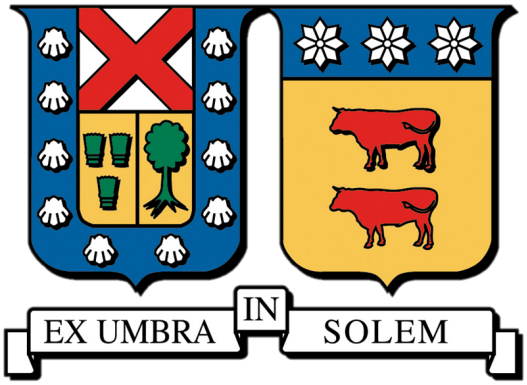
UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA

DEPARTAMENTO DE INGENIERÍA QUÍMICA Y AMBIENTAL



PREDICCIÓN DEL DESEMPEÑO DE PILA DE COMBUSTIBLE DE HIDRÓGENO VERDE EN LA REGIÓN DE ANTOFAGASTA MEDIANTE MODELOS DE APRENDIZAJE AUTOMÁTICO

FRANCISCO HENRÍQUEZ VIZCARRA

TESIS PARA OPTAR AL TÍTULO DE MAGÍSTER

EN CIENCIAS DE LA INGENIERÍA QUÍMICA

PROFESOR/A GUÍA: IVÁN CORNEJO GARCÍA

DICIEMBRE-2025

CONSTANCIA DE VALIDACIÓN Y CONFIDENCIALIDAD DE MONOGRAFÍA A REPOSITORIO ACADÉMICO

1.- IDENTIFICACIÓN DEL TRABAJO ACADÉMICO

Tipo de monografía (marcar una opción): ☐ Memoria o trabajo de título ☑ Tesis de Postgrado

Título del trabajo: Predicción del desempeño de pila de combustible de hidrógeno verde en la Región de Antofagasta mediante modelos de Aprendizaje Automático

Nombre del candidato(a): Francisco Vidal Henríquez Vizcarra

Carrera / Grado: Magíster en Ciencias de la Ingeniería Química

Campus: Casa Central Valparaíso **Departamento:** de Ingeniería Química y Ambiental

2.- VALIDACIÓN DEL PROFESOR GUÍA/DIRECTOR DE TESIS

Yo, Iván Andrés Cornejo García, en mi calidad de profesor(a) guía/director(a) del trabajo académico mencionado anteriormente **DEJO CONSTANCIA** que:

- He revisado esta versión del documento y corresponde a la versión final aprobada del trabajo.
- El trabajo cumple con los requisitos académicos y de formato establecidos por la institución.

3.- EVALUACIÓN DE CONFIDENCIALIDAD POR PROPIEDAD INDUSTRIAL (marcar una opción)

☑ El trabajo **NO contiene** información que amerite confidencialidad y puede ser publicado de inmediato en repositorio con acceso abierto.

☐ El trabajo **CONTIENE** información con potenciales implicancias de propiedad industrial o intelectual y requiere un periodo de confidencialidad (**embargo**) por **(marcar una opción)**:

☐ 6 meses ☐ 12 meses ☐ 2 años ☐ 3 años ☐ 5 años ☐ 10 años

Fundamentación de la necesidad de confidencialidad (obligatorio si se solicita embargo):

4.- FIRMAS

Profesor(a) guía o director(a) de memoria o tesis:

Fecha: 05-04-2026 **Firma:** 

Estudiante o Candidato(a):

Fecha: 02.04.2026 **Firma:** 

Este formulario debe ser insertado como página 2 de la memoria o tesis, completado y firmado por estudiante y profesor(a) antes de la entrega en portal PRISMA de Biblioteca USM.

Resumen

El despliegue de tecnologías de hidrógeno verde en la Región de Antofagasta requiere herramientas predictivas que permitan evaluar el desempeño de bancos de pilas de combustible de membrana de intercambio protónico (PEMFC) en un entorno con alta heterogeneidad climática y topográfica. En este contexto, se desarrolla un *pipeline* de modelado basado en Aprendizaje Automático para predecir con alta precisión y eficiencia la potencia eléctrica de un banco de tres pilas GenSure E-1100 de la planta piloto móvil de hidrógeno verde de CICITEM, desplegada desde la Cordillera de la Costa hasta la Precordillera y la Depresión Intermedia. La base de datos utilizada está compuesta por 1595 registros de variables ambientales (T_{amb}, p_{amb}, HR) y eléctricas (I, V, W) recopilados en distintos sitios, conformando un conjunto de tamaño acotado, pero altamente heterogéneo y multimodal, con desbalances en los estratos climático–operacionales y una marcada presencia de *outliers*, lo que exige esquemas de modelado robustos y coherentes con la física del sistema.

El flujo de trabajo integra tres componentes principales: (i) ingeniería de características acoplada a ablación incremental para construir una base de datos enriquecida con descriptores climáticos, temporales, operacionales y de reparto de combustible; (ii) una metodología probabilística de detección de *outliers* por familias de variables (FASEK5), que combina múltiples detectores en un esquema *leaky noisy-OR* para generar versiones depuradas de la base preservando su representatividad física y operativa; y (iii) la implementación de un *pipeline* robusto y reproducible para entrenar y comparar modelos de regresión (`RandomForestRegressor`, `XGBoost`, `CatBoost`, `MLPRegressor` y `SupportVectorRegressor`) con validación cruzada estratificada y optimización bayesiana de hiperparámetros. La calidad de los modelos se evalúa mediante un conjunto integrado de métricas de ajuste, sesgo y dispersión de residuos (RMSE, r, R^2, CCC, MEC, ME, SDE, Std), complementado con diagramas solares y de Taylor, además de métricas de generalización y ganancia relativa asociada a la depuración de la base de datos (GAP_{RMSE}, ΔRMSE, $Gain_{rel}$).

Los resultados muestran que la depuración moderada de *outliers* ($< 5\,\%$ de contaminación, < 80 registros) incrementa de forma sistemática la capacidad de generalización de todos los modelos, con reducciones de la diferencia entre el RMSE de entrenamiento y de prueba (GAP_{RMSE}) del orden de 4.6–9.0 W y ganancias relativas del RMSE de prueba en torno a 21–31 %, desplazando sus posiciones en los diagramas solar y de Taylor hacia regiones de mayor calidad. Bajo este marco

metodológico, la configuración XGB_v0.9950 emerge como el modelo con mejor compromiso entre precisión, robustez frente a *outliers* y capacidad de generalización, alcanzando un RMSE global $\approx$ 31 W ($\approx$ 11 % de la desviación estándar experimental, $\sigma \approx 275.6$ W), sesgo prácticamente nulo y valores de r ≈ 0.993 y $R^2 \approx 0.987$, con coeficientes de eficiencia de modelado (MEC) y de concordancia (CCC) del orden de 0.990. El análisis de importancia de características en los modelos basados en árboles indica que el comportamiento del banco está gobernado principalmente por el régimen de activación, la configuración interna y el reparto de combustible entre pilas y, en segundo lugar, por las condiciones ambientales y el contexto temporal, en concordancia con la influencia sobre el desempeño electroquímico predicha por modelos analíticos.

Este trabajo de tesis valida que un enfoque de Aprendizaje Automático integrado con ingeniería de características coherente con la física del sistema PEMFC y el contexto operativo local, junto con una etapa de depuración probabilística, constituye una herramienta robusta para predecir el desempeño de bancos PEMFC en campo. En consecuencia, el *pipeline* propuesto se configura como un marco reproducible extrapolable a otros bancos y contextos climáticos y sienta las bases para futuros trabajos sobre control, optimización e interpretabilidad de sistemas PEMFC bajo condiciones reales de operación.

Abstract

The deployment of green hydrogen technologies in the Antofagasta Region requires predictive tools that enable the assessment of the performance of proton exchange membrane fuel cells (PEMFC) in an environment with high climatic and topographic heterogeneity. In this context, a Machine Learning based modeling pipeline is developed to predict, with high accuracy and efficiency, the electrical power of a three cells GenSure E-1100 stack belonging to CICITEM's mobile green-hydrogen pilot plant, deployed from the Cordillera de la Costa to the Precordillera and the Depresión Intermedia. The database used is composed of 1595 records of environmental variables (T_{amb}, p_{amb}, HR) and electrical variables (I, V, W) collected at different sites, forming a dataset of limited size but highly heterogeneous and multimodal, with imbalances across climatic–operational strata and a marked presence of outliers, which demands robust modeling schemes consistent with the physics of the system.

The workflow integrates three main components: (i) feature engineering coupled with incremental ablation to build a database enriched with climatic, temporal, operational, and fuel-distribution descriptors; (ii) a probabilistic outlier-detection methodology by variable families (FASEK5), which combines multiple detectors in a leaky noisy-OR scheme to generate cleaned versions of the database while preserving its physical and operational representativeness; and (iii) the implementation of a robust and reproducible pipeline to train and compare regression models (`RandomForestRegressor`, `XGBoost`, `CatBoost`, `MLPRegressor`, and `SupportVectorRegressor`) with stratified cross-validation and bayesian hyperparameter optimization. Model quality is evaluated using an integrated set of goodness-of-fit, bias, and residual-dispersion metrics (RMSE, r, R^2, CCC, MEC, ME, SDE, Std), complemented with solar and Taylor diagrams, as well as generalization metrics and relative gain associated with database cleaning (GAP_{RMSE}, ΔRMSE, $Gain_{rel}$).

The results show that moderate outlier removal ($< 5\,\%$ contamination, < 80 records) systematically increases the generalization capacity of all models, with reductions in the gap between training and test RMSE (GAP_{RMSE}) on the order of 4.6–9.0 W and relative gains in test RMSE of about 21–31 %, shifting their positions in the solar and Taylor diagrams toward higher-quality regions. Within this methodological framework, the XGB_v0.9950 configuration emerges as the model with the best compromise between accuracy, robustness to outliers, and generalization capacity, achieving

a global RMSE $\approx$ 31 W ($\approx$ 11 % of the experimental standard deviation, $\sigma \approx 275.6$ W), practically zero bias, and values of r ≈ 0.993 and $R^2 \approx 0.987$, with modeling efficiency (MEC) and concordance (CCC) coefficients on the order of 0.990. The feature-importance analysis in tree-based models indicates that stack behavior is governed primarily by the activation regime, internal configuration, and fuel distribution among cells and, secondarily, by environmental conditions and temporal context, in agreement with the influence on electrochemical performance predicted by analytical models.

This thesis validates that a Machine Learning approach integrated with feature engineering consistent with PEMFC system physics and the local operating context, together with a probabilistic cleaning stage, constitutes a robust tool for predicting the performance of PEMFC stacks in the field. Consequently, the proposed pipeline is configured as a reproducible framework that is extrapolable to other stacks and climatic contexts and lays the groundwork for future work on control, optimization, and interpretability of PEMFC systems under real operating conditions.

Agradecimientos

A mi profesor guía, Dr. Iván Cornejo, por su paciencia y el apoyo moral brindado durante el programa de magíster. Las circunstancias no nos permitieron conocernos con la profundidad que hubiera deseado, pero sí lo suficiente para admirar su trabajo. Confío en que la vida le retribuirá su calidad humana y profesional.

Al Dr. Lindley Maxwell, por brindarme la oportunidad de colaborar en CICITEM y permitirme participar en el diseño y la puesta en marcha de la planta piloto móvil de hidrógeno verde. Este trabajo no habría sido posible sin la oportunidad que me dio y la confianza con la que me incentivó a emprender este programa.

A mi madre, cuyo esfuerzo cotidiano hizo posible que cursara la carrera de Ingeniería Civil Química y que sigue siendo mi principal referente para expandirme en la vida. Su resiliencia emocional y templanza continúan siendo luz. Me has dado la vida y espero honrar tu sacrificio con cada logro, de forma virtuosa. Eres la persona que más amo en este mundo.

Asimismo, agradezco a todo el equipo humano que sostiene el nivel de excelencia de la Universidad Técnica Federico Santa María. Me ha impresionado su estándar ético y profesional y el compromiso con que ejercen su labor formadora.

Agradezco haber descubierto la natación, disciplina que transformó mi vida al permitirme superar una fobia arraigada desde la niñez; al deporte, por ser prueba de que la transformación del espíritu también puede nacer de cultivar el cuerpo y utilizar la energía de forma consciente a diario; y a la meditación, por ayudarme a encontrar calma y paz interior en medio de las adversidades.

Finalmente, quiero agradecerme a mí mismo. La vida me ha puesto frente a pruebas complejas que en más de una ocasión me han hecho caer. Agradezco el aprendizaje de cada experiencia: son testimonio de una fortaleza interior que desconocía y maestras de la serenidad que mi alma necesitaba. Independientemente del desenlace formal de este proceso académico, este trabajo es prueba de que no me abandoné. La persistencia y la acción constante y consciente seguirán siendo mis guías espirituales.

Quiero dedicar este trabajo a la memoria de mi tía Tona. Su valiente batalla contra el cáncer me enseñó que la perseverancia trasciende el dolor y otorgó un nuevo significado a mi propósito y a mi sentido de vida.

Tabla de Contenidos

Lista de Tablas

Lista de Figuras

Nomenclatura

Símbolo	Significado, unidades
$a_{K,eff}$	Exponente efectivo de la familia K en el modelo FASEK5, que pondera su contribución al puntaje probabilístico global; adimensional.
$a_{K,max}$	Cota superior del exponente efectivo de la familia K en FASEK5; adimensional.
$a_{K,min}$	Cota inferior del exponente efectivo de la familia K en FASEK5; adimensional.
C	Hiperparámetro de complejidad de `SVR`; adimensional.
CCC	Coeficiente de Correlación de Concordancia; adimensional.
c_K	Coeficiente base de ponderación asignado a la familia K en la metodología FASEK5; adimensional.
CO_2	Dióxido de carbono; especie química asociada a emisiones de gases de efecto invernadero.
CO_2-eq	Dióxido de carbono equivalente; unidad utilizada para cuantificar emisiones de GEI en términos equivalentes de CO_2 (kg CO_2-eq, t CO_2-eq).
D	(i) Conjunto de datos supervisados $D = \{(\mathbf{x}_i, y_i)\}$ utilizado para entrenar y evaluar modelos; (ii) número de componentes de una composición $\mathbf{p} = (p_1, \ldots, p_D)$; adimensional; (iii) dimensión.
Day_cos	Componente cosenoidal de la codificación cíclica del día del año. Cumple la relación: $\text{Day\_cos} = \cos(2\pi d/m)$.
Day_sin	Componente sinusoidal de la codificación cíclica del día del año. Cumple la relación: $\text{Day\_sin} = \sin(2\pi d/m)$.
e	(i) Estrato operativo basal definido como e = Sitio × K_act × Config; índice categórico, adimensional; (ii) error absoluto de potencia eléctrica utilizado en EDA; W.
e_i	Errores o residuos de regresión entre observaciones y predicciones, $e_i = y_i - \hat{y}_i$; W.
E_{rev}	Potencial reversible ideal (termodinámico) de la celda PEM; V.
F_{iK}	Probabilidad de que el registro i sea *outlier* según la familia de detección K en FASEK5; adimensional.
$F_M(x)$	Modelo de *gradient boosting* como suma de M árboles de regresión, definido como $F_M(x) = \sum_{m=1}^{M} f_m(x)$; W.
f	Función objetivo o modelo de regresión que aproxima la relación entrada–salida del sistema; adimensional (función).

Símbolo	Significado, unidades
$\mathrm{f}(x)$	Modelo de regresión evaluado para el vector de entrada x; W.
$\mathrm{f}_m(x)$	Árbol de regresión individual dentro del *ensemble* de *gradient boosting*; W.
$\mathrm{f}_{\boldsymbol{\theta}}$	Familia de modelos de regresión parametrizados por el vector de parámetros $\boldsymbol{\theta}$; adimensional (función).
$\mathrm{GAP}_{\mathrm{RMSE}}$	Brecha de generalización $\mathrm{GAP}_{\mathrm{RMSE}} = \mathrm{RMSE}_{\mathrm{test}} - \mathrm{RMSE}_{\mathrm{train}}$; W.
$\mathrm{Gain}_{\mathrm{rel}}$	Ganancia relativa porcentual en $\mathrm{RMSE}_{\mathrm{test}}$ tras la depuración de *outliers*; %.
H	Entropía de Shannon de una composición $\mathbf{p}$; adimensional.
H_2	Hidrógeno molecular; vector energético utilizado como combustible.
$\mathrm{H}_2, \mathrm{H}_3$	Predictores continuos de entropía de regímenes bi y tricelda; $\mathrm{H_i} = \mathrm{K_i} \cdot \mathrm{H_{bal}}$; adimensional.
$\mathrm{H_2O}$	Agua producida en la reacción global de la celda PEM.
$\mathrm{H}_{\mathrm{active}}$	Entropía balanceada del reparto de corriente entre las PEMFC activas; adimensional.
$\mathrm{H}_{\mathrm{bal}}$	Entropía de Shannon normalizada, $\mathrm{H_{bal}} = \mathrm{H}/\ln(\mathrm{D})$, con valores entre 0 y 1; adimensional.
HR	Humedad relativa del aire ambiente; %.
h	Vector de hiperparámetros de un modelo, usado por la optimización bayesiana; adimensional.
h_{K}	Coeficiente de impacto de la familia K sobre el desempeño de los modelos (variación relativa de RMSE); adimensional.
I	Corriente eléctrica de la celda o del *stack* PEM; A.
$\mathrm{I}_1, \mathrm{I}_2, \mathrm{I}_3$	Corriente eléctrica de las pilas PEM FC_1, FC_2, FC_3; A.
IQR	Rango intercuartílico, $\mathrm{IQR} = \mathrm{Q}_3 - \mathrm{Q}_1$, usado en detección univariante de *outliers*; unidades de la variable analizada.
J	Función objetivo de *k-means* (distorsión intra–clúster).
J_{K}	Coeficiente de Jaccard del método de detección de la familia K; adimensional.
K	(i) Número de particiones en validación cruzada K-*fold*; (ii) número de clústeres en *k-means*; adimensional.
$\mathrm{K}_{\mathrm{act}}$	Número de pilas de combustible activas en el banco (regímenes K1, K2, K3); adimensional.
k	Hiperparámetro de *clustering* que hace referencia al *elbow method* (*k-distance graph*).
L_0	Parámetro de fuga global del modelo FASEK5; adimensional.
$\mathcal{L}_{\mathrm{reg}}(\boldsymbol{\theta})$	Pérdida regularizada que combina pérdida empírica y término de penalización $\lambda\Omega(\boldsymbol{\theta})$; unidades de la función de pérdida (por ejemplo, W^2).
l_{K}	Parámetro de fuga (*leakage*) de la familia K; probabilidad marginal de que un registro sea anómalo aunque ningún método de la familia lo detecte.
M	Dimensión del espacio de características en ciertos contextos (número de variables usadas, por ejemplo, en *k-means*); adimensional.
m	Período del ciclo anual en la codificación temporal; en este trabajo $\mathrm{m} = 365$; d.

Símbolo	Significado, unidades
ME	Error medio (sesgo) del modelo; W.
MEC	Coeficiente de Eficiencia de Modelado; adimensional.
MinPts	Hiperparámetro de DBSCAN: número mínimo de puntos requeridos en el vecindario ε; adimensional.
N	Número total de observaciones del conjunto de datos; adimensional.
n	Número de observaciones empleadas en el cálculo de métricas; adimensional.
n_K	Tamaño muestral efectivo de los estratos de la familia K en FASEK5; adimensional.
O_2	Oxígeno molecular; especie presente en el aire de alimentación al cátodo.
P	Potencia eléctrica instantánea de la celda o del *stack* PEM, típicamente $P = I\,V$; W.
p	(i) Dimensión del vector de características $\mathbf{x} \in \mathbb{R}^p$; (ii) vector composicional $\mathbf{p} = (p_1, \ldots, p_D)$ que representa repartos (por ejemplo, fracciones de corriente); adimensional.
p_{amb}	Presión ambiente en el sitio de operación de la planta piloto; bar.
p_d	d-ésima componente del vector composicional $\mathbf{p}$; adimensional.
$p_{m\|K}$	Probabilidad (peso) de eficacia de detección asignada al método m dentro de la familia K.
Q	Estadístico Q o SPE (*Squared Prediction Error*) asociado al residuo de reconstrucción en PCA.
Q_1, Q_2, Q_3	Primer, segundo y tercer cuartil de la distribución de una variable.
r	Coeficiente de correlación de Pearson entre observaciones y predicciones; adimensional.
r_K	Factor de confiabilidad del tamaño muestral de la familia K, cumple la relación: $r_K = n_K/(n_K + T)$; adimensional.
R^2	Coeficiente de determinación; proporción de la varianza de y explicada por el modelo; adimensional.
RMSE	Raíz del error cuadrático medio; W.
S	(i) Suma de corrientes de las pilas activas en el banco PEMFC, $S = I_1+I_2+I_3$; A. (ii) En *k-means*, conjunto de clústeres $S = \{S_1, \ldots, S_K\}$; adimensional.
SDE	Desviación estándar de los residuos; W.
SPE	*Squared Prediction Error*; residuo de reconstrucción en PCA.
Std	Desviación estándar muestral de una serie (por ejemplo, de la potencia observada); W.
S_i	Puntaje probabilístico global de FASEK5 para el registro i; adimensional.
S_k	k-ésimo clúster en *k-means*; adimensional.
T	(i) Temperatura (en contextos generales); °C. (ii) Parámetro de tamaño de referencia utilizado en la definición de r_K en FASEK5; adimensional.
T^2	Estadístico de Hotelling T^2 que mide la distancia de una observación a la media en el subespacio principal de PCA; adimensional.

Símbolo	Significado, unidades
T_{amb}	Temperatura ambiente en el sitio de operación de la planta piloto; °C.
V	Tensión eléctrica de salida de la celda o del *stack* PEM; V.
V_1, V_2, V_3	Tensión eléctrica de las pilas PEM FC_1, FC_2, FC_3; V.
V_{cell}	Tensión eléctrica real de una celda individual de la pila PEM, incluyendo pérdidas por activación, óhmicas y de concentración; V.
W	Potencia eléctrica del *stack* PEMFC (variable objetivo del problema de regresión); W.
W_1, W_2, W_3	Potencia eléctrica de las pilas PEM FC_1, FC_2, FC_3; W.
w_K	Peso compuesto de la familia K en FASEK5, $w_K = c_K^{1.25} r_K^{1.50} h_K^{2.50}$; adimensional.
$\tilde{w_K}$	Mediana de los pesos w_K, usada en la normalización del exponente $a_{K,eff}$; adimensional.
X	(i) Fracción de corriente asociada a la pila *i*-ésima: $X_i = I_i / \sum_j I_j$; (ii) vector de nodos padre binarios $X = (X_1, \ldots, X_m)$ en el modelo *(leaky) noisy-OR*.
X_K	Nodo padre binario individual en el modelo *noisy-OR* que indica si la familia K marca un registro como sospechoso; adimensional.
X_L	Nodo de fuga (*leak*) en el modelo *leaky noisy-OR*, que representa causas no modeladas; adimensional.
$\mathbf{x}$	Vector de características de entrada que describe el estado del sistema (variables ambientales, operativas, etc.); cada componente con sus propias unidades.
Y	Nodo hijo binario en el modelo *(leaky) noisy-OR* (indicador de "el registro es *outlier*"); adimensional.
y	Variable objetivo escalar (por ejemplo, potencia eléctrica); W.
y_i	Observación i-ésima de la variable objetivo; W.
$\hat{y}_i$	Predicción del modelo para la i-ésima observación; W.
$\bar{y}$	Media muestral de las observaciones y_i; W.
$\bar{\hat{y}}$	Media muestral de las predicciones $\hat{y}_i$; W.
Z	Coordenada ILR entre dos grupos de componentes de una composición; adimensional.
Z_1	Primera coordenada ILR del reparto de corrientes entre las pilas del banco PEMFC en régimen tricelda (K3); adimensional.
Z_2	Segunda coordenada ILR del reparto de corrientes entre las pilas del banco PEMFC en régimen tricelda (K3); adimensional.
$z_{i,m}$	Indicador binario de detección del método m sobre el registro *i* en FASEK5 ($z_{i,m} = 1$ si el método marca *outlier*); adimensional.

Símbolo	Significado, unidades
Δ	Variación; se define ganancia absoluta de RMSE (ΔRMSE), métrica de generalización.
ε	(i) Margen de insensibilidad de `SVR` (*epsilon*); (ii) radio de vecindad `eps` en DBSCAN; adimensional.
γ	(i) Parámetro del *kernel* RBF en `SVR`; (ii) parámetro de forma en la interpolación del exponente $a_{K,eff}$ en FASEK5; adimensional.
η_{act}	Sobrepotencial de activación en la celda PEM; V.
η_{conc}	Sobrepotencial de concentración asociado a limitaciones de transporte de masa en la celda PEM; V.
η_{ohm}	Sobrepotencial óhmico asociado a resistencias internas de la celda PEM; V.
λ	Hiperparámetro de regularización que pondera el término de penalización $\Omega(\boldsymbol{\theta})$ en la pérdida regularizada; en FASEK5, parámetro de calibración en la definición de ρ_K; adimensional.
$\ell(y, \hat{y})$	Función de pérdida que cuantifica la discrepancia entre valor observado y valor predicho (por ejemplo, pérdida cuadrática); unidades determinadas por la métrica (por ejemplo, W^2 para el error cuadrático medio).
ρ_K	Factor de redundancia de la familia K, definido a partir del coeficiente de Jaccard promedio, que atenúa la contribución de métodos solapados; adimensional.
σ_y	Desviación estándar muestral de las observaciones y_i; W.
$\sigma_{\hat{y}}$	Desviación estándar muestral de las predicciones $\hat{y}_i$; W.
$\boldsymbol{\theta}$	Vector de parámetros del modelo de regresión (por ejemplo, pesos de una red neuronal); adimensional.
θ_k	Parámetro de enlace del modelo *noisy-OR* para el padre k, $\theta_k = P(Y = 1 \mid X_k = 1, X_j = 0, j \neq k)$; adimensional.
θ_L	Probabilidad de fuga en el modelo *leaky noisy-OR*; adimensional.
$\Omega(\boldsymbol{\theta})$	Término de penalización o regularización aplicado a los parámetros del modelo (por ejemplo, normas L1 o L2); unidades compatibles con la pérdida.
Subíndices	
clean	Versión de la base de datos depurada de *outliers*.
CV	Valor medio de validación cruzada, utilizado para RMSE de K bloques (optimización bayesiana).
out	Versión de la base de datos con *outliers* (sin depuración).
test	Conjunto de prueba.
train	Conjunto de entrenamiento.
Superíndices	
$*$	Versión normalizada por la desviación estándar experimental de las métricas ME, SDE, RMSE y σ.

Sigla	Significado
0D, 1D, 2D, 3D	Modelos cero, uni, bi y tridimensionales utilizados en el modelado físico de PEMFC.
AEM	*Anion Exchange Membrane*; membrana de intercambio aniónico utilizada en los electrolizadores.
AI	*Artificial Intelligence*; Inteligencia Artificial.
Amb	Grupo de características ambientales (temperatura, presión y humedad relativa).
ANN	*Artificial Neural Network*; red neuronal artificial.
Baseline	Versión base de la base de datos compuesta por el conjunto de características que entrega un nivel de ajuste considerado aceptable en los ensayos preliminares (antes de aplicar depuración).
BD	Base de datos.
BiLSTM	*Bidirectional Long Short-Term Memory*; arquitectura de red neuronal recurrente para pronóstico de series temporales.
BO	*Bayesian Optimization*; optimización bayesiana de hiperparámetros.
CAT	Abreviatura utilizada para el modelo *CatBoost*, clase `CatBoostRegressor`.
CFD	*Computational Fluid Dynamics*; Dinámica de Fluidos Computacional.
CICITEM	Centro Científico Tecnológico de la Región de Antofagasta.
Cluster_ID	Etiqueta categórica del clúster climato–topográfico asignado a cada observación.
Cod_Season	Codificación categórica/*One-Hot* de la estación del año.
Config	Configuración del banco PEMFC (pilas encendidas/apagadas: ‘100’, ‘010’, ‘001’, ‘110’, ‘101’, ‘011’, ‘111’).
CPU	*Central Processing Unit*; unidad central de procesamiento; en el texto se usa para tiempos de cómputo (horas de CPU).
DBSCAN	*Density-Based Spatial Clustering of Applications with Noise*; algoritmo de *clustering* basado en densidad.
EDA	*Exploratory Data Analysis*; Análisis Exploratorio de Datos.
ERNC	Energías Renovables No Convencionales.
EZ-01–EZ-08	Banco de electrolizadores de membrana de intercambio aniónico (AEM) del sistema de producción de H_2 de la planta piloto.
F1–F5	Familias de detectores de *outliers*: F1 (univariantes V e I), F2 (corrientes y repartos por K_{act} y Config), F3 (curva I–V), F4 (climático–eléctrica multidimensional), F5 (potencia W).
FASEK5	Metodología de puntaje probabilístico para detección de *outliers* basada en cinco familias (F1–F5) y modelo *leaky noisy-OR*.
FC_ID	Identificador de la pila de combustible individual FC_i dentro del banco PEMFC con $i \in \{1, 2, 3\}$; variable categórica.
FC-01–FC-03	Pilas de combustible individuales del banco PEMFC de la planta piloto.
GEI	Gases de efecto invernadero.
H2, H3	Predictores continuo de entropía balanceada asociado a régimenes bi y tricelda.
HP	Hiperparámetros de modelos de regresión.

Sigla	Significado
HPS-01, HPS-02	Purificadores de hidrógeno encargados de secar y acondicionar la corriente de H_2 antes de su almacenamiento o uso.
ILR	*Isometric Log-Ratio*; transformación *log-ratio* isométrica para datos composicionales.
IQR	*Interquartile Range*; rango intercuartílico utilizado como criterio para la detección de valores atípicos en una dimensión.
k-means	Algoritmo de agrupamiento no supervisado (*k-means clustering*) aplicado a las variables ambientales.
LOF	*Local Outlier Factor*; algoritmo de detección de valores atípicos basado en densidad local.
MATLAB	Entorno de cálculo numérico y programación utilizado para los ensayos preliminares de redes neuronales (*Neural Net Fitting*).
MEA	*Membrane Electrode Assembly*; conjunto membrana–electrodo de una celda PEM.
ML	*Machine Learning*; Aprendizaje Automático.
MLP	*Multilayer Perceptron*; red neuronal de tipo perceptrón multicapa (*feedforward*), clase `MLPRegressor`.
Modo	Modo de operación eléctrica del banco (Float, Maintain); categórico, adimensional.
NFL	*No Free Lunch*; teoremas de optimización y aprendizaje supervisado.
Output	Valores predichos por el modelo (potencia estimada del banco PEMFC); W.
PC1, PC2, PC3	Primeras tres componentes de PCA.
PCA	*Principal Component Analysis*; Análisis de Componentes Principales.
P^3H_2V	Planta piloto móvil de hidrógeno verde de CICITEM.
PEM	*Proton Exchange Membrane*; membrana de intercambio protónico.
PEMFC	*Proton Exchange Membrane Fuel Cell*; pila de combustible de membrana de intercambio protónico.
PSDA	Sitio de medición PSDA correspondiente a uno de los emplazamientos de la campaña itinerante de la planta piloto móvil.
RBF	*Radial Basis Function*; tipo de *kernel* utilizado en SVR.
RFR	Clase `RandomForestRegressor`; implementación de regresión basada en bosques aleatorios de *scikit-learn* (Python).
Season	Estación del año (verano, otoño, invierno, primavera); variable categórica, adimensional.
Session_ID	Identificador de las sesiones operativas continuas de la planta piloto para cada sitio de medición.
Sitio	Etiqueta del sitio geográfico de operación de la planta (Tocopilla, Calama, SanPedro, Chacabuco, PSDA, etc.); categórica, adimensional.
`StratifiedKFold`	Esquema de validación cruzada estratificada K-fold.
SVM	*Support Vector Machine*; máquina de vectores de soporte.
SVR	*SupportVectorRegressor*; máquina de vectores de soporte para regresión, clase `SVR`.
Target	Valores medidos (potencia observada del banco PEMFC); W.

Sigla	Significado
v0.9950, v0.9975, v0.9990	Etiquetas de versiones depuradas de la base de datos mediante FASEK5 con umbrales de probabilidad media global ($S_i > 0.9950, 0.9975, 0.9990$)
WO	Base de datos natural (sin remoción de *outliers*).
WTM-01	Tanque de almacenamiento de agua de proceso de la planta piloto.
XGB	Abreviatura utilizada para el modelo de regresión `XGBRegressor` (XGBoost).
Z1_3	Codificación de la coordenada ILR Z_1 utilizada como predictor continuo en el modelado.
Z2_3	Codificación de la coordenada ILR Z_2 utilizada como predictor continuo en el modelado.

Capítulo 1

Introducción

El aumento sostenido de la demanda energética mundial, impulsado por el crecimiento económico, la urbanización y la industrialización, se ha sustentado históricamente en el uso intensivo de combustibles fósiles. Las emisiones de gases de efecto invernadero (GEI) asociadas a su combustión y a procesos industriales constituyen la fracción predominante de las emisiones antropogénicas globales y explican la mayor parte del calentamiento global observado desde la era preindustrial (1850–1900), así como los impactos ambientales asociados al cambio climático (Calvin et al., 2023). En 2019, las emisiones antropogénicas netas se estimaron en el orden de 59 $GtCO_2$-eq/año, de las cuales alrededor de un 65 % correspondió a CO_2 de origen fósil e industrial. Esta distribución sectorial se detalla en la Figura 1.1, donde se aprecia que el sistema energético concentra la mayor proporción de las emisiones globales de GEI.

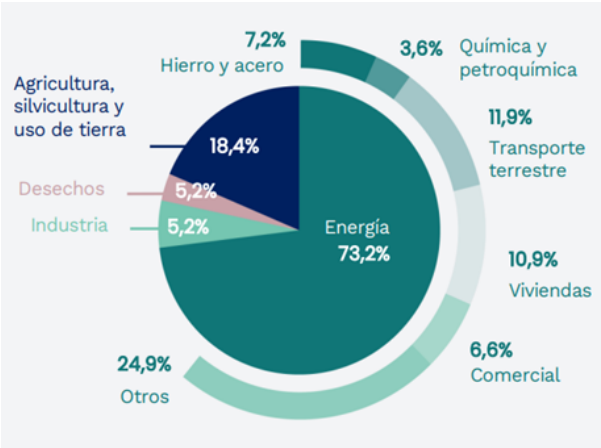


Figura 1.1: Emisiones globales de GEI por sector productivo (Ministerio de Energía, 2020).

A nivel global, la transición hacia sistemas energéticos de bajas emisiones de carbono combina la adopción progresiva de energías renovables no convencionales (ERNC) con el uso de vectores energéticos de bajas emisiones que permitan abastecer sectores donde la electrificación directa

resulta técnica o económicamente compleja. En este contexto, el hidrógeno verde, producido mediante electrólisis del agua con electricidad proveniente de ERNC, se perfila como vector clave para descarbonizar sectores industriales y de transporte intensivos en energía (Birol, 2019; Staffell et al., 2019). En 2018, alrededor del 96 % de la producción global de hidrógeno se realizaba a partir de combustibles fósiles, con apenas una fracción del 4 % de producción vía electrólisis (IRENA, 2018), como se ilustra en la Figura 1.2.

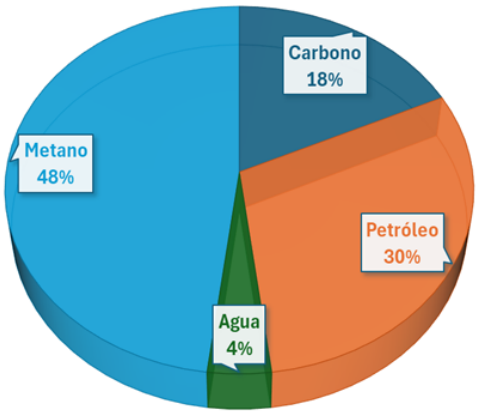


Figura 1.2: Fuentes de producción del hidrógeno a nivel mundial (Vásquez and Salinas, 2018).

Chile, y en particular la Región de Antofagasta, se han posicionado como actores estratégicos en este escenario debido a la combinación de un recurso solar de clase mundial y una política pública orientada al despliegue del hidrógeno verde (Ministerio de Energía, 2020). Se han reportado niveles excepcionales de irradiancia en el norte de Chile y condiciones favorables para la generación fotovoltaica a gran escala (Escobar et al., 2015). Este contexto ha impulsado el desarrollo de proyectos y pilotos que integran generación renovable, producción de hidrógeno y su uso en aplicaciones estacionarias y móviles. La Figura 1.3 sintetiza las aplicaciones finales potenciales del hidrógeno renovable (*Power-to-X*), destacando su versatilidad como vector energético en sectores como generación eléctrica, transporte, procesos industriales y almacenamiento de energía.

Dentro de la cadena de valor del hidrógeno, las pilas de combustible de membrana de intercambio protónico (PEMFC) cumplen un rol central como tecnologías de conversión electroquímica que transforman hidrógeno y oxígeno del aire en electricidad y calor, con eficiencias eléctricas del orden de 50–60 %, superiores a la eficiencia térmica de freno de motores de combustión interna en aplicaciones convencionales (30–36 % en motores de encendido por chispa y 42–43 % en motores de encendido por compresión) (Kreutz and Ogden, 2000; Dahham et al., 2022). Su arquitectura básica comprende un ánodo, un cátodo, una membrana polimérica conductora de protones y capas porosas para el transporte de reactivos y la evacuación de agua líquida. Son compatibles con aplicaciones de respaldo, microrredes y sistemas modulares asociados a hidrógeno verde (Wang et al., 2011).

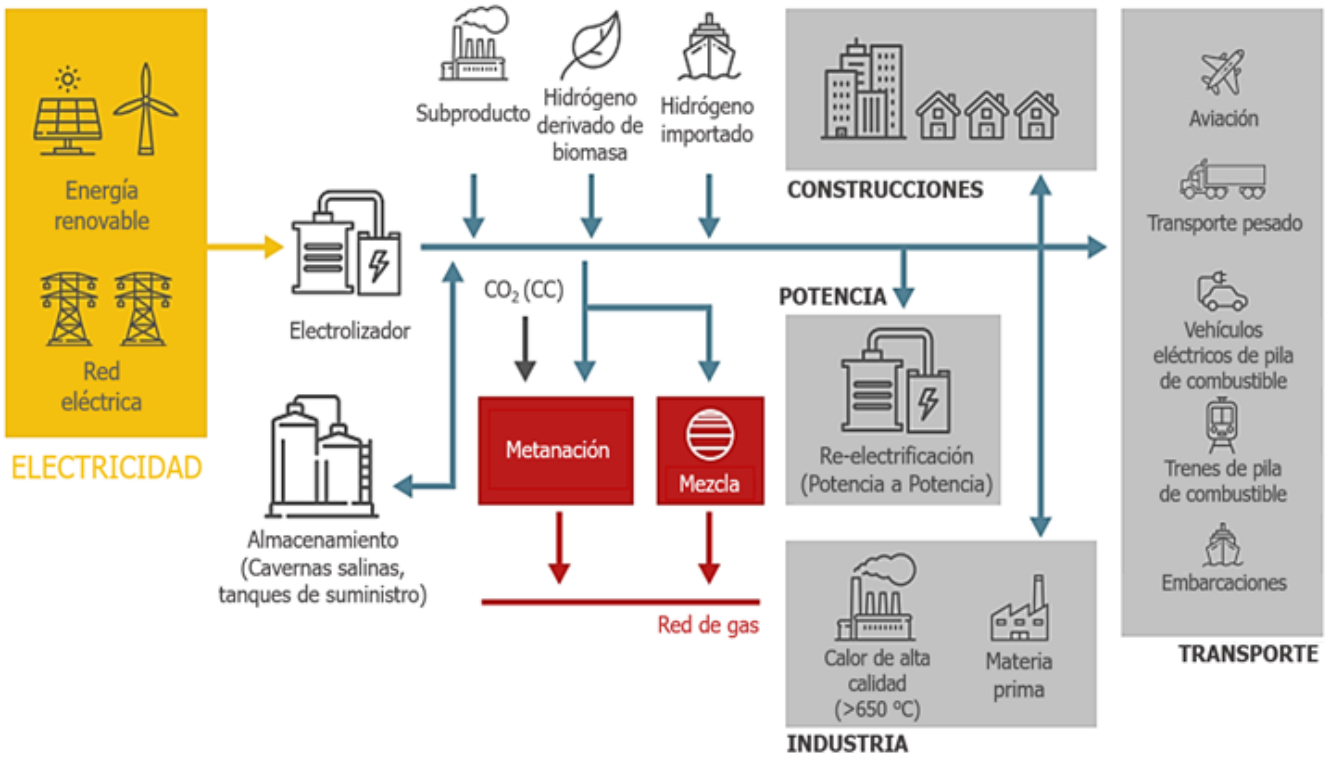


Figura 1.3: Diagrama de las potenciales aplicaciones finales del hidrógeno renovable (Traducido desde IRENA (2018)).

En este contexto, la Línea de Energía del Centro Científico Tecnológico de la Región de Antofagasta (CICITEM) desarrolló una planta piloto móvil de hidrógeno verde, contenerizada y transportable, que integra generación fotovoltaica, producción de hidrógeno por electrólisis, almacenamiento gaseoso y conversión electroquímica en un banco de pilas de combustible PEM, además de sistemas auxiliares de acondicionamiento y almacenamiento eléctrico (Henríquez, 2025). Esta planta se ha operado en campañas itinerantes *in situ* en distintos puntos de la Región de Antofagasta, exponiendo los equipos a condiciones desérticas contrastantes en términos de altitud, temperatura, humedad relativa y presión atmosférica. La Figura 1.4 muestra la topografía de la Región de Antofagasta, donde se aprecia el fuerte gradiente altitudinal entre el litoral y el altiplano que condiciona las variaciones climáticas que influyen sobre el desempeño de las tecnologías.

La configuración general de la planta se esquematiza en la Figura 1.5, mientras que la Figura 1.6 presenta el diagrama de flujo de proceso simplificado, destacando las principales corrientes de materia y energía entre los subsistemas fotovoltaico, de electrólisis, de almacenamiento de hidrógeno y el banco PEMFC. La Figura 1.7 ilustra la planta piloto móvil durante una de las campañas de medición en el desierto de Antofagasta, evidenciando las condiciones ambientales reales en las que se obtuvieron los datos utilizados en este estudio.

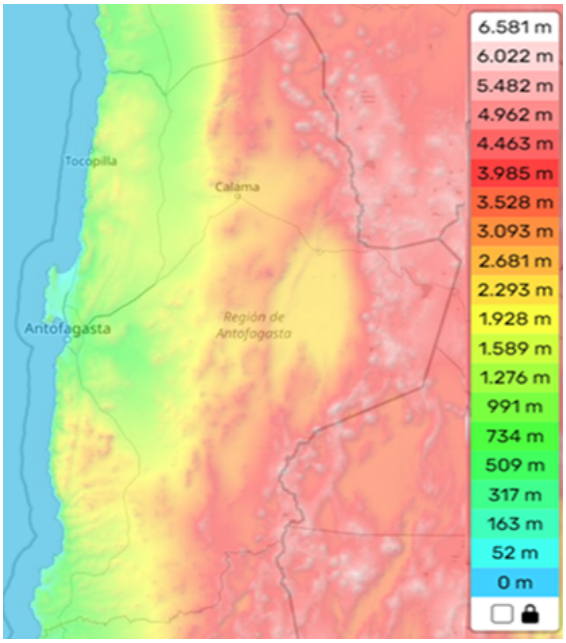


Figura 1.4: Mapa topográfico de la Región de Antofagasta (Topographic-Map.com, s. f.).

La base de datos empleada en esta tesis se construyó a partir de dichas campañas y recoge variables ambientales (temperatura ambiente, presión atmosférica y humedad relativa) y variables eléctricas del banco compuesto por tres pilas de combustible comerciales modelo GenSure E-1100 (corriente, tensión y potencia) (Plug Power Inc., 2018).

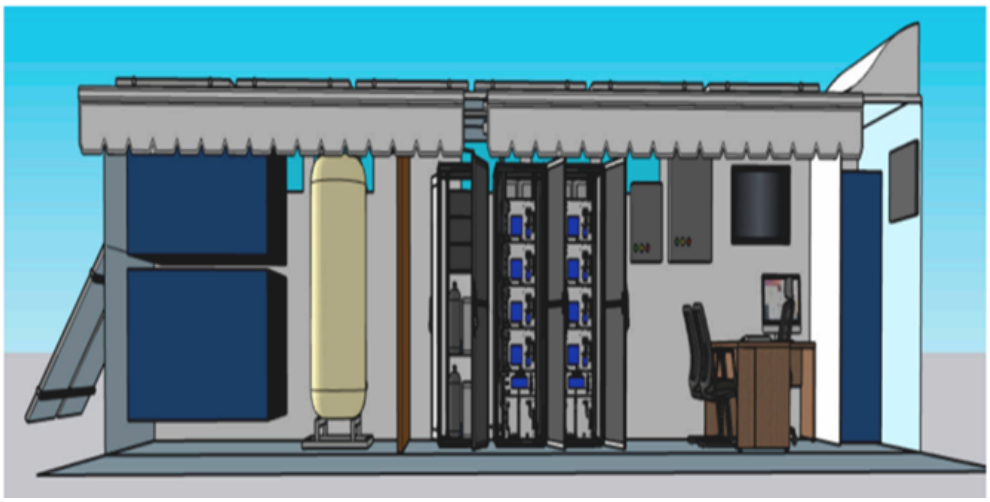


Figura 1.5: Esquema general de la planta piloto móvil de hidrógeno verde de CICITEM (Chavez-Angel et al., 2023).

La planta piloto móvil P^3H_2V está diseñada para producir del orden de 2.0–2.5 kg de H_2 por día y se organiza en tres subsistemas acoplados (Henríquez, 2025; Chavez-Angel et al., 2023). El primero corresponde a la generación fotovoltaica: un campo de paneles monocristalinos con una potencia instalada de 31.8 kW, montado al inicio de cada campaña experimental. De esta potencia, se considera una distribución nominal de 27.2 kW para la alimentación eléctrica de los electrolizadores del sistema de producción de hidrógeno y 5.4 kW para los consumos auxiliares de la planta, de acuerdo con la distribución eléctrica definida para la P^3H_2V (Henríquez, 2025; Chavez-Angel et al., 2023). El segundo subsistema es el de producción y acondicionamiento de H_2. El agua se almacena en el tanque WTM-01 y se trata mediante ósmosis inversa y desionización hasta alcanzar una conductividad compatible con los electrolizadores, conforme a especificaciones del sistema. El tren de producción integra ocho electrolizadores de membrana de intercambio aniónico AEM (EZ-01 a EZ-08), con una potencia total de 20 kW, en los que el agua se disocia en H_2 y O_2 en cámaras separadas. El hidrógeno producido se envía a los purificadores HPS-01 y HPS-02, donde se reduce el contenido de vapor de agua hasta lograr una pureza de 99.999 %; posteriormente se almacena en un estanque tipo IV a 35 bar o se deriva directamente al banco de pilas de combustible, mientras que el oxígeno se ventea al exterior del contenedor (Henríquez, 2025; Chavez-Angel et al., 2023).

El tercer subsistema corresponde al consumo de hidrógeno en el banco de pilas de combustible tipo PEM (FC-01, FC-02 y FC-03), con una capacidad nominal de 3.3 kW (Henríquez, 2025; Chavez-Angel et al., 2023). Las pilas se alimentan con H_2 proveniente del sistema de producción o del área de almacenamiento y con aire ambiente como fuente de O_2, impulsado mediante compresores integrados. En los electrodos tiene lugar una reacción electroquímica catalítica que convierte la energía química del hidrógeno en potencia eléctrica, utilizada para cargar un banco de baterías, abastecer los consumos auxiliares de la planta y respaldar la operación de los electrolizadores durante períodos de baja o nula radiación solar (Henríquez, 2025; Chavez-Angel et al., 2023). De este modo, la P^3H_2V implementa un esquema *Power-to-Hydrogen-to-Power*, en el cual los excedentes de energía fotovoltaica se transforman en H_2 y se recuperan posteriormente como electricidad gestionable mediante el banco de pilas de combustible (Henríquez, 2025). En este marco, el banco PEMFC se constituye en el componente central de análisis y modelado de la presente tesis.

El desempeño eléctrico de las PEMFC depende de manera no lineal y acoplada de las condiciones de operación, incluyendo temperatura, presión, humedad de entrada y caudales de alimentación, así como de estados internos de hidratación y transporte de masa en la membrana y las capas porosas de difusión (Wang et al., 2011). En ambientes de gran altitud y clima desértico, como los que enfrenta la planta piloto de CICITEM, la disminución de la presión atmosférica —y con ello de la presión parcial de oxígeno— junto con la variabilidad térmica diaria y la baja humedad, pueden afectar el desempeño y el margen para alcanzar la potencia nominal, tal como se ha reportado en

estudios de sensibilidad a las condiciones ambientales (Hordé et al., 2012; Saleh et al., 2018). Contar con modelos para predecir la potencia del banco PEMFC en función de variables ambientales y operativas es relevante para dimensionar aplicaciones de respaldo, planificar la operación bajo distintos entornos climáticos y evaluar la viabilidad técnica de soluciones basadas en hidrógeno verde en el norte de Chile.

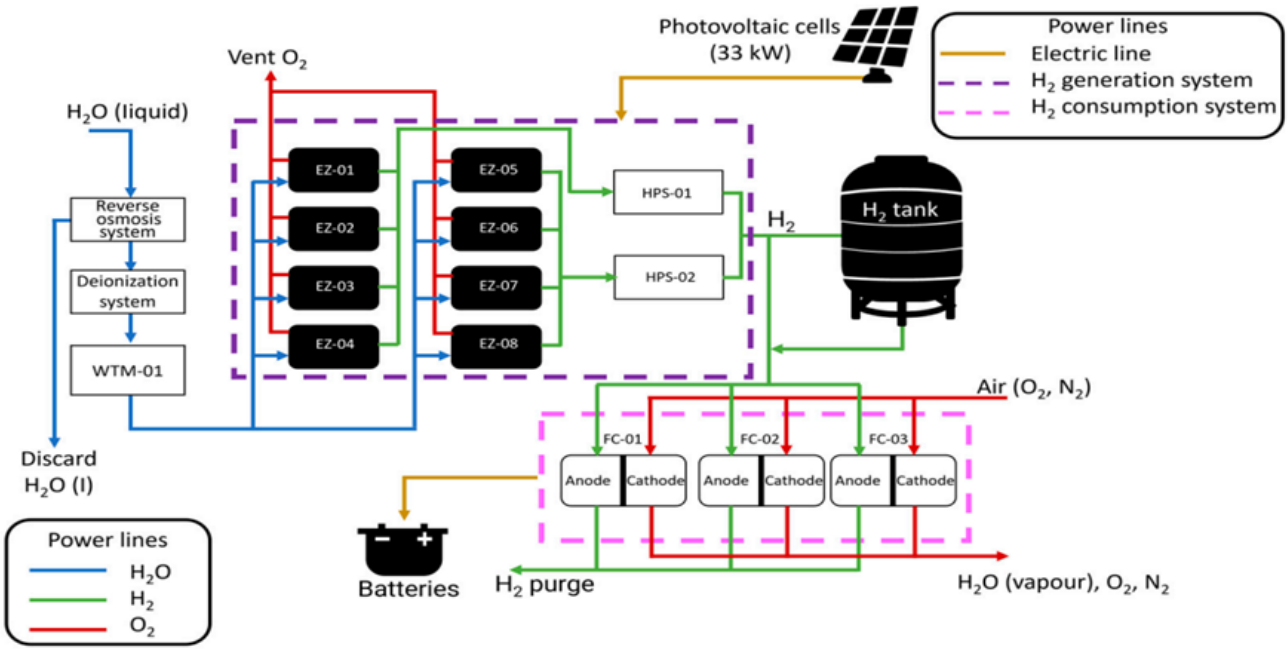


Figura 1.6: Diagrama de flujo simplificado de la P^3H_2V de CICITEM (Chavez-Angel et al., 2023).

Figura 1.7: Fotografía de la planta piloto móvil en operación de campo en el Desierto de Atacama (Chavez-Angel et al., 2023).

Tradicionalmente, el comportamiento de las pilas de combustible se ha modelado mediante enfoques fisicoquímicos, basados en balances de masa y carga (y, en formulaciones no isotérmicas, de

energía), cinética electroquímica y modelos de transporte de especies, resueltos mediante métodos numéricos o formulaciones de orden reducido (Bernardi and Verbrugge, 1991; Springer et al., 1991; Wang et al., 2011). Estos modelos permiten analizar con alta resolución fenómenos internos, pero su aplicación directa a bancos comerciales operando en terreno suele verse limitada por la cantidad de parámetros a calibrar, la disponibilidad de información geométrica y operativa detallada, y el costo computacional (en particular en formulaciones multidimensionales), además de la brecha entre los supuestos idealizados y los regímenes reales de operación bajo condiciones ambientales altamente variables (Hamdollahi and Jun, 2023).

Como complemento a estos enfoques, los métodos basados en datos y el Aprendizaje Automático (*Machine Learning*, ML) se han utilizado con éxito para aproximar la relación entre variables de entrada y el desempeño de las PEMFC a partir de registros experimentales, con buen desempeño predictivo en distintos regímenes de operación (Ding et al., 2022). En la literatura se han reportado aplicaciones de redes neuronales, máquinas de vectores de soporte, bosques aleatorios y métodos de *gradient boosting* para predecir tensión, potencia o indicadores de estado en celdas y bancos PEM, logrando modelos con alta capacidad de ajuste y menor necesidad de calibrar explícitamente parámetros fisicoquímicos internos que suelen ser difíciles de medir en terreno (Su et al., 2023; Ding et al., 2022).

Entre 2022 y 2024 se ha intensificado la incorporación de técnicas de *Machine Learning* (ML) en el modelado y la operación de sistemas PEMFC. Sharma et al. (2024) revisan su uso en diagnóstico y gestión del estado de salud, incluyendo tareas de predicción de desempeño y detección de fallas. En esa línea, Legala et al. (2022) desarrollan modelos basados en datos (ANN y SVR) para estimar variables de desempeño y estados internos —como tensión y variables asociadas al estado hídrico de la membrana— a partir de condiciones operacionales, mostrando alta precisión y reduciendo la necesidad de campañas experimentales extensas al apoyarse en modelos fisicoquímicos validados como fuente de datos. En trabajos posteriores, los mismos autores extienden el enfoque hacia la representación de degradación y de dinámicas térmicas y de presión bajo condiciones experimentales controladas (Legala et al., 2023). De forma complementaria, Ding et al. (2022) sintetizan aplicaciones de ML (ANN, SVM, RF, entre otros) orientadas a la optimización de PEMFC, abarcando desde el diseño de materiales hasta la operación y el control.

Sin embargo, muchos de estos trabajos se apoyan en configuraciones de laboratorio o bancos de prueba instrumentados, con condiciones de operación relativamente controladas y bases de datos con menor variabilidad exógena (ambiental y operacional) que las obtenidas en campañas *in situ* multi-sitio (Ding et al., 2022; Sharma et al., 2024). En contraste, la transferencia de modelos entrenados con datos de laboratorio hacia operación en terreno puede verse limitada por diferencias entre regímenes de operación, en particular bajo condiciones dinámicas, por la baja generalización entre sistemas y por la necesidad de históricos representativos, así como de control

de calidad y trazabilidad de datos (D'Silva et al., 2025; Yue et al., 2021). En consecuencia, los reportes de modelos basados en ML entrenados y validados directamente con series operacionales de plantas piloto en terreno—en particular, en configuraciones móviles, multi-sitio y expuestas a alta variabilidad climática, como la considerada en este estudio—resultan menos habituales en la literatura abierta, situando este caso en un escenario comparativamente poco documentado.

En este trabajo se dispone de una base de datos proveniente de campañas multi-sitio, con múltiples sesiones de operación, del banco PEMFC de la planta piloto móvil de CICITEM, con tamaño de muestra moderado, desbalances entre sitios, modos de operación y configuraciones internas, y presencia de ruido de medición y registros atípicos (*outliers*). El análisis exploratorio mostró que la base es acotada en tamaño, pero altamente heterogénea y multimodal, con desbalances marcados entre los estratos definidos por sitio, modo de operación e identificador de pila (FC_ID).

El tratamiento de *outliers* en datos de PEMFC se aborda con frecuencia mediante reglas de depuración determinísticas o mediante la aplicación de uno o pocos algoritmos de detección como etapa de preprocesamiento. Por ejemplo, Niu et al. (2025) remueven valores atípicos mediante el método del rango intercuartílico (IQR) antes de entrenar una red BiLSTM para predecir la vida útil restante. De forma análoga, Qin et al. (2024) exploran *Local Outlier Factor* (LOF) e *Isolation Forest* para depurar datos de curvas de polarización antes de ajustar modelos semi-empíricos.

En estos casos, los detectores suelen aplicarse de forma aislada o secuencial y con umbrales esencialmente deterministas, como parte de una etapa de preprocesamiento. En esta tesis se propone, en cambio, un esquema de depuración que integra explícitamente la evidencia de múltiples detectores y considera la estructura estratificada de la base de datos, con el objetivo de aumentar la consistencia de la limpieza bajo heterogeneidad operativa. Si bien en otros dominios se han propuesto estrategias de limpieza basadas en *ensembles* de detectores, en la literatura revisada para este estudio su adopción en datos experimentales de PEMFC aparece menos documentada (Li and Zhang, 2023).

En particular, para bancos PEMFC operando en condiciones reales del desierto de Atacama, abarcando ambientes costeros e interiores de distinta altitud, la literatura revisada para este estudio no reporta de forma sistemática un *pipeline* reproducible que permita integrar la depuración multi-detector de *outliers* con una evaluación controlada del efecto del nivel de depuración sobre el desempeño y la capacidad de generalización de modelos de Aprendizaje Automático, considerando además la estructura multi-sitio y multi-configuración de la base de datos. Esta brecha motiva el desarrollo de un enfoque de modelado que incorpore explícitamente el contexto climático local y la heterogeneidad operacional propia de campañas *in situ*.

Con el fin de contribuir a acortar esta brecha, en esta tesis se desarrolla y valida un *pipeline* de modelado de Aprendizaje Automático para predecir la potencia eléctrica de un banco de pilas de combustible PEM bajo las condiciones climáticas y operativas de la planta piloto móvil de

CICITEM en la Región de Antofagasta. El *pipeline* integra tres componentes principales: (i) un bloque de análisis exploratorio e ingeniería de características, que construye predictores ambientales y operativos coherentes con la física del sistema y los evalúa mediante una estrategia de ablación incremental; (ii) una metodología de puntaje probabilístico para la detección de *outliers*, basada en la combinación de familias de detectores univariantes, bivariantes y multidimensionales, que genera versiones depuradas de la base y permite estudiar la sensibilidad del modelado al nivel de limpieza; y (iii) la evaluación comparativa de modelos de Aprendizaje Automático —incluyendo *Random Forest*, *XGBoost*, *CatBoost*, redes neuronales *Multilayer Perceptron* y máquinas de vectores de soporte— entrenados sobre las distintas versiones de la base, cuantificando su ajuste y capacidad de generalización mediante métricas globales y diagramas solares y de Taylor (Wadoux et al., 2022). La Figura 1.8 resume este *pipeline*.

Figura 1.8: *Pipeline* de modelado basado en Aprendizaje Automático propuesto en esta tesis (Elaboración propia).

Capítulo 2

Objetivos y alcances

En este capítulo se presentan los objetivos y alcances de esta tesis.

2.1 Objetivo general

Desarrollar y validar un *pipeline* de modelado de Aprendizaje Automático para predecir la potencia eléctrica de un banco de pilas de combustible PEM bajo las condiciones climáticas y operativas de la planta piloto móvil de CICITEM en la Región de Antofagasta, evaluando el impacto de la ingeniería de características y de la depuración de *outliers* sobre el ajuste y la generalización de distintas configuraciones de modelos.

2.2 Objetivos específicos

1. Levantar el estado del arte sobre el modelado de pilas de combustible PEM mediante técnicas de Aprendizaje Automático.

2. Diseñar predictores de contexto ambiental y operativo coherentes con la física del sistema, evaluando su contribución a la capacidad predictiva mediante la estrategia de ablación incremental de características.

3. Desarrollar e implementar una metodología de puntaje probabilístico para la detección de *outliers* y cuantificar el impacto de distintos niveles de depuración de la base de datos sobre las métricas de ajuste y estabilidad de los modelos.

4. Evaluar y comparar modelos de Aprendizaje Automático sobre las distintas versiones de la base de datos, a fin de seleccionar el *pipeline* con mejor compromiso entre precisión y generalización.

2.3 Alcances

En la presente investigación se realizan predicciones de la potencia eléctrica de un banco de pilas de combustible PEM de hidrógeno verde mediante modelos de regresión basados en árboles de decisión y variantes de *gradient boosting*, redes neuronales artificiales y máquina de vectores de soporte. Conforme a las limitaciones de la base de datos, el trabajo contempla la incorporación de la ingeniería de características y depuración controlada de *outliers* para robustecer la calidad de los modelos.

La evaluación de desempeño se realiza mediante un enfoque integrado por métricas de ajuste, generalización, sesgo y dispersión del error y su representación en diagramas solares y de Taylor. A partir de los resultados obtenidos, se cuantifica el impacto de las técnicas implementadas y se comparan distintas configuraciones de los modelos.

Se adopta un enfoque de modelado pseudoestacionario, a pesar de que el sistema opera dinámicamente. Conforme a la caracterización del fabricante, las pilas de combustible experimentan derrateo de potencia eléctrica en función de la altitud y temperatura ambiente. Además, los tiempos de acondicionamiento iniciales son del orden de minutos, significativamente menor a la frecuencia de muestreo y a la extensión de varias horas de las sesiones operativas.

Dado que el enfoque es puramente basado en datos, no se aborda la simulación de modelos analíticos de la PEMFC para aumentar la cantidad de datos, ni la incorporación de términos de regularización basados en ecuaciones de continuidad sobre la función de pérdida de los modelos. Asimismo, no se diseñan estrategias de control u optimización de la operación de la planta.

El alcance se limita a validar la factibilidad del enfoque de modelado de Aprendizaje Automático en condiciones ambientales y operativas reales de la Región de Antofagasta y propone las técnicas a considerar para mejorar la calidad predictiva de los modelos, estableciendo un marco referencial reproducible para la planificación de campañas experimentales y su modelado en otras zonas de interés científico-tecnológico.

Capítulo 3

Materiales y Metodologías

Enfoque general

El flujo de trabajo se organizó en una secuencia de cuatro etapas principales:

1. Análisis exploratorio de datos (EDA) y ensayos preliminares con redes neuronales artificiales con implementación de ingeniería de características y ablación incremental.
2. Depuración de la base de datos mediante la metodología probabilística FASEK5, que integra 37 detectores organizados en cinco familias de detección de *outliers*.
3. Entrenamiento, calibración y comparación de los modelos predictivos sobre las versiones original y depuradas de la base de datos.
4. Evaluación del desempeño y selección de un modelo de referencia, considerando la capacidad de generalización y la robustez frente a las condiciones heterogéneas de operación.

Datos experimentales y variables de estudio

Se empleó una base de datos proveniente de la campaña itinerante de una planta piloto móvil de hidrógeno verde operada en la Región de Antofagasta. En esta campaña se registraron mediciones de variables ambientales y eléctricas del banco de pilas de combustible con una frecuencia de 15 minutos, se etiquetaron los sitios de medición (Tocopilla, Calama, SanPedro, Chacabuco y PSDA) y se reportaron la latitud y la longitud asociadas a cada emplazamiento.

Las variables nativas de la base de datos se clasificaron en los siguientes grupos:

- Ambientales: temperatura ambiente T_{amb}, presión atmosférica p_{amb} y humedad relativa (HR).
- Eléctricas: tensiones (V), corrientes (I) y potencias eléctricas (W) de las pilas de combustible.

En los experimentos de modelado supervisado, la variable objetivo del problema de regresión se definió como la potencia eléctrica de las pilas de combustible, W.

Entorno computacional y herramientas

La implementación de los modelos se realizó en dos entornos principales. A continuación, se listan las aplicaciones, librerías y clases utilizadas.

MATLAB

Aplicación *Neural Net Fitting* para ensayos preliminares con redes neuronales *feedforward* entrenadas con el algoritmo de regularización bayesiana.

Python 3.12

- Manejo y análisis de datos: `pandas`, `numpy`.
- Visualización: `matplotlib`, `seaborn`, `plotly`.
- Modelado y preprocesamiento (`scikit-learn`):
  - `sklearn.model_selection` (división entrenamiento–prueba, `StratifiedKFold`).
  - `sklearn.preprocessing` (`StandardScaler`, `RobustScaler`).
  - `sklearn.decomposition` (PCA).
  - `sklearn.cluster` (`DBSCAN`, `KMeans`).
  - `sklearn.ensemble` (`RandomForestRegressor`).
  - `sklearn.neural_network` (`MLPRegressor`).
  - `sklearn.svm` (`SVR`).
  - `sklearn.metrics` (RMSE, r, R^2, etc.).
- Otros modelos basados en árboles:
  - `xgboost.XGBRegressor`.
  - `catboost.CatBoostRegressor`.

El código se estructuró en módulos respetando la secuencia del flujo de trabajo general. Se incluyó versionamiento de las configuraciones de modelos y de las bases de datos depuradas para asegurar la reproducibilidad de los resultados.

Preprocesamiento de datos e ingeniería de características

Las operaciones de preprocesamiento tuvieron como objetivo obtener conjuntos de datos consistentes y adecuados para el entrenamiento y evaluación de los modelos. Como filtro inicial, se verificó la consistencia física de los registros de potencia eléctrica, con el fin de identificar registros erróneos o desajustes en el ingreso de datos.

Adicionalmente, se aplicó codificación *One-Hot* a las variables categóricas, transformándolas en columnas binarias, y se estandarizaron las variables numéricas. Estas operaciones son necesarias para estabilizar el entrenamiento de los modelos predictivos y evitar que diferencias de escala dominen el ajuste.

En paralelo, se diseñaron descriptores derivados a partir de la base de datos para incorporar contexto operacional, climático y temporal. El principio fundamental fue proporcionar información adicional que favoreciera el aprendizaje de los modelos. El diseño de estos nuevos descriptores requirió conocimiento de dominio sobre la tecnología y sobre la caracterización de la base de datos.

La contribución de estos predictores derivados se evaluó por bloques de información, aplicando una estrategia de ablación incremental en la aplicación *Neural Net Fitting* de MATLAB. Como resultado, se definió la base de datos *Baseline*, compuesta por el conjunto de características que, al entrenar, entregaban un nivel de ajuste considerado aceptable.

Detección de outliers y metodología FASEK5

Conforme al diagnóstico del EDA, se implementó una metodología de detección de outliers organizada en cinco familias de métodos (F1–F5), aplicadas sobre un estrato representativo del desbalance de grupos presentes en la base de datos. Estas familias abarcan desde reglas univariadas, bivariadas y multidimensionales sobre las variables nativas, e incluyen análisis por régimen operativo y configuración de las pilas de combustible, así como detección sobre la variable objetivo W.

Para esta tarea se emplearon técnicas estadísticas como diagramas de caja y bigote (*boxplots*), clusterización mediante DBSCAN y límites de control estadístico multidimensional (Hotelling's T^2 y SPE), entre otras. Otras consideraciones metodológicas se detallan en el capítulo respectivo.

La integración de las evidencias se realizó mediante la metodología FASEK5, que modela las evidencias como probabilidades de que cada registro sea un *outlier*, considerando factores como la fiabilidad de cada método, la jerarquización entre familias, entre otros factores.

Generación de versiones depuradas de la base de datos

A partir de los puntajes estimados por FASEK5 se generaron distintas versiones depuradas de la base de datos, definidas por umbrales de probabilidad bajo la restricción de una contaminación máxima del 5% respecto de la base sin depurar (WO). Estas versiones se nombraron según la proporción de datos retenidos: v0.9990, v0.9975 y v0.9950.

La comparación del desempeño de los modelos sobre la base original y sobre estas versiones depuradas permitió cuantificar la ganancia relativa en las métricas de desempeño asociada a la remoción controlada de *outliers*.

Modelos de Aprendizaje Automático y configuración de entrenamiento

Sobre las distintas versiones de la base de datos se entrenaron y compararon cinco modelos de regresión: `RandomForestRegressor` (RFR), `XGBRegressor` (XGB), `CatBoostRegressor` (CAT), `MLPRegressor` (MLP) y `SupportVectorRegressor` (SVR). Se aplicó un preprocesamiento específico en función de las características de cada uno de los modelos seleccionados.

Para asegurar la comparabilidad de los resultados, se utilizó el mismo esquema de validación cruzada `StratifiedKFold` y una semilla aleatoria común para garantizar la reproducibilidad. Los hiperparámetros de cada modelo se calibraron mediante el algoritmo de optimización bayesiana, utilizando como función objetivo el RMSE del conjunto de prueba.

Métricas de evaluación y criterios de selección

Para evaluar la calidad predictiva de los modelos se empleó un conjunto integral de métricas de ajuste, sesgo y dispersión del error, calculadas sobre el conjunto de prueba. Se seleccionó la raíz del error cuadrático medio (RMSE) como métrica principal.

Otras métricas, como el sesgo medio (ME), la desviación estándar de los errores (SDE), el coeficiente de correlación de Pearson (r), el coeficiente de determinación (R^2), entre otras; se utilizaron para cuantificar la calidad de las predicciones y para su representación mediante los diagramas solares y de Taylor.

Adicionalmente, se empleó la métrica $\mathrm{GAP}_{\mathrm{RMSE}}$ para cuantificar el sobreajuste entre los conjuntos de entrenamiento y prueba. Para evaluar la robustez de los modelos frente a *outliers*, se definieron la ganancia absoluta y la ganancia relativa en los datos de prueba al comparar las versiones depuradas de la base de datos con la base original.

Capítulo 4

Marco Teórico: Fundamentos del sistema PEMFC y enfoques de modelado

4.1 Fundamentos de los sistemas PEMFC

Una celda de combustible es un dispositivo electroquímico que genera energía eléctrica a partir de la oxidación de un combustible gaseoso, como hidrógeno o compuestos ricos en hidrógeno, sin etapas intermedias de conversión mecánica como ocurre en los motores de combustión interna (Barbir, 2012; O'hayre et al., 2016). A diferencia de las baterías, que almacenan internamente los reactivos, las celdas de combustible operan como sistemas abiertos: el combustible y el oxidante se suministran de forma continua mientras se entrega potencia eléctrica y calor. En el caso de las celdas de membrana de intercambio protónico (PEMFC), la reacción global corresponde a la oxidación electroquímica del hidrógeno:

$$\mathrm{H_2} + \frac{1}{2}\mathrm{O_2} \rightarrow \mathrm{H_2O} \tag{4.1}$$

Para sistemas PEMFC comerciales, la literatura reporta eficiencias eléctricas globales típicas de 40–60 %, en función del régimen de operación y del nivel de integración del sistema. Por su parte, la eficiencia termodinámica ideal asociada a la reacción global —definida como el cociente entre la variación de la energía de Gibbs y la entalpía de reacción— alcanza valores cercanos al 80–83 % bajo condiciones estándar (Barbir, 2012; O'hayre et al., 2016).

A nivel constructivo, el núcleo de una PEMFC se organiza en el ensamble membrana–electrodo (MEA, por sus siglas en inglés, *membrane electrode assembly*), conformado por una membrana polimérica conductora de protones que actúa como electrolito sólido y por las capas catalíticas del ánodo y del cátodo, donde ocurren las reacciones de oxidación y reducción. La membrana permite el transporte selectivo de protones y limita la mezcla directa de los reactantes gaseosos. La MEA

se intercala entre capas porosas de difusión de gases, que favorecen la canalización de reactivos y la evacuación de agua y calor. El conjunto se comprime entre placas bipolares, que incorporan canales de flujo y colectores de corriente eléctrica. La Figura 4.1 esquematiza esta configuración ánodo–membrana–cátodo y el transporte de especies y carga de la PEMFC.

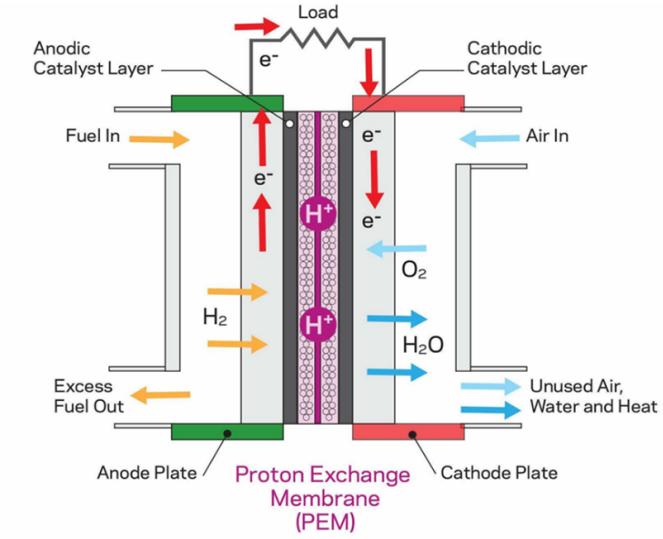


Figura 4.1: Esquema de configuración de una celda de combustible PEM operada con hidrógeno (Issa et al., 2025).

En operación, el hidrógeno se alimenta al ánodo, donde se oxida y genera protones y electrones; los protones atraviesan la membrana hacia el cátodo, mientras que los electrones circulan por el circuito externo, entregando potencia eléctrica. En el cátodo, el oxígeno del aire se reduce y reacciona con los protones para formar agua en fase líquida o vapor, según las condiciones de operación. De acuerdo con la ley de Faraday, la densidad de corriente se relaciona directamente con la tasa de reacción en la MEA, por lo que puede emplearse como medida indirecta del grado de conversión (Barbir, 2012; O'hayre et al., 2016).

Una celda PEM individual suele operar con tensiones entre 0.4 y 0.9 V y densidades de corriente del orden de 0.5–1.0 $\mathrm{A\,cm^{-2}}$, dependiendo del diseño de la MEA y de las condiciones de operación (Barbir, 2012; O'hayre et al., 2016). Para aplicaciones que requieren potencias superiores, múltiples celdas se conectan en serie formando un *stack*, de modo que el voltaje total del conjunto es aproximadamente la suma de los voltajes individuales. La respuesta tensión–corriente de una celda real puede expresarse como:

$$V_{\mathrm{cell}} = E_{\mathrm{rev}} - \eta_{\mathrm{act}} - \eta_{\mathrm{ohm}} - \eta_{\mathrm{conc}} \tag{4.2}$$

Donde $\mathrm{E_{rev}}$ es el potencial reversible y η_{act}, η_{ohm} y η_{conc} representan las pérdidas por activación, óhmicas y de concentración, respectivamente.

La Figura 4.2 muestra una curva de polarización típica, donde se aprecia la diferencia entre la tensión (o voltaje) ideal y el valor real de operación, ilustrando la forma funcional de contribución de las pérdidas de la ecuación 4.2 sobre la curva I-V.

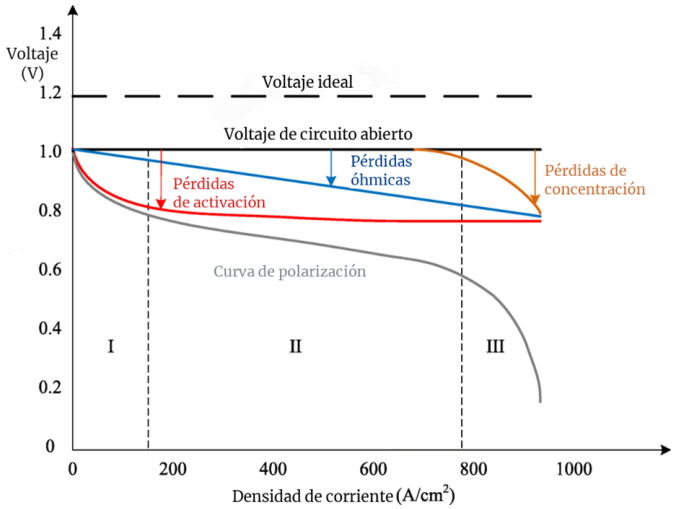


Figura 4.2: Curva de polarización típica de una celda de combustible tipo PEM (Li et al., 2019).

En este capítulo se presenta una descripción general del comportamiento físico del sistema. El contraste entre modelos analíticos y enfoques basados en datos se desarrolla en la Sección 4.3.

4.2 Curva de polarización y relación I-V-P

El desempeño eléctrico de una celda o *stack* PEM se caracteriza mediante la curva de polarización, que relaciona la tensión de salida V con la corriente I bajo condiciones de operación definidas (temperatura, presión, caudales de gas y estado de humidificación). La curva se determina fijando niveles sucesivos de corriente (o carga) y registrando la tensión en régimen cuasiestacionario. En una celda real, la tensión decrece con la corriente debido a pérdidas por activación, óhmicas y por transporte de masa (comúnmente agrupadas como pérdidas de concentración). A partir de V(I) se define un intervalo de operación que cumpla criterios de desempeño y restricciones (p. ej., tensión mínima por celda); en aplicaciones estacionarias se emplean a menudo valores de referencia del orden de 0.6–0.7 V por celda, según el sistema y el criterio adoptado (Barbir, 2012; O'hayre et al., 2016).

A partir de la curva de polarización se deriva la relación potencia–corriente. La potencia eléctrica se calcula como:

$$\mathrm{P} = \mathrm{I}\,\mathrm{V} \tag{4.3}$$

De este modo, para cada punto de la curva V(I) se define P(I) = I V(I). Típicamente, P(I) presenta un máximo a corrientes intermedias: a corrientes bajas la potencia es reducida por el bajo nivel de I, mientras que a corrientes altas la caída de V limita el incremento de potencia. En consecuencia, el rango de operación del *stack* se establece equilibrando potencia, margen de tensión (p. ej., una tensión mínima por celda) y criterios de durabilidad de la MEA (Barbir, 2012; O'hayre et al., 2016).

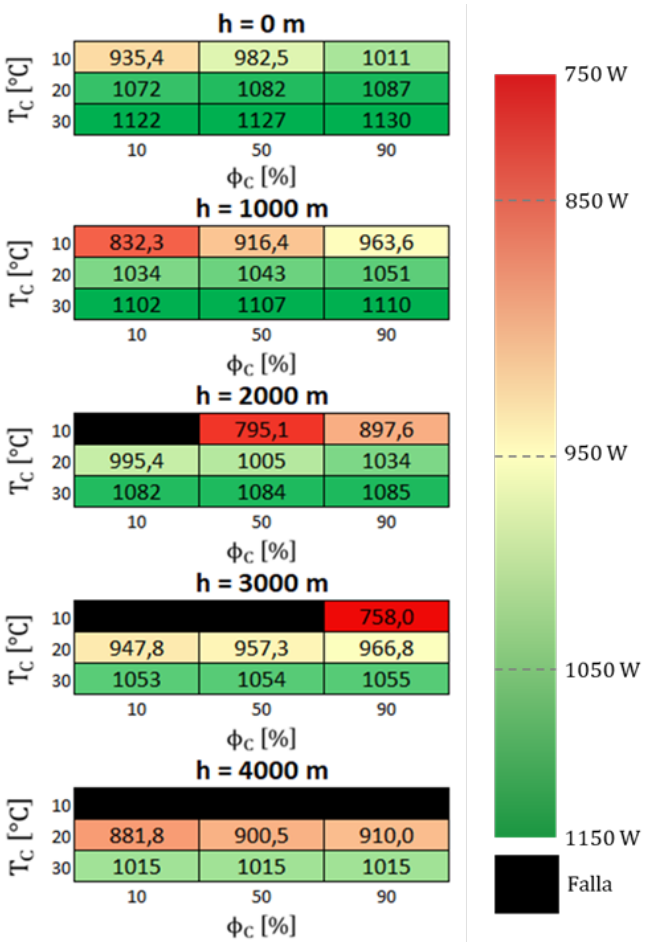


Figura 4.3: Mapa de calor de la potencia eléctrica de la pila PEM como función de la altitud, temperatura del cátodo y humedad relativa del aire bajo corriente constante (Henríquez, 2025).

La Figura 4.3 muestra un mapa de calor de la potencia eléctrica de un *stack* PEM de cátodo abierto operado a corriente fija, estimada a partir del modelo analítico estacionario presentado en Henríquez (2025). El modelo estima la potencia para combinaciones discretas de altitud, temperatura del aire del cátodo y humedad relativa del aire de alimentación. Cada matriz corresponde a una altitud; sus ejes representan la temperatura y la humedad relativa, y la escala cromática indica la potencia del *stack*. Las celdas negras indican condiciones fuera del rango operativo para la

corriente establecida, donde la tensión predicha cae por debajo del umbral mínimo y el *stack* no puede mantener ese régimen.

Bajo estos supuestos, la Figura 4.3 ilustra cualitativamente que la potencia disponible disminuye al aumentar la altitud y aumenta con la temperatura del aire del cátodo. Estos efectos se asocian principalmente a la menor presión atmosférica (y, por ende, menor presión parcial y disponibilidad de oxígeno) y a la influencia de la temperatura sobre la cinética electroquímica y las pérdidas de tensión internas del sistema. En este modelo la influencia de la humedad relativa del aire es marginal, puesto que no incorpora submodelos detallados de gestión de agua ni de cambio de fase en la MEA. Estudios experimentales y numéricos de mayor resolución muestran, sin embargo, que tanto la temperatura de operación como la humidificación de los gases de entrada son parámetros dominantes sobre el desempeño electroquímico de las PEMFC (Ozen et al., 2016; Saleh et al., 2018).

En este contexto, la curva de polarización y el mapa de operación I–V–P se emplean como referencia para interpretar el dominio de operación admisible y la respuesta de potencia frente a variaciones de temperatura, presión y humedad relativa.

4.3 Enfoques de modelado para sistemas PEMFC

El modelado de pilas de combustible PEM se ha desarrollado históricamente a partir de formulaciones analíticas y numéricas que resuelven balances de masa, cantidad de movimiento, carga y energía en una o más dimensiones espaciales. Los modelos 0D representan el sistema mediante balances globales en un volumen de control, mientras que los modelos 1D resuelven gradientes a lo largo de una coordenada espacial (por ejemplo, el espesor de la MEA). En contraste, los modelos 2D/3D del tipo CFD resuelven el acoplamiento entre transporte multifásico, electroquímica y transferencia de calor sobre la geometría de los canales y las capas porosas. Este enfoque permite estudiar fenómenos locales (arrastre de agua, distribución de corriente, gradientes térmicos) y apoyar el diseño de componentes, pero exige parametrización fisicoquímica y condiciones de borde que pueden ser difíciles de observar o medir en la práctica, además de un costo computacional elevado, especialmente en configuraciones 3D y/o transitorias (Wu, 2016).

Frente a este enfoque basado en leyes físicas, los modelos basados en datos utilizan mediciones experimentales u operacionales para aprender la relación entre variables de entrada (por ejemplo, corriente, temperatura, caudales, presiones y condiciones ambientales) y variables de salida como la tensión de celda o la potencia del *stack*. En este marco, los algoritmos de Aprendizaje Automático y Aprendizaje Profundo se entrenan sobre bases de datos experimentales o numéricas y se emplean como modelos sustitutos, evitando la resolución explícita de las ecuaciones de transporte durante la etapa de predicción. En particular, se ha reportado que la predicción de curvas I–V puede reducirse desde escalas de cientos de horas de CPU en modelos multidimensionales a tiempos del orden de

segundos en enfoques basados en datos, manteniendo una precisión adecuada dentro del dominio representado por los datos (Yang et al., 2023).

En aplicaciones orientadas a control, el desafío se centra en disponer de modelos con costo computacional reducido y precisión suficiente en un rango amplio de condiciones de operación, incluyendo regímenes transitorios y predicción de largo plazo. En este contexto, se han desarrollado tanto formulaciones mecanísticas de dimensionalidad reducida (0D y 1D) como alternativas empíricas y basadas en datos, evaluadas por su potencial de implementación en tiempo real (Zhao et al., 2021; Ding et al., 2022). En enfoques basados en datos, el desempeño depende de la distribución del conjunto de entrenamiento; por ello, su capacidad de generalización se restringe al dominio operacional representado por los datos disponibles (Zhao et al., 2021; Su et al., 2023).

A partir de lo anterior, en esta tesis se adopta un enfoque basado en datos para la predicción de la potencia eléctrica, utilizando principios físicos como marco de referencia para la interpretación de los resultados. El Capítulo 5 profundiza en el Estado del Arte sobre la aplicación de Aprendizaje Automático en PEMFC y describe, a nivel conceptual, las familias de modelos supervisados de regresión y los hiperparámetros más relevantes considerados en este trabajo.

4.4 Fundamentos generales de Aprendizaje Automático en regresión

El Aprendizaje Automático supervisado es un marco general para construir modelos predictivos a partir de datos empíricos, entendido como el estudio de algoritmos que mejoran su desempeño con la experiencia (Mitchell, 1997; Goodfellow et al., 2016). En problemas de regresión, el objetivo es aproximar una función f que asigne a cada vector de características $\mathbf{x} \in \mathbb{R}^p$ un valor escalar $y \in \mathbb{R}$, de modo que las predicciones reproduzcan el comportamiento observado del sistema físico dentro de un rango de operación relevante.

4.4.1 Fundamentos de los modelos de regresión

En un problema supervisado de regresión se dispone de un conjunto de datos $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, donde cada par $(\mathbf{x}_i, y_i)$ corresponde a una observación del sistema bajo ciertas condiciones de operación. Un modelo de regresión se describe como una familia de funciones $f_{\boldsymbol{\theta}}$, parametrizada por un vector de parámetros $\boldsymbol{\theta}$ que determina su estructura interna; por ejemplo, los pesos de una red neuronal o las reglas de partición y valores asociados a hojas en árboles de decisión (Bishop and Nasrabadi, 2006; Hastie et al., 2009).

El entrenamiento consiste en determinar los parámetros $\hat{\boldsymbol{\theta}}$ que minimizan el error entre las predicciones $\hat{y}_i = f_{\boldsymbol{\theta}}(\mathbf{x}_i)$ y los valores observados y_i. Este error se formaliza mediante una función

de pérdida $\ell(y, \hat{y})$, comúnmente basada en el error cuadrático. El problema de ajuste se expresa como:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \frac{1}{N} \sum_{i=1}^{N} \ell\big(y_i, f_{\boldsymbol{\theta}}(\mathbf{x}_i)\big) \tag{4.4}$$

En este planteamiento, el modelo se ajusta minimizando el error promedio sobre los datos disponibles (Ecuación 4.4), sin tener que fijar de antemano una única expresión para la relación entrada–salida (por ejemplo, una relación estrictamente lineal). En su lugar, es posible emplear modelos lineales y no lineales con distintos niveles de complejidad. En consecuencia, el enfoque permite capturar dependencias multivariables, siempre que los datos aporten información suficiente para identificar patrones representativos (Mitchell, 1997; Hastie et al., 2009).

En términos de ingeniería, la elección de la familia de modelos $f_{\boldsymbol{\theta}}$ implica un compromiso entre interpretabilidad, capacidad de representación y costo computacional. Modelos de baja complejidad entregan parámetros más interpretables, pero pueden ser insuficientes cuando existen interacciones no lineales marcadas; en cambio, arquitecturas de mayor complejidad requieren control de sobreajuste mediante regularización y validación para reducir la brecha de desempeño entre entrenamiento y validación/prueba (Bishop and Nasrabadi, 2006).

4.4.2 Parámetros, hiperparámetros y regularización

Es pertinente distinguir entre parámetros del modelo e hiperparámetros. Los parámetros $\boldsymbol{\theta}$ se estiman durante el entrenamiento mediante el procedimiento de optimización descrito en la subsección anterior. En cambio, los hiperparámetros $\boldsymbol{\lambda}$ definen la arquitectura del modelo y las condiciones de entrenamiento, tales como el número y la profundidad de árboles en un *ensemble*, la tasa de aprendizaje, el tamaño de la red neuronal y la magnitud de los términos de penalización, entre otros (Bishop and Nasrabadi, 2006; Goodfellow et al., 2016). Estos valores no se infieren directamente a partir del ajuste sobre los datos, sino que se determinan mediante la exploración de un espacio de configuraciones y su evaluación bajo esquemas de validación. En particular, los hiperparámetros de regularización son críticos, ya que controlan la complejidad efectiva del modelo y, en consecuencia, su desempeño de generalización.

La regularización reúne estrategias orientadas a controlar la complejidad efectiva del modelo y mitigar el sobreajuste. Una formulación estándar consiste en incorporar a la pérdida promedio un término de penalización $\Omega(\boldsymbol{\theta})$, ponderado por un hiperparámetro $\lambda > 0$, de modo que se minimiza una función objetivo regularizada del tipo:

$$\mathcal{L}_{\text{reg}}(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^{N} \ell\big(y_i, f_{\boldsymbol{\theta}}(\mathbf{x}_i)\big) + \lambda\, \Omega(\boldsymbol{\theta}) \tag{4.5}$$

La forma de $\Omega(\boldsymbol{\theta})$ determina el tipo de restricción que se impone en la regresión, favoreciendo soluciones con parámetros de menor magnitud, modelos más suaves o estructuras con menos grados de libertad efectivos (Hastie et al., 2009; Goodfellow et al., 2016). En *ensembles* basados en árboles, este control de complejidad se implementa mediante hiperparámetros estructurales, tales como la profundidad máxima y el número de árboles; y, en variantes de *gradient boosting*, mediante la tasa de aprendizaje, lo que reduce la propensión del modelo al sobreajuste.

Desde una perspectiva de ingeniería, los hiperparámetros y los mecanismos de regularización determinan el compromiso entre flexibilidad de ajuste bajo condiciones de operación heterogéneas y robustez frente a mediciones ruidosas o datos escasos en determinados rangos de operación.

4.4.3 Generalización, sesgo, varianza y compromiso de complejidad

La generalización se refiere a que el modelo mantenga un desempeño adecuado al evaluarse sobre observaciones no utilizadas en el entrenamiento, pero provenientes de la misma distribución subyacente. En términos conceptuales, el interés se centra en el riesgo esperado, definido como el valor esperado de la pérdida bajo la distribución verdadera de los datos, mientras que el entrenamiento solo proporciona una estimación del riesgo empírico a partir de una muestra finita (Bishop and Nasrabadi, 2006; Hastie et al., 2009).

El compromiso entre sesgo y varianza permite interpretar cómo la complejidad del modelo influye en su capacidad de generalización. En términos generales, modelos demasiado simples presentan un mayor error sistemático (sesgo) al no capturar patrones relevantes, lo que se traduce en errores altos tanto en entrenamiento como en validación o prueba (subajuste). En cambio, modelos excesivamente complejos pueden volverse sensibles al muestreo del conjunto de entrenamiento (varianza), ajustándose al ruido y obteniendo errores bajos en entrenamiento, pero con una pérdida de desempeño al evaluarse en datos no vistos (sobreajuste).

La Figura 4.4 ilustra de manera cualitativa cómo el error total se relaciona con las contribuciones de sesgo cuadrático y varianza, junto con un término irreducible asociado al ruido, en función de la complejidad del modelo. En el extremo izquierdo predominan modelos con alto sesgo y baja varianza (subajuste), mientras que en el extremo derecho se ubican modelos con bajo sesgo, pero varianza elevada (sobreajuste). En una zona intermedia se identifica una complejidad óptima aproximada, donde el error total se minimiza al lograr un compromiso adecuado entre estas contribuciones (Bishop and Nasrabadi, 2006; Hastie et al., 2009).

En la práctica, este compromiso se evalúa comparando métricas de desempeño entre el conjunto de entrenamiento y conjuntos de validación o prueba no utilizados durante el ajuste. La diferencia entre ambas evaluaciones puede interpretarse como una brecha de generalización y cuantificarse mediante indicadores de brecha de desempeño (GAP) definidos a partir de las métricas de error empleadas.

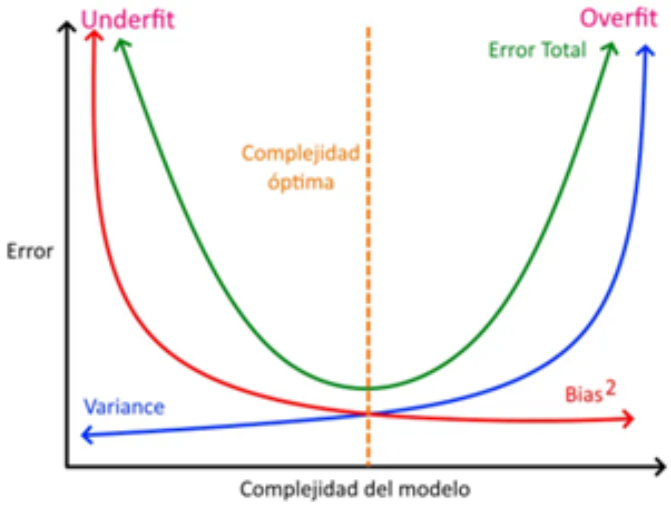


Figura 4.4: Esquema conceptual del compromiso sesgo–varianza (Cosio, 2022).

4.4.4 Validación cruzada y brecha de desempeño (GAP)

La estimación de la capacidad de generalización y la selección de hiperparámetros requieren particionar la base de datos. El esquema más simple consiste en separar un conjunto de entrenamiento y otro de prueba, pero esta división puede entregar estimaciones inestables del desempeño cuando el número de observaciones es limitado o cuando existe heterogeneidad marcada entre subgrupos de datos (Hastie et al., 2009). Para mitigar esta sensibilidad se utiliza la validación cruzada K-fold. En este esquema, el conjunto de datos se divide en K particiones de aproximadamente el mismo tamaño muestral, se entrena el modelo K veces utilizando en cada iteración $K - 1$ particiones para entrenamiento y la restante para validación; y se promedia la métrica de desempeño sobre todas las iteraciones.

Ciertos estudios comparativos clásicos han mostrado que variantes como la validación cruzada estratificada y, en particular, el esquema de 10 particiones estratificadas suele ofrecer un compromiso razonable entre sesgo, varianza y costo computacional (Kohavi et al., 1995; Arlot and Celisse, 2010). En contextos donde la base de datos se organiza en estratos bien definidos, las versiones estratificadas de la validación cruzada evitan que ciertos estratos queden ausentes en alguna iteración de entrenamiento o validación (Arlot and Celisse, 2010), como es el caso del esquema *StratifiedKFold*, que se ilustra en la Figura 4.5, donde cada estrato aporta una proporción de sus datos para ser evaluada en cada una de las iteraciones de la validación cruzada.

La comparación entre el error medido sobre los datos utilizados para entrenar el modelo y el error estimado mediante validación cruzada conduce al concepto de brecha de generalización (*generalization gap*), que en este trabajo se denota como GAP. De forma general, el GAP cuantifica la diferencia entre la métrica de error en entrenamiento y la métrica obtenida en prueba. Un GAP

reducido indica que el modelo mantiene un comportamiento similar en ambas etapas y sugiere una buena capacidad de generalización; un GAP elevado revela una discrepancia significativa que se puede asociar, en general, a sobreajuste.

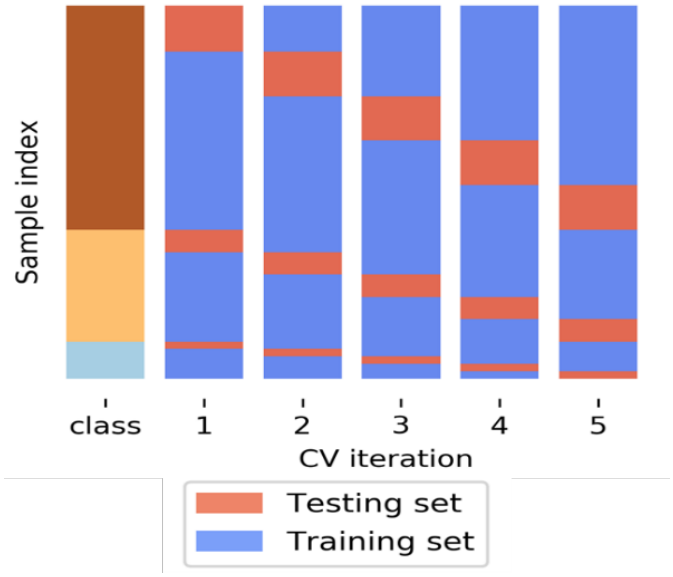


Figura 4.5: Esquema conceptual de validación cruzada *StratifiedKFold* con K=5 (Müller, 2020).

4.5 Ingeniería de características aplicada

La ingeniería de características es una etapa del flujo de trabajo en la que las variables medidas se transforman en representaciones más informativas para modelos de Aprendizaje Automático. En lugar de utilizar de forma directa todas las magnitudes disponibles, se construyen predictores derivados que concentran información relevante, reducen redundancias y facilitan el aprendizaje de relaciones entre variables (Kuhn and Johnson, 2019; Zheng and Casari, 2018). En aplicaciones de ingeniería, esta etapa permite incorporar conocimiento del proceso mediante transformaciones y combinaciones que mejoran la interpretabilidad operativa o la estabilidad numérica del modelado, y que reducen la complejidad efectiva del espacio de entrada.

4.5.1 Generalidades de la ingeniería de características

La ingeniería de características define un mapeo $\phi : \mathbb{R}^p \to \mathbb{R}^q$ que transforma el vector original de entrada $\mathbf{x}$ en un vector de características $\mathbf{z} = \phi(\mathbf{x})$. Esta transformación puede incluir reescalamientos, cambios de base, combinaciones no lineales y codificación de variables categóricas, con el objetivo de facilitar el entrenamiento y mejorar el desempeño del modelo bajo un esquema de evaluación consistente (Kuhn and Johnson, 2019). El propósito no es aumentar la dimensionalidad

sin criterio, sino obtener una representación más útil: más informativa y, cuando es pertinente, operativamente interpretable.

En problemas donde los predictores representan partes de un todo (por ejemplo, proporciones sujetas a una restricción de suma), es conveniente utilizar transformaciones que respeten las restricciones del espacio de datos y mitiguen dependencias inducidas por la propia parametrización de los componentes. De forma análoga, cuando existe heterogeneidad marcada entre regímenes de operación o condiciones de entorno, pueden emplearse descriptores agregados que resuman comportamientos típicos del sistema, reduciendo la complejidad efectiva del espacio de entrada y contribuyendo a una generalización más estable (Zheng and Casari, 2018).

4.5.2 Tipos de transformaciones y selección de características

Las operaciones de ingeniería de características se estructuran en tres bloques: transformación, creación y selección. La transformación modifica propiedades estadísticas de variables existentes —por ejemplo, estandarización, transformaciones logarítmicas o proyecciones a espacios con menor multicolinealidad— con el objetivo de mejorar su acondicionamiento numérico y facilitar el ajuste del modelo. La creación genera nuevas variables a partir de combinaciones o codificaciones de las originales, tales como variables indicadoras, términos de interacción, razones o índices derivados. La selección, por su parte, determina subconjuntos de predictores con aporte predictivo complementario y descarta aquellos redundantes o irrelevantes (Guyon and Elisseeff, 2003; Kuhn and Johnson, 2019).

Las estrategias de selección se clasifican en métodos tipo filtro, envolvente (*wrapper*) y embebidos, según si la evaluación de cada subconjunto se realiza con criterios independientes del modelo, a partir del desempeño de un modelo específico o mediante mecanismos internos del algoritmo durante el entrenamiento (Guyon and Elisseeff, 2003). En la práctica, suele combinarse la experiencia de dominio —por ejemplo, asegurar la inclusión de variables operacionales relevantes— con análisis de importancia de variables y multicolinealidad. En cualquier caso, la selección debe evitar la fuga de datos: ningún predictor ni criterio de filtrado puede incorporar información calculada a partir del conjunto de prueba, ni de la misma variable objetivo de la tarea de regresión.

4.6 Principio *No Free Lunch* y justificación del enfoque multimodelo

4.6.1 Principio *No Free Lunch* e implicancias en regresión

En el modelado predictivo de sistemas complejos, seleccionar un único algoritmo implica asumir que sus supuestos internos sobre la forma de la relación entrada–salida son adecuados para la tarea. Los teoremas *No Free Lunch* (NFL) establecen que, al promediar el desempeño sobre un conjunto suficientemente amplio de funciones objetivo o correspondencias entrada–salida, bajo distribuciones que no privilegian ninguna estructura particular, ningún algoritmo supera de forma sistemática a los demás; las mejoras en una clase de problemas se compensan con pérdidas de desempeño en otra (Wolpert and Macready, 1997; Wolpert, 2002). En consecuencia, cualquier ventaja observada en una tarea específica se atribuye a supuestos —explícitos o implícitos— sobre la clase de funciones que se desea aproximar.

En la práctica, el desempeño no es universal: depende de la coherencia entre: (i) los supuestos del algoritmo acerca del tipo de patrón que puede representar (por ejemplo, relaciones aproximadamente lineales, umbrales y particiones, o no linealidades complejas) y (ii) las propiedades estadísticas de los datos disponibles (por ejemplo, nivel de ruido, presencia de valores atípicos, colinealidad, desbalance entre regímenes de operación y heterogeneidad entre estratos) (Montgomery, 2002). Por este motivo, y dado que los teoremas NFL se formularon originalmente en el contexto de optimización tipo "caja negra" y luego se extendieron al aprendizaje supervisado, se adopta aquí este principio como fundamento metodológico: distintos modelos de regresión tienden a exhibir comportamientos diferenciados según la estructura del problema y el régimen de operación cubierto por los datos (Wolpert, 2002).

4.6.2 Enfoque multimodelo y *pipeline* comparativo

En esta tesis, este principio se operacionaliza mediante la comparación de varias familias de modelos con sesgos inductivos distintos: bosques aleatorios, variantes de *gradient boosting* basadas en árboles de decisión, redes neuronales multicapa y máquinas de vectores de soporte para regresión. La selección del modelo de referencia se sustenta en evidencia empírica obtenida bajo un esquema de evaluación homogéneo, en lugar de una preferencia previa por una única arquitectura.

Con este propósito, se adopta un enfoque multimodelo dentro de un *pipeline* común de Aprendizaje Automático. Este *pipeline* integra, de manera ordenada y reproducible, la preparación de datos, la ingeniería de características, la detección y tratamiento de valores atípicos, la partición en conjuntos de entrenamiento y validación, la búsqueda de hiperparámetros y la evaluación con métricas consistentes (Kuhn and Johnson, 2019). Un esquema usual consiste en reservar

del orden de 70 % de las observaciones para entrenamiento y 30 % para prueba, complementado con validación cruzada sobre el subconjunto de entrenamiento; de este modo, las diferencias de desempeño observadas entre modelos se atribuyen principalmente a su sesgo inductivo y a su estructura de regularización, y no a variaciones en las condiciones del procedimiento de evaluación.

La comparación entre algoritmos no se limita a identificar el menor error promedio; también incorpora criterios de robustez, estabilidad entre particiones y sensibilidad frente a distintas versiones de la base de datos. La literatura sobre evaluación de modelos supervisados enfatiza el uso de validación cruzada y propone pruebas estadísticas para reducir la probabilidad de concluir que un algoritmo es superior cuando las diferencias responden a variabilidad muestral (Dietterich, 1998).

En este marco, el enfoque multimodelo se justifica como una respuesta directa a las limitaciones señaladas por los teoremas NFL: dado que ningún algoritmo es óptimo en promedio bajo supuestos no informativos, resulta razonable explorar alternativas bajo un *pipeline* controlado y seleccionar el modelo de referencia más apropiado para la predicción de la potencia eléctrica del banco PEMFC. Los capítulos siguientes desarrollan, en primer lugar, los fundamentos teóricos de estas familias de modelos (Capítulo 5) y, posteriormente, su implementación concreta dentro del *pipeline* comparativo definido en el Capítulo 9.

Capítulo 5

Modelos y técnicas de Aprendizaje Automático

5.1 Estado del Arte: Modelos de Aprendizaje Automático aplicados a PEMFC

Entre 2022 y 2025, el modelado de pilas de combustible PEMFC mediante técnicas de Aprendizaje Automático (ML) se ha reportado como complemento a los enfoques fisicoquímicos, en particular cuando se requieren modelos sustitutos de bajo costo computacional para análisis paramétrico, control u optimización. Revisiones recientes y trabajos de referencia reportan aplicaciones orientadas a: (i) predecir curvas I–V y potencia eléctrica; (ii) estimar envejecimiento/degradación y vida útil; (iii) diagnosticar fallos; y (iv) optimizar condiciones de diseño y operación, tanto a nivel de celda como de *stack* (Ding et al., 2022; Su et al., 2023; Sharma et al., 2024; Legala et al., 2022). En conjunto, estas contribuciones abarcan redes neuronales artificiales (ANN), máquinas de vectores de soporte (SVM), modelos basados en árboles (*random forests* y variantes de *gradient boosting*) y esquemas híbridos con selección de variables y optimización, lo que motiva el uso de modelos capaces de capturar relaciones no lineales complejas a partir de datos experimentales y/o sintéticos.

Dentro de los modelos basados en árboles, los *random forests* y las familias de *gradient boosting* (por ejemplo, XGBoost y CatBoost) se reportan con capacidad para modelar no linealidades e interacciones entre variables a partir de datos experimentales y operacionales con ruido. Las revisiones de Su et al. (2023) y Sharma et al. (2024) documentan su aplicación, junto con ANN y SVM, en tareas de predicción de desempeño, diagnóstico y optimización. En 2023, Yuan et al. (2023) proponen un esquema XGBoost–Boruta para seleccionar variables del *balance of plant* y mejorar la predicción de la tensión de *stack*, reportando reducciones de RMSE de 23.8% (banco) y 14.1% (operación vehicular), con aumentos de R^2 de 0.06 y 0.04, respectivamente. En la misma línea, Zaveri

et al. (2023) emplean modelos supervisados para discriminar estados de deshidratación e inundación a partir de variables operacionales, vinculando estas condiciones con pérdidas de desempeño. Más recientemente, Zhang et al. (2025) combinan una capa convolucional unidimensional con CatBoost para predecir de manera multietapa la degradación de tensión en *stacks* PEMFC, mostrando el potencial de enfoques de *boosting* para pronóstico multietapa.

Las redes neuronales tipo perceptrón multicapa (MLP) y sus variantes constituyen una familia relevante en el modelado de PEMFC. La literatura de revisión reporta múltiples estudios en los que estas arquitecturas aproximan la relación entre condiciones operacionales y la tensión o potencia de salida, y se emplean tanto para predicción de envejecimiento/degradación como para construir modelos sustitutos de simulaciones de mayor costo computacional (Su et al., 2023; Ding et al., 2022). En paralelo, las máquinas de vectores de soporte para regresión (SVR) se han aplicado al modelado de la tensión bajo condiciones operacionales variables a partir de datos experimentales o datos sintéticos derivados de modelos fisicoquímicos, destacando su eficiencia cuando el problema se formula como regresión de una sola salida (Legala et al., 2022). En este marco, Zhong et al. (2006) constituye un antecedente temprano del uso de SVR (SVM en formulación de regresión) para modelar el desempeño de una PEMFC a partir de variables operacionales.

Las revisiones y trabajos de referencia publicados entre 2022 y 2024 indican que permanecen abiertos desafíos metodológicos en la aplicación de ML a PEMFC, especialmente en la gestión de bases de datos heterogéneas y ruidosas y en el diseño de *pipelines* reproducibles que integren de manera explícita el preprocesamiento, la selección de variables y el ajuste de hiperparámetros (Ding et al., 2022; Legala et al., 2022; Sharma et al., 2024). En este contexto, resultan especialmente relevantes modelos de regresión como *Random Forest*, XGBoost, CatBoost, redes neuronales tipo MLP y máquinas de vectores de soporte, con antecedentes tempranos (2006) y contribuciones recientes (2025) en aplicaciones de modelado de PEMFC (Zhong et al., 2006; Zhang et al., 2025). En conjunto, estos algoritmos ofrecen compromisos complementarios entre capacidad de representación, robustez frente a ruido y requerimientos de datos, y constituyen el núcleo de las técnicas supervisadas desarrolladas en las secciones siguientes.

5.2 Modelos supervisados de regresión

En este trabajo se consideran cinco modelos supervisados de regresión para predecir la potencia eléctrica del *stack* a partir de variables operacionales y ambientales: *Random Forest Regressor* (RFR), XGBoost (XGB), CatBoost (CAT), redes neuronales *feedforward* tipo perceptrón multicapa (MLP) y máquinas de vectores de soporte para regresión (SVR) (Breiman, 2001; Friedman, 2001; Chen, 2016; Prokhorenkova et al., 2018; Drucker et al., 1996; Goodfellow et al., 2016).

Todos estos modelos comparten el objetivo de aproximar una función desconocida f($\mathbf{x}$) que

mapea las variables de entrada $\mathbf{x}$ hacia la potencia eléctrica W, a partir de pares de entrenamiento $(\mathbf{x}_i, \mathrm{W}_i)$. Sin embargo, difieren en la forma en que controlan la capacidad del modelo, en los mecanismos de regularización y, cuando corresponde, en los elementos de estocasticidad durante el entrenamiento (p. ej., *bootstrap* o submuestreo), lo cual contribuye a reducir el sobreajuste (Goodfellow et al., 2016; Breiman, 2001; Chen, 2016).

En las subsecciones siguientes se describe, a nivel conceptual, el funcionamiento de cada modelo y el rol de los hiperparámetros más relevantes en términos de complejidad, regularización y estocasticidad.

5.2.1 Modelos basados en *ensembles* de árboles (RFR, XGB, CAT)

Los árboles de decisión particionan recursivamente el espacio de entrada mediante umbrales sobre las variables y, en cada hoja, aproximan la respuesta por un valor promedio (en regresión) (Hastie, 2009). Este tipo de modelo es flexible y captura no linealidades e interacciones entre variables; sin embargo, un árbol individual puede exhibir alta varianza. Para reducirla, es habitual combinar múltiples árboles en un *ensemble* mediante estrategias como *bagging* o *gradient boosting*, mejorando la estabilidad y el desempeño predictivo (Breiman, 2001; Friedman, 2001).

Random Forest (RFR)

Los bosques aleatorios combinan un gran número de árboles de decisión entrenados sobre muestras *bootstrap* (con reemplazo) del conjunto de entrenamiento y, en cada división, consideran solo un subconjunto aleatorio de características. La predicción final se obtiene promediando las salidas de todos los árboles, lo que reduce la varianza respecto de un solo árbol y mantiene, típicamente, un sesgo moderado (Breiman, 2001).

En implementaciones típicas (p. ej., `RandomForestRegressor` en *scikit-learn*), los hiperparámetros de complejidad incluyen (scikit-learn developers, 2025e):

- `n_estimators`: número de árboles del bosque; incrementarlo tiende a reducir la varianza del *ensemble* hasta una región de saturación.
- `max_depth`: profundidad máxima de los árboles; profundidades elevadas permiten ajustes más finos, pero aumentan el riesgo de sobreajuste.
- `min_samples_split` y `min_samples_leaf`: número mínimo de observaciones para dividir un nodo interno o constituir una hoja; valores mayores inducen árboles más "suaves" y con menor complejidad efectiva.

Como mecanismo de regularización estructural, en ciertas implementaciones puede emplearse la poda por complejidad (*cost–complexity pruning*), parametrizada, por ejemplo, mediante `ccp_alpha`,

que penaliza árboles con muchas hojas y favorece modelos más simples (scikit-learn developers, 2025e).

Los elementos de estocasticidad se controlan principalmente con `max_features` (proporción o número de variables candidatas en cada división) y el indicador `bootstrap` (activación del remuestreo con reemplazo). Reducir `max_features` o activar `bootstrap` aumenta la diversidad entre árboles y, en consecuencia, la robustez del *ensemble* (Breiman, 2001).

XGBoost (XGB)

El *gradient boosting* construye de manera aditiva un modelo compuesto por un conjunto de árboles de regresión de baja profundidad, donde cada árbol sucesivo se entrena para corregir los pseudo-residuos generados por el *ensemble* previo respecto de una función de pérdida dada. En su forma básica, el modelo se escribe como:

$$F_M(\mathbf{x}) = \sum_{m=1}^{M} f_m(\mathbf{x}) \tag{5.1}$$

Donde cada f_m corresponde a un árbol de regresión ajustado para aproximar los pseudo-residuos, definidos como el gradiente negativo de la función de pérdida evaluado en el *ensemble* actual (Friedman, 2001).

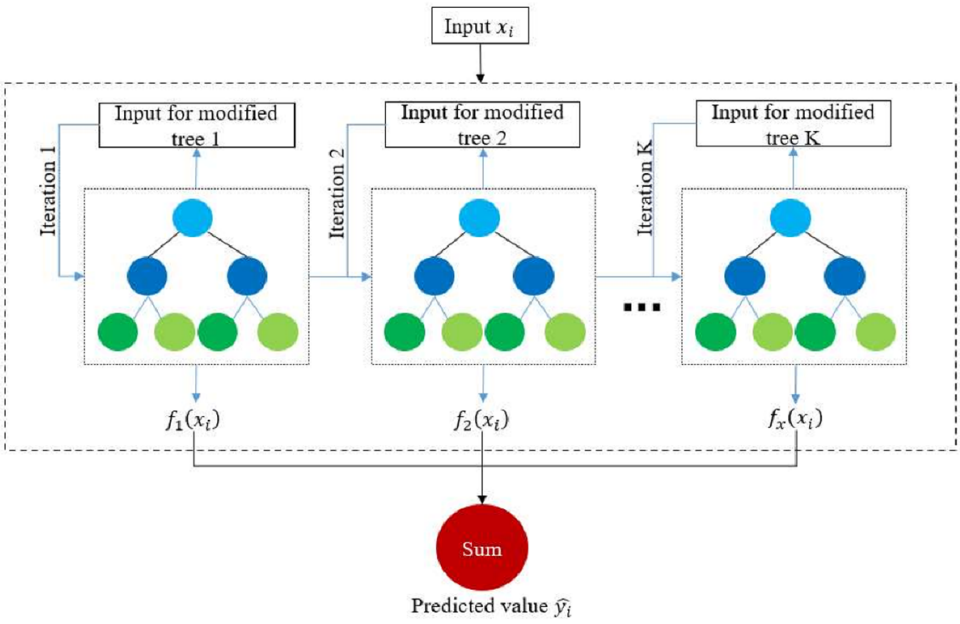


Figura 5.1: Diagrama conceptual del *ensemble* aditivo con *gradient boosting* de árboles de XGBoost (Zou et al., 2022).

La Figura 5.1 resume el esquema aditivo: cada árbol f_m se ajusta de forma secuencial a partir

del error residual del modelo construido hasta esa iteración, y su contribución se incorpora a la predicción. A diferencia del modelo *Random Forest*, cuyos árboles se entrenan de manera independiente y se combinan por promedio, en *gradient boosting* la construcción es iterativa y orientada a reducir el error acumulado.

XGBoost corresponde a una implementación optimizada de *gradient boosting* que incorpora una función objetivo regularizada, técnicas de submuestreo y optimizaciones computacionales para el manejo eficiente de grandes volúmenes de datos y matrices dispersas (Chen, 2016).

Los principales hiperparámetros de complejidad (según interfaces tipo *scikit-learn*) se resumen a continuación (scikit-learn developers, 2025a; Chen, 2016):

- `n_estimators`: número de árboles/iteraciones del *ensemble*.
- `max_depth`: profundidad máxima de cada árbol.
- `min_child_weight`: umbral mínimo de peso efectivo en el nodo hoja; valores altos restringen particiones en regiones con baja evidencia.

La regularización se controla mediante:

- `reg_lambda`: penalización L2 sobre los pesos de las hojas.
- `reg_alpha`: penalización L1 que favorece soluciones más escasas.
- `gamma`: ganancia mínima en la función objetivo requerida para aceptar una nueva partición.

Estos términos aparecen explícitamente en la función objetivo de XGBoost, que combina la pérdida de entrenamiento con un término que penaliza la complejidad del árbol (Chen, 2016).

La estocasticidad se introduce mediante submuestreo de observaciones (`subsample`) y de características (`colsample_bytree`); valores menores que 1.0 inducen diversidad y ayudan a reducir el riesgo de sobreajuste (Chen, 2016).

Finalmente, `eta` (o `learning_rate`) actúa como hiperparámetro de entrenamiento al atenuar la contribución de cada árbol mediante *shrinkage*, lo que favorece una optimización más estable a costa de un mayor número de iteraciones (Friedman, 2001; Chen, 2016).

CatBoost (CAT)

CatBoost es un algoritmo de *gradient boosting* que incorpora técnicas específicas para el tratamiento de variables categóricas y para reducir sesgos asociados a la fuga de información (*target leakage*) durante la codificación basada en la etiqueta. Dos componentes centrales son el *ordered boosting* y el uso de estadísticas de objetivo calculadas sobre permutaciones de los datos, de modo que la etiqueta de una observación no influya en la construcción de sus propias variables categóricas.

En CatBoost, los hiperparámetros de complejidad considerados habitualmente son:

- `iterations`: número máximo de iteraciones de *boosting* (o árboles en el *ensemble*).
- `depth`: profundidad máxima de los árboles de decisión internos.

La regularización se controla principalmente mediante `l2_leaf_reg`, que penaliza los valores de las hojas de los árboles de forma análoga a una regularización L2 sobre los parámetros.

Entre los elementos de estocasticidad destaca la configuración de remuestreo mediante el hiperparámetro `bootstrap_type` (por ejemplo, esquemas de *bayesian bootstrap*) y el submuestreo asociado, los cuales introducen aleatoriedad en la selección de observaciones durante el entrenamiento y contribuyen a mejorar la capacidad de generalización.

Por último, `learning_rate` controla el tamaño de paso de las actualizaciones de *boosting* y modula el compromiso entre velocidad de convergencia y estabilidad frente al sobreajuste (Prokhorenkova et al., 2018).

5.2.2 Redes neuronales *feedforward* (MLP)

Las redes neuronales *feedforward* de tipo perceptrón multicapa (*multilayer perceptron*, MLP) aproximan funciones no lineales mediante una estructura de capas de neuronas interconectadas, donde cada neurona aplica una transformación afín seguida de una función de activación no lineal. El ajuste de parámetros se realiza típicamente mediante descenso de gradiente (o variantes) y retropropagación del error (*backpropagation*), minimizando una función de pérdida sobre un conjunto de entrenamiento (Rumelhart et al., 1986; Goodfellow et al., 2016; scikit-learn developers, 2025d).

La Figura 5.2 presenta, de forma esquemática, una arquitectura MLP representativa del enfoque empleado en este trabajo, donde un conjunto de variables de entrada (ambientales y operacionales; p. ej., presión, temperatura y humedad relativa) se propaga a través de las capas ocultas de la red, y la capa de salida entrega la potencia eléctrica del sistema PEMFC.

En una MLP, la capacidad de representación está determinada principalmente por la arquitectura, es decir, por el número de capas ocultas y el número de neuronas por capa. Aumentar estas magnitudes incrementa el número de parámetros y, con ello, la capacidad para aproximar relaciones altamente no lineales, pero también el riesgo de sobreajuste y el costo computacional (Goodfellow et al., 2016).

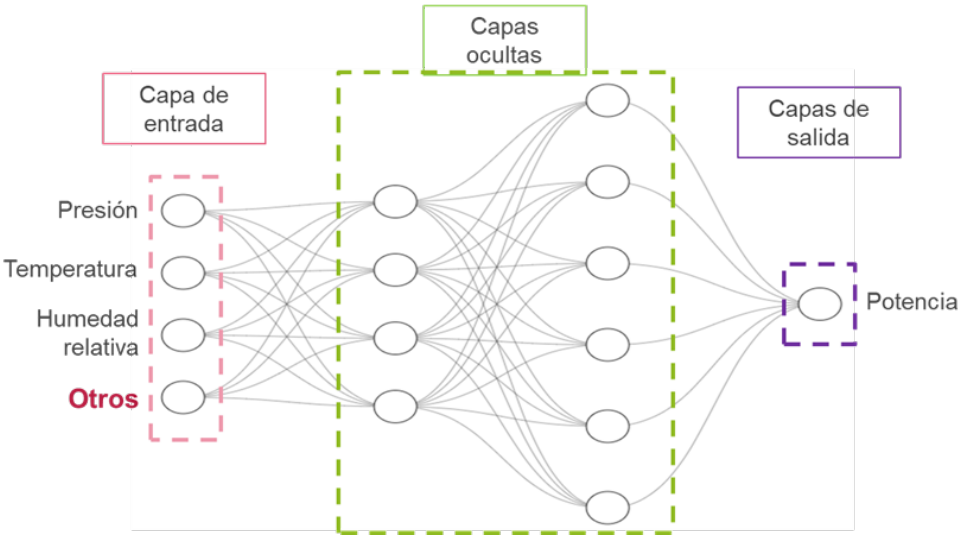


Figura 5.2: Arquitectura de las redes neuronales artificiales para predicción de potencia eléctrica del sistema PEMFC (Elaboración propia).

En implementaciones tipo *scikit-learn* (p. ej., `MLPRegressor`), la complejidad del modelo se controla principalmente mediante `hidden_layer_sizes`, que define el número de capas ocultas y el tamaño de cada una, junto con la elección de la función de activación (p. ej., `ReLU`), la cual influye tanto en la capacidad de predicción como en la estabilidad numérica del entrenamiento (scikit-learn developers, 2025c; Goodfellow et al., 2016).

La regularización explícita se incorpora habitualmente mediante un parámetro tipo `alpha`, que implementa una penalización L2 sobre los pesos de la red (*weight decay*). En términos prácticos, valores mayores de `alpha` reducen la magnitud de los pesos y favorecen soluciones más suaves, a costa de un posible aumento del sesgo (scikit-learn developers, 2025c; Goodfellow et al., 2016).

Finalmente, el proceso de entrenamiento se configura a partir de hiperparámetros como `learning_rate_init`, el número máximo de iteraciones (`max_iter`) y el uso de *early stopping* basado en el desempeño sobre un subconjunto de validación. En conjunto, estos controles permiten modular la velocidad de convergencia y mitigar el sobreajuste al interrumpir el entrenamiento cuando el error de validación deja de mejorar (scikit-learn developers, 2025c; Goodfellow et al., 2016).

5.2.3 Máquinas de vectores de soporte para regresión (SVR)

Las máquinas de vectores de soporte para regresión (*support vector regression*, SVR) extienden el principio de margen máximo de las SVM de clasificación al caso de regresión. En su formulación estándar, el objetivo es estimar una función $f(\mathbf{x})$ que mantenga la desviación $|f(\mathbf{x}_i) - \mathrm{W}_i|$ dentro de un tubo de insensibilidad de ancho 2ε para la mayor parte de las observaciones y, simultáneamente,

sea lo más "suave" posible. Estas propiedades se plantean como un problema de optimización convexa, con variables de holgura para modelar excedencias del tubo y un parámetro de penalización `C` que controla el compromiso entre suavidad del modelo y penalización por errores fuera del ε-tubo (Smola and Schölkopf, 2004; Hastie, 2009).

El principio de margen máximo se ilustra en la Figura 5.3 para el caso de clasificación binaria, donde el modelo busca el hiperplano que maximiza el margen, es decir, la distancia a las observaciones más cercanas. En SVR se mantiene el mismo criterio geométrico, sustituyendo las clases por un tubo de insensibilidad de semiancho ε alrededor de la función de regresión.

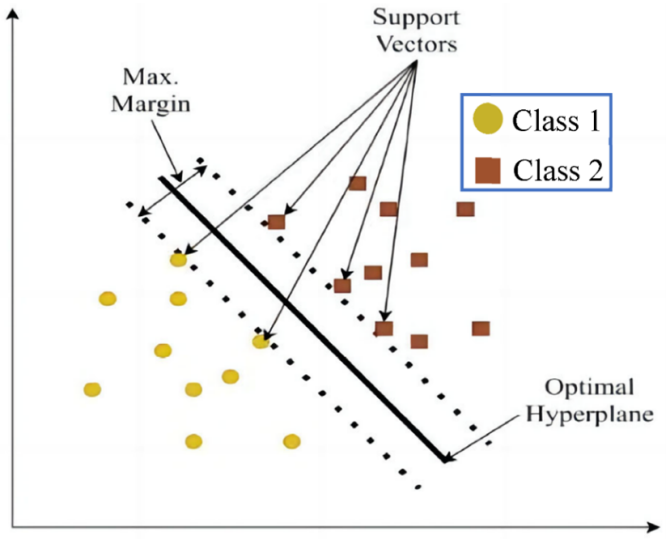


Figura 5.3: Diagrama conceptual del modelo SVR (Su et al., 2023).

Para capturar relaciones no lineales entre las variables de entrada y la variable objetivo, es habitual emplear un *kernel* de base radial (RBF), que induce una transformación implícita hacia un espacio de características de alta dimensionalidad, donde el ajuste se formula según el principio de margen máximo (Hastie, 2009). Los hiperparámetros más relevantes son `C` y `gamma` para el manejo de complejidad, y `epsilon` para regularización, cuyas descripciones se proporcionan a continuación (Smola and Schölkopf, 2004):

- `C`: controla la penalización asociada a exceder el ε-tubo; valores altos permiten ajustes más flexibles, pero pueden incrementar el riesgo de sobreajuste.
- `gamma`: parámetro del *kernel* RBF que determina el radio de acción de cada muestra; valores altos favorecen ajustes más locales y de mayor complejidad efectiva.
- `epsilon` (ε): define el ancho del tubo de insensibilidad alrededor de la función de regresión; valores mayores actúan como regularización adicional al admitir desviaciones pequeñas sin penalización explícita.

En la práctica, estos hiperparámetros se ajustan mediante estrategias de búsqueda combinadas con validación cruzada, con el fin de equilibrar desempeño de ajuste y capacidad de generalización en los modelos considerados (RFR, XGB, CAT, MLP y SVR). En la Tabla 5.1 se resumen los hiperparámetros más relevantes, clasificados por función, para las familias de regresión descritas en esta subsección.

Tabla 5.1: Resumen guía de hiperparámetros clasificados por función de la batería de modelos (clasificación indicativa; algunos hiperparámetros pueden cumplir más de un rol según implementación) (Elaboración propia).

Modelo	Complejidad	Regularización	Estocasticidad	Configuración

RFR	`n_estimators`
	`max_depth`
	`min_samples_split`
	`min_samples_leaf`
		`ccp_alpha`	`max_features`
			`bootstrap`
				–

XGB	`n_estimators`
	`max_depth`
	`min_child_weight`
		`reg_lambda`
		`reg_alpha`
		`gamma`
			`subsample`
			`colsample_bytree`
				`eta` (`learning_rate`)

CAT	`iterations` `depth`	`l2_leaf_reg`	`bootstrap_type`	`learning_rate`

MLP	`hidden_layer_sizes`	`alpha`	–	`activation` (`ReLU`)
				`solver` (`Adam`)
				`learning_rate_init`
				`max_iter`
				`early_stopping`

SVR	`C` `gamma`	`epsilon` (ε)	–	`kernel` (RBF)

5.3 Algoritmos de ingeniería de características

Las características derivadas, obtenidas a partir de las variables originales, permiten representar de forma más informativa las relaciones entre magnitudes, sintetizar la estructura de los datos y reducir redundancias. La ingeniería de características utiliza transformaciones y descriptores que incorporan conocimiento de dominio y mejoran el desempeño de los modelos supervisados, manteniendo coherencia con la interpretación física de las mediciones. En este contexto, las herramientas revisadas en esta sección se organizan en tres grupos principales orientados a capturar:

i. Estructuras composicionales y transformaciones *log-ratio*.

ii. Descriptores escalares de reparto (uniformidad o concentración).

iii. Identificación de regímenes ambientales mediante *clustering*.

El objetivo de esta sección es presentar de manera sintética estos enfoques, destacando su motivación y sus fundamentos, sin entrar en detalles formales extensos que ya se encuentran documentados en la literatura especializada.

5.3.1 Datos composicionales y transformación *isometric log-ratio* (ILR)

Los repartos de corriente entre las tres FC del banco pueden representarse como fracciones de la corriente total asociadas a cada unidad. Matemáticamente, estos vectores, con componentes no negativas y suma constante, se describen como datos composicionales. Para este caso de estudio, lo relevante son las relaciones relativas entre componentes (proporciones) y no sus valores absolutos. En presencia de ceros, las transformaciones *log-ratio* requieren un pretratamiento (por ejemplo, mediante esquemas de reemplazo o imputación) para trabajar en el dominio composicional. Aitchison propuso un tratamiento específico para este tipo de datos, basado en transformaciones *log-ratio*, que permiten aplicar técnicas estadísticas estándar mitigando correlaciones espurias asociadas a la restricción de suma constante (Aitchison, 1982; Filzmoser et al., 2018).

Dentro de esta familia, la transformación *isometric log-ratio* (ILR) mapea una composición a $\mathbb{R}^{D-1}$ mediante coordenadas ortonormales que representan "balances" entre grupos de componentes, permitiendo trabajar con geometría y métricas euclídeas equivalentes a las del espacio composicional de Aitchison (Egozcue et al., 2003).

De forma esquemática, para una composición normalizada $p = (p_1, p_2, \ldots, p_D)$, con $p_d > 0$ y $\sum_{d=1}^{D} p_d = 1$, y una partición de índices en dos grupos complementarios G_1 y G_2 de tamaños r y s ($r + s = D$), una coordenada ILR puede escribirse como:

$$z = \sqrt{\frac{rs}{r+s}} \ln \left[\frac{\left(\prod_{i \in G_1} p_i \right)^{1/r}}{\left(\prod_{j \in G_2} p_j \right)^{1/s}} \right] \tag{5.2}$$

En un banco PEMFC compuesto por tres FC ($D = 3$), la transformación ILR convierte el vector composicional en $D - 1$ coordenadas ortonormales, que cuantifican proporciones *log-ratio* normalizadas entre grupos de unidades del banco y pueden incorporarse como predictores en modelos supervisados. Los detalles de la formulación se encuentran en las referencias (Aitchison, 1982; Egozcue et al., 2003; Filzmoser et al., 2018).

5.3.2 Entropía balanceada de Shannon

Además de las coordenadas ILR, es útil disponer de un indicador escalar que resuma cuán uniforme o concentrado es un reparto. Para ello se adopta la entropía de Shannon, ampliamente utilizada como medida de dispersión o diversidad (Shannon, 1948; Cover, 1999).

Para una composición $p = (p_1, \ldots, p_D)$, la entropía de Shannon se define como:

$$H(p) = -\sum_{d=1}^{D} p_d \ln(p_d) \tag{5.3}$$

y toma valores bajos cuando la distribución se concentra en pocas componentes, alcanzando su máximo cuando el reparto es uniforme ($p_d = 1/D$ para todo d).

Con el fin de comparar repartos con distinto número de componentes, se utiliza una versión normalizada o "balanceada":

$$H_{\text{bal}}(p) = \frac{H(p)}{\ln(D)} \tag{5.4}$$

que toma valores entre 0 y 1. Valores cercanos a 1 indican un reparto homogéneo entre componentes, mientras que valores próximos a 0 reflejan una fuerte concentración en un subconjunto reducido.

En un banco PEMFC con D componentes, H_{bal} entrega una medida compacta de uniformidad del reparto entre unidades y facilita comparaciones entre condiciones de operación, complementando la información direccional provista por las coordenadas ILR (Shannon, 1948; Cover, 1999).

5.3.3 Clusterización *k-means* de variables ambientales

La clusterización es una técnica utilizada en el Análisis Exploratorio de Datos (EDA) para investigar la estructura de los datos, mediante la identificación de subgrupos (clústeres) de observaciones similares bajo una medida de distancia o similitud. En particular, k-means es un algoritmo no supervisado que particiona los datos en K clústeres, donde K se fija a priori.

En la Figura 5.4 se ilustra la aplicación de *k-means* sobre un conjunto de datos aleatorios, donde los clústeres asignados se distinguen por color.

Formalmente, busca una partición $S = \{S_1, \ldots, S_K\}$ de un conjunto de vectores $y_i \in \mathbb{R}^M$ en K clústeres no vacíos y no solapados, junto con sus centroides $C = \{c_1, \ldots, c_K\}$, minimizando la suma de distancias euclídeas cuadráticas intra-clúster (Hastie, 2009). La función objetivo, denominada distorsión o inercia, se expresa como:

$$J = \sum_{k=1}^{K} \sum_{i \in S_k} \|y_i - c_k\|_2^2 \tag{5.5}$$

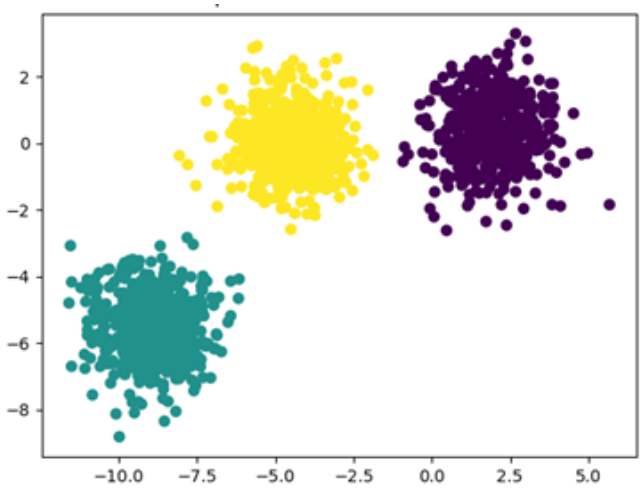


Figura 5.4: Visualización de *k-means* sobre grupos de datos generados aleatoriamente (scikit-learn developers, 2025b).

El algoritmo opera de forma iterativa bajo los siguientes pasos:

- Inicializar K centroides (por ejemplo, de forma aleatoria o mediante k-means++) (Arthur and Vassilvitskii, 2006).
- Asignar cada observación al centroide más cercano según distancia euclidiana.
- Recalcular cada centroide como la media de las observaciones asignadas al clúster.
- Repetir hasta que la asignación o la función objetivo J no cambien significativamente.

En muchos problemas prácticos, el valor de K no es conocido a priori y suele seleccionarse mediante criterios heurísticos. Una alternativa habitual es el *elbow method*, que consiste en evaluar J para distintos valores de K e identificar un punto a partir del cual la reducción marginal de la inercia se vuelve poco significativa, generando un "codo" visible en la curva (Kodinariya et al., 2013).

En el análisis de variables ambientales, k-means puede utilizarse para agrupar observaciones en regímenes similares, de modo que la etiqueta de clúster actúe como un descriptor discreto del contexto ambiental y facilite análisis comparativos entre condiciones.

5.4 Técnicas de diagnóstico de calidad de datos y detección de *outliers*

En bases de datos experimentales es habitual encontrar mediciones atípicas asociadas a errores instrumentales, fallas transitorias de operación o condiciones poco representativas del régimen

regular del proceso. Antes de entrenar modelos predictivos, conviene disponer de herramientas que permitan identificar estas observaciones de manera consistente y reproducible. Para abordar esta problemática pueden combinarse tres familias complementarias de métodos: (i) estadísticos robustos univariantes, apoyados en diagramas de caja y bigote; (ii) enfoques multivariantes basados en *PCA* y en los estadísticos de control T^2 y SPE; y (iii) algoritmos de *clustering* por densidad, en particular DBSCAN, que distinguen regiones densamente muestreadas de observaciones aisladas interpretadas como ruido.

5.4.1 Diagramas de caja y bigote (*boxplots*)

Los diagramas de caja y bigote (*boxplots*) fueron introducidos por Tukey como una herramienta de análisis exploratorio para resumir la distribución de una variable mediante la mediana, los cuartiles y los "bigotes", privilegiando medidas robustas frente a valores extremos y distribuciones asimétricas (Tukey, 1977). En comparación con resúmenes basados en la media y la desviación estándar, la mediana y el rango intercuartílico ($\mathrm{IQR} = \mathrm{Q}_3 - \mathrm{Q}_1$) reducen la influencia de *outliers* en la caracterización de la variable.

Para detección univariante se emplean con frecuencia las "vallas de Tukey": un dato se considera potencialmente atípico si cae fuera del intervalo $(\mathrm{Q}_1 - \mathrm{k}\cdot\mathrm{IQR},\ \mathrm{Q}_3 + \mathrm{k}\cdot\mathrm{IQR})$, donde $\mathrm{k} = 1.5$ se asocia a valores moderadamente alejados y $\mathrm{k} = 3.0$ a valores extremadamente atípicos (Tukey, 1977). Para distribuciones fuertemente sesgadas se han propuesto ajustes de estos umbrales, manteniendo la lógica de cuantiles robustos sin recurrir a supuestos de normalidad (Schwertman et al., 2004). En calidad de datos, los *boxplots* constituyen así un primer filtro para identificar observaciones incompatibles con el comportamiento central de una variable medida.

5.4.2 PCA y estadísticos de control multivariante

Cuando se analizan simultáneamente múltiples variables de proceso, suele existir correlación entre ellas, lo que limita la eficacia de criterios estrictamente univariantes. El análisis de componentes principales (PCA) proyecta el vector de observaciones $\mathbf{x} \in \mathbb{R}^{\mathrm{p}}$ sobre un subespacio de menor dimensión, definido por combinaciones lineales ortogonales (componentes principales) que capturan la mayor fracción de varianza total (Jolliffe, 2011). Esta representación permite describir el comportamiento dominante del conjunto de datos en un espacio reducido y distinguir variaciones sistemáticas de componentes residuales asociadas a ruido de medición.

A partir de un modelo PCA se definen estadísticos de control multivariante sobre las componentes principales seleccionadas. El estadístico de Hotelling T^2 cuantifica la distancia de una observación respecto de la media en el espacio de puntuaciones (*scores*), ponderando las varianzas asociadas a cada componente. En forma simplificada, si $\mathbf{t}$ es el vector de puntuaciones y $\mathbf{\Lambda}$ es la matriz

diagonal de varianzas de las componentes principales consideradas, T^2 es proporcional a $\mathbf{t}^\top \mathbf{\Lambda}^{-1} \mathbf{t}$. Bajo supuestos de normalidad, este estadístico puede compararse con umbrales derivados de distribuciones teóricas (por ejemplo, tipo F o aproximaciones χ^2) o mediante cuantiles empíricos, según el tamaño muestral y el ajuste del modelo.

El segundo estadístico es el error de predicción cuadrático, SPE (*Squared Prediction Error*), también conocido como estadístico Q. SPE cuantifica la magnitud del residuo de reconstrucción, es decir, la componente de $\mathbf{x}$ que no es explicada por las componentes principales seleccionadas. Valores altos de SPE sugieren comportamientos no capturados por la estructura multivariante dominante, incluso si T^2 es moderado.

En conjunto, T^2 y SPE sustentan esquemas de monitoreo multivariante en los que los *outliers* se asocian a observaciones que exceden los límites de confianza en uno o ambos estadísticos (Kourti and MacGregor, 1996).

5.4.3 DBSCAN *clustering*

Además de enfoques basados en estadísticos de dispersión, es posible detectar *outliers* examinando la estructura de densidad del conjunto de datos. DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) es un algoritmo de *clustering* no supervisado que agrupa puntos en regiones de alta densidad separadas por regiones de baja densidad, y marca explícitamente como ruido aquellos puntos que no pertenecen a ningún grupo denso (Ester et al., 1996).

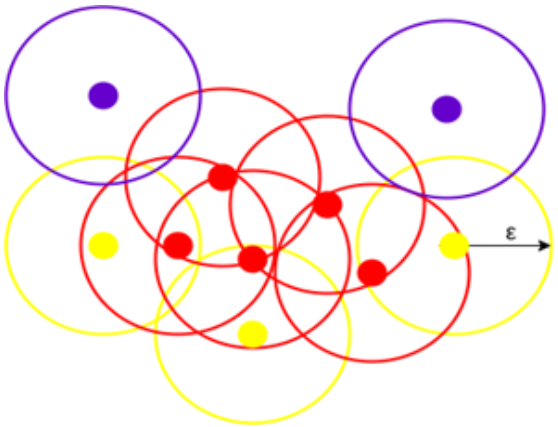


Figura 5.5: Esquema conceptual de DBSCAN para puntos núcleo, frontera y ruido (Sharma, 2020).

El algoritmo se basa en dos hiperparámetros: el radio de vecindad ε y el mínimo número de puntos MinPts requerido dentro de ese vecindario. Para cada observación se contabilizan sus vecinos dentro de un radio ε. Si el número de vecinos es al menos MinPts, el punto se clasifica como núcleo; si tiene menos vecinos pero se encuentra dentro del vecindario de un núcleo, se clasifica como frontera; y si no cumple ninguna de las anteriores, se etiqueta como ruido. Los clústeres se

construyen expandiendo recursivamente regiones conectadas a partir de puntos núcleo y asignando puntos frontera alcanzables, mientras que los puntos no alcanzables quedan etiquetados como ruido. En la Figura 5.5 se ilustra el criterio de clasificación con MinPts = 3.

Una ventaja de DBSCAN frente a algoritmos particionales como k-means es que no requiere fijar a priori el número de clústeres y puede capturar grupos de forma arbitraria; no obstante, su desempeño se ve afectado cuando coexisten regiones con densidades muy distintas (Ester et al., 1996; Schubert et al., 2017). En particular, el método es sensible a la elección de ε y MinPts (Schubert et al., 2017). En la práctica se utilizan heurísticas como el gráfico k-*distance* (ordenando la distancia al k-ésimo vecino más cercano) para seleccionar un rango razonable de ε, y reglas de tipo MinPts $\geq$ d + 1, donde d es la dimensión del espacio de características (Kumar, 2024; GeeksforGeeks, s. f.; Schubert et al., 2017). En bases de datos con ruido heterogéneo, la calibración de estos parámetros suele apoyarse en inspección visual y en el conocimiento del fenómeno físico observado (Schubert et al., 2017).

Desde la perspectiva de calidad de datos, DBSCAN aporta una visión geométrica basada en densidad que complementa los criterios univariantes y multivariantes: los puntos etiquetados como ruido corresponden a observaciones aisladas respecto de los regímenes densamente muestreados, por lo que constituyen candidatos naturales a ser tratados como *outliers* en etapas posteriores de depuración.

5.5 Modelo probabilístico *leaky noisy-OR*

En problemas donde se requiere integrar evidencia parcial proveniente de múltiples detectores sobre una misma observación, resulta conveniente evitar el uso de tablas de probabilidad condicional cuyo tamaño crece exponencialmente con el número de entradas. Para este tipo de combinación es habitual emplear el modelo *noisy-OR*, en el cual un nodo binario Y (por ejemplo, "la observación es *outlier*") depende de varios nodos binarios X_k ("el detector k marca la observación como sospechosa"), mediante parámetros que representan mecanismos de activación independientes asociados a cada padre (Koller and Friedman, 2009; Jianxing et al., 2021).

5.5.1 Modelo *noisy-OR*

Sea $Y \in \{0, 1\}$ el nodo hijo y $X = (X_1, \ldots, X_m)$ el conjunto de nodos padre binarios. Denotamos por $\mathrm{Pa}(Y) = \{1, \ldots, m\}$ el conjunto de índices de los padres de Y y por $A(x) = \{k \in \mathrm{Pa}(Y) : x_k = 1\}$ el conjunto de padres activos para la configuración $X = x$. El modelo *noisy-OR* supone que:

- Cada X_k es una causa potencial de Y.
- Los mecanismos de las causas actúan de forma independiente.

A cada padre se le asocia un parámetro de enlace $\theta_k \in [0, 1]$:

$$\theta_k = \mathrm{P}\big(Y = 1 \mid X_k = 1,\ X_j = 0\ \forall j \neq k\big) \tag{5.6}$$

que representa la probabilidad de que la causa k sea suficiente para producir el efecto cuando es la única activa. Bajo independencia de mecanismos, la probabilidad de que el efecto se active en presencia del conjunto de causas $A(x)$ viene dada por:

$$\mathrm{P}(Y = 1 \mid X = x) = 1 - \prod_{k \in A(x)} (1 - \theta_k) \tag{5.7}$$

Es decir, el efecto no se activa únicamente si todas las causas activas fallan simultáneamente, y la probabilidad de ese evento es el producto de los fallos individuales $1 - \theta_k$. Esta parametrización reemplaza una tabla condicional con 2^m configuraciones por solo m parámetros θ_k, manteniendo una interpretación directa del aporte de cada causa.

5.5.2 Extensión *leaky noisy-OR*

En el modelo básico, si ningún padre está activo ($A(x) = \varnothing$), se cumple $\mathrm{P}(Y = 1 \mid X = \mathbf{0}) = 0$. En sistemas reales suele existir una probabilidad residual de que el evento ocurra por causas no modeladas explícitamente. Para capturar estos factores se introduce un nodo de fuga X_L, que agrupa causas no representadas, parametrizado por una probabilidad de fuga $\theta_L \in [0, 1)$ (Jianxing et al., 2021).

Bajo el supuesto de que la fuga es independiente del resto de causas, la probabilidad de activación del efecto adopta la forma:

$$\mathrm{P}(Y = 1 \mid X = x) = 1 - (1 - \theta_L) \prod_{k \in A(x)} (1 - \theta_k) \tag{5.8}$$

El término $1 - \theta_L$ representa la probabilidad de que la fuga no active el efecto. Si no hay padres explícitos activos, la expresión se reduce a:

$$\mathrm{P}(Y = 1 \mid X = \mathbf{0}) = \theta_L \tag{5.9}$$

Esto refleja que el evento puede ocurrir aun cuando ninguno de los detectores modelados se active. Esta extensión, conocida como *leaky noisy-OR*, se ha utilizado en el modelado de fallos de sistemas complejos y redes bayesianas difusas, donde resulta inviable enumerar todas las causas posibles (Jianxing et al., 2021). La Figura 5.6 muestra esquemáticamente este concepto.

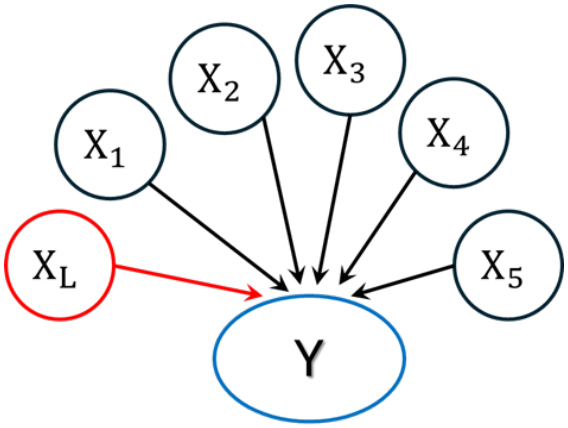


Figura 5.6: Esquema modelo *leaky noisy-OR* (Elaboración propia).

5.6 Optimización bayesiana de hiperparámetros

El desempeño de los modelos de regresión depende de hiperparámetros que no se ajustan durante el entrenamiento estándar. Su selección puede formularse como un problema de optimización de "caja negra", donde cada evaluación consiste en entrenar el modelo con una configuración h y estimar su desempeño mediante validación cruzada. En este marco, se define la función objetivo como:

$$f(h) = \mathrm{RMSE}_{\mathrm{CV}}(h) \tag{5.10}$$

Donde $\mathrm{RMSE}_{\mathrm{CV}}(h)$ corresponde al promedio, bajo un esquema de validación cruzada común a los modelos, de la métrica RMSE obtenida para la configuración h (Snoek et al., 2012; Shahriari et al., 2015; Frazier, 2018).

En enfoques clásicos como la búsqueda en malla (*grid search*), el número de evaluaciones crece rápidamente con la dimensión del espacio de búsqueda cuando se discretizan múltiples hiperparámetros, lo que puede volver prohibitivo el ajuste de modelos complejos. La búsqueda aleatoria (*random search*) mitiga este efecto al muestrear configuraciones al azar bajo un presupuesto fijo, aunque en espacios amplios (i.e. de mayor número de hiperparámetros) puede requerir muchas evaluaciones para localizar regiones competitivas. La optimización bayesiana ofrece una alternativa orientada a mejorar la eficiencia muestral: construye un modelo probabilístico sustituto (*surrogate model*) de la función objetivo y lo utiliza para seleccionar de forma adaptativa qué configuraciones evaluar, concentrando el presupuesto computacional en regiones prometedoras (Snoek et al., 2012; Shahriari et al., 2015).

El procedimiento parte de un conjunto inicial de evaluaciones $\{(h^{(n)}, f(h^{(n)}))\}$. Con estos datos se ajusta el sustituto —a menudo un proceso gaussiano u otro modelo que entregue una media predictiva y una medida de incertidumbre— y se define una función de adquisición $\alpha(h)$ que

equilibra exploración (zonas con alta incertidumbre) y explotación (bajo valor esperado de $f(h)$). Criterios como *expected improvement* o *upper confidence bound* son ejemplos habituales (Shahriari et al., 2015; Frazier, 2018). En cada iteración se maximiza $\alpha(h)$, se evalúa la función objetivo en la nueva configuración y se actualiza el sustituto, repitiendo el ciclo hasta alcanzar un número máximo de evaluaciones o hasta que la mejora se vuelve marginal.

En aplicaciones prácticas, la optimización bayesiana se utiliza para ajustar hiperparámetros dentro de espacios de búsqueda acotados, definidos por restricciones del modelo y consideraciones de estabilidad numérica. La calidad del resultado depende tanto del sustituto y la función de adquisición como de la especificación de rangos plausibles y del presupuesto de evaluaciones disponible (Snoek et al., 2012; Shahriari et al., 2015; Frazier, 2018).

5.7 Evaluación integrada de los modelos de regresión

5.7.1 Métricas de evaluación

Existe una amplia variedad de métricas de evaluación de la calidad y precisión de las predicciones de los modelos de regresión. A continuación, se presentan algunas definiciones claves y métricas utilizadas en este estudio (Legates and McCabe Jr, 1999; Willmott and Matsuura, 2005).

Los errores o residuos se calculan como la diferencia entre las observaciones y predicciones:

$$e_i = y_i - \hat{y}_i \tag{5.11}$$

El error medio (ME) indica la dirección y magnitud promedio del sesgo del modelo. Un ME > 0 implica subestimación y ME < 0, sobrestimación:

$$\text{ME} = \frac{1}{n}\sum_{i=1}^{n} e_i \tag{5.12}$$

La desviación estándar del error (SDE) cuantifica el grado de dispersión de los errores:

$$\text{SDE} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\big(e_i - \text{ME}\big)^2} \tag{5.13}$$

La raíz cuadrada del error cuadrático medio ($\text{RMSE} = \sqrt{\text{MSE}}$) cuantifica el tamaño típico de los errores de regresión:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2} \tag{5.14}$$

Se encuentra en la misma escala de la variable objetivo y es sensible a *outliers* (errores elevados al cuadrado en la sumatoria).

El Coeficiente de Correlación de Concordancia (CCC) mide el grado de acuerdo entre las observaciones y predicciones, considerando tanto el sesgo como la correlación (Lawrence and Lin, 1989). CCC penaliza la desviación respecto a la recta $y = \hat{y}$. Una forma típica es:

$$\text{CCC} = \frac{2r\,\sigma_y\,\sigma_{\hat{y}}}{\sigma_y^2 + \sigma_{\hat{y}}^2 + (\bar{y} - \bar{\hat{y}})^2} \tag{5.15}$$

Donde r es el coeficiente de correlación de Pearson, y $\bar{y}$ y σ son las medias aritméticas y desviaciones estándares muestrales, los cuales se definen según:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \qquad \bar{\hat{y}} = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i \tag{5.16}$$

$$\sigma_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}, \qquad \sigma_{\hat{y}} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2} \tag{5.17}$$

El coeficiente de correlación de Pearson r mide la magnitud y dirección de la relación lineal entre las variables en un rango de $[-1, 1]$:

$$\text{r} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\,\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}} \tag{5.18}$$

Al elevarlo al cuadrado se obtiene el coeficiente de determinación R^2, el cual cuantifica la proporción de la varianza explicada por el modelo a partir de los descriptores (bondad de ajuste).

El Coeficiente de Eficiencia de Modelado (MEC) cuantifica la eficiencia global del modelo por sobre utilizar la media de las observaciones como predictores (Nash and Sutcliffe, 1970):

$$\text{MEC} = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{5.19}$$

5.7.2 Diagrama solar

El diagrama solar es una representación visual del rendimiento de los modelos en términos de sesgo y dispersión de los errores, ambos estandarizados por la desviación estándar de las observaciones σ_y (escala invariante para comparabilidad). Para ello, aprovecha la relación pitagórica deducida a partir de la descomposición de RMSE (Wadoux et al., 2022):

$$(\text{RMSE}^*)^2 = (\text{ME}^*)^2 + (\text{SDE}^*)^2 \tag{5.20}$$

Donde:

$$(\text{RMSE}^*)^2 = \frac{\text{RMSE}}{\sigma_y}, \qquad (\text{ME}^*)^2 = \frac{\text{ME}}{\sigma_y}, \qquad (\text{SDE}^*)^2 = \frac{\text{SDE}}{\sigma_y} \tag{5.21}$$

En un sistema de coordenadas cartesianas, el eje X representa ME^*, el eje Y corresponde a SDE^* y la distancia de los puntos al origen, RMSE^*. El diagrama incluye bandas con rangos del coeficiente de correlación. Alternativamente, se pueden colorear los puntos según MEC o ajustar tamaño según valores de otra métrica suplementaria (por ejemplo, CCC), consolidando información de múltiples métricas.

De este modo, los modelos más cercanos al origen $(0, 0)$ tienen menor sesgo y menor dispersión del error, es decir, tienen mejor desempeño. En la Figura 5.7 se muestra el diagrama solar del trabajo de Wadoux et al. (2022).

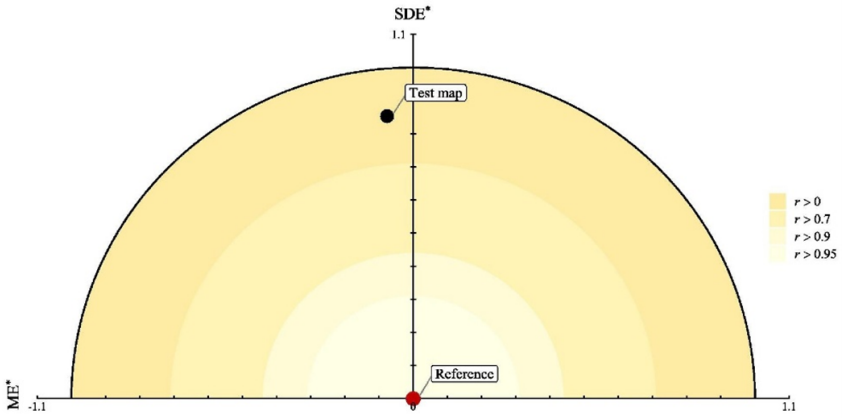


Figura 5.7: Ejemplo de diagrama solar (Wadoux et al., 2022).

5.7.3 Diagrama de Taylor

El Diagrama de Taylor representa visualmente el desempeño de un modelo en un sistema de coordenadas polares (Figura 5.8). El radio corresponde a la desviación estándar normalizada de las predicciones:

$$\sigma^* = \frac{\sigma_{\hat{y}}}{\sigma_y} \tag{5.22}$$

La posición angular se asigna al coeficiente de correlación r y se agregan contornos de RMSE^* centrado sustrayendo el sesgo ($\text{cRMSE}^* = \text{SDE}^*$), que mide el grado de predicción de la dispersión del modelo.

Los modelos más próximos al punto de referencia $(1, 0)$ tienen mejor desempeño en términos de estas métricas, lo que se traduce como mejor ajuste, mayor correlación y mejor reproducción de la dispersión de la curva de regresión (Taylor, 2001).

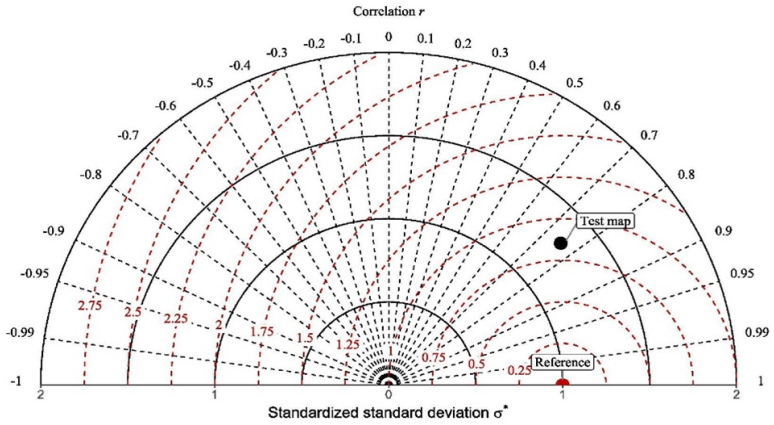


Figura 5.8: Ejemplo de diagrama de Taylor (Wadoux et al., 2022).

En este capítulo se han presentado los fundamentos teóricos de los modelos de regresión considerados, las herramientas de ingeniería de características, los esquemas de depuración de datos y combinación probabilística de detectores, así como los criterios de optimización de hiperparámetros y evaluación del desempeño. En conjunto, estos elementos constituyen el andamiaje metodológico sobre el cual se construye el *pipeline* de modelado desarrollado en los capítulos posteriores, donde se detallan su implementación específica y los resultados obtenidos para el banco PEMFC de la planta piloto móvil.

Capítulo 6

Caracterización de la base de datos y análisis exploratorio (EDA)

En este capítulo se describe la base de datos (BD) construida a partir de las mediciones de la campaña itinerante de la planta piloto móvil de CICITEM del desempeño de la PEMFC GenSure E-1100, la cual se desarrolló bajo el contexto climático y topográfico de la Región de Antofagasta. Se presenta un análisis exploratorio de los datos (EDA) univariante y bivariado con el propósito de caracterizar la BD. Este diagnóstico permitió definir las consideraciones metodológicas que condicionaron el diseño de la depuración de datos anómalos y adoptar las medidas necesarias para el modelado, coherentes con la estructura de los datos.

6.1 Descripción general y control de calidad básico

La campaña de CICITEM se desplegó en cinco sitios de la Región de Antofagasta: Tocopilla, SanPedro, Calama, Chacabuco y PSDA (se usan las etiquetas originales). Se realizaron mediciones de variables ambientales (temperatura ambiente T_{amb}, presión atmosférica p_{amb} y humedad relativa HR) y variables eléctricas del banco compuesto por 3 pilas de combustible (corriente I, tensión V y potencia W). La Tabla 6.1 proporciona un resumen de las estadísticas descriptivas elementales de la base de datos (1595 registros).

Con el fin de ilustrar la dinámica temporal de la operación típica de la planta, la Figura 6.1 presenta una serie de tiempo de las variables ambientales de Calama. Por otra parte, la Figura 6.2 muestra la evolución de las potencias individuales de las pilas de combustible (W_1, W_2, W_3), junto con la tensión y corriente de la primera PEMFC (V_1, I_1). La operación del sistema es principalmente nocturna con una mayor participación de la primera pila con ciclos de encendido y apagado. La tensión se mantiene constante, mientras que la corriente responde dinámicamente a las condiciones ambientales y demanda interna de la planta.

Tabla 6.1: Resumen estadístico de las variables ambientales y eléctricas de la base de datos.

Variable	Promedio	Desv. est.	Mín	Q1	Q2	Q3	Máx
T_{amb} (°C)	16.2	6.2	5.2	11.8	15.6	20.1	40.9
p_{amb} (bar)	0.870	0.078	0.764	0.782	0.898	0.906	0.974
HR (%)	32.0	22.8	1.4	14.0	21.0	42.0	80.0
I (A)	10.1	5.4	0.1	5.5	10.1	14.2	27.4
(V)	51.2	1.0	47.4	50.6	50.6	52.4	56.9
W (W)	519.4	275.6	6.1	278.0	511.8	736.8	1381.3

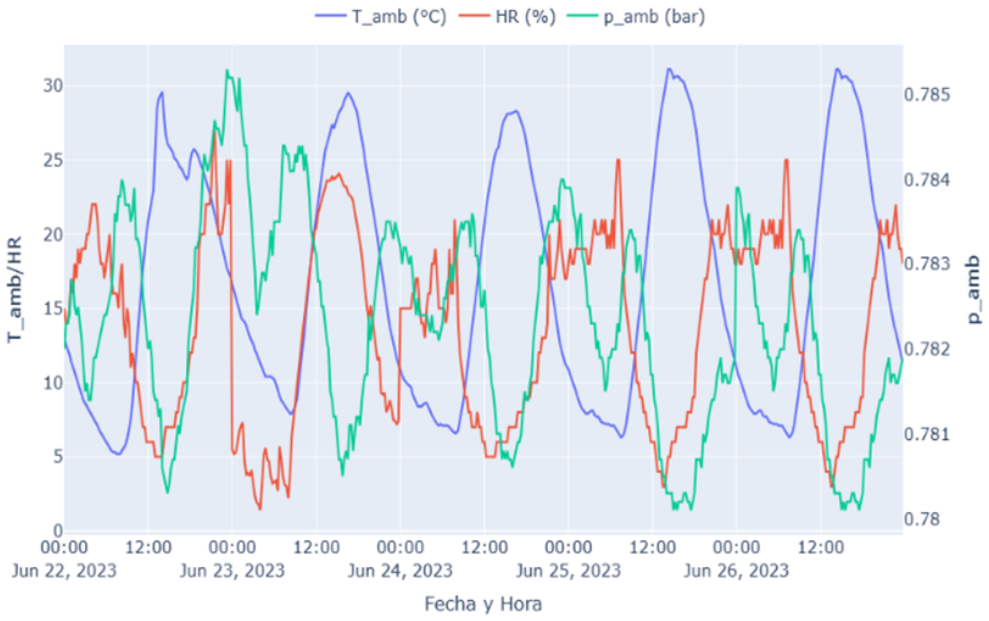
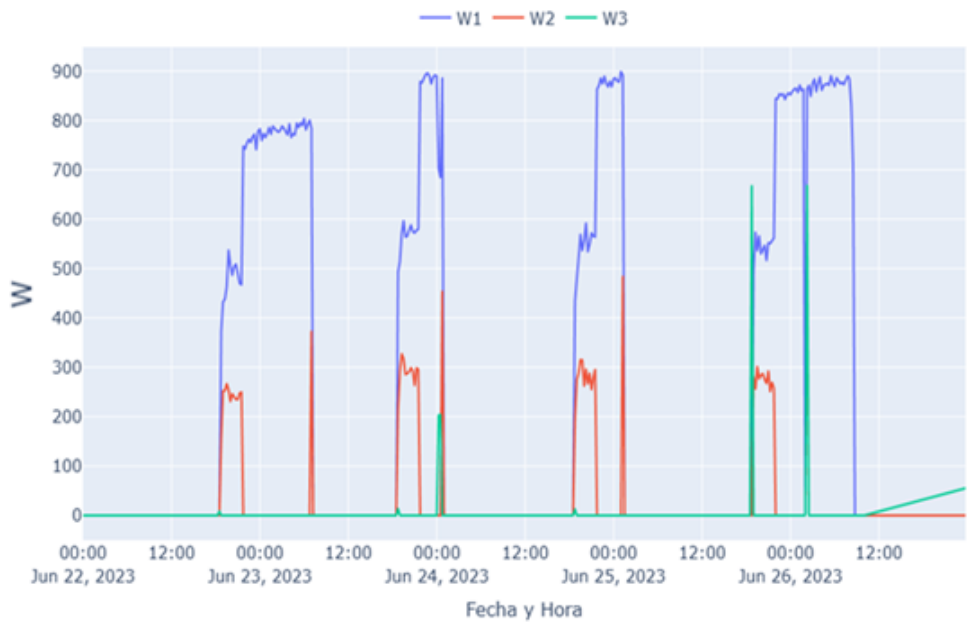
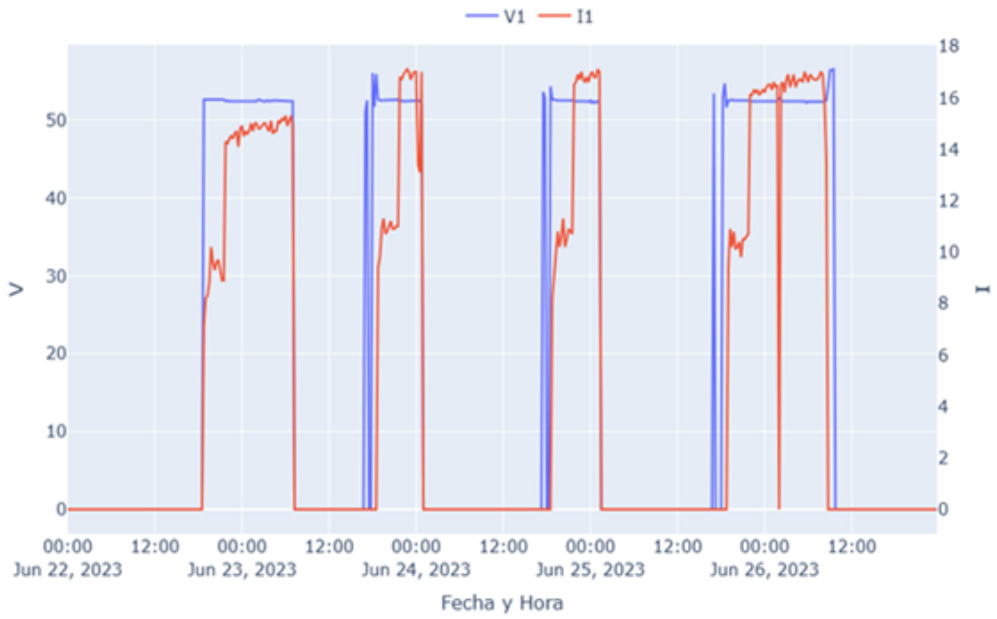


Figura 6.1: Serie de tiempo de la temperatura, humedad relativa y presión registradas en Calama.

Como control de calidad se verificó la consistencia eléctrica entre la potencia medida W y la potencia calculada $P = I \cdot V$. Esta tarea permitió verificar desajustes en la adquisición de datos y detectar mediciones inconsistentes. En la Figura 6.3 se muestra la distribución del error absoluto $e = |W - P|$ del banco de pilas de combustible en la campaña de Calama. En general, el error está acotado para los períodos de operación, lo cual se atribuye a la variabilidad instrumental de sensores. En contraparte, hay registros de la última sesión operativa que presentan un comportamiento que discrepa de la dinámica regular, por lo que fueron descartados al no pasar este primer filtro.



(a)



(b)

Figura 6.2: Series de tiempo de variables eléctricas: (a) potencias eléctricas de las pilas de combustible (W1, W2, W3); (b) tensión y corriente eléctricas medidas (V1, I1).

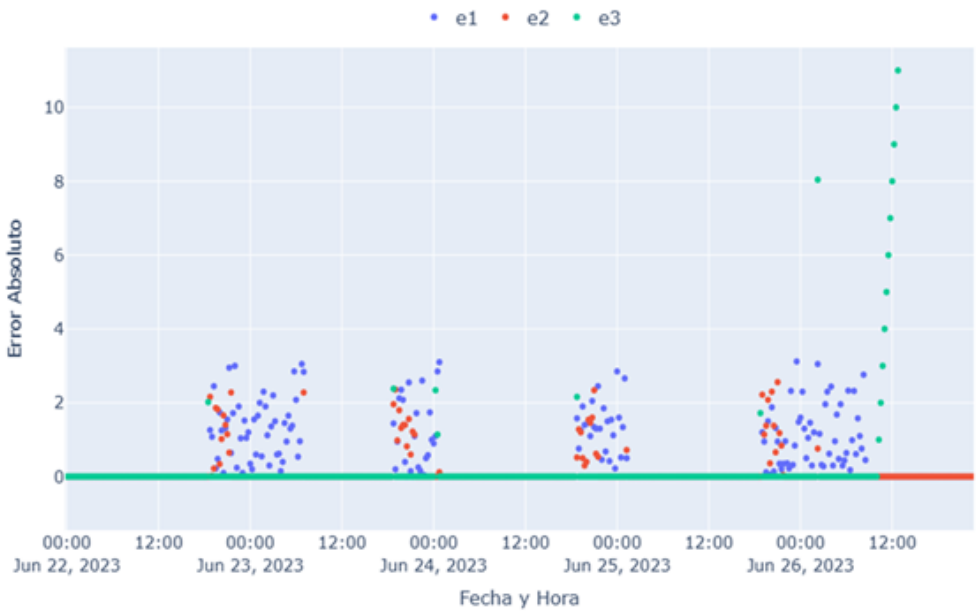


Figura 6.3: Distribución de los errores entre potencia eléctrica medida y calculada para la limpieza preliminar de la base de datos.

6.2 Contexto geográfico y heterogeneidad climática

La Figura 6.4 presenta el diseño experimental de la campaña itinerante de CICITEM, donde se muestra la localización geográfica y altitud de los cinco sitios de medición en la Región de Antofagasta y códigos de color por estación (verde: otoño, azul: invierno, naranja: primavera). El área de estudio se extiende desde la Cordillera de la Costa hasta zonas de la Precordillera y Depresión Intermedia con altitudes entre los 1400 y 3000 m s.n.m.

La Figura 6.5 presenta los *boxplots* por sitio de las variables ambientales (T_{amb}, p_{amb}, HR), mientras que la Tabla 6.2 detalla los cuartiles de estas variables. Se aprecia que Calama y SanPedro presentan climas áridos con variaciones de $\sim 10\,^{\circ}C$ y bajas presiones, coherente con las condiciones desérticas de altitud del altiplano; el sitio etiquetado como Tocopilla presenta temperatura y humedad moderadas con mayor variabilidad, condiciones típicas en la costa-cordillera; PSDA se caracteriza por la presencia de temperaturas máximas extremas ($\sim 40\,^{\circ}C$) y humedades variables (14-40 %); y Chacabuco se diferencia por un clima frío-húmedo con humedades relativas del orden de 70%, aproximadamente.

Esta marcada heterogeneidad climato-topográfica de la zona de estudio anticipa una influencia directa sobre el desempeño electroquímico del banco de PEMFC de cátodo abierto, las cuales son alimentadas con aire extraído desde el ambiente (Saleh et al., 2018). Estos hallazgos motivaron a la incorporación explícita de la variable Sitio como descriptor categórico en el modelado.

Figura 6.4: Diseño experimental de la campaña itinerante de CICITEM en la Región de Antofagasta (generado mediante Google Earth).

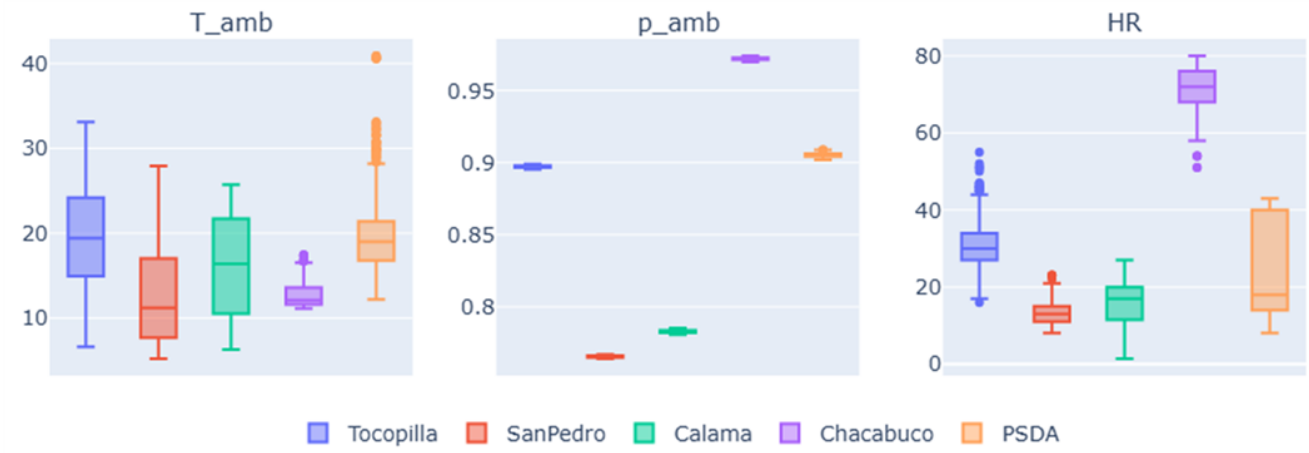


Figura 6.5: *Boxplots* de temperatura ambiente, presión atmosférica y humedad relativa por sitio.

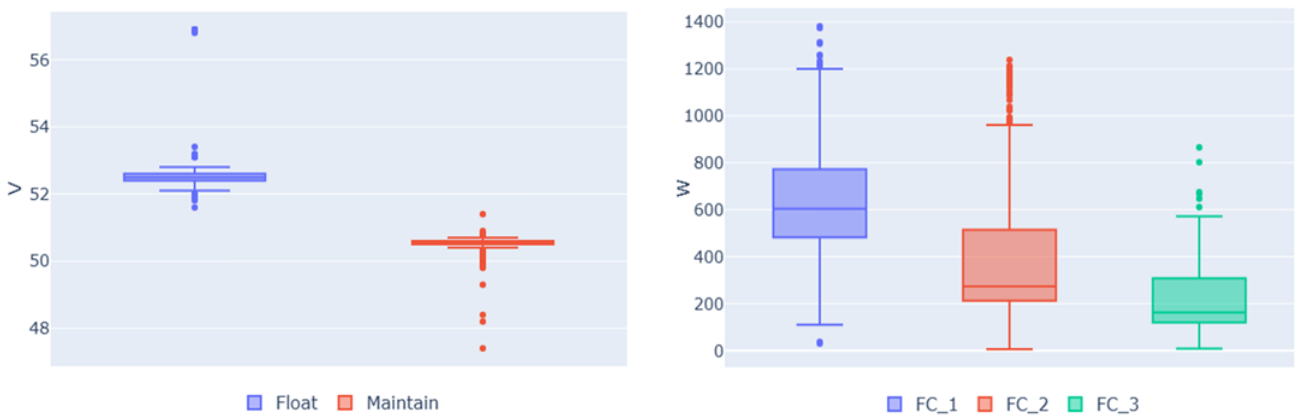
Tabla 6.2: Rangos típicos de condiciones ambientales de los sitios de medición.

Variable	T_{amb} (°C)			p_{amb} (bar)			HR (%)		
Sitio	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3
Calama	10.5	16.4	21.7	0.7820	0.7832	0.7836	11.6	17.0	20.0
Chacabuco	11.6	12.1	13.6	0.9714	0.9719	0.9726	68.0	72.0	76.0
PSDA	16.8	19.0	21.4	0.9043	0.9054	0.9061	14.0	18.0	40.0
SanPedro	7.7	11.2	17.0	0.7648	0.7655	0.7662	11.0	13.0	15.0
Tocopilla	15.0	19.4	24.2	0.8965	0.8973	0.8979	27.0	30.0	34.0

6.3 Modos de operación y curvas de polarización

De acuerdo con el manual de la PEMFC modelo GenSure E-1100, el sistema puede configurarse bajo dos modos operativos para la condición *Low Voltage Start* (Plug Power Inc., 2018). La lógica general de los modos es como sigue:

- Maintain: mantiene el bus DC en torno al umbral de baja tensión (*Low Voltage Threshold*) para entregar potencia a los sistemas auxiliares, minimizando la corriente de salida a las baterías.
- Float: recarga el banco de baterías de la planta además de proveer potencia de salida, ajusta la tensión hasta el valor configurado (*Float Voltage*) y la mantiene durante un tiempo predefinido por el usuario.



(a) (b)

Figura 6.6: *Boxplots* de variables eléctricas: (a) tensión eléctrica por Modo (Float/Maintain); (b) potencia eléctrica por FC_ID.

En la base de datos, esta configuración se registró mediante la variable categórica Modo, lo cual permitió describir estos regímenes en etapas posteriores de modelado. La Figura 6.6(a) muestra

los *boxplots* de tensión eléctrica por Modo, mostrando dos bandas concentradas y algunos valores extremos que serán tratados en la etapa de depuración. Por otra parte, la Figura 6.6(b) muestra los *boxplots* de W según identificador de pila de combustible (FC_ID), la cual refleja un orden descendiente de la potencia media (FC_1 > FC_2 > FC_3), las posiciones relativas de las medianas de W, así como las diferencias en los rangos intercuartílicos, asimetrías y dispersión entre las unidades.

Para explorar a mayor profundidad la interacción entre Sitio y Modo, en la Figura 6.7 se despliegan los *violinplots* de corriente eléctrica bajo esta estratificación. En estos diagramas se aprecian las diferencias entre distribuciones de los grupos y modos operativos por sitio, lo cual refuerza la alta heterogeneidad de la BD.

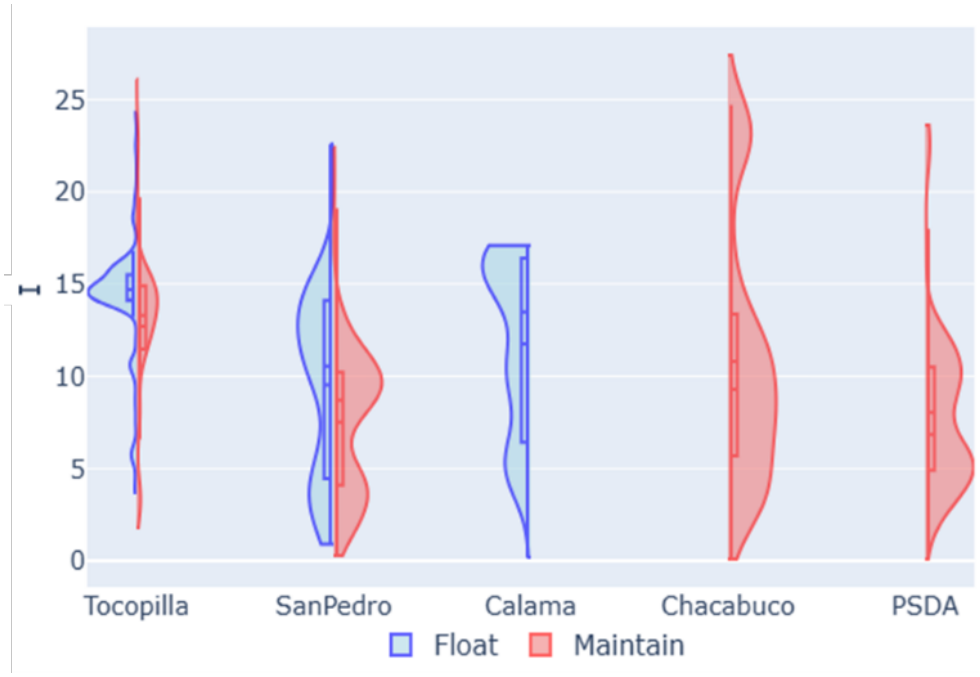


Figura 6.7: *Violinplots* de la corriente eléctrica por Sitio y Modo.

La Figura 6.8 resume las curvas I-V junto con *boxplots* marginales de tensión y corriente eléctricas de los sitios de medición. En esta representación se identifican dos bandas de tensión (50.5 V/52.5 V) y rangos de corriente de operación entre 0.1 – 27.4 A (Tabla 6.1). En la parte superior, los *boxplots* de corriente muestran las diferencias en las medianas y rangos intercuartílicos entre puntos de medición; y en el sector derecho, se aprecian la coexistencia de modos en Tocopilla y SanPedro y la dominancia para los sitios restantes. Por último, en la curva I-V se aprecian puntos aislados con tensiones muy bajas o altas para un mismo nivel de corriente que podrían ser considerados registros anómalos, lo cual se abordará dentro del esquema de detección probabilística en el Capítulo 8.

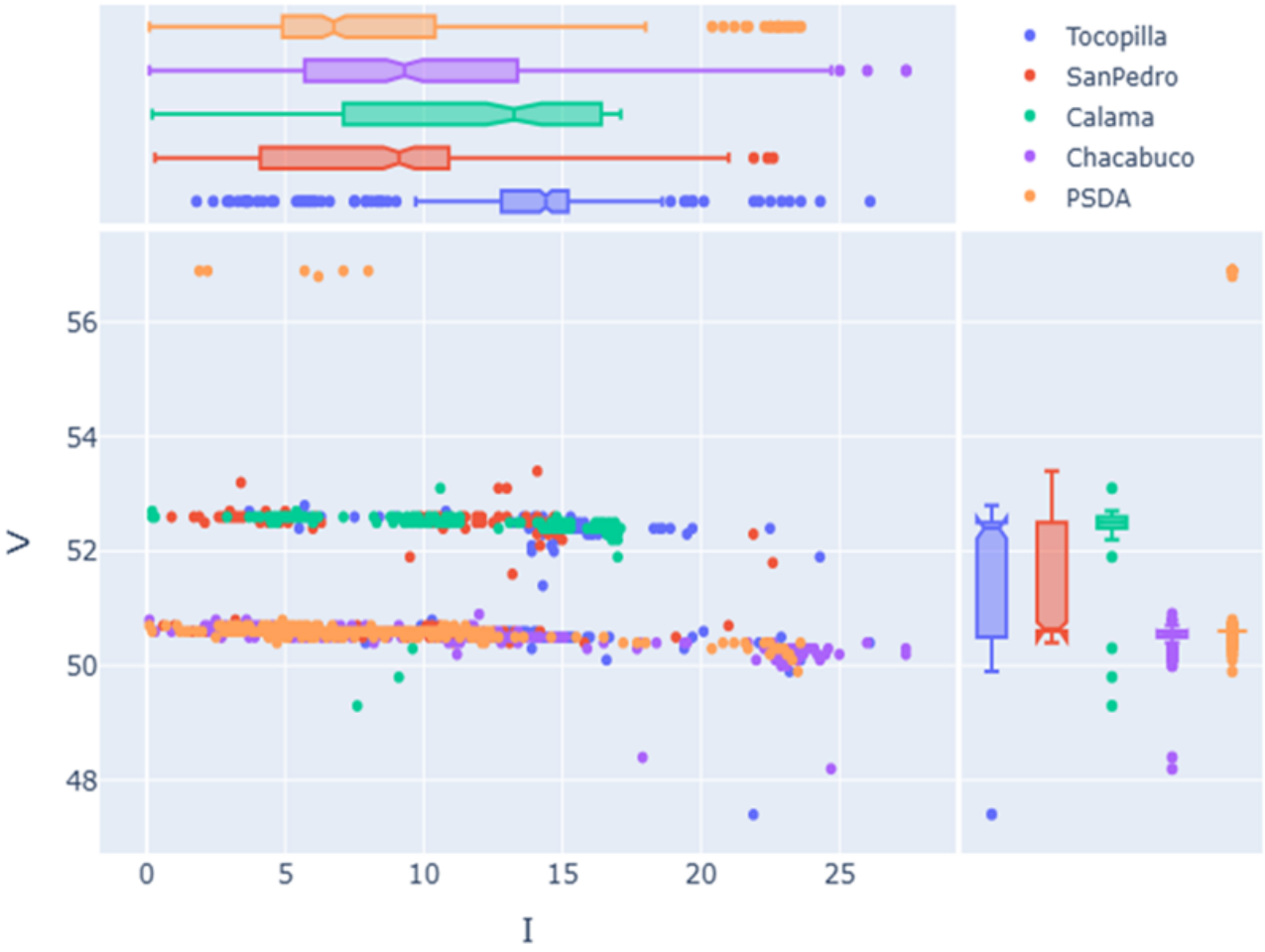


Figura 6.8: Curvas I–V de los sitios de medición y *boxplots* marginales de corriente y tensión.

6.4 Estructura y desbalance de grupos en la base de datos

La Tabla 6.3 resume la cantidad de registros por Sitio y Modo (filas) y FC_ID (columnas) de la base de datos original, compuesta por 1595 datos experimentales. El análisis por FC_ID indica desbalance claro: FC_1 concentra el 59.2 % de los registros, FC_2 aporta el 31.0 % y FC_3 sólo el 9.7 %. Más aún, existe desbalance por sitio: PSDA reúne el 25.2 % de los datos, SanPedro 22.1 %, Chacabuco 20.7 %, Tocopilla 18.5 % y Calama tan sólo un 13,5 %. Respecto a los recuentos por Modo, Maintain aporta un 68.7 % y Float, un 31.3 %, respectivamente.

Este hecho tiene implicancias directas tanto para el modelado como la depuración, puesto que hay grupos que, al contar con más datos, adquieren mayor representación, lo cual requiere utilizar un esquema de división de conjuntos de datos acorde a esta estructura, de modo que el aprendizaje de los modelos no sea sesgado hacia los grupos dominantes. Por este motivo, se debe adoptar un

esquema de estratificación compatible con el objetivo de mejorar el desempeño de los algoritmos y modelos predictivos utilizados.

Tabla 6.3: Distribución de los registros por Sitio, Modo y FC_ID.

Sitio	Modo	FC_1	FC_2	FC_3	Total (SitioxModo)
Calama	Float	156	49	6	211
	Maintain	0	3	0	3
Chacabuco	Maintain	201	92	38	331
PSDA	Float	3	1	2	6
	Maintain	159	194	43	396
SanPedro	Float	57	27	28	112
	Maintain	111	93	37	241
Tocopilla	Float	151	19	0	170
	Maintain	108	17	0	125
Total (FC_ID)		946	495	154	1595

6.5 Análisis bivariado entre variables ambientales y potencia eléctrica

La Figura 6.9 presenta un diagrama de pares de T_{amb}, p_{amb}, HR y W coloreado por Sitio. En la diagonal se observan las distribuciones de las variables con presencia de multimodalidad. La presión forma bandas discretas reflejando las diferencias de altitud entre sitios de medición, mientras que la temperatura y humedad relativa muestran comportamientos diferenciados por sitio, además de cierto grado de similitud climática entre SanPedro y Calama; y Tocopilla con PSDA.

En los paneles de dispersión entre W y las variables ambientales, la potencia eléctrica se distribuye en dominios parcialmente diferenciados por sitio. La correlación lineal es débil entre los pares ($|r| < 0.20$, calculado a partir de la mediana entre los sitios). Como hallazgo clave, se evidencia que bajo ciertas condiciones (T_{amb}, p_{amb}, HR) el sistema produce distintos valores de potencia, lo cual sugiere que el desempeño del sistema responde a efectos de la configuración operativa del sistema (por ejemplo, Modo u otros descriptores que se derivan más adelante) y de las condiciones climáticas locales que delimitan los rangos operativos, conforme a la fenomenología propia de la PEMFC descrita en la literatura. Este comportamiento no puramente determinístico refuerza la necesidad de utilizar modelos de Aprendizaje Automático, así como también, de análisis rigurosos para tareas como la detección de *outliers* presentes en la base de datos.

En síntesis, el EDA muestra que se dispone de una base de datos acotada en tamaño, pero altamente heterogénea y multimodal, con desbalances marcados entre los estratos definidos por Sitio, Modo y FC_ID. Este diagnóstico permitió adoptar medidas específicas para mejorar el

desempeño predictivo de los modelos de Aprendizaje Automático y justificó trabajar con una estructura explícita de grupos o estratos, tanto en la detección de *outliers* mediante la metodología FASEK5 como en los esquemas de validación del modelado. Bajo estas consideraciones, en el capítulo siguiente se presentan los ensayos preliminares de modelado en MATLAB, en los cuales se incorporan los predictores categóricos y descriptores de contexto definidos en este capítulo.

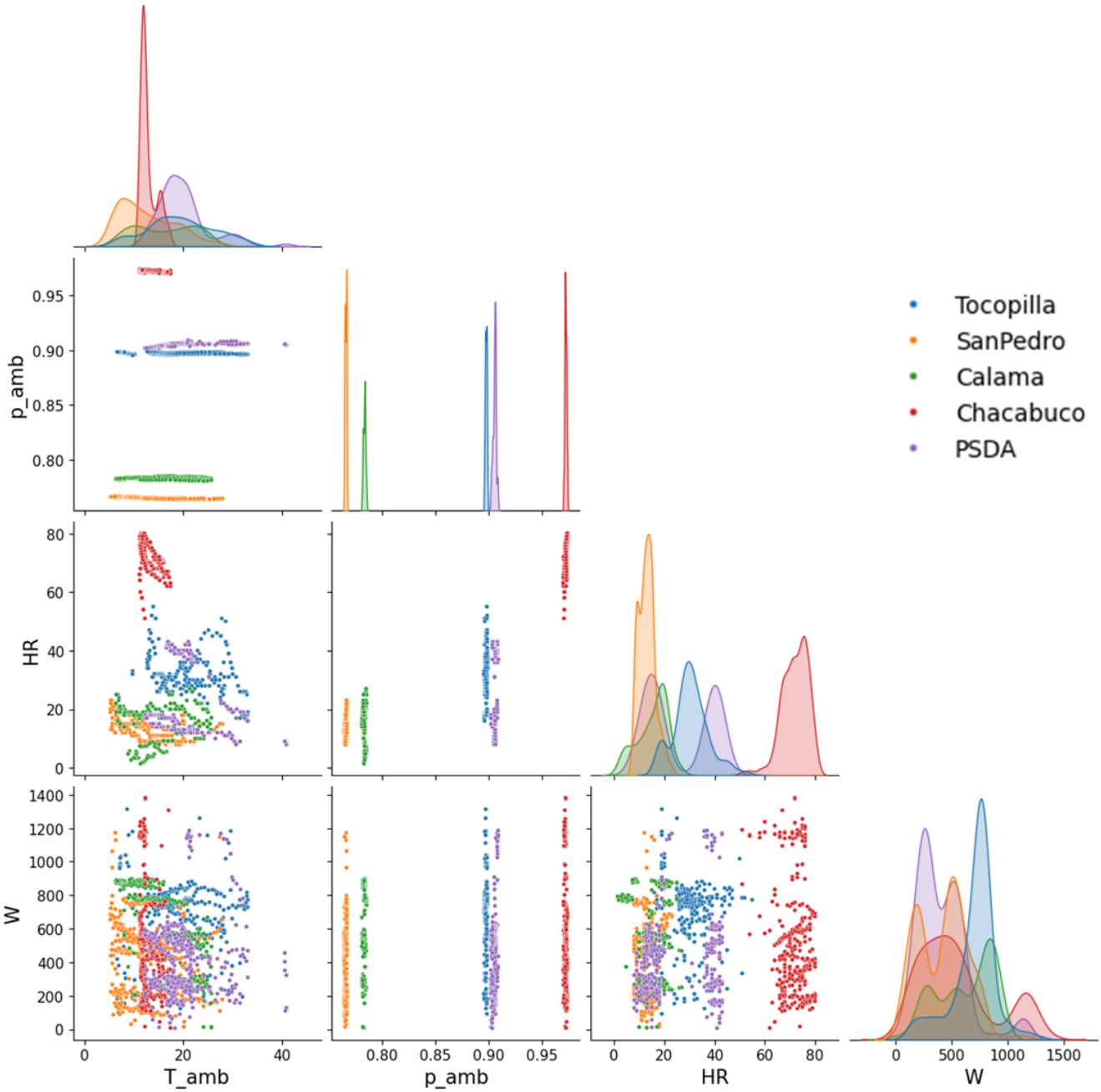


Figura 6.9: Diagrama de pares de correlación entre variables ambientales y potencia eléctrica.

Capítulo 7

Análisis Exploratorio de Datos y Modelado en MATLAB

7.1 Ensayos preliminares en MATLAB: *Neural Net Fitting*

Con el objetivo de validar que un modelo como las redes neuronales artificiales —en adelante, ANN— son compatibles para predecir con cierto grado de precisión el desempeño del banco de pilas de combustible de la planta piloto móvil, se realizaron experimentos preliminares con la aplicación *Neural Net Fitting* disponible en MATLAB. Este módulo permite crear y visualizar la arquitectura de red; y entrenar una red neuronal *feedforward* para resolver problemas de ajustes de datos (Zhou et al., 2022; MathWorks, s. f.). Conforme al flujo indicado en el sitio web oficial, se implementaron las siguientes tareas:

1. Importar la base de datos.
2. Dividir la base de datos en los conjuntos de entrenamiento, validación y prueba.
3. Configurar y entrenar la red.
4. Evaluar el desempeño de la red mediante la métrica error cuadrático medio (MSE).
5. Analizar los resultados mediante diagramas de dispersión e histograma de los residuos del modelo.

Para todas las pruebas se emplearon las proporciones de división 70/10/20 para los conjuntos de entrenamiento, validación y prueba. Se ensayaron distintos tamaños de redes, variando desde 5-10 capas ocultas, buscando un buen balance entre sesgo y varianza, i.e. la diferencia - en adelante, GAP - entre RMSE de prueba y entrenamiento, fijando 7 capas ocultas para todos los ensayos. En

cada una de las secuencias de entrenamiento se reentrenaron las ANN de forma iterativa hasta conseguir un balance satisfactorio. La aplicación recomienda para conjuntos de datos pequeños, ruidosos y de mayor complejidad, implementar el algoritmo de regularización bayesiana, compatible con la estructura de la base de datos. Esta configuración resultó eficaz en términos de desempeño, respecto al resto de algoritmos de optimización disponibles.

Inicialmente, la base de datos (BD) estaba compuesta por el grupo de variables ambientales (T_{amb}, p_{amb} y HR) y eléctricas (I, V, W). A esta BD se le asignó el nombre *Amb*. Tomando en cuenta la relación $W = I \cdot V$, existían cuatro posibles selecciones para la variable objetivo, de las cuales se presentan los valores de RMSE, coeficiente de correlación r y su diferencia (GAP) de la serie de experimentos en la Tabla 7.1. En la Figura 7.1 se presenta la arquitectura de red para el modelo de dos salidas (I, V).

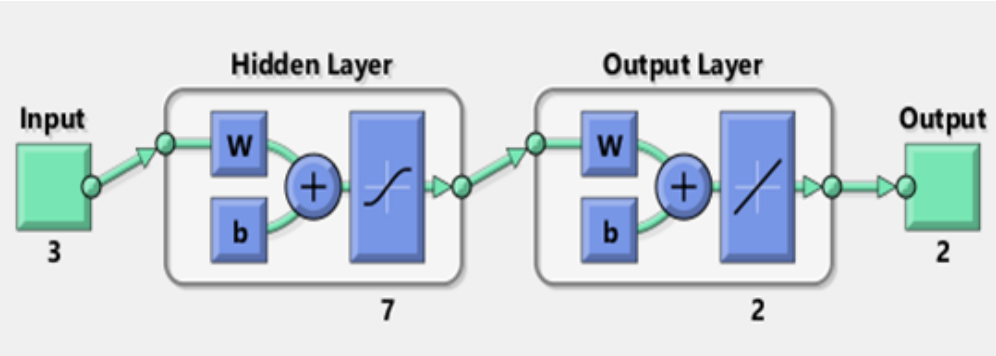


Figura 7.1: Arquitectura de las ANN (7 capas ocultas) para entrenamientos preliminares en aplicación *Neural Net Fitting* de MATLAB.

Referente a los resultados preliminares, un GAP_{RMSE} ($RMSE_{test} - RMSE_{train}$) positivo indicaría un sobreajuste de las predicciones de I y W. A su vez, los modelos de V y de (I, V) subajustaron los datos, indicando inconsistencias en el esquema de modelado, lo cual se originó por la poca dispersión y la presencia de dos bandas para la tensión eléctrica. Desde una perspectiva del Aprendizaje Automático, no es consistente que las predicciones en el conjunto de prueba sean superiores a las de entrenamiento, lo cual sugiere que el esquema de validación no está alineado con este tipo de problema. El análisis de signos de los GAP de los coeficientes r se obtiene por complementariedad. Cabe mencionar que las diferencias se encuentran en las escalas naturales de las variables y, por tanto, no son directamente comparables entre los experimentos.

Tabla 7.1: Resultados de la fase exploratoria en MATLAB – raíz del error cuadrático medio (RMSE), coeficiente de correlación (r) y GAP = Test - Train.

	I	V	I, V	W
RMSE Train	4.264	0.580	3.075	217.6
RMSE Test	4.434	0.552	3.050	222.3
RMSE Gap	0.170	-0.028	-0.025	4.7
r Train	0.607	0.809	0.989	0.613
r Test	0.575	0.812	0.989	0.597
r Gap	-0.032	0.003	0.000	-0.016

Basándose en el coeficiente de correlación, el mejor desempeño aparente correspondería al modelo de dos objetivos. Sin embargo, los diagramas de dispersión revelaron que esta clase de modelo, al tomar en cuenta ambas variables para calcular las métricas, mejora artificialmente el desempeño. En la Figura 7.2 se muestran las gráficas de ajuste entre las predicciones (*Output*) y mediciones (*Target*) para los conjuntos de entrenamiento y prueba, en las cuales se aprecian dos clústeres densos de datos: en el sector inferior izquierdo se ubican los datos de corriente eléctrica, exhibiendo una alta dispersión en las predicciones; mientras que, en el sector superior derecho, los datos de tensión eléctrica se encuentran compactados, lo cual se debe a las diferencias de orden de magnitudes entre las variables. En base a este análisis, se optó por descartar este enfoque de modelado para las etapas posteriores de la serie de experimentos.

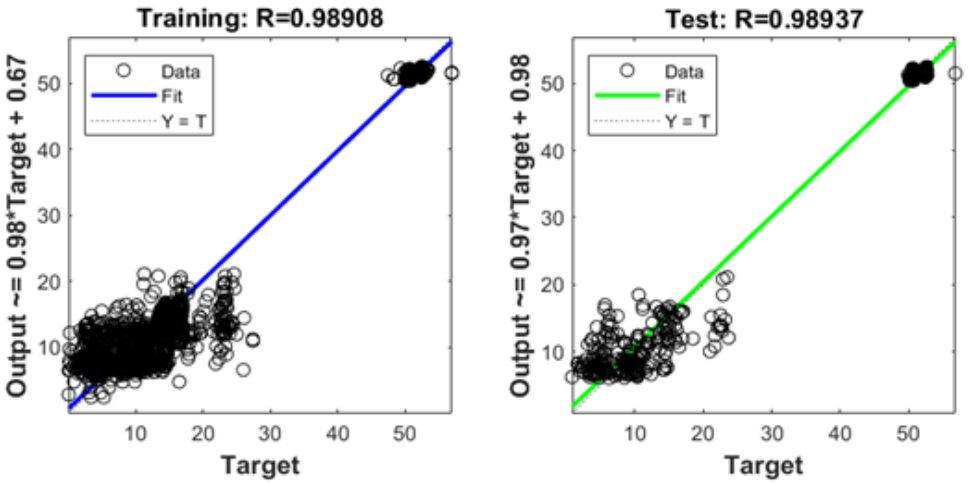


Figura 7.2: Diagramas de dispersión para el modelo de dos salidas.

De las variables candidatas a modelar (I y W), se optó por la potencia eléctrica, debido a que presentó un menor GAP para el coeficiente de correlación. Aunque este criterio no es completo para justificar esta decisión, en etapas posteriores se emplearon características derivadas de la corriente eléctrica, lo cual origina lo que se denomina fuga de datos, en el cual estos nuevos descriptores se

autocorrelacionan fuertemente con el objetivo, puesto que aportan señal de éste mismo para la tarea de regresión (JM et al., 2018).

La Figura 7.3 muestra el diagrama de dispersión de entrenamiento y prueba (o rectas de validación) de W utilizando el grupo de variables ambientales como descriptores. El desempeño de las ANN es notoriamente deficiente, lo cual se refleja en la marcada desalineación de las líneas de ajuste (Fit) de entrenamiento y prueba con la diagonal 1:1 (recta Y = T), con una marcada dispersión de las predicciones del modelo. Un resultado aceptable busca que, tanto la recta de regresión como los datos, se alineen progresivamente con esta diagonal y se reduzca la dispersión para todo el rango de valores de la variable objetivo (W). Esto se traduce en un modelo que acierta eficazmente con un bajo nivel de variabilidad de los residuos y generaliza a nuevos escenarios no vistos durante su etapa de entrenamiento.

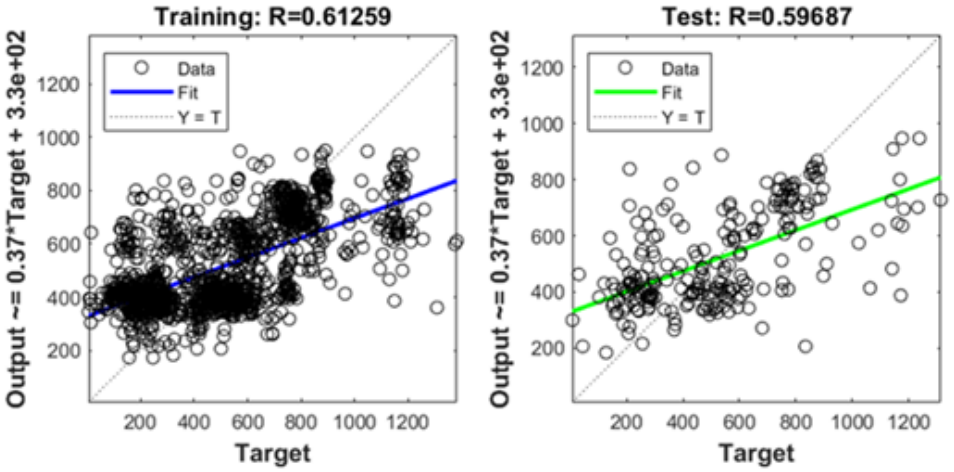


Figura 7.3: Rectas de validación de potencia eléctrica (Amb).

7.2 Implementación de la ablación incremental

Para mejorar el desempeño y la generalización del modelo, se implementó la estrategia de ablación incremental por bloques en conjunto con la ingeniería de características. La primera consiste en agregar sucesivamente grupos de variables y medir su contribución marginal al desempeño, un enfoque reportado en la literatura reciente (Lei et al., 2024; Cavagnero et al., 2022; Zhao et al., 2023). Por su parte, la ingeniería de características corresponde al proceso de extracción y/o transformación de predictores sustentada en el conocimiento de dominio para obtener representaciones latentes con mayor señal útil para la variable objetivo (Kuhn and Johnson, 2019).

La adopción efectiva de estas técnicas se fundamentó en el diagnóstico previo del EDA y en el conocimiento de los fundamentos teóricos que rigen el funcionamiento del sistema PEMFC. De

este modo, se trabajó bajo la hipótesis de que incorporar explícitamente información elemental sobre el contexto operativo y climato-topográfico local facilitaría el aprendizaje del modelo con respecto a sólo operar con datos crudos (Amb), reduciendo así la necesidad de inferir estos patrones y simplificando su proceso de aprendizaje.

Esta tarea se desplegó en cinco bloques agregación, los cuales se describen en las siguientes subsecciones.

7.2.1 Descriptores nativos de la base de datos

A partir del EDA, se verificó que existían variaciones operacionales en cada uno de los sitios de medición de la campaña. Asimismo, los *boxplots* de I indicaron diferencias entre las pilas de combustible del banco; mientras que los *boxplots* de V evidenciaban la presencia de dos bandas o modos de operación. Por consiguiente, se creó el primer grupo compuesto por los descriptores: Sitio, Modo y FC_ID.

Para su incorporación al esquema de modelado, se aplicó la codificación *One-Hot*: técnica en la cual se representan variables categóricas a un formato binario creando columnas para cada clase, donde el valor 1 indica que la categoría está activa; y 0 para el resto de los elementos (Qiu et al., 2022). Cabe mencionar que la agregación de las nuevas variables codificadas aumenta la dimensionalidad del conjunto, extendiendo los tiempos de cómputo de los experimentos.

Dado lo anterior, se crearon las siguientes columnas aplicando esta codificación:

- Sitio: {Tocopilla, Calama, SanPedro, Chacabuco, PSDA}
- Modo: {Mantain, Float}
- FC_ID: {FC_1, FC_2, FC_3}

La ablación incremental consideró la medición de la contribución sobre el RMSE por efecto independiente, por pares y por grupo compuesto. Los resultados se muestran en la Tabla 7.2 y su representación mediante gráfico de barras en la Figura 7.4. La contribución fue calculada según $\Delta = \text{Conjunto} - \text{Amb}$ utilizando las métricas globales. Se incorporaron los resultados del coeficiente de correlación r para indicar la ganancia relativa en el grado de alineación en las rectas de validación para cada conjunto ensayado.

Tabla 7.2: Trayectoria de ablación incremental de métricas RMSE y r para los descriptores nativos.

Conjunto	RMSE	Contr. RMSE	r	Contr. r
Amb	218.4	-	0.610	-
Sitio	213.6	-4.8	0.632	0.022
Modo	213.3	-5.0	0.638	0.028
FC_ID	195.0	-23.4	0.714	0.104
Sitio+Modo	208.3	-10.1	0.655	0.045
Sitio+FC_ID	171.8	-46.5	0.785	0.175
Modo+FC_ID	177.0	-41.4	0.761	0.151
Sitio+Modo+FC_ID	163.6	-54.8	0.805	0.195

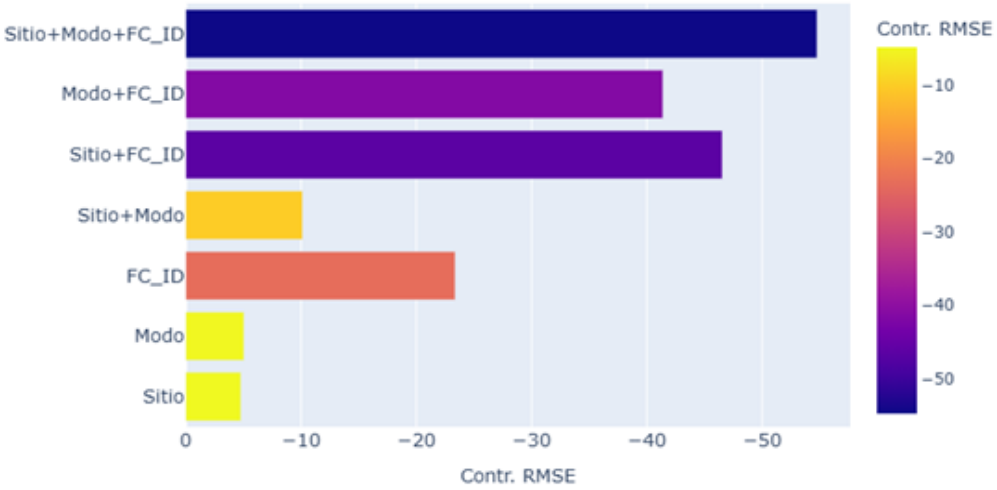


Figura 7.4: Contribución a RMSE de los descriptores nativos (Sitio, Modo, FC_ID).

Estos resultados muestran la trayectoria de ganancia marginal, tanto de RMSE como de r. Respecto a la agregación individual, el descriptor FC_ID superó ampliamente con una reducción de -23.4 W y 0.104 para RMSE y r, respectivamente. A su vez, el conjunto Sitio+FC_ID obtuvo el mayor grado de contribución para los grupos pareados, indicando sinergia entre estas variables. Si bien el descriptor Modo aparentemente produjo mejoras marginales, la contribución compuesta evidenció que el desempeño del modelo mejoró considerablemente, consiguiendo una ganancia para el RMSE de -54.8 W y de 0.195 para el coeficiente r.

En síntesis, la variable FC_ID es la que más utiliza la ANN para predecir la potencia eléctrica; mientras que Sitio aportó información secundaria y Modo mucho menor señal. El modelo se benefició al explicarle directamente la identidad de la PEMFC activa; al informar el sitio le ayudó a diferenciar combinaciones de temperatura, presión y humedades similares que producían distintos valores de W. Adicionalmente, el modo de operación informó la lógica de operación desplegada en cada localización.

En la Figura 7.5 se presentan las rectas de validación de entrenamiento y prueba para el grupo compuesto Sitio+Modo+FC_ID, en la cual se aprecia el progreso conseguido, traducido como un incremento del coeficiente de correlación r y un mayor grado de alineación respecto a la recta Y = T.

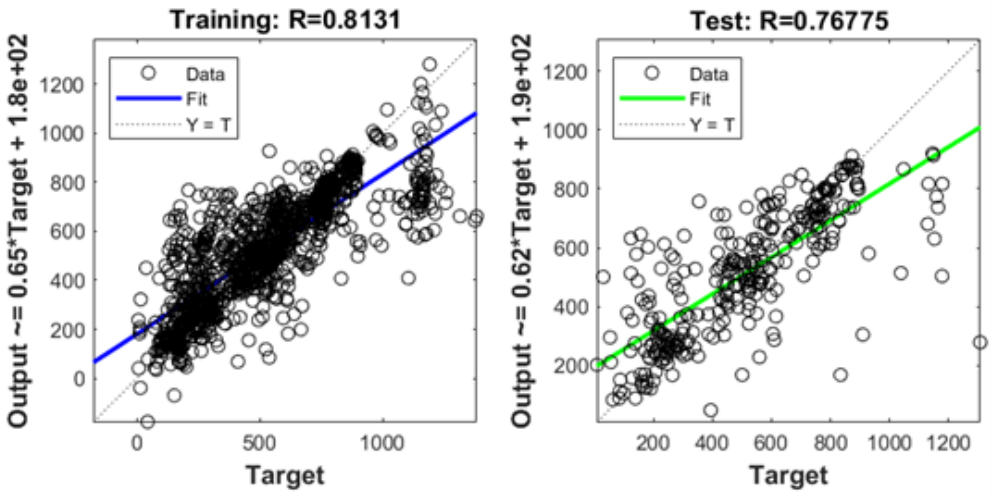


Figura 7.5: Rectas de validación de potencia eléctrica – Ablación incremental (Sitio+Modo+FC_ID).

7.2.2 Configuraciones operativas del banco PEMFC

Para incrementar el rendimiento de las ANN obtenido en la fase anterior, se procedió a crear un descriptor de configuración de las pilas de combustible, que informaba al modelo cuáles y en qué orden estaban encendidas. Como tal, añade señal complementaria a FC_ID, al describir de forma explícita la arquitectura de operación. A esta nueva característica se le denominó Config. Su selección surgió durante la etapa exploratoria de los datos, donde se verificó la presencia de regímenes específicos por sesión operativa en cada sitio de medición.

La creación de este descriptor requirió añadir columnas binarias para las PEMFC (*is_active_i*), que marcaban con un 1 si se encontraba encendida la FC_i. Luego, se concatenaron estos indicadores, originando 7 configuraciones admisibles cuando la planta se encontraba operativa. Así, este descriptor categórico estaba compuesto de los siguientes elementos: Config = {'100', '010', '001', '110', '101', '011', '111'}. A modo de ejemplo, la configuración '101' correspondería a la primera y tercera FC encendidas, y la segunda apagada.

Se procedió a agregar a la BD aplicando previamente la codificación *One-Hot* para el manejo de categóricas y se reentrenó la ANN. A continuación, se reportan las rectas de validación de los conjuntos de entrenamiento y prueba, verificándose el incremento en el desempeño de la red alimentada con el grupo ampliado de descriptores, tal como se aprecia en la Figura 7.6.

En lo sucesivo, no se realizarían experimentos ensayando conjuntos de 1 hasta N características,

lo cual supone reentrenar la ANN para un total de $2^N - 1$ combinaciones, lo cual rompe el esquema del enfoque utilizado y extiende desproporcionadamente el número de experimentos. Por lo que, solamente se evaluará el desempeño del modelo para cada actualización del grupo de características.

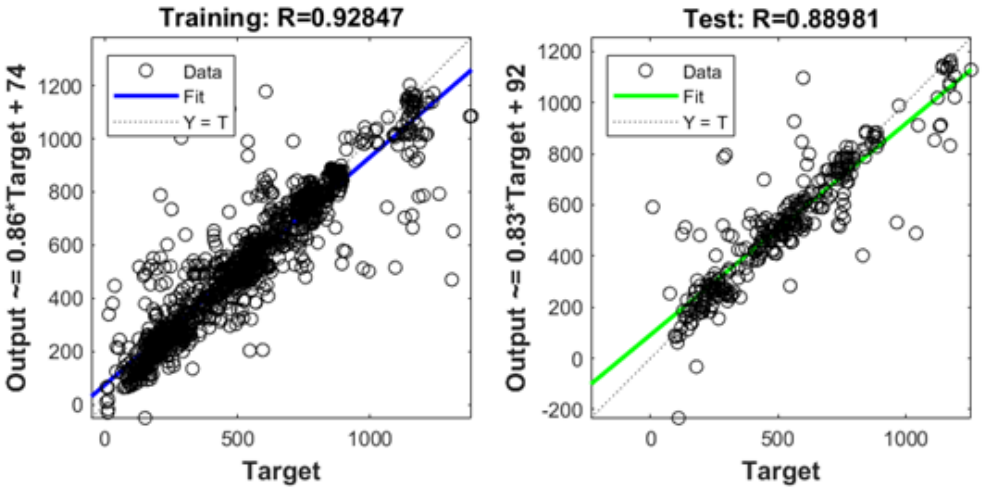


Figura 7.6: Rectas de validación de potencia eléctrica – Ablación incremental de 'Config'.

7.2.3 Contexto climato-topográfico: *k-means clustering*

La alta heterogeneidad topográfica de la Región de Antofagasta se correlaciona fuertemente con la distribución climática de la zona. De ahí que, los registros de (T_{amb}, p_{amb}, HR) tengan cierto grado de similitud entre los sitios de medición emplazados en una unidad de relieve común y, por consiguiente, éstos reciben la misma clasificación climática de Köppen-Geiger, aspectos explorados en un trabajo anterior sobre modelado de PEMFC (Henríquez, 2025).

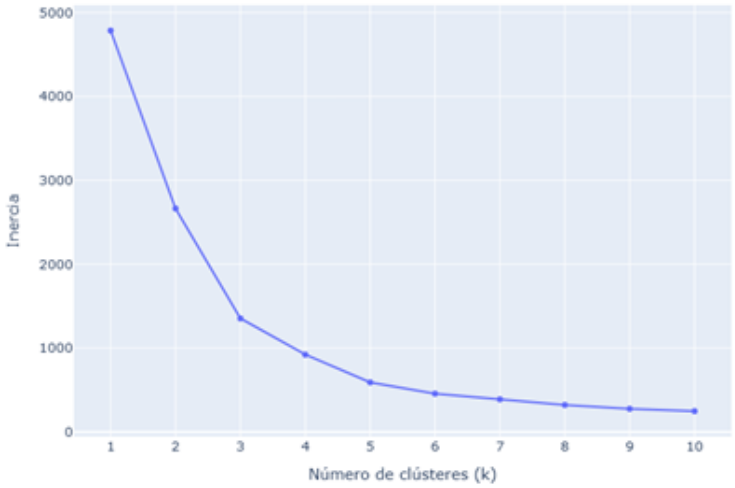
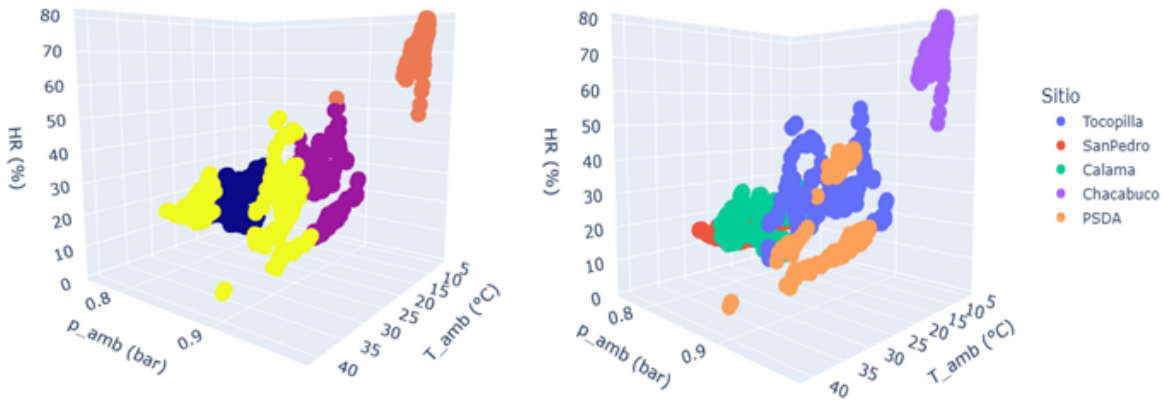


Figura 7.7: *Elbow method (k-means clustering* aplicado a variables ambientales).

En base a estos antecedentes, se propuso clusterizar las variables ambientales para resumir el régimen climático local. Con esto, se conseguiría fragmentar aún más la información, explicándole al modelo la clasificación climática latente presente en los datos.

Se aplicó *k-means clustering* fijando K = 4 clústeres a partir del *elbow method*, como se muestra en la Figura 7.7. Se presenta la visualización 3D de la clasificación obtenida a partir del algoritmo en conjunto con la representación de los datos coloreados por sitio para supervisar la asignación de los regímenes, los cuales se muestran en la Figura 7.8. A partir de las proyecciones 2D (Figura 7.9), se obtuvo la siguiente lectura:

- Clúster 0 (azul oscuro, p ≈ 0.76–0.78 bar, T ≈ 5–21 °C, HR ≈ 5–25%): agrupación de Calama y San Pedro con clima árido frío de altitud.
- Clúster 1 (morado, p ≈ 0.90–0.91 bar, T ≈ 13–22 °C, HR ≈ 15–45%): mezcla de PSDA y Tocopilla con condiciones mayormente templadas.
- Clúster 2 (naranja, p ≈ 0.97 bar, T ≈ 10–17 °C, HR ≈ 60–80%): sólo Chacabuco con clima húmedo de baja altitud por posible influencia de la costa.
- Clúster 3 (amarillo, p ≈ 0.90–0.91 bar, T ≈ 22–40 °C, HR ≈ 10–40%): mezcla de PSDA y Tocopilla con aportes de Calama y San Pedro con clima cálido y humedad de baja a moderada.



(a) (b)

Figura 7.8: (a) Visualización 3D *k-means clustering* (K = 4); (b) distribución espacial de variables ambientales coloreados por Sitio.

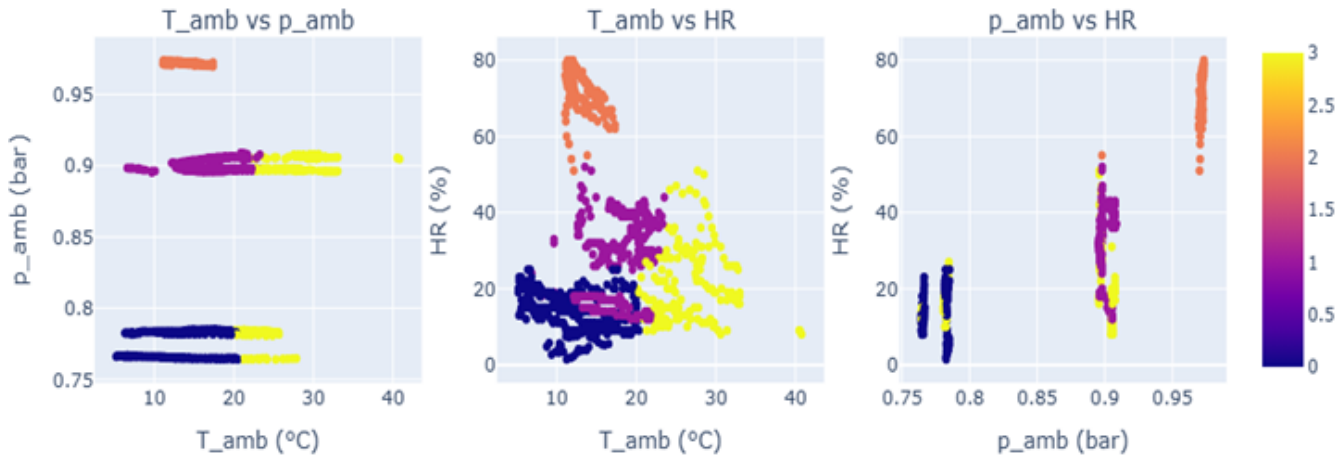


Figura 7.9: Proyección de los clústeres obtenidos con *k-means clustering*.

Se procedió a crear el descriptor categórico Cluster_ID y se empleó codificación *One-Hot*. Así, se obtuvieron los resultados que muestran en las rectas de validación de la Figura 7.10. En este caso, la ganancia marginal del ajuste es modesta, lo cual no implica que esta nueva característica no aporte señal complementaria a las ANN, puesto que el coeficiente de correlación ya se encuentra en un margen considerablemente alto, considerando las limitaciones observadas de nuestra BD.

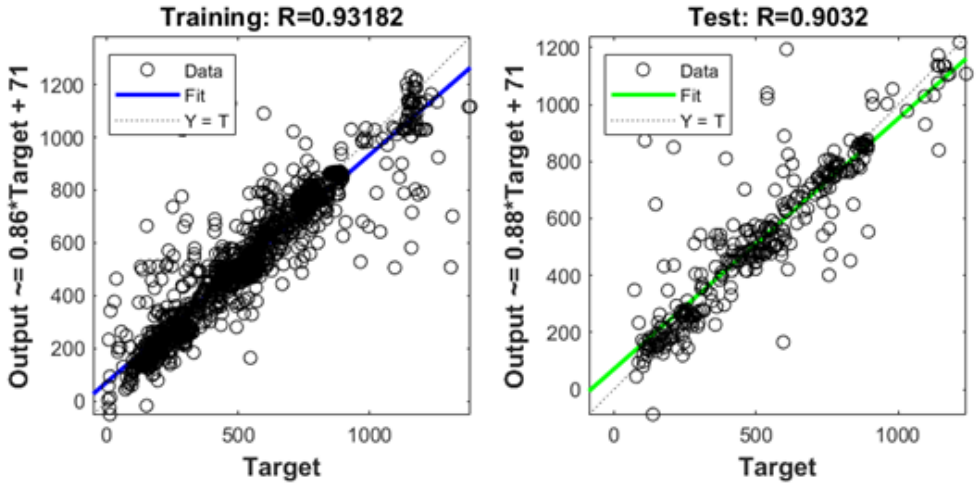


Figura 7.10: Rectas de validación de potencia eléctrica – Ablación incremental de Cluster_ID.

7.2.4 Codificación temporal y estacionalidad

Un enfoque alternativo consistió en derivar características desde los registros temporales de las mediciones. En primer lugar, se extrajo la estación del año en la cual se desarrollaron las campañas de medición en cada uno de los sitios, cuyo predictor se le asignó el nombre Season, compuesto por

los elementos: 'Invierno', 'Primavera' y 'Otono' (en lugar de otoño para el manejo semántico en Python). Como complemento, para informar el día en qué se efectuaron las mediciones al modelo, se empleó la codificación estándar para variables cíclicas en el manejo de series de tiempo, conforme a las siguientes relaciones (Hyndman and Athanasopoulos, 2018):

$$Day\_cos = \cos\left(\frac{2\pi d}{m}\right) \tag{7.1}$$

$$Day\_sin = \sin\left(\frac{2\pi d}{m}\right) \tag{7.2}$$

Donde d es el día del año y m = 365 (ciclo completo). Los nuevos predictores Day_cos y Day_sin corresponden a un vector dentro del círculo unitario, que establece un orden interpretable por esta clase de modelos. Al tomar estas medidas se incorporó contexto dinámico-temporal sin alterar la estructura de modelado, manteniendo el problema de regresión como pseudo-estacionario.

En la Figura 7.11 se ha representado esta codificación, diferenciando por color y tipo de marcador, el sitio y la estación del año respectivos. Se aprecia cómo el descriptor Season añade contexto temporal más general y, a su vez, Day_cos y Day_sin capturan transiciones suaves con cierta posición dentro de cada fase estacional.

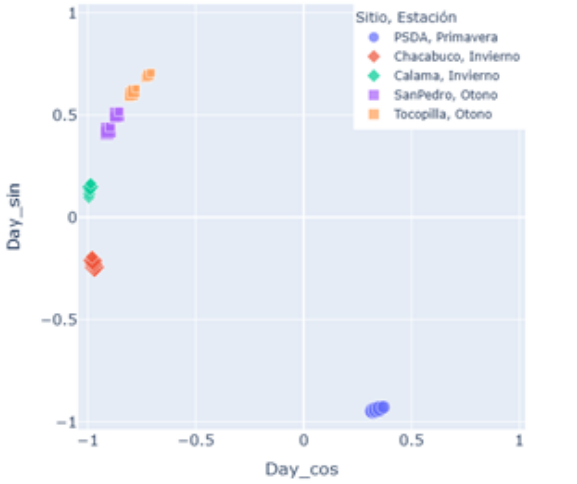


Figura 7.11: Codificación temporal y estación de las campañas de muestreo.

Al implementar la codificación *One-Hot* al predictor Season, se obtuvieron las rectas de validación para entrenamiento y prueba, como se muestran en la Figura 7.12. A pesar de la disminución aparente sobre el coeficiente r, el GAP obtenido es menor que el de los conjuntos anteriores, reflejando el beneficio de añadir estos descriptores temporales.

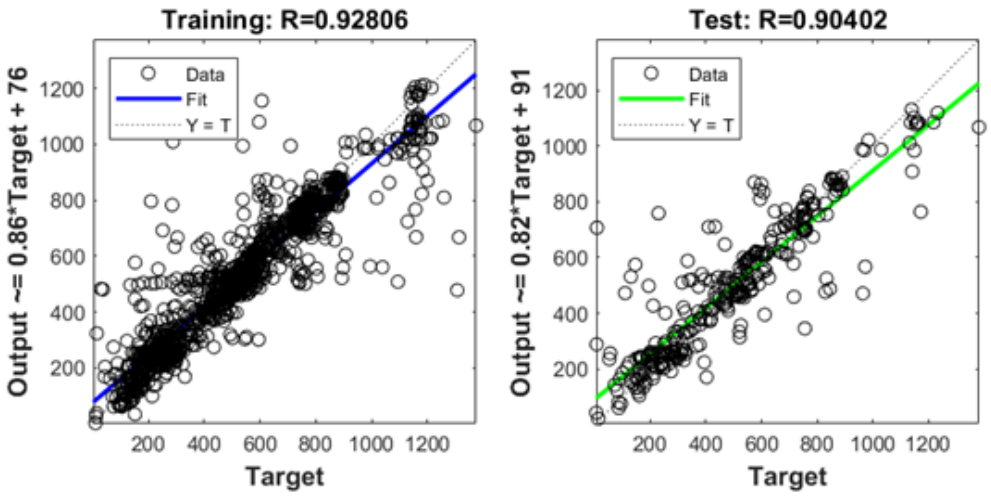


Figura 7.12: Rectas de validación de potencia eléctrica – Ablación incremental con predictores temporales.

7.2.5 Reparto de combustible y codificación mediante subespacios composicionales

Con el propósito de complementar al predictor Config, que tan sólo informaba al modelo la arquitectura de las PEMFC operativas, se crearon tres descriptores adicionales para describir la composición y reparto de combustible para los regímenes con 2 y 3 activas. Al tomar en cuenta que el sistema está montado en paralelo, las composiciones $\mathrm{X_i} = \mathrm{I_i} / \sum \mathrm{I_j}$, de acuerdo con la ley de Faraday, se toman de forma aproximada como la estequiometría de hidrógeno alimentado a cada PEMFC (O'hayre et al., 2016).

En primer lugar, se creó un descriptor de regímenes de activación (mono, bi y tricelda), sumando los elementos de 'Config', al cual le denominamos K_act = {'1', '2', '3'} y se codifican mediante *One-Hot*. Un descriptor adecuado de reparto para clases desbalanceadas es la entropía de Shannon normalizada que condensa la composición en único descriptor (Shannon, 1948), definida mediante la siguiente ecuación:

$$H = -\frac{\sum_i X_i \ln(X_i)}{\ln(D)}, \quad D \in \{2, 3\} \tag{7.3}$$

Mediante ingeniería de características, se amplificó por K_act codificado $(\mathrm{K_1}, \mathrm{K_2}, \mathrm{K_3})$, es decir:

$$H_2 = K_2 \cdot H \tag{7.4}$$

$$H_3 = K_3 \cdot H \tag{7.5}$$

Luego, se añadió un descriptor de reparto H = {'H2', 'H3'} focalizado para estos regímenes. Como señal complementaria, se aplicó la transformación ILR (*isometric log-ratio*) de los X_i para tricelda, generando vectores ortonormales que retienen la información de las tres partes en un subespacio de dos dimensiones (Egozcue et al., 2003), cuyas coordenadas vienen dadas por:

$$Z_1 = \sqrt{\frac{1}{2}} \cdot \ln\left(\frac{X_1}{X_2}\right) \tag{7.6}$$

$$Z_2 = \sqrt{\frac{2}{3}} \cdot \ln\left(\frac{\sqrt{X_1 X_2}}{X_3}\right) \tag{7.7}$$

Se completa el esquema multiplicando las coordenadas Z_1 y Z_2 por el identificador K_3. A este último descriptor, se le asignó el nombre Z = {'Z1_3', 'Z2_3'}.

Esta formulación buscaba aportar señal específica por régimen, lo cual apuntaba a conferir mayor grado de detalle que el modelo reconociera y utilizara para afinar sus predicciones. Los resultados obtenidos se presentan a continuación, entrenados sobre la BD denominada, en lo sucesivo, *Baseline*.

7.3 Resultados de la ablación incremental

Al reentrenar la ANN con el grupo compuesto por todos los descriptores, se obtuvo las rectas de validación de entrenamiento y validación que se muestra en la Figura 7.13, que muestra una mejora consistente en la capacidad predictiva del modelo con un coeficiente de correlación en entrenamiento de 0.9533. El incremento del GAP se justifica debido que se entrenó una cantidad limitada de veces en el entorno de MATLAB, por lo que estos resultados no reflejan el mejor rendimiento.

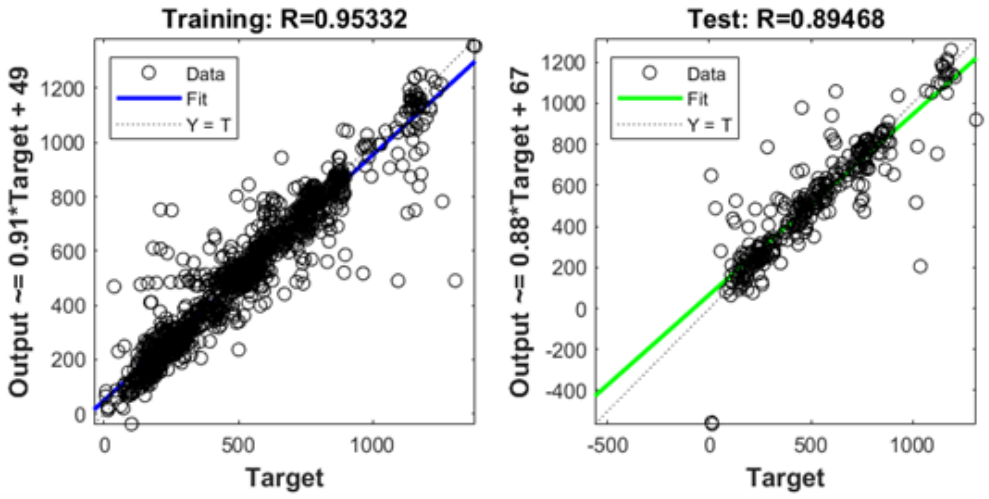
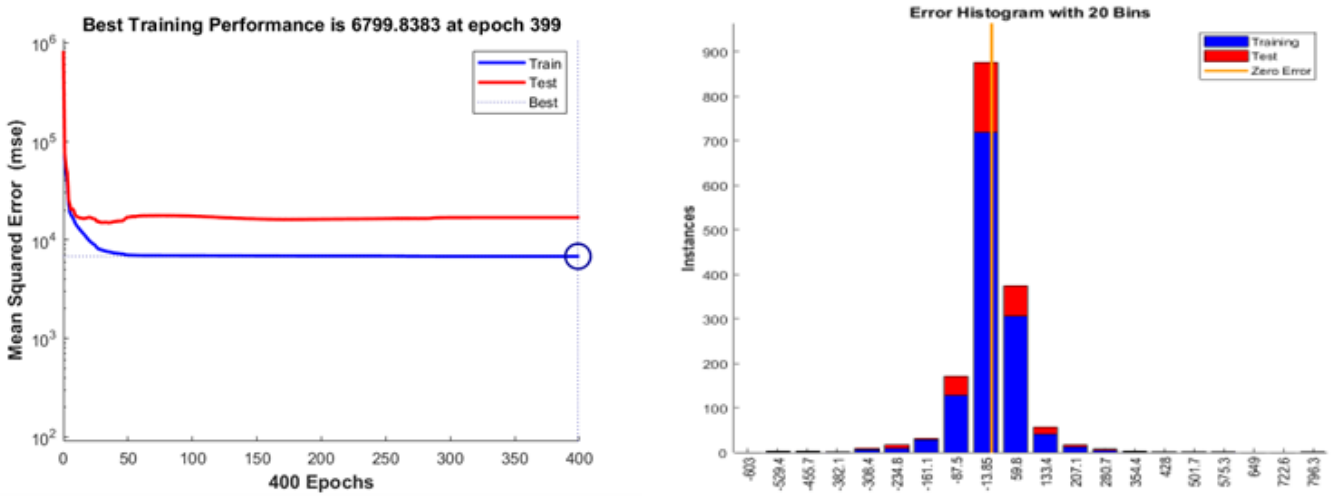


Figura 7.13: Rectas de validación de potencia eléctrica – Ablación incremental con todos los predictores.

Durante la serie de experimentos, se monitoreó la curva de aprendizaje del modelo y el histograma de los residuos, los que se muestran en las Figura 7.14. A partir de la curva de aprendizaje, se supervisó la estabilidad de la convergencia en cada iteración; el histograma (20 *bins* por defecto) indicó que la mayor proporción de datos fueron predichos dentro de un margen de error entre -87.5 y 59.8 W. Las colas extremas del histograma se atribuyeron producto de la BD ruidosa con marcado desbalance entre clases de los predictores propuestos. Por consiguiente, se tiene que el esquema de partición de los datos de MATLAB pudiera no ser la más afín para estos datos, entre otros factores que podrían repercutir sobre el rendimiento de las ANN.



(a) (b)

Figura 7.14: (a) Curva de aprendizaje de la ANN entrenada con todos los descriptores; (b) Histograma de errores (mediciones - predicciones) asociado.

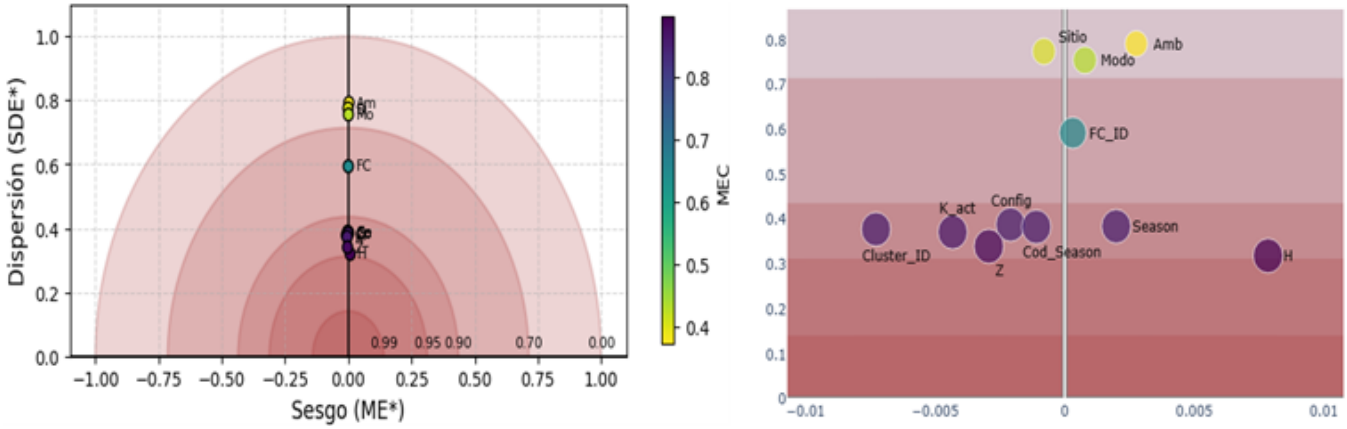
Se realizó una evaluación integral del desempeño de la serie de experimentos utilizando los diagramas solar y de Taylor. Las métricas globales se muestran en la Tabla 7.3 y resumen la secuencia de ablación incremental, en la cual se aprecia la contribución marginal en cada etapa.

La ruta del diagrama solar indicó una reducción consistente de la dispersión de los errores manteniendo una dirección hacia el origen (Figura 7.15(a)) y su detalle en Figura 7.15(b). El signo del sesgo medio del error (ME) indicó la dirección del ajuste y la alineación observada con el eje vertical se explica por la distribución de errores visualizados en el histograma. Los últimos conjuntos exhibieron mayor coeficiente MEC (color morado), indicando un mejor ajuste de los datos. El incremento de los coeficientes CCC se tradujeron en un mayor grado de alineación con la diagonal en los gráficos de dispersión.

Tabla 7.3: Conjunto de métricas del diagrama solar y de Taylor – Resultados de los experimentos de ablación incremental e ingeniería de características en la aplicación *Neural Net Fitting* de MATLAB.

Conjunto	Std	ME	SDE	RMSE	r	R^2	MEC	CCC
Observations	275.6							
Amb	166.5	0.762	218.4	218.4	0.610	0.372	0.372	0.540
Sitio	172.4	-0.220	213.6	213.6	0.632	0.399	0.399	0.568
Modo	176.5	0.215	208.3	208.3	0.655	0.429	0.429	0.595
FC_ID	221.7	0.087	163.6	163.6	0.805	0.648	0.648	0.786
Config	255.0	-0.576	107.3	107.3	0.921	0.848	0.848	0.918
Cluster_ID	258.3	-2.006	104.4	104.4	0.926	0.857	0.856	0.924
Season	251.7	0.548	106.2	106.2	0.923	0.852	0.852	0.919
Cod_Season	253.4	-0.301	105.8	105.8	0.923	0.853	0.853	0.920
K_act	259.7	-1.191	102.7	102.7	0.928	0.861	0.861	0.926
H	264.4	2.163	88.1	88.1	0.948	0.898	0.898	0.947
Z	263.6	-0.807	93.9	93.9	0.940	0.884	0.884	0.939

El diagrama de Taylor de la Figura 7.16 muestra el mayor grado de precisión conseguido, donde los últimos conjuntos se encuentran más cercanos al punto de referencia $(1, 0)$. Estos capturaban mayor proporción de la variabilidad de las observaciones (distancia radial) y un mayor coeficiente de correlación (posición angular). Ambas representaciones sustentan cuánto mejoró el desempeño de las ANN empleando la estrategia de ablación incremental e ingeniería de características.



(a) (b)

Figura 7.15: (a) Diagrama solar con resultados de los experimentos de la ablación incremental en MATLAB; (b) detalle de ubicación de los puntos.

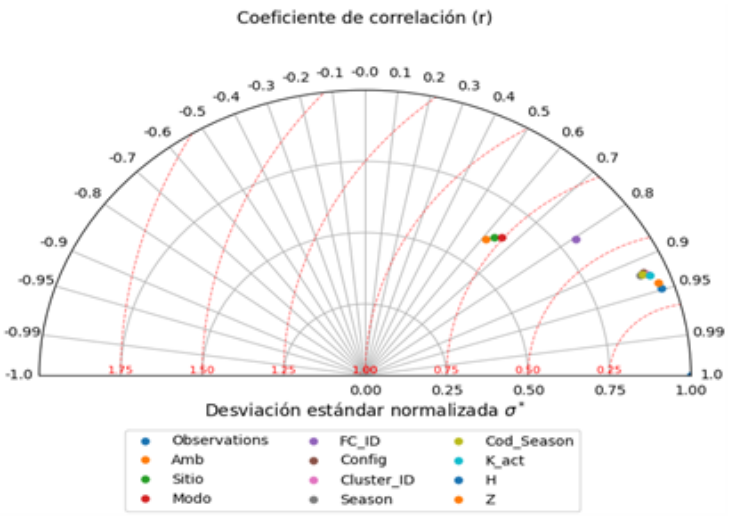


Figura 7.16: Diagrama de Taylor con resultados de los experimentos de ablación incremental en MATLAB.

En términos globales, el RMSE disminuyó de 218.4 W (Amb) a 93.9 W (conjunto Z, *Baseline*), mientras que el coeficiente de correlación r aumentó de 0.61 a 0.94. Estos resultados confirman la efectividad de la ingeniería de características para mejorar la predicción de la potencia eléctrica del banco PEMFC, pero también ponen de manifiesto limitaciones asociadas a la presencia de *outliers* y al desbalance de los estratos, expresadas como una elevada dispersión de los residuos. Este diagnóstico motivó la implementación de esquemas específicos para robustecer las predicciones. El siguiente capítulo presenta la depuración controlada de la base de datos mediante un esquema de detección de *outliers* compatible con la estructura específica de la BD.

Capítulo 8

Metodología de puntaje probabilístico: Detección y depuración de *outliers*

El desempeño de los modelos de Aprendizaje Automático depende fuertemente de la calidad de los datos. La operación del banco PEMFC de la planta piloto móvil se realizó en múltiples sitios, con modos de operación y configuraciones variables. Estas combinaciones generaron una BD limitada en cantidad de datos, pero altamente heterogénea, en la cual coexisten errores de registros, condiciones climáticas extremas e instancias operativas poco frecuentes. Bajo estas condiciones, una depuración somera basada solamente en criterios univariantes podría eliminar información relevante o, en el peor de los casos, admitir datos inconsistentes y no representativos que distorsionarían el ajuste y el desempeño de los modelos predictivos.

El objetivo de este capítulo es describir la metodología desarrollada para la detección y remoción controlada de *outliers* presentes en la BD. Para ello se diseñó una batería de 37 detectores organizados en cinco familias de evidencia (F1–F5), que actuaban sobre distintos subespacios físicos de variables y estadísticos del sistema: tensión, corrientes y repartos de combustible, curva I–V, grupo climático–eléctrico y potencia eléctrica. El aporte conjunto de los detectores se unificó mediante un modelo probabilístico estratificado, que se acuñó como "FASEK5", inspirado en la teoría de los modelos *noisy-OR*, que integraba los métodos por familia y concluía con un estimador de probabilidad global de *outlier*. Este se sometió a un margen de contaminación predefinido del 5% de la BD para evitar remover una cantidad excesiva de datos.

Como producto de su implementación, se buscó reducir la influencia de mediciones inconsistentes de la BD al eliminar datos ruidosos, conservando la estructura de información operativa, lo cual fue crítico para refinar el modelado de la variable objetivo (W), como se discutirá en los capítulos posteriores. El diagrama de la Figura 8.1 sintetiza la secuencia de pasos de esta metodología, donde se crearon 3 versiones de la BD depurada para posteriormente realizar análisis de sensibilidad sobre el desempeño de los modelos en función de la proporción de ruido eliminado de la base de datos.

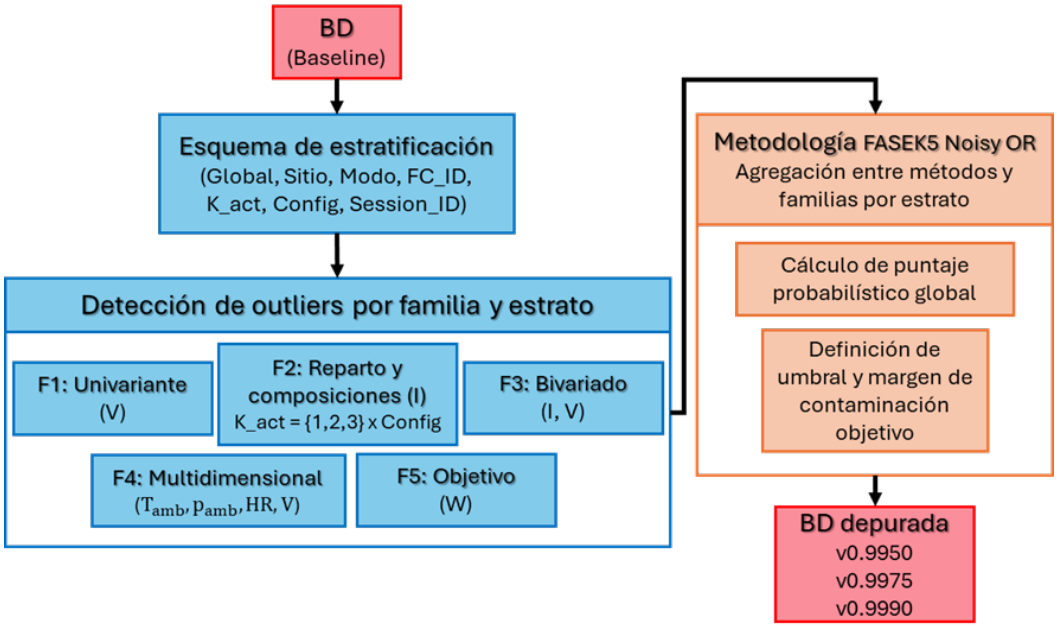


Figura 8.1: Esquema de la estrategia de detección de outliers y consolidación mediante metodología FASEK5, modelo *leaky noisy-OR*.

8.1 Esquema de estratificación para la detección de *outliers*

Como se mencionó anteriormente, se trabajó con una BD altamente heterogénea, lo cual se reflejaba en las distribuciones de las variables ambientales y eléctricas, originada por la mezcla de grupos de características categóricas (Sitio, Modo, Config, entre otras), de acuerdo con el diagnóstico obtenido mediante EDA complementario, al incluir los nuevos descriptores introducidos durante los experimentos de ablación incremental e ingeniería de características.

Para mitigar la influencia de estos múltiples grupos, la metodología de detección de *outliers* se diseñó e implementó en estratos operativos, con el fin de obtener distribuciones menos sesgadas y mantener la comparabilidad interna de las variables. Si bien la selección de los estratos se ajustó según el tipo de variable, grupo de características o familia de detección, el estrato operativo basal (e) corresponde a:

$$\mathrm{e} = \mathrm{Sitio} \times \mathrm{K\_act} \times \mathrm{Config} \tag{8.1}$$

Luego, se creó el identificador de sesión para cada uno de los sitios, en adelante Session_ID, que registraba períodos continuos de operación hasta el apagado del sistema. En particular, esta tarea fue clave para la detección de *outliers* de corriente de F2, puesto que esta variable presentaba distribuciones multimodales y con marcadas asimetrías que distorsionaban la detección mediante *boxplots*, inclusive en estratificaciones más finas. La asignación de Session_ID para cada sitio se

realizó según: Tocopilla (T1–T5), Calama (C1–C4), SanPedro (SP1–SP4), Chacabuco (CH1–CH4) y PSDA (P1–P4). Las sesiones se desarrollaron por las noches, iniciando entre las 19:00 hrs hasta las 08:00 hrs del día siguiente.

Durante la ejecución de esta tarea, se identificaron intermitencias en el sistema, períodos de operación aislados por un margen de 30–60 minutos previos o posteriores a la sesión completa, tanto para acondicionamiento del sistema como purga del hidrógeno excedente, u otros eventos específicos fuera de las condiciones regulares de operación. Todas estas instancias fueron marcadas con una columna adicional de 'Observaciones' para mantener la trazabilidad sobre su ubicación en el espacio de variables, además de ser candidatos a *outliers* a priori.

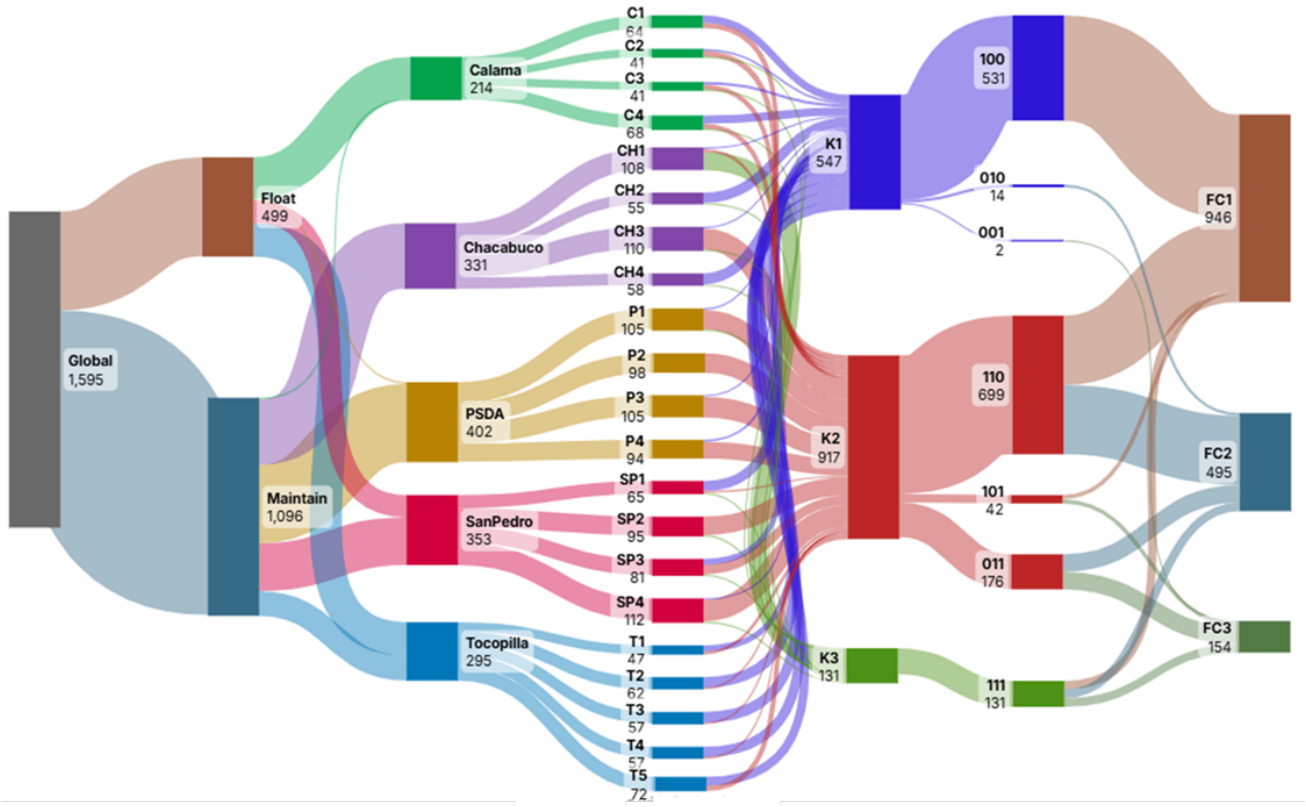


Figura 8.2: Diagrama de Sankey de la base de datos de CICITEM.

La Figura 8.2 despliega la arquitectura de la base de datos mediante un diagrama de Sankey, iniciando desde conjunto global de datos hacia las categorías Modo, Sitio, Session_ID, K_act (K1, K2, K3), Config y FC_ID. Este diagrama muestra la cantidad de datos de las categorías, donde la altura es proporcional al total general (1595 registros) con una marcada asimetría entre éstas. El grosor de los flujos refleja el desbalance en contribución relativa de información hacia el estrato subsecuente y cómo se fragmenta progresivamente esta señal a medida que se estratifica hacia K_act y Config.

Esto evidenció que algunos sitios y configuraciones concentraban la mayor parte de los registros, mientras otros quedaban escasamente representados dentro de la BD. Esta geometría de flujos justificó el diseño de detecciones de *outliers* mediante el esquema de estratificación ('Sitio×Config×K_act'). De este modo, se identificaba la contaminación interna de estratos más uniformes y se preservaba su representación en la BD, es decir, se impedía vaciar ramas completas del diagrama posterior a la depuración, a menos que el grupo de detectores justificaran esta acción.

8.2 Resultados globales de la detección de *outliers*

La estrategia de detección se implementó mediante 37 detectores organizados en cinco familias (F1–F5), cuyas características principales se resumen en la Tabla 8.1. La combinación de estas familias de detección buscó abordar desde errores instrumentales en una variable hasta configuraciones operativas anómalas en el espacio multidimensional.

Tabla 8.1: Características principales de las familias de detección de outliers.

Familia	Variable	Estrato de cálculo	Enfoque	Métodos aplicados	N° de detectores
F1	V	Global, Sitio, Modo	1D	Boxplot/1.5IQR + *offset*	2
	I	Sitio, Session_ID	1D	Boxplot	1
F2 - K1	I	Sitio, Session_ID	1D (FC_1)	Boxplot	2
F2 - K2	$\ln(I_1+I_2+I_3)$, H_{active}	Config: Sitio, FC_ID, Session_ID	1D, 2D	Boxplot (FC y PCA), DBSCAN en ln(S), H_{active}	13
F2 - K3	Z_1, Z_2	Global, Sitio	2D	Boxplot (FC y PCA), DBSCAN en ILR (Z_1, Z_2)	6
F3	(I, V)	Global, Sitio	2D	DBSCAN sobre curva I–V	2
F4	(T, p, HR, V)	Global, Sitio	Multi-D	DBSCAN (PC1, PC2, PC3); Hotelling's T^2 vs. SPE	8
F5	W	Global, Sitio, FC_ID	1D	Boxplot (3 IQR), DBSCAN, percentiles 1–99	3

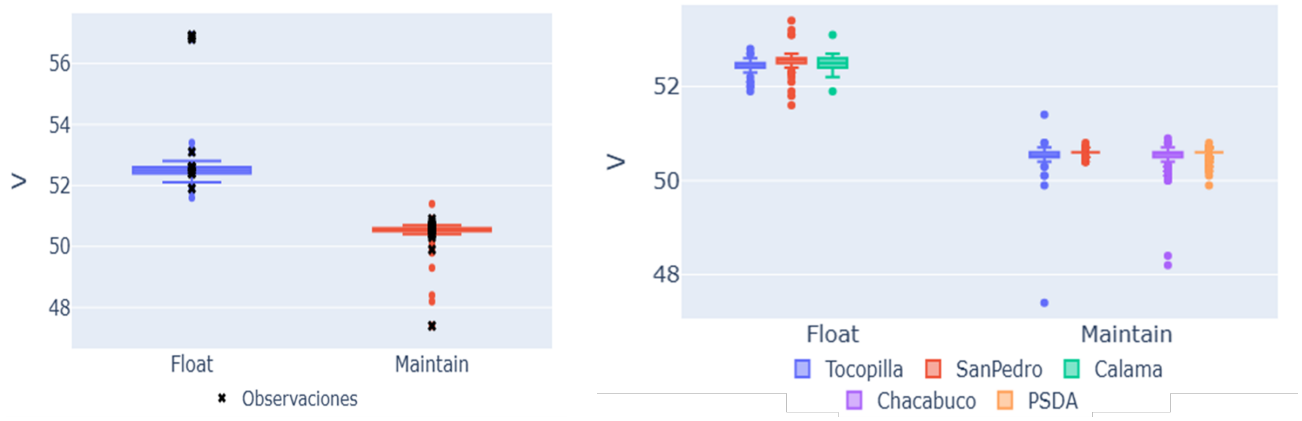
Este diseño buscaba concordancia con la lógica de operación de la planta piloto y la fenomenología del sistema PEMFC. Así, cada familia aportaba evidencia complementaria para ser analizada dentro del estrato basal (e). El detalle de configuración de los 37 detectores, tales como parámetros de DBSCAN, se encuentra disponible en el Apéndice A.2. En esta sección se presentan los hallazgos principales de cada familia con ejemplos representativos de referencia.

8.2.1 Familia 1: Detección univariante sobre tensión y corriente eléctricas

La familia F1 captura anomalías univariantes sobre las variables eléctricas fundamentales (V y I). Para la tensión eléctrica (V), se realizó la detección mediante *boxplots* por Modo y estratificado por Sitio, con umbrales de 1.5 veces el rango intercuartílico (IQR), agregando una corrección *offset* ($\approx 0.2 - 0.3$ V) debido a su distribución altamente compacta.

La Figura 8.3 muestra que las distribuciones de tensión se concentran en rangos estrechos con presencia de valores aislados de estas bandas principales. Estas anomalías podrían originarse por errores de medición, transición entre modos u operación fuera del horario regular, como es el caso del valor extremo en Float (~ 56.5 V), cuya medición se registró a las 14:00 hrs en Chacabuco.

A su vez, la detección de *outliers* de corriente I se realizó a nivel de Session_ID dentro de cada Sitio mediante *boxplots* (1.5 IQR). Por consiguiente, F1 aportó una primera capa de evidencia de anomalías instrumentales y operativas que se integra posteriormente con señales estructurales más amplias del resto de familias.



(a) (b)

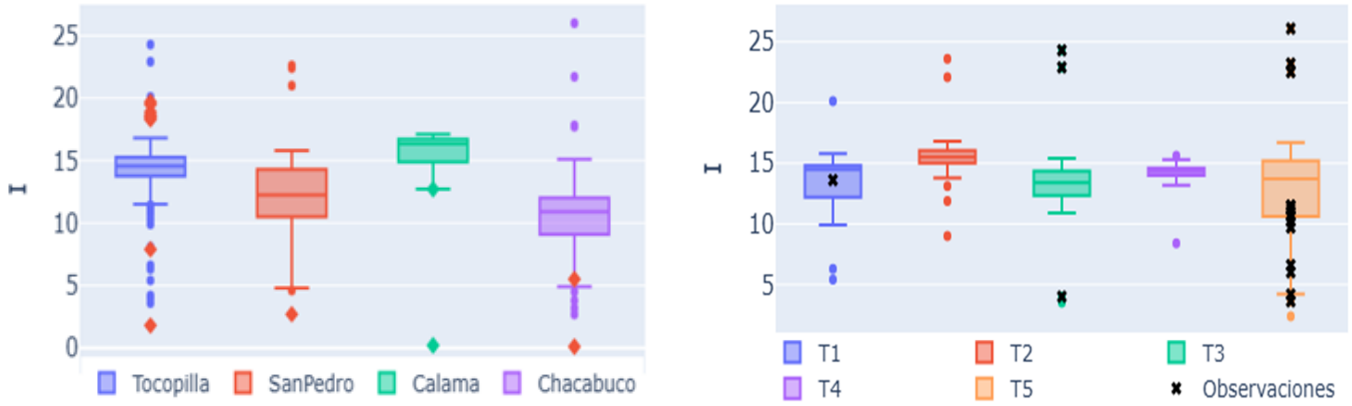
Figura 8.3: Familia 1: Detección de outliers de tensión eléctrica: (a) por Modo; y (b) bajo estratificación Sitio × Modo.

8.2.2 Familia 2: Corrientes, repartos y subespacios composicionales por K_act y Config

La familia F2 orientó la detección de *outliers* sobre el reparto de corrientes entre las PEMFC según el régimen de pilas activas (K_{act}) y su configuración (Config). Esto permitió extender el análisis de corrientes individuales de las FC, mediante el uso de variables derivadas como la suma de corrientes $S = I_1 + I_2 + I_3$, la entropía balanceada H_{active} y las coordenadas ILR de las composiciones de reparto (X_1, X_2, X_3) como indicador relativo del consumo de hidrógeno del banco PEMFC. Este análisis se estructuró para los tres regímenes operativos: monocelda (K1), bicelda (K2) y K3 (tricelda).

De acuerdo con el diagrama de Sankey global, el régimen K1 está compuesto principalmente por la FC_1 con 531 datos para la Config = '100'. Un análisis preliminar reveló que los puntos restantes de FC_2 y FC_3 (indicados con marcadores de diamante rojos y verdes) sesgaban las distribuciones a escala de 'Sitio', por lo que fueron descartados de este análisis (Figura 8.4(a)). Si bien la mayoría de los sitios presentaban distribuciones más estables para la detección, Tocopilla se caracterizó por un rango intercuartílico acotado y una mayor heterogeneidad en las colas extremas, lo cual no implicaba que estos puntos correspondieran a *outliers*, puesto que no eran eventos poco frecuentes que justificaran su eliminación.

Esta problemática se abordó al profundizar sobre las distribuciones por Session_ID, como se ilustra en la Figura 8.4(b). Para esta estratificación más fina, las sesiones muestran medianas y rangos más diferenciados y los *outliers* se encuentran aislados del IQR de los *boxplots*, tanto en las colas superiores e inferiores. Cabe mencionar que F2-K1 complementó al detector de corriente eléctrica de F1, el cual no realizaba esta segregación por regímenes y sesiones.



(a) (b)

Figura 8.4: Familia 2-K1: (a) Detección de *outliers* de FC_1 por Sitio; y (b) análisis por Session_ID en sitio Tocopilla.

En la Figura 8.5 se muestra el diagrama de Sankey para el régimen bicelda, el cual muestra el flujo de datos desde el régimen K2 hacia Session_ID y su contribución a Config. Esta arquitectura revela cuáles son las sesiones y configuraciones contienen la mayor parte de los datos y cuáles se conforman a partir de cantidades marginales como es el caso de la Config = '101'. Esta representación, complementaria a su tabla dinámica de origen, fue clave para definir los estratos de detección. Por ejemplo, para la Config = '110', se optó por realizar las detecciones por Sitio, mientras que PSDA contaba con sesiones con datos suficientes para su aplicación (P1–P4).

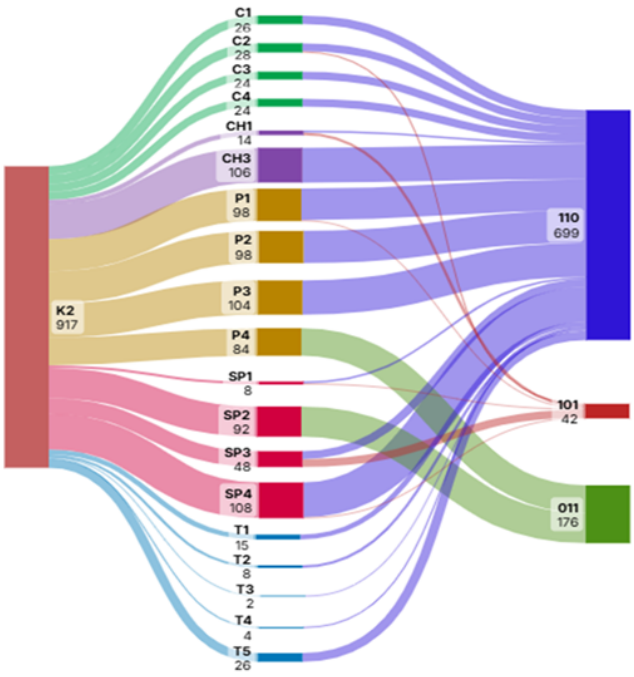


Figura 8.5: Diagrama de Sankey con cantidades de datos por Session_ID filtrados por K_act = 2.

Para el régimen bicelda (K2), se inició con la detección estratificando por FC_ID. Posteriormente, se recurrió a las variables ln(S) y H_{active} y se realizó la detección sobre PC1 empleando *boxplots* (criterio 1.5 IQR). En la Figura 8.6 se representan los resultados del DBSCAN sobre el plano $(H_{active}, \ln(S))$ para la configuración '110', donde se indican por color los clústeres formados. Se visualizan nubes principales más densas que corresponden a regímenes regulares de corriente total y reparto, en términos de la entropía; mientras que los puntos 'Outlier DBSCAN' se encuentran en regiones de baja densidad, y corresponderían a combinaciones inusuales dentro del régimen. Estos detectores se aplicaron para el resto de las configuraciones del régimen bicelda. Como resultado, se identificaron combinaciones composicionales poco frecuentes.

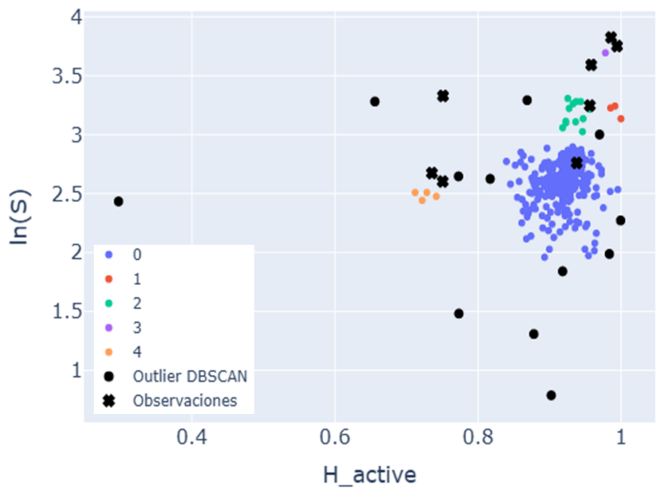


Figura 8.6: Familia 2-K2: Detección en coordenadas $(H_{active}, \ln(S))$ mediante algoritmo DBSCAN.

Por último, para el régimen tricelda (K3) se repitieron las detecciones de *boxplot* por FC_ID y análisis de PC1, tal como se aplicó para K2. Luego, se empleó la representación composicional ILR (Z_1, Z_2), sobre las cuales se aplicó nuevamente el algoritmo DBSCAN a escala global y para Chacabuco, con cantidad de puntos suficientes como indica el diagrama de Sankey global. Los resultados se muestran en la Figura 8.7, donde se evidencia un clúster principal en este subespacio de repartos relativamente equilibrados (más cercanos al origen) y los *outliers* fueron asignados a conjuntos de puntos periféricos con desbalances más marcados, en los cuales una o dos pilas operan a mayor capacidad y la tercera apenas produce corriente.

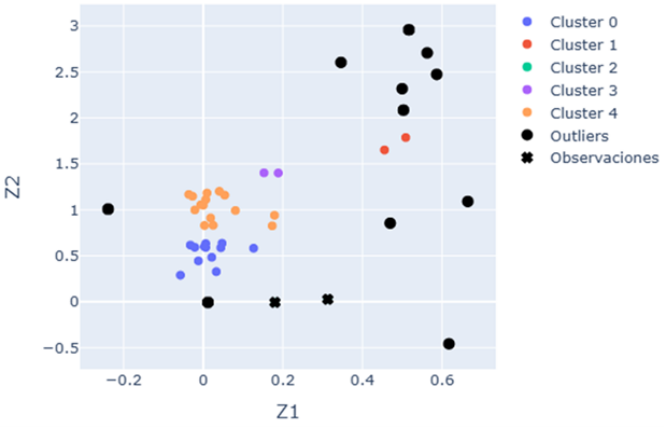
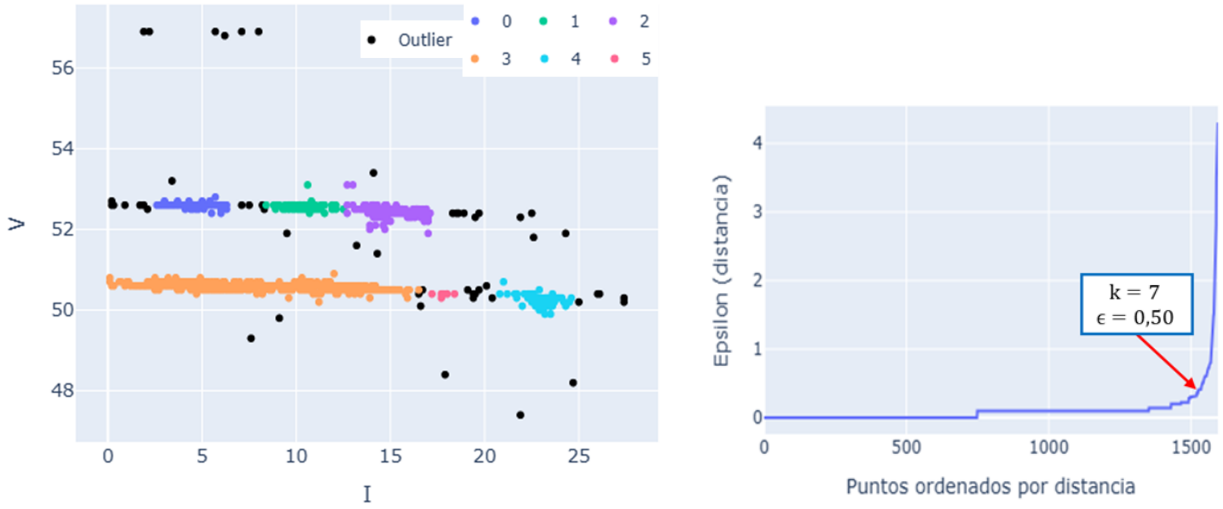


Figura 8.7: Familia 2-K3: Detección mediante DBSCAN sobre espacio composicional ILR.

8.2.3 Familia 3: Detección bivariada sobre la curva de operación I-V

La familia F3 abordó anomalías bidimensionales sobre la curva de polarización I–V de las PEMFC mediante la aplicación de DBSCAN a escala Global y por Sitio. Los parámetros `eps` y `min_samples` fueron calibrados a partir de las curvas *k-distance* mediante *elbow method*. La Figura 8.8 muestra un ejemplo de los clústeres principales y el ruido detectado ("Outlier" con color negro). Estas observaciones corresponden a puntos de operación con patrones irregulares, tales como baja tensión a un nivel de corriente moderada y son, por tanto, mediciones inestables o potenciales configuraciones inconsistentes. De este modo, la familia F3 detectó *outliers* que pudieran haberse pasado por alto durante las detecciones univariantes y composicionales (F1 y F2).



(a) (b)

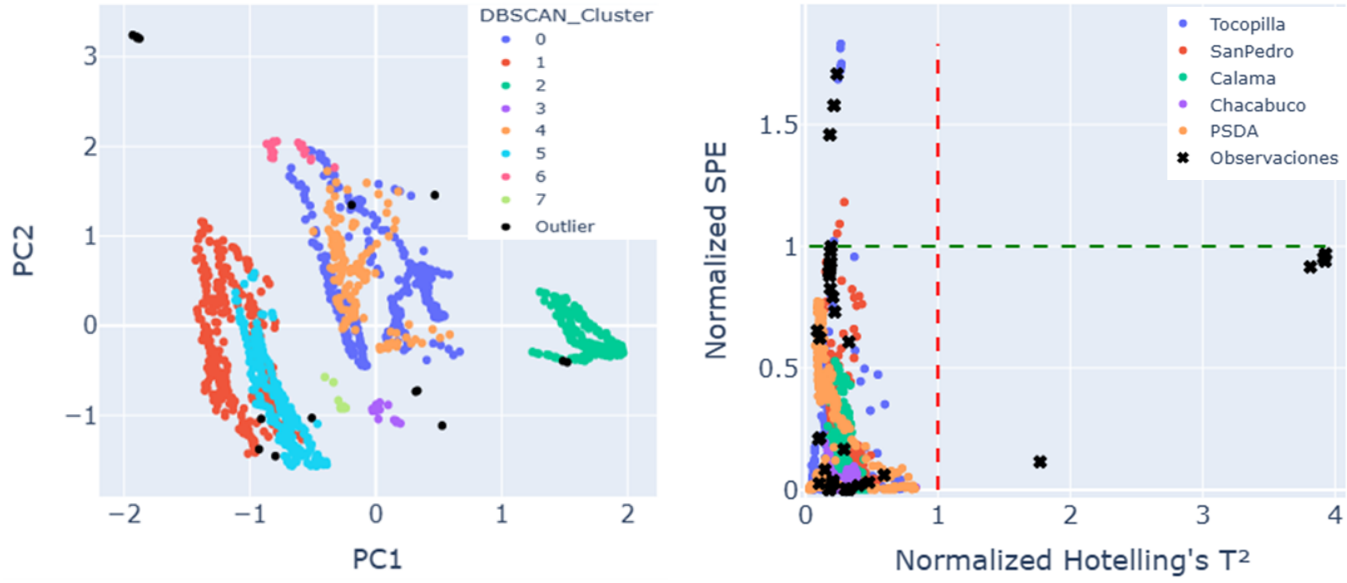
Figura 8.8: Familia 3: (a) DBSCAN sobre curva I–V; (b) gráfica *k-distance* para calibración del algoritmo (escala global de la BD).

8.2.4 Familia 4: Detección de *outliers* multidimensionales

La familia F4 realizó la detección multidimensional entre las variables ambientales y la tensión eléctrica mediante clusterización DBSCAN sobre la proyección de 3 componentes principales. A partir de este PCA, se calcularon los estadígrafos Hotelling's T^2 y SPE, definiendo límites de control del 99% de confianza para la detección de anomalías. Estos métodos se aplicaron a escala Global y por Sitio.

La Figura 8.9 presenta los resultados globales. En el panel (a), la proyección PC1 vs PC2 muestra clústeres bien definidos que agrupan las combinaciones climáticas y de tensión típicas de la campaña, mientras que los puntos periféricos en el espacio tridimensional fueron asignados como ruido. En el panel (b), los registros fuera de los límites de control para los estadígrafos

T^2–SPE normalizados (marcados con líneas segmentadas) corresponderían datos incompatibles estructuralmente, tanto por posibles errores de registros como por regímenes operativos extremos no representativos de la BD.



(a) (b)

Figura 8.9: Familia 4: Detección de outliers estructurales: (a) proyección en PC1–PC2 de DBSCAN sobre componentes principales; (b) T^2 versus SPE normalizados con umbrales de 99% de confianza y puntos fuera de control.

8.2.5 Familia 5: Variable objetivo (W)

La familia de detección F5 operó directamente sobre la variable objetivo W. En una primera fase, se analizaron los *boxplots* y percentiles a nivel Global y en el estrato Sitio×FC_ID, la cual se ilustra en la Figura 8.10. Este gráfico revela que las PEMFC presentaron diferencias en cuanto a magnitud y dispersión de potencia. Se puede notar que las cajas más amplias corresponden a grupos con mayor dispersión y las nubes de puntos contiguas muestran la multimodalidad de algunos grupos, por lo que se aplicaron umbrales de 3 IQR como criterio más restrictivo, dado el carácter crítico de esta variable.

Para la segunda fase, se aplicó DBSCAN sobre W estratificado por Sitio×FC_ID para identificar puntos de baja densidad local de potencia eléctrica. En la Figura 8.11 se ilustra la agrupación dentro de cada estrato, con nubes principales de color rojo y el ruido clasificado por el algoritmo en color negro, además de las instancias marcadas con 'Observaciones', con una buena tasa de coincidencia con los registros clasificados como ruido.

Con todo esto, el aporte de evidencia de la familia F5 contribuyó a penalizar potencias extremas y poco representativas de la BD y se procedió posteriormente a la integración de todas las familias mediante la metodología FASEK5 de cálculo de puntaje probabilístico que se discute a continuación.

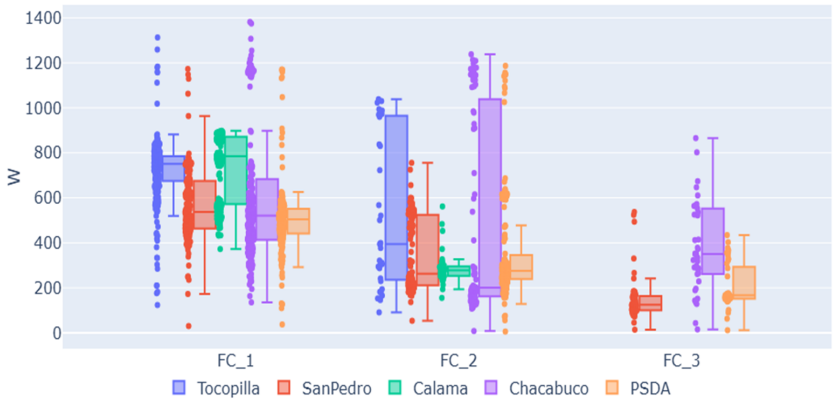


Figura 8.10: Familia 5: Detección mediante *boxplots* de potencia eléctrica por Sitio y FC_ID.

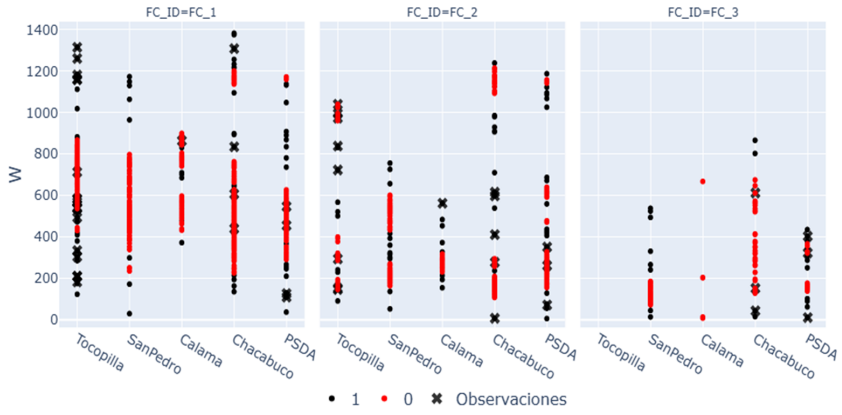


Figura 8.11: Familia 5: Detección aplicando DBSCAN sobre potencia eléctrica (Sitio × FC_ID).

8.3 Metodología de puntaje probabilístico de *outliers* FASEK5

La consolidación de las evidencias compuestas de los 37 detectores de las cinco familias (F1–F5) se desarrolló mediante un enfoque probabilístico basado en el modelo de compuertas *leaky noisy-OR* utilizado en redes bayesianas. Para esta tarea, se codificaron las detecciones mediante variables o *flags* binarios $z_{i,m} \in \{0, 1\}$, donde el valor 1 indica que el método m marcó el registro i como *outlier*.

La activación de cualquiera de estos *flags* ya sugería que el dato fuera candidato a *outlier*; sin embargo, su aporte debía ponderarse según la familia de origen y el esquema de estratificación, en lugar de simplemente sumar detecciones ignorando las diferencias entre métodos. Con esto, se

preservó la lógica de que rutas de evidencia alternativas converjan a la conclusión de que el registro podría ser ruido.

El esquema de detección requirió incorporar ajustes al modelo *leaky noisy-OR* para controlar factores como la redundancia entre métodos de una misma familia y su especificidad, la confiabilidad muestral de los estratos; además de asignar una jerarquía a cada familia de detección, en función de las variables físicas y el impacto sobre el modelado. A esta metodología se le denominó 'FASEK5' (puntaje global de las cinco familias según severidad y estratificación por K_act).

8.3.1 Formulación del modelo

La metodología FASEK5 se desarrolló a partir de dos niveles de agregación de evidencias:

1. *Noisy-OR* interno de agregación de los métodos $m \in M_K$ de la familia de detección $K \in \{1, \ldots, 5\}$.
2. *Noisy-OR* para cálculo del puntaje global S_i de agregación entre probabilidades de las familias.

En primer lugar, se modela la probabilidad conjunta de los métodos de cada familia F_{iK} como:

$$F_{iK} = 1 - (1 - l_K) \prod_{m \in M_K} (1 - p_{m|K})^{\rho_K z_{im}} \tag{8.2}$$

Donde:

- $p_{m|K}$ es una probabilidad de eficacia de detección del método m, la cual se asignó conforme a los siguientes reglas de jerarquización (detalle en el Apéndice A.3):
  - T^2/SPE con límites de control de 99% de confianza: máxima especificidad para desviaciones estructurales. Se le asignó mayor fiabilidad al ser una práctica estándar en control estadístico multivariante.
  - DBSCAN: alta especificidad al capturar clústeres de geometrías no convexas y ruido aislado, sin embargo, su efectividad es altamente sensible a la calibración de los hiperparámetros `eps` y `minPts`.
  - Puntos fuera de percentiles extremos y *boxplots* con límites de 3 IQR: especificidad media-baja. Son métodos útiles para EDA univariante, pero presentan sesgo y podrían arrojar falsos positivos para distribuciones con presencia de subgrupos.
  - *Boxplots* con umbrales de 1.5 IQR: baja especificidad, al ser criterios de detección más conservadores en una sola dimensión con influencia fuerte de distribuciones asimétricas.
- l_K es el parámetro de fuga de la familia K y corresponde a la probabilidad marginal de que el dato sea anómalo aun cuando ninguno de los *flags* lo detecte. Considera factores fuera de

control de naturaleza estocástica y aporta un margen de flexibilidad para posibles aspectos no cubiertos.

- ρ_{K} es el factor de redundancia de métodos por familia que se calcula a partir del coeficiente de Jaccard promedio $\bar{\mathrm{J}}_{\mathrm{K}}$ y un parámetro de calibración λ. Se definió como:

$$\rho_K = 1 - \lambda \bar{J}_K \tag{8.3}$$

El coeficiente de Jaccard $\mathrm{J_K}$ es una métrica del grado de similitud de dos conjuntos y se calcula a partir de la razón entre la cardinalidad de la intersección y la unión (Niwattanakul et al., 2013). Luego, se calcula el promedio de todos los pares por familia de cada estrato de cálculo. De forma que, si dos métodos coinciden en sus detecciones, se atenúa el efecto de solapamiento reduciendo ρ_{K}.

Sucesivamente, el puntaje probabilístico global se obtiene al combinar los $\mathrm{F_{iK}}$ de las familias. Este se modeló como:

$$S_i = 1 - (1 - L_0) \prod_{K=1}^{5} (1 - F_{iK})^{a_{K,\mathrm{eff}}} \tag{8.4}$$

$\mathrm{L_0}$ es el parámetro de fuga global y corresponde a la probabilidad marginal de que un registro sea *outlier* incluso si todos los métodos y familias no pudieran detectarlo, y el exponente $\mathrm{a_{K,eff}}$ se incorporó para modular la importancia por estrato y familia. Se definió como:

$$a_{K,\mathrm{eff}} = a_{K,\min} + (a_{K,\max} - a_{K,\min}) \left[1 - \exp\left(-\gamma \frac{w_K}{\tilde{w}_K}\right)\right] \tag{8.5}$$

El parámetro γ se utilizó para regular la interpolación de este exponente en el rango $[\mathrm{a_{K,min}}, \mathrm{a_{K,max}}]$, $\mathrm{w_K}$ es un peso compuesto y $\tilde{\mathrm{w}}_{\mathrm{K}}$ es la mediana por familia. Este se formuló como:

$$w_K = c_K^{1.25} \cdot r_K^{1.50} \cdot h_K^{2.50} \tag{8.6}$$

Donde $\mathrm{c_K}$ es el coeficiente de ponderación por familia, $\mathrm{r_K}$ es el factor de confiabilidad del tamaño muestral de los estratos $\mathrm{n_K}$ respecto a un tamaño de referencia T:

$$r_K = \frac{n_K}{n_K + T} \tag{8.7}$$

A partir de esta relación, si $\mathrm{n_K} \ll \mathrm{T}$, $\mathrm{r_K} \to 0$ regularizando la influencia de ese estrato y cuando $\mathrm{n_K} \gg \mathrm{T}$, $\mathrm{r_K} \to 1$ de forma asintótica.

Por último, $\mathrm{h_K}$ es el coeficiente de impacto de la familia K sobre los modelos, que cuantifica cuánto mejora o empeora el desempeño al remover el 3% de las mediciones de mayor probabilidad

F_{iK} ($\sim$ 50 datos). Se estimó mediante interpolación al calcular la variación relativa porcentual sobre la métrica RMSE.

Los exponentes de la ecuación (8.6) fueron asignados en función de los rangos de los factores c_K, r_K y h_K realizando ensayos paramétricos sobre la curva $a_{K,eff}$. Esto se discute brevemente en el Apéndice A.1. Los parámetros c_K, l_K, $a_{K,min}$, $a_{K,max}$ y h_K se encuentran disponibles en la Tabla A.10, mientras que los valores de los parámetros T, L_0, λ y γ en la Tabla A.11 del Apéndice A.5. Además, los resultados de la implementación de la metodología se encuentran disponibles en los Apéndices A.4 (detecciones por familia F1-F5) y A.6 (resultados integrados, probabilidad global S).

8.3.2 Resultados integrados del puntaje probabilístico FASEK5

La Figura 8.12 muestra la relación entre la cantidad de detectores activos y el puntaje probabilístico de FASEK5. Las barras verdes corresponden a la contaminación acumulada que se obtendría si el criterio fuese descartar todas las observaciones con al menos k detectores activos, y la curva roja corresponde a la probabilidad media global S para k detecciones. Se observa que el costo (i.e. cantidad de datos) sería excesivo si tan sólo se usaran los umbrales duros por cantidad de detecciones. Por ejemplo, para $k \geq 2$ se descartaría alrededor del 17% de la BD ($\approx$ 217 registros).

Esto confirma el beneficio de agregar múltiples familias y transformar a puntaje probabilístico, puesto que los registros con más de 5 evidencias originadas de múltiples familias contribuyen a una probabilidad global de *outlier* de 0.994, traduciéndose en un indicador más robusto que justifique su eliminación. Como medida adicional, se fijó un umbral máximo de contaminación del 5% ($\sim$ 80 datos) para preservar la representatividad de la BD, lo cual refuerza que la estrategia de estratificación para la estimación de S permitió concentrar la depuración sólo para la zona de mayor evidencia conjunta de las familias de detección sin exceder este umbral.

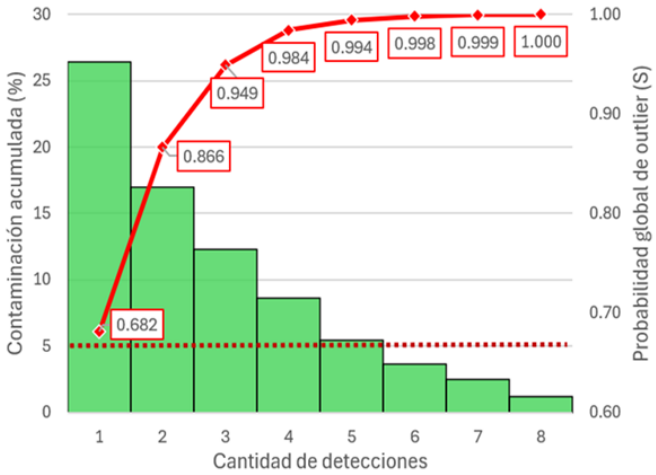


Figura 8.12: Contaminación acumulada (%) y probabilidad media por cantidad de detecciones.

Capítulo 9

Construcción y configuración de los modelos

Sobre la base del análisis exploratorio y la ingeniería de características descritos en los Capítulos 6 y 7, en conjunto con depuración de la BD acorde a la metodología probabilística FASEK5 discutida en el Capítulo 8, en éste se documenta la configuración de los modelos de regresión para predecir la potencia eléctrica del banco PEMFC.

Se describen de forma sintética las consideraciones generales del flujo de trabajo —en adelante, *pipeline*— común de modelado. Esto se refiere a la configuración específica de los modelos `RandomForestRegressor` (RFR), `XGBoost` (XGB), `CatBoost` (CAT), `MLPRegressor` (MLP) y `SupportVectorRegressor` (SVR); el esquema de validación cruzada utilizado en los entrenamientos y la metodología adoptada para el ajuste de los hiperparámetros utilizando el algoritmo de optimización bayesiana (BO), tarea crítica para reducir el sobreajuste e incrementar la capacidad de generalización de las predicciones.

El *pipeline* común incluye las siguientes tareas:

1. Ingesta de datos de la BD con outliers y sus versiones depuradas.
2. Preprocesamiento y división en conjuntos de entrenamiento y prueba bajo estratificación.
3. Construcción y entrenamiento basal de los modelos con validación cruzada.
4. Evaluación preliminar del desempeño, considerando ajuste y generalización.
5. Ajuste de hiperparámetros con refinamiento progresivo del espacio de búsqueda.
6. Entrenamiento y evaluación de los modelos calibrados.
7. Iteración para las tareas 5–6 hasta obtener desempeño aceptable.

Para efectos de comparabilidad, se empleó el RMSE como función de pérdida de la fase de entrenamiento y también como función objetivo de la BO. La división de los datos se realizó mediante el esquema *StratifiedKFold* con 5 grupos, utilizando el estrato basal 'SitioxK_actxConfig' y se utilizó la proporción 80/20 para entrenamiento y prueba. Para evaluar la calidad de los modelos se utilizaron las métricas RMSE, ME, SDE, r, R^2, MEC y CCC y su representación en los diagramas solar y de Taylor. El *pipeline* de modelado se resume en la Figura 9.1.

Adicionalmente, se definieron métricas complementarias para evaluar la capacidad de generalización de los modelos $\mathrm{GAP}_{\mathrm{RMSE}}$ y cuantificar la ganancia de la depuración de outliers, absoluta $\Delta\mathrm{RMSE}_{\mathrm{test}}$ y relativa porcentual $\mathrm{Gain}_{\mathrm{rel}}$. Estas se calcularon mediante las siguientes ecuaciones:

$$\mathrm{GAP}_{\mathrm{RMSE}} = \mathrm{RMSE}_{\mathrm{test}} - \mathrm{RMSE}_{\mathrm{train}} \tag{9.1}$$

$$\Delta\mathrm{RMSE}_{\mathrm{test}} = \mathrm{RMSE}_{\mathrm{test,out}} - \mathrm{RMSE}_{\mathrm{test,clean}} \tag{9.2}$$

$$\mathrm{Gain}_{\mathrm{rel}} = \frac{\mathrm{RMSE}_{\mathrm{test,out}} - \mathrm{RMSE}_{\mathrm{test,clean}}}{\mathrm{RMSE}_{\mathrm{test,out}}} \cdot 100 \tag{9.3}$$

Los subíndices 'train' y 'test' hacen referencia a los conjuntos de entrenamiento y prueba, mientras que 'out' y 'clean' corresponden a las versiones con outliers y depuradas de la BD. En cuanto a su interpretación, un $\mathrm{GAP}_{\mathrm{RMSE}}$ menor se traduce en un modelo con menor sobreajuste, mientras que valores mayores de $\Delta\mathrm{RMSE}_{\mathrm{test}}$ y $\mathrm{Gain}_{\mathrm{rel}}$ indican mayor sensibilidad de los modelos a los *outliers* detectados.

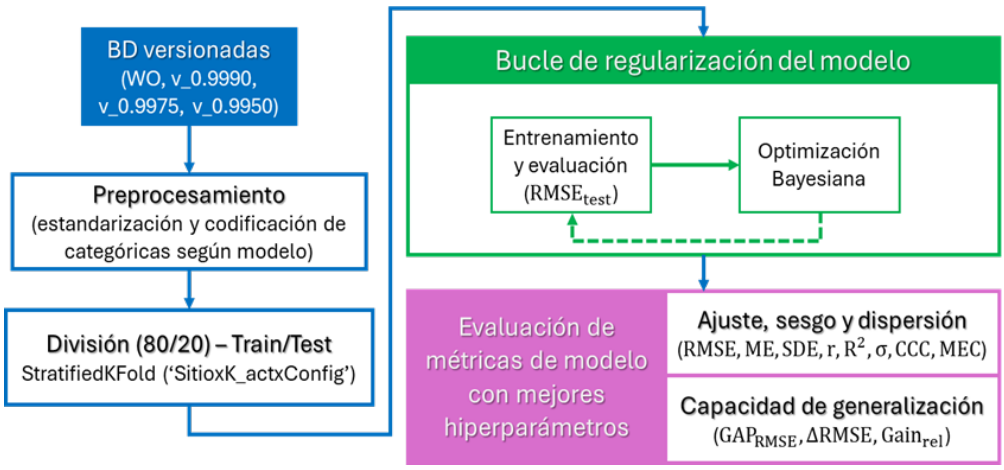


Figura 9.1: Esquema general del *pipeline* de modelado.

9.1 Arquitectura de los modelos y espacios de búsqueda

En cuanto a los modelos basados árboles y *gradient boosting*, para RFR y XGB se emplearon las codificaciones *One-Hot* para las variables categóricas. Por su parte, `CatBoost` no requirió este preprocesamiento de las variables, puesto que cuenta con manejo nativo de categorías. Además, CAT se configuró con condición de parada temprana y un máximo de 10 iteraciones sin variaciones del RMSE, que permiten detener el entrenamiento evitando sobreajuste. Esta clase de modelos no requiere que las variables continuas sean estandarizadas.

Tabla 9.1: Espacios de búsqueda de la optimización bayesiana y valores óptimos de hiperparámetros de cada modelo para las versiones WO y v_0.9950 de la base de datos.

Modelo	Hiperparámetro	Espacio de búsqueda	Valor óptimo (BD WO)	Valor óptimo (BD v_0.9950)
RFR	`n_estimators`	[100, 700]	100	639
	`max_depth`	[5, 15]	15	14
	`min_samples_split`	[10, 100]	10	10
	`min_samples_leaf`	[5, 30]	5	5
	`max_features`	{`'sqrt'`, `'log2'`, 0.5, 0.7, 0.9, 1.0}	0.7	0.7
	`ccp_alpha`	[0.0, 0.5]	0.425	0
XGB	`n_estimators`	[200, 500]	416	500
	`max_depth`	[4, 6]	6	6
	`eta`	[0.03, 0.07]	0.0584	0.0423
	`min_child_weight`	[3, 8]	3	3
	`subsample`	[0.75, 0.85]	0.85	0.85
	`colsample_bytree`	[0.75, 0.85]	0.75	0.75
	`reg_lambda`	[5.0, 50.0]	19.97	6.39
	`reg_alpha`	[5.0, 50.0]	5	5
	`gamma`	[5.0, 20.0]	20	5
CAT	`iterations`	1000	-	-
	`bootstrap_type`	`Bayesian`	-	-
	`learning_rate`	[0.01, 0.3]	0.1962	0.0887
	`depth`	[3, 10]	9	7
	`l2_leaf_reg`	[0.01, 10.0]	6.68	0.01
MLP	`hidden_layer_sizes`	(256, 128, 64, 32)	-	-
	`activation`	`ReLU`	-	-
	`solver`	`Adam`	-	-
	`alpha`	[1e-05, 1.0]	0.000021	0.000013
	`learning_rate_init`	[0.0001, 0.01]	0.00262	0.001121
SVR	`C`	[1e-06, 1000.0]	20.9	23.5
	`epsilon`	[1e-06, 10.0]	0.02669	0.00003
	`gamma`	[1e-06, 1000.0]	0.2059	0.1610

A partir de la Tabla 9.1, se observa que en la BD v_0.9950 los modelos RFR y XGB utilizaron un mayor número de iteraciones (`n_estimators`) y menor regularización, lo cual se aprecia en la reducción de HP como `ccp_alpha`, `eta` o `reg_lambda`, respectivamente y; este patrón se mantuvo para el modelo CAT, con `l2_leaf_reg` despreciable. Esto sugiere que el efecto de la depuración se tradujo en una regularización más suave al entrenar sobre una BD menos ruidosa.

El modelo SVR se implementó con *kernel* RBF para el manejo de no linealidades entre las variables de entrada y objetivo (W) y se seleccionaron rangos amplios para los HP, de modo de abarcar tanto configuraciones con regularización fuerte y mayor capacidad de ajuste. En conjunto con `MLPRegressor`, se estandarizaron las variables continuas (T_{amb}, p_{amb}, HR, H2, H3, Z1_3 y Z2_3) restando promedio y dividiendo por la desviación estándar; y se aplicó estandarización robusta a W sustrayendo la mediana y dividiendo por el rango intercuartílico, selección más apropiada en concordancia con la distribución observada.

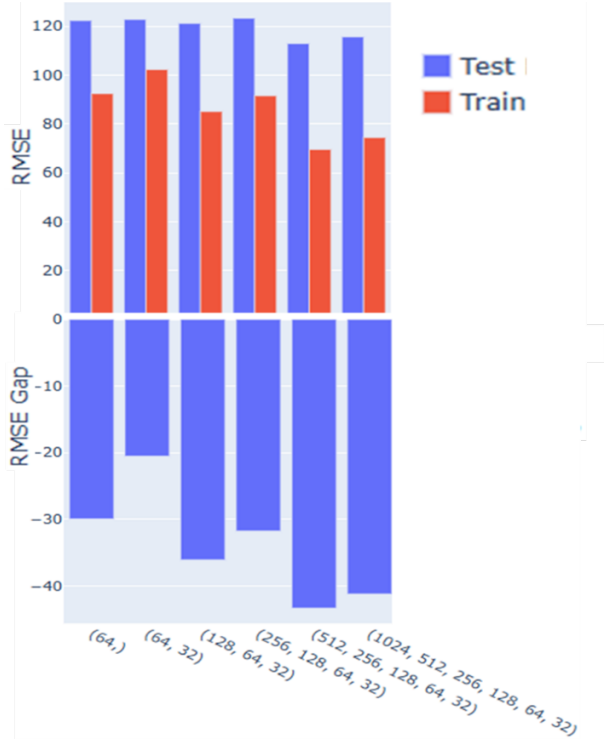


Figura 9.2: Desempeño base de `MLPRegressor` y GAP_{RMSE} por profundidad de las redes neuronales.

Por último, la configuración de MLP comenzó con la exploración preliminar (sin BO) de distintos tamaños de red sobre el desempeño en entrenamiento y prueba, abarcando desde 1–7 capas ocultas con una primera capa más densa y tamaños decrecientes para reducir los tiempos de entrenamiento. En la Figura 9.2 se despliegan los resultados de RMSE de entrenamiento y prueba en el panel superior y los GAP asociados en el panel inferior. Como solución de compromiso entre capacidad y brecha de generalización, se optó por la red de cuatro capas ocultas, que ofrecía un desempeño

razonable. Esta decisión metodológica se sustentó al considerar los RMSE obtenidos para el resto de los modelos, donde la arquitectura de 2 capas ocultas, que figuró como candidata potencial durante la fase exploratoria, se descartó al obtener RMSE de prueba mayor y menor capacidad en comparación con el resto de los modelos.

A partir de la arquitectura de 4 capas ocultas (256, 128, 64, 32) se aplicó optimización bayesiana con los hiperparámetros de la Tabla 9.1. Se utilizó la función de activación `ReLU`, optimizador `Adam` y se seleccionó un 10% de la base de datos para validación interna, opción disponible de `MLPRegressor` para control de sobreajuste. Además, se fijó condición de parada temprana bajo un máximo de 50 iteraciones sin variaciones del RMSE y tasa de aprendizaje constante a lo largo de toda la red.

9.2 Optimización bayesiana de hiperparámetros

Conforme a la literatura y ensayos preliminares, se definieron espacios de búsqueda relativamente amplios durante la primera iteración de la optimización bayesiana. Luego, se verificaban los resultados de RMSE y GAP. Si había margen potencial de mejora, se ajustaban los límites superior e inferior de los rangos preestablecidos, conforme al efecto esperado del HP sobre el sesgo y varianza de los modelos. En general, se requirieron hasta cuatro iteraciones para hallar regiones más promisorias.

La Figura 9.3 muestra los resultados de RMSE de las iteraciones del algoritmo de optimización bayesiana para el entrenamiento de `XGBoost` sobre la BD v_0.9950. Se observa una rápida disminución durante las primeras 20 evaluaciones y posteriormente converge paulatinamente. La presencia de saltos del RMSE se origina debido al esquema de exploración probabilístico de la BO. El comportamiento de esta curva fue similar para el resto de los modelos.

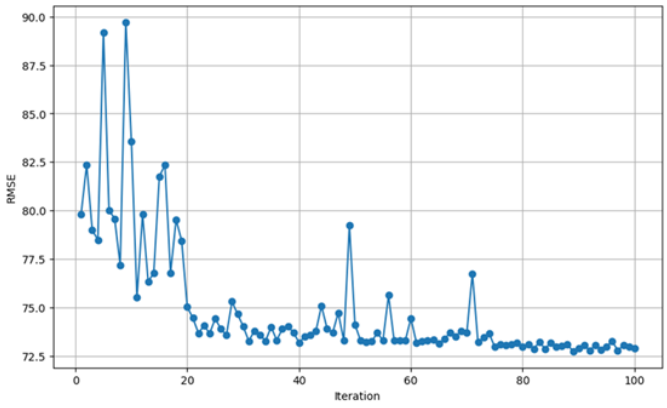


Figura 9.3: Evolución de RMSE durante la optimización bayesiana de hiperparámetros de la configuración XGB v_0.9950.

El consolidado de los espacios de búsqueda de hiperparámetros y configuraciones óptimas resultantes de la BO, tanto para la BD original (WO) y la versión depurada `v_0.9950`, se muestran en la Tabla 9.1. En ésta, se reporta el conjunto de hiperparámetros implementado para cada modelo (RFR, XGB, CAT, MLP, SVR).

En el capítulo siguiente se discuten a mayor profundidad los resultados de los modelos calibrados, evaluando su desempeño bajo las distintas versiones de la base de datos, utilizando el conjunto de métricas propuesto en este trabajo.

Capítulo 10

Resultados y Discusiones: Evaluación de los modelos predictivos

En los capítulos precedentes se desarrolló el *pipeline* de Aprendizaje Automático empleado para modelar la potencia eléctrica del banco de tres PEMFC a partir de los datos experimentales de la campaña itinerante multi-sitio en la Región de Antofagasta. Esta etapa incluyó: (i) la caracterización de la base de datos mediante EDA; (ii) la estrategia de ablación incremental e ingeniería de características implementada mediante experimentos con ANN en MATLAB; (iii) el desarrollo de la metodología FASEK5 para la depuración probabilística de *outliers*; y (iv) la descripción del *pipeline* de modelado y de las configuraciones específicas de la batería de modelos seleccionados (RFR, XGB, CAT, MLP, SVR), entrenados con validación cruzada estratificada (`StratifiedKFold`) y evaluados sobre la base de datos *Baseline*, compuesta por el conjunto de variables nativas y predictores derivados del contexto climato-topográfico, operativo y temporal.

El objetivo de este capítulo es evaluar cuantitativamente y comparar el desempeño de este conjunto de modelos en sus versiones entrenadas con la base de datos original con *outliers* (WO) y con las versiones depuradas (v0.9950, v0.9975, v0.9990), proporcionando los siguientes resultados:

- Seleccionar la configuración de modelo y la versión de la base de datos que ofrezcan el mejor compromiso entre precisión, robustez y capacidad de generalización.
- Analizar el desempeño del modelo de referencia mediante gráficas de validación y diagnóstico de los residuos de regresión.
- Identificar las características principales utilizadas para predecir la potencia eléctrica del banco PEMFC a partir de los gráficos de importancia generados para los modelos basados en árboles (RFR, XGB, CAT).

La evaluación se realiza mediante un conjunto de métricas de error y ajuste (RMSE, ME, SDE, r, R^2, MEC, CCC) y sus representaciones en el diagrama solar y el diagrama de Taylor, complementada con el análisis de métricas de generalización y de robustez frente a *outliers* (GAP_{RMSE}, $\Delta RMSE_{test}$, $Gain_{rel}$). Se incluye la desviación estándar de las predicciones (Std) utilizada en el diagrama de Taylor.

10.1 Comparación global de desempeño y efecto de remoción de *outliers*

La Tabla 10.1 detalla las métricas de desempeño estimadas en los entrenamientos sobre la BD con *outliers* (WO). Para esta versión, todos los modelos capturan una proporción significativa de la variabilidad de la potencia predicha, con coeficientes de determinación R^2 en el rango 0.91–0.97, eficiencias MEC entre 0.91–0.97 y coeficientes de concordancia CCC entre 0.95–0.98. Sin embargo, existen diferencias en la magnitud y la dispersión de los errores. XGB presenta el menor RMSE (50.6 W), equivalente a un 18 % de la desviación estándar de la potencia experimental (275.6 W), con sesgo medio despreciable (ME ≈ -0.07 W) y la SDE menor del grupo. MLP y CAT presentan RMSE intermedios, mientras que RFR y SVR registran RMSE cercanos a 81–83 W (≈ 30 % de la desviación estándar), lo que se traduce en residuos más dispersos (SDE mayor) y en valores de R^2 comparativamente menores que el resto del grupo cuando la BD posee registros anómalos.

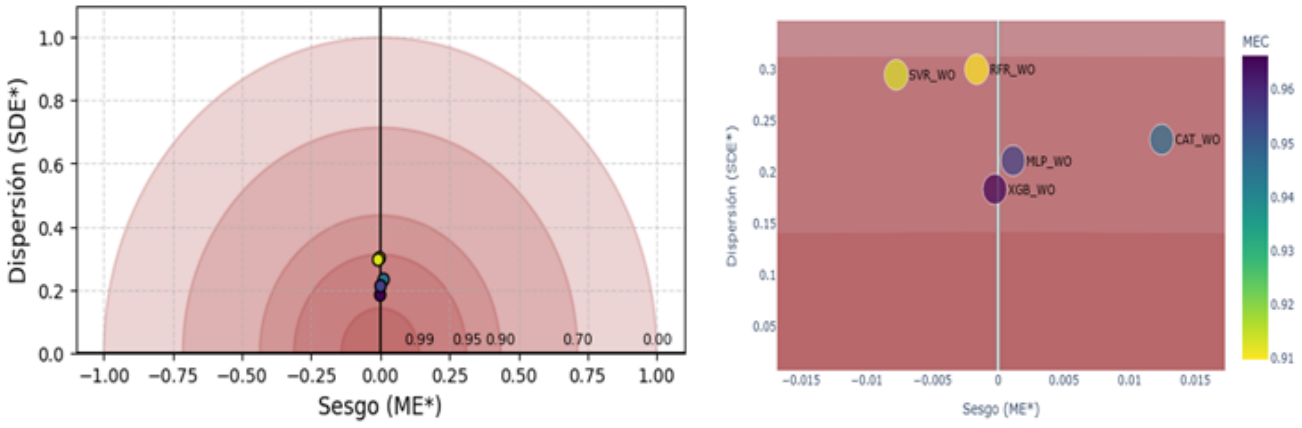
Tabla 10.1: Métricas globales de desempeño de los modelos – Base de datos con outliers (WO).

Conjunto	Std	ME	SDE	RMSE	r	R^2	MEC	CCC
Observations	275.6							
RFR_WO	251.2	-0.4485	82.760	82.761	0.9548	0.9117	0.9098	0.9508
XGB_WO	264.4	-0.0669	50.566	50.566	0.9833	0.9669	0.9663	0.9825
CAT_WO	263.9	3.4296	63.925	64.017	0.9729	0.9464	0.9461	0.9719
SVR_WO	264.7	-2.1339	81.281	81.309	0.9555	0.9131	0.9130	0.9547
MLP_WO	269.3	0.3151	58.241	58.242	0.9774	0.9553	0.9553	0.9772

De forma análoga, la Tabla 10.2 presenta las métricas para la BD v0.9950. La eliminación controlada de *outliers* mediante FASEK5 produce una mejora sistemática en todos los modelos, con reducciones del RMSE entre 8–20 W, sesgos medios despreciables y disminución de la SDE. En XGB, el RMSE se reduce hasta 30.7 W (≈ 11 % de la desviación estándar experimental), con r ≈ 0.994, $R^2 \approx 0.987$ y coeficientes MEC y CCC en torno a 0.990. La depuración beneficia también las métricas del resto de los modelos, manteniendo la jerarquía de desempeño observada en la base original. Estos patrones se representan gráficamente en los diagramas solar y de Taylor de las Figuras 10.1–10.4.

Tabla 10.2: Métricas globales de desempeño de los modelos – Base de datos depurada (v0.9950).

Conjunto	Std	ME	SDE	RMSE	r	R^2	MEC	CCC
Observations	275.6							
RFR_v0.9950	249.8	-0.3382	63.804	63.805	0.9715	0.9438	0.9426	0.9695
XGB_v0.9950	263.7	-0.0008	30.667	30.667	0.9936	0.9872	0.9876	0.9934
CAT_v0.9950	256.9	2.8746	44.079	44.173	0.9865	0.9731	0.9725	0.9858
SVR_v0.9950	260.0	-0.9560	61.069	61.077	0.9734	0.9475	0.9474	0.9731
MLP_v0.9950	259.1	0.9065	50.647	50.655	0.9818	0.9639	0.9638	0.9814



(a) (b)

Figura 10.1: Desempeño global de los modelos entrenados en la base de datos con outliers: (a) visión general del diagrama solar; (b) detalle ampliado con escala de color del coeficiente MEC.

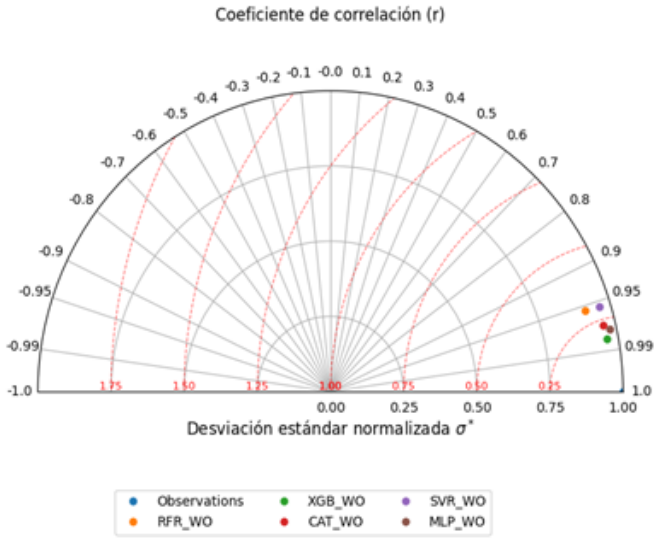
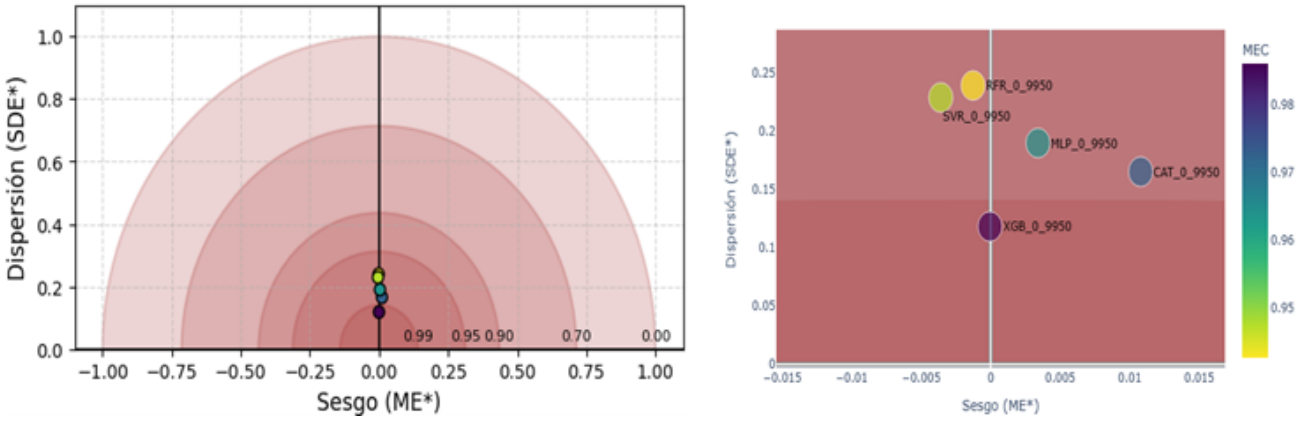


Figura 10.2: Desempeño global de los modelos en la base de datos con outliers: diagrama de Taylor.

En la BD con *outliers* (Figura 10.1), XGB se ubica más próximo al origen (SDE* ≈ 0.18, ME* ≈ 0), lo que se traduce en baja dispersión y sesgo despreciable de los residuos. CAT y MLP aparecen en posiciones algo más alejadas, mientras que RFR y SVR se sitúan en la zona superior, con SDE* ≈ 0.30. La escala de color (MEC) representa la eficiencia de modelado asociada a cada modelo. En el diagrama de Taylor (Figura 10.2), XGB combina una desviación estándar normalizada $\sigma^* \approx 0.96$ con el coeficiente de correlación mayor del grupo (r ≈ 0.98), ubicándose más próximo al punto de referencia (*Observations*).



(a) (b)

Figura 10.3: Desempeño global de los modelos en la base de datos depurada v0.9950: (a) visión general del diagrama solar; (b) detalle ampliado con escala de color del coeficiente MEC.

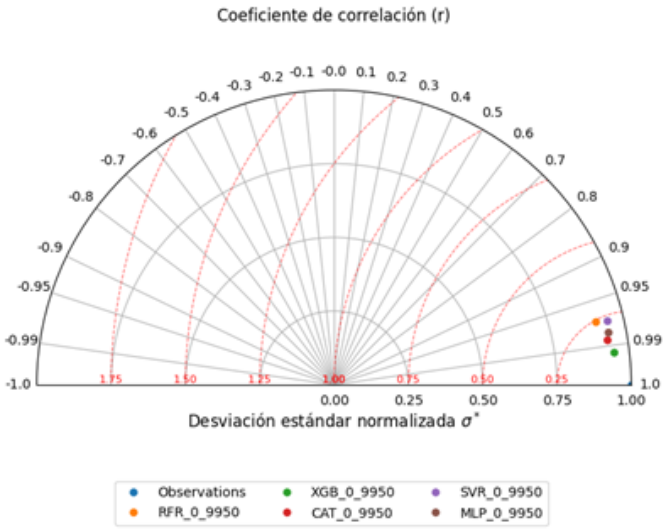


Figura 10.4: Desempeño global de los modelos en la base de datos depurada v0.9950: diagrama de Taylor.

Para la BD depurada (Figura 10.3), todos los modelos se desplazan hacia regiones de menor dispersión y mayor correlación, en coherencia con los resultados reportados en las Tablas 10.1 y 10.2. XGB es el que más se aproxima al punto de referencia: en el diagrama solar se sitúa con $\text{SDE}^* \approx 0.11$, $\text{ME}^* \approx 0$ y MEC ≈ 0.99, mientras que en el diagrama de Taylor (Figura 10.4) se ubica con $\sigma^* \approx 0.96$ y r ≈ 0.99. CAT y MLP también presentan mejoras apreciables, con SDE^* en torno a 0.16–0.18, aunque se ubican en una banda superior del coeficiente r, pero inferior a la de XGB. RFR y SVR, aunque más estables tras la depuración ($\text{SDE}^* \approx 0.22$–$0.23$), continúan por debajo del resto en términos de dispersión de errores y correlación.

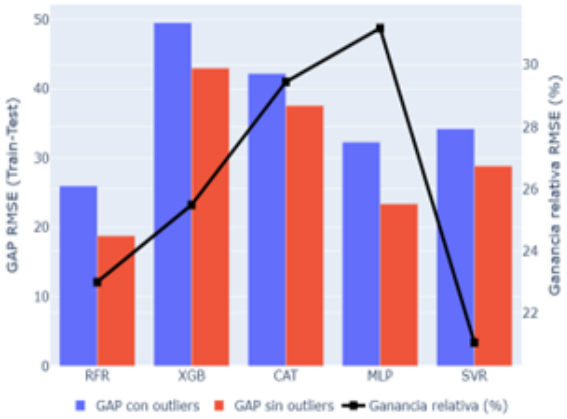


Figura 10.5: Efecto de la remoción de outliers sobre el GAP de RMSE y la ganancia relativa de los modelos.

La Figura 10.5 muestra la brecha de generalización y el efecto de la remoción de *outliers* sobre el RMSE de los modelos. Las barras corresponden a los valores de $\text{GAP}_{\text{RMSE}} = \text{RMSE}_{\text{test}} - \text{RMSE}_{\text{train}}$ para la BD con *outliers* (WO: azul) y depurada (v0.9950: rojo). La línea negra representa la ganancia relativa en $\text{RMSE}_{\text{test}}$ (Gain_{rel}) al remover los *outliers* bajo este umbral de probabilidad. El detalle de estas métricas se encuentra en la Tabla 10.3.

En todos los modelos, GAP_{RMSE} disminuye tras la depuración, indicando un mayor grado de generalización con variaciones entre $\sim$ 4.62 y 8.92 W ($\text{GAP}_{\text{out}} - \text{GAP}_{\text{clean}}$). XGB y CAT registran los mayores GAP_{RMSE} en ambas versiones de la BD, lo que sugiere mayor sobreajuste en entrenamiento; sin embargo, se debe considerar que son los modelos con mejor ajuste global, con $\text{RMSE}_{\text{test}}$ más bajos y coeficientes r y R^2 próximos a 1. RFR presenta los GAP_{RMSE} más reducidos, destacando por un comportamiento entrenamiento–prueba más estable, mientras que MLP y SVR exhiben resultados intermedios.

Tabla 10.3: GAP_{RMSE} (Test - Train) y ganancia absoluta (ΔRMSE) y relativa ($Gain_{rel}$) tras la depuración de *outliers* (conjunto de prueba).

Modelo	Train,out	Test,out	GAP	Train,clean	Test,clean	GAP	ΔRMSE	$Gain_{rel}$ (%)
RFR	87.24	113.2	25.97	68.39	87.18	18.79	26.03	22.99
XGB	48.11	97.63	49.53	29.77	72.76	42.99	24.88	25.48
CAT	59.75	102.0	42.20	34.36	71.93	37.58	30.02	29.45
MLP	84.70	117.0	32.30	57.13	80.52	23.38	36.48	31.18
SVR	72.84	107.1	34.23	55.69	84.54	28.85	22.53	21.04

Respecto de la sensibilidad a *outliers*, MLP y CAT alcanzan los valores más altos de $Gain_{rel}$ ($\approx$ 31 % y 29 %), seguidos por XGB con $\approx$ 25 % y, por consiguiente, su desempeño es más sensible a datos anómalos en el conjunto de prueba. En contraste, RFR y, en particular, SVR presentan ganancias menores ($\approx$ 21–23 %) y son, por tanto, más robustos al ruido de la BD utilizada, aunque con desempeño global en RMSE, SDE y métricas de ajuste inferior al de XGB.

En síntesis, los resultados de las Tablas 10.1–10.3 y las Figuras 10.1–10.5 posicionan a XGBoost como el modelo con mejor compromiso entre precisión, robustez frente a *outliers* y capacidad de generalización. Por este motivo se selecciona XGB_v0.9950 como modelo de referencia para el análisis detallado de las gráficas de validación, el comportamiento de los residuos y la identificación de las características más relevantes utilizadas por el algoritmo.

10.2 Análisis del modelo de referencia

10.2.1 Desempeño global y rectas de validación

La Figura 10.6 muestra las rectas de validación del modelo de referencia XGB_v0.9950 para los conjuntos de entrenamiento, prueba y global. En los tres paneles se observa una alineación de las predicciones en torno a la diagonal 1:1, indicativa de una buena concordancia entre observaciones y predicciones en todo el rango operativo de potencia del banco PEMFC. En el conjunto de entrenamiento se aprecia una dispersión reducida y pocas instancias con errores elevados. En el conjunto de prueba se registra un aumento moderado en la dispersión de los residuos hacia los extremos de potencia eléctrica, aunque sin pérdida relevante de linealidad ni evidencias de sesgos sistemáticos en rangos específicos. Este comportamiento es coherente con las métricas globales y de generalización reportadas en las Tablas 10.2 y 10.3.

A escala global, XGB_v0.9950 presenta un error típico RMSE $\approx$ 30.7 W, equivalente a $\approx$ 11 % de la desviación estándar experimental ($\approx$ 275,6 W), con valores de ME y SDE que indican un balance adecuado entre sesgo y varianza y ausencia de patrones marcados de error sistemático. De acuerdo con la Tabla 10.3, los RMSE de entrenamiento y prueba en la base depurada son $\approx$ 29.8 W

y $\approx$ 72.8 W, respectivamente ($GAP_{RMSE} \approx 43.0$ W). Este patrón es consistente con los coeficientes r, R^2, MEC y CCC estimados, y ratifica que las predicciones del modelo explican la mayor parte de la variabilidad de la potencia eléctrica del banco PEMFC bajo la heterogeneidad climática y las configuraciones operativas específicas de la planta piloto móvil.

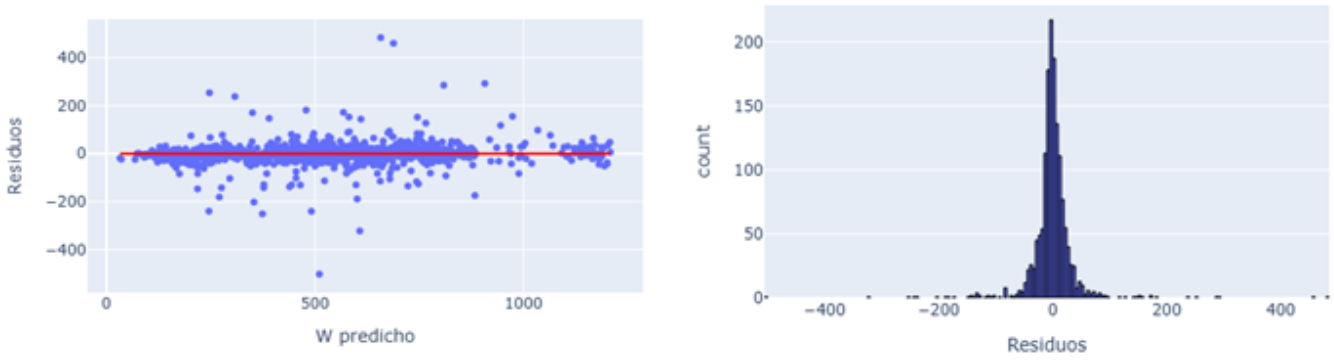
Figura 10.6: Desempeño del modelo de referencia XGB_v0.9950: rectas de validación en conjuntos de entrenamiento, prueba y global.

10.2.2 Análisis de residuos de regresión

Complementariamente al análisis de las rectas de validación, la Figura 10.7 muestra el análisis de residuos de XGB_v0.9950, a partir del gráfico de residuos en función de la potencia predicha y el histograma asociado.

En el gráfico de dispersión, los residuos se concentran en torno a cero en todo el rango de potencia, sin curvaturas ni patrones para valores específicos de W que sugieran errores sistemáticos presentes en el modelo. La banda principal se mantiene aproximadamente en el intervalo $\pm$100 W, con ciertas instancias que alcanzan desviaciones del orden de $\pm$400 W. Se aprecia un leve incremento de la dispersión en algunos rangos de potencia, atribuible a cierto grado de heterocedasticidad moderado y acotado.

El histograma de residuos muestra una distribución aproximadamente unimodal y centrada en cero, consistente con los resultados de ME y SDE reportados. En conjunto, esto respalda la calidad del ajuste obtenido a partir de las técnicas de ablación incremental, ingeniería de características y depuración de *outliers*, así como la validez de las predicciones generadas por el modelo.



(a) (b)

Figura 10.7: (a) Análisis de residuos; (b) histograma del modelo XGB_v0.9950.

10.2.3 Sensibilidad del modelo XGBoost a la remoción de *outliers*

La Tabla 10.4 resume el desempeño de XGBoost frente a distintos niveles de depuración, comparando la configuración entrenada sobre la base original (XGB_WO) con las variantes XGB_v0.9990, XGB_v0.9975 y XGB_v0.9950. Para estas versiones, los RMSE disminuyen hasta 30.7–32.5 W, lo que representa una reducción de 35–40 % respecto a XGB_WO, junto con disminuciones de magnitud similar para la SDE y un sesgo medio despreciable (ME ≈ 0 W). Los coeficientes r, R^2, MEC y CCC muestran incrementos marginales hasta alcanzar los valores de XGB_v0.9950, lo que refuerza el ajuste superior y mayor robustez frente a *outliers*.

Cabe mencionar que XGB_v0.9950 se adopta como referencia para mantener un criterio uniforme de comparación con el resto de los modelos, los cuales registraron variaciones más marcadas en sus métricas tras las depuraciones sucesivas de la BD.

Tabla 10.4: Efecto de remoción de *outliers* sobre las métricas globales del modelo XGBoost.

Conjunto	Std	ME	SDE	RMSE	r	R^2	MEC	CCC
Observations	275.6							
XGB_WO	264.4	-0.0669	50.566	50.566	0.9833	0.9669	0.9663	0.9825
XGB_v0.9990	260.9	-0.0142	31.532	31.532	0.9931	0.9862	0.9869	0.9929
XGB_v0.9975	262.4	-0.0482	32.468	32.468	0.9927	0.9855	0.9861	0.9925
XGB_v0.9950	263.7	-0.0008	30.667	30.667	0.9936	0.9872	0.9876	0.9934

10.3 Importancia de características de los modelos basados en árboles

Las Figuras 10.8 (XGB_v0.9950) y 10.9 (CAT_v0.9950, RFR_v0.9950) muestran que los tres modelos comparten un núcleo de variables asociadas a la arquitectura del banco PEMFC y al reparto de hidrógeno. En XGBoost, K_act y Config concentran la mayor parte de la importancia relativa (≈ 0.4 y 0.2), seguidas por FC_ID y H que, en conjunto, explican más de la mitad de la importancia acumulada. En `RandomForestRegressor` el *ranking* está dominado por H, FC_ID y K_act, mientras que en `CatBoost` destacan FC_ID, Amb y H, con una contribución relevante de Cod_Season. Esta estructura es consistente con los resultados de ablación incremental y sugiere que la potencia eléctrica está controlada, en primer lugar, por el régimen de activación, la configuración interna del banco y el reparto de combustible entre las pilas PEMFC y, en segundo lugar, por las condiciones ambientales locales y el contexto temporal específico.

Las variables Z, Modo y Cod_Season presentan importancias intermedias —salvo en CAT, donde Cod_Season se sitúa entre las más influyentes—, lo que indica que aportan señal complementaria para discriminar regímenes operativos. En contraste, los predictores Sitio y Season, así como Cluster_ID en RFR y CAT, muestran importancias de bajo orden. Esto sugiere que la variabilidad regional y estacional no actúa de forma directa a través de la etiqueta geográfica, sino que se transmite principalmente mediante las combinaciones de clima (Amb) y configuración interna. Un ejemplo de este comportamiento es XGBoost, que asigna más peso a Cluster_ID que a las variables ambientales continuas, utilizando la representación latente de regímenes climato–topográficos para realizar las predicciones.

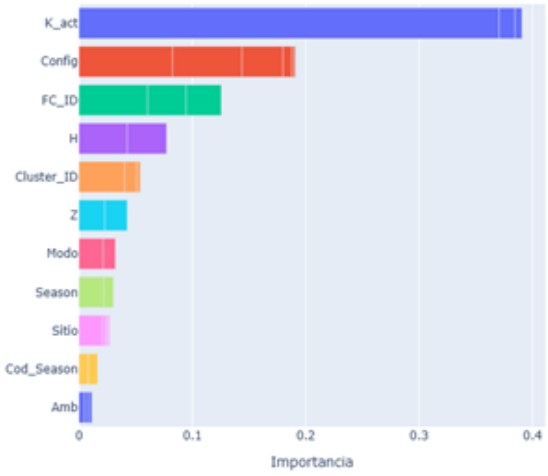
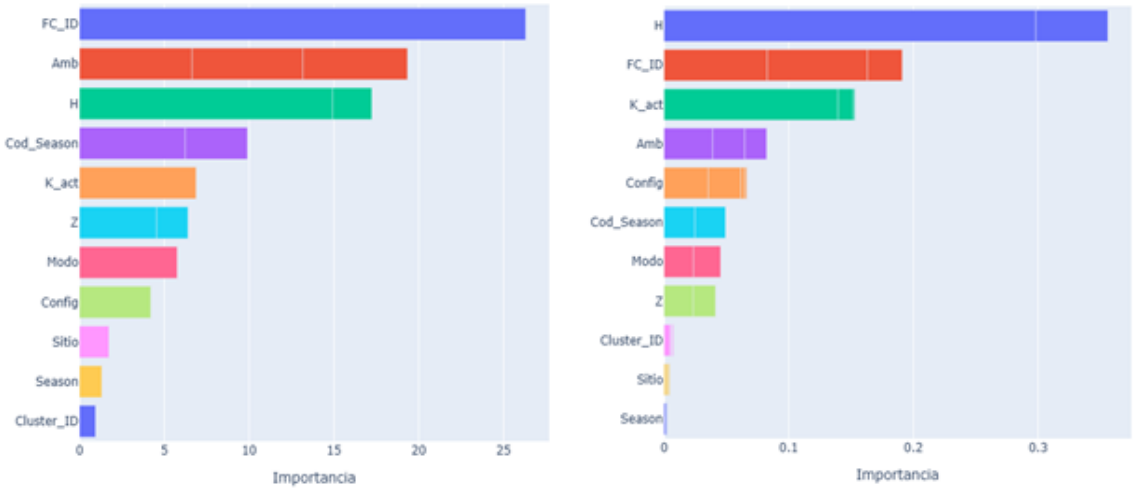


Figura 10.8: Gráfico de importancia de características del modelo de referencia XGB_v0.9950.



(a) (b)

Figura 10.9: Gráfico de importancia de características para los modelos basados en árboles: (a) CAT_v0.9950; (b) RFR_v0.9950.

10.4 Desempeño del resto de modelos: BD v.09950

La Figura 10.10 despliega las rectas de validación en el conjunto de prueba para los modelos RFR, CAT, MLP y SVR entrenados sobre la BD v0.9950. Las rectas de regresión (en color verde) y sus ecuaciones se muestran en el eje vertical, correspondientes a la relación entre los valores predichos (*Output*) y medidos (*Target*) de potencia eléctrica. Adicionalmente, se reportan los coeficientes de correlación r y de determinación R^2.

En general, se observa un alineamiento consistente respecto de la diagonal 1:1 (en rojo), y la magnitud de r y R^2 refleja precisamente el grado de desviación respecto de esa referencia. Bajo esta lectura, la jerarquía de calidad de ajuste sigue el orden CAT, MLP, SVR y RFR, con diferencias contrastantes en la dispersión y en las predicciones en los rangos extremos de potencia. Según la Tabla 10.3, los errores típicos $RMSE_{test,clean}$ se encuentran en el rango aproximado 71.9–87.2 W para estos modelos, orden de magnitud que se refleja en el comportamiento gráfico: CAT y MLP muestran nubes más compactas con algunos puntos de marcada dispersión, mientras que SVR y RFR presentan nubes relativamente más abiertas, con desviaciones notables en los rangos extremos de potencia.

Los patrones observados en las rectas de validación y las diferencias en $RMSE_{test}$ y en la dispersión de los residuos reafirman la calidad relativa superior de XGB_v0.9950 discutida en las secciones precedentes. Cabe destacar que los resultados de las rectas de validación corresponden exclusivamente a una de las permutaciones del esquema de validación cruzada `StratifiedKFold`, lo que explica diferencias menores frente a los valores reportados en las Tablas 10.2 y 10.3.

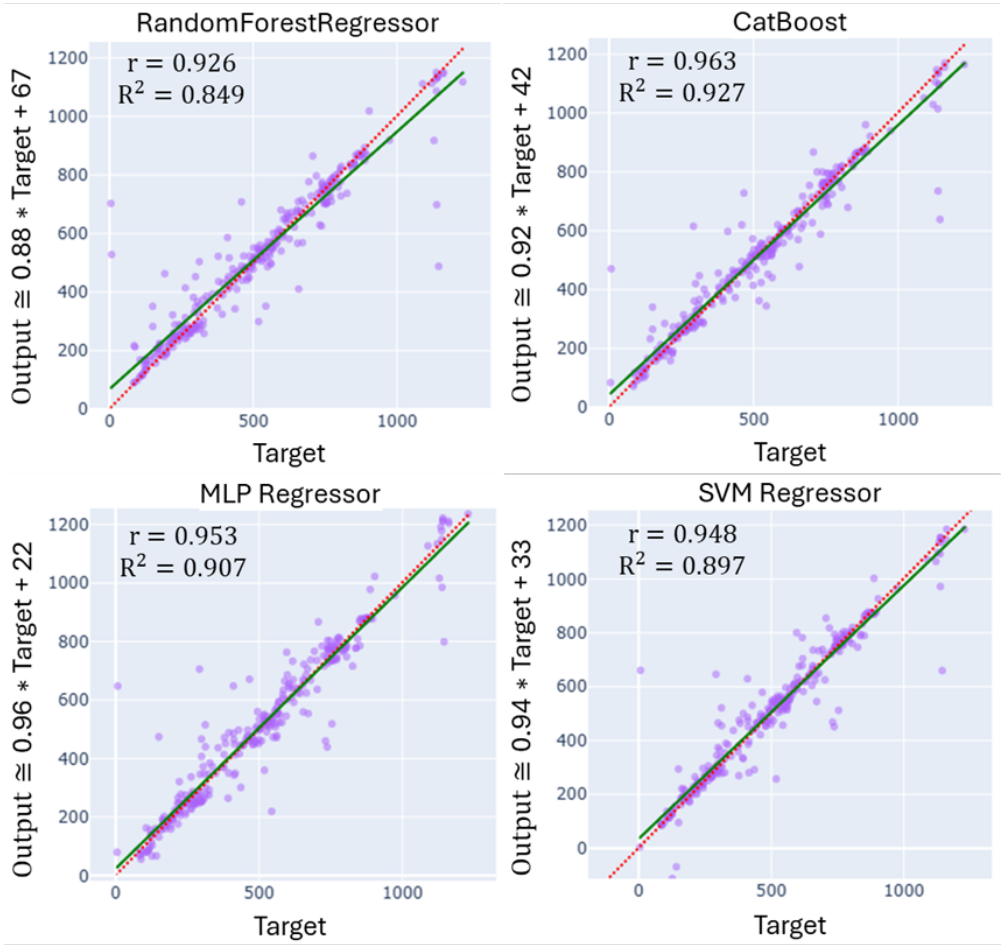


Figura 10.10: Rectas de validación del conjunto de prueba de los modelos RFR, CAT, MLP y SVR_v0.9950.

En síntesis, este capítulo presenta la evaluación comparativa de los modelos de Aprendizaje Automático seleccionados para predecir la potencia del banco PEMFC E-1100 de la planta piloto móvil. El análisis considera el efecto de la depuración probabilística de *outliers* sobre las métricas de ajuste, sesgo, dispersión y generalización, así como la distribución de importancia de las características para interpretar los modelos basados en árboles y sus variantes de *gradient boosting* (RFR, XGB y CAT). Estos resultados justifican la elección de XGB_v0.9950 como la configuración con mejor compromiso global entre desempeño e interpretación física de las variables utilizadas y derivadas mediante ingeniería de características, y constituyen la base empírica sobre la cual se formulan las conclusiones generales y las proyecciones de este trabajo.

Capítulo 11

Conclusiones y Recomendaciones

En este trabajo de tesis se desarrolló un *pipeline* de modelado basado en Aprendizaje Automático para predecir, con alta precisión y eficiencia, la potencia eléctrica de un banco PEMFC bajo condiciones de elevada heterogeneidad climática y topográfica en la Región de Antofagasta. Los datos experimentales provienen de una campaña itinerante de una planta piloto móvil de hidrógeno verde de CICITEM, que generó una base de datos limitada en tamaño, altamente heterogénea y multimodal, con desbalances en los estratos climático–operacionales y una marcada presencia de registros anómalos, lo que exigió robustecer los esquemas de modelado en coherencia con la física del sistema. Se registraron 1595 mediciones de variables ambientales (T_{amb}, p_{amb}, HR) y eléctricas (I, V, W) de tres pilas de combustible GenSure E-1100. El área de estudio se extiende desde la Cordillera de la Costa hasta zonas de la Precordillera y la Depresión Intermedia, con altitudes entre 1400 y 3000 m s.n.m.

El flujo de trabajo propuesto integra: (i) ingeniería de características acoplada a ablación incremental para construir una base de datos enriquecida con descriptores categóricos e híbridos de contexto climático, temporal, operacional y de reparto de combustible; (ii) una metodología probabilística de detección de *outliers* por familias (FASEK5), que combina múltiples detectores sobre distintas proyecciones físicas de las variables mediante un modelo *leaky noisy-OR* para generar versiones depuradas de la base de datos; y (iii) un *pipeline* robusto y reproducible que incorpora validación cruzada estratificada y optimización bayesiana para el ajuste de hiperparámetros de los modelos de regresión `RandomForestRegressor` (RFR), `XGBoost` (XGB), `CatBoost` (CAT), `MLPRegressor` (MLP) y `SupportVectorRegressor` (SVR), utilizando el RMSE del conjunto de prueba como función de pérdida. La calidad de los modelos se evaluó mediante un conjunto integrado de métricas de ajuste, sesgo y dispersión de residuos (RMSE, r, R^2, CCC, MEC, ME, SDE, Std), complementado con diagramas solares y de Taylor, y con el análisis de la capacidad de generalización y de la ganancia relativa asociada a la remoción de *outliers* (GAP_{RMSE}, ΔRMSE, $Gain_{rel}$).

La implementación de los experimentos de ablación incremental en la aplicación *Neural Net Fitting* de MATLAB permitió validar que, al integrar predictores diseñados a partir del conocimiento de la física del sistema y de la caracterización de la base de datos, es posible alcanzar un alto ajuste y desempeño global en redes neuronales artificiales (MLP) entrenadas con regularización bayesiana (RMSE $\approx$ 94 W, r $\approx$ 0.940, $R^2 \approx 0.884$). Estos experimentos mostraron que la potencia eléctrica del banco PEMFC en entornos reales no solo responde a la variabilidad climática, sino que depende directamente de las decisiones operativas que determinan la configuración interna del sistema. En particular, el modelo se benefició al incorporar la identidad, cantidad y orden de las pilas de combustible, así como el reparto de combustible utilizado en los regímenes bi y tricelda. No obstante, al tratarse de una etapa exploratoria, se requirió abordar el tratamiento del ruido y el ajuste de hiperparámetros para reducir la brecha de generalización entre los conjuntos de entrenamiento y prueba.

La depuración de *outliers* mediante FASEK5 permitió establecer umbrales de remoción coherentes con la estructura estratificada de la base de datos, evitando una eliminación indiscriminada de registros. El esquema probabilístico opera como un estimador menos sesgado por criterios aislados de detección, concentrando la depuración en la región de mayor evidencia conjunta, con probabilidades globales superiores a 0.9950 y una tasa de contaminación acotada al 5 % ($\approx$ 80 datos), preservando la representatividad física y operativa del conjunto. Bajo este esquema, la remoción controlada de *outliers* aumentó la capacidad de generalización de todos los modelos, con reducciones de $\mathrm{GAP}_{\mathrm{RMSE}}$ del orden de 4.6–9.0 W y ganancias relativas de RMSE de prueba en torno a 21–31 %. En particular, la mejora en la capacidad predictiva de XGBoost y CatBoost se tradujo en $\mathrm{GAP}_{\mathrm{RMSE}} \approx 43.0$ y $\approx$ 37.6 W, respectivamente. En términos globales, el desempeño de los modelos mejoró, desplazando sus posiciones en los diagramas solar y de Taylor hacia regiones de mayor calidad, con reducciones del sesgo medio y de la dispersión de los errores y una mayor predicción de la variabilidad observada en las mediciones experimentales de la potencia eléctrica del banco PEMFC.

Sobre este cimiento metodológico, la configuración XGB_v0.9950 se seleccionó como modelo con mejor compromiso entre precisión, robustez frente a *outliers* y capacidad de generalización. Esta configuración alcanzó el menor RMSE global ($\approx$ 31 W, 11 % de Std experimental), sesgo prácticamente nulo y valores de r ≈ 0.993 y $R^2 \approx 0.987$, junto con MEC y CCC del orden de 0.990, indicadores de buena alineación en las rectas de validación y de alta eficiencia global del modelo. El resto de los modelos (RFR, CAT, MLP y SVR) obtuvo desempeños de muy buena calidad; los valores promedio y desviaciones estándar fueron: RMSE $\simeq 50.1 \pm 13.4$ W, r $\simeq 0.981 \pm 0.01$ y $R^2 \simeq 0.963 \pm 0.02$. La desviación estándar media de las potencias predichas por estos modelos (Std $\simeq$ 257.9 W) equivale al 93.6 % del valor experimental σ ($\approx$ 275.6 W), lo que evidencia una buena reproducción de la variabilidad observada de la potencia en la base de datos depurada v0.9950. Estos resultados no contradicen el principio *No Free Lunch*, sino que lo operacionalizan:

bajo un *pipeline* común de preprocesamiento, depuración y evaluación, XGBoost emergió como la mejor alternativa para esta base de datos y este contexto operativo específicos.

El análisis comparativo de importancia de características para los modelos XGB, RFR y CAT reforzó la relevancia de la ablación incremental y de la ingeniería de características. Los resultados sugieren que la potencia eléctrica del sistema PEMFC está controlada, en primer lugar, por el régimen de activación, la configuración interna del banco y el reparto de combustible entre las pilas y, en segundo lugar, por las condiciones ambientales locales y el contexto temporal específico. Estos patrones son coherentes con la influencia esperada de la temperatura, la presión y la humedad relativa sobre el desempeño electroquímico y, al mismo tiempo, respaldan la selección de modelos de Aprendizaje Automático para predecir el desempeño de PEMFC en entornos reales.

Por lo tanto, el flujo de trabajo propuesto constituye un marco reproducible cuyos principios pueden extrapolarse a otros bancos PEMFC, a campañas de medición más amplias y a otros contextos climáticos. Los resultados de esta investigación muestran que es factible desarrollar modelos robustos para predecir la potencia eléctrica con un ajuste elevado. A partir de los principales hallazgos de este proyecto, se proponen las siguientes líneas de investigación para trabajos futuros de interés científico-tecnológico:

1. Desarrollar estrategias de control y optimización de la potencia eléctrica centradas en decisiones de configuración interna y en la gestión del reparto de flujo de hidrógeno entre *stacks* bajo distintas condiciones ambientales.

2. Entrenar y evaluar un metamodelo construido a partir de los modelos implementados en este trabajo (modelos "estudiantes"), explorando el esquema de agregación más adecuado y aplicando técnicas de interpretabilidad reportadas en la literatura (valores SHAP, diagramas ALE y método local LIME) para verificar la consistencia física de los resultados.

3. Reproducir y refinar el *pipeline* propuesto en otros entornos climáticos, a partir de datos experimentales de desempeño de un sistema PEMFC.

Estos avances pueden utilizarse como herramienta para la operación y planificación de sistemas PEMFC en entornos climato-topográficos desafiantes, contribuyendo a consolidar soluciones energéticas limpias y descentralizadas para la Región de Antofagasta, núcleo del despliegue de las tecnologías de hidrógeno verde en Chile.

Referencias

John Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160, 1982.

Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. 2010.

David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. Technical report, Stanford, 2006.

Frano Barbir. *PEM fuel cells: theory and practice*. Academic press, 2012.

Dawn M Bernardi and Mark W Verbrugge. Mathematical model of a gas diffusion electrode bonded to a polymer electrolyte. *AIChE journal*, 37(8):1151–1163, 1991.

Fatih Birol. The future of hydrogen: seizing today's opportunities. *IEA Report prepared for the G*, 20:442, 2019.

Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Katherine Calvin, Dipak Dasgupta, Gerhard Krinner, Aditi Mukherji, Peter W Thorne, Christopher Trisos, José Romero, Paulina Aldunce, Ko Barrett, Gabriel Blanco, et al. Ipcc, 2023: Climate change 2023: Synthesis report. contribution of working groups i, ii and iii to the sixth assessment report of the intergovernmental panel on climate change [core writing team, h. lee and j. romero (eds.)]. ipcc, geneva, switzerland. 2023.

Niccolò Cavagnero, Fernando Dos Santos, Marco Ciccone, Giuseppe Averta, Tatiana Tommasi, and Paolo Rech. Fault-aware design and training to enhance dnns reliability with zero-overhead. *arXiv preprint arXiv:2205.14420*, 2022.

Emigdio Chavez-Angel, Alejandro Castro-Alvarez, Nicolas Sapunar, Francisco Henríquez, Javier Saavedra, Sebastián Rodríguez, Iván Cornejo, and Lindley Maxwell. Exploring the potential of

green hydrogen production and application in the antofagasta region of chile. *Energies*, 16(11): 4509, 2023.

Tianqi Chen. Xgboost: A scalable tree boosting system. *Cornell University*, 2016.

Nicolás Arrioja Landa Cosio. Guía definitiva a bias-variance tradeoff. `https://medium.com/@nicolasarrioja/gu%C3%ADa-definitiva-a-bias-variance-tradeoff-94fb5c118d0f`, April 2022. Medium blog post.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

Rami Y Dahham, Haiqiao Wei, and Jiaying Pan. Improving thermal efficiency of internal combustion engines: recent progress and remaining challenges. *Energies*, 15(17):6222, 2022.

Thomas G Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923, 1998.

Rui Ding, Shiqiao Zhang, Yawen Chen, Zhiyan Rui, Kang Hua, Yongkang Wu, Xiaoke Li, Xiao Duan, Xuebin Wang, Jia Li, et al. Application of machine learning in optimizing proton exchange membrane fuel cells: A review. *Energy and AI*, 9:100170, 2022.

Harris Drucker, Christopher J Burges, Linda Kaufman, Alex Smola, and Vladimir Vapnik. Support vector regression machines. *Advances in neural information processing systems*, 9, 1996.

Greg D'Silva, Eashaal Mahmood, Rhodri Jervis, and Shangwei Zhou. Generalised fault diagnostics of polymer electrolyte fuel cells using machine learning. 2025.

Juan José Egozcue, Vera Pawlowsky-Glahn, Glòria Mateu-Figueras, and Carles Barcelo-Vidal. Isometric logratio transformations for compositional data analysis. *Mathematical geology*, 35(3): 279–300, 2003.

Rodrigo A Escobar, Cristián Cortés, Alan Pino, Marcelo Salgado, Enio Bueno Pereira, Fernando Ramos Martins, John Boland, and José Miguel Cardemil. Estimating the potential for solar energy utilization in chile by satellite-derived data and ground station measurements. *Solar Energy*, 121:139–151, 2015.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996.

Peter Filzmoser, Karel Hron, and Matthias Templ. Applied compositional data analysis. *Cham: Springer*, 2018.

Peter I Frazier. Bayesian optimization. In *Recent advances in optimization and modeling of contemporary problems*, pages 255–278. Informs, 2018.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

GeeksforGeeks. Dbscan clustering in ml – density based clustering. `https://www.geeksforgeeks.org/machine-learning/dbscan-clustering-in-ml-density-based-clustering/`, s. f.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

Sahra Hamdollahi and Luo Jun. A review on modeling of proton exchange membrane fuel cell. *Chemical Industry and Chemical Engineering Quarterly*, 29(1):61–74, 2023.

Trevor Hastie. The elements of statistical learning: data mining, inference, and prediction, 2009.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. An introduction to statistical learning. 2009.

Francisco Henríquez. Simulación multiescala de una pila de combustible tipo membrana de intercambio protónico para evaluar el efecto de parámetros operacionales sobre el desempeño bajo las condiciones climáticas de la región de antofagasta. Tesis para optar al grado de licenciado en ciencias de la ingeniería y al título de ingeniero civil químico, Universidad Católica del Norte, Antofagasta, Chile, jul 2025.

Théophile Hordé, Patrick Achard, and Rudolf Metkemeijer. Pemfc application for aviation: Experimental and numerical study of sensitivity to altitude. *International Journal of Hydrogen Energy*, 37(14):10818–10829, 2012.

Rob J Hyndman and George Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2018.

IRENA. Hydrogen from renewable power: Technology outlook for the energy transition. Technical report, International Renewable Energy Agency (IRENA), Abu Dhabi, United Arab Emirates, 2018. URL `https://www.irena.org/-/media/Files/IRENA/Agency/Publication/2018/Sep/IRENA_Hydrogen_from_renewable_power_2018.pdf`. Accessed: 2025-11-29.

Mohamed Issa, Mohamed Abd Elaziz, and Sameh I Selem. Enhanced hunger games search algorithm that incorporates the marine predator optimization algorithm for optimal extraction of parameters in pem fuel cells. *Scientific Reports*, 15(1):4474, 2025.

Yu Jianxing, Wu Shibo, Yu Yang, Chen Haicheng, Fan Haizhao, Liu Jiahao, and Ge Shenwei. Process system failure evaluation method based on a noisy-or gate intuitionistic fuzzy bayesian network in an uncertain environment. *Process Safety and Environmental Protection*, 150:281–297, 2021.

Balajee JM et al. Data wrangling and data leakage in machine learning for healthcare. 2018.

Ian Jolliffe. Principal component analysis. In *International encyclopedia of statistical science*, pages 1094–1096. Springer, 2011.

Trupti M Kodinariya, Prashant R Makwana, et al. Review on determining number of cluster in k-means clustering. *International Journal*, 1(6):90–95, 2013.

Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada, 1995.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Theodora Kourti and John F MacGregor. Multivariate spc methods for process and product monitoring. *Journal of quality technology*, 28(4):409–428, 1996.

Thomas G Kreutz and Joan M Ogden. Assessment of hydrogen-fueled proton exchange membrane fuel cells for distributed generation and cogeneration. In *proceedings of the 2000 US DOE hydrogen program review*, pages 1–43, 2000.

Max Kuhn and Kjell Johnson. *Feature engineering and selection: A practical approach for predictive models*. Chapman and Hall/CRC, 2019.

Rajesh Kumar. A guide to the dbscan clustering algorithm. `https://www.datacamp.com/tutorial/dbscan-clustering-algorithm`, September 2024. Accessed: 2025-11-02.

I Lawrence and Kuei Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, pages 255–268, 1989.

Adithya Legala, Jian Zhao, and Xianguo Li. Machine learning modeling for proton exchange membrane fuel cell performance. *Energy and AI*, 10:100183, 2022.

Adithya Legala, Samaneh Shahgaldi, and Xianguo Li. Data-based modelling of proton exchange membrane fuel cell performance and degradation dynamics. *Energy Conversion and Management*, 296:117668, 2023.

David R Legates and Gregory J McCabe Jr. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. *Water resources research*, 35(1):233–241, 1999.

Bowen Lei, Dongkuan Xu, Ruqi Zhang, and Bani Mallick. Embracing unknown step by step: Towards reliable sparse training in real world. *arXiv preprint arXiv:2403.20047*, 2024.

Zhongliang Li, Zhixue Zheng, Liangfei Xu, and Xiaonan Lu. A review of the applications of fuel cells in microgrids: opportunities and challenges. *BMC energy*, 1(1):8, 2019.

Zihao Li and Liumei Zhang. An ensemble outlier detection method based on information entropy-weighted subspaces for high-dimensional data. *Entropy*, 25(8):1185, 2023.

MathWorks. Neural net fitting, s. f. URL `https://www.mathworks.com/help/deeplearning/ref/neuralnetfitting-app.html`. Deep Learning Toolbox, MATLAB.

Ministerio de Energía. Gobierno presenta la estrategia nacional para que chile sea líder mundial en hidrógeno verde. `https://www.gob.cl/noticias/gobierno-presenta-la-estrategia-nacional-para-que-chile-sea-lider-mundial-en-hidrogeno-verde/`, 2020. Comunicado de prensa.

Tom Mitchell. Machine learning. *Publisher: McGraw Hill*, page 31, 1997.

James Montgomery. The no free lunch theorems for optimisation: An overview. *The No Free Lunch Theorems, Evolution and Evolutionary Algorithms*, 2002.

Andreas C. Müller. Data splitting strategies. `https://amueller.github.io/aml/04-model-evaluation/1-data-splitting-strategies.html`, 2020. Accessed: 2025-12-02.

J Eamonn Nash and Jonh V Sutcliffe. River flow forecasting through conceptual models part i—a discussion of principles. *Journal of hydrology*, 10(3):282–290, 1970.

Wenxu Niu, Xiaokang Li, Haobin Tian, and Caiping Liang. Remaining useful life prediction of pemfc based on 2-layer bidirectional lstm network. *World Electric Vehicle Journal*, 16(9):511, 2025.

Suphakit Niwattanakul, Jatsada Singthongchai, Ekkachai Naenudorn, and Supachanun Wanapu. Using of jaccard coefficient for keywords similarity. In *Proceedings of the international multiconference of engineers and computer scientists*, volume 1, pages 380–384, 2013.

Ryan O'hayre, Suk-Won Cha, Whitney Colella, and Fritz B Prinz. *Fuel cell fundamentals*. John Wiley & Sons, 2016.

Dilek Nur Ozen, Bora Timurkutluk, and Kemal Altinisik. Effects of operation temperature and reactant gas humidity levels on performance of pem fuel cells. *Renewable and Sustainable Energy Reviews*, 59:1298–1306, 2016.

Plug Power Inc. *GenSure E-Series Fuel Cell System: Operator's Manual*, May 2018. Document No. 640-111960, Rev. 12.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31, 2018.

Jiahang Qin, Yongping Hou, and Liying Ma. Research on automatic removal of outliers in fuel cell test data and fitting method of polarization curve. Technical report, SAE Technical Paper, 2024.

Shilin Qiu, Qihe Liu, Shijie Zhou, and Wen Huang. Adversarial attack and defense technologies in natural language processing: A survey. *Neurocomputing*, 492:278–307, 2022.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

Ibrahim Saleh, Rashid Ali, and Hongwei Zhang. Environmental impact of high altitudes on the operation of pem fuel cell based uas. *Journal of Energy and Power Engineering*, 10(3):87–105, 2018.

Erich Schubert, Jörg Sander, Martin Ester, Hans Peter Kriegel, and Xiaowei Xu. Dbscan revisited, revisited: why and how you should (still) use dbscan. *ACM Transactions on Database Systems (TODS)*, 42(3):1–21, 2017.

Neil C Schwertman, Margaret Ann Owens, and Robiah Adnan. A simple more general boxplot method for identifying outliers. *Computational statistics & data analysis*, 47(1):165–174, 2004.

scikit-learn developers. GradientBoostingRegressor — scikit-learn 1.8.0 documentation. `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html`, 2025a. Accessed: 2025-09-08.

scikit-learn developers. Demonstration of k-means assumptions, 2025b. URL `https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_assumptions.html`. scikit-learn 1.7.2 documentation.

scikit-learn developers. MLPRegressor — scikit-learn 1.8.0 documentation. `https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html`, 2025c. Accessed: 2025-10-08.

scikit-learn developers. Neural network models (supervised) — scikit-learn 1.8.0 documentation. https://scikit-learn.org/stable/modules/neural_networks_supervised.html, 2025d. Accessed: 2025-10-08.

scikit-learn developers. RandomForestRegressor — scikit-learn 1.8.0 documentation. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html?utm_source=chatgpt.com, 2025e. Accessed: 2025-09-12.

Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P Adams, and Nando De Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1): 148–175, 2015.

Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Amit Sharma. How dbscan clustering works: A comprehensive guide with implementations in python. https://www.analyticsvidhya.com/blog/2020/09/how-dbscan-clustering-works/, September 2020. Accessed: 2025-10-06.

Priynka Sharma, Maurizio Cirrincione, Ali Mohammadi, Giansalvo Cirrincione, and Rahul R Kumar. An overview of artificial intelligence-based techniques for pemfc system diagnosis. *IEEE Access*, 2024.

Alex J Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004.

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, 25, 2012.

Thomas E Springer, TA Zawodzinski, and Shimshon Gottesfeld. Polymer electrolyte fuel cell model. *Journal of the electrochemical society*, 138(8):2334, 1991.

Iain Staffell, Daniel Scamman, Anthony Velazquez Abad, Paul Balcombe, Paul E Dodds, Paul Ekins, Nilay Shah, and Kate R Ward. The role of hydrogen and fuel cells in the global energy system. *Energy & Environmental Science*, 12(2):463–491, 2019.

Danqi Su, Jiayang Zheng, Junjie Ma, Zizhe Dong, Zhangjie Chen, and Yanzhou Qin. Application of machine learning in fuel cell research. *Energies*, 16(11):4390, 2023.

Karl E Taylor. Summarizing multiple aspects of model performance in a single diagram. *Journal of geophysical research: atmospheres*, 106(D7):7183–7192, 2001.

Topographic-Map.com. Mapa topográfico región de antofagasta, altitud, relieve. https://es-cl.topographic-map.com/map-4tvmt/Regi%C3%B3n-de-Antofagasta/, s. f. Mapa topográfico interactivo.

John W Tukey. Exploratory data analysis. *Reading/Addison-Wesley*, 1977.

Rodrigo Vásquez and Felipe Salinas. *Tecnologías del hidrógeno y perspectivas para Chile*. Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH, Santiago, Chile, 2018. URL https://4echile.cl/wp-content/uploads/2020/07/LIBRO-TECNOLOGIAS-H2-Y-PERSPECTIVAS-CHILE.pdf. En colaboración con el Ministerio de Energía de Chile.

Alexandre MJ-C Wadoux, Dennis JJ Walvoort, and Dick J Brus. An integrated approach for the evaluation of quantitative soil maps through taylor and solar diagrams. *Geoderma*, 405:115332, 2022.

Yun Wang, Ken S Chen, Jeffrey Mishler, Sung Chan Cho, and Xavier Cordobes Adroher. A review of polymer electrolyte membrane fuel cells: Technology, applications, and needs on fundamental research. *Applied energy*, 88(4):981–1007, 2011.

Cort J Willmott and Kenji Matsuura. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate research*, 30(1):79–82, 2005.

David H Wolpert. The supervised learning no-free-lunch theorems. *Soft computing and industry: Recent applications*, pages 25–42, 2002.

David H Wolpert and William G Macready. No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, 1(1):67–82, 1997.

Horng-Wen Wu. A review of recent development: Transport and performance modeling of pem fuel cells. *Applied energy*, 165:81–106, 2016.

Jinrong Yang, Yichun Wu, and Xingyang Liu. Proton exchange membrane fuel cell power prediction based on ridge regression and convolutional neural network data-driven model. *Sustainability*, 15 (14):11010, 2023.

Xinjie Yuan, Fujun Chen, Zenggang Xia, Linlin Zhuang, Kui Jiao, Zhijun Peng, Bowen Wang, Richard Bucknall, Konrad Yearwood, and Zhongjun Hou. A novel feature susceptibility approach for a pemfc control system based on an improved xgboost-boruta algorithm. *Energy and AI*, 12: 100229, 2023.

Meiling Yue, Zeina Al Masry, Samir Jemei, and Noureddine Zerhouni. An online prognostics-based health management strategy for fuel cell hybrid electric vehicles. *International Journal of Hydrogen Energy*, 46(24):13206–13218, 2021.

Jaydev Chetan Zaveri, Shankar Raman Dhanushkodi, C Ramesh Kumar, Jan Taler, Marek Majdak, and Bohdan Weglowski. Predicting the performance of pem fuel cells by determining dehydration or flooding in the cell using machine learning models. *Energies*, 16(19):6968, 2023.

Zehui Zhang, Tianhang Dong, Xiaobin Xu, Weiwei Huo, Bin Zuo, and Leiqi Zhang. Multi-step performance degradation prediction method for proton-exchange membrane fuel cell stack using 1d convolution layer and catboost. *International Journal of Adaptive Control and Signal Processing*, 39(7):1434–1450, 2025.

Bowen Zhao, Huanlai Xing, Xinhan Wang, Fuhong Song, and Zhiwen Xiao. Rethinking attention mechanism in time series classification. *Information Sciences*, 627:97–114, 2023.

Jian Zhao, Xianguo Li, Chris Shum, and John McPhee. A review of physics-based and data-driven models for real-time control of polymer electrolyte membrane fuel cells. *Energy and AI*, 6:100114, 2021.

Alice Zheng and Amanda Casari. *Feature engineering for machine learning: principles and techniques for data scientists*. " O'Reilly Media, Inc.", 2018.

Zhi-Dan Zhong, Xin-Jian Zhu, and Guang-Yi Cao. Modeling a pemfc by a support vector machine. *Journal of Power Sources*, 160(1):293–298, 2006.

Xichuan Zhou, Haijun Liu, Cong Shi, and Ji Liu. Chapter 2 — the basics of deep learning, 2022.

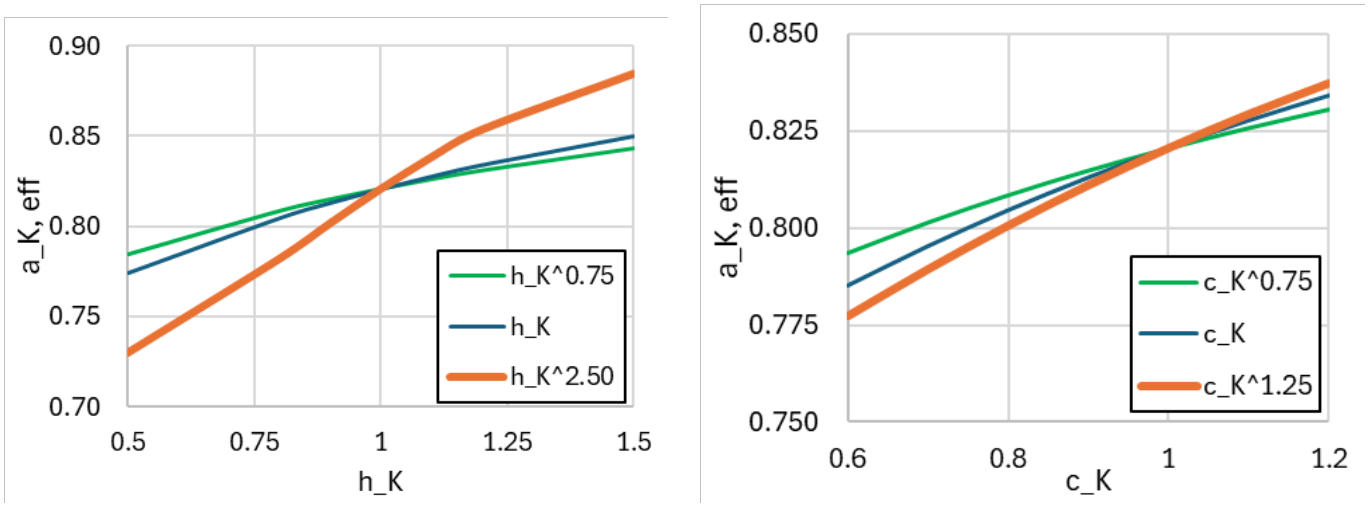
Miao Zou, Wu-Gui Jiang, Qing-Hua Qin, Yu-Cheng Liu, and Mao-Lin Li. Optimized xgboost model with small dataset for predicting relative density of ti-6al-4v parts manufactured by selective laser melting. *Materials*, 15(15):5298, 2022.

Apéndice A

Diseño de la metodología FASEK5: Compendio de gráficas y tablas

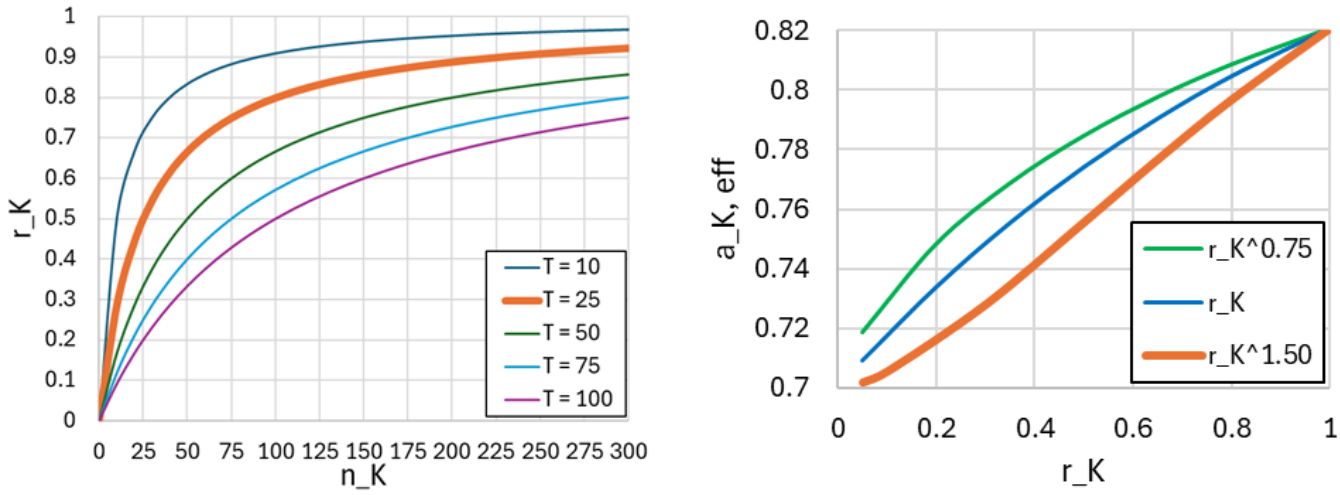
A.1 Calibración de exponentes de $a_{K,eff}$

Las Figuras A.1–A.2 muestran el análisis de sensibilidad efectuado para determinar los exponentes de los parámetros h_K, c_K y r_K sobre $a_{K,eff}$. De forma complementaria, en la Figura A.2(a) se presenta el estudio preliminar para definir el valor de T que permite ajustar el comportamiento asintótico deseado de n_K sobre r_K. Las curvas resaltadas en color naranja corresponden a las selecciones de exponentes utilizadas para el cálculo de w_K y a la selección del parámetro T. En líneas generales, la calibración se realizó para producir variaciones más marcadas dentro del rango efectivo de cada parámetro analizado.



(a) (b)

Figura A.1: Análisis paramétrico de sensibilidad de los exponentes sobre $a_{K,eff}$: (a) h_K; (b) c_K.



(a) (b)

Figura A.2: Análisis de sensibilidad: (a) efecto de T sobre r_K; (b) exponente de r_K sobre $a_{K,eff}$.

A.2 Parámetros calibrados de DBSCAN por familia (F2, F3 y F4)

Las Tablas A.1–A.3 presentan los valores de los parámetros `eps` (`MinPts`) y `k` seleccionados durante las calibraciones de las clusterizaciones mediante DBSCAN, además del porcentaje de ruido o contaminación detectada por el algoritmo para el grupo analizado (Config, Global, Sitio, Session_ID, según corresponda).

Tabla A.1: Calibración DBSCAN - Detección Familia 2: *outliers* de reparto (K_act × Config).

Config	Estrato	`eps`	`k`	Ruido (%)
011	Global	0.400	5	12.5
	PSDA	0.400	5	21.4
	SanPedro	0.400	5	0.70
101	Global	0.500	5	28.5
110	Global	0.550	5	4.00
	Tocopilla	0.400	5	10.9
	Calama	0.800	5	13.2
	SanPedro	0.500	5	12.2
	Chacabuco	0.600	5	21.4
	PSDA	0.700	5	10.7
110	P1 (PSDA)	0.600	5	14.6
	P2 (PSDA)	0.600	5	18.4
	P3 (PSDA)	0.700	5	25.0

Tabla A.2: Calibración DBSCAN - Detección Familia 3: *outliers* bivariados (I, V).

Conjunto	`eps`	`k`	Ruido (%)
Global	0.500	7	3.45
Tocopilla	0.600	6	12.5
Calama	0.500	9	8.87
SanPedro	0.600	9	7.08
Chacabuco	0.600	7	5.13
PSDA	0.500	7	6.96

Tabla A.3: Calibración DBSCAN - Detección Familia 4: *outliers* multidimensionales (PCA).

Config	Estrato	`eps`	`k`	Ruido (%)
011	Global	0.400	5	12.5
	PSDA	0.400	5	21.4
	SanPedro	0.400	5	0.70
101	Global	0.500	5	28.5
110	Global	0.550	5	4.00
	Tocopilla	0.400	5	10.9
	Calama	0.800	5	13.2
	SanPedro	0.500	5	12.2
	Chacabuco	0.600	5	21.4
	PSDA	0.700	5	10.7
110	P1 (PSDA)	0.600	5	14.6
	P2 (PSDA)	0.600	5	18.4
	P3 (PSDA)	0.700	5	25.0

A.3 Calibración FASEK5: Parámetros $p_{m|K}$ seleccionados por método en cada familia de detección (F1-F5)

La Tabla A.4 resume los valores de $p_{m|K}$ de los 37 métodos utilizados para la agregación interna (*leaky noisy-OR*) de la metodología FASEK5. Además, se incluyen los nombres de los detectores tipo *flags*.

Tabla A.4: Valores utilizados de $p_{m|K}$: *leaky noisy-OR* interno.

Identificador de columna (detectores)	$p_{m\|K}$
F1_V_Modo	0.65
F1_V_Sitio_Modo	0.63
F1_I_Session	0.58
F2_K1_I_Sitio	0.57
F2_K1_I_Session	0.62
F2_K2_011_FC_ID_Session	0.63
F2_K2_011_PCA_Global	0.65
F2_K2_011_DBSCAN_General	0.74
F2_K2_011_DBSCAN_Session	0.72
F2_K2_101_FC_ID_Global	0.63
F2_K2_101_PCA_Global	0.65
F2_K2_101_DBSCAN_General	0.74
F2_K2_110_FC_ID_Sitio	0.63
F2_K2_110_FC_ID_Session(PSDA)	0.64
F2_K2_110_PCA_Global	0.65
F2_K2_110_DBSCAN_General	0.74
F2_K2_110_DBSCAN_Sitio	0.73
F2_K2_110_DBSCAN_Session(PSDA)	0.72
F2_K3_I-FC_ID_Global	0.63
F2_K3_I-FC_ID_Chacabuco	0.65
F2_K3_PCA_Global	0.63
F2_K3_PCA_Chacabuco	0.64
F2_K3_DBSCAN_Global	0.77
F2_K3_DBSCAN_Chacabuco	0.74
F3_2D_DBSCAN_Global	0.78
F3_2D_DBSCAN_Sitio	0.76
F4_PCA_DBSCAN_Global	0.75
F4_PCA_T2_Global	0.82
F4_PCA_SPE_Global	0.79
F4_PCA_T2+SPE_Global	0.90
F4_PCA_DBSCAN_Sitio	0.73
F4_PCA_T2_Sitio	0.83
F4_PCA_SPE_Sitio	0.80
F4_PCA_T2+SPE_Sitio	0.90
F5_W_Sitio_FCID	0.71
F5_W_DBSCAN_Sitio_FCID_refined	0.76
F5_W_Percentile_Sitio_FCID	0.70

A.4 Resultados por familia de detección: Metodología FASEK5

Las Tablas A.5–A.9 resumen las cantidades de detecciones en cada estrato de detección por familia con totales parciales por cantidad de detecciones (columnas) y por Sitio, Modo, Config y FC_ID (filas), según corresponda.

Tabla A.5: Cantidad de detecciones en F1 (Sitio × Modo).

Sitio × Modo	1	2	Total
Calama	2		2
Maintain	2		2
Chacabuco	8	2	10
Maintain	8	2	10
PSDA	12	1	13
Float	6		6
Maintain	6	1	7
SanPedro	10	3	13
Float	2	3	5
Maintain	8		8
Tocopilla	17	4	21
Float	8	1	9
Maintain	9	3	12
Total	49	10	59

Tabla A.6: Cantidad de detecciones en F2 (Sitio × Config).

Sitio × Config	1	2	3	4	5	6	Total
Calama	5	14					19
100	1						1
110	4	8					12
111		6					6
Chacabuco	27	11	11	6	2		57
100	9	1					10
101		1	3				4
110	17	6	1				24
111	1	3	7	6	2		19
PSDA	35	22	22	10	3	4	96
011	9	7	11	1			28
100	1						1
101		1	1				2
110	25	11	10	9	3	4	62
111		3					3
SanPedro	29	20	8	4			61
011	3	5	2				10
100	4	5					9
101	3	3					6
110	13	4	6	4			27
111	6	3					9
Tocopilla	22	15	5	1			43
100	9	12					21
110	13	3	5	1			22
Total	118	82	46	21	5	4	276

Tabla A.7: Cantidad de detección en F3 (Sitio).

Sitio	1	2	Total
Calama	10	9	19
Chacabuco	10	8	18
PSDA	21	7	28
SanPedro	15	11	26
Tocopilla	35	15	50
Total	91	50	141

Tabla A.8: Cantidad de detección en F4 (Sitio).

Sitio	1	2	3	4	5	Total
Calama	24	1	2		1	28
Chacabuco	34	5	2			41
PSDA	17	1		6		24
SanPedro	21	2	2			25
Tocopilla	24	3	2			29
Total	120	12	8	6	1	147

Tabla A.9: Cantidad de detecciones en F5 (Sitio × FC_ID).

Sitio × FC_ID	1	2	3	Total
Calama	12	4	1	17
FC_1	8	1		9
FC_2	4	3	1	8
Chacabuco	26	8		34
FC_1	13	4		17
FC_2	11	2		13
FC_3	2	2		4
PSDA	36	19	3	58
FC_1	14	7	2	23
FC_2	13	10	1	24
FC_3	9	2		11
SanPedro	16	11	1	28
FC_1	4	4		8
FC_2	8	4		12
FC_3	4	3	1	8
Tocopilla	18	11	6	35
FC_1	9	10	6	25
FC_2	9	1		10
Total	108	53	11	172

A.5 Tablas auxiliares – Implementación metodología FASEK5

La Tabla A.10 resume los valores utilizados para los parámetros c_K, h_K (post-entrenamiento eliminando 3% de ruido de la BD, detalle no se muestra por efectos de extensión del trabajo) y l_K, así como las cotas $a_{K,min}$ y $a_{K,max}$ del parámetro $a_{K,eff}$ de cada familia, incluidas las variantes F2-K1, F2-K2 y F2-K3.

Tabla A.10: Parámetros de las familias de detección.

Familia (K)	c_K	h_K	l_K	$a_{K,min}$	$a_{K,max}$
F1	0.70	1.345	0.050	0.40	0.70
F2-K1	0.60	1.475	0.040	0.30	0.60
F2-K2	0.80	1.475	0.040	0.50	0.80
F2-K3	0.90	1.475	0.040	0.60	0.90
F3	1.00	1.185	0.020	0.70	1.00
F4	1.10	0.949	0.015	0.80	1.10
F5	1.20	1.508	0.010	0.90	1.20

La Tabla A.11 incluye los valores de los parámetros T, L_0, λ y γ para completar la implementación de la metodología de detección probabilística de *outliers* FASEK5 de este trabajo.

Tabla A.11: Parámetros auxiliares.

Parámetro	Valor
T	25
L_0	0.0075
λ	0.3
γ	ln(2)

A.6 Resultados integrados de probabilidad global de *outlier* (S_i) - FASEK5

La Tabla A.12 resume los resultados de la implementación global de FASEK5, en la cual se muestra la cantidad de detecciones (o registros) por k métodos con *flag* activo y la probabilidad media S asociada. Se incluyen métricas agregadas como % total de la BD y cantidades acumuladas absoluta y porcentual.

Tabla A.12: Resultados de la implementación de la metodología FASEK5.

k (métodos activos)	Cantidad de detecciones	% total (cont. BD)	Cantidad de removidos ($\geq$ k)	% removido con umbral $\geq$ k	Probabilidad media (S)
1	150	9.40	421	26.4	0.68161
2	75	4.70	271	17.0	0.86616
3	59	3.70	196	12.3	0.94901
4	50	3.13	137	8.59	0.98410
5	29	1.82	87	5.45	0.99430
6	18	1.13	58	3.64	0.99779
7	21	1.32	40	2.51	0.99932
> 8	19	1.20	19	1.20	0.99989