Predicción de equilibrios líquido-vapor con aprendizaje de máquina sin código

Nicolás Iglesias y Alejandro Valencia

1 INTRODUCCIÓN

Los diagramas de equilibrio líquido- vapor son esenciales tanto en la industria como en la academia para investigar procesos de separación. Los modelos de predicción existentes (UNIQUAC, Wilson, NTRL) dependen de datos experimentales de los compuestos y dejan de ser confiables a presiones altas. El aprendizaje de máquina permite generar un modelo probabilístico que relaciona estos diagramas a información estructural de las moléculas. **A pesar de los avances en este frente, estos siempre se limitan a herramientas codificadas** (lo que dificulta la reproducibilidad de los resultados, la diversificación de modelos y la interdisciplinaridad), **así como a tipos de moléculas restringidos**

OBJETIVOS

Predecir las fracciones de vapor de un sistema binario con base en descriptores moleculares para un rango amplio de moléculas

LIMITACIONES

Dado que se trata de un avance inicial, **el modelo propuesto depende de los datos proporcionados para las fracciones líquidas del sistema a una presión y temperatura específicas.**

2 METODOLOGÍA

INFORMACIÓN MOLECULAR

Por medio de “scraping” se recolectó información molecular de más de 147,300 moléculas y sus correspondientes descriptores moleculares (208)

EQUILIBRIOS LÍQUIDO-VAPOR

Se recolectaron 27,000 datos de equilibrio experimentales que comprendían 562 pares de compuestos únicos.

PROCESAMIENTO

Se utilizó la información molecular para reducir la dimensionalidad del problema mediante un análisis de correlación y análisis de componentes principales (80 % varianza). Se realizó un aumento de los datos con base en entendimiento termodinamico

MODELOS

Se realizó una separación de los datos que permitiera independencia en la evaluación (por parejas de compuestos). Adicionalmente, se realizó una optimización de hiperparámetros con validación cruzada para evaluar su capacidad de generalización.

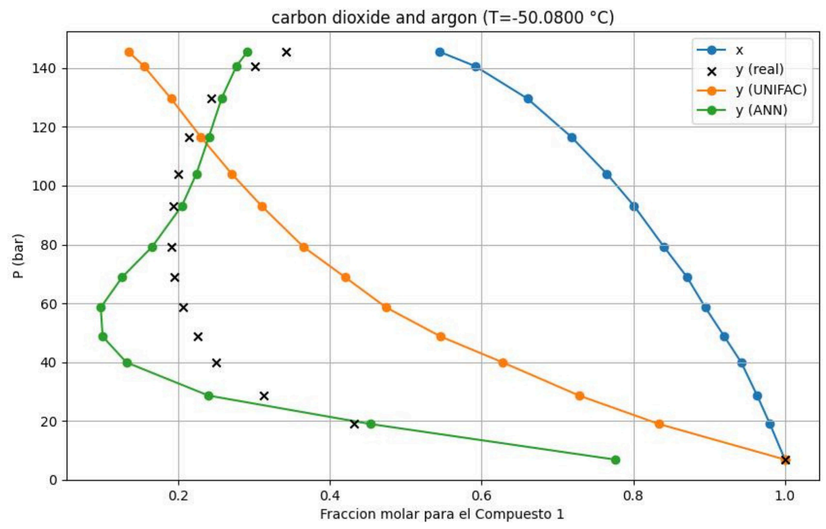
Se evaluaron los siguientes modelos:

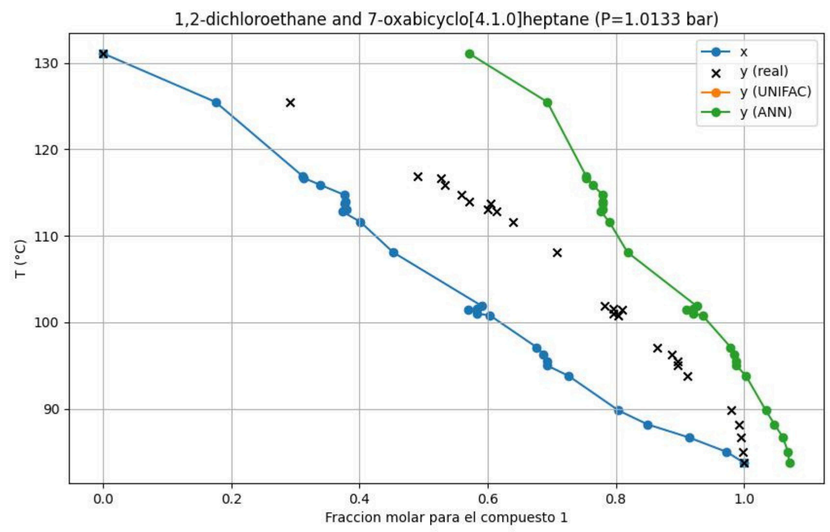
- **Regresión lineal**
- **Random Forest**
- **XGBoost**
- **Redes neuronales**
- **UNIFAC**

3 RESULTADOS

Las redes neuronales tuvieron el mejor comportamiento entre los modelos evaluados con un R^2 de 0.811 en el set de prueba (en comparación al 0.868 de UNIFAC en esos mismos datos). Adicionalmente, se observó que los resultados dependían de los tipos de compuestos, lo cual indica que **el modelo predice mejor ciertos tipos de moléculas.**

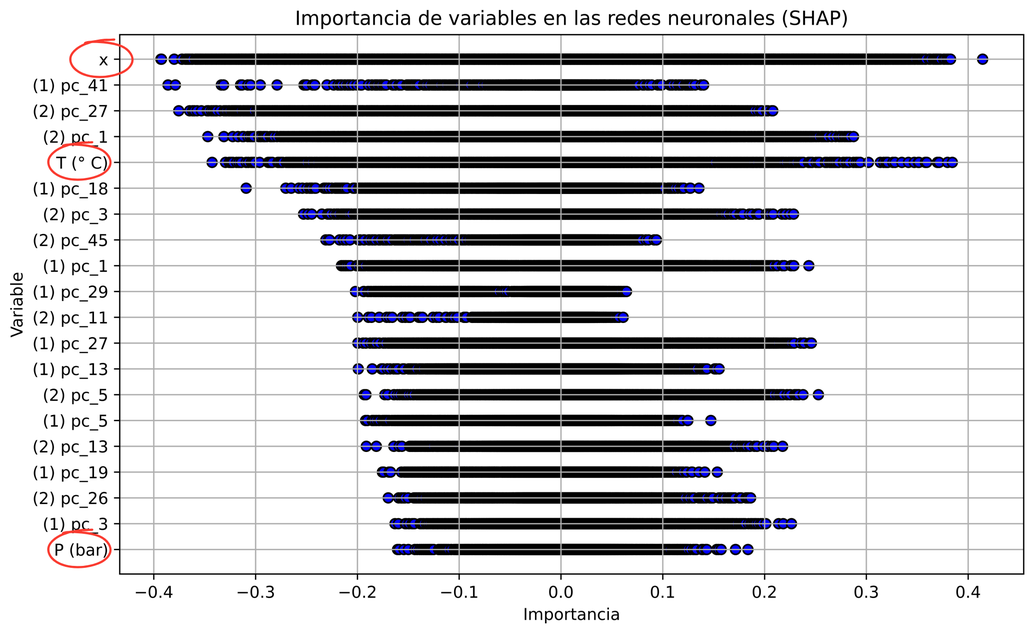
Matriz de resultados	Regresión lineal	XGBoost	Random Forest	Redes Neuronales	UNIFAC
RMSE (Test)	0.389	0.188	0.232	0.156	0.130
MAE (Test)	0.259	0.139	0.187	0.091	0.054
R^2 (Test)	**0.200**	**0.732**	**0.609**	**0.811**	**0.868**





Se visualizaron equilibrios de fases que el modelo predijo mejor que UNIFAC **y uno, incluso, que no puede calcularse con UNIFAC debido a la falta de parámetros de sus grupos funcionale**s. Adicionalmente, se compararon más de 200 equilibrios isobáricos e isotérmicos en un alto rango de condiciones de operación.

Un análisis de sensibilidad realizado por el método de SHAP encontró que **el modelo, en efecto, utiliza la información molecular para sus predicciones** (incluso algunas veces más que la presión y la temperatura)



4 CONCLUSIONES

El modelo basado en redes neuronales demostró una buena capacidad de predicción que asemeja modelos robustos como UNIFAC.

El modelo a futuro necesita incorporar restricciones termodinámicas que mejoren su rendimiento y permitan tanto una mayor independencia de datos experimentales como una mejora en la precisión