

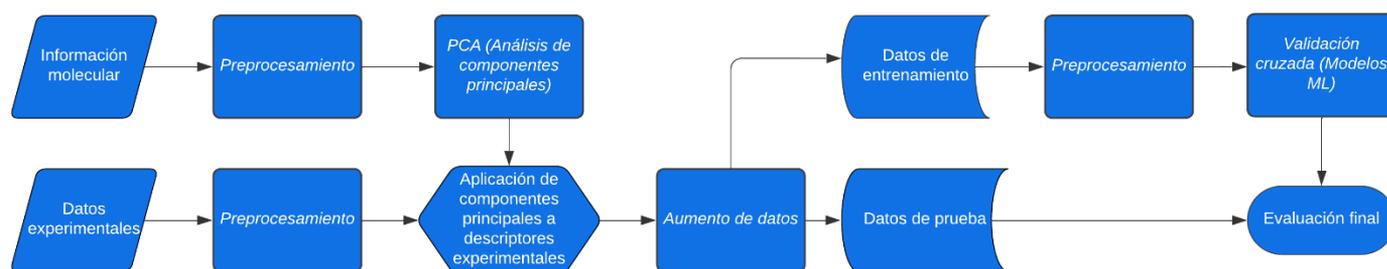
ELV-ML: Predicción de diagramas de equilibrio líquido-vapor con Aprendizaje de Máquina sin Código

Nicolás Iglesias Giraldo¹, Alejandro Valencia Virquez*

¹Departamento de Ingeniería Química, Universidad de los Andes, Bogotá, Colombia

Resumen gráfico

A continuación, se encuentra un flujograma del manejo de datos propuesto



Resumen

El auge reciente del aprendizaje automático, particularmente mediante modelos sofisticados como el "aprendizaje profundo", ha permitido resolver problemas y desarrollar nuevos enfoques en la ingeniería química. Esta investigación se centra en la predicción de diagramas de equilibrio vapor-líquido (ELV) utilizando modelos de aprendizaje automático. La investigación presente se concentró en predecir diagramas para todo tipo de moléculas, a diferentes presiones y temperaturas y sin la suposición de ningún modelo físico (a diferencia de los recientes avances en literatura al respecto [1]–[4]). A través de un procedimiento experimental detallado y la evaluación de múltiples modelos, se buscó correlacionar descriptores moleculares (obtenidos de la librería RDKit [5]) con los diagramas ELV (en particular con las fracciones de vapor) mediante el uso de herramientas sin código como RapidMiner [6] para facilitar el proceso.

Para llevar a cabo esta investigación, se siguieron los siguientes pasos experimentales. Primero, se realizó una búsqueda exhaustiva en bases de datos robustas como ChERIC [7], artículos científicos revisados por pares y el motor ThermoData de NIST [8] en Aspen Plus. En total, se recopiló datos experimentales de 562 pares de compuestos, con 27,349 puntos de equilibrio únicos.

Para mejorar el rendimiento del modelo, se aplicaron técnicas de reducción de dimensionalidad como el Análisis de Componentes Principales (PCA) y la selección de características mediante correlación basados en información obtenida por "scraping" de bases de datos moleculares de un conjunto amplio de moléculas diversas. Esto redujo las características de entrada de 208 a 53, preservando el 80% de la varianza original de los datos. Posteriormente, los datos se dividieron en conjuntos de entrenamiento (80%) y prueba (20%). Para evitar sobreajustes, se utilizó la validación cruzada en lotes (para optimizar los modelos) que garantizara que los datos de prueba fueran completamente independientes de los datos de entrenamiento.

Se entrenaron y optimizaron varios modelos, los cuales incluyen una regresión lineal (como modelo base), Random Forest, XGBoost y redes neuronales artificiales (ANNs). Se utilizó una búsqueda de malla para optimizar los hiperparámetros de cada modelo. Los modelos se evaluaron utilizando métricas del error cuadrático medio (RMSE), el error absoluto medio (MAE) y el coeficiente de determinación (R^2). Además, se compararon los resultados con el modelo UNIFAC, comúnmente utilizado en la predicción de diagramas VLE.

Los resultados preliminares mostraron que las redes neuronales artificiales (ANNs) ofrecen el mejor rendimiento entre los modelos evaluados (modelos que se encuentran todavía en proceso de optimización). El modelo UNIFAC por su parte, aunque eficiente para compuestos comunes, mostró grandes errores en datos de moléculas poco comunes. La regresión lineal tuvo el peor desempeño, indicando la falta de linealidad en el problema. Random Forest mostró un bajo rendimiento, mientras que XGBoost tuvo una mejora significativa respecto a Random Forest. Las redes neuronales superaron a todos los modelos, manejando mejor las no linealidades del problema.

La investigación demostró que el uso de modelos de aprendizaje automático, particularmente las redes neuronales, puede mejorar significativamente la predicción de diagramas VLE. Además, la utilización de herramientas sin código como RapidMiner facilitó la implementación y optimización de los modelos, lo cual hace el proceso accesible a la comunidad académica.

Palabras clave: Aprendizaje de máquina, descriptores moleculares, equilibrio líquido-vapor

Reconocimientos

Agradecemos a la Universidad de los Andes y a los profesores Andrés González Barrios y Pablo Ortiz Herrera por su apoyo

Referencias

- [1] C. A. Faúndez, F. A. Quiero, and J. O. Valderrama, “Phase equilibrium modeling in ethanol + congener mixtures using an artificial neural network,” *Fluid Phase Equilib.*, vol. 292, no. 1–2, pp. 29–35, 2010, doi: 10.1016/j.fluid.2010.01.001.
- [2] S. Chakraborty, Y. Sun, G. Lin, and L. Qiao, “Vapor–liquid equilibrium estimation of n-alkane/nitrogen mixtures using neural networks,” *J. Comput. Appl. Math.*, vol. 408, p. 114059, 2022, doi: 10.1016/j.cam.2021.114059.
- [3] A. Carranza-Abaid, H. F. Svendsen, and J. P. Jakobsen, “Thermodynamically consistent vapor-liquid equilibrium modelling with artificial neural networks,” *Fluid Phase Equilib.*, vol. 564, no. August 2022, p. 113597, 2023, doi: 10.1016/j.fluid.2022.113597.
- [4] G. Sun, Z. Zhao, S. Sun, Y. Ma, H. Li, and X. Gao, “Vapor-liquid phase equilibria behavior prediction of binary mixtures using machine learning,” *Chem. Eng. Sci.*, vol. 282, no. October, p. 119358, 2023, doi: 10.1016/j.ces.2023.119358.
- [5] “The RDKit Documentation — The RDKit 2023.09.6 documentation.” <https://www.rdkit.org/docs/> (accessed Mar. 15, 2024).
- [6] “Data Analytics and AI Platform | Altair RapidMiner.” <https://altair.com/altair-rapidminer> (accessed Mar. 15, 2024).
- [7] ChERIC, “KDB Binary Vapor Liquid Equilibrium.” <https://www.cheric.org/research/kdb/hcvle/listvle.php>.
- [8] “ThermoData Engine | NIST.” <https://www.nist.gov/mml/acmd/trc/thermodata-engine> (accessed Mar. 15, 2024).