TECS CDT: ML Rotation Assessment

Now you have completed the training component of your ML rotation, you will be working in teams to use machine learning to tackle a real-world problem in chemistry. You will be asked to run a classification task to determine what molecules are toxic, and you should use the skills you have developed in all the notebooks to produce a comprehensive piece of work.

Problem Context

The **Tox21** database contains nearly 8000 molecules, where a qualitative measurement of their toxicity has been conducted on two different targets. The first target includes 7 nuclear receptor bioassays and the second target 5 stress response pathway bioassays.

More information can be found here:

General: https://tox21.gov/overview/

Assays: https://ncats.nih.gov/research/research-activities/Tox21/assays

For each molecule you are provided with a SMILES string and then a binary value on if the molecule is deemed toxic in the corresponding assay. The data is located in the file:

Tox21.csv

It is your goal to come up with a machine learning algorithm, in which if a new molecule is provided it can predict if it will be toxic or not. You will also need to assess how accurate your model is, and if it could be used as a resource for scientists working on new unseen molecules.

In the following some suggestions are made on aspects that will be important to consider. It is important to ensure that you thoroughly plan as part of a team.

Suggestion 1

The data contains multiple assays, with multiple binary values. Formally this can be treated as a *multi-task classification problem*. The formulation of this can be quite advanced (even on Scikit-Learn). Therefore, you need to decide if you will set this up as a multi-task problem, focus on a single bioassay (does this then translate to other receptors?) or combine specific assays into a total toxicity score.

As you examine this aspect of the problem, you will notice that the dataset has entries missing, will you automatically define these as 0 or 1, remove them completely or use an algorithm with *masking*.

Suggestion 2

Performing descriptive statistics on your dataset is a useful way to understand if the dataset is balanced and provide useful visuals. RDKit will be very useful here – for example you can define substructures and see their prevalence in the dataset *etc*.. Examining other infographics may be useful here.

Suggestion 3

You are only provided with the SMILES strings for this problem. Therefore, you will need to perform the *featurisation*. You can use RDKit descriptors, which are cheminformatic descriptors – which are information rich, and may aid in the early investigation and visualisation. Alternatively, you may wish to use a molecular fingerprint generated from the SMILES strings (what other featurisation methods are there).

Suggestion 4

Consider the balance of your dataset, how many molecules are toxic and non-toxic. This result may impact your train-(validate)-test split. For example, with the train-test split you can pass class labels to perform a stratified split, is this useful? What size splits will you choose considering the amount of data.

Suggestion 5

Consider which models you will examine (many of the models in the regression exercise have analogous classification algorithms). You will need to consider the hyperparameters of each model and which metric(s) you are using to evaluate your models.

Suggestion 6

Once you have created a final model, interrogate how well it is performing. Some prompt questions include: 1. Are certain functionalities always deemed as toxic? 2. What are the

false positives/negatives and what implication does this have? 3. Does toxicity translate over bioassays? 4. What does the model say about out-of-sample predictions?

Suggestion 7

Consider how you will display the information in a concise way, results, information on ML and implications in chemistry.

Suggestion 8

Spend time thinking about how you will create the infographic, some useful tools include (you can use animations!):

- Powerpoint
- Piktochart (https://piktochart.com/formats/infographics/)
- Snappa (https://snappa.com/create/infographics)
- Venngage (https://venngage.com/)
- Canva (https://www.canva.com/create/infographics/)
- Visme (https://www.visme.co/make-infographics/)
- Freepik (https://www.freepik.com/free-photos-vectors/infographic)
- Adobe Express (https://www.adobe.com/express/create/infographic)
- https://buffer.com/library/infographic-makers/

Conclusion

This is an open-ended task (it is an actual problem in chemistry), therefore there is no correct answer, and many directions can be taken in the project. It is important to ensure that you have a clear plan on how, why and what you will be doing, so you have achievable goals to produce a good infographic for the press conference.