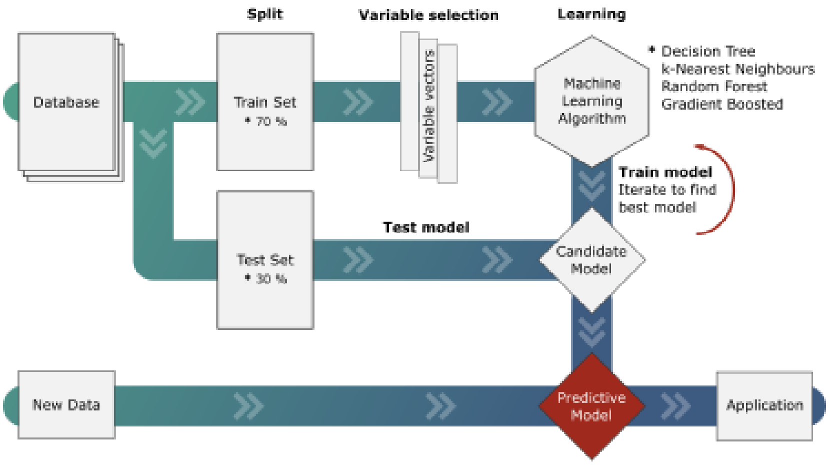
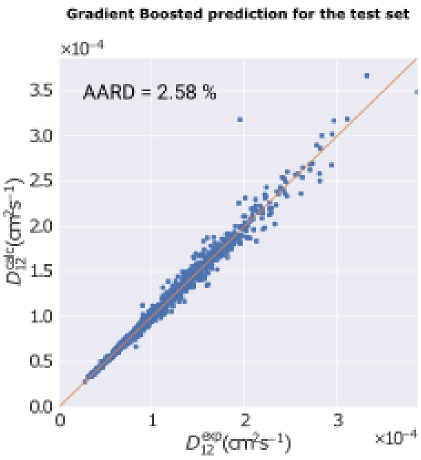
Prediction of diffusivities in supercritical carbon dioxide using machine learning models

José P.S. Aniceto, Bruno Zêzere, Carlos M. Silva

CICECO, Department of Chemistry, University of Aveiro, 3810-193 Aveiro, Portugal





The knowledge of transport properties is required for the design, simulation and scale-up of separations and chemical reactions. In the case of the molecular diffusion coefficient, D_{12}, it is fundamental to estimate dispersion coefficients, convective mass transfer coefficients, and catalysts efficiency factors. Many such processes make use of supercritical carbon dioxide (SC-CO_2) as it enables fine tuning the affinity of the solvent mixture to specific solutes. Since experimental D_{12} data in SC-CO_2 is not widely available, there is a significant demand for accurate models capable of providing reliable D_{12} estimations. Currently, the Wilke-Chang equation, which is a modification of the Stokes–Einstein equation, is the most well-known and most used model to calculate solute diffusivities in SC-CO_2. Recently, Artificial Intelligence models have been applied to the prediction of diffusivities of gases at atmospheric pressure and binary diffusion coefficients of liquids.

In this work, machine learning algorithms were applied to develop predictive models to estimate diffusivities of solutes in supercritical carbon dioxide. A large database of experimental data containing 21 properties for 174 solutes and 4917 data points (covering small and large molecules), was used in the training of the machine learning models. The database was randomly split 70/30 % into training and testing sets, respectively. Four machine learning models were evaluated: a *k*-Nearest Neighbors model, a Decision Tree algorithm, and two Ensemble Method (Random Forest and Gradient Boosted). The results were compared with a simple multi-linear regression and with the Wilke-Chang equation.

The best results were found using the Gradient Boosted algorithm which showed an average absolute relative deviation (AARD) of 2.58 % (see Figure) for the 1476 points in the test set (not used in model training). This model has six parameters including temperature, pressure, density, solute molar mass, solute critical pressure, and solute acentric factor. The *k*-Nearest Neighbors, Decision Tree and Random Forest models presented overall results between 4.1 % and 5.5 %. By comparison the multi-linear regression obtained an AARD of 15.86 % and the Wilke-Chang equation resulted in an AARD of 12.41 % for the same test set.

Acknowledgements

This work was developed within the scope of the project CICECO-Aveiro Institute of Materials, UIDB/50011/2020 & UIDP/50011/2020, financed by national funds through the FCT/MEC and when appropriate co-financed by FEDER under the PT2020 Partnership Agreement. Bruno Zêzere thanks FCT for PhD grant SFRH/BD/137751/2018.