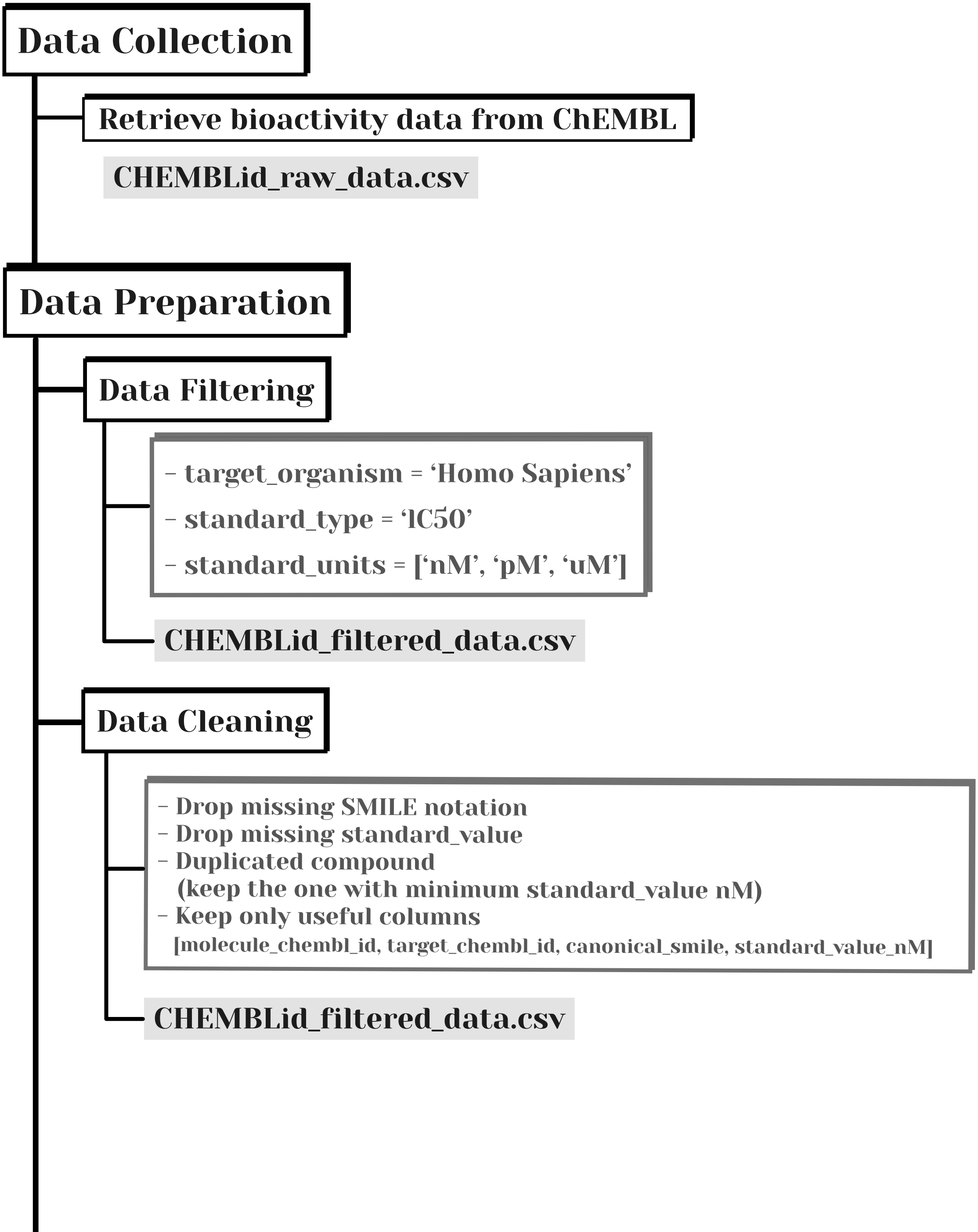
Pipeline



Data Collection

Retrieve bioactivity data from ChEMBL

CHEMBLid_raw_data.csv

Data Preparation

Data Filtering

- target_organism = ‘Homo Sapiens’
- standard_type = ‘IC50’
- standard_units = [‘nM’, ‘pM’, ‘uM’]

CHEMBLid_filtered_data.csv

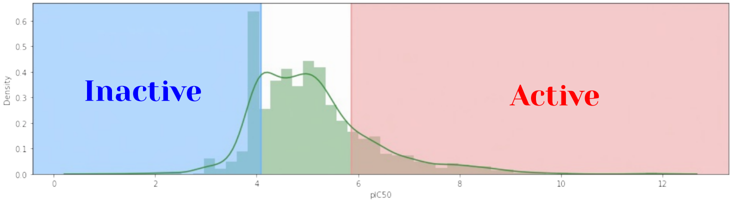
Data Cleaning

- Drop missing SMILE notation
- Drop missing standard_value
- Duplicated compound
  (keep the one with minimum standard_value nM)
- Keep only useful columns
  [molecule_chembl_id, target_chembl_id, canonical_smile, standard_value_nM]

CHEMBLid_filtered_data.csv

- **Data Labeling**
  - **Scaling IC50 to pIC50**
  - **Labeling**

    < 20th percentile = ‘inactive’ = 0
    > 80th percentile = ‘active’ = 1



    CHEMBLid_3cls_labeled_data.csv

  - **Remove ‘intermediate’ class**

    CHEMBLid_binary_2cls_labeled_data.csv

- **Data Transformation** (For multi-labeled only)
  - **Load existing all bioactivity data** original-mlc-data.csv
  - **Data Aggregation**

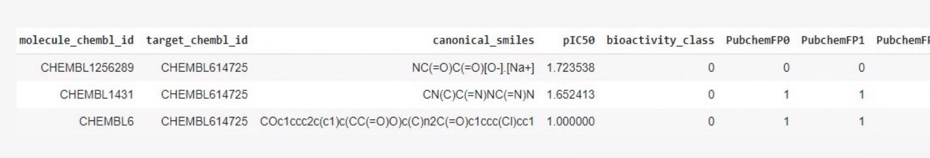
    CHEMBLid_multi_2cls_labeled_data.csv

PubChem Fingerprint Calculation

PaDEL-Descriptor

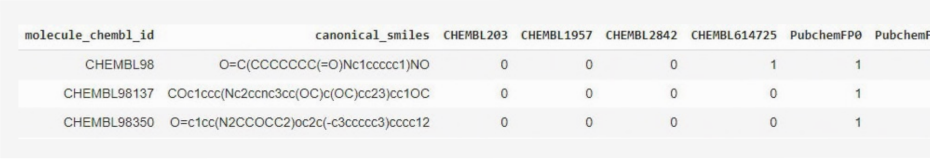
http://pubmed.nobility.nvm.nih.gov/214252294/

CHEMBLid_binary_FP_all_data.csv



molecule_chembl_id	target_chembl_id	canonical_smiles	pIC50	bioactivity_class	PubchemFP0	PubchemFP1	PubchemFI
CHEMBL1256289	CHEMBL614725	NC(=O)C(=O)[O-].[Na+]	1.723538	0	0	0	
CHEMBL1431	CHEMBL614725	CN(C)C(=N)NC(=N)N	1.652413	0	1	1	
CHEMBL6	CHEMBL614725	COc1ccc2c(c1)c(CC(=O)O)c(C)n2C(=O)c1ccc(Cl)cc1	1.000000	0	1	1	

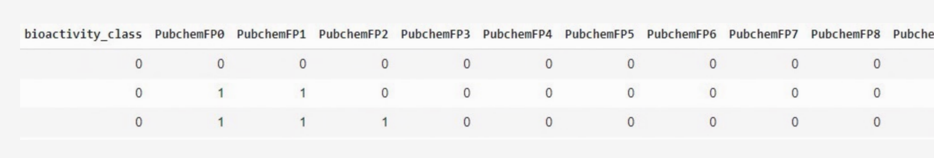
CHEMBLid_multi_FP_all_data.csv



molecule_chembl_id	canonical_smiles	CHEMBL203	CHEMBL1957	CHEMBL2842	CHEMBL614725	PubchemFP0	PubchemF
CHEMBL98	O=C(CCCCCCC(=O)Nc1ccccc1)NO	0	0	0	1	1	
CHEMBL98137	COc1ccc(Nc2ccnc3cc(OC)c(OC)cc23)cc1OC	0	0	0	0	1	
CHEMBL98350	O=c1cc(N2CCOCC2)oc2c(-c3ccccc3)cccc12	0	0	0	0	1	

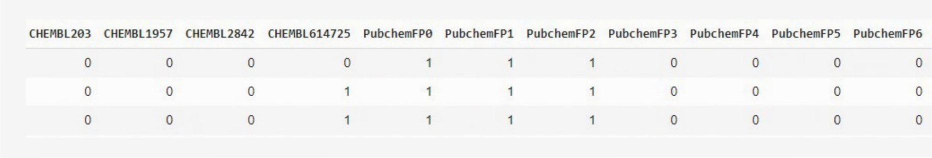
Dataset for Modeling

CHEMBLid_Binary_binary_dataset.csv

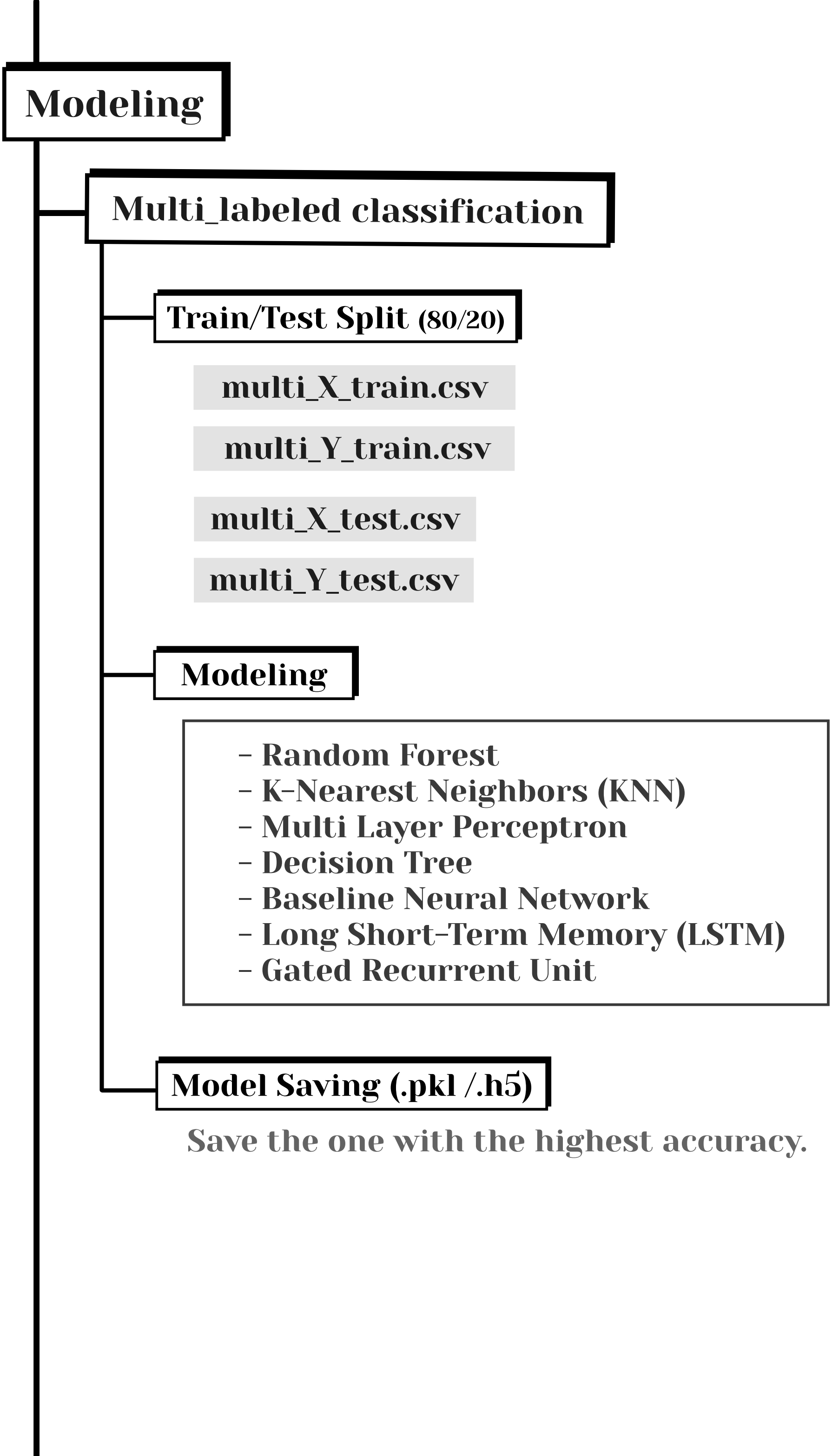


bioactivity_class	PubchemFP0	PubchemFP1	PubchemFP2	PubchemFP3	PubchemFP4	PubchemFP5	PubchemFP6	PubchemFP7	PubchemFP8	Pubche
0	0	0	0	0	0	0	0	0	0	
0	1	1	0	0	0	0	0	0	0	
0	1	1	1	0	0	0	0	0	0	

CHEMBLid_Multi_modeling_dataset.csv



CHEMBL203	CHEMBL1957	CHEMBL2842	CHEMBL614725	PubchemFP0	PubchemFP1	PubchemFP2	PubchemFP3	PubchemFP4	PubchemFP5	PubchemFP6
0	0	0	0	1	1	1	0	0	0	0
0	0	0	1	1	1	1	0	0	0	0
0	0	0	1	1	1	1	0	0	0	0



Modeling

Multi_labeled classification

Train/Test Split (80/20)

multi_X_train.csv

multi_Y_train.csv

multi_X_test.csv

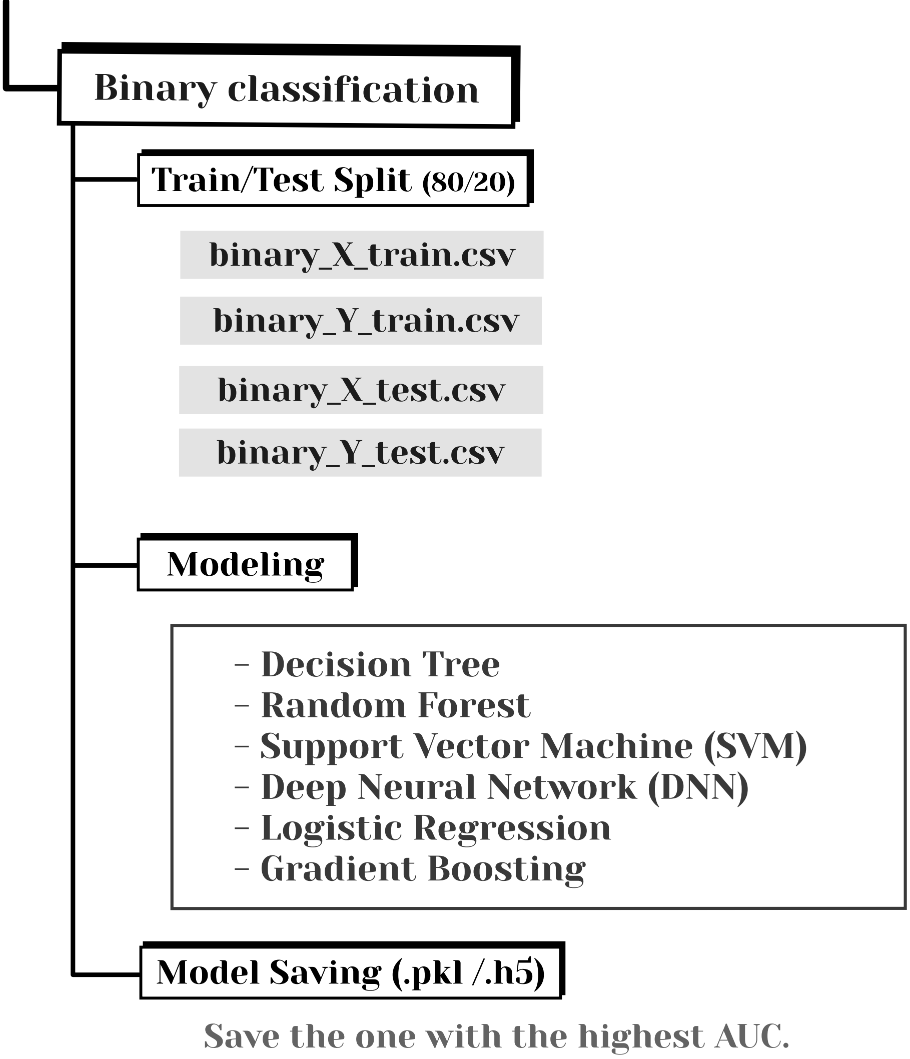
multi_Y_test.csv

Modeling

- Random Forest
- K-Nearest Neighbors (KNN)
- Multi Layer Perceptron
- Decision Tree
- Baseline Neural Network
- Long Short-Term Memory (LSTM)
- Gated Recurrent Unit

Model Saving (.pkl /.h5)

Save the one with the highest accuracy.



Binary classification

Train/Test Split (80/20)

binary_X_train.csv

binary_Y_train.csv

binary_X_test.csv

binary_Y_test.csv

Modeling

- Decision Tree
- Random Forest
- Support Vector Machine (SVM)
- Deep Neural Network (DNN)
- Logistic Regression
- Gradient Boosting

Model Saving (.pkl /.h5)

Save the one with the highest AUC.