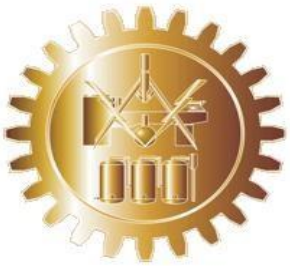
Instituto Tecnológico de Aguascalientes



“Implementación de ciencia de datos y aprendizaje automático para la formulación y validación de modelos predictivos aplicados a propiedades termodinámicas de sistemas electroliticos ”

T E S I S

Para Obtener el Grado de: Maestro en Ciencias en Ingeniería Química

PRESENTA:

I.Q. Mateo Agudelo Jaramillo

Asesor

Dr. José Enrique Jaime Leal

Aguascalientes, Ags., Agosto del 2023

Instituto Tecnológico

de Aguascalientes

Tesis

Implementación de ciencia de datos y aprendizaje automático para la formulación y validación de modelos predictivos aplicados a propiedades termodinámicas de sistemas electroliticos

Presenta

I.Q. Mateo Agudelo Jaramillo

Asesor

Dr. José Enrique Jaime Leal

Sinodales

Dra. Norma Aurea Rangel Vasquez

Dr. Alejandro Meza De Luna

Aguascalientes, Ags., Agosto del 2023

Aguascalientes, Ags, 02/agosto/2023
Oficio DEPI No. 0432/2023

Asunto: Autorización de tesis

C. MATEO AGUDELO JARAMILLO
PRESENTE

Los abajo firmantes, Miembros del Jurado para Examen de Grado de Maestría, hacemos CONSTAR que, habiendo revisado el Trabajo de Tesis desarrollado por Usted, bajo el título: **"Implementación de ciencia de datos y aprendizaje automático para la formulación y validación de modelos predictivos aplicados a propiedades termodinámicas de sistemas electrolíticos"**, hemos dictaminado que éste es de aceptarse, por lo que se autoriza su impresión.

ATENTAMENTE
Excelencia en Educación Tecnológica®
Ingenio, Cultura y Saber, que conducen a la Excelencia

DR. JOSÉ ENRIQUE JAIME LEAL

DRA. NORMA AUREA RANGEL VÁZQUEZ

DR. ALEJANDRO MEZA DE LUNA

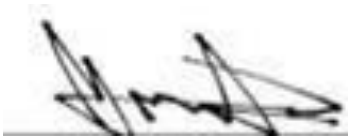
EDUCACIÓN | TECNOLÓGICO NACIONAL DE MÉXICO
INSTITUTO TECNOLÓGICO DE AGUASCALIENTES
DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN

c.c.p. Archivo
JEJL/yevo

CARTA CESION DE DERECHOS

En la Ciudad de Aguascalientes, el que suscribe Mateo Agudelo Jaramillo estudiante de la Maestría en Ciencias en Ingeniería Química, adscrito al Tecnológico Nacional de México-Instituto Tecnológico de Aguascalientes, manifiesta que es autor (a) intelectual del presente trabajo de Tesis bajo la dirección de José Jaime Enrique leal y cede los derechos del trabajo "Implementación de ciencia de datos y aprendizaje automático para la formulación y validación de modelos predictivos aplicados a propiedades termodinámicas de sistemas electrolíticos", al Tecnológico Nacional de México – Instituto Tecnológico de Aguascalientes para su difusión, con fines académicos y de investigación.

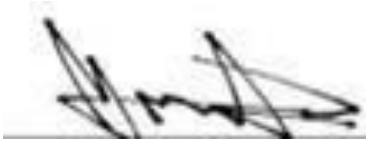
Los usuarios de la información no deben reproducir el contenido textual, gráficas o datos del trabajo sin el permiso expreso del autor y/o director del trabajo. Si el permiso se otorga, el usuario deberá dar el agradecimiento correspondiente y citar la fuente del mismo.



Mateo Agudelo Jaramillo

DECLARACIÓN DE AUTENTICIDAD Y DE NO PLAGIO

En la Ciudad de Aguascalientes, el que suscribe Mateo Agudelo Jaramillo estudiante de la Maestría en Ciencias en Ingeniería Química, adscrito al Tecnológico Nacional de México-Instituto Tecnológico de Aguascalientes, manifiesta que se responsabiliza de la autenticidad y originalidad del presente trabajo "Implementación de ciencia de datos y aprendizaje automático para la formulación y validación de modelos predictivos aplicados a propiedades termodinámicas de sistemas electrolíticos", el cual ha sido presentado para la obtención del grado correspondiente.



Mateo Agudelo Jaramillo

Agradecimientos

A Dios por permitirme concluir una etapa más.

A mis padres por estar siempre conmigo y soportarme todo este tiempo. A ellos se los agradezco principalmente

A mis hermanos que han sido la base para seguir motivado y querer continuar.

A mi Asesor Dr. José Enrique Jaime Leal por su apoyo y paciencia en la realización de este proyecto.

A mis amigos que me apoyaron verdaderamente durante todo este tiempo.

Al Consejo Nacional de Ciencia y Tecnología (Conacyt) por brindarme el apoyo economico para poder culminar este proyecto.

Índice

Lista de Figuras.

Pag.

Lista de Tablas

Pag.

Capítulo 1. Introducción.

1.1 Las soluciones electrolíticas y sus propiedades termodinámicas.

Las soluciones electrolíticas cumplen diferentes aplicaciones en diversas áreas de la ciencia e ingeniería. Por sus características y propiedades, estos sistemas son de amplia aplicación industrial, por ejemplo, la extracción comercial de aluminio, refinación de cobre, procesos de galvanoplastia y producción de fertilizantes. Asimismo, un área donde este tipo de sistemas ha tenido una amplia aplicación corresponde a los procesos de separación en la industria petroquímica, ya que debido a sus propiedades físico-químicas estas soluciones pueden funcionar como solventes, catalizadores, y permiten desplazar puntos de azeotropía, entre otros fenómenos que favorecen a los procesos donde se incorporan.

Dado lo anterior, y de acuerdo a Belvéze *et al*. (2004) es necesaria una adecuada identificación de las propiedades termodinámicas de este tipo de sistemas, ya que esto permitirá una caracterización de las mismas con fines de modelación, diseño, optimización y control de un proceso en los que se desee incorporarlas. Esta caracterización de las propiedades termodinámicas de los sistemas electrolíticos parte de propiedades tales como la densidad, la viscosidad, el coeficiente de actividad ó fugacidad, entre otras y donde es necesario proponer y desarrollar modelos matemáticos que permitan una adecuada predicción de estas propiedades a partir de datos cuantificables del sistema que se desee estudiar (Karimzadeh & Hosseini, 2019).

Acorde a la literatura, los modelos que se han desarrollado hasta el momento, con fines de predicción de propiedades termodinámicas en sistemas electrolíticos, se fundamentan basados en ecuaciones de estado o ecuaciones termodinámicas, y en donde estos modelos involucran diferente número de parámetros. La mayoría de estos modelos existentes se clasifican como Modelos de Composición Local, y los cuales se desarrollaron considerando las interacciones producto de las fuerzas de corto y largo alcance entre iones y moléculas del sistema. Este tipo de modelos se han empleado en mayor medida por su fácil aplicación ya que requieren un manejo menos complicación en comparación con otro tipo de modelos tales como las ecuaciones de estado. Existen investigaciones más recientes en las que se asumen nuevas interacciones para mejorar modelos existentes, como el caso del modelo de Pitzer modificado por Lach *et al*. (2018). Así, el modelado termodinámico de este tipo de sistemas ha sido estudiado bastante tiempo, y sigue siendo un tema interesante debido a la alta desviación inherente de la idealidad.

Si bien, y de acuerdo a los estudios reportados en la literatura, este tipo de modelos ha presentado resultados aceptables para la mayoría de los sistemas en los que se han implementado, existen casos donde la capacidad del modelo queda limitada y esto puede obedecer a las siguientes razones:

i) Debido a la complejidad y no idealidad de los sistemas electrolíticos estos tienden a ser altamente no lineales lo que limita el rango de operación de los actuales modelos que buscan predecir sus propiedades, además que puede darse el caso de que el modelo no pueda operar con algunos tipos de sistemas complejos.

ii) La capacidad de predicción de todo modelo termodinámico está en función de un conjunto de parámetros de ajuste característicos para cada sistema que se estudie, y donde estos son generados a partir de un proceso de correlación de datos experimentales, por lo que, si la cantidad de estos es limitada o nula, de la misma forma será la capacidad predictiva del modelo utilizado.

iii) La estructura propia del modelo, es decir los elementos que constituyen al modelo al momento de su diseño pueden robustecer o no su capacidad de modelación y esto se basa en las teorías, fundamentos físicos y químicos y simplificaciones adoptadas para su construcción.

Por lo anterior, es imposible establecer un modelo o conjunto de estos capaz de modelar cierta propiedad para cualquier tipo de sistema de solución electrolítica y esto ocurre en todas las áreas de la ciencia, donde el proceso de generación de modelos capaces de abordar cualquier sistema o proceso sigue siendo motivo de investigación. Actualmente, las investigaciones enfocadas a la generación de modelos se basan en nuevas propuestas o bien, en la modificación de modelos actuales con el fin de robustecer a estos.

Por otro lado, con el desarrollo de la ciencia computacional se ha desarrollado una diversidad de estrategias y metodologías que buscan resolver problemas donde las teorías y metodologías actuales quedan limitadas. Conforme a esto, tal es el caso del área de la Inteligencia Artificial en la cual se presentan diversas herramientas y/o estrategias que pueden ayudar a resolver el problema de la modelación de datos en sistemas complejos como lo son la predicción de propiedades termodinámicas en sistemas de soluciones electrolíticas y donde el siguiente apartado da una breve descripción de estas herramientas computacionales.

1.2 Aplicación de la Inteligencia Artificial en problemas de Ingeniería Química.

Dentro del área de la Inteligencia Artificial existe una amplia diversidad de campos donde ésta es aplicable como lo es la Robótica, la Computación paralela, Procesamiento de lenguaje e imágenes, Identificación de patrones, Aprendizaje profundo y Aprendizaje automático, por citar los más significativos, y todos ellos buscan resolver problemas donde los métodos o herramientas actuales están limitados.

Particularmente, en el campo del aprendizaje automático se aplica la ciencia de datos, la cual consiste en generar conocimiento o resolver un problema a partir de una base de datos inherente y asociada al sistema o proceso que se estudia. Así, en el aprendizaje automático los datos existentes son utilizados dentro de una estrategia numérica perteneciente a este campo, y los cuales son tratados conforme a la estructura del algoritmo para resolver un problema en particular. Entre los algoritmos existentes y pertenecientes al campo del lenguaje automático están las Redes Neuronales, Arboles de Decisión, Inferencia Bayesiana, entre otros.

En particular, la aplicación de la ciencia de datos y el aprendizaje automático es diverso, cubriendo áreas como la medicina, la biología, la economía, las finanzas, las ciencias políticas y sociales, la Ingeniería Química, entre otras; Y su aplicación en estas áreas se ha centralizado en la modelación y predicción de datos y propiedades con fines de diseño, determinación de patrones y toma de decisiones. Particularmente, en el área de la Ingeniería Química su aplicación se ha centrado en la modelación de propiedades termodinámicas de componentes y mezclas, de parámetros de sistemas reactivos, de procesos de transferencia de masa-energía, entre otros.

Con base a lo anterior, la modelación y/o predicción de propiedades termodinámicas en soluciones electrolíticas es crucial con fines de diseño, operación, optimización y control de los sistemas o procesos donde este tipo de soluciones son aplicables debido a las ventajas inherentes que sus propiedades físico-químicas aportan sobre un sistema en particular. En este contexto y conforme a lo referido previamente, la ciencia de datos y el aprendizaje automático pueden proporcionar herramientas poderosas para el modelado y la predicción de estas propiedades.

Por citar algunos estudios donde se han aplicado alguna de las herramientas del aprendizaje automático para resolver casos específicos está el trabajo de Palmer *et al*. (2007) quienes desarrollaron modelos de *QSPR* para predecir la solubilidad acuosa de moléculas orgánicas a través de varios modelos de regresión como bosque aleatorio y redes neuronales artificiales, y el cual se destacó por su precisión en la predicción de la solubilidad acuosa. Por su parte, Stenzel *et al*. (2017) aplicaron redes neuronales y árboles aleatorios para relacionar datos de microestructuras en 3D para predecir conductividades efectivas de cualquier material.

En otra investigación relevante, Asensio-Delgado *et al*. (2022) aplicaron redes neuronales artificiales para predecir la solubilidad de los gases F en líquidos iónicos (LI´s) a partir de las propiedades fácilmente accesibles de los compuestos puros. Por su parte Azadfar *et al*. (2022), desarrollaron un nuevo modelo basado en redes neuronales artificiales (*ANN*) para estimar la capacidad calorífica de líquidos iónicos puros.

Si bien a la fecha existen ya algunos estudios que han incorporado alguna estrategia del aprendizaje automático y la ciencia de datos aplicados a los sistemas electrolíticos en sistemas acuosos, estos solo han abordado un conjunto limitado de sistemas aplicándolos específicamente para predecir una determinada propiedad, además de que a la fecha no se ha localizado algún trabajo que considere generar un modelo hibrido que considere a alguno de los modelos existentes que ha dado buenos resultados e incorpore alguna estrategia de aprendizaje automático con el fin de robustecer y mejorar la capacidad predictiva del modelo.

De esta manera, este estudio pretende aprovechar estas áreas de oportunidad enfocado en la predicción de propiedades termodinámicas en sistemas electrolíticos en solución acuosa.

1.3 Objetivos.

1.3.1 Objetivo General.

Desarrollar una metodología que permita el acoplamiento de una estrategia de aprendizaje automático junto con un modelo de composición local para generar un modelo hibrido para la modelación y/o predicción de propiedades termodinámicas en sistemas electrolíticos en solución.

1.3.2 Objetivos específicos

- Aplicar las bases teóricas de la ciencia de datos para la ubicación, extracción, depuración y estandarización de los datos concernientes a sistemas electrolíticos en solución, y que fungirán como entradas y salidas a las herramientas de aprendizaje automático.
- Desarrollar la programación de un conjunto de estrategias o modelos de aprendizaje automático que serán evaluados y comparadas sus capacidades de modelación de datos, así como establecer las métricas estadísticas con que se realizará dicha evaluación cuantitativa.
- Evaluar el desempeño de los modelos de aprendizaje automático programados incorporando los datos de entrada-salida depurados para evaluar la capacidad predictiva de propiedades termodinámicas de sistemas electrolíticos en solución, así como determinar los estadísticos correspondientes producto del proceso de correlación.
- Evaluar el efecto e impacto que tiene alguna de las propiedades de los sistemas electrolíticos, sobre la capacidad predictiva de uno de los modelos de aprendizaje automático en la predicción de propiedades termodinámicas en estos sistemas.

- Acoplar un modelo de composición local y un modelo de aprendizaje automático para generar un modelo hibrido y evaluar su capacidad predictiva sobre un conjunto de sistemas de electrolíticos en solución.

1.4 Hipótesis.

El desarrollo de una metodología que permita implementar y evaluar distintos modelos de aprendizaje automático, permitirá seleccionar aquel modelo que presenta mejor capacidad predictiva para la modelación de propiedades termodinámicas de sistemas electrolíticos en solución. A su vez, la generación de un modelo hibrido presentará un comportamiento más robusto en comparación al modelo original para la predicción de parámetros termodinámicos en este tipo de sistemas electrolíticos en solución.

1.5 Justificación

Una etapa importante de la ingeniería de procesos es el área de diseño, la cual busca estructurar y modelar matemáticamente la dinámica de todo un sistema o proceso para conocer, a priori, su comportamiento y los fenómenos presentes previo a su construcción, operación y control; Y donde para ello, es necesario conocer los valores de las variables operativas y/o parámetros que intervienen en las ecuaciones gobernantes del sistema.

Dependiendo del tipo de proceso o sistema que se desee diseñar, algunas de estas variables o parámetros deben obtenerse previo cálculo a través de algún modelo matemático y donde, de acuerdo a su valor calculado, podrá impactar favorablemente o negativamente al modelo del sistema en el cual se incorpore. Por ende, la confiabilidad o capacidad predictiva del modelo que se utilice jugará un valor preponderante en la calidad de los resultados finales.

Por ejemplo, los procesos del área de Ingeniería Química en su gran mayoría, requieren parámetros termodinámicos que son modelados o calculados a través de modelos termodinámicos o ecuaciones de estado. Tal es el caso de los sistemas electrolíticos los cuales debido a sus características físico-químicas tienen una amplia aplicabilidad en los procesos de corrosión, se emplean como solventes para determinadas mezclas y tienen un impacto considerable en procesos de separación ya que pueden desplazar puntos de azeotropía. Por ende, es importante contar con información confiable de sus parámetros termodinámicos como lo es el coeficiente de actividad y el coeficiente osmótico ya que con estos se puede determinar el potencial químico de un sistema o bien, medir la desviación del comportamiento de un disolvente respecto de su idealidad, respectivamente.

Por lo anterior, si bien a la fecha existen propuestas de modelos para predecir este tipo de parámetros termodinámicos, su confiabilidad sigue siendo limitada a cierto rango de operación, altamente dependiente de la existencia y calidad de los valores de los parámetros de interacción que intervienen en el modelo y, al tipo y complejidad del sistema electrolítico que se desee modelar. Por estas características, los actuales modelos siguen presentando restricciones en su operatividad.

Una alternativa que se ha presentado en fechas recientes es la aplicación de herramientas o estrategias de la Inteligencia Artificial, particularmente del área del aprendizaje automático junto con las metodologías de la ciencia de datos para desarrollar herramientas de modelación basadas en un conjunto de datos de entrada-salida. A la fecha, este tipo de estrategias enfocadas a esta finalidad se ha extendido en muchos campos y sus resultados han sido alentadores.

Por ende, este tipo de estrategias se presentan como una alternativa atractiva para ser aplicadas en este estudio, donde se propone generar una metodología que permita incorporar modelos de aprendizaje automático y generar un modelo hibrido para la modelación y/o predicción de propiedades termodinámicas de sistemas electrolíticos en solución. Cabe indicar que este sería el primer estudio en que se genera un modelo hibrido aplicado en la modelación de propiedades en este tipo de sistemas.

Capítulo 2. Marco Teórico.

En este capítulo se describe el fundamento teórico de las soluciones electrolíticas y sus interacciones, así como los modelos que representan a éstas y a sus principales propiedades termodinámicas. Asimismo, se presenta las bases teóricas de los modelos y métodos de aprendizaje automático que se implementaron con el objetivo de modelar las principales propiedades termodinámicas de este tipo de sistemas electrolíticos.

2.1 Definición de solución electrolítica.

Las soluciones electrolíticas son compuestos cuyos constituyentes (iones) están inmersos o disueltos en un solvente, y esto se da debido a la reacción de disociación que sucede cuando una sustancia de naturaleza iónica entra en contacto con un solvente de naturaleza polar (ej. agua). Este fenómeno cambia las propiedades del soluto-solvente y se genera un sistema que permite la fluidez de electrones debido al potencial eléctrico que forman los cationes y aniones. La Figura 2.1, presenta un esquema ilustrativo de este tipo de soluciones (Panagiotopoulos & Yue, 2023).

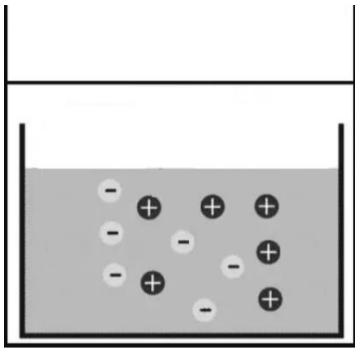


Figura 2.1 *Esquema general de una solución electrolítica.*

Tal y como se mencionó en la sección 1.1, los sistemas electrolíticos poseen una amplia gama de aplicaciones dentro de los procesos del área de la Ingeniería Química, particularmente en las operaciones de extracción y separación.[2] Asimismo, para una adecuada utilización de este tipo de sistemas es necesario contar con modelos robustos que sean capaces de predecir sus propiedades termodinámicas ya que al conocer su dinámica real pueden generarse y/o modelarse procesos de forma más adecuada y confiable. Entre las principales propiedades termodinámicas que suelen utilizarse para este tipo de sistemas están las que se describen a continuación.

2.2 Principales propiedades termodinámicas de los sistemas electrolíticos.

2.2.1 Coeficiente de actividad medio iónico (γ±).

El coeficiente de actividad medio iónico (γ±) en soluciones electrolíticas, corresponde a una medida de desviación del comportamiento de un electrólito en solución respecto a la ley de acción de masas. Es decir, corresponde a una ponderación de la corrección que se debe aplicar a la actividad iónica calculada a partir de la concentración de iones en solución para tener en cuenta las interacciones entre estos (Saravi & Panagiotopoulos, 2022).

En conformidad a la literatura, el coeficiente de actividad medio iónico se puede expresar matemáticamente mediante la Ecuación (2.1) y su cálculo para soluciones muy diluidas con base en la teoría de Debye-Hückel, se puede realizar con las Ecuaciones (2.2) y (2.3):

$$\gamma_{\pm} = \frac{\sum |z_i| c_i \gamma_i}{\sum |z_i| c_i} \quad \{2.1\}$$

$$\log(\gamma_{\pm}) = -A * \sqrt{I} \quad \{2.2\}$$

$$I = 0.5 * \Sigma\, ci * z_i^{\,2} \quad \{2.3\}$$

Donde $\gamma\pm$ es el coeficiente de actividad medio iónico, γ_i es el coeficiente de actividad del ion *i*, z_i es la carga del ion *i*, c_i es la concentración del ion *i*, *A* es el coeficiente de Debye-Hückel, que es una constante que depende de la temperatura y el solvente e *I* es la fuerza iónica de la solución.

Particularmente, el coeficiente de actividad medio iónico se emplea en química analítica y en electroquímica para calcular constantes de equilibrio en soluciones de alta concentración y para predecir el comportamiento de los iones en soluciones complejas. También es importante en la química de soluciones, donde se utiliza para calcular la actividad efectiva de los iones en soluciones concentradas (Attias *et al*., 2022).

Por otra parte, el coeficiente de actividad medio iónico también es importante en la determinación de constantes de equilibrio en soluciones electrolíticas, donde dicha constante de equilibrio es una medida de la fuerza de un ácido o una base en solución y se puede calcular a partir de la actividad iónica de los iones en la solución. Al tener en cuenta las desviaciones de la ley de acción de masas, se pueden obtener valores más precisos de las constantes de equilibrio en soluciones de alta concentración (Kuramochi *et al*., 2005).

Otra aplicación del coeficiente de actividad medio iónico es en la interpretación de mediciones electroquímicas en soluciones complejas de alta concentración, y donde éste se utiliza para corregir las mediciones y obtener valores precisos de las propiedades electroquímicas de la solución.

2.2.2 Coeficiente Osmótico (Φ).

El coeficiente osmótico de soluciones electrolíticas (Φ) corresponde a una medida de la desviación del comportamiento osmótico ideal de una solución debido a la presencia de iones disueltos. En una solución ideal, la presión osmótica es proporcional a la concentración total de solutos en la solución, sin embargo, en una solución electrolítica los iones disueltos pueden interactuar entre sí y con el solvente de una manera que afecta la presión osmótica de la solución. Así, el coeficiente osmótico se define como la relación entre la presión osmótica experimental y la presión osmótica calculada a partir de la concentración total de solutos en la solución. La expresión matemática que define a esta propiedad está dada por la Ecuación (2.4).

$$\Phi = \frac{\pi_{exp}}{\pi_{cal}} \qquad \{2.4\}$$

Donde π_{exp} es la presión osmótica experimental y π_{calc} es la presión osmótica calculada a partir de la concentración total de solutos en la solución. Para una solución ideal el coeficiente osmótico tendrá un valor de 1, mientras que un valor distinto indicará desviación del comportamiento ideal.

El coeficiente osmótico se utiliza en diversos procesos entre los que están la separación de iones mediante técnicas de electroforesis y cromatografía de intercambio iónico. También es importante en la comprensión del transporte de iones a través de membranas biológicas y en la formulación de soluciones electrolíticas para aplicaciones en electroquímica y electroterapia (Delač Marion *et al*., 2015).

Tal y como se mencionó anteriormente, estos parámetros termodinámicos forman parte de modelos más elaborados que definen alguna otra propiedad o comportamiento de un sistema o proceso y que son utilizados con fines de diseño en procesos industriales más complejos. En el siguiente apartado se aborda una explicación general de la complejidad de estos modelos y sus aplicaciones en sistemas donde intervienen las soluciones electrolíticas.

2.3 Modelación en sistemas electrolíticos y sus interacciones.

Desde la perspectiva de la termodinámica, es posible observar que los sistemas electrolíticos presentan una complejidad mayor y que va más allá de lo que la teoría clásica puede explicar, y esto se debe a que a nivel macroscópico se presentan fenómenos que no pueden ser comprendidos sin recurrir a la termodinámica estadística. En estos sistemas, se pueden encontrar tanto especies iónicas como moleculares,

lo que implica que existen tres tipos de interacciones definidas como ión-ión, molécula-molécula y molécula-ión (Renon, 1996).

La presencia de electrolitos fuertes que se disocian completamente en solución o de electrolitos débiles que lo hacen en menor proporción, genera cambios significativos en el comportamiento y las propiedades del sistema. Estos cambios tienen su origen en las fuerzas de Coulomb que son de largo alcance y que se producen entre los iones.

Como se mencionó, existen diferentes tipos de modelos que se utilizan con fines de predicciones o para correlacionar datos termodinámicos, pero los más comunes que se encuentran en la literatura son los modelos de ecuaciones de estado y los modelos termodinámicos. A continuación. se describen estos brevemente.

2.3.1 Modelos termodinámicos.

Dentro del cálculo de equilibrio de fase para las soluciones electrolíticas, una propiedad de interés que puede describir la capacidad de separación que se presenta en un sistema es el potencial químico. El cálculo del potencial químico por constituyente del sistema está representado por la Ecuación (2.5).

$$\mu_i = \mu_i^{ref}(T_0, P_0,) + RT\frac{lnf_i}{f_i^{ref}} \quad \{2.5\}$$

Donde T es la temperatura en kelvin, R es la constante universal de los gases, f_i es el coeficiente de fugacidad del componente i, f_i^{ref} es el coeficiente de fugacidad en el estado de referencia y μ_i^{ref} es el potencial químico en el estado de referencia.

Puesto que el estado de referencia generalmente se toma como el solvente líquido puro o como la mezcla fluida en el estado de gas ideal a la misma presión y temperatura, la Ecuación (2.5) suele manejarse como:

$$\mu_i = \mu_i^*(T, P) + RTLnx_i\gamma_i^{\pm} \quad \{2.6\}$$

Donde x_i es la fracción molar del componente i en la mezcla, $\gamma_i^{\pm}$ es el coeficiente de actividad medio iónico por componente i, por lo que para su determinación se requiere el empleo de un modelo de composición local adecuado como lo son el modelo de Pitzer (Lach *et al*., 2018) , el modelo *e-NRTL* (Chen *et al*., 2001), el modelo *e-UNIQUAC* (Mazloumi, 2016) o el modelo *MSA* (Soares *et al*., 2022). El estado de referencia (μ_i^*) se toma, para moléculas neutras, a su presión de vapor, mientras que, para iones, se considera dilución infinita en el solvente.

2.3.2 Modelos basados en ecuaciones de estado (EOS).

Para el caso de modelos cuyo cálculo del potencial químico está en función de ecuaciones de estado su representación está dada por.

$$\mu_i = \mu_i^{\#}(T,P,x) + RTLnfi \quad \{2.7\}$$

Donde f_i es el coeficiente de fugacidad del componente *i* en la mezcla, el cual suele determinarse a través de una ecuación de estado. El estado de referencia ($\mu_i^{\#}$) corresponde a la mezcla fluida tomada como un gas ideal a la misma temperatura *T* y presión *P* que la mezcla fluida y *x* corresponde al vector de composición de la mezcla. Para el cálculo del logaritmo del coeficiente de fugacidad, éste se obtiene utilizando la derivada del número de moles de la energía de Helmholtz residual basada en el volumen y lo cual está representado a través de la siguiente expresión (Michelsen & Mollerup, 2007).

$$RTLnfi = \frac{\partial A^{Res}(T,P)}{\partial n} - RTLn(Z) \quad \{2.8\}$$

Donde *Z* es el factor de compresibilidad (Ver Ecuación 2.8). Una relación para los dos enfoques (EOS *vs* modelos termodinámicos), se puede representar usando la definición de los coeficientes de actividad (Ecuación 2.9):

$$\gamma i = \frac{fi}{fi^*} \quad \{2.9\}$$

Donde fi es el coeficiente de fugacidad del componente *i* en la mezcla y fi^* es el coeficiente de fugacidad del componente *i* en la misma mezcla, pero en una condición de referencia, que generalmente es a una presión y temperatura específicas.

Como puede observarse, el cálculo del potencial químico (μ_i) se facilita a través de modelos que operen utilizando el coeficiente de actividad medio iónico ($\gamma_i^{\pm}$) como se indica en la Ecuación (2.6) y esto es debido a que en este tipo de modelos los parámetros/componentes son de fácil determinación ya que corresponden a variables de estado dentro del sistema o bien, en el caso del coeficiente de actividad este es calculado mediante modelos que están en función de parámetros medibles y de parámetros de interacción que pueden ser conocidos para un sistema en particular o bien, pueden ser determinados a través de un proceso de correlación de datos.

A la fecha, existen diversos modelos termodinámicos empleados para el cálculo del coeficiente de actividad medio iónico reportados en la literatura como son el modelo de Pitzer (Lach *et al*., 2018) , el modelo *e-NRTL* (Chen *et al*., 2001), el modelo *e-UNIQUAC* (Mazloumi, 2016) o el modelo *MSA* (Soares *et al.*, 2022) entre otros. Todos estos modelos están diseñados o construidos considerando las contribuciones

de las interacciones y fuerzas presentes en este tipo de sistemas. En el siguiente apartado se abordará brevemente una explicación de este tipo de interacciones presentes en los sistemas electrolíticos.(Kontogeorgis, 2010)

2.3.3 Interacciones presentes en un sistema electrolítico.

Dado que los sistemas de solución de iones tienen la condición de una elevada no idealidad termodinámica, esto hace que sea atractivo buscar modelar sus diferentes propiedades termodinámicas. Sin embargo, para este tipo de sistemas se requiere un modelado de las diferentes interacciones electroestáticas de corto alcance como lo es la polaridad, las fuerzas de atracción o las asociadas a Van Der Waals; Asimismo, es importante analizar las interacciones de largo alcance que se basan en la teoría de Debye Hückel que se fundamenta en la ley de Coulomb y las interacciones de mediano alcance entre componentes del sistema.

En la práctica, los modelos termodinámicos suelen construirse utilizando un ciclo termodinámico y donde cada transformación, que corresponde a una interacción específica, aporta una contribución aditiva a la energía total de Gibbs (para el caso de los modelos con coeficiente de actividad) o a la energía de Helmholtz (para el caso de los modelos que integran ecuaciones de estado). En el caso de una ecuación de estado, se calcula la energía residual de Helmholtz a un volumen y una temperatura. Conforme a lo anterior, la Ecuación (2.10) muestra la adición de las contribuciones correspondientes para el cálculo de la propiedad de exceso de la energía libre de Gibbs (Rozmus *et al*., 2012).

$$\frac{g^{exc}}{RT} = \frac{g^{exc}_{L.R}}{RT} + \frac{g^{exc}_{M.R}}{RT} + \frac{g^{exc}_{S.R}}{RT} \qquad \{2.10\}$$

Donde $g^{exc}_{L.R}$, $g^{exc}_{M.R}$ y $g^{exc}_{S.R}$ corresponden a las interacciones de largo, medio y corto alcance, respectivamente y que son referidas a la propiedad de exceso de la energía libre de Gibbs. De acuerdo a la termodinámica clásica, el coeficiente de actividad se define como.

$$\ln \gamma_i = \left[\frac{\partial \left(n \frac{g^{ex}}{RT} \right)}{\partial n_i} \right]_{T,P,n_j \neq n_i} \qquad \{2.11\}$$

Donde γ_i es el coeficiente de actividad del componente *i* en solución, *R* es la constante universal de los gases, *T* es la temperatura del sistema en grados Kelvin, n_i es el número de moles del componente *i* y *n* es el número de moles totales del sistema, respectivamente. Así, considerando la expresión anterior y la

Ecuación (2.10), los coeficientes de actividad para el caso de los *Modelos de Composición Local* están representados por.

$$\ln\gamma_i = \ln\gamma_i^{L.R.} + \ln\gamma_i^{M.R.} + \ln\gamma_i^{S.R.} \qquad \{2.12\}$$

Donde $\gamma_i^{L.R.}$ representa la contribución de las fuerzas electroestáticas de largo alcance del componente *i* debido a fuerzas de coulomb, mientras que $\gamma_i^{M.R.}$ y $\gamma_i^{S.R.}$ representa la contribución derivada de las interacciones de mediano y corto alcance del componente *i*.

Es importante mencionar que estas contribuciones están asociadas a fuerzas de descarga, de repulsión y de dispersión, y donde este tipo de interacciones también pueden ser representadas a través de una ecuación de estado tales como la Ecuación *SRK* (Maribo-Mogensen, 2014), o la Ecuación de Peng-Robinson con Myers (Myers *et al*., 2002). Existen trabajos recientes donde el cálculo de estas contribuciones se fundamenta en ecuaciones de estado construidas a partir de la teoría estadística de asociación de fluidos (*SAFT*), y la cual se describe brevemente en el siguiente apartado (Shahriari & Dehghani, 2018). La Tabla 2.1, presenta las interacciones que suelen presentarse en los sistemas electrolíticos, el modelo que lo describe y el tipo de fuerza a la que corresponde la interacción.

Tabla 2.1 *Interacciones comunes en soluciones electrolíticas* (Derbenev et al., 2018).

Interacción	***Descripción***	***Modelo***	***Alcance***
Interacción iónica de Coulomb	*Fuerza atractiva o repulsiva entre iones cargados*	*Ley de Coulomb*	*Corto alcance*
Interacción de van der Waals	*Fuerza atractiva o repulsiva entre átomos o moléculas neutras debido a dipolos eléctricos y de dispersión*	*Teoría de London o de van der Waals*	*Corto alcance*
Interacción de dispersión electrostática	*Fuerza atractiva o repulsiva entre partículas cargadas o dipolares en presencia de un medio dispersante*	*Modelo de Debye-Hückel*	*Largo alcance*

Tabla 2.1 cont…. *Interacciones comunes en soluciones electrolíticas* (Derbenev et al., 2018).

Interacción	***Descripción***	***Modelo***	***Alcance***
Interacción de hidratación	*Fuerza atractiva o repulsiva entre iones y moléculas de agua debido a la polaridad y las interacciones dipolo-dipolo*	*Modelo de solvatación*	*Corto alcance*
Interacción estérica	*Fuerza repulsiva entre iones o moléculas debido a la sobreposición de sus nubes electrónicas o estéricas*	*Modelo de esfera dura o modelo de exclusiones duras*	*Corto alcance*

2.3.4 Teoría estadística de asociación de fluidos (*SAFT*).

Esta se basa en la teoría de perturbaciones la cual emplea la termodinámica estadística para explicar cómo los fluidos complejos y la mezcla de estos forman asociaciones a través de enlaces de hidrógeno (Kontogeorgis, 2010). Desde su propuesta en 1990 se ha empleado en diversos modelos de ecuaciones de estado de base molecular para describir la contribución de la energía de Helmholtz debido a la asociación, permitiendo describir propiedades termodinámicas y de equilibrio de fase de fluidos puros y mezclas de fluidos (Gubbins, 2016; Nezbeda, 2020).

SAFT se ha aplicado a una amplia variedad de fluidos, incluidos fluidos supercríticos, polímeros, cristales líquidos, electrolitos, soluciones de tensioactivos y refrigerantes (Economou, 2001). *SAFT* es una de las primeras teorías en describir con precisión (en comparación con la simulación molecular) los efectos sobre las propiedades de los fluidos del tamaño y la forma molecular, además de la asociación entre moléculas (Chapman et al., 1989, 1990; Gil-Villegas *et al*., 1997).

Entre los trabajos reportados en la literatura enfocados en la modelación de propiedades termodinámicas están la aportación de Lu y Maurer (1993) quienes trabajaron la predicción del coeficiente de actividad en sistemas de alta concentración a través de un modelo de alta interacción. Sun *et al*. (2020) realizaron un estudio de modelos para predecir el coeficiente de actividad. Bollas *et al*. (2008) proponen un modelo *e-NRTL* refinado en sistemas multi-electrolíticos. Por su parte, Hagtalab y Peyvandi (2009) proponen un modelo *e-UNIQUAC-NRF* para la predicción del coeficiente de actividad en sistemas electrolíticos, entre otros estudios.

Conforme a lo anterior, si bien ya existen diversas teorías y modelos que abordan la predicción de propiedades termodinámicas en sistemas electrolíticos y los cuales han aportado resultados satisfactorios, su confiabilidad sigue siendo limitada en sistemas complejos. Por ejemplo, los modelos termodinámicos basados en el cálculo del coeficiente de actividad operan solo a bajas presiones y suelen requerir el conocimiento previo de parámetros de interacción, los cuales en caso de no existir requieren un proceso de correlación de datos con respecto a datos experimentales, y lo cual se complica si estos últimos no existen. Para los modelos basados en la utilización de ecuaciones de estado también se presentan limitaciones en su capacidad predictiva ya que estos modelos presentan desviaciones considerables ante sistemas multi-componentes en disolución o bien, son incapaces de abordar sistemas multi-solventes.

Si bien es correcto que la mayoría de los procesos industriales que operan con sistemas electrolíticos pueden ser diseñados con los modelos termodinámicos existentes, también lo es que existen procesos más complejos cuyos sistemas electrolíticos son nuevos y donde las teorías y modelos existentes son incapaces de generar resultados satisfactorios. Esto se debe a las complicaciones/limitaciones antes expuestas por parte de los actuales modelos termodinámicos.

Ante este tipo de sistemas electrolíticos complejos, surge la necesidad de generar teorías y modelos capaces de mejorar la capacidad predictiva de las propiedades termodinámicas de los mismos. O bien, aplicar metodologías y/o herramientas numéricas capaces de predecir comportamientos, dinámicas y/o parámetros de los sistemas donde las teorías o modelos tradicionales son incapaces de hacerlo. Ante esto, las herramientas basadas en la Inteligencia Artificial (AI), particularmente las estrategias de Aprendizaje Automático, se presentan como herramientas atractivas para ser implementadas para tal fin. En los siguientes apartados se abordará la teoría base de este tipo de herramientas y su aplicación en la predicción de datos en sistemas.

2.4 Herramientas de Aprendizaje Automático (ML).

Dentro del campo de la Inteligencia Artificial existen diversas herramientas entre las que se encuentran las herramientas de Aprendizaje Automático, y donde estas pueden operar como estrategias predictivas de datos o dinámicas de un sistema o proceso. Los siguientes apartados describen brevemente aquellas herramientas de mayor utilización y que están clasificadas dentro de las herramientas de Aprendizaje Automático.

2.4.1 Redes Neuronales Artificiales (*ANN*).

Este tipo de herramienta consiste de un conjunto de unidades denominadas neuronas artificiales interconectadas entre sí para transmitirse señales. De esta manera la información entrante atraviesa la red neuronal sometiéndose esta información a diversas operaciones para generar una información de salida.

Cada neurona esta interconectada con otras a través de enlaces y en cada enlace el valor de salida de la neurona previa es multiplicada por un valor de peso, el cual puede incrementar o inhibir el estado de activación de las neuronas adyacentes. A su vez, a la salida de cada neurona puede existir una función limitadora o umbral la cual modifica el valor resultado o bien, acota el resultado dentro de un límite para ser aceptado por la siguiente neurona, a esta función se le denomina función de activación. A modo de ejemplo, la Figura 2.2 muestra una estructura base de las entradas y salida que puede presentar una neurona dentro de la estructura total de una red neuronal.

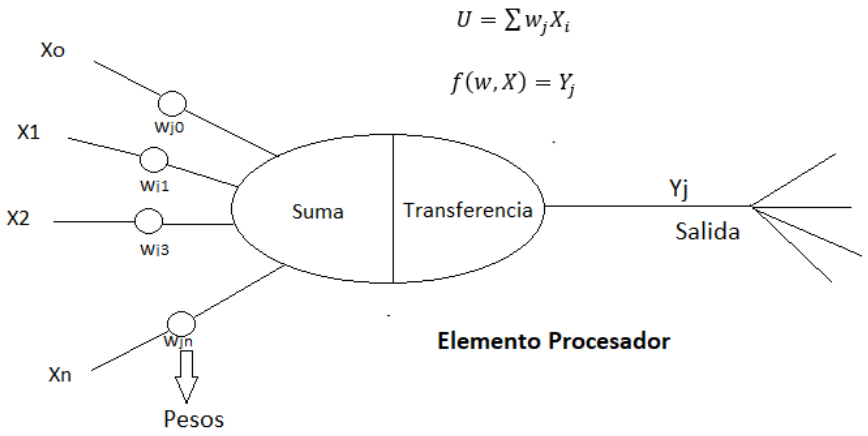


Figura 2.2 *Estructura base de entradas-salidas de una neurona en una estructura general de una red neuronal.*

Así, las redes neuronales artificiales son sistemas que aprenden y se forman, asimismo, en lugar de ser programadas de forma explícita. Este aprendizaje automático se lleva a cabo buscando minimizar una función de perdida que evalúa a la red totalmente, por lo que los valores de los pesos se van ajustado buscando minimizar dicha función. Este último proceso se lleva a cabo a través de la propagación hacia atrás (retroalimentación) dentro de la red neuronal artificial.

Todas las redes neuronales artificiales cuentan con al menos tres capas, una capa de neuronas de entrada, una capa de neuronas intermedias u ocultas y una capa de neuronas de salida, cada capa puede tener desde 1 hasta *n* neuronas y cada neurona tendrá interconexión con las neuronas de las capas previa y posterior a esta. La Figura 2.3 muestra una representación de las capas presentes dentro de una red neuronal artificial y sus inter-conexiones entre capas.

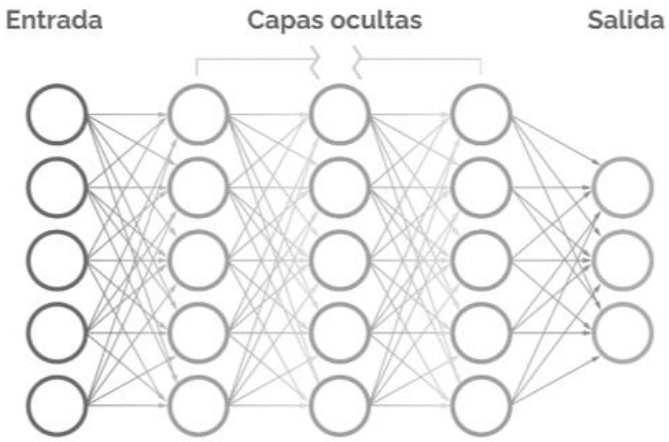


Figura 2.3 *Representación de una red neuronal artificial multi-capas y sus inter-conexiones* .

A la fecha, existen diversas clases y tipos de estructuras de redes neuronales artificiales donde por citar algunas están la red neuronal dinámica (*DNN*), red neuronal recurrente (*RNN*), red neuronal recurrente simple (*SRNN*), red neuronal estocástica (*SNN*), red neuronal modular (*MNN*), entre otras tantas. Independientemente del tipo de red neuronal artificial que se trate, todas estas buscan resolver problemas complejos que van desde cuestiones de control y automatización, toma de decisiones, predicción y modelación de datos, entre otros, Particularmente, en el área de la Ingeniería Química esta herramienta se ha aplicado en tareas de predicción de datos y modelación de sistemas termodinámicos, generación de rutas óptimas en redes de intercambio, entre otras. De manera más puntual, en el caso de la predicción y/o modelación de parámetros termodinámicos en sistemas electrolíticos está el trabajo de Dajnowicz *et al*. (2022) quienes realizaron simulaciones de líquidos electrolíticos mediante una red neuronal de alta dimensión, por su parte Chen *et al*. (2023) emplearon las redes neuronales para la modelación de polímeros-electrolitos en sistemas acuosos involucrando biomoléculas; Yan *et al*. (2021) aplicaron redes neuronales convolucionarias para identificar nuevos electrolitos aplicados en baterías; Pagotto *et al*. (2022) por su parte trabajaron en la predicción de propiedades de sistemas electrolíticas combinando redes neuronales aceleradas y la teoría de solvente continuo. También, Benimam *et al*. (2020) emplearon redes neuronales para la modelación del coeficiente de actividad de soluciones electrolíticas en dilución infinita, por citar algunos estudios.

2.4.2 Proceso Gaussiano Regresor (*GPR*).

Otra herramienta dentro de las técnicas del aprendizaje automático está el Proceso Gaussiano Regresor (*GPR*). Esta es una estrategia de regresión bajo un enfoque bayesiano no paramétrico y cuya funcionalidad se basa en proporcionar predicciones con cierta incertidumbre, es decir, modela funciones aleatorias como

distribuciones de probabilidad. En el contexto de la regresión, se utiliza un *GPR* para modelar una función desconocida que se asume que sigue una distribución Gaussiana (T. Chen & Guestrin, 2016).

En esta estrategia, se parte de un conjunto de datos que se consideran de entrenamiento y que consiste en pares de entrada-salida para entrenar una función que pueda predecir la salida para cualquier entrada. El modelo del *GPR* se compone por una función de media y una función de covarianza, esta última empleada para definir la distribución de probabilidad del *GPR*. La función de la media se representa por la siguiente expresión.

$$\mu(x) = E[f(x)] \qquad \{2.13\}$$

Donde μ (x) corresponde a un vector de entrada de datos transformados que se determina mediante $E[f(x)]$ que corresponde a la función que se desea modelar, y donde *x* corresponde a los datos puntuales de entrada. La función de covarianza está representada mediante.

$$k(x_i, x_j) = E[(f(x_i) - \mu(x_i))\left(f\left((x_j) - \mu(x_j)\right)\right] \qquad \{2.14\}$$

Esta función de covarianza mide la similitud entre los valores de la función en dos puntos diferentes, $f((x_i)$ y $f(x_j)$, además se utiliza para definir la matriz de covarianza la cual es necesaria para calcular la distribución de probabilidad del *GPR*. La función de distribución se define por.

$$p(f(x)|X, y, x|) = N\left(\mu(x), \sigma^2(x)\right) \qquad \{2.15\}$$

Donde $p\left(f(X)\right)$ y $p\left(f(y)\right)$ son los datos de entrenamiento y $p\left(f(x)\right)$ es el punto en el que se desea realizar una predicción. La varianza de la distribución se calcula utilizando la función de covarianza y los datos de entrenamiento. Las funciones de media y de covarianza se utilizan para definir la distribución de probabilidad del GPR y ajustar el modelo a los datos de entrenamiento (Stephenson *et al*., 2018).

De esta manera, el modelo *GPR* se ajusta a los datos mediante la optimización de los hiper-parámetros de la función de covarianza, lo que permite ajustar la complejidad del modelo a los datos experimentales. Además, el *GPR* proporciona una estimación de la incertidumbre asociada a las predicciones del modelo que se establece mediante la distribución de probabilidad, lo que lo convierte en una herramienta útil para el proceso de predicción de datos (Chen *et al*., 2022).

Entre los trabajos reportados en la literatura donde se ha implementado esta herramienta está el de Deringer *et al*. (2021) quienes emplearon *GPR* para la modelación de materiales y moléculas. Kim *et al*. (2011) lo aplicó en el análisis de trayectorias y movimiento. Yu (2012) la aplicó para la predicción de

procesos químicos no lineales, particularmente para modelar mezclas químicas. Por su parte Zhou, *et al*. (2015) lo implemento para la modelación de procesos batch, entre otros estudios.

2.4.3 Aumento de gradiente extremo (*XGBoost*)

Otra herramienta dentro de las estrategias del aprendizaje automático esta la denominada eXtreme Gradient Boosting (*XGBoost*) la cual es un algoritmo de aprendizaje automático basado en árboles de decisión. Esta herramienta emplea el método de refuerzo para construir modelos predictivos. Su operación parte de crear un conjunto de árboles de decisión, donde cada árbol se construye sobre los errores del árbol anterior y donde se busca la minimización de una función objetivo. La función objetivo de *XGBoost* se define como la suma de una función de pérdida y una función de regularización y se optimiza normalmente utilizando técnicas de descenso de gradiente. La función objetivo de *XGBoost* está definida por.

$$obj^{(t)} = \sum_{i=1}^{n} L\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_{t)}) \qquad \{2.16\}$$

Donde *t* es el número de iteraciones $L\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right)$ es la función de perdida que mide la diferencia entre la predicción actual y la verdadera salida, $y_i, \hat{y}_i^{(t-1)}$es la predicción actual, $f_t(x_i)$ es la predicción del nuevo árbol en la iteración para el ejemplo *i*, por su parte $\Omega(f_{(t)})$ es la función de regularización, que mide la complejidad del árbol (Chen & Guestrin, 2016).

Entre los estudios reportados en la literatura en los que se ha implementado esta estrategia está el trabajo de Said *et al*. (2023) quienes emplearon las herramientas *XGBoost* y *GPR* para modelar la conductividad termina en nano-fluidos. Zhong *et al*. (2021) aplicaron *XGBoost* en la predicción de reactividad de radicales HO junto con componentes orgánicos. Grols y Domínguez (2021) aplicaron esta herramienta en la predicción del proceso de co-cristalización de compuestos, por citar algunos estudios.

Si bien existe una mayor cantidad de estudios en los que se han implementado estas o alguna otra herramienta del área del aprendizaje automático, es particularmente en el campo de la Ingeniería Química donde su aplicación se presenta como herramientas atractivas para su implementación, particularmente para la modelación/predicción de parámetros y/o propiedades químicas o termodinámicas y principalmente, en sistemas o procesos complejos que se caracterizan ya sea por su alta no linealidad o bien, sistemas donde los modelos o teorías actuales tienen limitación en su capacidad predictiva pues operan ya sea bajo un rango operativo limitado o donde se carece de información de ciertos parámetros para nuevos sistemas.

Por lo anterior, este estudio aborda la implementación de este tipo de herramientas de aprendizaje automático para la modelación de propiedades termodinámicas en sistemas electrolíticos, primeramente, para mejorar la capacidad de modelación con respecto a los modelos actuales los cuales presentan

desviaciones en sistemas complejos y en una segunda parte, se propone la creación de un modelo hibrido el cual parte de un modelo de predicción actual y se incorpora esta herramienta de inteligencia artificial con el fin de mejorar la capacidad predictiva del primero.

En el siguiente capítulo se describe de forma general la metodología que se siguió en la presente investigación y como se cuantificaron los resultados para el logro de los objetivos trazados de inicio.

Capítulo 3. Metodología.

En este capítulo se describe el procedimiento que se siguió para el desarrollo de la presente investigación. Cada etapa del proyecto es descrita a detalle y cada una abonará para el logro de los objetivos trazados.

3.1 Recolección de datos.

Puesto que el principal objetivo de esta investigación consiste en desarrollar una metodología que permita mejorar la capacidad predictiva o de modelación de las propiedades termodinámicas de los sistemas electrolíticos en solución con respecto a los modelos actuales, y dado que esto se busca mediante la implementación de herramientas de aprendizaje automático, es necesario contar con suficientes datos como base de estos sistemas, principalmente con respecto a la propiedad termodinámica que se desea modelar y que como se ha mencionado previamente corresponde al coeficiente de actividad y al coeficiente osmótico los cuales fungirán como los elementos de salida, También deben tenerse suficientes parámetros asociados a estos sistemas que permitan ser alimentados como los elementos de entrada para las estrategias de aprendizaje automático que se implementen.

Si bien es cierto que la literatura ofrece información de numerosos sistemas de electrolitos, su recolección tiende a ser complicada y llevar un tiempo considerable. Ante esto, algunos autores han trabajado para generar bases de datos que involucran distintos sistemas de electrolitos, en su mayoría en solución acuosa puesto que industrialmente son los más utilizados. Entre las bases de datos disponibles que presentan información referente a los sistemas electrolíticos está una que fue la que se utilizó y que a continuación se referencia. Todo esto y los apartados subsecuentes forman parte del área de ciencia de datos y es necesaria antes del empleo de los modelos de aprendizaje automático.

3.1.1 Base de datos Idaho de soluciones termodinámicas (*IDST*)

Una base de datos de acceso abierto y con un amplio rango de sustancias soluto-solvente, disponible desde el 2020 es la base de datos termodinámica de soluciones de Idaho (*IDST*). Esta base de datos cuenta con más de 300 conjuntos de datos acuosos y 121 conjuntos de datos no acuosos, considerando 9 solventes diferentes al agua y 119 referencias, entre las cuales en su mayoría son sistemas electrolíticos. Los datos que se tienen en esta fuente de datos se caracterizan por entregar 10 o más puntos de datos en un rango de concentraciones para una condición isotérmica (Feeley *et al*., 2021).

3.1.2 Extracción de datos iniciales.

Previo a la recopilación de la información precedente de la base de datos elegida, se hace una evaluación de que datos existentes son necesarios e importantes para la investigación, esto debido a que la base *IDST* ofrece una variedad de propiedades a distintas concentraciones para cada tipo de solución, además de presentar una referencia de estos datos experimentales. Particularmente, los datos de interés para este estudio corresponden a las propiedades referidas al coeficiente de actividad y coeficiente osmótico.

Dado que se busca la predicción de las propiedades mencionadas anteriormente, se seleccionaron variables que, de inicio serán evaluadas para conformar el conjunto de datos de entrada de los modelos de aprendizaje automático a implementar.

La Tabla 3.1 ilustra la clasificación propuesta de las diferentes variables que se extrajeron de la base *IDST* y que en conjunto fungirán como los datos de entrada-salida en los modelos de aprendizaje automático. Estos datos corresponden a un conjunto de 334 sistemas electrolíticos.

Tabla 3.1 *Clasificación de variables entrada-salida seleccionadas.*

Variables de entrada			***Variable de salida (interés)***
*Peso molecular solvente**	*Concentración molal (mol/kg solvente) **	*Fuerza iónica* *fracción molar iónica*	*Coeficiente de actividad medio iónico, Coeficiente osmótico*
Carga del anión	*Anión (Categ.)**	*Núm. de partículas*	
*Catión (Categ.)**	*Núm. Cationes**	*Núm. Aniones**	
*Peso molecular Soluto**	*Carga del catión*	*Temperatura(K)**	

(*) Variables predictoras para entrada de los modelos.

Conforme a la Tabla 3.1, de los datos extraídos estos se clasificaron en 13 variables de entrada de las cuales tres están relacionadas con la concentración, cinco referidas a la cantidad y carga de los iones, dos referidas a pesos moleculares y una a la temperatura del sistema; Además se consideraron dos variables categóricas que representan la carga del anión y catión.

Todas estas variables se recolectaron para los diferentes sistemas existentes, de forma que para cada dato de concentración de cada solución se tienen datos completos de estas 14 variables y de las dos variables de salida de interés. Cabe hacer mención que dada la cantidad de sistemas dentro de la base de datos *IDST* se hizo uso de librerías y programación en Python®. La Figura 3.1 muestra la secuencia que se siguió para la extracción de los datos antes mencionados.

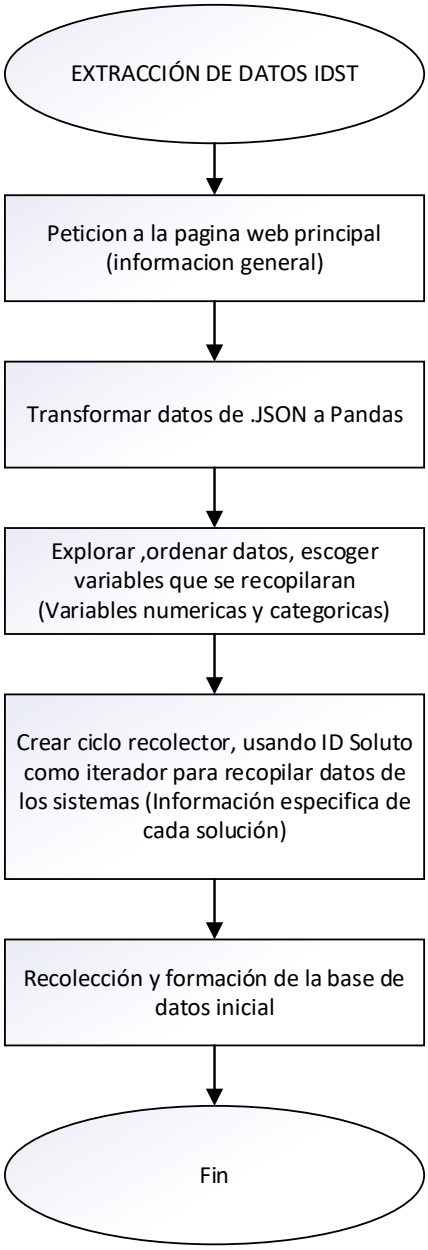


Figura 3.1 *Secuencia extracción de datos de manera remota usando como base de datos (IDST).*

3.2 Pre-tratamiento de datos.

Una vez que los datos son extraídos de la base de datos *IDST* estos deben ser analizados y depurados previo a su utilización y esto es debido a que bastantes sistemas suelen presentar información incompleta o sesgada. Así, para manejar esta información se realizó un pre-tratamiento que consistió en localizar datos inconsistentes o vacíos que presente la base de datos inicial. A su vez, se realizó una depuración de datos tipo outliers (datos alejados numéricamente del resto de datos), esto con el fin de dar consistencia a los datos numéricos y evitar datos fuera de la tendencia o secuencia (Nisbet *et al*., 2018).

De esta forma, el pre-tratamiento de los datos se llevó a cabo empleando librerías que operan para localizar sesgos o desviaciones en los datos tales como pandas y SK-Learn. A su vez, uno de los métodos que permitió la identificación de datos anómalos fue el código denominado Isolation Forest (IF). Una vez realizada la depuración se eliminó el 1% de los datos anómalos, esto con el fin de evitar perdida de información (Ding & Fei, 2013).

Posterior a la depuración y con los programas indicados, se llevó a cabo otro pre-tratamiento de datos con el fin de ubicar soluciones con cationes y aniones que generaban información repetida para diferentes sistemas. Esta depuración de datos atípicos, valores nulos e inconsistentes y sistemas repetidos, permitió obtener un conjunto de datos completos que servirán como daos de entrada-salida en los algoritmos a emplear.

3.2.1 Codificación de las variables de entrada categóricas.

Como se mencionó anteriormente, se identificaron dos variables categóricas (carga del anión y catión) con el fin de buscar que estas variables mejoren la capacidad predictiva de las herramientas de aprendizaje automático que se implementen. Cebe hacer mención que las variables categóricas no se pueden introducir directamente en una red neuronal artificial ya que estas variables al ser de tipo cualitativo y no cuantitativo, requieren una transformación previa para ser representados por valores numéricos (Hastie *et al*., 2009).

Por ejemplo, Serrano *et al*. (2020) emplearon variables categóricas para mejorar un modelo de gasificación. En su estudio el material del lecho para un gasificador de lecho fluidizado se consideró una variable categórica y por ende, codificaron esto como una variable ordinal, es decir, adaptando un orden cuantitativo en las variables. Sin embargo, los valores de las variables de esta investigación no siguen un orden o escala intrínsecos y por tanto, sería más apropiado aplicar una estrategia de trasformación más robusta.

Conforme a lo anterior, las variables categóricas consideradas en este estudio (catión y anión), fueron tomadas como variables nominales, es decir se procede a que una variable categórica la cual puede tomar

n valores sea transformada en una variable binaria, donde los valores a tomar sean de 1 o 0, indicando la presencia o ausencia de la variable, respectivamente; Esto significa que una variable categórica de entrada con cinco categorías requerirá 5 columnas de entrada. Para el caso de las variables categóricas del presente estudio, se tienen más de 50 categorías para cada una de estas, Por lo tanto, la codificación generó más de 50 columnas nuevas que corresponden a los diferentes tipos de cationes y aniones. Para llevar a cabo la codificación antes mencionada se utilizó la librería Sk-Learn (Zhang & Wallace, 2017).

3.3 Análisis preliminar de datos de entrada.

Una vez que las variables categóricas fueron transformadas para otorgarles un valor numérico y puedan integrarse al conjunto de las variables de entrada, las 11 variables restantes y que se indican en la Tabla 3.1, recibieron un análisis preliminar de datos de entrada. Con este análisis se pretende determinar el nivel de relación entre variables con el fin de poder descartar aquellas cuyo nivel sea alto y reducir así el número de variables de entrada (variables predictoras) al algoritmo de aprendizaje automático. Para esto, se aplicó un proceso de correlación de Spearman (*SCC*).

. Es importante mencionar que en diversos estudios se propone emplear la correlación de Pearson, pero esta aplica siempre y cuando los lleven una distribución gaussiana, lo que no es el caso para este estudio. La expresión que define la correlación de Spearman está dada por.

$$SCC = \frac{\sum_{i=1}^{N}(R(xi)-\bar{R}(x))*(R(yi)-\bar{R}(y))}{\sqrt{\sum_{i=1}^{N}\left(R(xi)-\bar{R}(x)\right)^2}*\sqrt{\sum_{i=1}^{N}\left(R(yi)-\bar{R}(y)\right)^2}} \qquad \{3.1\}$$

Donde *N* es el tamaño de la muestra, $R(x_i)$ y $R(yi)$ son los rangos de las muestras individuales de las dos variables a comparar, respectivamente. $R(x)$ y $R(y)$ representa el rango promedio de las dos variables a comparar, respectivamente.

Para un *SCC* de magnitud nula se interpreta que dos parámetros o variables no están correlacionadas, y cuanto más se acerque el valor a 1 o -1, existirá una relación lineal. Para este estudio, y en conformidad con la literatura, se considera que las variables tendrán correlación si su *SCC* es >0.8 o <-0.8 y ello conducirá a eliminar a una de las variables involucradas antes de su utilización en los modelos de aprendizaje automático. Este proceso es necesario con el fin de eficientar el proceso de modelación de las herramientas (Lu *et al*., 2019).

En una primera fase, se cuantificó la correlación entre las 11 variables de entrada antes mencionadas y las dos variables de salida, lo cual permitió determinar el nivel las relaciones lineales que existen en el conjunto de variables. La Figura 3.2 presenta una matriz de correlación entre las variables de entrada-salida y su nivel de ponderación (*SCC*).

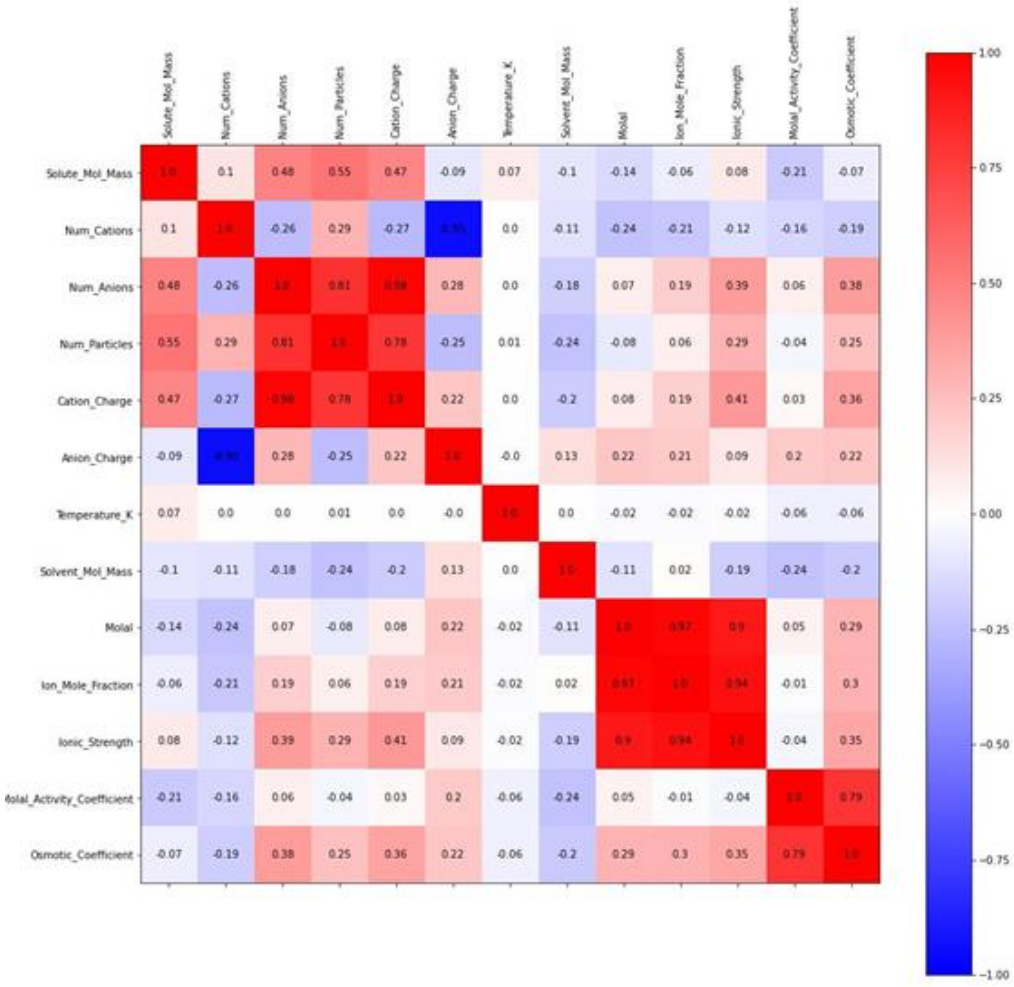


Figura 3.2 *Matriz del análisis de correlación entre variables de entrada-salida.*

Así, y de acuerdo a los resultados con que se construyó la Figura 3.2, se logró identificar a las variables que presentan un nivel de correlación considerable y que por ende pueden ser susceptibles de eliminarse. Se observa que las variables que presentaron mayor correlación entre ellas fueron el número de catión, el número de anión, la carga del catión, la carga del anión y el número de partículas. Además, otra relación con fuerte correlación identificada corresponde entre las tres variables relacionadas a la concentración, siendo la concentración molal, la fracción molar iónica y la fuerza iónica.

Con base a estos resultados, para este estudio en particular se descartaron las variables fracción molar iónica y la fuerza iónica, dejando solo a la concentración molal como variable predictora y por otra parte, se descartaron las variables número de partículas, la carga del catión y la carga del anión, dejando a las variables número de catión, el número de anión, para mantenerlas como variables predictoras junto con las

variables que no presentaron una fuerte correlación. Así, de tener 13 variables de inicio, solo 8 son consideradas como variables predictoras (seis del análisis de correlación y dos correspondientes a las variables categóricas).

Una vez realizada la depuración de aquellas variables que presentaron un nivel de correlación importante, el proceso de análisis de correlación se realizó nuevamente, pero ahora solo con las seis variables que no presentaron correlación, esto con el fin de garantizar la ausencia de correlación entre estas. La Figura 3.3 muestra la matriz del análisis *SCC* obtenida y la cual corrobora lo antes mencionado.

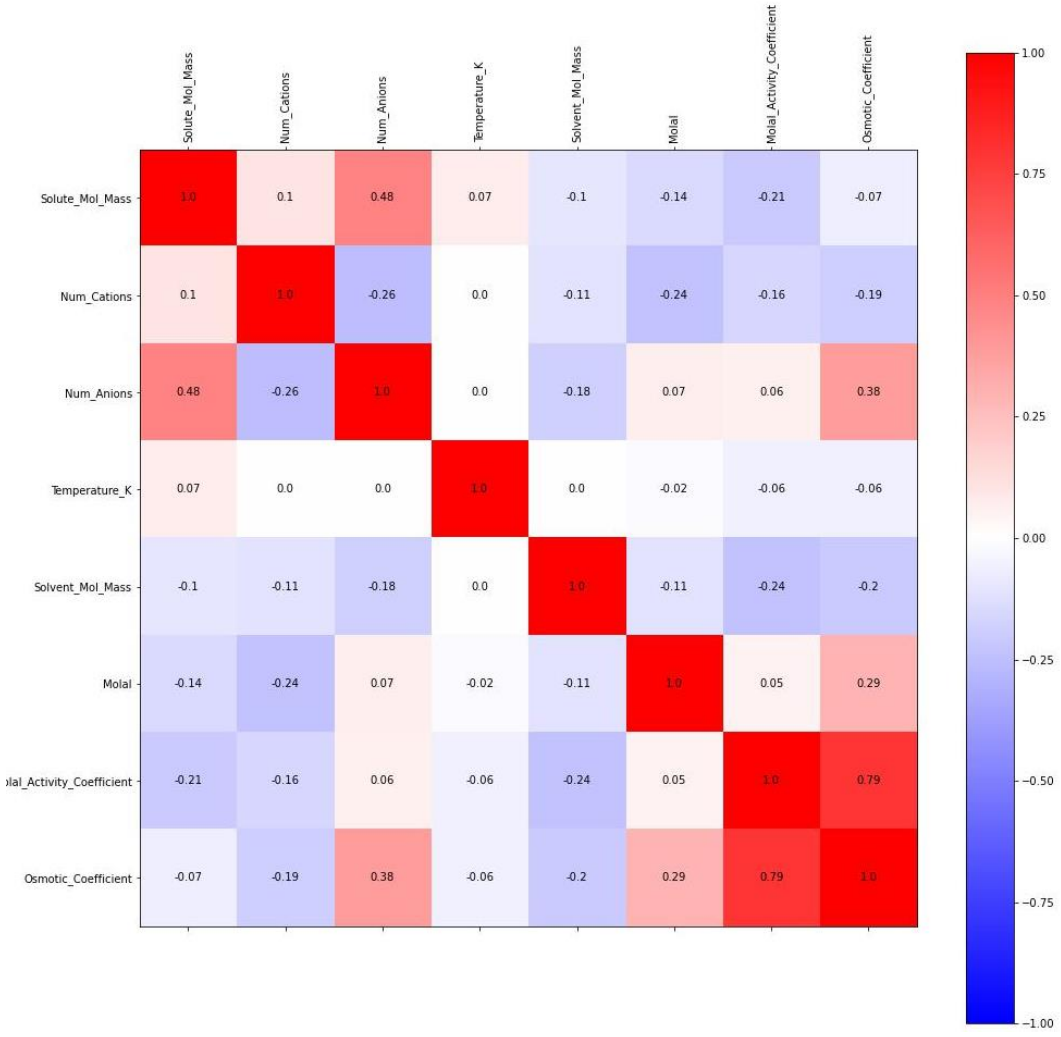


Figura 3.3 *Matriz del análisis de correlación entre variables predictoras de entrada-salida.*

3.4 Reestructuración de la base de datos de entrada.

Definidas las variables predictoras que serán las variables de entrada a los algoritmos de aprendizaje automático se realizó una clasificación-cuantificación final de los datos de ingreso. La Tabla 3.2 muestra la cuantificación de datos. Esto no significa una eliminación de datos o transformación de los mismos, sino solo es una actualización de la información existente quitando a las variables con alta correlación.

Tabla 3.2 *Cuantificación final de datos de entrada.*

Datos de entrada a los algoritmos	
Número Total de datos	*8604*
Número de sistemas electrolíticos Total	*334*
Número de puntos por sistema	*10-50*
Numero de variables predictoras	*8 (2 categóricas)*
Numero de propiedades a predecir	*2*
Soluciones electrolíticas básicas	*4*
Soluciones electrolíticas acidas	*27*
Soluciones electrolíticas salinas	*303*

Cabe destacar la importancia que tiene el contar con la mayor cantidad de datos al emplear herramientas de laprendizaje automático con fines de predicción y/o modelación ya que a mayor cantidad de datos el algoritmo mejorará su capacidad predictiva y los resultados obtenidos serán más consistentes acorde a lo real. Con respecto a los sistemas salinos estos se clasificaron por familia de la sal donde la Tabla 3.3 muestra esta clasificación y la cantidad de sistemas por tipo de sistema salino.

Tabla 3.3 *Clasificación de sales presentes.*

Clasificación soluciones salinas	
Tipo de anión	*Numero de sistemas*
Cloruros	*85*
Fluoruros	*12*
Bromuros	*56*
Yoduros	*19*
Nitratos	*32*
Sales de azufre	*53*
Otro	*46*

Cada una de las variables predictoras que operan como variables de entrada, deben contar con un valor dentro de un rango definido, y el cual es importante establecer al utilizar una herramienta de aprendizaje autónomo. Esto con la finalidad de que el entrenamiento previo de la herramienta pueda operar con conocimiento de los valores puntuales y así esta pueda mejorar la capacidad predictiva de lo que se busca. La Tabla 3.4 muestra los intervalos y cantidad de valores con que opera cada una de las variables predictoras o variables de entrada hacia los algoritmos de aprendizaje automático que se utilizaran en este estudio.

Tabla 3.4 *Intervalos o cantidad de datos con que operan las variables predictoras.*

Variables predictoras	***Rango- cantidad***
Peso molecular soluto	*20 - 1000 g/mol*
Peso molecular solvente	*20 - 90 g/mol*
Temperatura	*0 - 60°C*
Número de cationes	*1-4*
Número de aniones	*1-4*
Concentración	*0.001 - 30 mol/L*
Tipo Catión	*86*
Tipo Anión	*77*

Es importante resaltar que todas las variables de entrada consideradas presentan diferentes escalas, por lo que es importante realizar una transformación de los datos antes de incorporarlos en los códigos de aprendizaje automático, y este proceso se realiza con la finalidad de estandarizar las escalas en cada una de las variables de entrada. El siguiente apartado aborda este proceso de estandarización de los datos de ingreso a las herramientas de inteligencia artificial.

3.5 Proceso de estandarización de los datos de entrada.

Dada la variabilidad de rangos y valores con que operan las diferentes variables predictoras y que serán las variables de entrada a los algoritmos de aprendizaje automático, se procede a realizar una normalización o estandarización de los datos con la finalidad de que las variables operen todas dentro de una misma escala. Esto es beneficioso para los algoritmos de aprendizaje automático ya que de acuerdo a la teoría tienden a trabajar de manera más robusta con datos estandarizados. De esta forma, la estandarización de los datos se realizó aplicando la siguiente expresión.

$$Xestand = \frac{X-\mu}{\sigma} \qquad \{3.2\}$$

Donde *X* es el valor de la variable de entrada, μ es la media del conjunto de datos de la variable y σ es la desviación estándar del conjunto de datos. Así, una vez realizada la normalización de datos para todo el conjunto de variables de entrada se procedió a desarrollar las herramientas de inteligencia artificial que se utilizarían en este estudio y las cuales se explican a continuación.

3.6 Desarrollo de modelos de aprendizaje automático.

3.6.1 Desarrollo de red neuronal artificial (*ANN-MLP*)

Para el desarrollo del modelo de la red neuronal que se implementará, se dividió la base de datos final en dos partes, el primer conjunto de entrenamiento y el segundo conjunto de prueba en una proporción 80/20, respectivamente. La selección de los datos para conformar ambos conjuntos se realizó de manera aleatoria, con el fin de mejorar el proceso de entrenamiento y se generen resultados adecuados. Dentro del conjunto de datos de entrenamiento se tomó el 10% de éste para emplearlos como conjunto de validación y el cual se aplicó en el ajuste de los hiper-parámetros de la red. La ecuación que refleja la base matemática de la red neuronal artificial que se desarrollo está definida por.

$$yk = f_0(\sum_h w_{hk}\, f_h(\sum_i w_{ih} x_i)) \qquad \{3.3\}$$

Donde *yk* corresponde a la respuesta o salida *y* por nodo *k* en la red neuronal, f_0 es la función para los nodos de salida de la red, x_i: es la variable del dato de entrada *i*, f_h es la función para los nodos ocultos de la red, el índice *h* corresponde al número de nodos ocultos en la red, el índice *i* corresponde al número del dato de entrada, el índice *k* corresponde al número o dato del nodo de salida, w_{hk} es el peso otorgado al nodo oculto *h* por nodo de salida *k* y w_{ih} es el peso otorgado por ejemplo *i* en el nodo oculto *h*.

En particular, para este estudio, se desarrolló un tipo de red neuronal artificial de tipo perceptrón multicapa (***ANN-MLP***), la cual contó con una única capa de salida para las dos variables de interés (Coeficiente de actividad y Coeficiente osmótico). En su diseño se empleó la biblioteca Keras contenida en Python® y su modelo secuencial para crear el modelo de la red y el cual fue entrenado con los datos correspondientes. Este tipo de red presenta diferentes hiper-parámetros que se ajustaron mediante un proceso de valorización. Para ello, se emplearon herramientas específicas como Keras Tunner y el método de búsqueda aleatoria (O'Malley *et al*., 2019).

Para el ajuste de los hiper-parámetros de la red neuronal se identificó primero el tipo de parámetro y el rango o valores con que este podría operar y una vez definidos estos, se procedió a su valorización a partir del empleo de técnicas como validación cruzada. Con ello, se logró determinar la configuración óptima para la red neuronal (***ANN-MLP***) (Géron, 2019). La Tabla 3.5 muestra los hiper-parámetros, sus rangos/valores estudiados y el valor tomado como óptimo para la red neuronal a utilizar.

Tabla 3.5 *Datos de los hiper-parámetros y sus valores óptimos para el modelo (ANN-MLP).*

Hiper parámetros variables	***Rango/valores***	***Valor elegido***
Numero capas ocultas	*1-3*	*3*
Numero de neuronas por capa	*5-15-20*	*20 capa 1- 15 capa 2 y 3*
Optimizador	*Adam - Adamax*	*Adam*
Tasa de aprendizaje	*0.1-0.01-0.001*	*0.01*
Tamaño del lote	*32-64-128-256*	*256*
Tipo de función de activación	*Relu – Swish- Smish*	*Smish*
Fracción de ponderación de la función MAPE. (α)	*0-1*	*0.08*
Fracción ponderada del peso en la función de rendimiento para (w_γ)	*0-1*	*0.6*
Hiper parámetros Fijos		
Función de rendimiento	*Combinación (MAPE Y MSE)*	*Combinación (MAPE Y MSE)*
Parámetro de detención temprana	*12*	*12*
% datos de prueba	*20%*	*20%*

Los parámetros α y w_γ hacen referencia a la ponderación dada de la función *MAPE* y la ponderación dada al peso en la función de rendimiento de la variable de salida del coeficiente de actividad, respectivamente.

3.6.2 Funciones de rendimiento empleadas en la *ANN-MLP*.

Las funciones de rendimiento son una parte fundamental en la construcción y optimización de los modelos de redes neuronales artificiales y su elección influye significativamente en el rendimiento del modelo. Estas funciones también conocidas como funciones de pérdida, son implementadas en las redes neuronales artificiales para evaluar la diferencia entre la salida esperada y la salida real durante la etapa de entrenamiento del modelo. Estas funciones permiten que la estructura del modelo ajuste sus parámetros para minimizar el error y mejorar la precisión de las predicciones.

Conforme a la literatura, existen diferentes tipos de funciones de rendimiento que se utilizan según el tipo de problema que se esté abordando, tales como la regresión o la clasificación. Por ejemplo, en la regresión se puede utilizar el error cuadrático medio (*MSE*) o el error absoluto medio porcentual (*MAPE*), mientras que en la clasificación se puede utilizar la entropía cruzada (cross-entropy) o el índice de Gini. Para el caso de este estudio se utilizó la función de rendimiento de tipo regresión.

Las Ecuaciones (3.4) y (3.5) muestran las ecuaciones de rendimiento $MAPE_{\gamma,\Phi}$ y $MSE_{\gamma,\Phi}$, respectivamente, mientras que la Ecuación (3.6) muestra la función combinada de ambas expresiones y la cual fue la que se utilizó en este trabajo.

$$MAPE_{\gamma\Phi} = \left(\frac{1}{N}\sum_i^N \frac{|y_i^0 - y_i^p|}{y_i^0} * 100\%\right) wA + (1 + wA) \left(\frac{1}{N}\sum_i^N \frac{|y_i^0 - y_i^p|}{y_i^0} * 100\%\right) \quad \{3.4\}$$

$$MSE_{\gamma\Phi} = \frac{1}{N}\sum_{i=1}^N (y_i^0 - y_i^p)^2 * wA + (1 - wA) * \frac{1}{N}\sum_{i=1}^N (y_i^0 - y_i^p)^2 \quad \{3.5\}$$

$$F_{combinada} = \frac{1}{N}((1-\alpha)\sum_{i=1}^N (y_i^0 - y_i^p)^2 + \alpha\left(\sum_i^N \frac{|y_i^0 - y_i^p|}{y_i^0} * 100\right) \quad \{3.5\}$$

Donde N es el número de datos de entrenamiento, y_i^0 y y_i^p corresponden al valor de salida esperada y real, respectivamente para el dato *i*, respectivamente.

De esta manera al tener definidos el tipo de red neuronal a utilizar, la ecuación que define la interacción nodal, los valores de los hiper-parámetros a utilizar y la función de rendimiento junto con los datos de entrenamiento y prueba, se puede evaluar la capacidad predictiva de la red neuronal.

3.6.3 Desarrollo del modelo Proceso Gaussiano Regresor (*GPR*).

Otro de los modelos de aprendizaje automático que se abordó en este trabajo fue el del modelo Proceso Gaussiano Regresor (*GPR*), el cual se llevó a cabo mediante la librería SK-learn de Python® y en donde se utilizó la misma agrupación de los datos de entrada que se utilizaron para las redes neuronales.

De igual forma, se identificaron los hiper-parámetros del modelo, así como sus rangos o valores de operación y se procedió a su valoración con el fin de determinar los valores apropiados con que operaria el modelo y los cuales permitirán obtener los valores mínimos de los errores de los conjuntos de entrenamiento y de validación, esto mediante las herramientas de Keras Tunner con su método de búsqueda aleatorio y validación cruzada con un cv=5. La Tabla 3.6 muestra los hiper-parametros estudiados, asi como sus rangos o valores de operación y el dato elegido para operar el modelo *GPR*.

Tabla 3.6 *Datos de los hiper-parámetros y sus valores óptimos para el modelo (GPR).*

Hiper parámetros (GPR)	***Rango-valor***	***Valor elegido***
λ_G *Longitud de escala*	*0-1*	*0.8*
v (Parámetro de suavidad)	*1-2*	*2*
α_G *(Parámetro de regularización)*	*0.1-0.01-0.001*	*0.1*
Numero de reinicios	*2-10*	*4*
Hiper parámetros Fijos (GPR)		
Kernel	*Mattern*	*Mattern*
Datos de prueba	*20%*	*20%*

El hiper-parámetro denominado *Kernel* corresponde al tipo de función de covarianza donde el modelo tipo *Mattern* se caracteriza por su capacidad para modelar procesos con patrones de fluctuaciones irregulares y discontinuos, que son necesarias en el caso de este estudio. La expresión que define este modelo de covarianza se presenta a continuación (Gómez-Bombarelli *et al*., 2018).

$$k(xi, xj) = \sigma^2 * [(2^{(1-v)/\gamma G\,(v))} * \left(\sqrt{2v} * \frac{d(xi,xj)}{\lambda}\right)^{v}] * B_v\left(\sqrt{2v} * \frac{d(xi,j)}{\lambda}\right) \quad \{3.6\}$$

Donde σ^2: es la varianza del proceso, $d(xi, xj)$ es la distancia euclidiana entre los puntos *xi* y *xj*, $\gamma G\,(v)$ es la función gamma y B_v es la función de Bessel modificada de orden *v*. Así, dado que el modelo *GPR* es una herramienta de predicción de datos basado en la minimización de la anterior función y dado que está ya se tiene definida, se procedió a su evaluación para analizar su capacidad predictiva y cuyos resultados se presentan el apartado respectivo.

3.7 Desarrollo del modelo hibrido (*Wilson^{XM}-XGB*).

Tal como se mencionó anteriormente, si bien existen propuestas de modelos que tratan de predecir alguna variable termodinámica de interés y donde sus resultados son aceptables, estos en diversas ocasiones presentan desviaciones al operar fuera de un rango, con valores de parámetros de interacción no adecuados o nulos o bien, complejidad misma de la solución/mezcla a estudiar, entre otros. Así, con el fin de superar estas limitaciones que tienen los modelos actuales, se propone el desarrollo y evaluación de un modelo hibrido que combine un modelo de composición local con un modelo de aprendizaje automático, específicamente el modelo eXtreme Gradient Boosting (*XGBoost*), para la determinación del parámetro termodinámico coeficiente de actividad.

El objetivo de esta hibridación es emplear los resultados de la modelación del modelo de composición local como característica base para el modelo *XGBoost*, lo que se conoce como transferencia de aprendizaje. De esta manera, el modelo híbrido se beneficia de la capacidad del modelo de composición local para capturar las interacciones a nivel molecular, así como de la capacidad del modelo de aprendizaje automático para generalizar a partir de estos datos patrones complejos que permitan mejorar la capacidad predictiva de la propiedad de interés. En conjunto, esta metodología híbrida ofrece una forma prometedora de predecir el coeficiente de actividad de soluciones electrolíticas con mayor precisión. En los siguientes apartados se describe brevemente el modelo de composición local utilizado, así como la herramienta de aprendizaje automático que se utilizó en el proceso de hibridación.

3.7.1 Modelo de composición local Wilson de Xu y Macedo (*Wilson^{XM}*)

Si bien existen diversos modelos de composición local, para este estudio se establece emplear el modelo de Wilson propuesto por Xu y Macedo (2003). Inicialmente el modelo de Wilson fue desarrollado para representar la no idealidad de mezclas no electrolíticas, sin embargo, su extensión a sistemas electrolíticos requiere de varios parámetros ajustables. Como alternativa, Renon y Prausnitz (1969) propusieron una forma de derivar el modelo de Wilson (1964) donde se requiere de varios parámetros ajustables. Como alternativa, Renon y Prausnitz (1969) generaron una propuesta a partir del concepto de composición local, asumiendo que el parámetro de energía es equivalente a la entalpía local de las celdas (Zhao *et al*., 2000). En este sentido, se empleó una extensión del modelo de Wilson propuesto inicialmente para soluciones acuosas de polímeros, donde la energía libre de Gibbs de exceso se representa a través de.

$$\frac{g^{EX}_{S.R.}}{RT} = \frac{1}{\alpha}(X_S \ln[X_S + (X_a + X_c)H_{em}] + X_c \ln(X_a + X_s H_{em}) + X_a \ln(X_c + X_s H_{me})) \quad \{3.7\}$$

Y donde las expresiones que definen el cálculo del coeficiente de actividad por catión y anión, así como la determinación del coeficiente de actividad medio iónico estarán definidas, respectivamente por.

$$ln\gamma_c^{wilson} = -Z_c\left[\frac{X_s H_{em}}{X_S+(X_a+X_c)H_{em}} + \frac{X_a}{X_c+X_s H_{me}} + \ln(X_a + X_s H_{me}) - H_{em} - lnH_{me}\right] \quad \{3.8\}$$

$$ln\gamma_a^{wilson} = -Z_c\left[\frac{X_s H_{em}}{X_S+(X_a+X_c)H_{em}} + \frac{X_c}{X_a+X_s H_{me}} + \ln(X_c + X_s H_{me}) - H_{em}X_a - lnH_{me}\right] \quad \{3.9\}$$

$$\ln(\gamma_{\pm}) = \frac{1}{v}[v_+ \ln\gamma_c + v_- \ln\gamma_a] - \ln(1 + 0.0001 M_{w\ Solvente}\ mv) \quad \{3.10\}$$

Donde α es un parámetro que caracteriza la tendencia de la especie *i* y *j* a no distribuirse al azar y cuyo valor puede fijarse ó ajustarse, variando de 0.008 a 7.0, aunque Xu y Macedo (2003) recomiendan emplear un valor de 0.1. En este estudio se consideró los escenarios donde α es constante y también donde es un parámetro de ajuste. Los parámetros H_{em} y H_{me} corresponden a las energías de interacción (solvente/ión, ión/solvente) y se obtienen del ajuste de los datos experimentales. Estos parámetros son una función de la diferencia de entalpías entre especies de la celda local. La variable X_j es la fracción molar efectiva para un sistema electrolítico conteniendo un ión *j* con carga Z_j y se calcula mediante la expresión $X_j = x_j C_j$ donde $C_j = Z_j$ para iones y $C_j = 1$ para moléculas.

Con respecto al coeficiente de actividad promedio iónico, este es una función de la molalidad (m), del peso molecular del solvente (M_w solvente), de los coeficientes estequiométricos (v_+) y (v_-) que se obtienen de la disociación de la sal $M_{v_+}X_{v_-} \leftrightarrow v_+M^{z+}v_-X^{z-}$, y de la suma de estos coeficientes (v).

Referente a los sistemas abordados para el modelo híbrido se consideraron 34 soluciones electrolíticas de sales de amonio acuosas a 25 grados Celsius correspondientes a 743 registros. Estos sistemas, así como los valores de los parámetros de ajuste se reportan en el estudio de Jaime-Leal y Bonilla-Petriciolet, (2008) quienes aplicaron una estrategia de optimización para el ajuste de parámetros en estos sistemas. La Tabla 3.7 indica por tipo de anión la cantidad de sales estudiadas.

Tabla 3.7 *Clasificación por tipo de anión de las sales de amonio estudiadas por el modelo híbrido.*

Clasificación de familias para la base de datos de sales de amonio	
Tipo de anion	***Numero de sistemas***
Cloruros	*8*
Sales de azufre	*11*
Bromuros	*8*
Yoduros	*5*
Otras	*2*

El ajuste de los parámetros se aplicó para cada sistema, para lograr obtener los resultados del coeficiente de actividad promedio del modelo de composición local ($\gamma_{+-\ (Wilson}^{XM}$), estos resultados se utilizaron como una entrada adicional para el modelo de aprendizaje automático *XGB*, con el fin de combinar los modelos.

3.7.2 Modelo eXtreme Gradient Boosting (*XGBoost*)

Definido el modelo base con el cual se generará el modelo hibrido, se procedió a la selección de la herramienta de aprendizaje automático que se incorporará. Para ello, se decidió trabajar con el método eXtreme Gradient Boosting (*XGBoost*) dentro de la librería de Pythton® y cuya operación previamente ya ha sido descrita.

Las variables de entrada y salida del modelo *XGBoost* se describen en la Tabla 3.8, en donde se observan las seis variables de entrada por parte del algoritmo de aprendizaje automático y la entrada por parte del modelo de composición local como parte de la transferencia de aprendizaje y donde la salida corresponde al parámetro coeficiente de actividad medio iónico.

Tabla 3.8 *Clasificación de variables Modelo hibrido.*

<table>
<tr><th colspan="3">Variables de entrada</th><th>Variable de salida</th></tr>
<tr><td>Concentración molal (mol/kg solvente)</td><td>Anión (Categ.)</td><td>Catión (Categ.)</td><td rowspan="3">Coeficiente de actividad medio iónico ($\gamma_{+-\ (exp)}$)</td></tr>
<tr><td>Peso molecular Soluto</td><td>Núm. Cationes</td><td>Núm. Aniones</td></tr>
<tr><td colspan="3">Coeficiente de actividad medio iónico calculado $\gamma_{+-\ (Wilson}^{XM}$)</td></tr>
</table>

Como puede notarse en la Tabla 3.8, se, presentan dos variables categóricas, las cuales siguieron el mismo proceso de estandarización descrito en el apartado 3.2.1 con el fin de pasarlas a variables binarias y posteriormente, se procedió con la normalización de todas las variables de entrada descrita en el apartado 3.5.

También se realizó la identificación de los hiper-parametros correspondientes al modelo *XGboost*, los rangos/valores de operación y valores adecuados encontrados durante el proceso de prueba lo cual permitirá encontrar valores de predicción mejores en comparación al modelo base al operar solo éste. Estos datos se muestran en la Tabla 3.10 la cual se detalla a continuación. Se empleó la misma proporción de datos 80/20 para el proceso de entrenamiento-prueba de la herramienta *XGBoost*.

Tabla 3.9 *Datos de los hiper-parámetros y sus valores óptimos para el modelo (XGBoost).*

Hiperparámetro	***Rango/valor***	***Valor Definido***
Numero de arboles	*1000-5000*	*2000*
Profundidad por árbol	*1-20*	*10*
Tasa de aprendizaje	*0-1*	*0.005*
Peso mínimo por nodo	*1-10*	*1*
Proporción muestras para entrenar cada árbol	*0-1*	*0.3*
Proporción de variables para entrenar cada árbol	*0-1*	*0.9*
Parámetro de regularización L1	*0-1*	*0.06*
Función de perdida	*MSE, Tweddie*	*Tweddie*

La función de perdida *Tweedie* elegida, se emplea en modelos de regresión para minimizar el índice de varianza del modelo y corresponde a una medida de la dispersión de la variable de respuesta. La regresión *Tweedie* es útil para el problema en cuestión, ya que los datos de entrada presentan una distribución sesgada, además de que los resultados del modelo son menos propensos a presentar sobre-ajuste. Los resultados obtenidos en un comparativo entre el modelo base de composición local y el modelo hibrido se abordan en el apartado de resultados de este trabajo (Halder *et al*., 2021).

3.8 Métricas estadísticas de evaluación de los modelos implementados.

Una vez definidas las estrategias de aprendizaje automático y la arquitectura del método hibrido que se implementará en el proceso de modelación se procedió con el proceso de entrenamiento y prueba para analizar sus capacidades predictivas o de modelación de los parámetros termodinámicos a estudiar. Esta evaluación de la capacidad predictiva de las estrategias implementadas corresponde a un proceso de correlación entre los datos experimentales y los datos obtenidos por cada uno de los modelos por lo que existen diversas métricas estadísticas que pueden ser indicadores de la calidad o desviación de los resultados modelados y los datos experimentales. La Tabla 3.10 muestra las métricas estadísticas utilizadas en el presente estudio para medir la calidad de las predicciones por parte de los modelos desarrollados.

Tabla 3.10 *Métricas estadísticas para la evaluación de los modelos de aprendizaje automático y modelo hibrido.*

Métricas estadísticas para la evaluación de los modelos. ↓	***Ecuaciones*** ↓
Desviación porcentual (PDi)	$PDi = (\frac{y_i^0 - y_i^p}{y_i^0}) * 100\%$
Error medio porcentual (MAPE)	$MAPE = \frac{1}{N}\sum_{i}^{N} \frac{\lvert y_i^0 - y_i^p \rvert}{y_i^0} * 100\%$
Raíz del error cuadrático medio (RMSE)	$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i^0 - y_i^p)^2}$
Coeficiente de correlación	$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i^0 - y_i^p)^2}{\sum_{i=1}^{N}(y_i^0 - \bar{y}^0)^2}$
Media ($\bar{\bar{\mu}}$)	$\bar{\bar{\mu}} = \frac{1}{N}\sum_{i=1}^{N} PDi$
Desviación estándar (SD)	$SD = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\lvert PDi \rvert - \overline{\lvert \mu \rvert}}$

Tabla 3.10 cont…. *Métricas estadísticas para la evaluación de los modelos de aprendizaje automático y modelo hibrido.*

***Métricas estadísticas para la evaluación de resultados validación cruzada*↓**	***Ecuaciones*** ↓
Media $\bar{\bar{\mu}}x$	$\bar{\bar{\mu}}x = \frac{1}{k}\sum_{i=1}^{N} MAPE_{\gamma\Phi}i$
Desviación estándar (SDx)	$SDx = \sqrt{\frac{1}{k-1}\sum_{i=1}^{N}\left\|MAPE_{\gamma\Phi}i\right\| - \overline{\left\|\bar{\bar{\mu}}x\right\|}}$

3.9 Resumen del esquema metodológico.

Si bien los pasos que se siguieron para el logro del objetivo general que corresponde a generar una metodología que permita incorporar herramientas de inteligencia artificial para la predicción o modelamiento de propiedades termodinámicas en sistemas electrolíticos ya se ha descrito a detalle en los anteriores apartados, es conveniente desarrollar y presentar una estructura que generalice las etapas de este proceso. Así, la Figura 3.4 presenta mediante un diagrama de bloques la secuencia metodológica que se siguió y que define las etapas a seguir para alcanzar el objetivo antes mencionado. Este diagrama representa las etapas posteriores a la extracción de datos y que se representa a través de la Figura 3.1.

La Figura 3.4 se divide en una zona de preparación de los datos; en donde se eliminaron los datos atípicos y donde se realiza una limpieza a valores nulos y atípicos; A su vez, se codificaron las variables categóricas a variables cuantificables, como fueron el tipo de anión y catión, se eliminaron variables co-dependiente con la correlación de Spearman y finalmente, se realizó la normalización de todos los datos de entrada a los modelos de aprendizaje automático.

En otra zona del diagramase se especifican los pasos para la creación de cada uno de los modelos implementados. donde se realizó el diseño y su evaluación respectiva, logrando la mejor configuración a partir del ajuste de los híper-parámetros del modelo. Para el desarrollo del modelo hibrido, se realizaron los mismos pasos de la Figura 3.4 pero la base de datos de entrada difiere pues se integra el modelo de composición local directamente lo cual elimina la etapa de limpieza y correlación de datos.

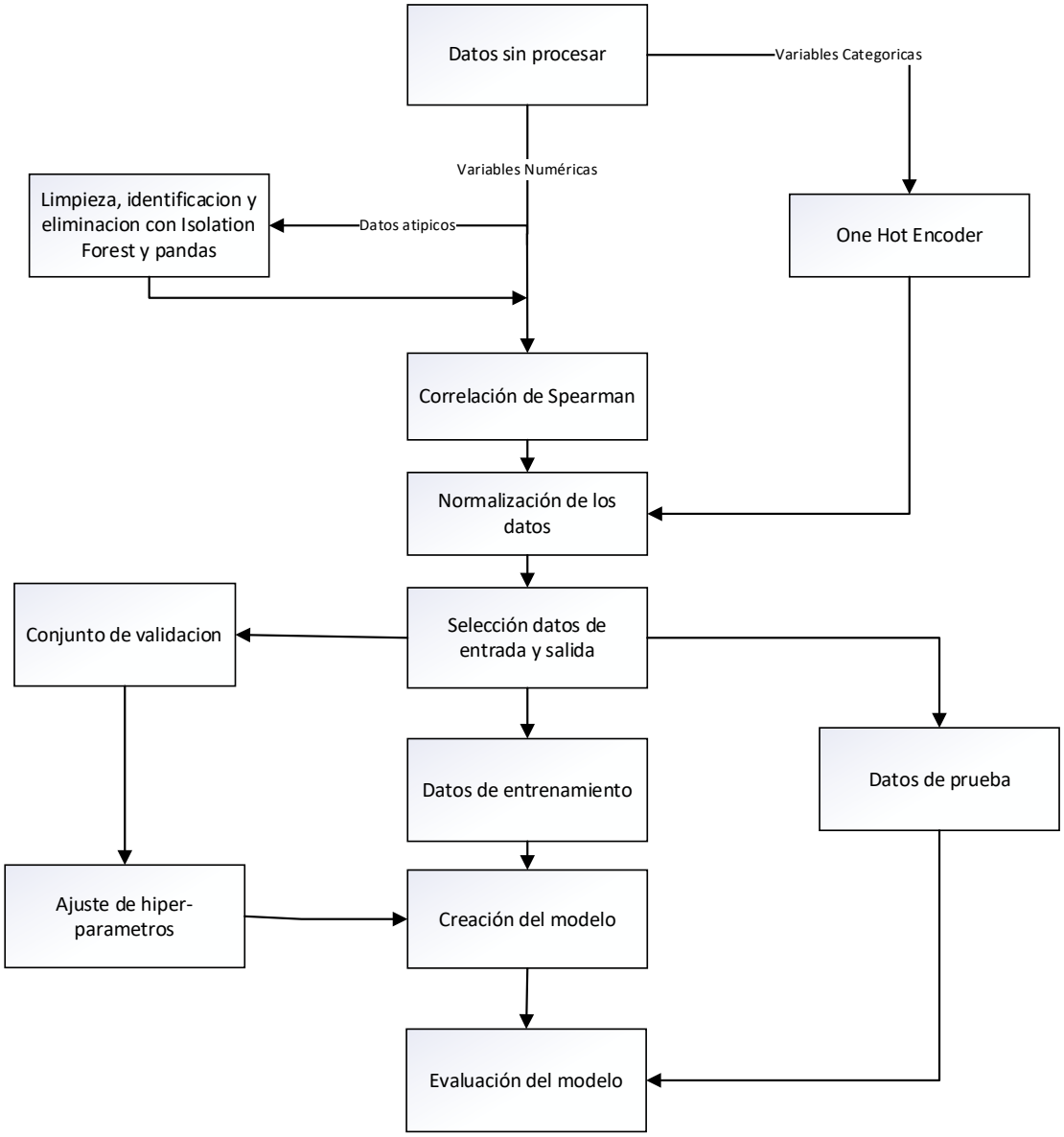


Figura 3.4 *Diagrama de flujo de la secuencia metodológica para implementar herramientas de aprendizaje automático en un proceso de predicción de datos termodinámicos en sistemas electrolíticos*

Capítulo 4. Resultados.

En este capítulo se presentan los resultados obtenidos de la aplicación de los modelos de inteligencia artificial desarrollados para la predicción de propiedades termodinámicas en sistemas electrolíticos en solución. En una primera parte se presenta la evaluación de los modelos ***ANN-MLP*** y ***GPR*** para la estimación del coeficiente de actividad medio iónico y el coeficiente osmótico del conjunto depurado total de sistemas electrolítico tomado de la base de datos *IDST* y en una segunda etapa. se realizó el mismo análisis pero sobre las diferentes familias clasificadas en el conjunto de datos. Finalmente, se presentan los resultados generados al aplicar el modelo hibrido desarrollado ***Wilson^{XM}-XGB*** para la estimación del coeficiente de actividad medio iónico para las sales de amonio consideradas.

4.1 Evaluación de los modelos *ANN-MLP* y k sobre el conjunto total de datos *IDST*.

En esta primera etapa se evaluó el desempeño y capacidad predictiva de los modelos de aprendizaje automático ***ANN-MLP*** y ***GPR*** descritos en el anterior capítulo. Ambos modelos consideraron a los 8,604 datos correspondientes a los 334 sistemas electrolíticos y que se definieron posterior a la depuración de la base de datos *IDST* inicial donde se tenían alrededor de 10,000 datos de estos sistemas. También, la evaluación de los modelos considero tanto a los datos de entrenamiento como los a datos de prueba en la relación 80/20, respectivamente y que se había establecido de inicio.

Una vez realizadas las ejecuciones de los códigos de ambos modelos, los resultados generados se recolectaron y organizaron para su procesamiento y presentación, esto con el fin de llevar a cabo un análisis de los mismos que permita efectuar un estudio comparativo entre ambos modelos y que permita establecer objetivamente el nivel de confiabilidad o robustez en la predicción de los datos termodinámicos.

Así, en un primer análisis de los resultados, cada dato modelado por el algoritmo de aprendizaje automático fue comparado con su respectivo dato experimental a través de un gráfico de correlación el cual permite de manera gráfica ver el nivel de dispersión existente entre pares de datos. Mientras más alejado este un punto que representa un par de datos (modelado *vs* experimental) de la línea diagonal, menor es la confiabilidad del algoritmo para modelar un dato. Esto permitirá de inicio evaluar visualmente la capacidad predictiva de ambos modelos.

La Figura 4.1 presenta el gráfico de correlación del modelo ***ANN-MLP*** para las variables termodinámicas coeficiente de actividad medio iónico y el coeficiente osmótico, respectivamente. En este se puede observar que el nivel de dispersión de los puntos fuera de la diagonal es considerablemente menor

para la modelación del coeficiente osmótico en comparación para la modelación del coeficiente de actividad medio iónico, donde la dispersión es mayor (datos alejados de la diagonal).

Por su parte, la Figura 4.2 muestra a través de gráficos de histogramas las desviaciones porcentuales de los errores o desviaciones del proceso de modelación por parte del modelo de aprendizaje automático. En estas figuras se corrobora que la ***ANN-MLP*** presenta mejor capacidad predictiva sobre el coeficiente osmótico ya que presenta menor desviación porcentual en comparación al modelamiento del coeficiente de actividad medio iónico, tanto para los datos de entrenamiento como de prueba.

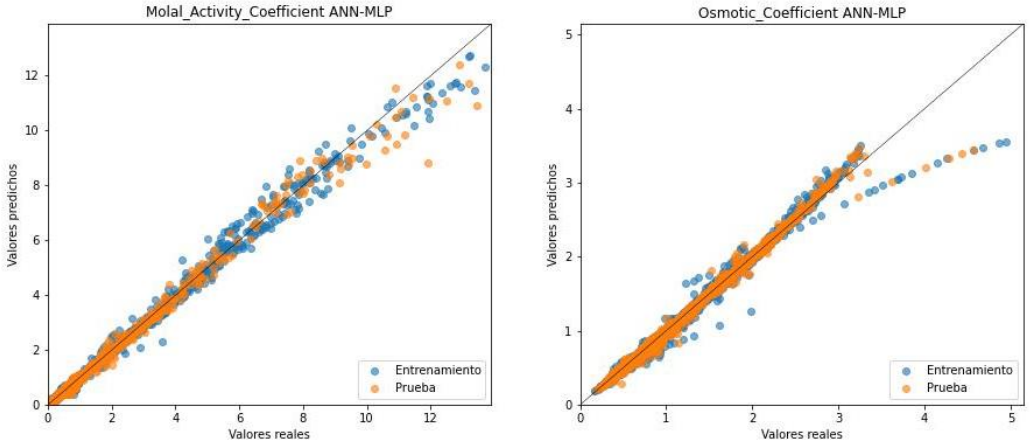


Figura 4.1 *Grafico de correlación entre datos experimentales y datos modelados por el modelo ANN-MLP para la estimación del coeficiente de actividad medio iónico y el coeficiente osmótico en sistemas electrolíticos.*

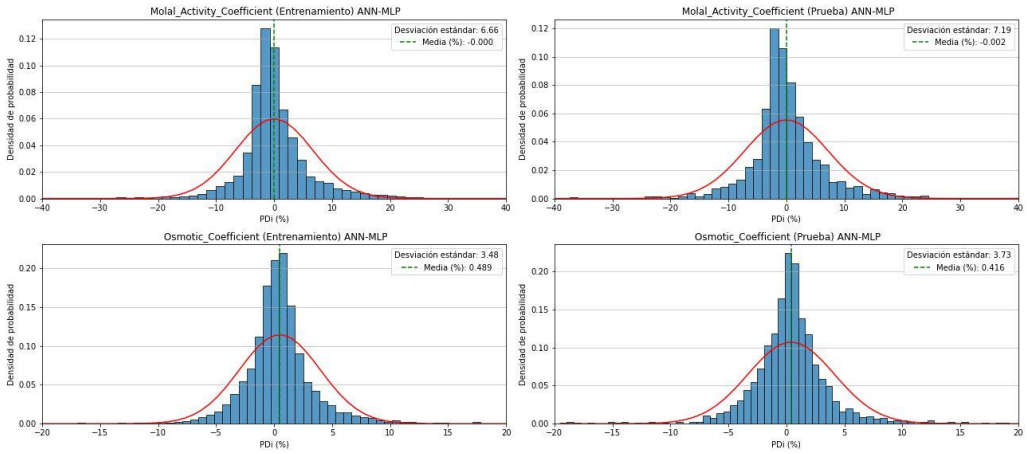


Figura 4.2 *Histograma de desviaciones porcentuales (PDi, %) para el modelo ANN-MLP para la estimación del coeficiente de actividad medio iónico y el coeficiente osmótico en sistemas electrolíticos.*

Este mismo analisis se realizó para el modelo de aprendizaje automatico ***GPR***, donde la Figura 4.3 muestra el gráfico de correlación generado para ambas variables termodinámicas, tanto para los datos de entrenamiento como los datos de prueba del modelo empleado. En esta figura se observa que el modelo utilizado presenta menor dispersion de datos (dato experimental *vs* dato modelado) para el coeficiente osmótico en comparación al nivel de dispersión de datos para el coeficiente de actividad medio iónico. Por su parte la Figura 4.4 corresponde a la presentación de histogramas de las desviaciones porcentuales de los errorres para ambas propiedades, tanto para los datos de entrenamiento como los datos de prueba; Y estos muestran que el modelo ***GPR*** presenta menor grado de desviacion de los errores para el coeficiente osmótico en contraste con el coeficiente de actividad medio iónico, por tanto se ratifica lo presentado en los gráficos de correlación respectivos.

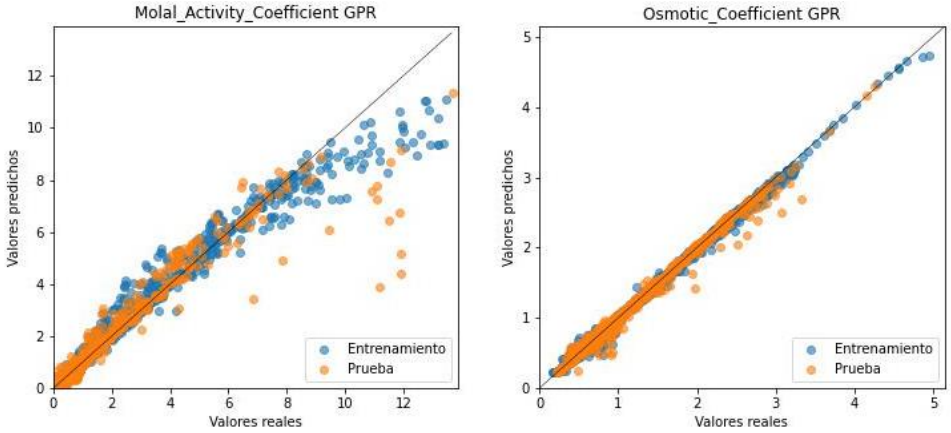


Figura 4.3 *Grafico de correlación entre datos experimentales y datos modelados por el modelo GPR para la estimación del coeficiente de actividad medio iónico y el coeficiente osmótico en sistemas electrolíticos.*

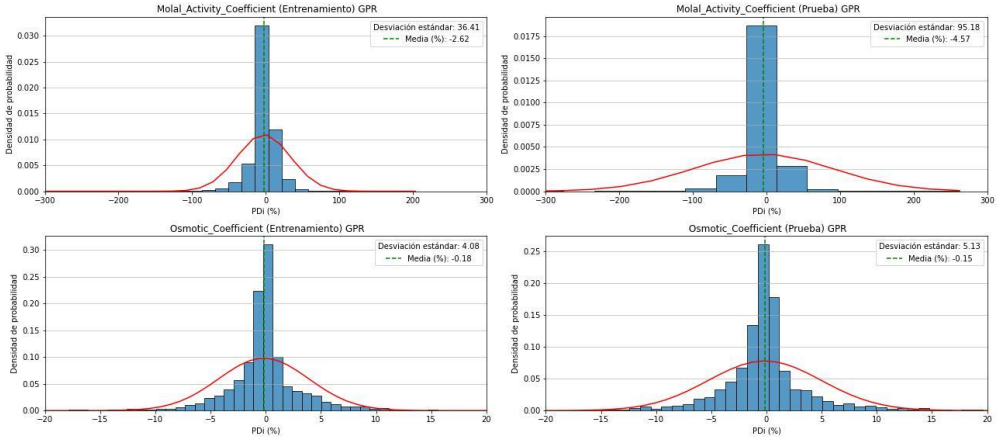


Figura 4.4 *Histograma de desviaciones porcentuales (PDi, %) para el modelo GPR para la estimación del coeficiente de actividad medio iónico y el coeficiente osmótico en sistemas electrolíticos.*

Como pudo observarse, ambos modelos de aprendizaje automático, ***ANN-MLP*** y ***GPR*** presentan mejor capacidad de modelamiento de la propiedad termodinámica definida como el coeficiente osmótico en comparación con el coeficiente de actividad para todos los sistemas electrolíticos contemplados en esta parte del estudio. Sin embargo, para poder efectuar un comparativo entre ambas estrategias y determinar cuál de estas presenta un mayor nivel de predicción de las propiedades termodinámicas estudiadas se aplicaron, sobre los errores o desviaciones de cada punto, las métricas estadísticas descritas en el apartado 3.8. Así, las Tablas 4.1 y 4.2 muestran los resultados del cálculo de las métricas para el coeficiente de actividad y coeficiente osmótico, de los datos de entrenamiento y datos de prueba tanto para el modelo ***ANN-MLP*** y el modelo ***GPR***, respectivamente.

Tabla 4.1 *Métricas estadísticas de las desviaciones generadas en la modelación del coeficiente de actividad medio iónico y coeficiente osmótico mediante la ANN-MLP para todos los sistemas electrolíticos estudiados.*

Métrica	*Propiedad termodinámica*			
	γ^{+-}, Datos Entrenamiento	γ^{+-}, Datos Prueba	Φ, Datos Entrenamiento	Φ, Datos Prueba
R^2	0.994	0.990	0.989	0.990
RMSE	0.105	0.148	**0.055**	**0.055**
MAPE	4.107	4.503	**2.163**	**2.295**
$\bar{\bar{\mu}}$	-0.0002	0.002	0.48	0.41
SD	6.66	7.19	**3.48**	**3.73**
±1%, PDi	22.74	18.83	40.54	38.12
± 5%, PDi	76.24	71.76	90.31	90.06
±10%, PDi	90.42	88.67	97.95	97.62
PDi $_{Max}$	104.92	72.04	35.91	39.14
PDi $_{Min}$	0.003	0.01	0.00016	0.0008
Media $\bar{\bar{\mu}}x$	4.67			
Desviación estándar (SDx)	0.12			

Las métricas ±1, 5 y 10 % PDi representan el porcentaje de los errores de modelación en el rango especificado

Tabla 4.2 *Métricas estadísticas de las desviaciones generadas en la modelación del coeficiente de actividad medio iónico y coeficiente osmótico mediante la GPR para todos los sistemas electrolíticos estudiados.*

Métrica	*Propiedad termodinámica*			
	γ^{+-}, Datos Entrenamiento	γ^{+-}, Datos Prueba	Φ, Datos Entrenamiento	Φ, Datos Prueba
R^2	0.973	0.913	0.996	0.99
RMSE	0.241	0.409	**0.032**	**0.052**
MAPE	12.2	16.7	**2.2**	**2.7**
$\bar{\bar{\mu}}$	-2.62	-4.57	-0.18	-0.15
SD	36.41	95.18	**4.08**	**5.13**
±1%, PDi	15	12.55	50.03	44.96
± 5%, PDi	47.88	43.51	87.93	90.06
±10%, PDi	65.08	58.86	97.06	95.26
PDi $_{Max}$	1608.3	3876.89	45.99	49.72
PDi $_{Min}$	0.00052	0.000231	0.000002	0.000007
Media $\bar{\bar{\mu}}x$	12.72			
Desviación estándar (SDx)	0.34			

Las métricas ±1, 5 y 10 % PDi representan el porcentaje de los errores de modelación en el rango especificado

Partiendo del cálculo y análisis de las métricas estadísticas mostradas en las Tablas 4.1 y 4.2 se puede establecer que en particular las métricas definidas como la *Raíz del error cuadrático medio (RMSE),* el *Error medio porcentual (MAPE)* y la *Desviación estándar (SD)* son las métricas que mejor reflejan el comportamiento de las desviaciones o errores determinados en el proceso de correlación de datos y que pueden definir de forma más clara la capacidad predictiva de los modelos de aprendizaje automático.

Con base a lo antes expuesto, se puede ratificar de manera cuantitativa que ambos modelos (***ANN-MLP*** y ***GPR***) presentan una mejor capacidad predictiva para el coeficiente osmótico pues sus respectivos valores del *RMSE*, *MAPE* y *SD*, son menores en comparación a los obtenidos para la modelación del coeficiente de actividad promedio iónico, respectivamente para cada modelo.

Por otra parte, al comparar estas métricas entre modelos, se observa que si bien las métricas *RMSE*, *MAPE* entre ambos modelos son muy similares, no así para la métrica *SD* donde el modelo ***GPR*** presenta un valor mayor en comparación al modelo ***ANN-MLP***, por lo que este último modelo presenta mejor

capacidad predictiva para el cálculo del coeficiente osmótico. De igual forma, si se realiza el mismo análisis con respecto a la capacidad predictiva de los modelos para el coeficiente de actividad promedio iónico, puede observarse que las tres métricas *RMSE*, *MAPE* y *SD*, presentan un valor considerablemente menor para el modelo de la ***ANN-MLP*** en comparación con los resultados obtenidos para el modelo ***GPR***.

4.2 Evaluación del modelo *ANN-MLP* para las diferentes familias de soluciones estudiadas.

Definido el modelo de aprendizaje automático que presenta mayor robustez para modelar alguna de las propiedades termodinámicas estudias y el cual fue la ***ANN-MLP***, se evaluó ésta sobre las distintas familias de soluciones electrolíticas estudiadas, esto para analizar la influencia o efecto que pudiese tener el tipo de solución sobre la capacidad predictiva del modelo. Los resultados y observaciones se detallan a continuación.

Las Figuras 4.5 a la 4.13 muestran los respectivos gráficos de correlación de los datos experimentales *vs* datos modelados, correspondientes a los datos de entrenamiento y de prueba del modelo, así como los histogramas de las desviaciones estándar de los errores obtenidos durante el proceso de correlación. Estos gráficos están presentados para la modelación del coeficiente de actividad y el coeficiente osmótico, respectivamente y por tipo de conjunto de sistemas electrolíticos estudiados.

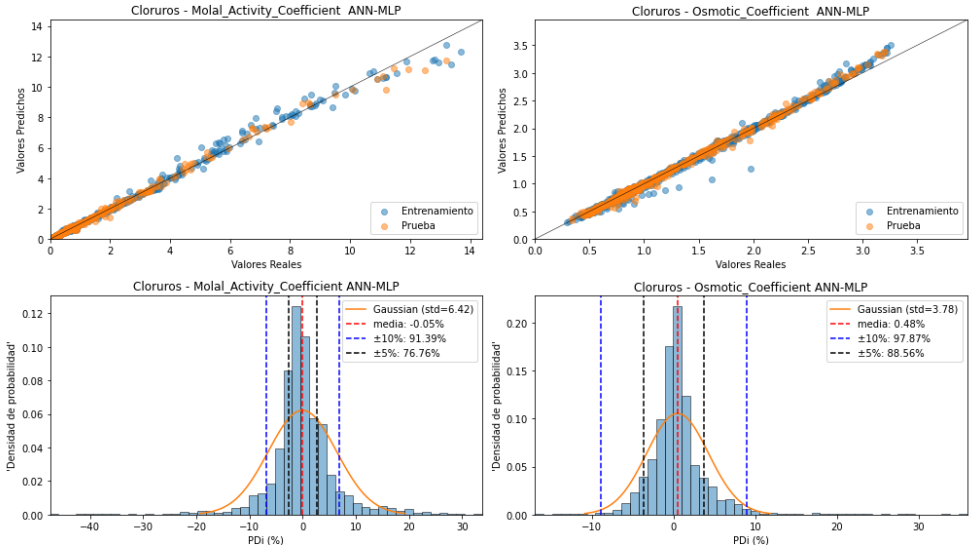


Figura 4.5 *Gráfico de Correlación entre datos experimentales vs datos modelados e Histograma de desviaciones porcentuales para el modelo ANN-MLP para la familia de* ***SALES DE CLORURO.***

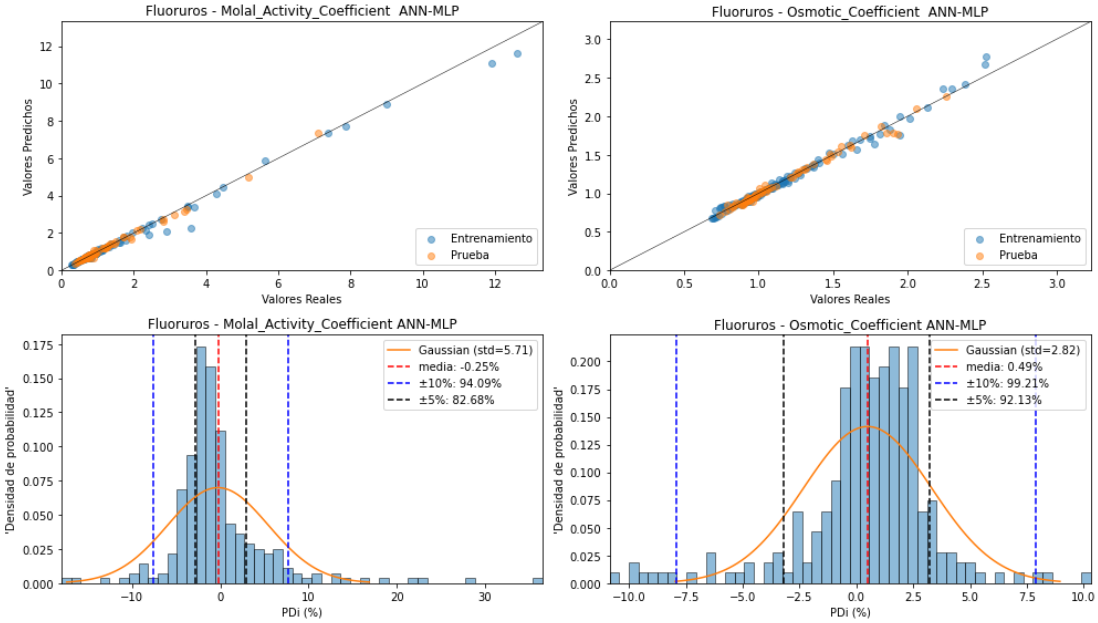


Figura 4.6 *Gráfico de Correlación entre datos experimentales vs datos modelados e Histograma de desviaciones porcentuales para el modelo ANN-MLP para la familia de* ***SALES DE FLUORURO.***

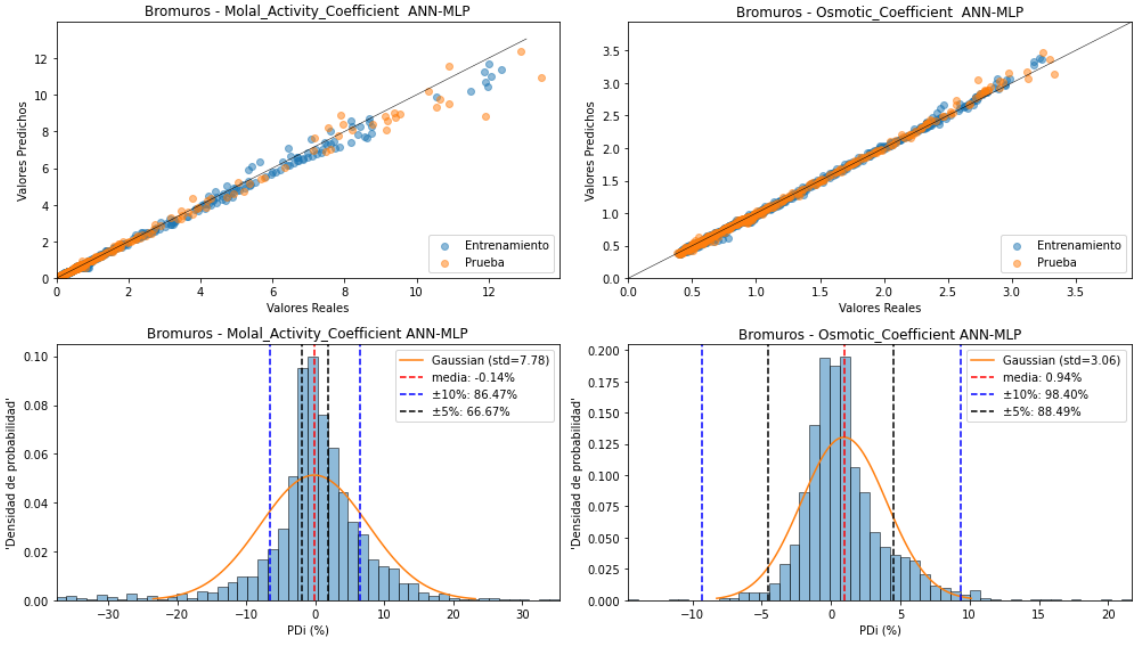


Figura 4.7 *Gráfico de Correlación entre datos experimentales vs datos modelados e Histograma de desviaciones porcentuales para el modelo ANN-MLP para la familia de* ***SALES DE BROMURO.***

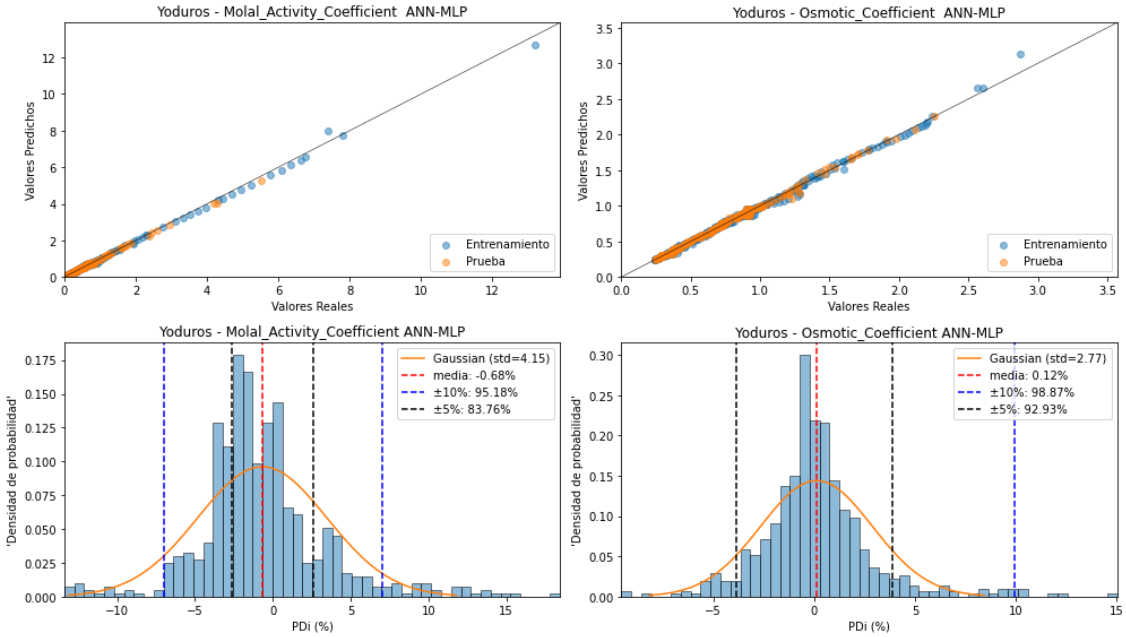


Figura 4.8 *Gráfico de Correlación entre datos experimentales vs datos modelados e Histograma de desviaciones porcentuales para el modelo ANN-MLP para la familia de* ***SALES DE IODURO.***

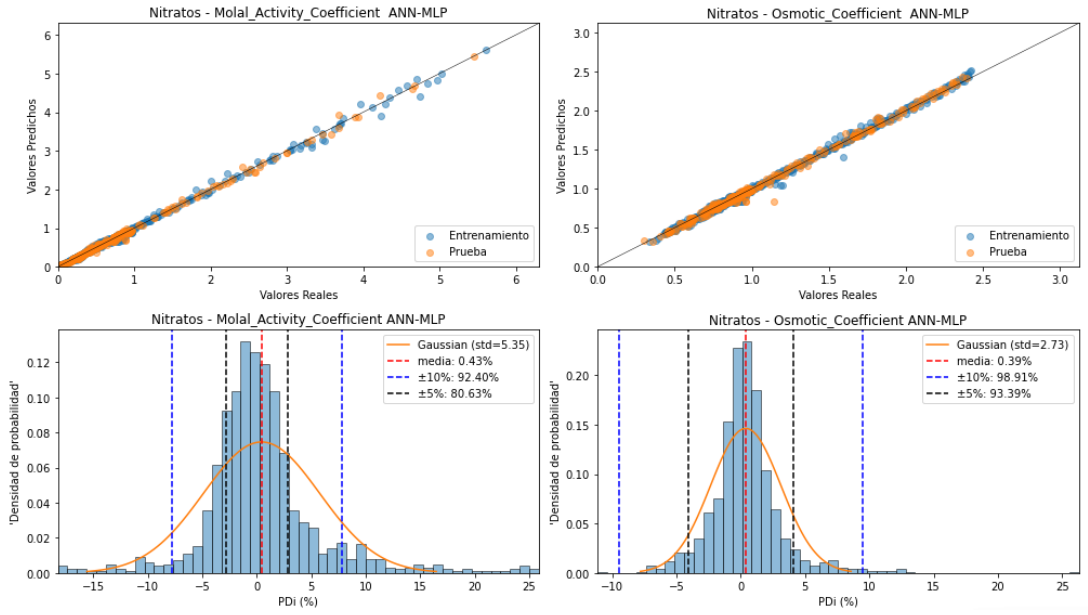


Figura 4.9 *Gráfico de Correlación entre datos experimentales vs datos modelados e Histograma de desviaciones porcentuales para el modelo ANN-MLP para la familia de* ***SALES DE NITRATO.***

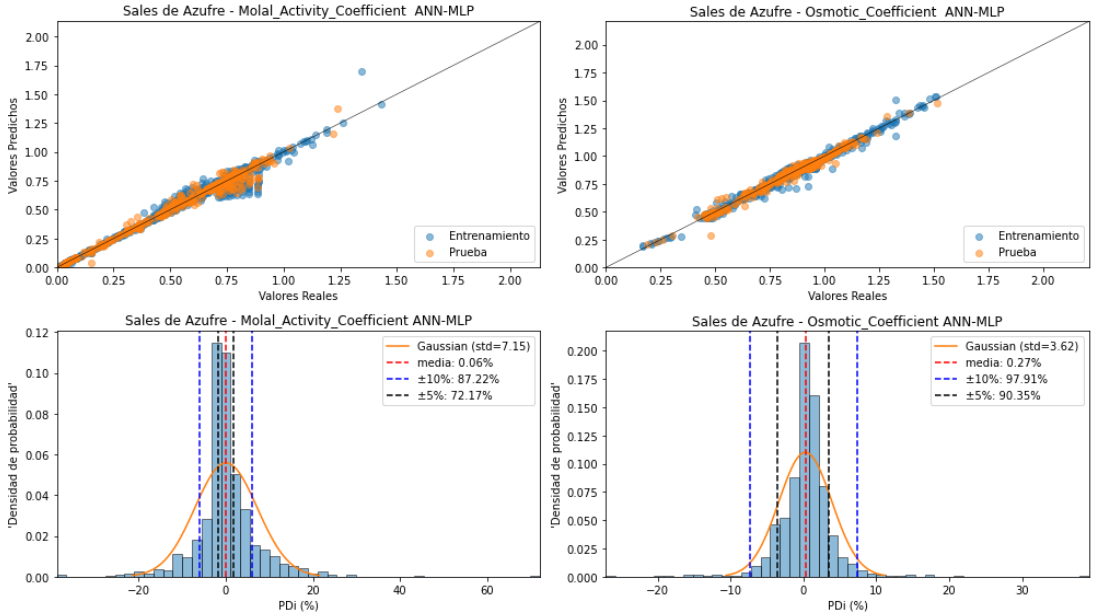


Figura 4.10 *Gráfico de Correlación entre datos experimentales vs datos modelados e Histograma de desviaciones porcentuales para el modelo ANN-MLP para la familia de* ***SALES DE AZUFRE.***

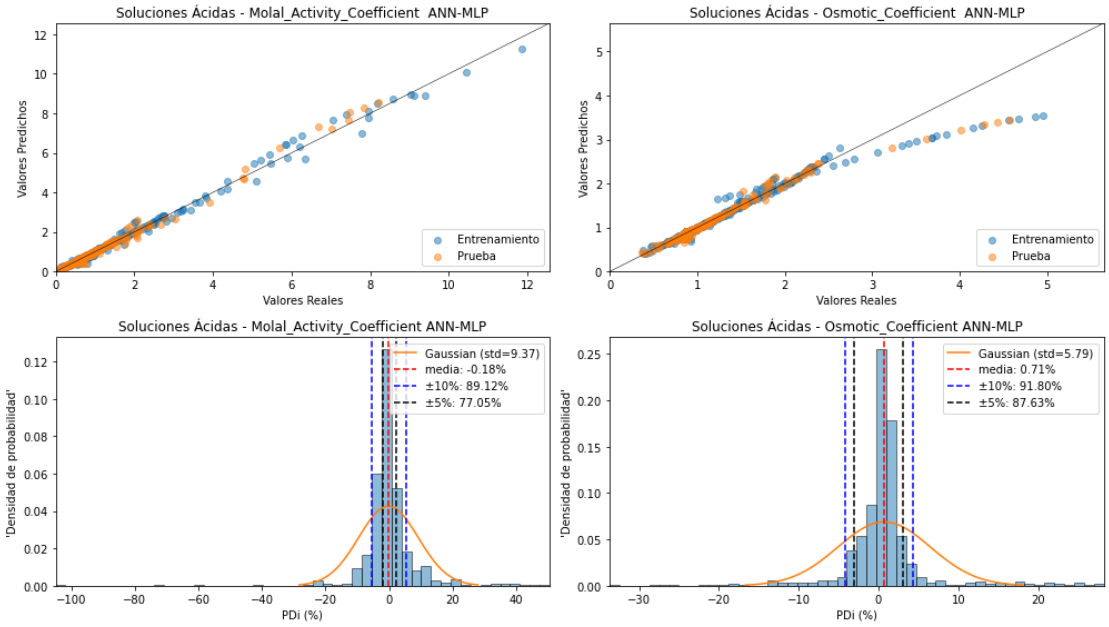


Figura 4.11 *Gráfico de Correlación entre datos experimentales vs datos modelados e Histograma de desviaciones porcentuales para el modelo ANN-MLP para la familia de* ***SOLUCIONES ACIDAS.***

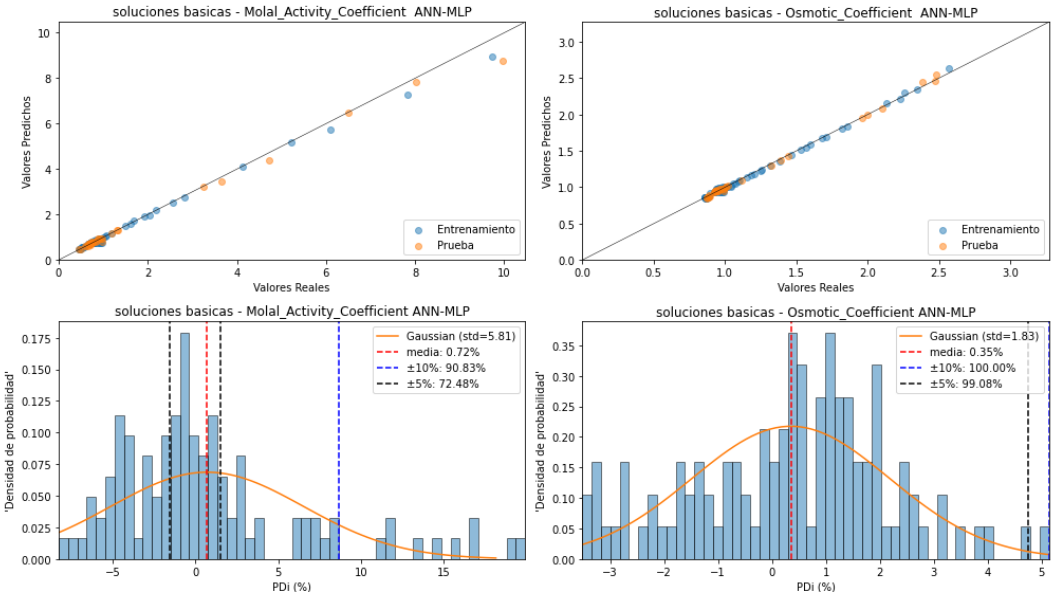


Figura 4.12 *Gráfico de Correlación entre datos experimentales vs datos modelados e Histograma de desviaciones porcentuales para el modelo ANN-MLP para la familia de* ***SOLUCIONES BÁSICAS.***

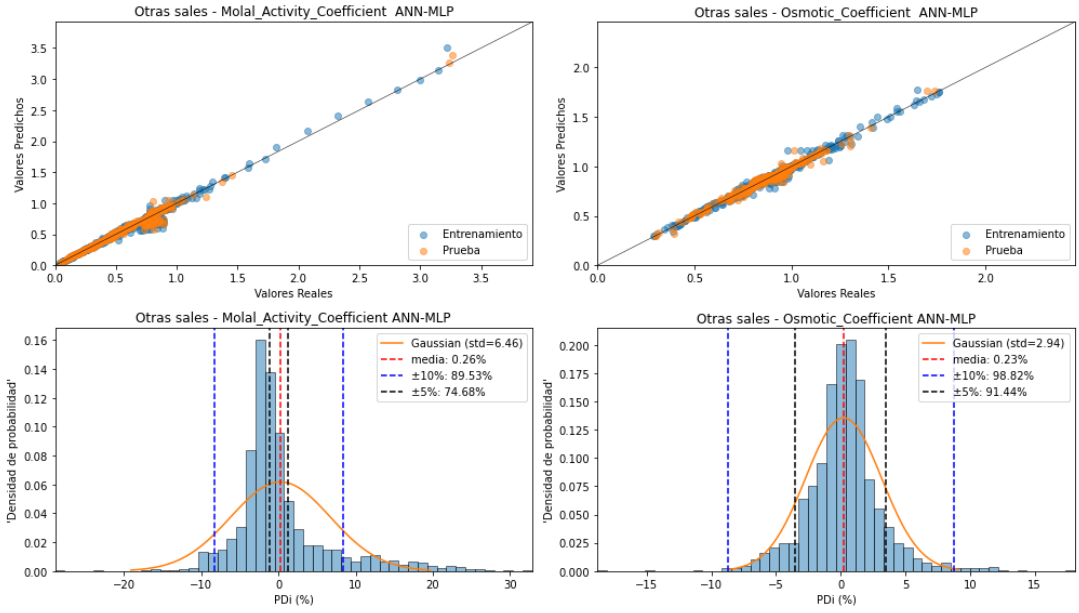


Figura 4.13 *Gráfico de Correlación entre datos experimentales vs datos modelados e Histograma de desviaciones porcentuales para el modelo ANN-MLP para la familia de* ***OTRAS SALES.***

Si bien las figuras de correlación previas ilustran la dispersión de los puntos (dato experimental *vs* dato modelado) para ambas propiedades y cada conjunto de sistemas electrolíticos estudiado, es complicado poder cuantificar directamente el grado de dispersión o desviación para poder determinar la capacidad de predicción del modelo ***ANN-MLP*** en ambas propiedades, y lo mismo ocurre con los histogramas de desviaciones estándar de los errores de cada grupo de sistemas.

Por ende, a partir de las desviaciones o errores producto del proceso de correlación de datos en cada grupo de sistema electrolítico abordado, se determinó de igual forma que en el apartado previo las métricas *Raíz del error cuadrático medio (RMSE),* el *Error medio porcentual (MAPE)* y la *Desviación estándar (SD)* con el fin de generar un análisis comparativo de forma cuantitativa sobre estas métricas para medir de forma objetiva e impacto de los grupos sobre la capacidad predictiva del modelo de aprendizaje automático abordado.

Así, la Tabla 4.3 muestra los valores de las métricas *RMSE* y *MAPE* por conjunto de sistemas electrolíticos y para cada propiedad termodinámica analizada. Por datos de entrenamiento y datos de prueba. Solo la métrica *SD* se determinó para el conjunto total de datos en cada propiedad termodinámica evaluada.

Tabla 4.3 *Métricas estadísticas por grupo de sistemas electrolíticos para el cálculo del coeficiente de actividad medio iónico y coeficiente osmótico mediante el modelo* ***ANN-MLP****.*

Grupo de sistema	*Métrica*	*Propiedad termodinámica*			
		γ_{+-}, Datos Entrenamiento	γ_{+-}, Datos Prueba	Φ, Datos Entrenamiento	Φ, Datos Prueba
Cloruros	*RMSE*	0.14	0.13	0.05	0.04
	MAPE	3.91	*4.33*	2.29	*2.39*
	SD	6.42		3.78	
Fluoruros	*RMSE*	0.16	0.08	0.04	0.03
	MAPE	3.53	3.80	2.301	2.312
	SD	5.71		2.82	
Bromuros	*RMSE*	0.14	0.26	0.03	0.04
	MAPE	5.05	5.04	2.15	2.18
	SD	7.78		3.06	
Ioduros	*RMSE*	0.05	0.04	0.03	0.02
	MAPE	3.07	2.99	1.82	1.93

Tabla 4.3 (Continuación…) *Métricas estadísticas por grupo de sistemas electrolíticos para el cálculo del coeficiente de actividad medio iónico y coeficiente osmótico mediante el modelo* ***ANN-MLP****.*

Ioduros	*SD*	4.15		2.77	
Nitratos	*RMSE*	0.04	0.04	0.03	0.03
	MAPE	3.34	3.81	1.73	2.04
	SD	5.35		2.73	
Azufre	*RMSE*	0.04	0.04	0.03	0.03
	MAPE	4.40	4.62	2.25	2.34
	SD	7.15		3.62	
Soluciones Básicas	*RMSE*	0.13	0.23	0.02	0.02
	MAPE	4.23	3.60	1.58	1.38
	SD	5.81		1.83	
Soluciones Acidas	*RMSE*	0.13	0.13	0.16	0.16
	MAPE	4.33	5.52	2.95	3.23
	SD	9.37		5.79	
Otras sales	*RMSE*	0.05	0.05	0.03	0.03
	MAPE	4.21	4.35	2.00	2.11
	SD	6.46		2.94	

Analizando los datos de la tabla previa, de un primer análisis se tiene que independientemente de la métrica que se tome, el modelo ***ANN-MLP*** siempre presenta mejor capacidad predictiva con respecto al coeficiente osmótico seguido del coeficiente de actividad medio iónico y esto va en concordancia con lo obtenido cuando se analizaron los datos en general sin considerar clasificación alguna.

Para poder analizar el efecto que tiene cada grupo de sistema electrolítico sobre la capacidad predictiva del modelo ***ANN-MLP*** se analizó la métrica *SD* ya que esta refleja que tan alejado están los puntos de desviación respecto a un error nulo. Así, para el cálculo del coeficiente osmótico la capacidad predictiva del modelo presentará el siguiente orden: ***Sales básicas > Nitratos ≈ Ioduros ≈ Fluoruros ≈ Otras sales*** > ***Bromuros*** > ***Azufre ≈ Cloruros > Sales ácidas***. Por otra parte, si se requiere la determinación del coeficiente de actividad medio iónico la capacidad de modelación por parte del modelo será. ***Ioduros > Nitratos ≈ Fluoruros ≈ Sales básicas > Cloruros ≈ Otras sales > Azufre*** > **Bromuro > Sales ácidas.** Por tanto, el tipo de sistema electrolítico si influye sobre la capacidad predictiva del modelo de aprendizaje automático.

4.3 Evaluación del modelo híbrido *Wilson^{XM}-XGB* para modelar el coeficiente de actividad en sistemas electrolíticos.

Tal como se mencionó en el apartado de la Metodología, si bien ya existen modelos desarrollados que buscan predecir alguna propiedad termodinámica y los cuales han dado resultados aceptables para los sistemas en los que se han abordado, estos modelos aun presentan limitaciones en su capacidad predictiva fuera de ciertos rangos operativos o al aplicarlos a sistemas complejos de los cuales no se conoce o se tiene información limitada de los mismos. Ante esto, se busca trabajar sobre los modelos actuales ya sea modificando estos o bien, hibridandolos a través de incorporar alguna otra estrategia que mejore su robustez.

En este último apartado, se presentan los resultados generados por el modelo de composición local de ***Wilson^{XM}*** y el modelo ***Wilson^{XM}-XGB*** generado al hibridar el primero con el modelo de aprendizaje automático ***XGB*** para predecir exclusivamente el coeficiente de actividad medio iónico sobre un conjunto de 34 sistemas electrolíticos de sales de amonio y realizar su comparativo. Asimismo, el análisis de los resultados sigue el mismo esquema que en los dos apartados previos.

La Figura 4.14 muestra el gráfico de correlación de datos donde se ve la dispersión de puntos (dato experimental *vs* dato modelado) tanto para el modelo original como para el modelo de Wilson^{XM} como para el modelo hibrido ***Wilson^{XM}-XGB***, y donde se observa que este último presenta una mejor correlación respecto al modelo no hibrido.

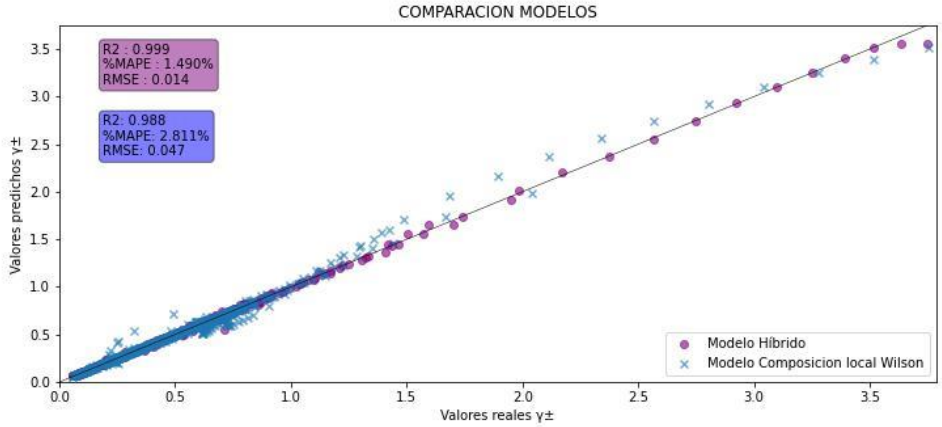


Figura 4.14 *Grafico de Correlación entre datos experimentales vs datos modelados por el modelo* ***Wilson^{XM}*** *y el modelo* ***Wilson^{XM}-XGB*** *sobre un conjunto de sistemas electrolíticos de-sales de amonio.*

Con el fin de poder realizar un análisis y comparativo cuantitativo entre ambos modelos se determinó de igual forma que en el apartado previo la *Raíz del error cuadrático medio (RMSE),* el *Error medio porcentual (MAPE)* y la *Desviación estándar (SD)* de las desviaciones o errores generados durante el proceso de correlación de datos de los sistemas estudiados, y donde la Tabla 4.4 presenta dichos resultados de estas métricas.

Así, con base a los resultados presentados en la Tabla 4.4, puede observarse que las tres métricas consideradas (*RMSE, MAPE y SD*) presentan un valor mínimo considerable para el modelo ***Wilson^{XM}-XGB*** en comparación a los valores obtenidos para el modelo ***Wilson^{XM}***, lo cual indica un mayor grado de desviaciones para este último modelo. Así, se puede establecer y confirmar que el modelo hibrido tiene mejor capacidad predictiva para el coeficiente de actividad medio iónico en este tipo de sistemas electrolítico de sales de amonio.

Tabla 4.4. *Métricas estadísticas obtenidas para los sistemas electrolíticos de sales de amonio generadas en el cálculo del coeficiente de actividad medio iónico mediante modelo híbrido* ***(Wilson^{XM}-XGB)*** *y el modelo* ***Wilson^{XM}***.

Modelo	*Métrica*	*Propiedad termodinámica*		
		γ^{+-}, Datos Entrenamiento	γ^{+-}, Datos Prueba	γ^{+-}, Datos Totales
Wilson^{XM}	*RMSE*	*…*	*…*	*0.046*
	MAPE	*…*	*…*	*2.800*
	SD	*…*	*…*	*5.91*
Wilson^{XM}-XGB	*RMSE*	*0.013*	*0.018*	*…*
	MAPE	*1.397*	*1.861*	*…*
	SD	*2.73*	*3.73*	*…*

Con el fin de ilustrar como el modelo hibrido ***Wilson^{XM}-XGB*** presenta una mejor capacidad predictiva en comparación al modelo original ***Wilson^{XM}***, se tomaron aleatoriamente cuatro sistemas de las sales de amonio estudiadas y se comparó su capacidad predictiva respecto a los datos experimentales de cada sistema, donde la Figura 4.15 muestra este comparativo. Puede observarse que en algunos sistemas ambos modelos presentan una competencia similar en sus capacidades de predicción, pero en otros sistemas donde el modelo ***Wilson^{XM}*** presenta cierta dificultad para ajustarse con los datos experimentales, el modelo ***Wilson^{XM}-XGB*** presenta una mejor capacidad de predicción en relación a los mismos datos experimentales. Estos resultados van en conformidad con lo observado anteriormente basándose en los datos cuantitativo de las métricas estadística calculadas.

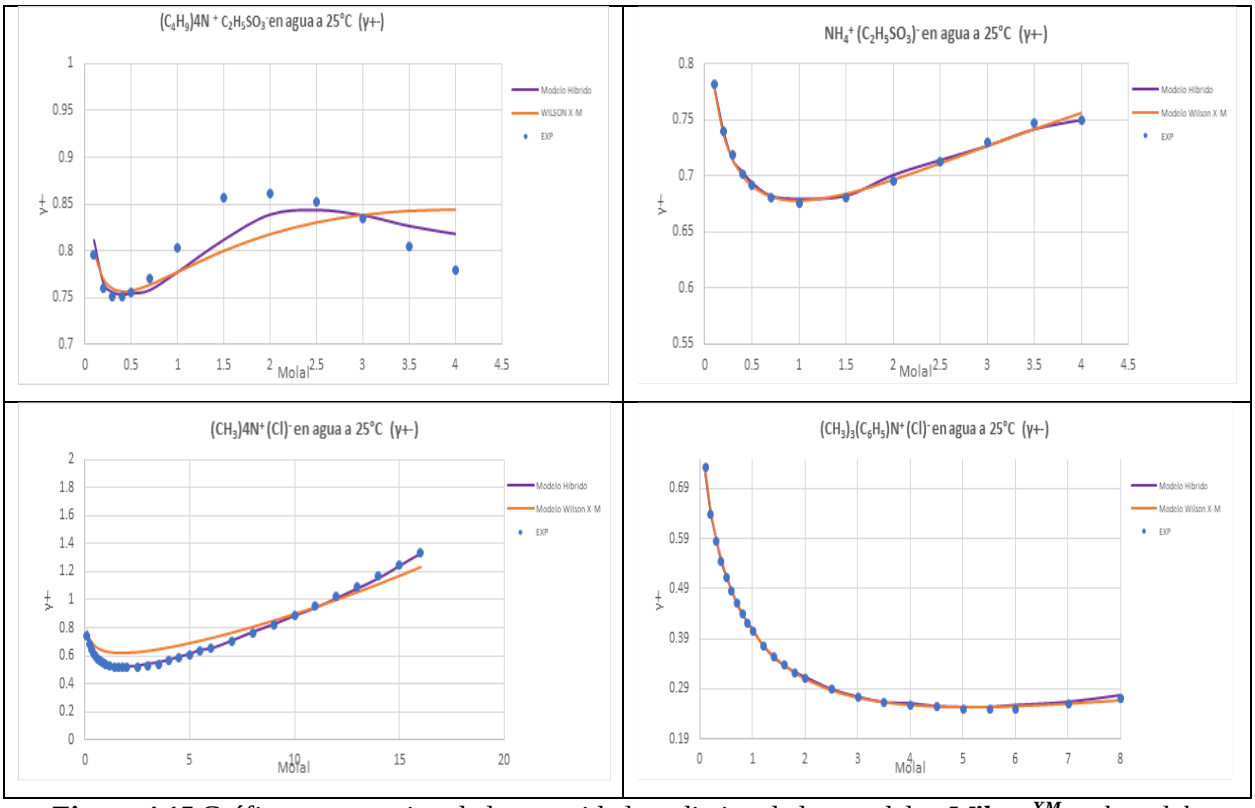


Figura 4.15 Gráfico comparativo de la capacidad predictiva de los modelos ***Wilson^{XM}*** *y el modelo* ***Wilson^{XM}-XGB*** sobre cuatro sistemas electrolíticos de sales de amonio estudiadas.

Finalmente, concentrando los resultados y análisis de los tres apartados previos se puede establecer un conjunto de observaciones y aportaciones que se presentan en el siguiente capítulo, junto con un conjunto de recomendaciones que desembocan de dicho análisis

Capítulo 5. Conclusiones y recomendaciones.

Con base en los resultados generados en el presente estudio y que permitieron cumplir con los objetivos particulares y general, se puede establecer de manera general los siguientes puntos.

A partir de la aplicación de la teoría y metodología correspondiente a la ciencia de datos, se efectuó una adecuada extracción, depuración y clasificación de datos correspondientes a los distintos sistemas electrolíticos abordados en este estudio. Particularmente, se incorporaron dos variables categóricas las cuales requirieron un proceso de transformación para integrarlas al resto de los datos de entrada-salida de los modelos de aprendizaje automático.

Se realizó la codificación en Python® de dos modelos de aprendizaje automático, siendo estos las ***ANN-MLP*** y ***GPR*** y ambos fueron evaluados y comparadas sus capacidades para la predicción del coeficiente de actividad medio iónico y el coeficiente osmótico de los sistemas electrolíticos en solución estudiados. En particular, ambos modelos presentan una mayor capacidad para la modelación del coeficiente osmótico con respecto al coeficiente de actividad medió iónico, pero es el modelo ***ANN-MLP*** aquel que presenta mayor robustez predictiva con respecto al modelo ***GPR***, Por tanto, es el modelo ***ANN-MLP*** aquel que se recomienda emplear para este tipo de proceso de modelación de datos termodinámicos en sistemas electrolíticos en solución.

Definido el modelo de aprendizaje automático que favorece la predicción de datos (***ANN-MLP***), este se empleó para realizar un estudio comparativo entre grupos de sistemas electrolíticos los cuales se clasificaron en nueve categorías distintas. De esta forma, y de acuerdo a los resultados, para la modelación del coeficiente osmótico la capacidad predictiva del modelo presentará el siguiente orden: ***Sales básicas*** > ***Nitratos*** ≈ ***Ioduros*** ≈ ***Fluoruros*** ≈ ***Otras sales*** > ***Bromuros*** > ***Azufre*** ≈ ***Cloruros*** > ***Sales ácidas***. Si se requiere la determinación del coeficiente de actividad medio iónico la capacidad de modelación por parte del modelo será. ***Ioduros*** > ***Nitratos*** ≈ ***Fluoruros*** ≈ ***Sales básicas*** > ***Cloruros*** ≈ ***Otras sales*** > ***Azufre*** > **Bromuro** > ***Sales ácidas*****.** Por tanto, el tipo de sistema electrolítico si presenta un efecto sobre la capacidad predictiva del modelo de aprendizaje automático que se utiliza.

Para la última etapa de esta investigación donde se realizó la integración de un modelo de composición local denominado ***WILSON^{XM}*** y el modelo ***XGB*** para generar el modelo hibrido ***Wilson^{XM}-XGB*** para la predicción del coeficiente de actividad medio iónico sobre un grupo de sistemas electrolíticos de sales de amonio en solución y con base a los resultados obtenidos, se puede establecer que el modelo hibrido presenta una mejor capacidad de modelación de la propiedad que con respecto al modelo de composición local y esto se evidencia al presentarse un menor error y desviación estándar de las desviaciones generadas

durante el proceso de correlación entre los datos experimentales y los datos modelados por ambos modelos evaluados.

De esta manera, se puede establecer que los objetivos específicos y el objetivo general que correspondió al desarrollo de una metodología que acople una estrategia de aprendizaje automático junto con un modelo de composición local para generar un modelo hibrido para la modelación y/o predicción de propiedades termodinámicas en sistemas electrolíticos en solución se cumplió.

RECOMENDACIONES

1. Analizar y evaluar distintas configuraciones de las redes neuronales artificiales para este tipo de sistemas, esto con el fin de determinar aquella estructura de la red que favorece el proceso de modelación de las propiedades termodinámicas en sistemas electrolíticos.
2. Realizar un estudio a profundidad sobre el efecto que tienen ciertos grupos de sistemas electrolíticos sobre la capacidad de modelación o predicción que presentan los modelos de aprendizaje automático.
3. Continuar con el estudio del proceso de hibridación de modelos para localizar aquel que presenta mayor capacidad predictiva de propiedades termodinámicas en la mayoría de los sistemas electrolíticos en solución.

Capítulo 6. Bibliografía.

Asensio-Delgado, S., Pardo, F., Zarca, G., & Urtiaga, A. (2022). Machine learning for predicting the solubility of high-GWP fluorinated refrigerants in ionic liquids. *Journal of Molecular Liquids*, *367*, 120472. https://doi.org/10.1016/j.molliq.2022.120472

Attias, R., Dlugatch, B., Blumen, O., Shwartsman, K., Salama, M., Shpigel, N., & Sharon, D. (2022). Determination of Average Coulombic Efficiency for Rechargeable Magnesium Metal Anodes in Prospective Electrolyte Solutions. *ACS Applied Materials & Interfaces*, *14*(27), 30952–30961. https://doi.org/10.1021/acsami.2c08008

Azadfar, R., Shaabanzadeh, M., Hashemi-Moghaddam, H., & Nafchi, A. M. (2022). Estimation of Heat Capacity of 143 Pure Ionic Liquids Using Artificial Neural Network. *International Journal of Thermophysics*, *43*(6), 81. https://doi.org/10.1007/s10765-022-03003-2

Belvèze, L. S., Brennecke, J. F., & Stadtherr, M. A. (2004). Modeling of Activity Coefficients of Aqueous Solutions of Quaternary Ammonium Salts with the Electrolyte-NRTL Equation. *Industrial & Engineering Chemistry Research*, *43*(3), 815–825. https://doi.org/10.1021/ie0340701

Benimam, H., Si-Moussa, C., Laidi, M., & Hanini, S. (2020). Modeling the activity coefficient at infinite dilution of water in ionic liquids using artificial neural networks and support vector machines. *Neural Computing and Applications*, *32*(12), 8635–8653. https://doi.org/10.1007/s00521-019-04356-w

Bollas, G. M., Chen, C. C., & Barton, P. I. (2008). Refined electrolyte-NRTL model: Activity coefficient expressions for application to multi-electrolyte systems. *AIChE Journal*, *54*(6), 1608–1624. https://doi.org/10.1002/aic.11485

Chapman, W. G., Gubbins, K. E., Jackson, G., & Radosz, M. (1989). SAFT: Equation-of-state solution model for associating fluids. *Fluid Phase Equilibria*, *52*, 31–38. https://doi.org/10.1016/0378-3812(89)80308-5

Chapman, W. G., Gubbins, K. E., Jackson, G., & Radosz, M. (1990). New reference equation of state for associating liquids. *Industrial & Engineering Chemistry Research*, *29*(8), 1709–1721. https://doi.org/10.1021/ie00104a021

Chen, C.-C., Bokis, C. P., & Mathias, P. (2001). Segment-based excess Gibbs energy model for aqueous organic electrolytes. *AIChE Journal*, *47*(11), 2593–2602. https://doi.org/10.1002/aic.690471122

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Chen, T.-C., Hammid, A. T., Akbarov, A. N., Shariati, K., Dinari, M., & Ali, M. S. (2022). Estimating the Physical Properties of Nanofluids Using a Connectionist Intelligent Model Known as Gaussian Process Regression Approach. *International Journal of Chemical Engineering*, *2022*, e1017341. https://doi.org/10.1155/2022/1017341

Chen, Y., Liang, X., & Kontogeorgis, G. M. (2023). Artificial neural network modeling on the polymer-electrolyte aqueous two-phase systems involving biomolecules. *Separation and Purification Technology*, *306*, 122624. https://doi.org/10.1016/j.seppur.2022.122624

Dajnowicz, S., Agarwal, G., Stevenson, J. M., Jacobson, L. D., Ramezanghorbani, F., Leswing, K., Friesner, R. A., Halls, M. D., & Abel, R. (2022). High-Dimensional Neural Network Potential for Liquid Electrolyte Simulations. *The Journal of Physical Chemistry B*, *126*(33), 6271–6280. https://doi.org/10.1021/acs.jpcb.2c03746

Delač Marion, I., Grgičin, D., Salamon, K., Bernstorff, S., & Vuletić, T. (2015). Polyelectrolyte Composite: Hyaluronic Acid Mixture with DNA. *Macromolecules*, *48*(8), 2686–2696. https://doi.org/10.1021/ma502090x

Derbenev, I. N., Filippov, A. V., Stace, A. J., & Besley, E. (2018). Electrostatic interactions between charged dielectric particles in an electrolyte solution: Constant potential boundary conditions. *Soft Matter*, *14*(26), 5480–5487. https://doi.org/10.1039/C8SM01068D

Deringer, V. L., Bartók, A. P., Bernstein, N., Wilkins, D. M., Ceriotti, M., & Csányi, G. (2021). Gaussian Process Regression for Materials and Molecules. *Chemical Reviews*, *121*(16), 10073–10141. https://doi.org/10.1021/acs.chemrev.1c00022

Ding, Z., & Fei, M. (2013). An Anomaly Detection Approach Based on Isolation Forest Algorithm for Streaming Data using Sliding Window. *IFAC Proceedings Volumes*, *46*(20), 12–17. https://doi.org/10.3182/20130902-3-CN-3020.00044

Economou, I. (2001). Statistical Associating Fluid Theory: A Successful Model for the Calculation of Thermodynamic and Phase Equilibrium Properties of Complex Fluid Mixtures. *Industrial & Engineering Chemistry Research*, *41*. https://doi.org/10.1021/ie0102201

Feeley, B. P., Overton, M. A., Galloway, M. M., Lecrivain, T. J., & Wilson, A. D. (2021). Idaho database of solution thermodynamics. *Journal of Molecular Liquids*, *338*, 116574. https://doi.org/10.1016/j.molliq.2021.116574

Georgios M. Kontogeorgis. (2010). The Statistical Associating Fluid Theory (SAFT). In *Thermodynamic Models for Industrial Applications* (pp. 221–259). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470747537.ch8

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems* (2nd ed.).

Gil-Villegas, A., Galindo, A., Whitehead, P., Mills, S., Jackson, G., & Burgess, A. (1997). Statistical Associating Fluid Theory for Chain Molecules with Attractive Potentials of Variable Range. *The Journal of Chemical Physics*, *106*, 4168–4186. https://doi.org/10.1063/1.473101

Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T. D., Adams, R. P., & Aspuru-Guzik, A. (2018). Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Central Science*, *4*(2), 268–276. https://doi.org/10.1021/acscentsci.7b00572

Gröls, J. R., & Castro-Dominguez, B. (2021). Mechanochemical co-crystallization: Insights and predictions. *Computers & Chemical Engineering*, *153*, 107416. https://doi.org/10.1016/j.compchemeng.2021.107416

Gubbins, K. E. (2016). Perturbation theories of the thermodynamics of polar and associating liquids: A historical perspective. *Fluid Phase Equilibria*, *416*, 3–17. https://doi.org/10.1016/j.fluid.2015.12.043

Haghtalab, A., & Peyvandi, K. (2009). Electrolyte-UNIQUAC-NRF model for the correlation of the mean activity coefficient of electrolyte solutions. *Fluid Phase Equilibria*, *281*(2), 163–171. https://doi.org/10.1016/j.fluid.2009.04.013

Halder, A., Mohammed, S., Chen, K., & Dey, D. K. (2021). Spatial Tweedie exponential dispersion models. *Scandinavian Actuarial Journal*, *2021*(10), 1017–1036. https://doi.org/10.1080/03461238.2021.1921017

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer. https://doi.org/10.1007/978-0-387-84858-7

Hüllen, G., Zhai, J., Kim, S. H., Sinha, A., Realff, M. J., & Boukouvala, F. (2020). Managing uncertainty in data-driven simulation-based optimization. *Computers & Chemical Engineering*, *136*, 106519. https://doi.org/10.1016/j.compchemeng.2019.106519

Jaime-Leal, J. E., & Bonilla-Petriciolet, A. (2008). Correlation of Activity Coefficients in Aqueous Solutions of Ammonium Salts Using Local Composition Models and Stochastic Optimization Methods. *Chemical Product and Process Modeling*, *3*(1). https://doi.org/10.2202/1934-2659.1237

Karimzadeh, Z., & Hosseini, S. A. A. (2019). Mixed Electrolyte in Mixed Solvent: Activity Coefficient Measuring and Modeling for the HCl + NaCl + Methanol + Water System. *Journal of Chemical & Engineering Data*. https://doi.org/10.1021/acs.jced.8b01033

Kim, K., Lee, D., & Essa, I. (2011). Gaussian process regression flow for analysis of motion trajectories. *2011 International Conference on Computer Vision*, 1164–1171. https://doi.org/10.1109/ICCV.2011.6126365

Kontogeorgis, G. (2010). *Thermodynamic Models for Industrial Applications: From Classical and Advanced Mixing Rules to Association Theories | Wiley*. https://www.wiley.com/en-us/Thermodynamic+Models+for+Industrial+Applications%3A+From+Classical+and+Advanced+Mixing+Rules+to+Association+Theories-p-9780470697269

Kontogeorgis, G. M., Tsivintzelis, I., von Solms, N., Grenner, A., Bøgh, D., Frost, M., Knage-Rasmussen, A., & Economou, I. G. (2010). Use of monomer fraction data in the parametrization of association theories. *Fluid Phase Equilibria*, *296*(2), 219–229. https://doi.org/10.1016/j.fluid.2010.05.028

Kuramochi, H., Osako, M., Kida, A., Nishimura, K., Kawamoto, K., Asakuma, Y., Fukui, K., & Maeda, K. (2005). Determination of Ion-Specific NRTL Parameters for Predicting Phase Equilibria in Aqueous Multielectrolyte Solutions. *Industrial & Engineering Chemistry Research*, *44*(9), 3289–3297. https://doi.org/10.1021/ie049377u

Lach, A., André, L., Guignot, S., Christov, C., Henocq, P., & Lassin, A. (2018). A Pitzer Parametrization To Predict Solution Properties and Salt Solubility in the H–Na–K–Ca–Mg–NO3–H2O System at 298.15 K. *Journal of Chemical & Engineering Data*, *63*(3), 787–800. https://doi.org/10.1021/acs.jced.7b00953

Lee, L., Sun, S., & Lin, C. (2008). Predictions of thermodynamic properties of aqueous single-electrolyte solutions with the two-ionic-parameter activity coefficient model. *Fluid Phase Equilibria*, *264*(1–2), 45–54. https://doi.org/10.1016/j.fluid.2007.10.015

Liu, Y., Hong, W., & Cao, B. (2019). Machine learning for predicting thermodynamic properties of pure fluids and their mixtures. *Energy*, *188*, 116091. https://doi.org/10.1016/j.energy.2019.116091

Lu, H., Hu, X., Cao, B., Chai, W., & Yan, F. (2019). Prediction of liquidus temperature for complex electrolyte systems Na3AlF6-AlF3-CaF2-MgF2-Al2O3-KF-LiF based on the machine learning methods. *Chemometrics and Intelligent Laboratory Systems*, *189*, 110–120. https://doi.org/10.1016/j.chemolab.2019.03.015

Lu, X., & Maurer, G. (1993). Model for describing activity coefficients in mixed electrolyte aqueous solutions. *AIChE Journal*, *39*(9), 1527–1538. https://doi.org/10.1002/aic.690390912

Maribo-Mogensen, B. (2014). *Development of an Electrolyte CPA Equation of state for Applications in the Petroleum and Chemical Industries*. AIChE.

Mazloumi, S. H. (2016). Correlation of the mean activity coefficient of aqueous electrolyte solutions using an equation of state. *Chinese Journal of Chemical Engineering*, *24*(10), 1456–1463. https://doi.org/10.1016/j.cjche.2016.04.002

Michelsen, M. L., & Mollerup, J. M. (2007). *Thermodynamic models: Fundamentals & computational aspects*. Tie-Line Publications.

Myers, J. A., Sandler, S. I., & Wood, R. H. (2002). An Equation of State for Electrolyte Solutions Covering Wide Ranges of Temperature, Pressure, and Composition. *Industrial & Engineering Chemistry Research*, *41*(13), 3282–3297. https://doi.org/10.1021/ie011016g

Nezbeda, I. (2020). On molecular-based equations of state: Perturbation theories, simple models, and SAFT modeling. *Frontiers in Physics*, *8*, 287. https://doi.org/10.3389/fphy.2020.00287

Nisbet, R., Miner, G., & Yale, K. (2018). Chapter 4—Data Understanding and Preparation. In R. Nisbet, G. Miner, & K. Yale (Eds.), *Handbook of Statistical Analysis and Data Mining Applications (Second Edition)* (pp. 55–82). Academic Press. https://doi.org/10.1016/B978-0-12-416632-5.00004-9

O'Malley, Tom and Bursztein, Elie and Long, James and Chollet, Fran\c{c}ois and Jin, Haifeng and Invernizzi, Luca and others. (2019). *KerasTuner* [Computer software]. \url{https://github.com/keras-team/keras-tuner

Pagotto, J., Zhang, J., & Duignan, T. T. (2022, October 21). *Predicting electrolyte solution properties by combining neural network accelerated molecular dynamics and continuum solvent theory*. NeurIPS 2022 AI for Science: Progress and Promises. https://openreview.net/forum?id=nIYyCVP3X6e

Palmer, D. S., O'Boyle, N. M., Glen, R. C., & Mitchell, J. B. O. (2007). Random forest models to predict aqueous solubility. *Journal of Chemical Information and Modeling*, *47*(1), 150–158. https://doi.org/10.1021/ci060164k

Panagiotopoulos, A. Z., & Yue, S. (2023). Dynamics of Aqueous Electrolyte Solutions: Challenges for Simulations. *The Journal of Physical Chemistry B, 127*(2), 430–437. https://doi.org/10.1021/acs.jpcb.2c07477

Renon, H. (1996). Models for excess properties of electrolyte solutions: Molecular bases and classification, needs and trends for new developments. *Fluid Phase Equilibria*, *116*(1), 217–224. https://doi.org/10.1016/0378-3812(95)02890-0

Roduner, E., & Krüger, T. P. J. (2022). The origin of irreversibility and thermalization in thermodynamic processes. *Physics Reports*, *944*, 1–43. https://doi.org/10.1016/j.physrep.2021.11.002

Rozmus, J., Brunella, I., Mougin, P., & Hemptinne, J.-C. (2012). Isobaric Vapor–Liquid Equilibria of Tertiary Amine and n-Alkane/Alkanol Binary Mixtures: Experimental Measurements and Modeling with GC-PPC-SAFT. *Journal of Chemical & Engineering Data*, *57*, 2915–2922. https://doi.org/10.1021/je300568h

Rupp, M., Tkatchenko, A., Müller, K.-R., & von Lilienfeld, O. A. (2012). Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Physical Review Letters*, *108*(5), 058301. https://doi.org/10.1103/PhysRevLett.108.058301

Said, Z., Sharma, P., Bora, B. J., & Pandey, A. K. (2023). Sonication impact on thermal conductivity of f-MWCNT nanofluids using XGBoost and Gaussian process regression. *Journal of the Taiwan Institute of Chemical Engineers*, *145*, 104818. https://doi.org/10.1016/j.jtice.2023.104818

Saravi, S. H., & Panagiotopoulos, A. Z. (2022). Activity Coefficients and Solubilities of NaCl in Water–Methanol Solutions from Molecular Dynamics Simulations. *The Journal of Physical Chemistry B, 126*(15), 2891–2898. https://doi.org/10.1021/acs.jpcb.2c00813

Serrano, D., Golpour, I., & Sánchez-Delgado, S. (2020). Predicting the effect of bed materials in bubbling fluidized bed gasification using artificial neural networks (ANNs) modeling approach. *Fuel*, *266*, 117021. https://doi.org/10.1016/j.fuel.2020.117021

Shahriari, R., & Dehghani, M. R. (2018). New electrolyte SAFT-VR Morse EOS for prediction of solid-liquid equilibrium in aqueous electrolyte solutions. *Fluid Phase Equilibria*, *463*, 128–141. https://doi.org/10.1016/j.fluid.2018.02.006

Soares, E. do A., Vernin, N. S., Santos, M. S., & Tavares, F. W. (2022). Real Electrolyte Solutions in the Functionalized Mean Spherical Approximation: A Density Functional Theory for Simple Electrolyte Solutions. *The Journal of Physical Chemistry B*, *126*(32), 6095–6101. https://doi.org/10.1021/acs.jpcb.2c00816

Stenzel, O., Pecho, O., Holzer, L., Neumann, M., & Schmidt, V. (2017). Big data for microstructure-property relationships: A case study of predicting effective conductivities. *AIChE Journal*, *63*(9), 4224–4232. https://doi.org/10.1002/aic.15757

Stephenson, D., Kermode, J. R., & Lockerby, D. A. (2018). Accelerating multiscale modelling of fluids with on-the-fly Gaussian process regression. *Microfluidics and Nanofluidics*, *22*(12), 139. https://doi.org/10.1007/s10404-018-2164-z

Sun, L., Liang, X., Solms, N. von, & Kontogeorgis, G. M. (2020). Analysis of Some Electrolyte Models Including Their Ability to Predict the Activity Coefficients of Individual Ions. *Industrial & Engineering Chemistry Research*, *59*(25), 11790–11809. https://doi.org/10.1021/acs.iecr.0c00980

Wang, Y., Qiu, L. P., & Hu, M. F. (2018). Magnesium Ammonium Phosphate Crystallization: A Possible Way for Recovery of Phosphorus from Wastewater. *IOP Conference Series: Materials Science and Engineering*, *392*(3), 032032. https://doi.org/10.1088/1757-899X/392/3/032032

Wilson, G. M. (1964). Vapor-Liquid Equilibrium. XI. A New Expression for the Excess Free Energy of Mixing. *Journal of the American Chemical Society*, *86*(2), 127–130. https://doi.org/10.1021/ja01056a002

Yan, A., Sokolinski, T., Lane, W., Tan, J., Ferris, K., & Ryan, E. M. (2021). Applying transfer learning with convolutional neural networks to identify novel electrolytes for metal air batteries. *Computational and Theoretical Chemistry*, *1205*, 113443. https://doi.org/10.1016/j.comptc.2021.113443

Young, A. F., Ahón, V. R. R., & Pessoa, F. L. P. (2021). Vapour-liquid equilibrium for mixed-solvent—Strong electrolytes mixtures with EoS/GE models. *The Journal of Chemical Thermodynamics*, *154*, 106339. https://doi.org/10.1016/j.jct.2020.106339

Yu, J. (2012). Online quality prediction of nonlinear and non-Gaussian chemical processes with shifting dynamics using finite mixture model based Gaussian process regression approach. *Chemical Engineering Science*, *82*, 22–30. https://doi.org/10.1016/j.ces.2012.07.018

Zhang, Y., & Wallace, B. (2017). A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification. *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 253–263. https://aclanthology.org/I17-1026

Zhao E., Yu M., Sauvé R.E., Khoshkbarchi M.K. (2000). *Extension of the Wilson Model to Electrolyte Solutions.* 173:161-175.

Zhong, S., Zhang, K., Wang, D., & Zhang, H. (2021). Shedding light on "Black Box" machine learning models for predicting the reactivity of HO radicals toward organic compounds. *Chemical Engineering Journal*, *405*, 126627. https://doi.org/10.1016/j.cej.2020.126627

Zhou, L., Chen, J., & Song, Z. (2015). Recursive Gaussian Process Regression Model for Adaptive Quality Monitoring in Batch Processes. *Mathematical Problems in Engineering*, *2015*, e761280. https://doi.org/10.1155/2015/761280