ENRIQUECIMIENTO AUTOMÁTICO DE DATOS EXTRAÍDOS MEDIANTE EL PROCESAMIENTO DE LENGUAJE NATURAL EN LA MINERÍA DE INFORMACIÓN DE CELDAS SOLARES DE PEROVSKITA

Juan David Arroyave Zapata

Sebastian Camilo Toscano Higuera

Trabajo de Grado para optar al título de Ingeniera Electrónica

Director

Cristian David Camacho Parra

Ingeniero Electricista

Codirector

Franklin Alexander Sepúlveda Sepúlveda

Doctorado en Ingeniería Electrónica

Universidad Industrial de Santander

Facultad de Ingenierías Fisicomecánicas

Escuela de Ingenierías Eléctrica, Electrónica y de Telecomunicaciones

Bucaramanga

2022

Agradecimientos

Agradecer primero a Dios que siempre me guió en el camino de tomar las mejores decisiones para mi desarrollo personal. A mis padres, por el amor y el apoyo incondicional que siempre me han demostrado, que me permitieron culminar con este nuevo objetivo de mi vida personal. A mis hermanos, que de alguna u otra manera siempre estuvieron presentes con su apoyo. Al resto de mi familia, que también han sido un pilar fundamental en este proceso. A mi novia que me acompañó gran parte del camino brindándome paz y mucha calma. A mi compañero de proyecto y en especial a nuestro director y codirector que estuvieron presentes en este crecimiento intelectual. Por último y no menos importante, a mis mascotas que me brindaron ese apoyo emocional cuando más lo necesité.

Sebastian Camilo Toscano Higuera

Agradezco al señor Jesucristo por apoyarme en los momentos que más lo necesite, por guiarme de la forma más provechosa e inteligente para mí. Me agradezco por trabajar duro por lo que quiero y siempre creer en mi mismo, a mi familia por enseñarme lo que me faltaba para mejorar y sobrepasar cada adversidad presentada. También a mí abuelita Elida por llenarme de amor cuándo más lo necesite, este trabajo se lo dedico a ella que me cuida todo el tiempo.

Juan David Arroyave Zapata

Tabla de Contenido

Lista de Figuras

Lista de Tablas

Resumen

Título: Enriquecimiento automático de datos extraídos mediante el procesamiento de lenguaje natural en la minería de información de celdas solares de perovskita. [*]

Autores: Sebastian Camilo Toscano Higuera & Juan David Arroyave Zapata [**]

Palabras Clave: Imputación, Mezclas gaussianas, Celdas solares de Perovskita

Descripción: En los últimos años se identifica un desarrollo notorio en el campo de la ciencia de los materiales logrando grandes cantidades de información. Dicha información se hace compleja de recopilar debido a distintos factores que dificultan obtener una base de datos confiable con una estructura bien definida. Por esta razón, las técnicas de PLN ofrecen una solución a una robusta recopilación de datos. Sin embargo, este conjunto de datos extraído con PLN presentan incongruencias y/o instancias incompletas. Por ello, se hace necesario aplicar estrategias para manejar estos problemas, las cuales en este trabajo de investigación se usa un único modelo de función de probabilidad para el enriquecimiento de un conjunto de datos de parámetros de celdas solares de Perovskita. El desarrollo de esta investigación se lleva a cabo en 2 fases. En la primera se acondiciona la base de datos para eliminar valores atípicos y en la segunda se entrena y evalúa el modelo. En consecuencia a este procedimiento se consigue una precisión media mayor al 50 % y en el caso de la variable de eficiencia de conversión de energía se logra una precisión del 84 %. En conclusión, esta investigación presenta resultados prometedores dada que resulta ser una propuesta innovadora en el campo de las celdas solares de Perovskita. Por último para obtener un mejor entrenamiento del modelo se recomienda hacer uso de una base de datos con una cantidad de datos más significativa.

---

[*] Trabajo de grado

[**] Facultad de Ingenierías Físico-Mecánicas. Escuela de Ingenierías Eléctrica, Electrónica y telecomunicaciones. Director: Cristian David Camacho Parra, Ingeniero electricista.
Codirector: Franklin Alezander Sepúlveda Sepúlveda, Doctorado en ingeniería electrónica.

Abstract

Title: Automatic enrichment of extracted data using natural language processing in mining of information of solar cells of perovskite. [*]

Author: Sebastian Camilo Toscano Higuera & Juan David Arroyave Zapata [**]

Keywords: Imputation, Gaussian mix, Solar cells, Perovskite.

Description: In recent years, a notable development has been identified in the field of materials science, achieving large amounts of information. This information becomes complex to collect due to different factors that make it difficult to obtain a reliable database with a well-defined structure. For this reason, PLN techniques offer a solution to robust data collection. However, this dataset extracted with PLN presents inconsistencies and/or incomplete instances. Therefore, it is necessary to apply strategies to handle these problems, which in this research work we use a single probability function model for the enrichment of a dataset of Perovskite solar cell parameters. The development of this research is carried out in 2 phases. In the first, the database is conditioned to eliminate outliers, and in the second, the model is trained and evaluated. As a result of this procedure, an average accuracy greater than 50 % is achieved and in the case of the energy conversion efficiency variable, an accuracy of 84 % is achieved. In conclusion, this research presents promising results given that it turns out to be an innovative proposal in the field of Perovskite solar cells. Finally, to obtain a better training of the model, it is recommended to use a database with a more significant amount of data.

---

[*] Bachelor Thesis

[**] Facultad de Ingenierías Físico-Mecánicas. Escuela de Ingenierías Eléctrica, Electrónica y telecomunicaciones.
Director: Cristian David Camacho Parra, Ingeniero electricista.
Codirector: Franklin Alezander Sepúlveda Sepúlveda, Doctorado en ingeniería electrónica.

Introducción

En los últimos años la investigación de celdas solares de Perovskita ha crecido de forma exponencial, donde algunos esfuerzos se han centrado en la mejora de la eficiencia, la histéresis, la reproducibilidad y la toxicidad de los materiales Olivetti et al. (2020). Estas investigaciones contienen gran cantidad de información valiosa. Sin embargo, esta información no es fácil de extraer. Yılmaz and Yıldırım (2021) Çağla Odabaşı and Yıldırım (2020).

Para recopilar esta información a gran escala y obtener un conjunto de datos confiable con estructuras definidas y claras para su procesamiento se requiere de inversiones significativas en tiempo y dinero. Además, algunos autores no reportan la misma información dado el enfoque de cada uno en su investigación, aumentando la probabilidad de hallar sesgos en las bases de datos Allahyari et al. (2017) Arora and Kansal (2019).

Es aquí donde las estrategias de procesamiento del lenguaje natural (PLN) dan solución al problema de la recopilación de información masiva dado que la extracción de datos se complementa con el aprendizaje automático. Estas estrategias de PLN optimizan la extracción de información en términos de dinero y tiempo Kononova et al. (2019) Cole (2020) gracias principalmente a que permite procesar grandes densidades de información haciendo uso de investigaciones previas como principal fuente de datos para así generar reportes concisos referentes a información especificada al momento de su implementación Shetty and Ramprasad (2021).

No obstante, este conjunto de datos extraídos mediante las técnicas de aprendizaje automático presentan problemas y errores en los datos extraídos ya que se encuentran datos incompletos que complican la interpretación de la información. Estos errores se atribuyen a diferentes razones, como el mal funcionamiento del dispositivo y/o fallas en la medición en la caracterización de las celdas solares Swain and Cole (2016). Las estrategias para manejar datos vacíos se pueden dividir en 3 grupos distintos: a) eliminación de casos incompletos, b) imputación de valores y c) diseño de métodos de aprendizaje que manejen datos faltantes de forma directa Arora and Kansal (2019).

En la caracterización física del comportamiento de las celdas solares se tiene que las variables de interés más comunes son el voltaje de circuito abierto, la corriente de cortocircuito, el factor de llenado y la eficiencia de conversión de energía. Estas se relacionan bajo expresiones matemáticas que no tienen en cuenta las particulares relaciones intrínsecas y extrínsecas de la celda solar, lo que se traduce en mayor complejidad del cálculo de estas variables Kim et al. (2017). En este trabajo de grado se propone y aplica el uso de un único modelo de mezclas gaussianas (GMM) representando modelos probabilísticos, con los cuáles se realiza el enriquecimiento de parámetros de rendimiento en la base de datos de celdas solares de Perovskita, consiguiendo que independientemente de la variable se logre generar una imputación artificial. La metodología adaptada es comprendida de 3 etapas, preprocesamiento de datos, preparación del modelo y aplicación de la investigación con la que se busca enriquecer conjuntos de datos que faciliten el desarrollo científico de nuevas tecnologías relacionadas con las Perovskitas fotovoltaicas . Sang et al. (2020) Jiang et al. (2021)

1. Objetivos

Objetivo general

- Implementar un modelo de aprendizaje automático basado en estimación de función de densidad de probabilidad para el enriquecimiento de datos extraídos por medio de herramientas de procesamiento de lenguaje natural.

Objetivos específicos

- Implementar un algoritmo para estimar funciones de densidad de probabilidad basados en mezclas Gaussianas.
- Evaluar el modelo de aprendizaje automático en términos precisión y funcionabilidad.
- Imputar parámetros de rendimientos faltantes de la base de datos de celdas solares de Perovskita.

2. Marco teórico y estado del arte

2.1. Marco teórico

2.1.1. Imputación con Mezclas Gaussianas. En esta sección se explicará cómo un modelo de mezclas Gaussianas puede ser usado para imputar valores faltantes. Asumamos que X es la entrada del modelo e Y es la variable objetivo. Ambas son definidas como variables aleatorias. Por simplicidad asumiremos un sistema MISO (Múltiples entradas, una sola salida), pero el análisis es fácilmente extendido a más variables de salida.

Definiendo $f_{x,y}(x,y)$ como la función de probabilidad de densidad (FDP), la salida Y puede ser estimada usando el concepto de valor esperado de la siguiente manera:

$$y = m_x = E[Y|X = x] = \int y f_y(y|X) dy \tag{1}$$

Añadiendo un variable auxiliar $Z = [X,Y]$, entonces $f_z = f_{X,Y}(x,y)$, donde

$$f_Z(Z) = \sum_{j=1}^{K} \pi_j \cdot \mathcal{N}\left(Z; \mu_j, C_j\right) \tag{2}$$

De la ecuación anterior $\mathcal{N}\left(Z; \mu_j, C_j\right)$ es una función de densidad de probabilidad de dimensión $d+1$ con media μ y matriz de covarianza C donde d es la dimensión de X.

$$C_j = \begin{bmatrix} C_j^{YY} & C_j^{YX} \\ \\ C_j^{XY} & C_j^{XX} \end{bmatrix} \tag{3}$$

Dado que estamos asumiendo una sola variable de salida, C_j^{YY} es un valor escalar; C_j^{XX} es la matriz de covarianza de entrada con dimensiones $d \times d$; C_j^{YX} es un vector de dimensiones $1 \times d$; y, $C_j^{XY} = \left(C_j^{YX}\right)^T$ es un vector columna de dimensiones $d \times 1$. Complementariamente tenemos qué:

$$m_j(X) = \mu_j^Y + C_j^{YX} \cdot \text{inv}\left(C_j^X\right) \cdot \left(x - \mu_j^X\right) \tag{4}$$

Entonces la estimación de Y dado un valor x:

$$E[Y \mid X = x] = \sum_{j=1}^{K} \beta_j(x) \cdot m_j(x) \tag{5}$$

Donde

$$\beta_j = \frac{\pi_j \cdot \mathcal{N}\left(x; \mu_j^X, C_j^X\right)}{\sum_{j=1}^{K} \pi_j \cdot \mathcal{N}\left(x; \mu_j^X, C_j^X\right)} \tag{6}$$

La expresión $\beta_j(\Pr(j \mid X = x)$) también se denomina responsabilidades, la cuál juega el papel principal en la estimación de valores desconocidos. Para más profundización visitar Sepúlveda (2020).

2.1.2. Modelos de Mezclas Gaussianas. Los modelos de mezclas Gaussianas son modelos flexibles, con los cuáles se puede manejar valores atípicos, heterogeneidad y asimetría. Donde cualquier distribución continua puede aproximarse mediante una distribución de mezcla Gaussianas finitas Sang et al. (2020) Pilania et al. (2016) con pesos, medias y covarianzas desconocidas. Los parámetros de las mezclas que componen un Modelo Gaussiano pueden ser estimados mediante la aplicación del algoritmo Expectation-Maximization (EM) Jiang et al. (2021). Es característico de cada GMM que cada mezcla gaussiana es conocida como una componente y representa un agrupamiento diferente.

Cada componente del modelo es lineal e independiente entre si, y al sumarlas, su totalidad representa la función de densidad de probabilidad (FDP) del GMM

$$P(x_i) = \sum_{k=1}^{K} \pi_k f_k(x_i \mid \mu_k, \varepsilon_k) \tag{7}$$

En la fórmula anterior, π_i representa el peso de la componente k en el GMM; μ_i y ε_i representan el vector medio y la matriz de covarianza de la componente k; $P(x_i)$ representa la probabilidad de x_i generada por el GMM. Además, π_k cumple las siguientes condiciones:

$$\pi_k \geq 0, \quad \sum_{k=1}^{G} \pi_k = 1 \tag{8}$$

La FDP k de la componente $f_k(x_i \mid \mu_k, \varepsilon_k)$ es expresada cómo:

$$f_k(x_i \mid \mu_k, \varepsilon_k) = \frac{\exp\left\{-(1/2)(x_i-\mu_k)^T \sum_k^{-1}(x_i-\mu_k)\right\}}{(2\pi)^{p/2}|\varepsilon_k|^{1/2}} \tag{9}$$

2.1.3. Criterio de información bayesiano. El criterio de información bayesiano (BIC) es utilizado para estimar del factor de Bayes con el cuál se evalúa el desempeño de dos o más modelos. Este criterio es adecuado para la elección del número óptimo de mezclas y el tipo de matriz de covarianza de las mezclas del GMM Jiang et al. (2021). Definido como:

$$BIC = ln(n)d - ln(L\hat{}) \tag{10}$$

De la ecuación anterior, n es el número de datos, d es el número de parámetros del modelo, y $L\hat{}$ es el valor maximizado de la verosimilitud del modelo ajustado. Si $L\hat{}$ aumenta, el puntaje BIC disminuye lo que traduce un modelo mejor ajustado.

2.1.4. Celdas solares de Perovskita. Para hablar de celdas solares, hasta el momento se tienen en cuenta tres generaciones: primera generación (silicio cristalino), segunda generación (película delgada) y tercera generación (tecnologías emergentes) Sahoo et al. (2018).

La primera generación de celdas solares es la tecnología basada en obleas de silicio donde más del 85 % del mercado fotovoltaico usa este tipo de celdas solares Sahoo et al. (2018). En la segunda generación de celdas solares la eficiencia de estas ya superó el 20 % gracias a su estabilidad térmica y química Sahoo et al. (2018).

Con la creciente demanda energética el desarrollo de la tercer generación se enfoca con el objetivo de mejorar las relaciones eficiencia/costo y que al mismo tiempo, el impacto ambiental en comparación con las generaciones anteriores sea menor. Dentro de las alternativas contempladas como tecnologías emergentes, se incluyen las celdas solares orgánicas, puntos cuánticos, Perovskitas, entre otros Sahoo et al. (2018).

El mineral de Perovskita tiene una gran capacidad de absorber luz y utiliza menos de un 1μm para obtener una cantidad similar de luz solar con respecto a otras celdas solares. Dada su versatilidad la perovskita se ha utilizado como reemplazo en algunas celdas solares de película delgada siendo usada por primera vez en 2009 por Miyasaka. El Laboratorio Nacional de Energías Renovables (NREL) reporta las mejores eficiencias de las tecnologías de celdas fotovoltaicas desde 1976 NREL (2022). Para el caso particular de las celdas solares de perovskita, en la figura 1 tomada de NREL (2022) se observa su rápido crecimiento, pasando de eficiencias cercanas al 3,8 % en 2009 a valores del 25,7 % a en Junio 2022 Sahoo et al. (2018) NREL (2022).

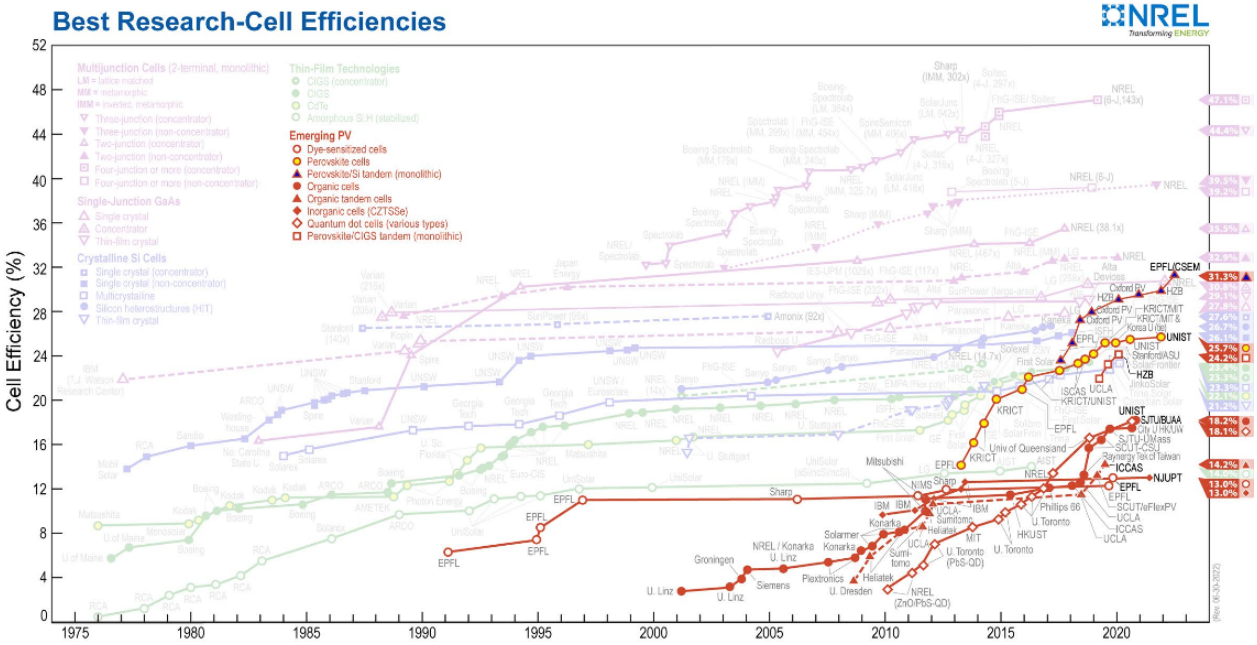


Figura 1. Diagrama de eficiencias en celdas solares de Perovskita.

2.1.5. Parámetros de desempeño. En términos generales, las características de rendimiento de celdas solares de Perovskita pueden describirse en función de la curva voltaje-corriente, la cuál puede tener una forma como la de la imagen 2. Con esta gráfica se consigue una aproximación inicial para describir el voltaje de circuito abierto (Voc), la corriente de cortocircuito(Jsc), el factor de llenado (FF) y la eficiencia de de conversión de energía (PCE), de una celda solar Bisquert (2017).

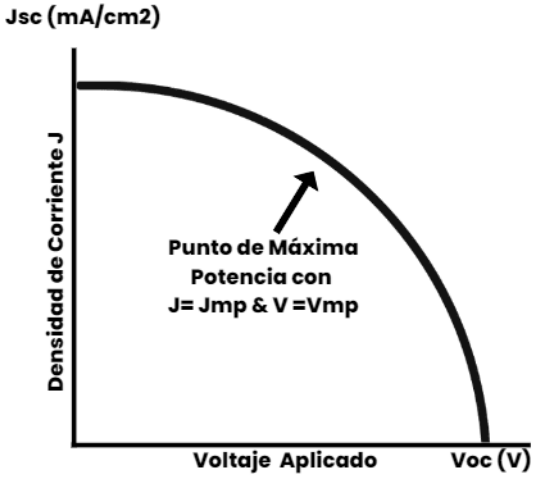


Figura 2. Diagrama J-V representativo de una celda solar.

Estas variables se relacionan a través del modo de operación de la celda solar, donde con barridos con diferentes tensiones y en diferentes sentidos, se logra registrar un comportamiento como el mostrado en la figura 2. Este diagrama no solo facilita el cálculo en los parámetros de rendimiento, sino que también, permite determinar el grado de histéresis del material, lo cual se presenta cuando la curva que se hace con el barrido de izquierda a derecha es muy diferente a cuando se realiza al contrario Fonash (2012).

Dentro de los puntos de interés se resalta el de máxima potencia, el cual se obtiene al encontrar la J_{mp} y el V_{mp}, que son la corriente de máxima potencia y el voltaje de máxima potencia respectivamente. Este punto es de los más relevantes, pues determina uno de los parámetros con el que se evalúa el desempeño general de la celda solar: La eficiencia de conversión de energía, la

cual se describe en la ecuación siguiente 11 Fonash (2012).

$$\eta_{PCE} = \frac{J_{mp} \times V_{mp}}{P_{in}} \tag{11}$$

Donde el total de potencia de entrada P_{in} por área que incide en una celda para un fotón dado ϕ_o es la integral de la energía entrante por tiempo por área por ancho de banda en todo el espectro de fotones Fonash (2012).

$$P_{in} = \int_{\lambda} \frac{hc}{\lambda} \phi_o(\lambda) d\lambda \tag{12}$$

Idealmente la curva de corriente-tensión sería rectangular, pero dados los factores externos se está lejos de este comportamiento. Para estudiar que tan lejos se está del comportamiento ideal se tiene el factor de forma, el cual se usa para medir que tan cercano está el sistema experimental del sistema ideal. Este parámetro se encuentra descrito en la ecuación 13 donde la J_{sc} es la corriente de corto circuito y el V_{oc} es el voltaje de circuito abierto Fonash (2012).

$$FF = \frac{J_{mp} \times V_{mp}}{J_{sc} \times V_{oc}} \tag{13}$$

Con una correlación fuerte entre variables, se podría suponer que la caracterización de las celdas solares resultaría simple por medio de las relaciones matemáticas, pero aún así, no suele ser este el procedimiento estándar al momento de caracterizar las propiedades físicas de celdas solares de Perovskita. Al contrario de esto, en la práctica se suelen llevar a cabo complicados procedi-

mientos experimentales en dónde se busca mejorar la precisión del cálculo de los parámetros de rendimientocheol Kim et al. (2021).

Lo anterior se fundamenta en la degradación continua del material debido a factores únicos intrínsecos y extrínsecos, lo que implica que la forma de caracterización se debe adaptar al tipo de tecnología estudiada y que paralelamente, tenga en cuenta el posible grado de histéresis del material cheol Kim et al. (2021) Kim et al. (2015).

2.2. Estado del arte

El desarrollo progresivo de nuestra sociedad es impactado directamente por la innovación en la adaptabilidad de nuevos materiales Correa-Baena et al. (2018). Paralelamente la familiarización con el aprendizaje automático es acelerado por este desarrollo ya que se hace popular en la comunidad científica al emplear técnicas optimizadas para cumplir con la eficiencia, estabilidad y costos Çağla Odabaşı and Yıldırım (2020).

En trabajos de investigación anteriores se usa el enfoque de aprendizaje automático para determinar la estabilidad energética de celdas solares de Perovskita Çağla Odabaşı and Yıldırım (2020). Otros investigadores, con enfoques similares, consiguen estimar el comportamiento de la banda de conducción, el cuál es fundamental para determinar el campo de aplicabilidad de las celdas solares Stanley and Gagliardi (2019). Pero si de imputar información faltante en celdas solares de Perovskita se trata, esta investigación resulta innovadora, aun cuando la aplicabilidad del aprendizaje automático en el campo de materiales crece directamente con los reportes investigativos sobre esta nueva generación de celdas solares Yılmaz and Yıldırım (2021) Cole (2020). Para manejar los datos vacíos las estrategias se dividen en 3 grupos: a) eliminación de casos incompletos,

b) imputación de valores y c) diseño de métodos de aprendizaje.

El caso de eliminación por lista debido a su simplicidad es el más común. Este descarta la fila o columna según el caso donde falte información. El grupo de investigadores enfocado en el diseño de métodos de aprendizaje automático consiste en adaptar métodos existentes o crear nuevos con la capacidad de manejar estos datos faltantes. Estos métodos de aprendizaje automático evitan el uso de eliminación por lista dado que induce una pérdida considerable de información pues hay casos en dónde los datos faltantes son significativos para las investigaciones Mesquita et al. (2019).

Un método de imputación de valores es el método de los K vecinos más cercanos (KNN por sus siglas en inglés) donde se predice un nuevo dato basado en observaciones conocidas o pasadas. Éste método puede ser considerado extremo debido a que si hay algún caso desfavorable o no consistente, este afecta directamente la posible predicción Raschka (2018). Otro método es la imputación por regresión lineal la cual predice el valor de una variable según el valor de otra u otras. La regresión lineal se ajusta a una linea recta o a una superficie para minimizar las discrepancia entre los valores previstos y los reales Limeres (2011).

Lee and Kim (2022) aplican un modelo de mezclas gaussianas tradicional a una base de datos de ingresos y gastos a los hogares de Corea del Sur donde la tasa de coincidencia fue del 85 %, lo cual sugiere errores de medición en los ingresos reportados y Xue et al. (2019) proponen un modelo de imputación que captura tanto información transversal como correlaciones tempora-

les con datos de pacientes ingresados a hospitales donde el mejor modelo puede proporcionar una imputación más precisa que los puntos de referencia en todos los conjuntos de datos.

3. Enriquecimiento automático de parámetros de desempeño de celdas solares de Perovskitas

El proceso de enriquecimiento propuesto en esta investigación contempla dos procesos generales, en el primero se prepara la base de datos de forma que las inconsistencias de almacenado de información no interfieran con el cálculo del modelo, mientras que en la fase dos se llevan a cabo los pasos necesarios para entrenar y evaluar el desempeño del modelo calculado. En la Figura 3 se presenta una vista general de la investigación y el flujo que se lleva a cabo para enriquecer información de manera automática.

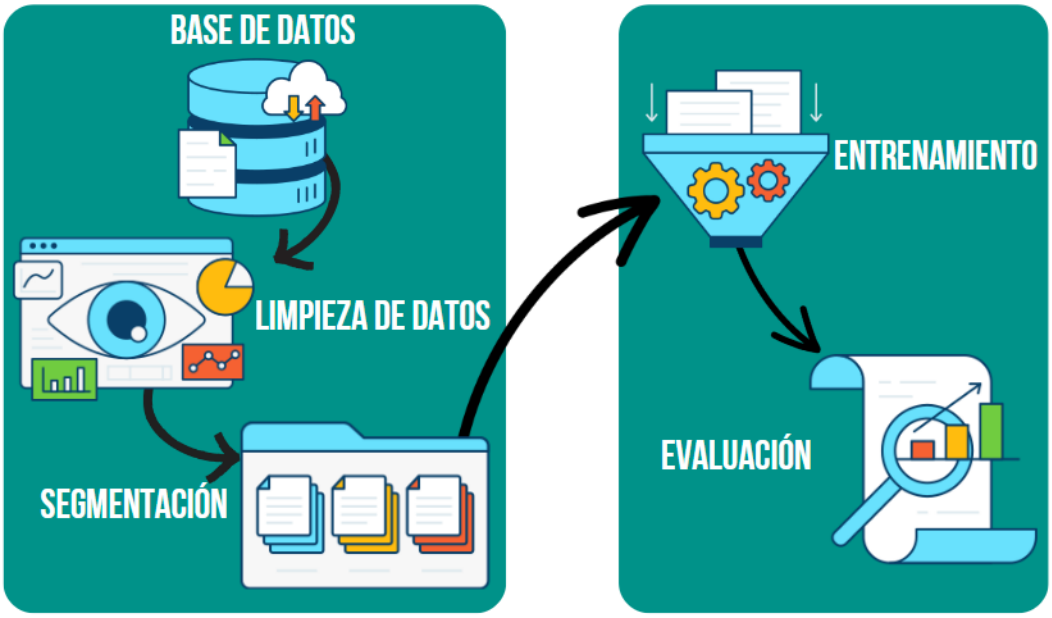


Figura 3. Sistema de enriquecimiento automático de parámetros de rendimiento de celdas solares de Perovskita.

3.1. Preprocesamiento de los Datos

El preprocesamiento es dividido en 2 etapas diferentes, éstas fueron elegidas luego de estudiar los enfoques en la literatura seleccionada, dónde la metodología de esta investigación recopila los casos más provechosos y flexibles con respecto a nuestro conjunto de datos. Inicialmente se depura la base de datos, consiguiendo mayor control sobre los rangos de cada variable, esto se hace con la limitación de que la omisión de datos no afecte significativamente la cantidad de información que se usará en el entrenamiento del modelo. Luego, dado que en la etapa de preparación se necesitan dos conjuntos de datos uno para el entrenamiento y otro para la evaluación del modelo, segmentamos la base datos, produciendo 2 diferentes, adicional se estandariza la base de datos. A continuación, se entra en más detalle sobre las actividades mencionadas.

3.1.1. Depuración de la base de datos. En capítulos anteriores hablamos que los artículos científicos son presentados en formatos diferentes, lo cual hace que las bases de datos tengan inconsistencias en sus valores. A esto sumamos que la base de datos usada en esta investigación brindada por CMIB (2020), la cuál es de uso público, fue extraída mediante técnicas de Web Scraping. Estas técnicas pueden aumentar la aparición de errores si la estructura y/o nomenclatura de las páginas web difieren sutilmente.

Con la base de datos extraída, seguimos con la identificación de columnas no significativas para nuestra investigación las cuáles decidimos filtrar, dado que no aportan nada para el posterior entrenamiento, además también se filtra el área, puesto que esta variable es seleccionada de forma

arbitraria por los investigadores. De esto quedan las variables de interés definidas cómo: eficiencia de conversión de energía (PCE), corriente de corto circuito (Jsc), voltaje de circuito abierto (Voc) y el factor de forma (F.F).

Con las variables de interés definidas, se hace necesario entender el comportamiento de la base datos, para ello realizamos una descripción inicial, la cuál es calculada con el comando *describe* de la librería *Sklearn*. En esta descripción se calculan los hiperparámetros que dictan el comportamiento de cada variable como la media y la desviación estándar, además de información inicial para entender la distribución de los datos. Esta información se presenta en la tabla 1.

Tabla 1
Descripción de datos sin tratar.

	Jsc(mA/cm2)	Voc(V)	F.F.	PCE(%)
Total	2917	2923	2913	2936
media	469,412	232,932	0,637	94,844
std	12502,304	3316,189	0,130	2847,723
min	700e-06	9.9e-03	0,06	2,54e-09
max	635444	103046	0,933	151812

La información de la tabla 1 nos presenta una desviación estándar demasiado alta en las variables *JSC*, *VOC* y *PCE*. Esto indica que hay una dispersión de datos elevada con respecto a la media, la cual en estos casos también se observa que tiene una magnitud elevada. Paralelamente,

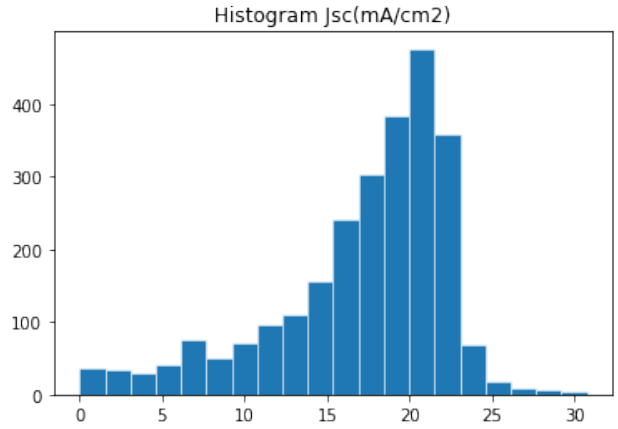
los máximos de las variables son considerablemente mayores en magnitud, con respecto de los valores mínimos, esto se traduce en mayor complejidad para el hallazgo del modelo por la gran variabilidad entre las variables.

Luego de identificar las inconsistencias en la base de datos establecemos una estrategia para la limitación de los rangos de las variables, pues las inconsistencias de la información inicial en el calculo del modelo aumenta la probabilidad de sobre ajuste. Esta limitación de datos se hace acotando las variables en diferentes rangos con el fin de disminuir los datos atípicos sin perder una gran cantidad de información. Estos rangos son: a) J_{sc} de 0 a 35, b) V_{oc} de 0 a 2, c) FF de 0 a 1 y d) PCE de 0 a 40. De manera concreta el sistema de nuestra investigación contempla tres entradas y una única salida.

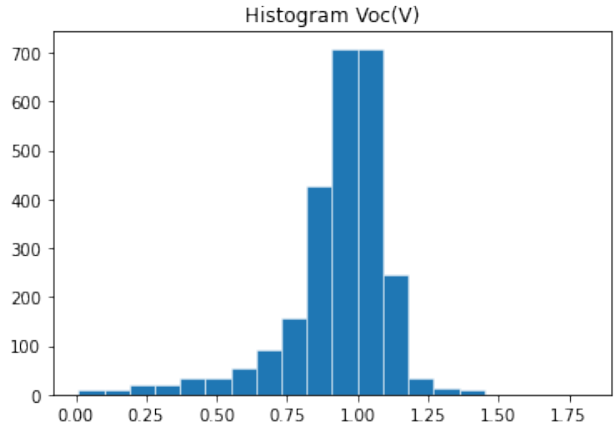
La eliminación de datos atípicos implica qué en esta investigación no se tengan en cuenta alrededor de 350 celdas solares de Perovskita, esto porque su aporte al cálculo del modelo podría ser negativo. En la tabla 2 presentamos la base de datos tratada, aquí la nube de datos gira en torno a valores menores y más comunes en la comunidad científica. Además en la figura 4 se presentan los histogramas de los parámetros de rendimiento. Estos cambios en los datos ofrecen mayor resolución la cual permite mayor control al momento del cálculo del modelo.

Tabla 2
Descripción de la base de datos después de la limpieza.

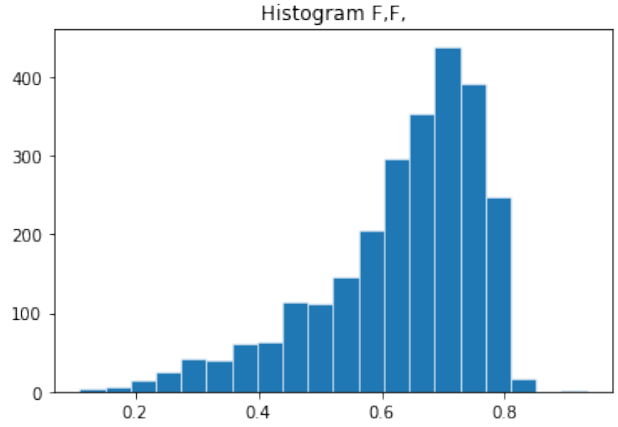
	Jsc(mA/cm2)	Voc(V)	F.F.	PCE(%)
total	2569	2569	2569	2569
media	17,14	0,92	0,63	10,86
std	5,33	0,18	0,13	5
min	7e-04	9,9e-03	0,11	2,54e-09
max	30,75	1,81	0,93	22,19



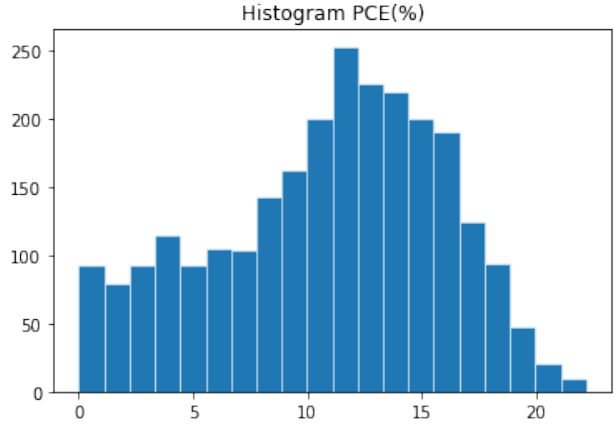
(a) Corriente de corto circuito.



(b) Voltaje de circuito abierto.



(c) Factor de forma.



(d) Eficiencia de conversión de energía.

Figura 4. Histogramas de la base de datos después de la limpieza.

3.1.2. Estandarización y Segmentación. Tras la limpieza de valores atípicos se hace necesario estandarizar cada una de las variables, pues la diferencia de magnitudes es alta y esto aumenta el sesgo en el hallazgo de los modelos. Para conseguir este escalamiento hacemos uso de la librería preprocessing de *Sklearn* donde se estandarizan los datos por columna empleando los

valores de media y desviación estándar del conjunto de datos de entrenamiento consiguiendo una media de 0 y una desviación estándar de 1.

Por último, el conjunto de datos a usar en la etapa de aplicación es extraído de la base de datos original, en donde definimos como concepto de selección, que haya una sola columna vacía por fila, es decir que la información disponible sobre una celda solar en particular en cuánto a sus parámetros de rendimiento sea del 75 %. Este conjunto de datos se presenta en la tabla 3:

Tabla 3
Datos que cumplen el criterio de selección para imputación.

Jsc(mA/cm2)	Voc(V)	F.F.	PCE(%)
NaN	10,100	0,637	12,900
NaN	1,060	0,744	14,300
21,80	1,050	NaN	15,10
21,96	1,100	NaN	18,90
16,60	0,918	NaN	12,40
12,00	0,800	NaN	2,900
10,90	1,120	NaN	6,690
10,90	1,060	NaN	6,800
0,70	0,01	0,06	NaN

3.2. Preparación

En esta sección, se explicará el procedimiento para el entrenamiento y evaluación del modelo con el cual se realiza el trabajo de predicción de valores desconocidos. En orden de no extender mucho la explicación, se presenta el cálculo de una variable en particular, el procedimiento es fácilmente extensible para los otros casos.

3.2.1. Entrenamiento. Para el entrenamiento de modelos a partir de mezclas Gaussianas, primero se deben definir los hiperparámetros de las componentes del modelo, referidos como el número de mezclas y el tipo de matriz de covarianza a usar. Esto se logra tomando el conjunto de datos de entrenamiento y aplicando el criterio Bayesiano mediante la librería de *Sklearn*, la cuál ofrece una estimación óptima para el número de mezclas y el tipo de matriz de covarianza óptimo para el GMM. En la imagen 5 se presenta el criterio Bayesiano de uno de las 10 estimaciones realizadas para la variable PCE, el mismo cálculo se llevó a cabo para las demás variables, dónde los valores presentado en la tabla 4 son el promedio total de esa etapa del entrenamiento.

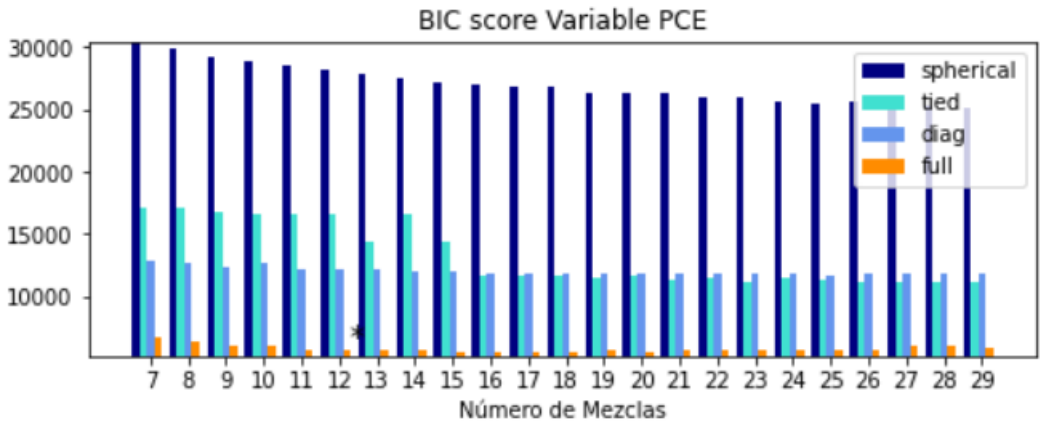


Figura 5. BIC por modelo para estimar cada variable.

Es común a todos los modelos que el tipo de matriz sea 'Full', esto se confirma con la teoría, ya que es el tipo de matriz que reúne los mejores atributos de los demás tipos para elección. Por otra parte, el número de mezclas si cambia y para el modelo final se toma el promedio total entre las variables del sistema.

Tabla 4
Promedio de Número de mezclas óptimo.

Modelo	Número de Mezclas
JSC	7
VOC	9
F.F	8
PCE	8
Promedio	8

3.2.2. Evaluación. Para la evaluación del modelo se busca medir la precisión y exactitud de las predicciones conseguidas ante un conjunto de datos que el modelo no conoce, los datos de prueba. En esta evaluación se hace uso del concepto de validación cruzada por divisiones k, en dónde definimos una división k de 10 para el conjunto de 2569 celdas solares, de las cuáles 2313 fueron para entrenamiento y 256 para evaluación. La validación cruzada es realizada paralelamente para las 4 variables por medio de la librería *Kfold* de *Sklearn*. Complementariamente, la evaluación de precisión se lleva a cabo por el estadístico R^2, con el que se comparan los resultados obtenidos por el modelo con los valores reales del conjunto de evaluación. Este método estadístico es calculado mediante la función *r*2*_score* de la librería *Sklearn* de *Python*. Si R^2 es igual a 1 implicaría una predicción perfecta, lo contrario sería el obtener un valor de 0. Dada la función de la librería *Sklearn* este valor podría ser negativo y esto se debe a que el modelo puede ser arbitrariamente peor. En la tabla 5 se observa el valor de precisión en cada una de las divisiones y en la tabla 6 se

muestra la descripción más a detalle de la tabla 5.

Tabla 5
Valores de R^2 para cada K.

Jsc(mA/cm2)	Voc(V)	F.F.	PCE(%)
0,729	0,278	0,180	0,894
0,573	-0,387	-0,323	0,792
0,633	0,648	0,311	0,848
0,801	0,401	0,494	0,913
0,609	0,284	0,340	0,846
0,388	0,338	-0,187	0,734
0,804	0,615	0,598	0,949
0,793	0,586	0,569	0,883
0,537	-0,084	0,099	0,882
0,594	0,253	0,233	0,821

Tabla 6
Descripción del estadístico R^2.

	Jsc(mA/cm2)	Voc(V)	F.F.	PCE(%)
total	10	10	10	10
media	0,646	0,293	0,232	0,856
std	0,135	0,322	0,305	0,062
min	0,388	-0,387	-0,323	0,734
max	0,804	0,648	0,597	0,949

3.3. Aplicación

Posteriormente al entrenamiento y evaluación se aplica dicho modelo para imputar los valores vacíos de la base de datos que cumplen con el criterio de selección. Tal y como se habló en la 3.1.1 el sistema es de 3 entradas y una salida. Para ello se hace uso de los datos de la tabla 3 donde se procesan los datos de acuerdo a la sección 3.1.2. Una vez el modelo identifica los datos de entrada procede a hacer la predicción. Esta predicción se puede ver más a detalle en la sección 4.3.

4. Resultados

En el capitulo anterior desarrollábamos la aplicabilidad de nuestra investigación y cómo llevarla a cabo para un conjunto de datos específico, ahora para esta sección entramos en detalle evaluando el desempeño del modelo calculado para cada una de las variables. El conjunto de evaluación

corresponde a un total de 256 celdas solares de Perovskita. Además se presentan y revisan los resultados obtenidos en la imputación de datos vacíos para los casos donde la perdida de información corresponde al 25 %. Los casos dónde la pérdida de datos es superior a este valor están por fuera del alcance de esta investigación.

4.1. Comparativas de evaluación

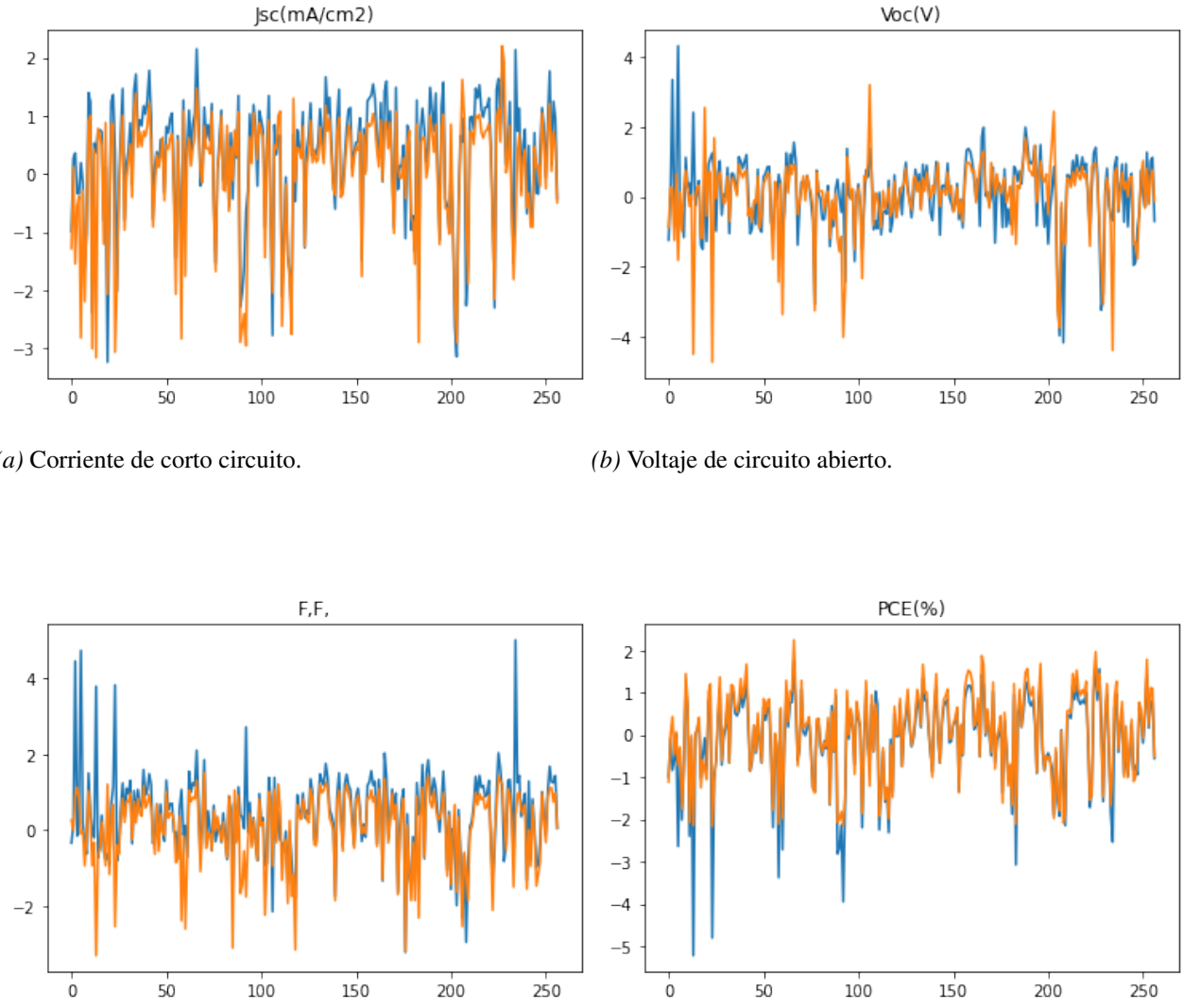
Para la evaluación de los modelos hallados se hace la comparativa entre el conjunto de parámetros reales y las estimaciones calculadas. Del conjunto de datos en general se tiene una precisión media superior al 50 % con respecto al conjunto de datos reales. El conjunto de datos que presenta menor rendimiento es el voltaje de corto circuito, se tiene la hipótesis qué esto sucede gracias a que es el valor más complejo de manipular. Adicionalmente, el factor de forma también presenta un bajo rendimiento, siendo el segundo con menor. Esto se atribuye que el factor de forma y el voltaje de corto circuito presentas una relación física directa.

Ahora, con respecto a los datos de la corriente de corto circuito y la eficiencia energética, la adaptación del algoritmo cubre la nube de datos con buena precisión, consiguiendo aciertos en el 60 % para el JSC y del 84 % para la PCE. Estos datos son resumidos en la tabla 7. Estos valores del estadístico R^2 son los valores más cercanos a la media de los resultados del mismo en la validación cruzada hecha en el capítulo 3.2.2

Tabla 7
Media del Estadístico R^2 para conjunto de 10 Pruebas.

	Jsc(mA/cm2)	Voc(V)	F,F,	PCE(%)
R^2Score	0,609	0,284	0,340	0,846

En la figura 6 se presenta este comportamiento por cada parámetro individual de la totalidad de los datos, donde las gráficas en azul corresponden a los valores estimados y las gráficas en naranja los valores reales.



(a) Corriente de corto circuito.

(b) Voltaje de circuito abierto.

(c) Factor de forma.

(d) Eficiencia de conversión de energía.

Figura 6. Comparativas de datos reales vs datos estimados

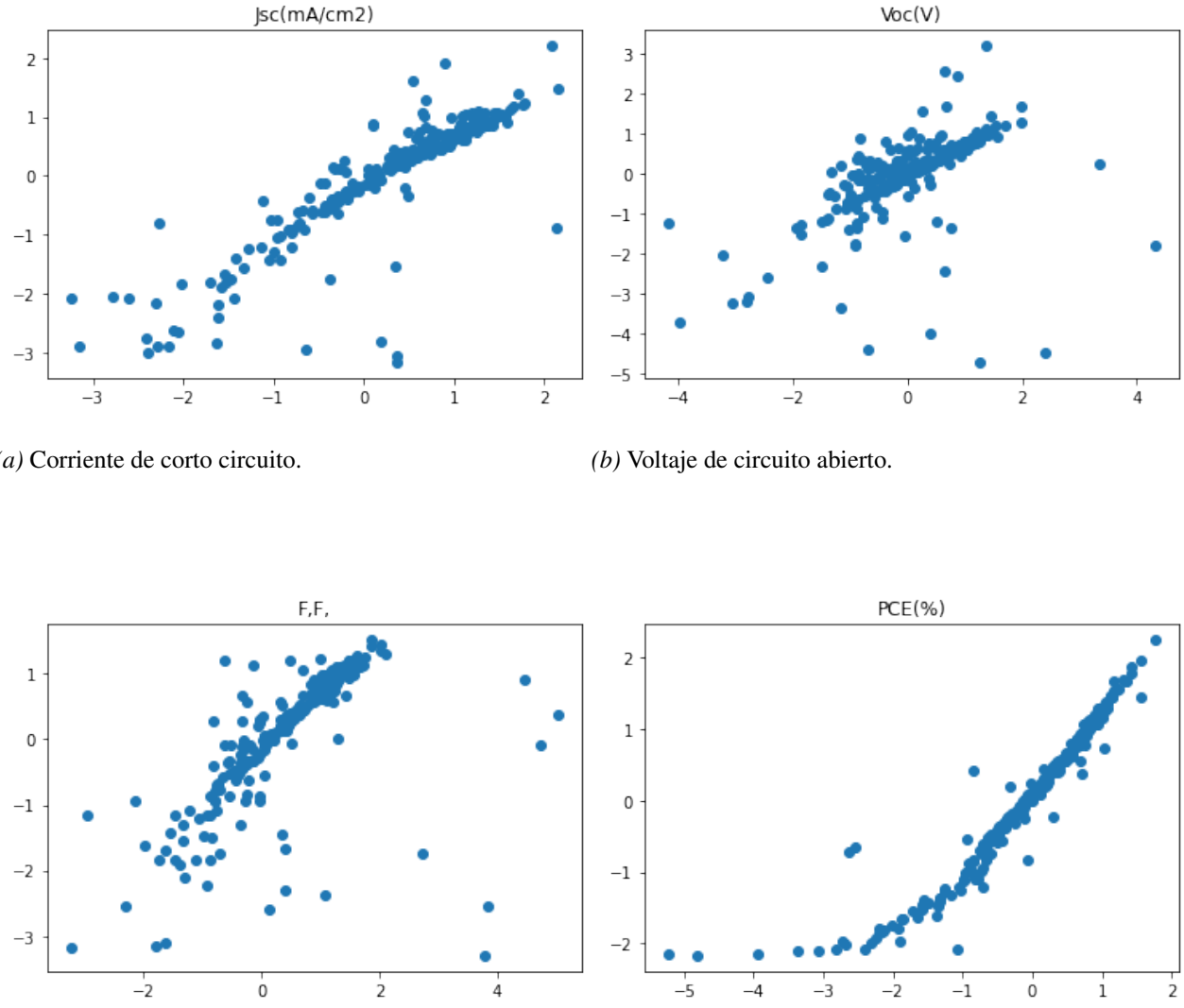
Por otro lado, en la tabla 8 se muestran 10 datos tomados aleatoriamente de la base de datos estimados de cada uno de los parámetros y se comparan con su respectivo valor real con el fin de validar la tendencia de los resultados anteriormente mostrados.

Tabla 8
Comparación de datos reales y estimados.

JSC		VOC		FF		PCE	
Datos reales	Datos estimados	Datos reales	Datos estimados	Datos reales	Datos estimados	Datos reales	Datos estimados
15	15,1525	0,829	0,8352	0,342	0,3673	4,26	4,2385
5,96	4,5791	0,855	0,8195	0,236	0,3452	1,2	1,4849
16,3	16,3133	0,68	0,7493	0,44	0,4923	5,23	5,0415
18,2	18,2255	1,04	1,0381	0,69	0,6922	13,1	13,0983
11,2	11,286	0,865	0,8737	0,61	0,6118	5,9	6,2758
14,13	14,1529	1,08	1,0045	0,58	0,5832	8,94	9,1459
5,89	5,8132	0,65	0,6644	0,63	0,5862	2,4	2,3484
10,03	10,0947	0,58	0,6245	0,58	0,5681	3,4	3,6082
18,3	18,2645	1	0,9988	0,793	0,7893	14,5	14,4984
15	15,0371	0,9	0,8979	0,58	0,5796	7,9	7,8939

4.2. Correlación de datos

Como medida adicional para evaluar el rendimiento del modelo, en la figura 7 se presenta la correlación entre los datos reales y las estimaciones. Teóricamente el máximo desempeño se representa con una linea recta de pendiente 1, en nuestro caso los resultados obtenidos permiten apreciar esta tendencia, especialmente en la corriente de corto circuito y la eficiencia energética. Si bien las gráficas de el factor de forma y el voltaje de corto circuito dan indicios de ser una recta, no lo son del todo confirmando los resultados mostrados en la tabla 7.



(a) Corriente de corto circuito.

(b) Voltaje de circuito abierto.

(c) Factor de forma.

(d) Eficiencia de conversión de energía.

Figura 7. Correlación de datos reales vs datos estimados

4.3. Imputación de datos faltantes

Usualmente las mezclas Gaussianas se emplean para el agrupamiento de conjuntos de datos, pero haciendo uso del algoritmo explicado en la sección 2.1.1 se pueden imputar datos de forma individual luego de entrenado un modelo que se acople a la nube de datos. A continuación en la tabla 9

presentamos los resultados obtenidos tras imputar los datos faltantes en la tabla 3:

Tabla 9
Imputación de datos faltantes.

Jsc(mA/cm2)	Voc(V)	F,F,	PCE(%)
11,219	10,100	0,637	12,900
19,008	1,060	0,744	14,300
21,80	1,050	0,658	15,10
21,96	1,100	0,782	18,90
21,96	1,100	0,782	18,90
12,00	0,800	0,327	2,900
10,90	1,120	0,620	6,690
10,90	1,060	0,657	6,800
0,70	0,01	0,06	0,255

El voltaje de corto circuito no es estimado en la fase de aplicación ya que en la base de

datos estudiada no existen conjuntos de datos que cumplan con el criterio de selección.

5. Conclusiones

En este trabajo de grado se consigue usa un modelo de mezclas Gaussianas para enriquecer los datos faltantes de un conjunto de datos sobre parámetros de rendimiento de celdas solares de Perovskitas. Investigando la literatura este trabajo al ser innovador no se pudo comparar con otros resultados obtenidos. Aún así, podemos afirmar, que este primer intento se ve prometedor.

Según la distribución de la información, al remover datos, es posible sesgar aún más (por parte del humano) los resultados, dado que al limitar la limpieza a rangos establecidos manualmente se pueden perder datos que no son atípicos y pueden aportar al entrenamiento del modelo.

El área no es tomada en cuenta como parámetro a estimar o influyente dentro de esta investigación, ya que este es seleccionado de manera arbitraria por los investigadores del campo solar alrededor de la Perovskita como material principal.

El voltaje de circuito abierto es la variable con mayor sesgo estadístico, dado que la diferencia entre magnitudes es significativa para el hallazgo de un modelo que estime esta variable. La tendencia encontrada se atribuye a que el VOC es una variable que al momento, es complicada de manipular y que presenta desafíos en la investigación de las celdas solares de Perovskita.

En la estimación del eficiencia energética se presenta mayor exactitud, lo anterior se atribuye a la distribución de los datos y la dependencia física con las demás variables del conjunto de datos investigado.

Por otra parte se tiene que el modelo responde con mejor desempeño cuando su matriz de covarianza es del tipo "full", dado que esta es cuadrada y simétrica donde sus dimensiones son iguales al número de variables.

6. Recomendaciones

Con el fin de maximizar la oportunidad de replica o mejora de los resultados obtenidos con este trabajo se recomienda tener en cuenta las siguientes observaciones prácticas:

- Hacer uso de un conjunto de datos con al menos 2500 instancias para lograr un correcto entrenamiento del modelo. Si dicho conjunto de datos es mayor la probabilidad de mejora en el entrenamiento aumenta considerablemente.

Referencias Bibliográficas

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., and Kochut, K. (2017). A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques.

Arora, M. and Kansal, V. (2019). Character level embedding with deep convolutional neural network for text normalization of unstructured data for Twitter sentiment analysis. *Social Network Analysis and Mining*, 9(1):0.

Bisquert, J. (2017). *The physics of solar cells: perovskites, organics, and photovoltaic fundamentals*. CRC press.

cheol Kim, M., Ham, S. Y., Cheng, D., Wynn, T. A., Jung, H. S., and Meng, Y. S. (2021). Advanced characterization techniques for overcoming challenges of perovskite solar cell materials.

CMIB (2020). Perovskite solar cells db.

Cole, J. M. (2020). A Design-to-Device Pipeline for Data-Driven Materials Discovery. *Accounts of Chemical Research*, 53(3):599–610.

Correa-Baena, J. P., Hippalgaonkar, K., van Duren, J., Jaffer, S., Chandrasekhar, V. R., Stevanovic, V., Wadia, C., Guha, S., and Buonassisi, T. (2018). Accelerating materials development via automation, machine learning, and high-performance computing.

Fonash, S. (2012). *Solar cell device physics*. Elsevier.

Jiang, D., Lin, W., and Raghavan, N. (2021). A gaussian mixture model clustering ensemble regressor for semiconductor manufacturing final test yield prediction. *IEEE Access*, 9:22253–22263.

Kim, E., Huang, K., Saunders, A., McCallum, A., Ceder, G., and Olivetti, E. (2017). Materials Synthesis Insights from Scientific Literature via Text Extraction and Machine Learning. *Chemistry of Materials*, 29(21):9436–9444.

Kim, H. S., Jang, I. H., Ahn, N., Choi, M., Guerrero, A., Bisquert, J., and Park, N. G. (2015). Control of i-v hysteresis in ch3nh3pbi3 perovskite solar cell. *Journal of Physical Chemistry Letters*, 6:4633–4639.

Kononova, O., Huo, H., He, T., Rong, Z., Botari, T., Sun, W., Tshitoyan, V., and Ceder, G. (2019). Text-mined dataset of inorganic materials synthesis recipes. *Scientific Data*, 6(1):1–11.

Lee, D. and Kim, J. K. (2022). Semiparametric imputation using conditional gaussian mixture models under item nonresponse. *Biometrics*, 78(1):227–237.

Limeres, C. C. (2011). RegresiÓn lineal simple. In *Universidad de Santiago de Compostela*.

Mesquita, D., Gomes, J., and Rodriguez, L. (2019). Artificial neural networks with random weights for incomplete datasets. *Neural Processing Letters*, 50:2345.

NREL (2022). *National renewable energy laboratory transforming energy*.

Olivetti, E., Cole, J., Kim, E., Kononova, O., Ceder, G., Han, T., and Hiszpanski, A. (2020). Data-

driven materials research enabled by natural language processing and information extraction. *Applied Physics Reviews*, 7:041317.

Pilania, G., Balachandran, P. V., Kim, C., and Lookman, T. (2016). Finding new perovskite halides via machine learning. *Frontiers in Materials*, 3:1–7.

Raschka, S. (2018). Machine learning lecture notes. In *University of Wisconsin–Madison*.

Sahoo, S. K., Manoharan, B., and Sivakumar, N. (2018). Chapter 1 - introduction: Why perovskite and perovskite solar cells? In Thomas, S. and Thankappan, A., editors, *Perovskite Photovoltaics*, pages 1–24. Academic Press.

Sang, H., Kim, J. K., and Lee, D. (2020). Semiparametric fractional imputation using gaussian mixture models for handling multivariate missing data. *Journal of the American Statistical Association*, 0(0):1–10.

Sepúlveda, A. (2020). Gaussian mixture regression. In *Universidad Industrial de Santander*.

Shetty, P. and Ramprasad, R. (2021). Automated knowledge extraction from polymer literature using natural language processing. *iScience*, 24(1):101922.

Stanley, J. and Gagliardi, A. (2019). Machine learning bandgaps of inorganic mixed halide perovskites. volume 2018-July.

Swain, M. C. and Cole, J. M. (2016). ChemDataExtractor: A Toolkit for Automated Extraction

of Chemical Information from the Scientific Literature. *Journal of Chemical Information and Modeling*, 56(10):1894–1904.

Xue, Y., Klabjan, D., and Luo, Y. (2019). Mixture-based multiple imputation model for clinical data with a temporal dimension. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 245–252.

Yılmaz, B. and Yıldırım, R. (2021). Critical review of machine learning applications in perovskite solar research. *Nano Energy*, 80:105546.

Çağla Odabaşı and Yıldırım, R. (2020). Machine learning analysis on stability of perovskite solar cells. *Solar Energy Materials and Solar Cells*, 205:110284.