Random forest (Gini,Entropy)

# 라이브러리 링크

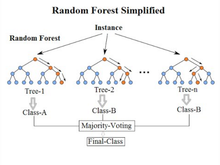
**install :**

<https://scikit-learn.org/stable/install.html>

**github :**

https://github.com/scikit-learn/scikit-learn/blob/82df48934eba1df9a1ed3be98aaace8eada59e6e/sklearn/ensemble/\_forest.py

# 기초 설명



random forest학습법은 분류, 회기 분석 등에 사용되는 앙상블 학습 방법의 일종으로.

훈련 과정에서 다수의 결정 트리를 구성하여 이 다수의 결정 트리들을 학습하는 방법으로 random forest는 검출, 분류 또는 회기 등 다양한 분야에서 활용중이다.

# 버전 정보

* NumPy >= 1.14.6 (pip install numpy)
* Scipy >= 1.1.0 (pip install scipy)
* Joblib >= 0.11 (pip install joblib
* Threadpoolctl >= 2.0.0 (pip install threadpoolctl)
* pandas >= 1.2.4 (pip install pandas)

# 데이터셋 설명 및 출처

* Housing.scv: the Data set of housing price, based on certain factors like house area, bedrooms, furnished, nearness to mainroad, etc.
* Harrison, D. and Rubinfeld, D.L. (1978) Hedonic prices and the demand for clean air. J. Environ. Economics and Management 5, 81–102.  
  Belsley D.A., Kuh, E. and Welsch, R.E. (1980) Regression Diagnostics. Identifying Influential Data and Sources of Collinearity. New York: Wiley.
* Sources : <https://www.kaggle.com/yasserh/housing-prices-dataset>

# 코드 설명

* 데이터셋으로 부터 불러온 데이터를 기반으로 이를 Random Forest 학습법을 통해 모델에 학습시켜 이를 통해 집의 가격 및 화장실수,창고수, 등등의 정보를 기반으로 침실수를 예측하는 모델.
* 이에대해 Random Forest 학습 과정에서의 gini,entropy 방식의 차이 및 Random Forest 내부에서 사용되는 트리 갯수의 차이 등에 따라 변화하는 모델의 정확도를 출력함.

# 검증 방법

* 코드 내부에서 데이터셋을 학습용 데이터셋과 검증용 데이터셋으로 나누어 이를 검증함.

(test\_size = 0.2, random\_state = 42)

* 추가적으로 sklearn의 accuracy\_score 함수를 사용하여 실제값과 예측값 사이의 적합도를 평가.