AI Chemist Challenge - Lipophilicity Prediction

Objective

The goal of this challenge is to assess your machine learning and Python programming skills for a potential AI Chemist role. You will design and implement a regression model to predict a critical drug property: lipophilicity.

Background

ANYO Labs is at the forefront of revolutionizing early-stage drug discovery with cutting-edge AI solutions. As an AI Chemist, you'll be a crucial part of our team, developing and exploring new software innovations that directly impact the field of computational chemistry.

This challenge is designed to assess your qualifications for the AI Chemist role at ANYO Labs. It will specifically evaluate your understanding of Machine Learning (ML) principles and how they can be applied to solve problems in drug discovery. The task involves building a machine learning model to predict a key drug property: lipophilicity.

Task

1. **Dataset:** You are provided with a file containing the training data to use. Familiarize yourself with its content and format.
2. **Environment:** Python is the required programming language. You are free to select appropriate ML/AI frameworks (e.g., scikit-learn, TensorFlow, PyTorch).
3. **Modeling Approach:**
   - **Preprocessing:** Clean and prepare the data as needed.
   - **Feature Engineering:** Decide how to represent molecular structures numerically (e.g., molecular descriptors, fingerprints).
   - **Algorithm:** Select a regression algorithm suitable for the problem. Experiment with different choices, if desired.
4. **Training:**
   - Utilize appropriate dataset splitting techniques (random or scaffold).
   - Train your model, optimizing hyperparameters for best performance.
5. **Evaluation:**
   - Employ relevant regression metrics (e.g., Root Mean Squared Error, R-squared).

Deliverables:

- **GitHub Repository:** Create a **public** repo containing:
  - Your training code with clear comments explaining steps and decisions.

    - The trained model (saved in a standard format).
    - A brief README file with instructions on how to use your model to make predictions on new molecules.
- **Live Performance Evaluation of Models**:
    - You will be invited to a live virtual meeting where the performance of their models can be assessed through a test run.
    - Testset will be given during the meeting.

Evaluation Criteria

- **Python Proficiency:** Clean code structure, meaningful variable names.
- **ML Expertise:** Sound understanding of regression concepts, algorithm selection, and performance evaluation.
- **Documentation:** Ability to explain your thought process and code choices.
- **Creativity:** Exploration of innovative feature engineering or modeling techniques.

Important Notes

- Focus on a well-explained solution, even if relatively simple, rather than an overly complex model.
- We value a clear, rational approach. Document the reasoning behind your decisions.

We look forward to seeing your work!