Dopamine dataset consists of 479 molecules represented with ChEMBL identifiers and is labeled with pIC50 molecular property. The task is to implement machine learning approach of predicting the pIC50 molecular property by training on molecular descriptors retrieved from the SMILES representation.

Preprocessing:

* Retrieve SMILES representation for each molecule using the ChEMBL REST API through your Python environment.
* Make SMILES canonical by using rdkit Python package.
* Calculate molecular descriptors by using Mordred Python package or some other available molecular descriptor calculator. Use all available 0D, 1D, 2D and 3D molecular descriptors.
* Clean the transformed data of missing values and outliers. Remove duplicates.
* Split the preprocessed dataset into subset for training and subset for testing in ratio 75:25. If using deep learning, split into subset for validation as well (ratio 70:15:15)
* Define and initialize scaler by using Scikit-learn Python package. Scale the data appropriately.

Modeling:

* Use whichever machine learning model you like for supervised learning of the regression problem alongside with random forest and support vector machine as benchmark models.
* Use Random Forest and Support Vector Machine models’ implementation from the Scikit-learn Python package.
* Optimize each model’s hyperparameters.

Model selection:

* Define methods for calculating RMSE, MAE and r2.
* Define any cross-validation method.
* Implement the whole preprocessing and modeling data pipeline and estimate each model’s prediction performance by cross validating. Use the defined metrics. Compare and discuss results.