Assignment 08 - Machine Learning QSAR

Assignment Name:	**Week 8 Assignment - Machine Learning QSAR**
Weight:	6.25% of final grade
Due Date:	Tuesday, Midnight
Associated Learning Outcomes:	Train and validate machine learning models for cheminformatics-based drug discovery.
Assignment Prompt:	We will use the BBB data set again from Week 5. This week we will use it to create a ML model of BBB penetration in knime, and predict BBB penetrance likelihood for melatonin receptor ligands. You may use any available machine learning methods, but I recommend Random Forest. Melatonin receptor ligands can be found here: (https://www.ebi.ac.uk/chembl/target_report_card/CHEMBL1946/). Generate a 4 page report that includes the following: 1) Getting / Cleaning the Data (**8 points**) a) Overview of the known compounds (what property space do they cover? Perform a property analysis and compare the training molecules and the test molecules. Is using this training set appropriate for the use case? b) Is there possibility for leakage? Meaning are there training molecules either identical to the proposed test set, or similar? c) What properties and/or fingerprints will you use to train and test your models? Why? 2) Model building (**10 points**) a) How did you split the data - show your work in as much detail as possible indicating all code or knime workflows. Recall from the discussion this week that random-splitting is almost never appropriate for chemical data, which contains a lot of biases. Consider your training data carefully

	b) Use clustering you learned last week and generate a 2D plot of the chemical space, and label your training vs test sets. Should you use cross-validation? c) Model validation - what metric did you use to train your model? 3) Prediction / Conclusions (**7 points**) a) Predict the likelihood of BBB penetrance for the melatonin receptor ligands. Analyse the predictions and comment on a few of the case studies. b) Find a chemical space (e.g. use previous weeks chemical sets, or find your own on ChemBl - for example use results from your VS from week 6) and predict BBB penetrance using your model. Does this seem to make sense? What were your prior assumptions?
Context/Purpose:	The identification of new chemical matter, and optmizing lead compounds is one of the main jobs for a cheminformatician. This week, we will explore some practical applications.
Requirements and Logistics:	Students should turn in a typed word or google docs document converted to PDF. All figures should be pasted into the document to generate a one-file submission. Handwritten materials will not be accepted.

Assignment Rubric Template

	Excellent Full Value	**Acceptable** Half Points	**Incorrect / Absent** No Points
Organization (3 Points)	Assignment is written in a clear and interpretable manner (3 points)	Answers are there, but either disorganized, or grammar errors present (1 points)	Multiple grammar errors, or very disorganized. (0 points)
Assignment Questions (22 Points)	Questions are answered and workflows screenshots are provided.	Partially correct, or supporting workflows are not provided.	Incorrect, or not answered