In [1]:

```
import numpy as np
from sklearn.pipeline import Pipeline
from sklearn.neighbors import KNeighborsRegressor
from sklearn.impute import SimpleImputer
from sklearn.feature_selection import SelectKBest, f_regression
from numpy.random import randint
import pandas as pd
my_NIA = 100419401
np.random.seed(my_NIA)
#3.Selecting the first four points,300 attributes)
train = pd.read_csv('train.csv')
test = pd.read_csv('test.csv')
X_train = train.iloc[:,0:300].values
y_train = train.iloc[:,1200:1201].values
X_test = test.iloc[:,0:300].values
y_test = test.iloc[:,1200:1201].values
```

In [2]:

```
#4 .10% of columns300*10%=30
selectedcolumns=np.random.choice(300, 30, replace=False)
selectedcolumns
```

Out[2]:

```
array([203,  63,  46, 199,  40,  34,  65, 180, 120,  77, 174, 232,  49,
        70,  88,  47, 136, 186, 118, 258, 107, 129, 248, 282, 108, 111,
        55,  31,  32,  89])
```

In [3]:

```
for column in selectedcolumns:
    rplace1=np.random.choice(4380, 438, replace=False)
    for place in  rplace1:
        X_train[place,column]=np.nan
for column in selectedcolumns:
    rplace2=np.random.choice(733, 73, replace=False)
    for place in  rplace2:
        X_test[place,column]=np.nan
```

In [4]:

```
from sklearn.pipeline import Pipeline
from sklearn.feature_selection import VarianceThreshold
from sklearn import preprocessing
from sklearn.decomposition import PCA
from sklearn.neighbors import KNeighborsRegressor
from sklearn.model_selection import GridSearchCV
#4.validation using PredefinedSplit in part1
from sklearn.model_selection import PredefinedSplit
validation_indices = np.zeros(X_train.shape[0])
validation_indices[:round(10/12*X_train.shape[0])] = -1
tr_val_partition = PredefinedSplit(validation_indices)

imputer=SimpleImputer(strategy='median')
remove=VarianceThreshold(threshold=1)
scaler = preprocessing.MinMaxScaler()
pca=PCA()
knn=KNeighborsRegressor()
estimators = [('impute', imputer), ('remove', remove),('scaler',scaler),("pca",pca), (
'knn', knn)]
pcapipeline=Pipeline(estimators)
param_grid={'n_components':range(1,40,1)}
```

In [5]:

```
from sklearn.model_selection import GridSearchCV
pca_grid=GridSearchCV(pcapipeline,param_grid,scoring='neg_mean_squared_error',cv=tr_val
_partition,n_jobs=-1, verbose=1)
```

In [6]:

```
pca_grid=pca_grid.fit(X_train,y_train)
y_test_pca_pred=pca_grid.predict(X_test)
pca_grid.best_params_
pca_grid.best_score_
```

```
Fitting 1 folds for each of 39 candidates, totalling 39 fits

[Parallel(n_jobs=-1)]: Using backend LokyBackend with 8 concurrent worker
s.
```

```
---------------------------------------------------------------------------
-
_RemoteTraceback                          Traceback (most recent call las
t)
_RemoteTraceback:
"""
Traceback (most recent call last):
  File "C:\Users\15096\Anaconda3\lib\site-packages\joblib\externals\loky\p
rocess_executor.py", line 418, in _process_worker
    r = call_item()
  File "C:\Users\15096\Anaconda3\lib\site-packages\joblib\externals\loky\p
rocess_executor.py", line 272, in __call__
    return self.fn(*self.args, **self.kwargs)
  File "C:\Users\15096\Anaconda3\lib\site-packages\joblib\_parallel_backen
ds.py", line 567, in __call__
    return self.func(*args, **kwargs)
  File "C:\Users\15096\Anaconda3\lib\site-packages\joblib\parallel.py", li
ne 225, in __call__
    for func, args, kwargs in self.items]
  File "C:\Users\15096\Anaconda3\lib\site-packages\joblib\parallel.py", li
ne 225, in <listcomp>
    for func, args, kwargs in self.items]
  File "C:\Users\15096\Anaconda3\lib\site-packages\sklearn\model_selection
\_validation.py", line 503, in _fit_and_score
    estimator.set_params(**parameters)
  File "C:\Users\15096\Anaconda3\lib\site-packages\sklearn\pipeline.py", l
ine 164, in set_params
    self._set_params('steps', **kwargs)
  File "C:\Users\15096\Anaconda3\lib\site-packages\sklearn\utils\metaestim
ators.py", line 50, in _set_params
    super().set_params(**params)
  File "C:\Users\15096\Anaconda3\lib\site-packages\sklearn\base.py", line
 224, in set_params
    (key, self))
ValueError: Invalid parameter n_components for estimator Pipeline(memory=N
one,
         steps=[('impute',
                 SimpleImputer(add_indicator=False, copy=True, fill_value=
None,
                               missing_values=nan, strategy='median',
                               verbose=0)),
                ('remove', VarianceThreshold(threshold=1)),
                ('scaler', MinMaxScaler(copy=True, feature_range=(0, 1))),
                ('pca',
                 PCA(copy=True, iterated_power='auto', n_components=None,
                     random_state=None, svd_solver='auto', tol=0.0,
                     whiten=False)),
                ('knn',
                 KNeighborsRegressor(algorithm='auto', leaf_size=30,
                                     metric='minkowski', metric_params=Non
e,
                                     n_jobs=None, n_neighbors=5, p=2,
                                     weights='uniform'))],
         verbose=False). Check the list of available parameters with `esti
mator.get_params().keys()`.
"""

The above exception was the direct cause of the following exception:

ValueError                                Traceback (most recent call las
t)
```

```
<ipython-input-6-76d96ee34844> in <module>
----> 1 pca_grid=pca_grid.fit(X_train,y_train)
      2 y_test_pca_pred=pca_grid.predict(X_test)
      3 pca_grid.best_params_
      4 pca_grid.best_score_

~\Anaconda3\lib\site-packages\sklearn\model_selection\_search.py in fit(se
lf, X, y, groups, **fit_params)
    686                 return results
    687
--> 688             self._run_search(evaluate_candidates)
    689
    690         # For multi-metric evaluation, store the best_index_, best
_params_ and

~\Anaconda3\lib\site-packages\sklearn\model_selection\_search.py in _run_s
earch(self, evaluate_candidates)
   1147     def _run_search(self, evaluate_candidates):
   1148         """Search all candidates in param_grid"""
-> 1149         evaluate_candidates(ParameterGrid(self.param_grid))
   1150
   1151

~\Anaconda3\lib\site-packages\sklearn\model_selection\_search.py in evalua
te_candidates(candidate_params)
    665                                for parameters, (train, test)
    666                                in product(candidate_params,
--> 667                                           cv.split(X, y, groups)))
    668
    669                 if len(out) < 1:

~\Anaconda3\lib\site-packages\joblib\parallel.py in __call__(self, iterabl
e)
    932
    933             with self._backend.retrieval_context():
--> 934                 self.retrieve()
    935             # Make sure that we get a last message telling us we a
re done
    936             elapsed_time = time.time() - self._start_time

~\Anaconda3\lib\site-packages\joblib\parallel.py in retrieve(self)
    831             try:
    832                 if getattr(self._backend, 'supports_timeout', Fals
e):
--> 833                     self._output.extend(job.get(timeout=self.timeo
ut))
    834                 else:
    835                     self._output.extend(job.get())

~\Anaconda3\lib\site-packages\joblib\_parallel_backends.py in wrap_future_
result(future, timeout)
    519         AsyncResults.get from multiprocessing."""
    520         try:
--> 521             return future.result(timeout=timeout)
    522         except LokyTimeoutError:
    523             raise TimeoutError()

~\Anaconda3\lib\concurrent\futures\_base.py in result(self, timeout)
    433                 raise CancelledError()
    434             elif self._state == FINISHED:
--> 435                 return self.__get_result()
```

```
    436             else:
    437                 raise TimeoutError()

~\Anaconda3\lib\concurrent\futures\_base.py in __get_result(self)
    382     def __get_result(self):
    383         if self._exception:
--> 384             raise self._exception
    385         else:
    386             return self._result

ValueError: Invalid parameter n_components for estimator Pipeline(memory=N
one,
         steps=[('impute',
                 SimpleImputer(add_indicator=False, copy=True, fill_value=
None,
                               missing_values=nan, strategy='median',
                               verbose=0)),
                ('remove', VarianceThreshold(threshold=1)),
                ('scaler', MinMaxScaler(copy=True, feature_range=(0, 1))),
                ('pca',
                 PCA(copy=True, iterated_power='auto', n_components=None,
                     random_state=None, svd_solver='auto', tol=0.0,
                     whiten=False)),
                ('knn',
                 KNeighborsRegressor(algorithm='auto', leaf_size=30,
                                     metric='minkowski', metric_params=Non
e,
                                     n_jobs=None, n_neighbors=5, p=2,
                                     weights='uniform'))],
         verbose=False). Check the list of available parameters with `esti
mator.get_params().keys()`.
```

In [7]:

```
#pipeline2 SelectkBest
from sklearn.feature_selection import SelectKBest,f_regression
selector=SelectKBest(f_regression)
estimators2 = [('impute', imputer), ('remove', remove),('scaler',scaler),("select",sele
ctor), ('knn', knn)]
selectpipeline=Pipeline(estimators2)
param_grid2={'select_k':range(1,40,1)}
from sklearn.model_selection import GridSearchCV
select_grid=GridSearchCV(selectpipeline,param_grid2,scoring='neg_mean_squared_error',cv
=tr_val_partition,n_jobs=-1, verbose=1)
select_grid=select_grid.fit(X_train,y_train)
y_test_select_pred=select_grid.predict(X_test)
select_grid.best_params_
select_grid.best_score_
```

```
Fitting 1 folds for each of 39 candidates, totalling 39 fits

[Parallel(n_jobs=-1)]: Using backend LokyBackend with 8 concurrent worker
s.
```

```
---------------------------------------------------------------------------
-
_RemoteTraceback                          Traceback (most recent call las
t)
_RemoteTraceback:
"""
Traceback (most recent call last):
  File "C:\Users\15096\Anaconda3\lib\site-packages\joblib\externals\loky\p
rocess_executor.py", line 418, in _process_worker
    r = call_item()
  File "C:\Users\15096\Anaconda3\lib\site-packages\joblib\externals\loky\p
rocess_executor.py", line 272, in __call__
    return self.fn(*self.args, **self.kwargs)
  File "C:\Users\15096\Anaconda3\lib\site-packages\joblib\_parallel_backen
ds.py", line 567, in __call__
    return self.func(*args, **kwargs)
  File "C:\Users\15096\Anaconda3\lib\site-packages\joblib\parallel.py", li
ne 225, in __call__
    for func, args, kwargs in self.items]
  File "C:\Users\15096\Anaconda3\lib\site-packages\joblib\parallel.py", li
ne 225, in <listcomp>
    for func, args, kwargs in self.items]
  File "C:\Users\15096\Anaconda3\lib\site-packages\sklearn\model_selection
\_validation.py", line 503, in _fit_and_score
    estimator.set_params(**parameters)
  File "C:\Users\15096\Anaconda3\lib\site-packages\sklearn\pipeline.py", l
ine 164, in set_params
    self._set_params('steps', **kwargs)
  File "C:\Users\15096\Anaconda3\lib\site-packages\sklearn\utils\metaestim
ators.py", line 50, in _set_params
    super().set_params(**params)
  File "C:\Users\15096\Anaconda3\lib\site-packages\sklearn\base.py", line
 224, in set_params
    (key, self))
ValueError: Invalid parameter select_k for estimator Pipeline(memory=None,
         steps=[('impute',
                 SimpleImputer(add_indicator=False, copy=True, fill_value=
None,
                               missing_values=nan, strategy='median',
                               verbose=0)),
                ('remove', VarianceThreshold(threshold=1)),
                ('scaler', MinMaxScaler(copy=True, feature_range=(0, 1))),
                ('select',
                 SelectKBest(k=10,
                             score_func=<function f_regression at 0x000001
586387F828>)),
                ('knn',
                 KNeighborsRegressor(algorithm='auto', leaf_size=30,
                                     metric='minkowski', metric_params=Non
e,
                                     n_jobs=None, n_neighbors=5, p=2,
                                     weights='uniform'))],
         verbose=False). Check the list of available parameters with `esti
mator.get_params().keys()`.
"""

The above exception was the direct cause of the following exception:

ValueError                                Traceback (most recent call las
t)
<ipython-input-7-b86b2a3c146c> in <module>
```

```
      7 from sklearn.model_selection import GridSearchCV
      8 select_grid=GridSearchCV(selectpipeline,param_grid2,scoring='neg_m
ean_squared_error',cv=tr_val_partition,n_jobs=-1, verbose=1)
----> 9 select_grid=select_grid.fit(X_train,y_train)
     10 y_test_select_pred=select_grid.predict(X_test)
     11 select_grid.best_params_

~\Anaconda3\lib\site-packages\sklearn\model_selection\_search.py in fit(se
lf, X, y, groups, **fit_params)
    686                 return results
    687
--> 688             self._run_search(evaluate_candidates)
    689
    690         # For multi-metric evaluation, store the best_index_, best
_params_ and

~\Anaconda3\lib\site-packages\sklearn\model_selection\_search.py in _run_s
earch(self, evaluate_candidates)
   1147     def _run_search(self, evaluate_candidates):
   1148         """Search all candidates in param_grid"""
-> 1149         evaluate_candidates(ParameterGrid(self.param_grid))
   1150
   1151

~\Anaconda3\lib\site-packages\sklearn\model_selection\_search.py in evalua
te_candidates(candidate_params)
    665                                for parameters, (train, test)
    666                                in product(candidate_params,
--> 667                                           cv.split(X, y, groups)))
    668
    669                 if len(out) < 1:

~\Anaconda3\lib\site-packages\joblib\parallel.py in __call__(self, iterabl
e)
    932
    933             with self._backend.retrieval_context():
--> 934                 self.retrieve()
    935             # Make sure that we get a last message telling us we a
re done
    936             elapsed_time = time.time() - self._start_time

~\Anaconda3\lib\site-packages\joblib\parallel.py in retrieve(self)
    831             try:
    832                 if getattr(self._backend, 'supports_timeout', Fals
e):
--> 833                     self._output.extend(job.get(timeout=self.timeo
ut))
    834                 else:
    835                     self._output.extend(job.get())

~\Anaconda3\lib\site-packages\joblib\_parallel_backends.py in wrap_future_
result(future, timeout)
    519         AsyncResults.get from multiprocessing."""
    520         try:
--> 521             return future.result(timeout=timeout)
    522         except LokyTimeoutError:
    523             raise TimeoutError()

~\Anaconda3\lib\concurrent\futures\_base.py in result(self, timeout)
    433                 raise CancelledError()
    434             elif self._state == FINISHED:
```

```
--> 435                 return self.__get_result()
    436             else:
    437                 raise TimeoutError()

~\Anaconda3\lib\concurrent\futures\_base.py in __get_result(self)
    382     def __get_result(self):
    383         if self._exception:
--> 384             raise self._exception
    385         else:
    386             return self._result

ValueError: Invalid parameter select_k for estimator Pipeline(memory=None,
         steps=[('impute',
                 SimpleImputer(add_indicator=False, copy=True, fill_value=
None,
                               missing_values=nan, strategy='median',
                               verbose=0)),
                ('remove', VarianceThreshold(threshold=1)),
                ('scaler', MinMaxScaler(copy=True, feature_range=(0, 1))),
                ('select',
                 SelectKBest(k=10,
                             score_func=<function f_regression at 0x000001
586387F828>)),
                ('knn',
                 KNeighborsRegressor(algorithm='auto', leaf_size=30,
                                     metric='minkowski', metric_params=Non
e,
                                     n_jobs=None, n_neighbors=5, p=2,
                                     weights='uniform'))],
         verbose=False). Check the list of available parameters with `esti
mator.get_params().keys()`.
```

In [ ]: