**Drug-Likeness Prediction (pChEMBL Value Estimation)**

**Overview**

This project implements a **machine learning model** to predict the **pChEMBL values** of compounds based on their SMILES representation. The dataset consists of **bioactivity data** for acetylcholinesterase inhibitors. The project involves data preprocessing, molecular descriptor calculation, and model training to estimate drug-likeness.

**Features**

* **SMILES-based Prediction**: Uses molecular descriptors derived from SMILES strings.
* **Machine Learning Model**: Random Forest Regressor with train-test splitting.
* **Performance Metrics**:
  + Train R²: ~86%
  + Test R²: ~61%
  + Slight overfitting, but Random Forest performed better than other models.
* **Hyperparameter Tuning**: Attempts to improve model performance resulted in degradation.

**Workflow**

1. **Data Preprocessing**
   * Load the dataset (acetylcholinesterase\_01\_bioactivity\_data\_processed.csv).
   * Extract molecular features (descriptors) from SMILES representations.
2. **Feature Engineering**
   * Convert SMILES into numerical descriptors.
   * Prepare data for machine learning models.
3. **Model Training**
   * Train-test split.
   * Train a **Random Forest Regressor** to predict pChEMBL values.
   * Evaluate model performance.
4. **Results & Interpretation**
   * Random Forest Regressor provided the best fit.
   * No significant improvement from hyperparameter tuning.

**How to Use**

1. Install dependencies:

pip install numpy pandas scikit-learn rdkit

1. Run the notebook and load the dataset.
2. Input a SMILES string to get a predicted pChEMBL value.

**Future Improvements**

* Use advanced models like **XGBoost** or deep learning.
* Optimize feature selection to reduce overfitting.
* Expand dataset for better generalization.

**Source of data**: **DATA PROFESSOR** (YouTube/Git)