**III.인공지능 기반 물질 개발 및 거동 분석**

1. 유기용매 막분리 소재 분석

유기용매 나노여과분리막 (Organic solvent nanofiltration, OSN)은 용질의 선택적 분리 기능과 효율적인 물질 투과 능력을 갖고있어 난분해성 폐수의 분리에 탁월한 성능을 보인다. 이로 인해 유기용매 나노여과분리막은 화학, 환경, 의료, 제약, 식품공업 등 다양한 응용이 가능한 까닭에 산업적 수요 기대가 높아지고 있다.

유기용매 나노여과 분리막은 친환경적이고 효율적인 분리 기술이므로, 분리막의 성능 향상을 통해 분리 공정 효율을 개선시키려는 시도가 많이 진행된다. 하지만 분리막 성능은 분리막의 특성 (Property)과 사용 환경 (Operating condition)에 따라 큰 차이가 발생하여 예측이 상당히 어려운 실정이다.

따라서, 고성능 분리막 설계를 위해서는 반복적인 실험에 의존하는 것이 보편적이다. 하지만, 시행착오법은 많은 예산과 시간을 소모하여 다양한 조건에서의 분리막 설계를 지연시키는 한계점을 갖는다. 따라서 새로운 분리막 개발을 통한 분리 공정 개선은 상당히 소극적이게 이루어지며 낮은 성공률을 갖는다. 결국, 분리막 R&D의 효율을 증가시키기 위해서는 분리막 성능 예측이 필수적이다. 특히, 분리막의 주요 변수를 규명하여 정형화된 분리막 성능의 화학적 차원 (Chemical dimension)을 정의한다면, 분리막 성능 예측 모델 개발의 초석이 될 수 있다.

본 실습에서는 차원 축소 방법론 기반 분리막 성능의 경향 파악과 주요 변수 규명을 목적으로한다.

**예제 1: 주성분 분석을 이용한 설계 인자 차원 축소**

예제는 Python 3.5 프로그래밍 언어를 기준으로 Jupyter 개발환경에서 작성되었다. 예제 실습을 위해 Anaconda 프로그램을 아래 URL로부터 다운로드할 수 있다.

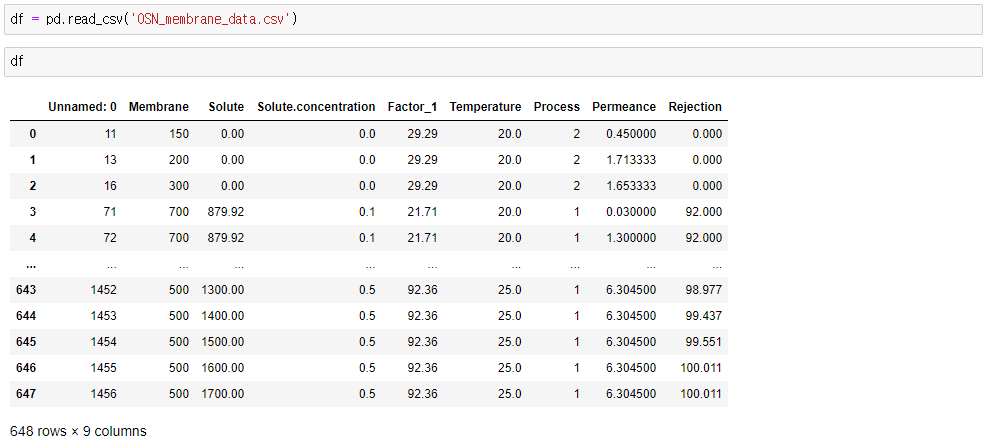
https://www.anaconda.com/

* **데이터 링크**

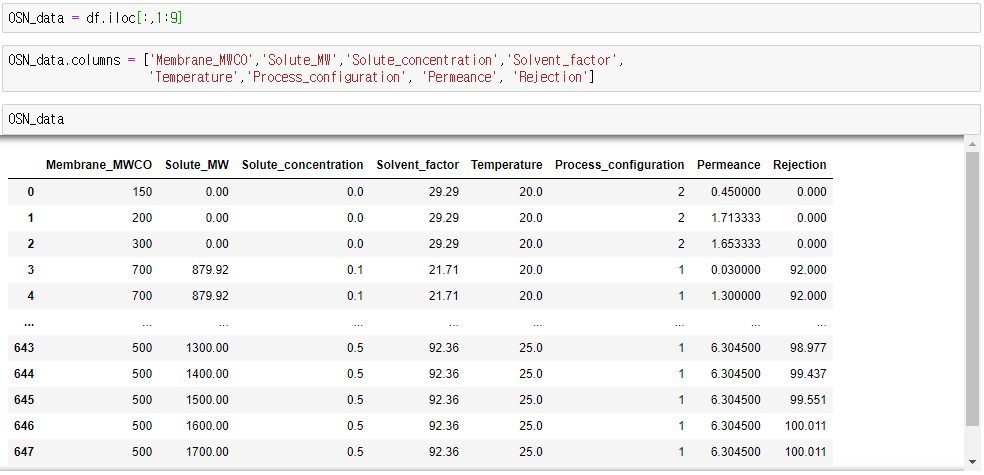
(OSN 데이터 링크 삽입)



* ‘pandas’ package는 데이터 편집을 위한 도구로, python 언어에서 매우 유용하게 사용된다. OSN 데이터 불러오기 및 데이터 편집 (컬럼 및 이름 설정)을 위해 본 실습에서 사용한다.
* ‘numpy’ package는 수학적 기능이 탑재된 도구로, 난수 발생과 대수적 계산을 위해 사용된다. 본 실습에서는 데이터의 형태 변환 및 벡터 계산에 사용한다.
* ‘matplotlib.pyplot’ package는 시각화를 위한 그래프 도구로, 다양한 그래프를 그리기 위해 사용된다.
* ‘pandas’ package를 import하고 ‘pd’로 축약해 사용한다.
* ‘numpy’ package를 import하고 ‘np’로 축약해 사용한다.
* ‘matplotlib.pyplot’ package를 import하고 ‘plt’로 축약해 사용한다.



* 데이터는 ‘OSN\_membrane\_data.csv’로 Jupyter 환경에 저장됐다. ‘pd’에 내장된 read.csv 함수를 사용하여 해당 데이터를 Jupyter script 환경에 불러온다.
* 해당 데이터는 Jupyter script 환경에서 df로 명명되었으며, df를 타이핑하여 데이터를 읽어온다.



* 행 번호 (row name)가 포함된 첫번째 열을 제거하고 사용한다. 해당 열은 첫번째 열이고, python에서는 0번째 열이다.
* 행 번호가 삭제된 데이터는 ‘OSN\_data’로 명명하여 사용한다. ‘.colums’ 함수를 사용하여 ‘OSN\_data’의 열 이름을 변수 이름으로 설정한다.

**실습: 기계학습 기반 막분리 소재 설계 인자 규명**

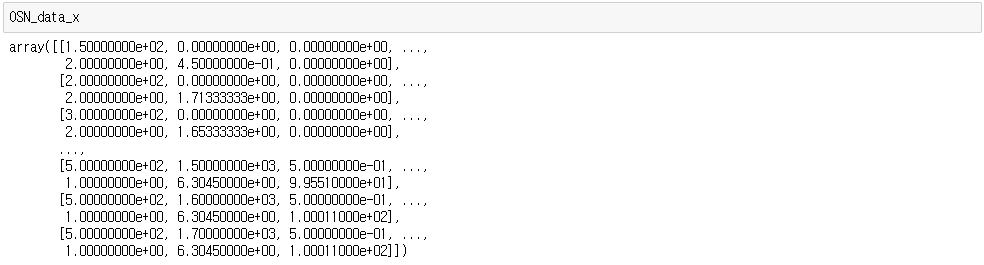
* **주성분 분석 방법론 실습**
* 데이터 전처리 (Data preprocessing)



* ‘sklearn’ package는 계산을 위한 유용한 함수를 다수 내장하고 있다. ‘sklearn.preprocessing’의 ‘StandardScaler’ 함수를 import하도록 한다.

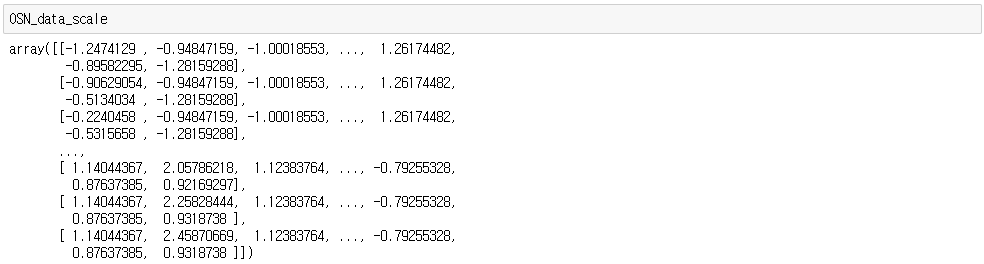


* ‘OSN\_data’의 값만을 ‘OSN\_data\_x’로 가져온다. 이때 ‘.loc’ 함수를 사용하여 데이터를 지정하고, ‘.value’함수를 사용하여 값을 가져온다.



* ‘OSN\_data\_x’를 확인한다. 데이터는 Array 형식으로 편집됐다.





* ‘StandardScaler’ 함수를 사용하여 ‘OSN\_data\_x’ 데이터의 표준화를 진행한다.
* 화학적차원을 정보차원으로 투영 (Chemical space to Information space)



* ‘sklearn.decomposition’의 ‘PCA’함수를 import하도록 한다.

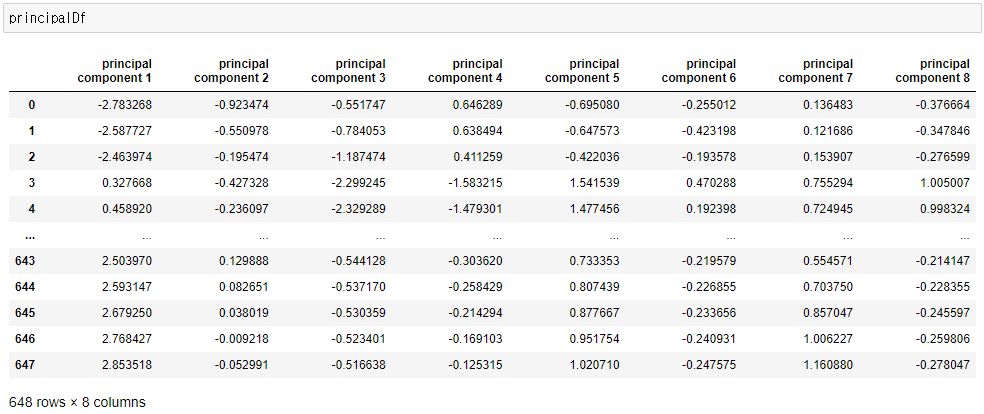


* ‘PCA’함수를 사용하여 8개의 성분을 갖는 주성분을 생성한다.

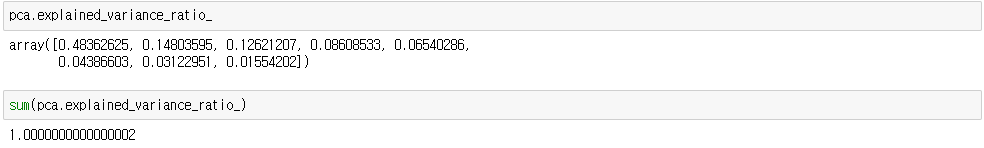


* ‘pca.fit\_transform’ 함수를 사용하여 정규화된 ‘OSN\_data\_scale’ 데이터를 정보차원으로 변환한다. 정보 차원으로 변환된 데이터는 ‘principalComponents’로 명명한다.



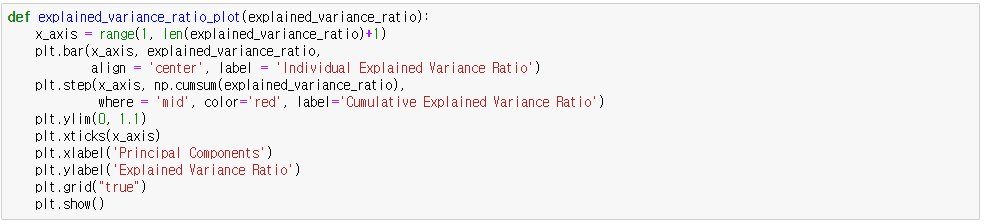


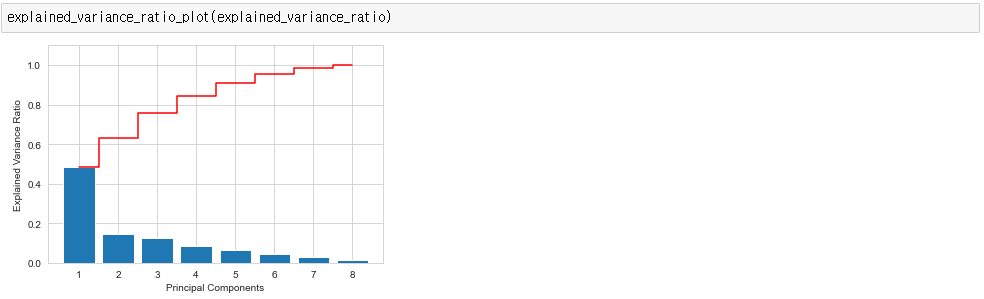
* ‘pd’ 패키지의 ‘DataFrame’ 함수를 사용하여 ‘principalComponents’ 데이터을 ‘principalDf’로 명명하고 열 이름 (Column name)을 수정한다.



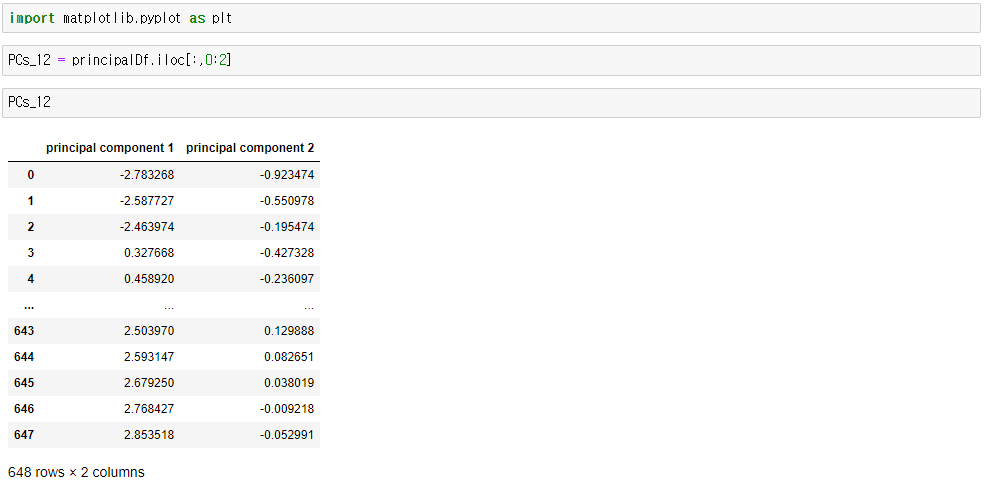
* ‘.explained\_variance\_ratio’ 함수를 통해 미리 생성해둔 ‘pca’의 각 주성분의 데이터 설명도 (dimension coverage)를 확인할 수 있다.
* 앞에서부터 첫번째 주성분 (PC1)이며 혼자서 전체 데이터의 48.3% 정도 설명할 수 있는 것을 확인한다.
* 전체 주성분 (PC1~8)의 데이터 설명도 총합은 100%이다.
* 시각화를 위한 차원 개수 선택 (Cumulative variance plot)



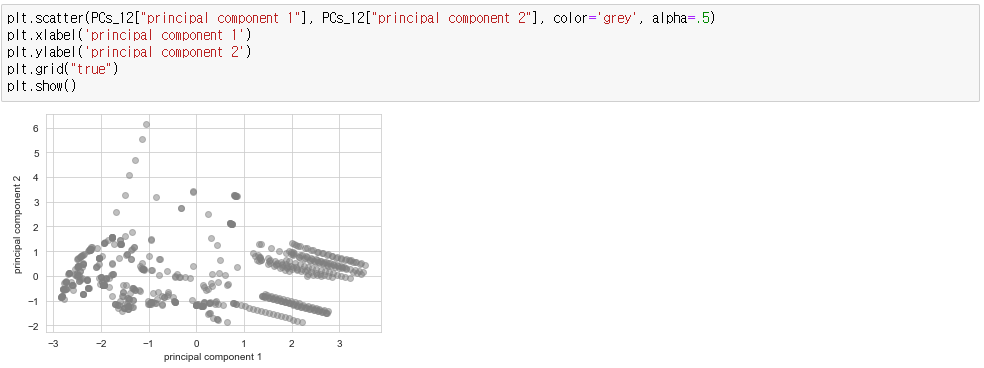




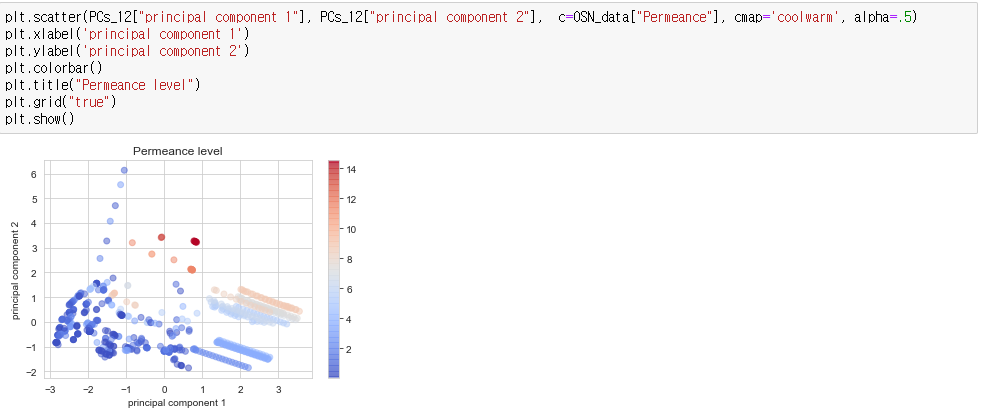
* 차원 축소는 전체 데이터의 80~90% 이상의 설명력을 갖는 주성분 개수를 선택하면 된다. 여기서 누적 그래프를 확인하면 최소 4개이상의 주성분이 선택되야한다는 것을 알 수 있다.
* 시각화를 위해 함수를 정의하여 사용한다. ‘def’를 사용하여 함수를 정의할 수 있고, ‘plt’ package를 사용해 시각화 한다.
* 나노여과 분리막 정보 차원 시각화



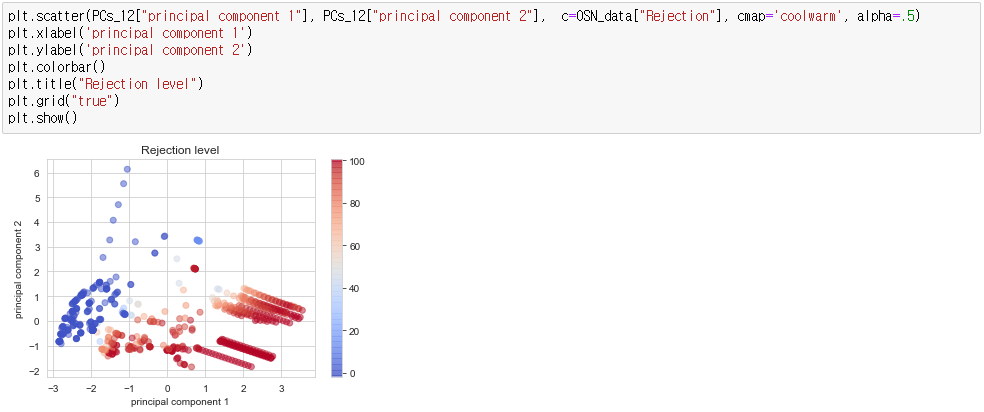
* ‘pd’ package의 ‘iloc’ 함수를 활용하여 ‘principalDf’ 데이터의 1~2번째 열을 추출하여 ‘PC\_12’로 명명한다.
* 나노여과 분리막 정보 차원 시각화 (2D visualization)



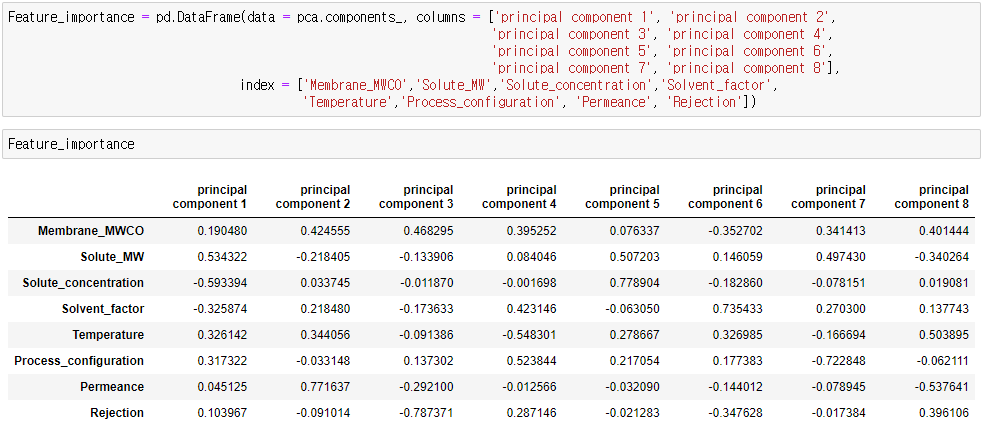
* ‘plt’ package의 ‘scatter’ 함수를 활용하여 시각화한다. ‘PC\_12’의 첫번째 열과 두번째 열을 좌표로 사용한다.



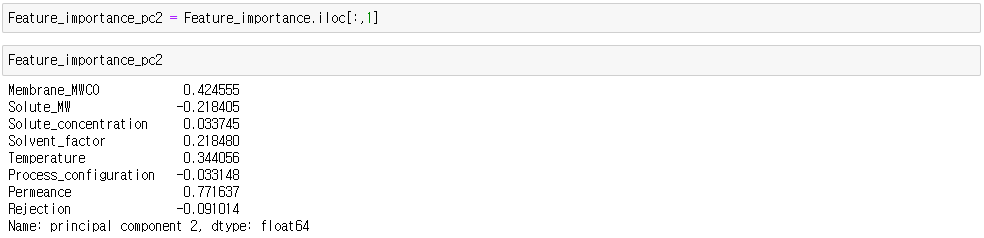
* ‘plt’ package의 ‘scatter’ 함수를 활용하여 시각화한다. ‘c‘ 함수 인자 (Argument)를 사용하여 ‘OSN\_data’의 Permeance 열을 색으로 표현할 수 있다.



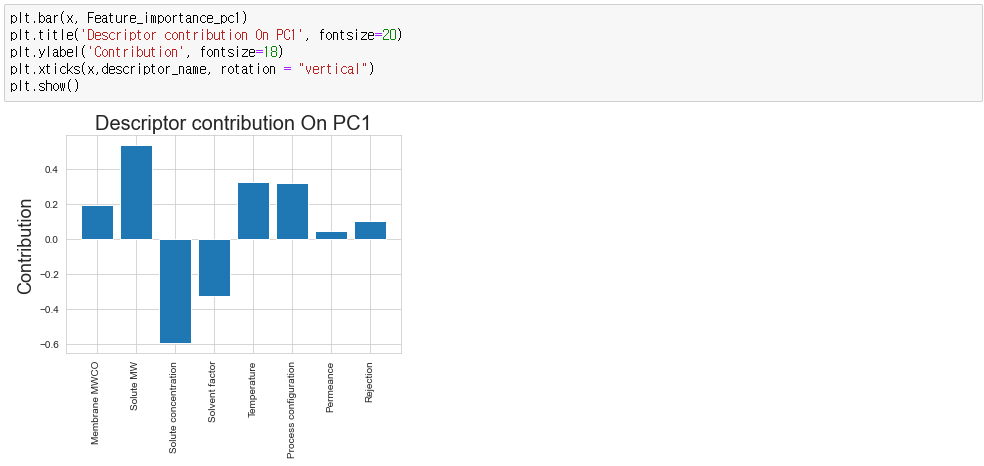
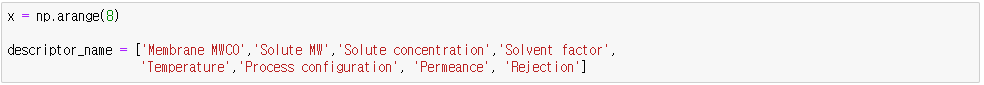
* ‘plt’ package의 ‘scatter’ 함수를 활용하여 시각화한다. ‘c‘ 함수 인자 (Argument)를 사용하여 ‘OSN\_data’의 Rejection 열을 색으로 표현할 수 있다.
* 나노여과 분리막 변수 중요도 파악 (Descriptor importance identification)

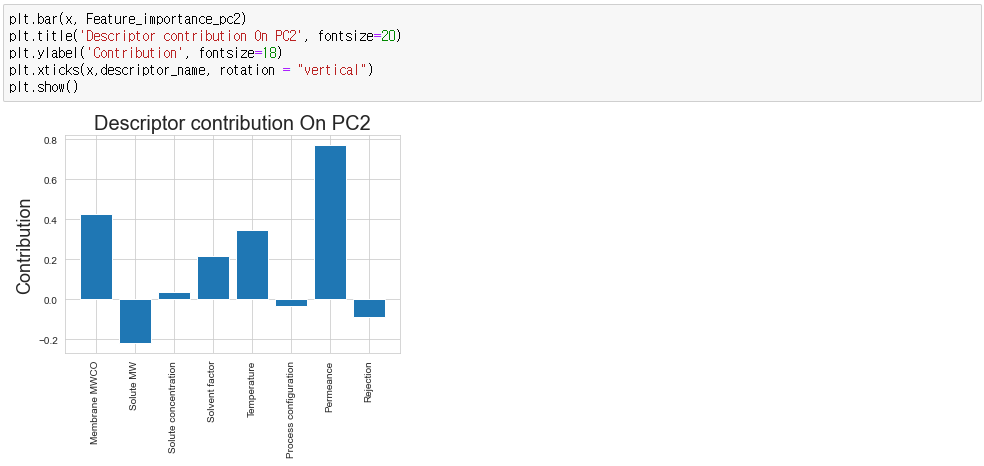


* ‘.components\_’ 함수를 사용하면 ‘pca’의 변수들의 주성분에 대한 기여도를 확인할 수 있다.
* 각 성분의 이름을 열 이름으로, 행은 변수 이름으로 설정하고 데이터 를 ‘Feature\_importance’로 명명한다.



* 가장 데이터 설명력이 좋은 두 성분 (PC1, PC2)를 추출해 사용한다.
* 각 PC1과 PC2에대한 기여도 정보를 추출하여 ‘Feature\_importance\_pc1’과 ‘Feature\_importance\_pc2’로 명명한다.





* ‘np’ package의 ‘array’ 함수를 사용하여 8개의 name space를 갖는 ‘x’ array를 생성한다.
* ‘plt’ package의 ‘bar’ 함수를 사용하여 ‘Feature\_importance\_pc1’과 ‘Feature\_imporatnce\_pc2’를 시각화한다.
* 결과로 PC1에서 Solute M.W와 Solute Concentration, PC2에서는 MWCO가 가장 주요한 변수로 식별되었다.

**Reference**

G. Modla and P. Lang, “*Removal and Recovery of Organic Solvents from Aqueous Waste Mixtures by Extractive and Pressure Swing Distillation*”, Industrial & Engineering Chemistry Research, Vol. 51, Issue 35, 11473-11481, 2012

우창화, “*특허 및 논문 게재 분석을 통한 수처리용 분리막의 연구동향*”, Appl. Chem. Eng., Vol. 28, No. 4, 2017

Dimitar Peshev, Andrew G. Livingston, “*OSN Designer, a tool for predicting organic solvent nanofiltration technology performance using Aspen One, MATLAB and CAPE OPEN*”, Chemical Engineering Science, Vol. 104, 975-987, 2013

BCC Research, “*Global Markets and Technologies for Nanofiltration*”, 2014

김경호, “*나노여과 막의 기술 및 산업시장 동향*”, 한국과학기술정보연구원 중소기업혁신본부, 2015

MECO, “Industrial reverse osmosis products”

코오롱 멤브레인, “Cleanfill®”

도레이첨단소재, “NE8040”

Jiahui Hu, Changsu Kim, Peter Halasz, Jeong F. Kim, Jiyong Kim, Gyorgy Szekely, “*Artificial intelligence for performance prediction of organic solvent nanofiltration membranes*”, Journal of Membrane Science, Vol. 619, 118513, 2021

Patrizia Marchetti, Andrew G. Livingston, “*Predictive membrane transport models for Organic Solvent Nanofiltration: How complex do we need to be?*”, Journal of Membrane Science, vol. 476, 530-533, 2015

Herve Abdi, Lynne J. Williams, “Principal component analysis”, WIREs Computational statistics, 433-459, vol.2, 2010