Predicting the Hazardous Qualities of Compounds Based on NFPA 704 Labels Using Deep Learning for QSAR/QSPR Modeling

Almutairi, Hashim (School: Dar Althikr Private School)

Thousands of chemical substances are synthesized and introduced to the science field every day. The hazardous properties for these chemicals need to be established before they are used. The identification of hazardous substances is a tedious and time-consuming process. The process involves much lab testing to identify physical (e.g. flammability, reactivity) and health hazards (e.g. carcinogenicity, sensitization). In this project the QSAR/QSPR model was developed using the information of 6,538 chemicals, including SMILES identifiers, the image of the chemical along with the NFPA 704 label, and CACTVS. Logistic regressions, Multi-Linear Perceptron, Decision Trees, Random Forest and Neural Networks (NN) are the learning models, which have been used in modeling the hazardous features. The accuracy of CACTVS voting classifier evaluated by 10-fold cross validation was 81%. Using train/test split evaluator CACTVS-NN displayed a 80.73% accuracy, SMILES-NN 72.81%, IMAGE-NN 74.73% and IMAGE-CACTVS- SMILES-NN 81.05%. In addition, the important features, which have the highest association with the hazardous properties of the chemical, have been extracted from the possible models. The representation identifiers of the chemical has a great impact on the model accuracy since CACTVS, SMILES and IMAGE represent the same chemical in a variety of ways. The model can be used productively for chemical hazard predictions and as an assistance tool for researchers and scientist in laboratories.

Awards Won:
Third Award of $1,000
American Statistical Association: Certificate of Honorable Mention
King Abdulaziz &amp
his Companions Foundation for Giftedness and Creativity: Award of $1,000 for research in Innovative Technology.