Task 1: Baseline Modeling — The Standard ML Pipeline

Weeks: 1–2
Title: *"Let's Build a Model!"*

Objectives:

- Clean and standardize chemical data
- Apply a standard cheminformatics ML pipeline
- Train multiple models and evaluate their performance

Deliverables:

- Presentation covering data preparation, model performance, and model comparison
- Jupyter notebook containing well-commented code and proper documentation of analysis logic
    - Make sure to run your notebook in one go. This ensures that the sequence of the code is correct and anyone can run it after you.
    - **Push your code to the GitHub repo before the next session**

Description

You are a chechinformatics research and development (R&D) team in a pharmaceutical company. The CEO wants to reduce the experimental costs by embracing the promising new trend of machine learning (ML) and artificial intelligence (AI).

One of the main experiments run by the company is to measure the aqueous solubility of a molecule. This means that a molecule in solid state is stirred in water and monitored to see when it will precipitate.

Knowing the concentration needed before a molecule precipitates is important because it helps the pharmacists in deciding whether a molecule can work as an oral drug or does it need to be intravenous. Also, it helps them adjust the dose needed to make sure it will arrive at the place it is meant to reach in the body before it dissolves completely.

So, it would truly be great if there is an ML model that can accurately predict the aqueous solubility of new molecules and skip this experimental step already.

Knowing that ML and AI are all about data, you ask your CEO to give you a dataset of molecules with their corresponding aqueous solubility measurements to use for training an ML model.

The CEO gives you a dataset of ~ 9 thousand molecules (see attached). It consists of two pieces of information, a molecule represented as a SMILES string and a numerical value of the LogS unit.

1st internal meeting

You held an internal meeting together to discuss how you can approach this task.

A team member mentioned the need to make sure that these **SMILES** strings are **sanitized** and **standardized.** Another mentioned that it is important to have a numerical representation of each SMILES string because ML models can only process numbers.

But how can you convert SMILES into numbers?

A team member read that a tool called **Mordred** provides useful descriptors that can describe a molecule numerically. This can be interesting to explore.

Another team member mentioned that **MACCS** structural keys can also be used. This descriptor was designed by chemists. So, hopefully it will be chemically relevant.

Now, the only thing left is to train an ML model. But which one?

You decided to try your luck with two powerful families of modeling, **Random Forests (RF)** and **Support Vector Machines (SVM)**.

You split the tasks between you, and you go modeling.

- **You will split the dataset into 80% training and 20% testing**
- **You will train four models (each representation with each model)**
- **You will check the performance of each model on the test set and decide which model and representation pair works best.**

Meeting the CEO

You request a meeting with the CEO to show her your results. She is not that experienced with ML, so you need to communicate your results with her in an intuitive way.

Remember, she will make a big decision to stop an experimental branch based on what you will be showing. So, do it wisely, and nicely!

If you did not manage to convince her, your contract will be terminated because AI will be failing her and she needs to focus on what is working.

You split the task between you, and you prepare a presentation that explains the following:

- **What was your training scheme? How and why did you split the data?**
- **What were the representations and models you used? Why did you use them?**
- **How did you decide which representation and model to use eventually? What analysis help you pick?**
- **How well will this model perform if the CEO gives you new molecules?**

The CEO challenge

The CEO was quite impressed, but before making a final decision, she gives you two new internal datasets and asks you to show your model performance on each one of them.

If the results remained the same, this would make her very confident

- **You will need to save your trained model in a format that allows you to load it and use them for testing new molecules**
- **You will receive these two new datasets 3 days before the next lecture**
- **You will report the results of your chosen model in your presentation**

Important notes for your deliverables

The bare minimum you can do is to have a functional Jupyter notebook that

1. Sanitizes and standardizes SMILES
2. Represent SMILES in Mordred and MACC descriptors
3. Train the RF and SVM models
4. Use different evaluation metrics to assess the performance of your models

If you have questions alongside your implementation journey (and I hope you do), simply note them down in this spreadsheet and move on.

Once you have your bare minimum product and you want to learn more, go back to your questions and answer them one by one.

Each answer will help you understand your work better, and therefore, present it and explain it better.

A cheat hint!

If you copy and paste this document to a language model, it should give you such bare minimum functional code, notebook, and presentation outline.

I am saying this to emphasize that this is NOT the important part. It's your questions and digging deeper to understand what is happening...