

6dp7yxy4f

June 6, 2024

```
[ ]: import numpy as np
import pandas as pd

!pip install rdkit-pypi
import rdkit
from rdkit import Chem
from rdkit.Chem import AllChem
```

Requirement already satisfied: rdkit-pypi in /usr/local/lib/python3.10/dist-packages (2022.9.5)  
Requirement already satisfied: numpy in /usr/local/lib/python3.10/dist-packages (from rdkit-pypi) (1.25.2)  
Requirement already satisfied: Pillow in /usr/local/lib/python3.10/dist-packages (from rdkit-pypi) (9.4.0)

```
[ ]: train = pd.read_csv('/content/X_train_AX2CWD7.csv')
Y_train = pd.read_csv('/content/y_train_hCIvDMj.csv')['y']
test = pd.read_csv('/content/X_test_0mB4D2v.csv')
```

```
[ ]: # Convert SMILES to mols
train_mols = [AllChem.MolFromSmiles(smile) for smile in
train['smiles']]
test_mols = [AllChem.MolFromSmiles(smile) for smile in test['smiles']]
```

```
[ ]: # Convert Mol to fingerprints
train_fps = np.array([AllChem.GetMorganFingerprintAsBitVect(mol,
radius=2, nBits=2048) for mol in train_mols])
test_fps = np.array([AllChem.GetMorganFingerprintAsBitVect(mol,
radius=2, nBits=2048) for mol in test_mols])
```

```
[ ]: X_train = train_fps
X_test = test_fps
```

```
[ ]: from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import cross_validate
```

```
[ ]: model = RandomForestRegressor()

cv = cross_validate(estimator=model,
                    X=X_train,
                    y=Y_train,
                    cv=3,
                    scoring=["neg_median_absolute_error"],
                    n_jobs=-1,
                    verbose=0)
print(f"Mean MAE over 10 folds = {-cv['test_neg_median_absolute_error'].mean():.
    ↪4f}")
```

Mean MAE over 10 folds = 0.4431

```
[ ]: #opti rf

parameters ={'max_depth':[3,5,10,None],
             'n_estimators':[10,100,200],
             'max_features':[1,3,5,7],
             'min_samples_leaf':[1,2,3],
             'min_samples_split':[1,2,3]
            }

from sklearn.model_selection import GridSearchCV

scoring = "neg_median_absolute_error"

rfGS = GridSearchCV(model,parameters,cv=3,scoring=scoring)
rfGS.fit(X_train,Y_train)
```

/usr/local/lib/python3.10/dist-packages/sklearn/model\_selection/\_validation.py:378: FitFailedWarning:  
432 fits failed out of a total of 1296.  
The score on these train-test partitions for these parameters will be set to nan.  
If these failures are not expected, you can try to debug them by setting error\_score='raise'.

Below are more details about the failures:

```
-----
432 fits failed with the following error:
Traceback (most recent call last):
  File "/usr/local/lib/python3.10/dist-packages/sklearn/model_selection/_validation.py", line 686, in _fit_and_score
    estimator.fit(X_train, y_train, **fit_params)
  File "/usr/local/lib/python3.10/dist-packages/sklearn/ensemble/_forest.py", line 340, in fit
    self._validate_params()
```

```

File "/usr/local/lib/python3.10/dist-packages/sklearn/base.py", line 600, in
_validate_params
    validate_parameter_constraints(
File "/usr/local/lib/python3.10/dist-
packages/sklearn/utils/_param_validation.py", line 97, in
validate_parameter_constraints
    raise InvalidParameterError(
sklearn.utils._param_validation.InvalidParameterError: The 'min_samples_split'
parameter of RandomForestRegressor must be an int in the range [2, inf) or a
float in the range (0.0, 1.0]. Got 1 instead.

```

```

warnings.warn(some_fits_failed_message, FitFailedWarning)
/usr/local/lib/python3.10/dist-packages/sklearn/model_selection/_search.py:952:
UserWarning: One or more of the test scores are non-finite: [      nan
nan      nan -1.00964122 -1.01173039 -1.01315898
-1.01638276 -1.01017684 -1.0111199      nan      nan      nan
-1.00722442 -1.0093324  -1.0126763  -1.01493452 -1.01048898 -1.01178652
      nan      nan      nan -1.01105657 -1.01489673 -1.01471783
-1.01486522 -1.01117206 -1.01480303      nan      nan      nan
-0.99665491 -1.00107675 -1.00856028 -1.00455239 -1.00225164 -1.00444011
      nan      nan      nan -1.00899647 -1.00517213 -1.00486342
-1.00072866 -1.00402508 -1.00131622      nan      nan      nan
-0.99223362 -1.00618676 -1.00321872 -1.00499367 -1.00507436 -1.00456394
      nan      nan      nan -0.99936056 -0.99114586 -0.9946129
-0.98698777 -0.99600227 -0.99741215      nan      nan      nan
-1.00069509 -0.99359601 -0.99419528 -0.98314616 -0.99397611 -0.99226767
      nan      nan      nan -0.99400115 -0.9967248  -0.99571412
-0.98648272 -0.99501075 -0.9950583      nan      nan      nan
-0.99537057 -0.98881321 -0.98758654 -0.98188308 -0.98424843 -0.98469798
      nan      nan      nan -0.98597738 -0.99023375 -0.98660466
-0.99781923 -0.98497484 -0.99008784      nan      nan      nan
-0.97774038 -0.98554891 -0.99060387 -0.98659612 -0.99088888 -0.9884243
      nan      nan      nan -0.99826392 -1.00096341 -1.00598466
-1.00852455 -1.00617339 -1.00290436      nan      nan      nan
-0.99741646 -1.00287652 -1.00690661 -1.01083615 -1.01022214 -1.00609651
      nan      nan      nan -1.01076914 -1.01215305 -1.00917272
-1.01274437 -1.01140202 -1.01272844      nan      nan      nan
-0.99177416 -0.98687184 -0.98920987 -0.99699992 -0.98555028 -0.98647671
      nan      nan      nan -0.99070153 -0.98522247 -0.99057508
-0.99840808 -0.98612232 -0.9901768      nan      nan      nan
-0.99250648 -0.98785992 -0.98870023 -0.99490161 -0.99410626 -0.98777354
      nan      nan      nan -0.97373623 -0.9768996  -0.96945902
-0.97475577 -0.97068112 -0.97599031      nan      nan      nan
-0.97684577 -0.98022315 -0.97327327 -0.98587098 -0.9804268  -0.9736871
      nan      nan      nan -0.97860345 -0.97537614 -0.97612237
-0.97893128 -0.98207809 -0.98002379      nan      nan      nan
-0.97442443 -0.96370993 -0.96152808 -0.97041491 -0.95665881 -0.9627602
      nan      nan      nan -0.95651837 -0.97162401 -0.96195726

```

-0.97419118	-0.96343492	-0.96058	nan	nan	nan
-0.97012908	-0.96345323	-0.96253443	-0.96627177	-0.96909131	-0.96749902
nan	nan	nan	-0.97065031	-0.9794422	-0.97466559
-0.98074452	-0.97916785	-0.97742774	nan	nan	nan
-0.98945704	-0.99084754	-0.99506058	-0.99250443	-0.9990644	-0.99514071
nan	nan	nan	-1.0049589	-1.00147757	-1.00235474
-1.00782257	-1.0009751	-1.00117789	nan	nan	nan
-0.9351918	-0.93815285	-0.94254149	-0.94252843	-0.9433914	-0.93958461
nan	nan	nan	-0.95017639	-0.95673617	-0.95235906
-0.95064829	-0.96270051	-0.9601322	nan	nan	nan
-0.95028545	-0.96165974	-0.9579557	-0.96143062	-0.95987349	-0.96402122
nan	nan	nan	-0.92293275	-0.9110397	-0.91215598
-0.92980228	-0.91036465	-0.90943679	nan	nan	nan
-0.9336994	-0.93349685	-0.928822	-0.92515951	-0.92549546	-0.92321741
nan	nan	nan	-0.92829495	-0.93282517	-0.9408489
-0.93652479	-0.93827831	-0.93072434	nan	nan	nan
-0.91473202	-0.89507912	-0.88561606	-0.91877099	-0.88501402	-0.889217
nan	nan	nan	-0.90980277	-0.89888302	-0.90531233
-0.92260099	-0.89448536	-0.8991795	nan	nan	nan
-0.91335217	-0.90685475	-0.90042839	-0.91550596	-0.9100975	-0.90674792
nan	nan	nan	-0.56956892	-0.53902522	-0.53901771
-0.58283131	-0.56269175	-0.57692495	nan	nan	nan
-0.96112435	-0.97414996	-0.98034772	-0.98182746	-0.97157749	-0.98017605
nan	nan	nan	-0.99421945	-0.99790681	-0.99812681
-1.00600127	-0.99286924	-0.99961286	nan	nan	nan
-0.55039478	-0.52587619	-0.52404813	-0.57244385	-0.55133926	-0.54984079
nan	nan	nan	-0.87595007	-0.86338215	-0.86722478
-0.87631411	-0.8708154	-0.86020432	nan	nan	nan
-0.89121296	-0.89019832	-0.8919875	-0.90911339	-0.89164889	-0.88809536
nan	nan	nan	-0.53324125	-0.51764146	-0.51654393
-0.53934416	-0.54215466	-0.53674965	nan	nan	nan
-0.80015329	-0.81601176	-0.80905986	-0.79458024	-0.80383012	-0.8075757
nan	nan	nan	-0.84981835	-0.84896925	-0.84671533
-0.87228239	-0.84561224	-0.84755564	nan	nan	nan
-0.52467533	-0.50351085	-0.51001228	-0.55808162	-0.52565886	-0.53149721
nan	nan	nan	-0.75544529	-0.76924283	-0.77389269
-0.77579908	-0.76363736	-0.76545975	nan	nan	nan
-0.80969299	-0.81163375	-0.81392648	-0.80945711	-0.80846864	-0.80839187]

warnings.warn(

```
[ ]: GridSearchCV(cv=3, estimator=RandomForestRegressor(),
                  param_grid={'max_depth': [3, 5, 10, None],
                              'max_features': [1, 3, 5, 7],
                              'min_samples_leaf': [1, 2, 3],
                              'min_samples_split': [1, 2, 3],
                              'n_estimators': [10, 100, 200]},
                  scoring='neg_median_absolute_error')
```

```
[ ]: best_rf = rfGS.best_estimator_
```

```
[ ]: model = best_rf

cv = cross_validate(estimator=model,
                    X=X_train,
                    y=Y_train,
                    cv=3,
                    scoring=["neg_median_absolute_error"],
                    n_jobs=-1,
                    verbose=0)
print(f"Mean MAE over 10 folds = {-cv['test_neg_median_absolute_error'].mean():.
    ↪4f}")
```

Mean MAE over 10 folds = 0.5155

```
[ ]: from sklearn.tree import DecisionTreeRegressor

model = DecisionTreeRegressor()

cv = cross_validate(estimator=model,
                    X=X_train,
                    y=Y_train,
                    cv=3,
                    scoring=["neg_median_absolute_error"],
                    n_jobs=-1,
                    verbose=0)
print(f"Mean MAE over 10 folds = {-cv['test_neg_median_absolute_error'].mean():.
    ↪4f}")
```

Mean MAE over 10 folds = 0.4785

```
[ ]: parameters={"splitter": ["best", "random"],
                  "max_depth" : [1,3,5,7,9,11,12],
                  "min_samples_leaf": [1,2,3,4,5,6,7,8,9,10],
                  "min_weight_fraction_leaf": [0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9],
                  "max_features": ["auto", "log2", "sqrt", None],
                  "max_leaf_nodes": [None,10,20,30,40,50,60,70,80,90] }

scoring = "neg_median_absolute_error"

dtGS = GridSearchCV(model,parameters,cv=3,scoring=scoring)
#dtGS.fit(X_train,Y_train) trop long
```

Le flux de sortie a été tronqué et ne contient que les 5000 dernières lignes.

/usr/local/lib/python3.10/dist-packages/sklearn/tree/\_classes.py:277:

FutureWarning: `max\_features='auto'` has been deprecated in 1.1 and will be removed in 1.3. To keep the past behaviour, explicitly set `max\_features=1.0`.





















































































































































































































































































































































































































































































```
warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/tree/_classes.py:277:
FutureWarning: `max_features='auto'` has been deprecated in 1.1 and will be
removed in 1.3. To keep the past behaviour, explicitly set `max_features=1.0`.
warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/tree/_classes.py:277:
FutureWarning: `max_features='auto'` has been deprecated in 1.1 and will be
removed in 1.3. To keep the past behaviour, explicitly set `max_features=1.0`.
warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/tree/_classes.py:277:
FutureWarning: `max_features='auto'` has been deprecated in 1.1 and will be
removed in 1.3. To keep the past behaviour, explicitly set `max_features=1.0`.
warnings.warn(
```

```
-----
KeyboardInterrupt                                Traceback (most recent call last)
<ipython-input-34-bbf97a8f1f31> in <cell line: 11>()
      9
     10 dtGS = GridSearchCV(model,parameters,cv=3,scoring=scoring)
----> 11 dtGS.fit(X_train,Y_train)

/usr/local/lib/python3.10/dist-packages/sklearn/model_selection/_search.py in _
fit(self, X, y, groups, **fit_params)
    872         return results
    873
--> 874         self._run_search(evaluate_candidates)
    875
    876         # multimetric is determined here because in the case of a _
callable

/usr/local/lib/python3.10/dist-packages/sklearn/model_selection/_search.py in _
_run_search(self, evaluate_candidates)
    1386     def _run_search(self, evaluate_candidates):
    1387         """Search all candidates in param_grid"""
-> 1388         evaluate_candidates(ParameterGrid(self.param_grid))
    1389
    1390

/usr/local/lib/python3.10/dist-packages/sklearn/model_selection/_search.py in _
evaluate_candidates(candidate_params, cv, more_results)
    819         )
    820
--> 821         out = parallel(
    822             delayed(_fit_and_score)(
    823                 clone(base_estimator),
```

```

/usr/local/lib/python3.10/dist-packages/sklearn/utils/parallel.py in
↳ __call__(self, iterable)
    61         for delayed_func, args, kwargs in iterable
    62     )
--> 63     return super().__call__(iterable_with_config)
    64
    65

/usr/local/lib/python3.10/dist-packages/joblib/parallel.py in __call__(self,
↳ iterable)
    1916         output = self._get_sequential_output(iterable)
    1917         next(output)
-> 1918         return output if self.return_generator else list(output)
    1919
    1920         # Let's create an ID that uniquely identifies the current call.
↳ If the

/usr/local/lib/python3.10/dist-packages/joblib/parallel.py in
↳ _get_sequential_output(self, iterable)
    1845         self.n_dispatched_batches += 1
    1846         self.n_dispatched_tasks += 1
-> 1847         res = func(*args, **kwargs)
    1848         self.n_completed_tasks += 1
    1849         self.print_progress()

/usr/local/lib/python3.10/dist-packages/sklearn/utils/parallel.py in
↳ __call__(self, *args, **kwargs)
    121         config = {}
    122         with config_context(**config):
--> 123         return self.function(*args, **kwargs)

/usr/local/lib/python3.10/dist-packages/sklearn/model_selection/_validation.py
↳ in _fit_and_score(estimator, X, y, scorer, train, test, verbose, parameters,
↳ fit_params, return_train_score, return_parameters, return_n_test_samples,
↳ return_times, return_estimator, split_progress, candidate_progress,
↳ error_score)
    684         estimator.fit(X_train, **fit_params)
    685     else:
-> 686         estimator.fit(X_train, y_train, **fit_params)
    687
    688     except Exception:

/usr/local/lib/python3.10/dist-packages/sklearn/tree/_classes.py in fit(self, X
↳ y, sample_weight, check_input)
    1245         """
    1246
-> 1247         super().fit(
    1248             X,

```

```

1249         y,

/usr/local/lib/python3.10/dist-packages/sklearn/tree/_classes.py in fit(self, X,
↳ y, sample_weight, check_input)
    184         check_X_params = dict(dtype=DTYPE, accept_sparse="csc")
    185         check_y_params = dict(ensure_2d=False, dtype=None)
--> 186         X, y = self._validate_data(
    187             X, y, validate_separately=(check_X_params,
↳ check_y_params)
    188         )

/usr/local/lib/python3.10/dist-packages/sklearn/base.py in _validate_data(self,
↳ X, y, reset, validate_separately, **check_params)
    577         if "estimator" not in check_X_params:
    578             check_X_params = {**default_check_params,
↳ **check_X_params}
--> 579         X = check_array(X, input_name="X", **check_X_params)
    580         if "estimator" not in check_y_params:
    581             check_y_params = {**default_check_params,
↳ **check_y_params}

/usr/local/lib/python3.10/dist-packages/sklearn/utils/validation.py in
↳ check_array(array, accept_sparse, accept_large_sparse, dtype, order, copy,
↳ force_all_finite, ensure_2d, allow_nd, ensure_min_samples,
↳ ensure_min_features, estimator, input_name)
    877         array = xp.astype(array, dtype, copy=False)
    878         else:
--> 879         array = _asarray_with_order(array, order=order,
↳ dtype=dtype, xp=xp)
    880         except ComplexWarning as complex_warning:
    881             raise ValueError(

/usr/local/lib/python3.10/dist-packages/sklearn/utils/_array_api.py in
↳ _asarray_with_order(array, dtype, order, copy, xp)
    183         if xp.__name__ in {"numpy", "numpy.array_api"}:
    184             # Use NumPy API to support order
--> 185         array = numpy.asarray(array, order=order, dtype=dtype)
    186         return xp.asarray(array, copy=copy)
    187     else:

KeyboardInterrupt:

```

```
[ ]: #best_dt = dtGS.best_estimator_
```

```
[ ]: #model = best_dt
model_dt = DecisionTreeRegressor().fit(X_train, Y_train)
best_dt = model_dt
```



```
[ ]: #cv = cross_validate(estimator=model, X=X_train, y=Y_train, cv=3,
    ↪scoring=["neg_median_absolute_error"], n_jobs=-1, verbose=0)
    #print(f"Mean MAE over 10 folds = {-cv['test_neg_median_absolute_error'].mean():.
    ↪.4f}")
```

```
[ ]: from sklearn.neighbors import KNeighborsRegressor

model = KNeighborsRegressor()

cv = cross_validate(estimator=model,
    X=X_train,
    y=Y_train,
    cv=3,
    scoring=["neg_median_absolute_error"],
    n_jobs=-1,
    verbose=0)
print(f"Mean MAE over 10 folds = {-cv['test_neg_median_absolute_error'].mean():.
    ↪.4f}")
```

Mean MAE over 10 folds = 0.4669

```
[ ]: parameters = {'n_neighbors': list(range(1,31))}

scoring = "neg_median_absolute_error"

knnGS = GridSearchCV(model,parameters,cv=3,scoring=scoring)
knnGS.fit(X_train,Y_train)
```

```
[ ]: GridSearchCV(cv=3, estimator=KNeighborsRegressor(),
    param_grid={'n_neighbors': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12,
    13, 14, 15, 16, 17, 18, 19, 20, 21, 22,
    23, 24, 25, 26, 27, 28, 29, 30]},
    scoring='neg_median_absolute_error')
```

```
[ ]: best_knn = knnGS.best_estimator_
```

```
[ ]: model = best_knn

cv = cross_validate(estimator=model,
    X=X_train,
    y=Y_train,
    cv=3,
    scoring=["neg_median_absolute_error"],
    n_jobs=-1,
    verbose=0)
print(f"Mean MAE over 10 folds = {-cv['test_neg_median_absolute_error'].mean():.
    ↪.4f}")
```

Mean MAE over 10 folds = 0.4014

```
[ ]: import sklearn.metrics as metrics

rf_acc = metrics.median_absolute_error(Y_train, best_rf.predict(X_train))
knn_acc = metrics.median_absolute_error(Y_train, best_knn.predict(X_train))
dt_acc = metrics.median_absolute_error(Y_train, best_dt.predict(X_train))
```

```
[ ]: model_acc = pd.DataFrame({
    'Model': ['KNN', 'Decision Tree',
              'Random Forest'],

    'median_absolute_error': [knn_acc, dt_acc,
                              rf_acc]})

model_acc.sort_values(by='median_absolute_error', ascending=False)
```

```
[ ]:
      Model  median_absolute_error
2  Random Forest          0.175092
0           KNN           0.000000
1  Decision Tree           0.000000
```

```
[ ]: mean_acc = np.mean([rf_acc, knn_acc, dt_acc])
weight_rf = rf_acc / mean_acc
weight_knn = knn_acc / mean_acc
weight_dt = dt_acc / mean_acc
```

```
[ ]: from sklearn.ensemble import VotingRegressor

ensemble = VotingRegressor(estimators=[("rf", best_rf), ("knn", best_knn),
    ↪("dt", best_dt)], weights=[weight_rf, weight_knn, weight_dt])
```

```
[ ]: ensemble.fit(X_train, Y_train)

ens_prediction = ensemble.predict(X_train)
metrics.median_absolute_error(Y_train, ens_prediction)
```

```
[ ]: 0.17546683859685075
```

```
[ ]: ensemble.fit(X_train, Y_train)

y_pred = ensemble.predict(X_test)

submission_df = pd.DataFrame()
submission_df['id'] = test['id']
submission_df['y'] = y_pred
submission_df.to_csv("y_benchmark.csv", index=False)
```

```
[ ]: from google.colab import files  
files.download('y_benchmark.csv')
```

<IPython.core.display.Javascript object>

<IPython.core.display.Javascript object>