de múltiples entradas. Utiliza trayectorias simuladas de partículas que siguen una dinámica Hamiltoniana ficticia. Durante la simulación, se combinan la posición actual de los parámetros con muestras aleatorias de momentum. Esta combinación permite ajustar la trayectoria de la simulación y buscar regiones de alta densidad de probabilidad posterior.
Este método emplea el criterio de aceptación y rechazo de Metropolis-Hasting para decidir si acepta o rechaza una propuesta de nuevos valores. Este criterio compara las densidades de probabilidad de los puntos de la trayectoria simuladas con los puntos anteriores de la cadena MCMC. Si la propuesta mejora la densidad de probabilidad, se acepta con una probabilidad determinada. En caso contrario, se rechaza y se mantiene el último valor.
La función “sample” recibe como parámetros de entrada 4 valores, denominados “draws”

Tabla 6. Parametros del sampleo

Parámetro	Valor
Draws	1000
Tune	500
Chains	5
Cores	4

Parámetros de funcionamiento de la función "sample"del proceso de inferencia bayesiana del modelo.

que es el número de muestras que se desean generar de la distribución posterior, “tune” se refiere a la cantidad de iteraciones de ajuste que se realizan antes de comenzar las muestras, lo que permite adaptaciones para mejorar la convergencia del muestreo, “chains” como su nombre lo indica es el número de cadenas de MCMC que se ejecutarán en paralelo, por último “cores”, este parámetro especifica el número de núcleos de CPU que se utilizarán para ejecutar las cadenas en paralelo. Para este proyecto se utilizaron como parámetros los siguientes valores. La Figura 6 consiste en una composición de 12 imágenes que representan las distribuciones de densidad posterior de los coeficientes del modelo. Estas distribuciones se obtienen mediante el proceso de muestreo de la función sample implementando el algoritmo NUTS, tal como se mencionó anteriormente. Cada gráfica muestra múltiples curvas que representan las distribuciones posteriores de los diversos parámetros del modelo. Al observar estas gráficas, es posible identificar cómo el modelo converge hacia una estimación final y comprender la incertidumbre asociada a estos coeficientes. Además, las colas de las distribuciones pueden indicar la posible presencia de valores atípicos o extremos. Estas visualizaciones resultan fundamentales para comprender el funcionamiento y la confiabilidad de la inferencia de los coeficientes realizada por el modelo bayesiano.

Evaluación El proceso de evaluación de un modelo de predicción juega un papel fundamental en la validación y comprensión de sus resultados. En el caso de modelos bayesianos, como el que se presenta en este estudio, es importante considerar métricas de rendimiento que sean adecuadas para capturar las características únicas de este enfoque estadístico.

En este estudio, se ha optado por utilizar el Mean Absolute Percentaje Error (MAPE) como una métrica inicial para comparar el rendimiento del modelo bayesiano propuesto. MAPE Representa el promedio porcentual de la diferencia absoluta entre los valores reales y los valores predichos en relación a los valores reales. Un MAPE cercano a cero indica que las predicciones del modelo son muy precisas en relación a los valores reales, lo cual es deseable. Por otro lado, un MAPE alto indica que las predicciones del modelo tienen una discrepancia significativa con los valores reales, lo cual indica una menor precisión en las estimaciones, la ecuación 33, donde n es el número total de observaciones, Y_i son los valores reales u observados y $\hat{Y}_i$ son los valores predichos por el modelo. Sin embargo, es importante tener en cuenta que esta métrica no captura completamente las ventajas de un modelo bayesiano en términos de estimaciones probabilísticas, ya que este enfoque difiere del modelo frecuentista tradicional. Dado que el modelo bayesiano incorpora la incertidumbre en las estimaciones, es necesario considerar otras métricas que reflejen la distribución posterior y la incertidumbre asociada, dado que el modelo no entrega a su salida una estimación puntual, si no un función de densidad de probabilidad. Además del MAPE, se tendrá en cuenta el Intervalo de Credibilidad de Mayor Densidad (HDI, por sus siglas en inglés) del 90 %, la cual es una métrica esencial para medir la incertidumbre en los modelos estadísticos, especialmente en el contexto de modelos bayesianos. El HDI no se limita a proporcionar una estimación puntual; en su lugar, define un rango de valores que engloba la mayoría de las posibles predicciones del modelo. Por ejemplo, un HDI del 90 % indica que estamos buscando un intervalo que contenga el 90 % de las predicciones del modelo, lo que nos brinda un alto grado de confianza en nuestras estimaciones. Cuan-

to más estrecho sea el HDI, más precisa será la predicción y mayor será la confianza en el modelo. En resumen, el HDI refleja la dispersión de los valores predichos y proporciona una medida sólida de la incertidumbre del modelo [25].

Adicional a lo anterior es importante tener en cuenta la baja densidad de datos con la que se trabaja, lo que implica que el conjunto de datos de testeo corresponda a un total de 13 celdas solares de perovskita, además de utilizar la media de la distribución de densidad de probabilidad generada por el modelo, como estimación frecuentista, y de este modo integrar el MAPE a la evaluación de dicho modelo. También cabe mencionar que no se genera un modelo de imputación para las variables de entrada Ion_A, Ion_B y Ion_X, dado que provienen de una relación que implica que la imputación no entregue información significativa a dichos datos faltantes.

Tabla 7. Valores de MAPE [%] para cada K-fold *"modelo imputación"*

fold	Band gap	DeltaH	DeltaL	Grain size	Thickness	PCE	Voc	Jsc	FF
1	3.32	66.15	30.11	98.54	29.03	2.78	2.17	2.32	1.67
2	0.69	14.36	24.99	76.04	24.72	1.93	1.40	2.03	2.92
3	2.84	69.24	303.74	313.64	99.25	13.80	2.13	17.69	4.25
4	3.23	110.51	214.14	91.53	35.71	1.92	3.41	1.27	4.96
5	21.10	108.09	120.58	131.36	37.02	37.22	5.70	30.49	7.66

Métrica de rendimiento MAPE [%], para cada fold respecto a la variable imputada.

Las Tablas 7 y 9 muestran el valor porcentual del MAPE obtenido para cada uno de los 5 modelos generados por variable en el "modelo de imputación" el "modelo de imputación por estimación", respectivamente. Estas tablas proporcionan una visión general de los resultados frecuentistas del algoritmo de imputación y los modelos resultantes. Por otro lado, las Tablas 8 y 10 resumen los resultados del MAPE para los modelos mencionados

[25] Craig K Enders. “Applied Missing Data Analysis”. En: ed. por Todd D Little. New York London: The Guilford Press, 2010, pág. 165.

Tabla 8. Descripción del MAPE [%] *"modelo imputación"*

	Band gap	DeltaH	DeltaL	Grain size	Thickness	PCE	Voc	Jsc	FF
total	5	5	5	5	5	5	5	5	5
media	6.36	73.67	138.71	142.22	55.94	11.49	2.92	10.76	4.29
std	7.45	35.03	107.68	87.58	32.29	13.53	1.50	11.61	2.02
min	0.69	14.36	24.04	76.04	24.72	1.92	1.40	1.27	1.67
max	21.10	110.51	303.74	313.64	99.25	37.02	5.70	30.49	7.66

Resumen de meta-datos de los parámetros de rendimiento MAPE [%].

Tabla 9. Valores de MAPE [%] para cada K-fold *"modelo imputación por estimación"*.

fold	PCE	Voc	Jsc	FF
1	13.68	7.10	8.29	4.76
2	8.54	3.91	4.22	5.08
3	27.14	6.36	26.87	3.74
4	9.44	4.88	5.07	8.40
5	64.87	17.53	136.67	9.45

Métrica de rendimiento MAPE [%], para cada fold respecto a la variable imputada.

Tabla 10. Descripción del MAPE [%] *"modelo imputación por estimación"*.

	PCE	Voc	Jsc	FF
total	5	5	5	5
media	24.74	7.96	36.22	6.28
std	21.14	4.91	50.89	2.22
min	8.54	3.91	4.22	3.74
max	64.87	17.53	136.67	9.45

Resumen de meta-datos de los parámetros de rendimiento MAPE [%].

anteriormente. La media ofrece una idea del resultado promedio del algoritmo, la desviación estándar indica cuán dispersos están los resultados de los modelos, y el mínimo y el máximo representan el mejor y el peor modelo obtenido, respectivamente.

Tabla 11. Valor Porcentual [%] del intervalo de credibilidad para cada fold *"modelo imputación"*.

fold	Band gap	DeltaH	DeltaL	Grain size	Thickness	PCE	Voc	Jsc	FF
1	50.69	46.80	38.71	40.95	25.64	69.46	55.96	65.13	44.31
2	35.20	37.45	32.26	41.67	40.64	80.84	58.33	69.09	56.83
3	32.11	55.65	42.15	50.93	35.37	79.82	58.32	85.44	44.58
4	43.47	40.05	23.15	32.21	46.25	67.66	47.94	51.23	45.78
5	35.24	37.56	16.18	14.63	19.64	39.26	24.23	34.77	16.68

Representación del valor porcentual del intervalo de credibilidad respecto al rango de valores de la predicción, para cada fold generado por el modelo de cada variable imputada.

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{Y_i - \hat{Y}_i}{Y_i} \right| \times 100\,\% \tag{33}$$

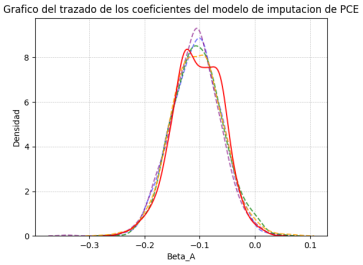
Como medida adicional para generar una representación gráfica del rendimiento del modelo, en la Figura 7 consta de 12 gráficos que representan los resultados de las predicciones del "modelo de imputación", mientras que la Figura 8 se compone de 4 gráficos que muestran los resultados de las predicciones del "modelo de imputación por estimación". Estas figuras se centran en analizar la correlación entre los valores reales y las estimaciones generadas por los modelos mencionados. En un escenario ideal, un desempeño óptimo se representaría mediante una línea recta con pendiente igual a 1. Esto implicaría que el modelo genera estimaciones que se correlacionan perfectamente con los valores reales, manteniéndose cerca de esta línea de referencia ideal. Al observar la proximidad de los puntos de datos a esta línea de referencia, podemos evaluar qué tan bien el modelo se ajusta a los datos reales y cómo se comportan sus predicciones en comparación con los valores observados. En estas es apreciable que las variables de parámetros de rendimiento de síntesis son las que mejor se ajustan a línea recta con pendiente igual a 1 con una baja dispersión para el “modelo imputación”, y de igual manera para el ”modelo imputación por estimación” los resultados se ajustan a la línea recta, pero con una mayor dispersión que el modelo anterior.

Tabla 12. Valor Porcentual [%] del intervalo de credibilidad para cada fold *"modelo imputación por estimación"*

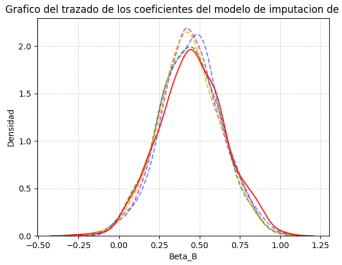
fold	PCE	Voc	Jsc	FF
1	51.70	38.50	56.57	29.17
2	48.35	32.18	53.18	44.03
3	48.82	30.65	53.87	42.04
4	42.14	36.61	45.64	40.40
5	15.72	15.27	38.64	17.64

Representación del valor porcentual del intervalo de credibilidad respecto al rango de valores de la predicción, para cada fold generado por el modelo de cada variable imputada.

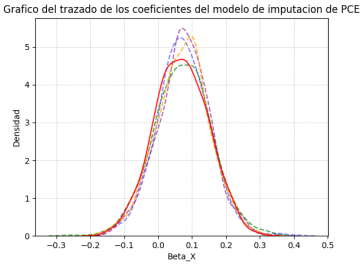
Figura 6. Gráfica del trazado de los coeficientes del modelo de imputación de PCE



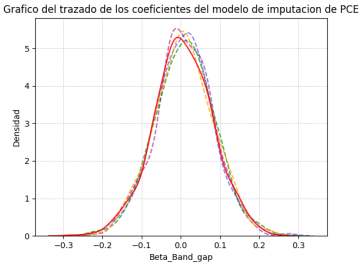
(a) Beta Ion$_A$



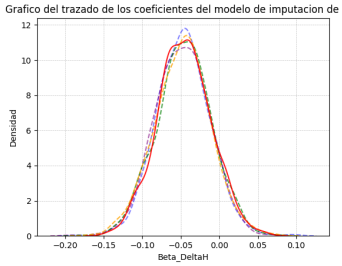
(b) Beta Ion$_B$



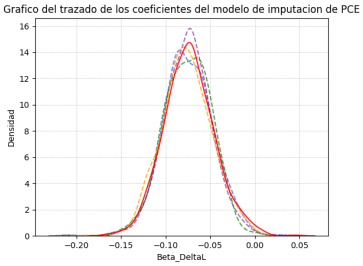
(c) Beta Ion$_X$



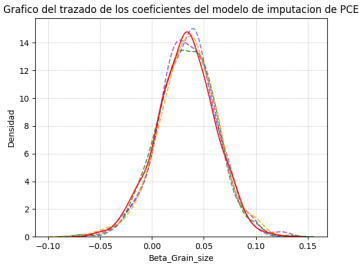
(d) Beta Band gap



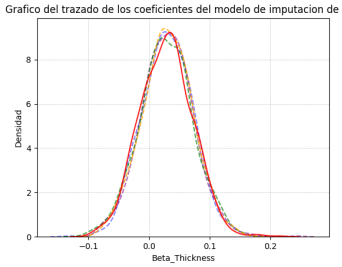
(e) Beta DeltaH



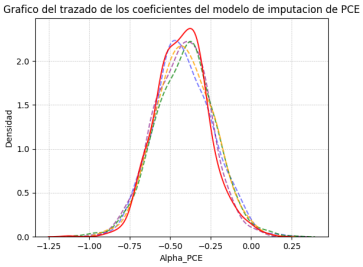
(f) Beta DeltaL



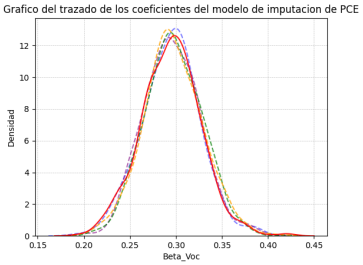
(g) Beta Grain size



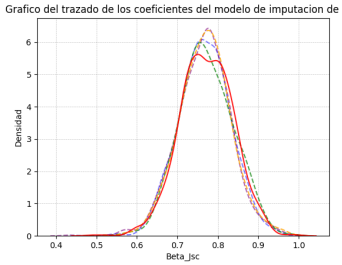
(h) Beta Thickness



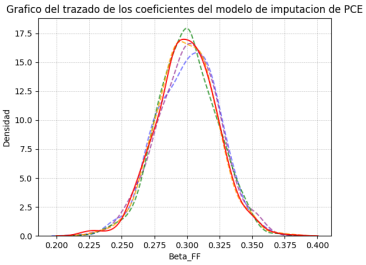
(i) Alpha PCE



(j) Beta Voc



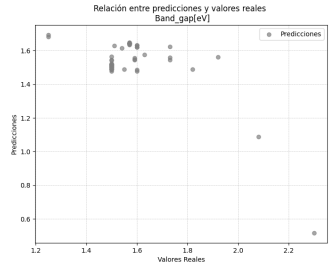
(k) Beta Jsc



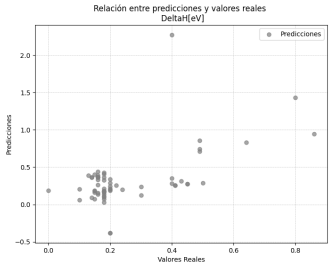
(l) Beta FF

La gráfica muestra el comportamiento del trazado de la PDF posterior de los coeficientes asociados a los parámetros de predicción del modelo, generada en la etapa de sampleo.

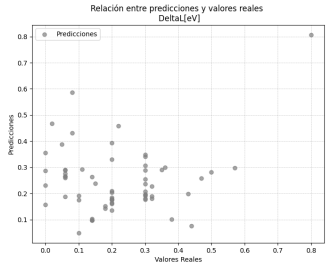
Figura 7. Correlación entre datos reales vs datos estimados



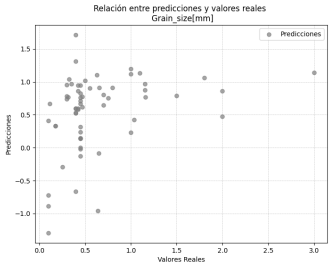
(a) Band gap



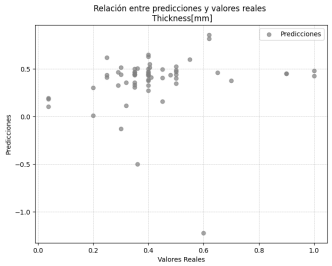
(b) DeltaH



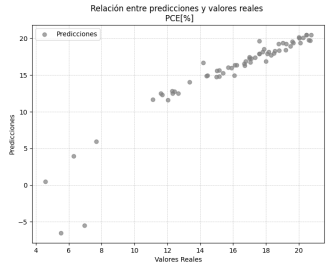
(c) DeltaL



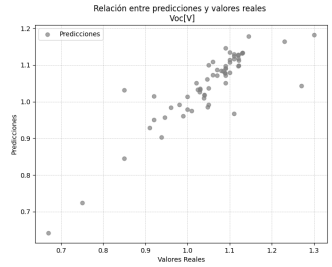
(d) Grain size



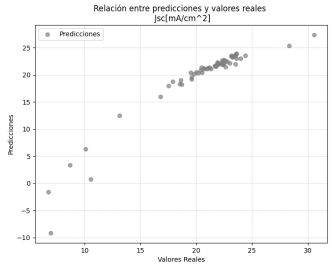
(e) Thickness



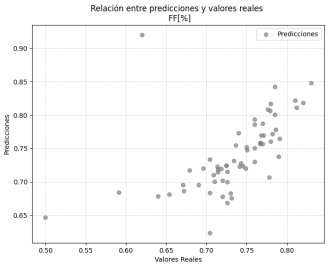
(f) PCE



(g) Voc



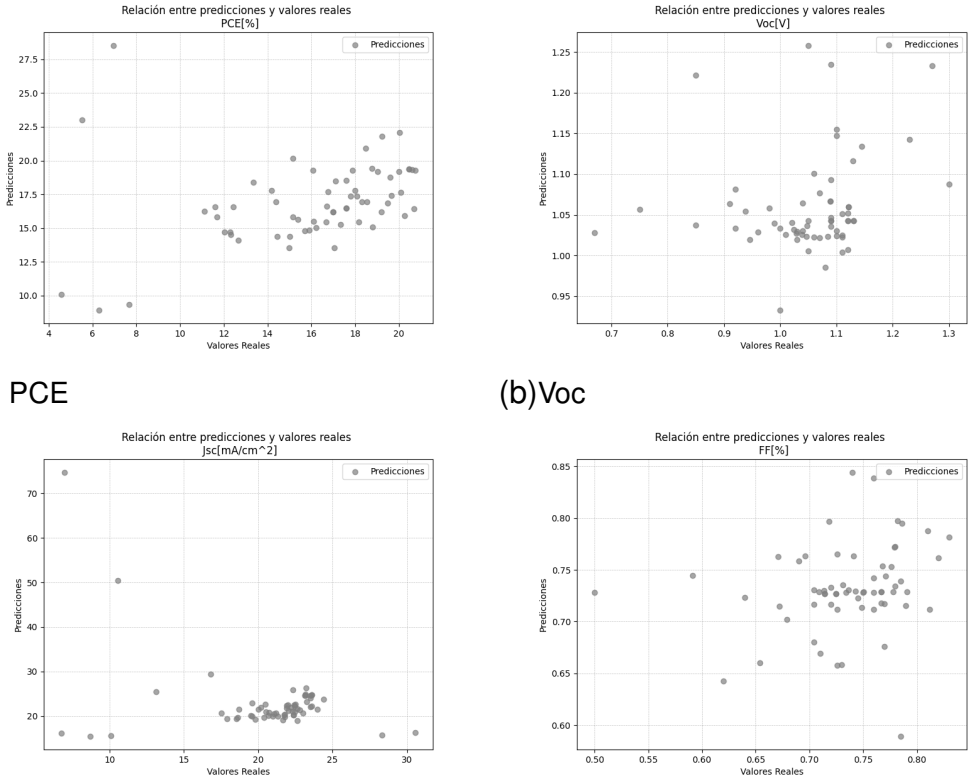
(h) Jsc



(i) FF

La figura expone las gráficas asociadas a las correlaciones de los datos reales frente a las predicciones generadas para cada variable imputada.

Figura 8. Correlación entre datos reales vs datos estimados *“modelo imputación por estimación”*



(a) PCE (b)Voc

(c) Jsc (d) FF

La figura presenta las gráficas asociadas a las correlaciones de los datos reales frente a las predicciones generadas para cada variable imputada.

2. RESULTADOS

En el capítulo anterior, se exploró la aplicabilidad de la investigación y se describió cómo llevarla a cabo en el contexto de un conjunto de datos específico. Ahora, en esta sección, nos adentraremos en un análisis detallado del rendimiento del modelo calculado para cada una de las variables. El conjunto de datos de evaluación se tuvo en cuenta 13 celdas solares de perovskita por cada k-fold generado.

2.1. Comparativa de evaluación

Para la evaluación de los modelos generados se hace la comparativa entre el conjunto de parámetros reales y las predicciones obtenidas. Parte importante de la comprensión de este proyecto es comprender la naturaleza de la estadística bayesiana, para esto es importante comprender que tipo de resultados generan los modelos de imputación. La Figura 9 muestra la función de densidad probabilidad posterior obtenida como predicción resultante del modelo de imputación por estimación, con la intención de no entorpecer la explicación se muestran solo uno los resultados de dicho modelo para cada variable de interés, además las variable de entrada asociadas a dichos resultados se muestran en la Tabla 13.

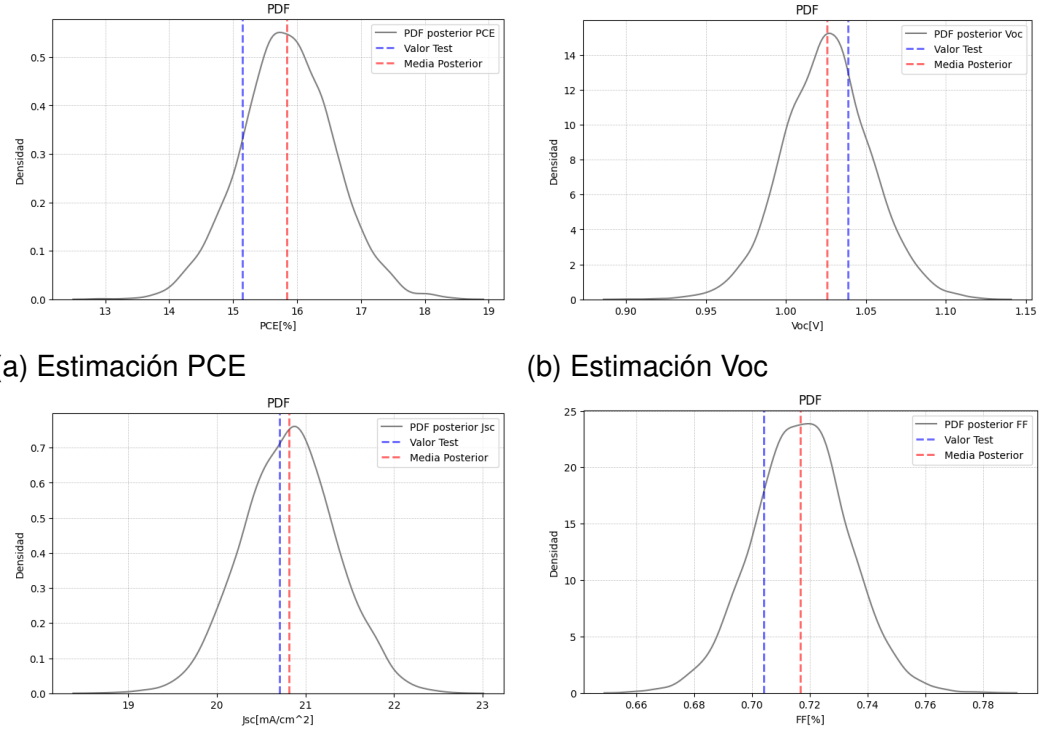
Tabla 13. Variables de entrada para el ejemplo de predicción *"modelo imputación por estimación"*

Ion$_A$	Ion$_B$	Ion$_X$	Band gap	DeltaH	DeltaL	Grain size	Thickness
1.0	1.0	1.0	1.5	0.2	0.0	0.425	0.35

Esta tabla muestra las variables de entrada para el ejemplo de la predicción generado por el *"modelo imputación por estimación"* mostrado en la Figura 9.

Del conjunto de datos las variables que presentan menor rendimiento son los denomina-

Figura 9. Estimación *“modelo imputación por estimación”*



(a) Estimación PCE

(b) Estimación Voc

(c) Estimación Jsc

(d) Estimación FF

El gráfico muestra el resultado de las predicciones o salidas del modelo de imputación por estimación, dados los datos de entrada de la Tabla 13.

dos entradas de la síntesis de la celda solar de perovskita, se genera la hipótesis de que su relación con su entorno de variables es más compleja que una relación lineal, variables tales como DeltaH, DeltaL, Grain Size y Thickness poseen un MAPE medio superior al 50 % incluso algunos superan el 100 % lo que indica que el modelo no logra una precisión aceptable en sus predicciones para estas variables.

Ahora con respecto a los datos denominados métricas de rendimiento de las celdas solares, tales como PCE, Voc, Jsc y FF, adicional a estos también la entrada de síntesis Band gap. En estos la media del estadístico MAPE es inferior al 10 %, en algunos casos es incluso inferior al 5 % tal como lo es para Voc y FF como se puede apreciar en la Tabla 8.

Esto indica que los modelos generados logran adaptarse con buena precisión a la nube de datos, además de una mejor relación entre estos y su entorno de variables. Por otra parte, los modelos generados por medio de la imputación por estimación de las salidas las variables Voc y FF, la media del estadístico MAPE es inferior al 10 %, para PCE es superior al 20 % y Jsc es superior al 35 % lo que muestra que en comparativa con el método inicial el método de imputación por estimación logra resultados inferiores en todas las salidas, y con bastante diferencia para la salida PDE y Jsc, donde los resultados desmejoraron en valor superior al 10 %.
En la Figura 10 y 11 se presenta el comportamiento por cada parámetro de forma individual en la totalidad de los datos, lo que permite generar una comparativa entre los datos reales y los estimados.

2.2. Correlación de datos

En los modelos generados es apreciable que las denominadas entradas de la síntesis del material no logran una correlación lineal deseada entre sus predicciones y los valores reales, por otra parte los parámetros de rendimiento de la síntesis tales como PCE, Voc, Jsc, y FF, logran establecer sus correlaciones al rededor de la linea recta de pendiente 1, lo que respalda los resultados mostrados en la Tabla 7. Por otra parte, en el modelo imputación por estimación es poco apreciable la correlación lineal entre los valores predichos y los vales reales, dado que existe una mayor dispersión de la gráfica al rededor de la recta de pendiente uno, lo que muestra de forma gráfica una desmejora en la incertidumbre y precisión de los resultados obtenidos de los parámetros de rendimiento de la síntesis con el modelo de imputación inicial.

2.3. Análisis de resultados no frecuentistas

Como se mencionó en el capitulo anterior, en este proyecto es importante tener en cuenta no solo los resultados frecuentistas, si no los parámetros bayesianos que nos indican la

incertidumbre asociada en las estimaciones. La Figura12 y 13 muestra una representación gráfica del intervalo de credibilidad asociada a cada variable en cada k-fold generado. Un intervalo estrecho indica una menor incertidumbre asociado a las estimaciones.
Además complementando la información gráfica mostrada la Tabla 11 y 12 que nos muestra una representación porcentual del intervalo de credibilidad en el rango total de valores posibles de la PDF generada para las predicciones, para el modelo de imputación y el modelo imputación por estimación respectivamente.
A simple vista puede ser apreciable que en la gran mayoría de variables del intervalo de credibilidad, no es muy estrecho en la mayoría de casos dicho intervalo se encuentra en un rango entre el 30 y el 50 %, además es superior el número de casos donde este porcentaje sobrepasa el intervalo anterior respecto aquellos que son inferiores a dicho intervalo, este resultado es respaldado por los valores porcentuales dados por la Tabla 11 y 12, esto podría interpretarse como un factor de mediana o alta incertidumbre de los modelos generados, pero es necesario hacer un análisis adicional antes de poder afirmar dicha conclusión.
En la Figura 14 y 15 se muestra la densidad de probabilidad posterior de las predicciones por cada k-fold generado. Estas gráficas nos brindarán una idea del comportamiento de las PDF asociadas a las predicciones, permitiendo una vista más amplia y un mejor entendimiento de los resultados. En ellas es apreciable lo estrecho de dichas distribuciones de probabilidad, indicando que el modelo logra ajustar sus estimaciones a un rango estrecho en comparación con el rango de valores que pueden tomar esas variables. Esto significa que, aunque el intervalo de credibilidad señale una alta incertidumbre, la densidad de probabilidad posterior nos permite re-interpretar dicho señalamiento, ya que, aunque el intervalo de credibilidad sea amplio, lo es en una estrecha densidad de probabilidad posterior de las estimaciones, en comparación con el rango de valores posibles de las variables imputadas

2.4. Comparación de resultados

Durante el planteamiento de nuestro proyecto, se planteo el objetivo de comparar el desempeño del algoritmo de imputación planteado frente a algoritmos basados en Modelo de Mezclas Gaussianas (GMM) y Máximo Likelihood [5] en escenarios de pequeños y medianos volumenes de datos. Sin embargo, en el proceso de ejecución y evaluación de los resultados, nos encontramos con desafíos y consideraciones que hicieron que esta comparación no fuera factible.

El enfoque bayesiano que se empleo siempre fue considerado como el enfoque mas adecuado para escenarios de pocos datos, un aspecto que habíamos anticipado y que había dado forma y razón de ser a nuestro proyecto. Una consideración clave que influyó en nuestra decisión fue la inadecuación de los GMMs y Maximo Likelihood para escenarios de pocos datos. En contraste el modelo de referencia se basó en un conjunto de datos más densamente poblado y con un conjunto limitado de variables de imputación que solo tenia en cuenta cuatro variables correspondientes a los parámetros de rendimiento de síntesis, nuestro enfoque involucró un conjunto de variables más amplio que incorporaba ocho variables adicionales de parámetros del proceso de síntesis. Dividiendo la investigación en dos modelos distintos, uno de nuestros modelos abordaba la imputación de todas las variables del conjuntos de datos partiendo del entorno de variable de dicho conjunto, mientras que otro se enfocaba en predecir los parámetros de rendimiento de síntesis por medio de los parámetros del proceso de síntesis. Esta diferencia hizo que la aplicación de estos modelos fuera impráctica e inadecuada en nuestro contexto especifico.

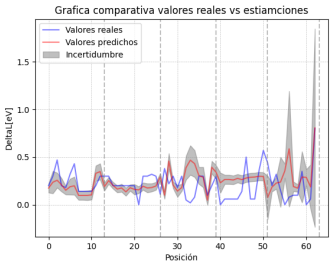
En resumen, la falta de homogeneidad en los conjuntos de datos implementados, la efectividad del enfoque bayesiano en el escenario planteado de pocos datos y la inadecuación de GMMs para dichos escenarios, fueron factores determinantes en nuestra elección de no llevar a cabo la comparación con los algoritmos mencionados. Nuestro enfoque se centro en la aplicación efectiva de técnicas estadísticas bayesianas, que se adaptaron de manera robusta a las características únicas de nuestros datos y proporcionaron resulta-

dos prometedores en nuestro contexto particular.

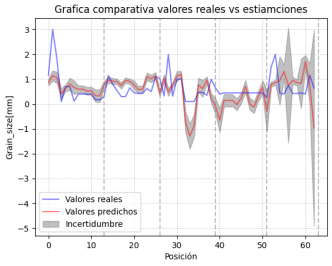
Figura 10. Comparativas de datos reales vs datos estimados

(a) Band gap

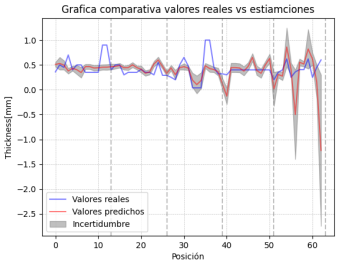
(b) DeltaH



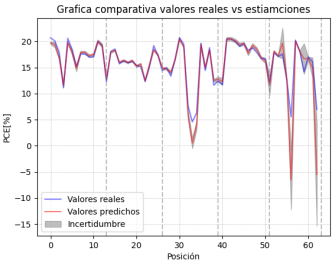
(c) DeltaL



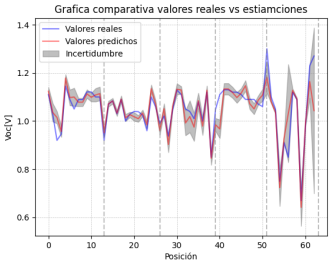
(d) Grain size



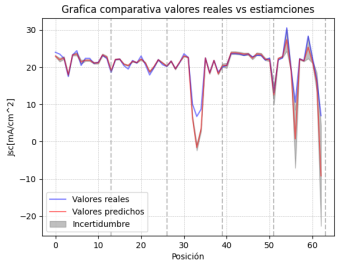
(e) Thickness



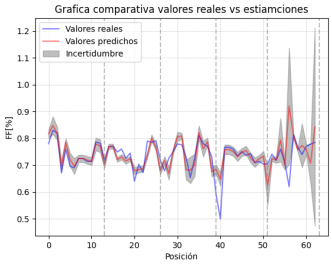
(f) PCE



(g) Voc



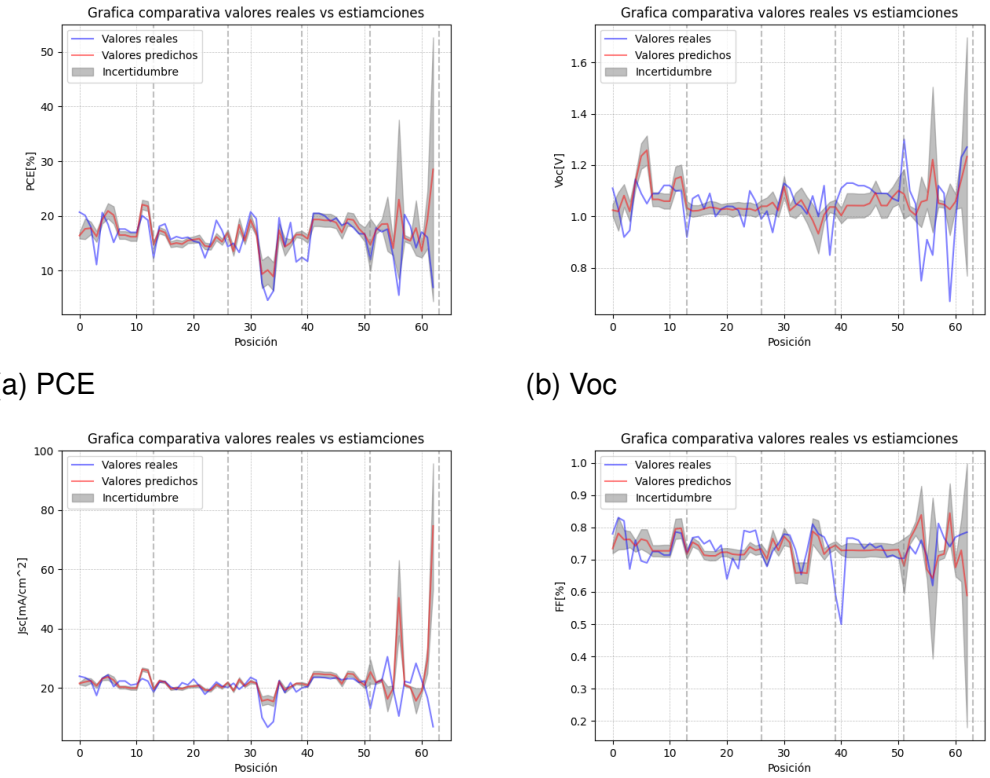
(h) Jsc



(i) FF

Mediante esta figura se presenta el comportamiento que siguen las predicciones frente a los datos reales y la incertidumbre que se asocia a las mencionadas predicciones.

Figura 11. Comparativas de datos reales vs datos estimados *“modelo imputación por estimación”*



(a) PCE

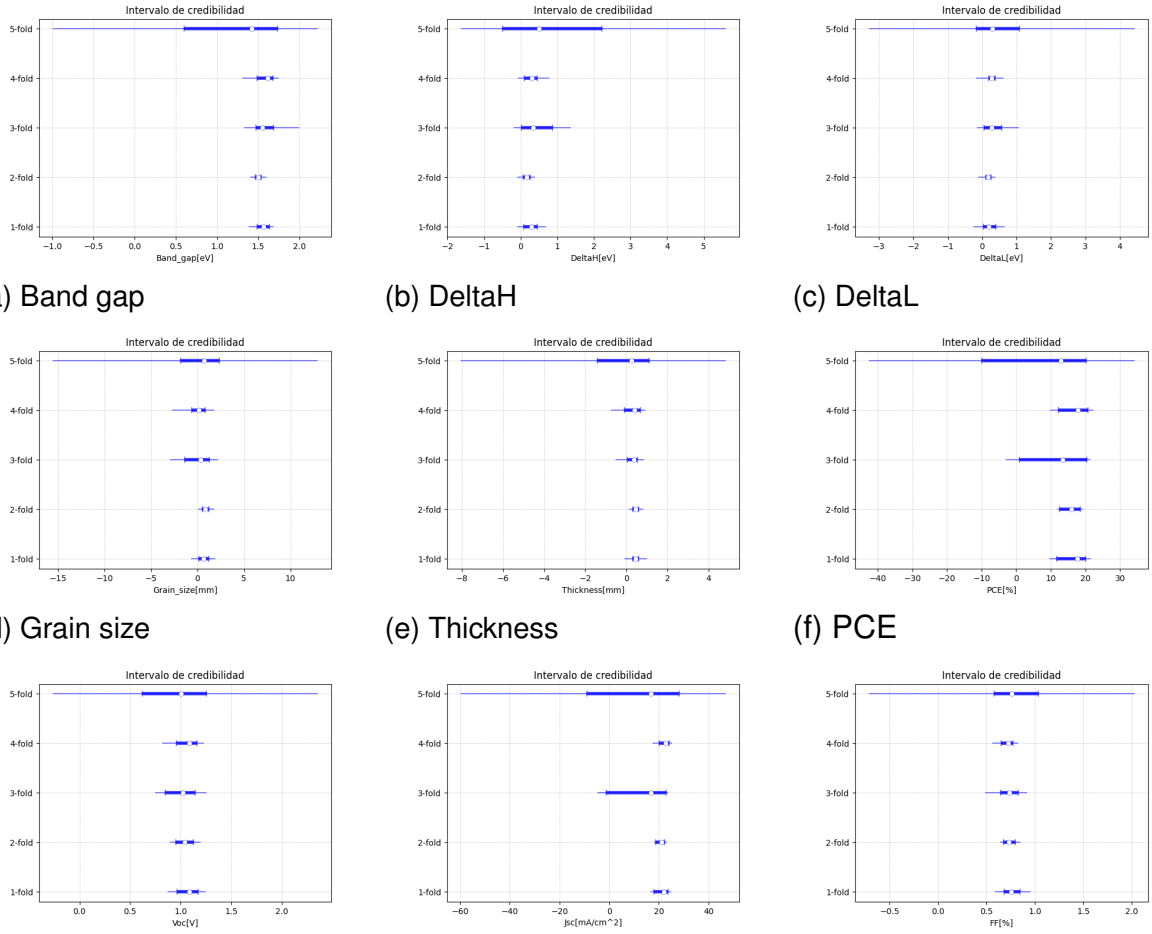
(b) Voc

(c) Jsc

(d) FF

Mediante esta figura se presenta el comportamiento que siguen las predicciones frente a los datos reales y la incertidumbre que se asocia a las mencionadas predicciones.

Figura 12. Representación gráfica del intervalo de credibilidad



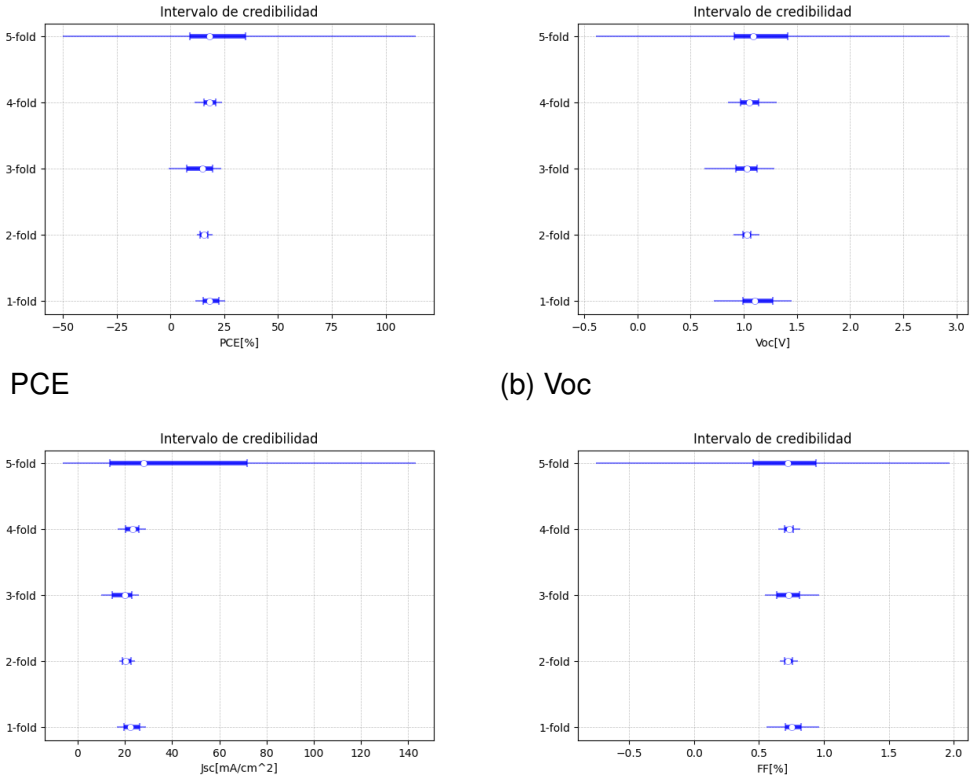
(a) Band gap (b) DeltaH (c) DeltaL

(d) Grain size (e) Thickness (f) PCE

(g) Voc (h) Jsc (i) FF

En esta figura se presentan las gráficas relacionadas a los intervalos de credibilidad por cada fold de las predicciones para cada modelo de variable imputada.

Figura 13. Representación gráfica del intervalo de credibilidad *“modelo imputación por estimación”*



(a) PCE

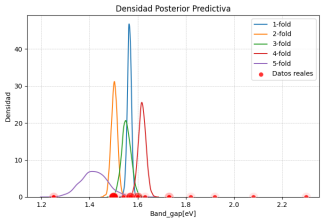
(b) Voc

(c) Jsc

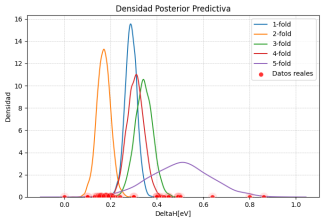
(d) FF

En esta figura se presentan las gráficas relacionadas a los intervalos de credibilidad por cada fold de las predicciones para cada modelo de variable imputada.

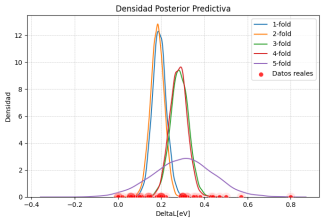
Figura 14. Densidad de probabilidad posterior



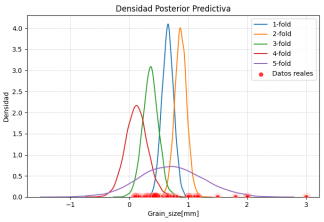
(a) Band gap



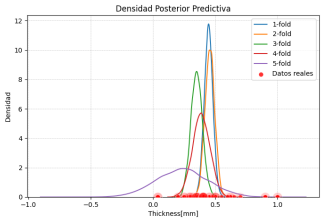
(b) DeltaH



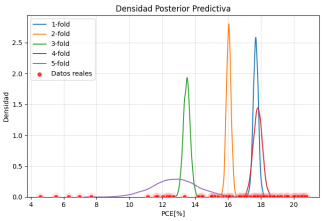
(c) DeltaL



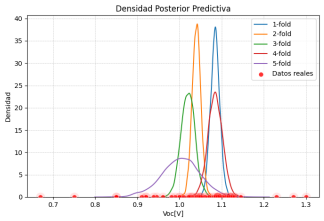
(d) Grain size



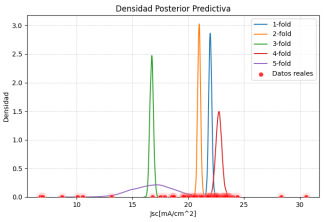
(e) Thickness



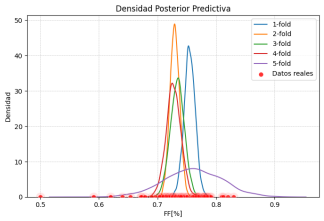
(f) PCE



(g) Voc



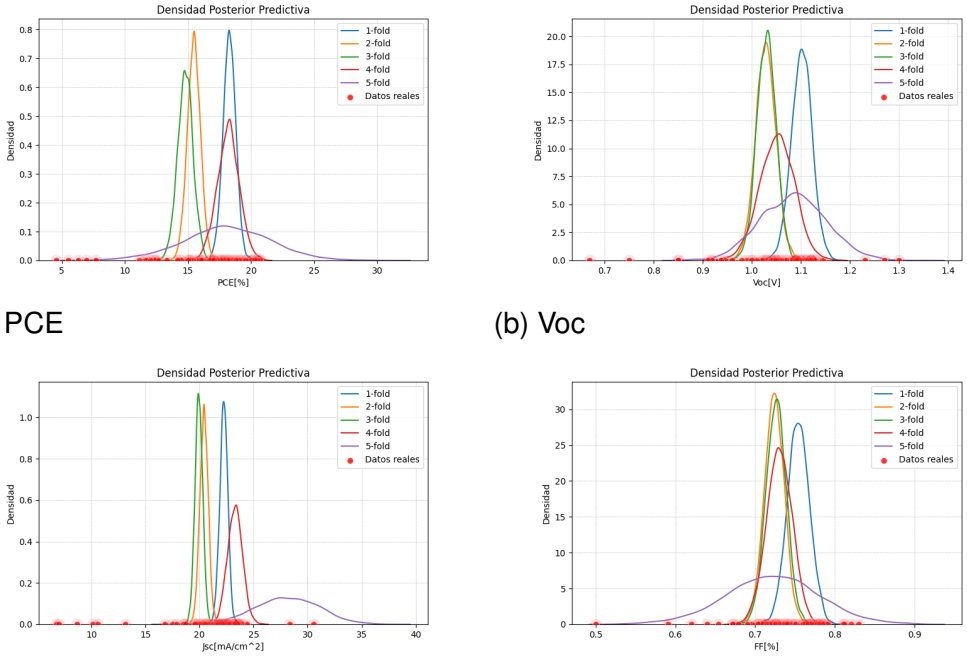
(h) Jsc



(i) FF

En la presente figura se muestran las gráficas de "Densidad de probabilidad posterior"por cada fold realizado para cada modelo de variable imputada, que muestra el comportamiento de las PDF asociadas a las predicciones.

Figura 15. Densidad de probabilidad posterior *“modelo imputación por estimación”*.



(a) PCE

(b) Voc

(c) Jsc

(d) FF

En la presente figura se muestran las gráficas de "Densidad de probabilidad posterior"por cada fold realizado para cada modelo de variable imputada, que muestra el comportamiento de las PDF asociadas a las predicciones.

3. CONCLUSIONES

- En este trabajo de grado se construye un modelo de Regresión Lineal Bayesiana que permite imputar los datos faltantes de un conjunto de datos sobre parámetros de síntesis y rendimiento de celdas solares de perovskita. Este mismo esquema sirve para la estimación de variables eléctricas de desempeño de celdas solares. Investigando en la literatura científica, no se evidencian trabajos previos sobre estos datos, ni en el uso de la Regresión Lineal Bayesiana como modelo de análisis. El enfoque innovador y los resultados obtenidos demuestran el potencial de combinar datos novedosos con metodologías estadísticas avanzadas y no convencionales como lo es la estadística bayesiana para abordar desafíos y generar conocimientos en áreas emergentes.

- Dados los resultados obtenidos por medio del modelo de imputación por estimación es apreciable que existe una relación entre los parámetros de síntesis o variables de entrada con los parámetros de rendimiento de las celdas solares de perovskita o variables de salida. Aunque los resultados obtenidos por este método son prometedores, los mejores resultados se obtienen para el caso en el que se estiman las variables eléctricas de desempeño de celdas solares a partir de todos los parámetros del conjunto de datos, es decir los parámetros de síntesis y el entorno de parámetros de rendimiento de síntesis.

- El uso del algoritmo de Regresión Lineal Bayesiano en este proyecto ha demostrado ser una herramienta valiosa para abordar problemas de análisis en casos en los que se tienen pocos datos. El uso de BLR permitió obtener estimaciones probabilísticas de los parámetros del modelo, lo que brinda una ventaja significativa en la interpretación y la evaluación de los resultados; Además, al entregar una función de probabilidad asociada a la estimación, el científico de ciencia de materiales se

puede formar una idea acerca de la incertidumbre asociada a esa estimación.

- La combinación de la información a priori con la información observada mediante el modelo bayesiano es una técnica eficaz para la estimación de variables aleatorias. Mediante este trabajo, se ha analizado la construcción de funciones de densidad de probabilidad a priori, lo que ha permitido obtener estimaciones acertadas de los parámetros estudiados en las celdas solares de perovskita.

- La utilización de transformaciones de variables ha demostrado ser una estrategia efectiva para mejorar el rendimiento de los resultados de la corriente de cortocircuito (Jsc) en el análisis de las celdas solares de perovskita. Esta técnica ha permitido abordar la falta de linealidad en la relación entre las variables predictoras y la variable objetivo, proporcionando una mejora significativa en la precisión de las estimaciones.

4. RECOMENDACIONES

Con el fin de maximizar la oportunidad de replica o mejora, de los resultados presentados en este trabajo con este modelo en particular y basado en las conclusiones derivadas de los resultados obtenidos, se recomienda tener en cuenta las siguientes practicas a la hora de abordar un proyecto con características similares.

- Prestar atención a la calidad de los datos con los cuales se dispone para trabajar es crucial. Es importante construir una base de datos de alta calidad. Esto implica elegir variables que tengan algún tipo de correlación entre sí, aunque no sea precisamente lineal, o realizar transformaciones para llevarlas a un contexto donde el modelo pueda explicarlas con más facilidad. Además, es fundamental contar con una cantidad suficiente de datos de entrenamiento y validación. Esto garantizará una evaluación más precisa y confiable del desempeño del modelo y permitirá obtener conclusiones más sólidas.

- Explorar y utilizar técnicas de transformación de variables basadas en algún criterio de correlación, para de esta manera mejorar no solo la relación entre las variables implícitas en el problema si no también mejorar el rendimiento de los modelos.

- Es fundamental considerar la aplicación de técnicas de regularización. Estas técnicas desempeñan un papel crucial al permitir que el modelo converja hacia una solución más precisa en un tiempo menor, evitando problemas comunes en el aprendizaje automático, como el sobreajuste y el subajuste.

- En cualquiera de los casos que se considere implementar un modelo de predicción o imputación, es importante utilizar la validación cruzada, esta permite un vistazo más amplio de los resultados, y el como se pueden ajustar los modelos a distintos subconjuntos de datos de entrenamiento y testeo.

- Considerar la replica y validación de los resultados en diferentes conjuntos de datos con diferentes condiciones experimentales. Esto permitirá verificar la generalización de los modelos y asegurar la confiabilidad de los resultados.
- Para las variables que no obtuvieron resultados satisfactorios en el proyecto presentado, se recomienda explorar algoritmos de Machine Learning basados en estadística bayesiana. Estos algoritmos ofrecen un amplio espectro de modelos predictivos, muchos de los cuales son más complejos que el modelo de Regresión Lineal Bayesiana utilizado en el proyecto. Por lo tanto, es posible que estos algoritmos generen mejores resultados para variables con relaciones complejas.

BIBLIOGRAFÍA

Ala-Luhtala, Juha y Robert Piché. “Gaussian Scale Mixture Models for Robust Linear Multivariate Regression with Missing Data”. En: *Communications in Statistics - Simulation and Computation* 45.3 (2016), págs. 791-813. DOI: 10.1080/03610918.2013.875565 (vid. pág. 13).

Arroyave Zapata, Juan David y Sebastian Camilo Toscano Higuera. *Enriquecimiento automático de datos extraídos mediante el procesamiento de lenguaje natural en la minería de información de celdas solares de perovskita*. Trabajo de Grado para optar al título de Ingeniera Electrónica. Universidad Industrial de Santander. 2022 (vid. págs. 10, 52).

Belenguer-Llorens, Albert et al. “A Novel Bayesian Linear Regression Model for the Analysis of Neuroimaging Data”. En: *Applied Sciences* 12.5 (2022), pág. 2571. DOI: 10.3390/app12052571 (vid. pág. 14).

Bisquert, Juan. “Basic Operation of Solar Cells”. En: *The Physics of Solar Cells Perovskites, Organics, and Photovoltaic Fundamentals*. Boca Raton, FL: CRC Press, 2018, págs. 139-143 (vid. págs. 20, 23).

— “Basic Operation of Solar Cells”. En: *The Physics of Solar Cells Perovskites, Organics, and Photovoltaic Fundamentals*. Boca Raton, FL: CRC Press, 2018, pág. 89 (vid. pág. 22).

Chan, Stanley H. “Regression”. En: *Introduction to Probability for Data Science*. Michigan Publishing, 2021, págs. 389-390 (vid. pág. 15).

Deisenroth, Marc Peter, A. Aldo Faisal y Cheng Soon Ong. *Mathematics for Machine Learning*. This version is free to view and download for personal use only. Not for re-distribution, re-sale, or use in derivative works. https://mml-book.com. Cambridge University Press, 2020, págs. 309-321 (vid. págs. 16-19).

Ding, Yaohui y Arun Ross. "A comparison of imputation methods for handling missing scores in biometric fusion". En: *Pattern Recognition* 45.3 (2012), págs. 919-933. DOI: https://doi.org/10.1016/j.patcog.2011.08.002 (vid. pág. 13).

Enders, Craig K. "Applied Missing Data Analysis". En: ed. por Todd D Little. New York London: The Guilford Press, 2010, págs. 39-49 (vid. pág. 13).

— "Applied Missing Data Analysis". En: ed. por Todd D Little. New York London: The Guilford Press, 2010, pág. 165 (vid. pág. 40).

Garcia-Aunon, Pablo y Antonio Barrientos Cruz. "Control optimization of an aerial robotic swarm in a search task and its adaptation to different scenarios". En: *Journal of Computational Science* 29 (2018), págs. 107-118. DOI: https://doi.org/10.1016/j.jocs.2018.10.004 (vid. pág. 11).

Jacobsson, T. Jesper et al. "An open-access database and analysis tool for perovskite solar cells based on the FAIR data principles". En: *Nature Energy* 7.1 (2022), págs. 107-115. DOI: 10.1038/s41560-021-00941-3 (vid. pág. 26).

Kononova, Olga et al. "Text-mined dataset of inorganic materials synthesis recipes". En: *Sci Data* 6 (2019), pág. 203. DOI: 10.1038/s41597-019-0224-1 (vid. pág. 10).

Lee, Michael D. y Eric-Jan Wagenmakers. "The Basics of Bayesian Analysis". En: *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press, 2014, pág. 3 (vid. pág. 15).

Lee, Sun-Yong et al. “Semi-GAN: An Improved GAN-Based Missing Data Imputation Method for the Semiconductor Industry”. En: *IEEE Access* 10 (2022), págs. 328-338. DOI: 10.1109/ACCESS.2022.3188871 (vid. pág. 10).

Olivetti, Elsa A. et al. “Data-driven materials research enabled by natural language processing and information extraction”. En: *Applied Physics Reviews* 7 (2020). DOI: https://doi.org/10.1063/5.0021106 (vid. pág. 9).

Ortega-San-Martin, Luis. “Introduction to Perovskites: A Historical Perspective”. En: *Revolution of Perovskite Synthesis, Properties and Applications*. Ed. por Narayanasamy Sabari Arul y Vellalapalayam Devaraj Nithya. Advanced Structured Materials. Springer, 2020. DOI: 10.1007/978-981-15-1267-4_12 (vid. pág. 19).

Rasifaghihi, N., S.S. Li y F. Haghighat. “Forecast of urban water consumption under the impact of climate change”. En: *Sustainable Cities and Society* 52 (2020), pág. 101848. DOI: https://doi.org/10.1016/j.scs.2019.101848 (vid. pág. 14).

Sahoo, Sarat Kumar, Balamurugan Manoharan y Narendiran Sivakumar. “Introduction: Why Perovskite and Perovskite Solar Cells?” En: *PEROVSKITE PHOTOVOLTAICS: Basic to Advanced Concepts and Implementation*. Ed. por Sabu Thomas y Aparna Thankappan. Academic Press, Elsevier, 2018, págs. 1-22. DOI: https://doi.org/10.1016/B978-0-12-812915-9.00001-0 (vid. pág. 19).

ScienceDirect. *Estadísticas de publicación de la base de datos, respecto a celdas solares de perovskita*. [Sitio web]. Consulta realizada el 2 de septiembre 2023. Disponible en. URL: https://www-sciencedirect-com.bibliotecavirtual.uis.edu.co (vid. pág. 12).

Shetty, Pranav y Rampi Ramprasad. “Automated knowledge extraction from polymer literature using natural language processing”. En: *iScience* 24.1 (2021), pág. 101922. DOI: 10.1016/j.isci.2020.101922 (vid. pág. 10).

Singh, Rohit y Santosh Singh Rathore. “Linear and non-linear bayesian regression methods for software fault prediction”. En: *International Journal of System Assurance Engineering and Management* 13 (2022), págs. 1864-1884. DOI: https://doi-org.bibliotecavirtual.uis.edu.co/10.1007/s13198-021-01582-1 (vid. pág. 14).

Thoppil, George Stephen y Alankar Alankar. “Predicting the formation and stability of oxide perovskites by extracting underlying mechanisms using machine learning”. En: *Computational Materials Science* 211 (2022). DOI: https://doi.org/10.1016/j.commatsci.2022.111506 (vid. págs. 9, 10).

Velez, Jeisson et al. “Absorber layer thickness as a new feature in statistical learning tools of Perovskite solar cells”. En: *Journal of Applied Research and Technology* In Press (2023) (vid. págs. 11, 26).

Wu, Rui et al. “Data Imputation for Multivariate Time Series Sensor Data With Large Gaps of Missing Data”. En: *IEEE Sensors Journal* 22.11 (2022), págs. 671-683. DOI: 10.1109/JSEN.2022.3166643 (vid. pág. 10).

Wu, Zhifang et al. “Passivation strategies for enhancing device performance of perovskite solar cells”. En: *Nano Energy* 115 (2023). DOI: https://doi.org/10.1016/j.nanoen.2023.108731 (vid. pág. 20).