

Introduction to Machine Learning

Kyungsik Han

본 영상에서 다룰 내용

- 머신러닝 기초 개념
 - 회귀(Regression)
 - 분류(Classification)
 - 군집(Clustering)
- scikit-learn 라이브러리 소개

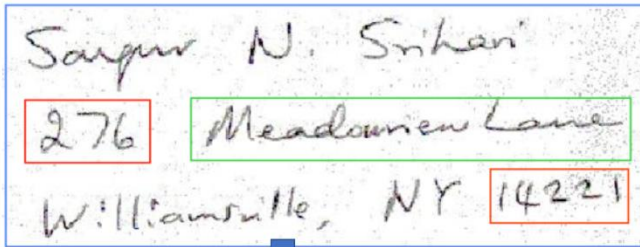
Machine Learning is

- ... learning from data
- ... on its own
- ... discovering hidden patterns
- ... data-driven decisions

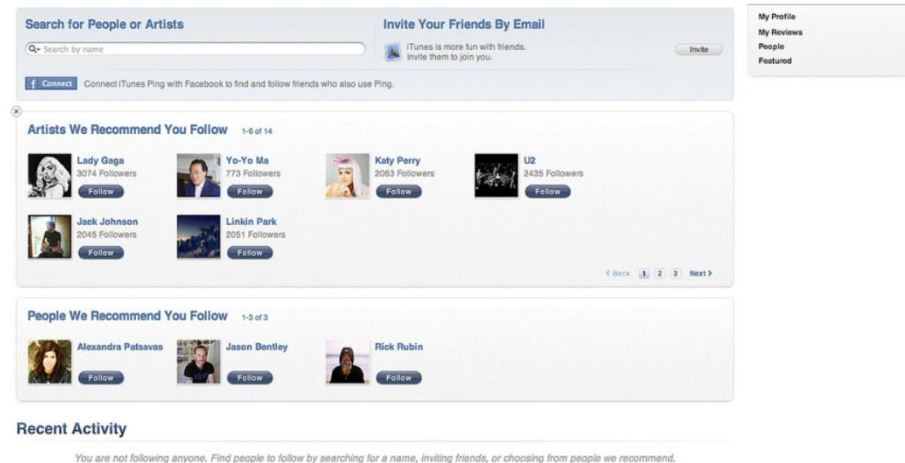


Machine Learning 예제

- 신용카드 이상 사용 감지
- 손글씨 감지
- 웹사이트에서의 추천기능



ZIP Code: 14221
Primary number: 276



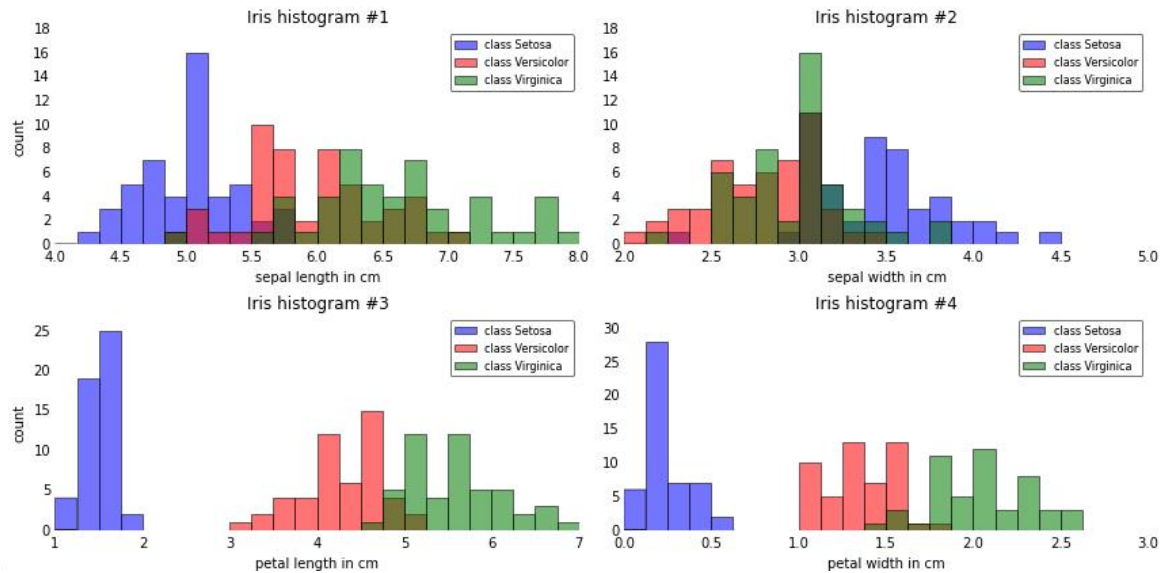
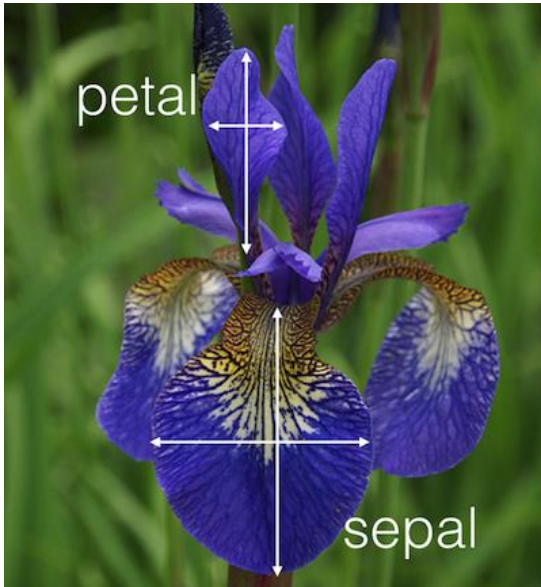
Machine Learning Models

- Learn from data
- Discover patterns and trends
- Allow for data-driven decisions
- Used in many different applications

머신 러닝의 종류

Classification (분류)

- 목적: 어떤 데이터에 대한 category를 예측하는 것



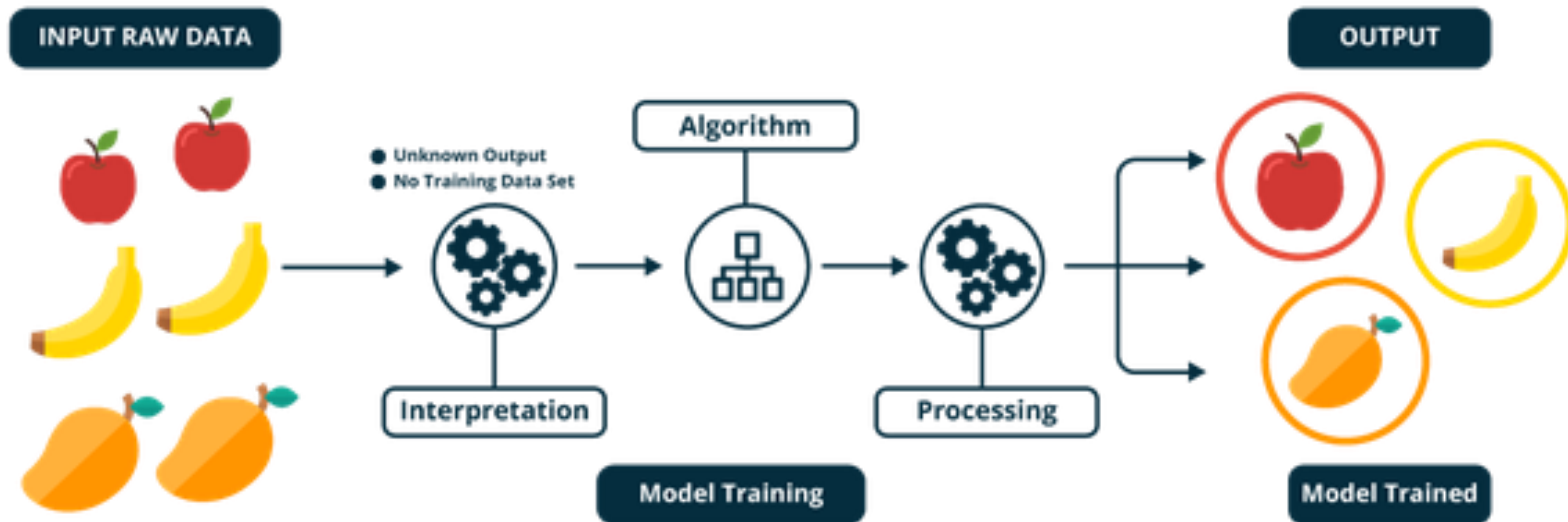
Regression (회귀)

- 목적: 숫자화된 데이터로 예측하는 것



Clustering (군집화)

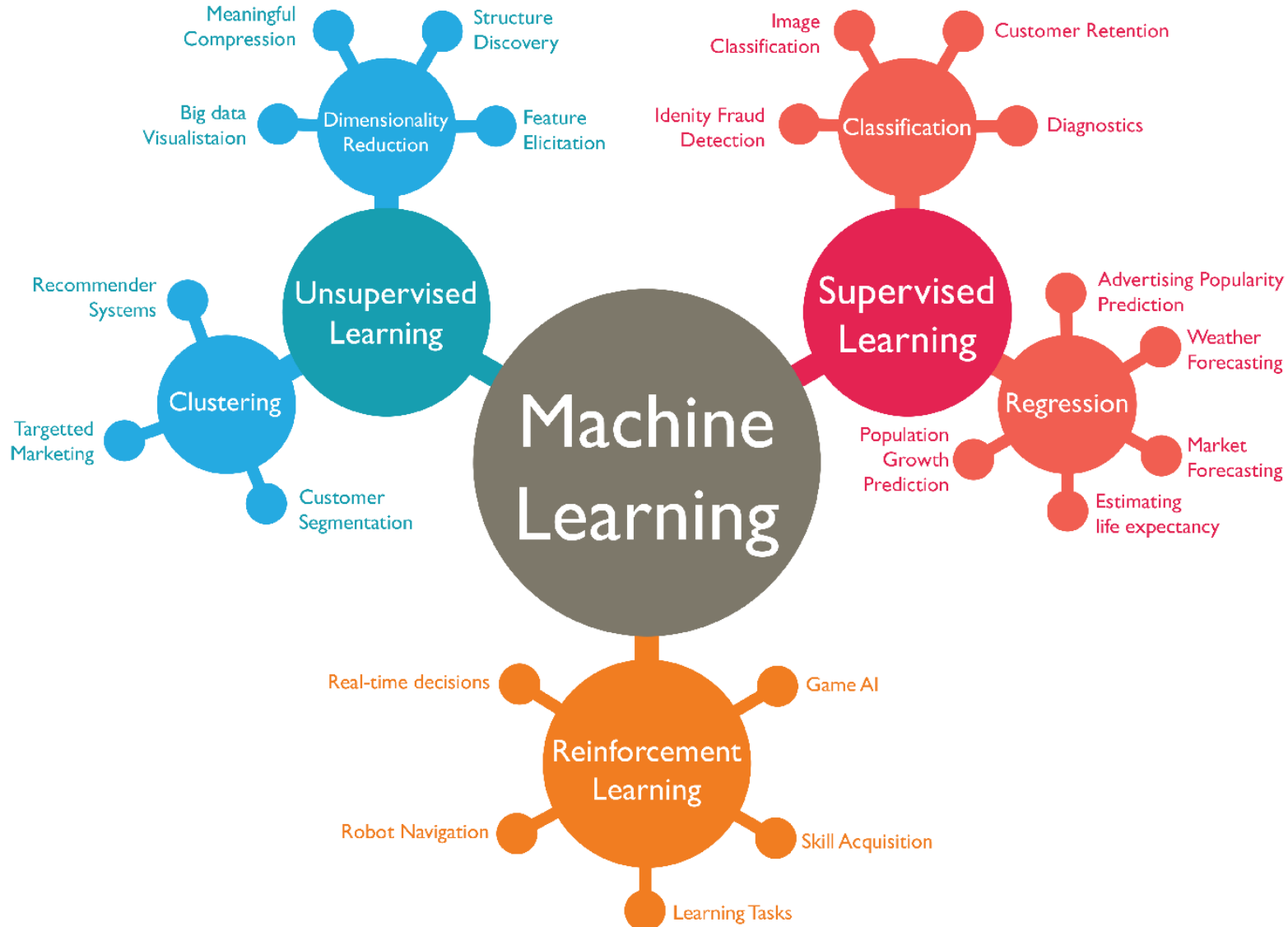
- 목적: 비슷한 특징을 가진 아이템을 그룹화



Supervised vs Unsupervised

- 지도 학습(Supervised Approaches)
 - 정답값 제공: Target (what model is predicting) is provided
 - 'Labeled' data
 - 분류(Classification) & 회귀(regression) are supervised
- 비지도 학습(Unsupervised Approaches)
 - 정답값 미제공: Target is unknown or unavailable
 - 'Unlabeled' data
 - 군집 분석(Cluster analysis) is unsupervised

Supervised vs Unsupervised



Machine Learning 용어 (Terminology)

Samples		Variable				
		ID	Date	MinTemp	MaxTemp	RainFall
		1	2018-03-01	1	10	0.1
		2	2018-03-02	2	16	0.0
		3	2018-03-03	-1	12	0.0
		4	2018-03-04	3	15	0.0

Other names for 'sample'

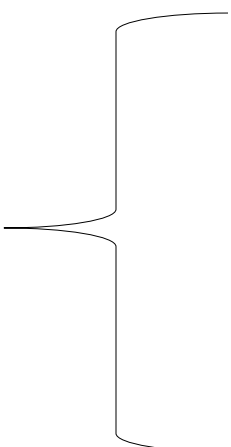
row

instance

record

example

observation



ID	Date	MinTemp	MaxTemp	RainFall
1	2018-03-01	1	10	0.1
2	2018-03-02	2	16	0.0
3	2018-03-03	-1	12	0.0
4	2018-03-04	3	15	0.0

Other names for 'variable'

variable

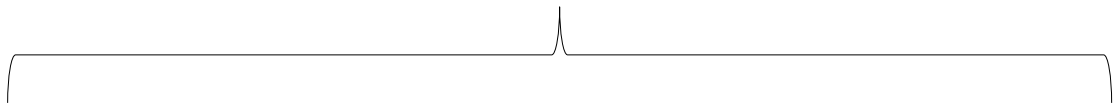
feature

column

dimension

attribute

field



ID	Date	MinTemp	MaxTemp	RainFall
1	2018-03-01	1	10	0.1
2	2018-03-02	2	16	0.0
3	2018-03-03	-1	12	0.0
4	2018-03-04	3	15	0.0

Sample and Variable

Sample



- Instance
- Record
- Row
- Observation
- ...

Variable



- Feature
- Field
- Column
- ...

ID	Date	MinTemp	MaxTemp	RainFall
1	2018-03-01	1	10	0.1
2	2018-03-02	2	16	0.0
3	2018-03-03	-1	12	0.0
4	2018-03-04	3	15	0.0

Numeric

Categorical

scikit-learn (or sklearn) library

scikit-learn 라이브러리 (1/4)

- 파이썬에서 기계 학습을 위한 오픈 소스 라이브러리
- NumPy, SciPy, matplotlib 기반으로 만들어짐
- 활발한 개발 커뮤니티
- 라이브러리의 지속적인 발전

scikit-learn

Machine Learning in Python

[Getting Started](#)[Release Highlights for 0.24](#)[GitHub](#)

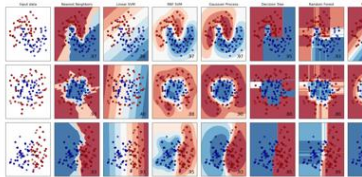
- Simple and efficient tools for predictive data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

Identifying which category an object belongs to.

Applications: Spam detection, image recognition.

Algorithms: SVM, nearest neighbors, random forest, and more...



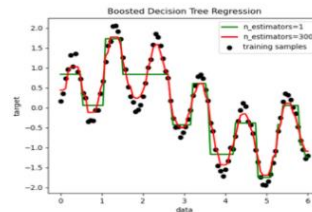
Examples

Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.

Algorithms: SVR, nearest neighbors, random forest, and more...



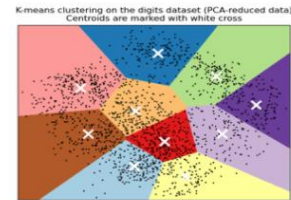
Examples

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes

Algorithms: k-Means, spectral clustering, mean-shift, and more...



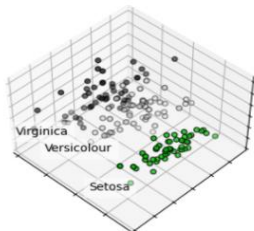
Examples

Dimensionality reduction

Reducing the number of random variables to consider.

Applications: Visualization, Increased efficiency

Algorithms: k-Means, feature selection, non-negative matrix factorization, and more...

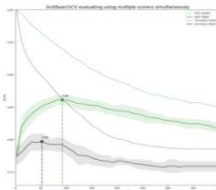


Model selection

Comparing, validating and choosing parameters and models.

Applications: Improved accuracy via parameter tuning

Algorithms: grid search, cross validation, metrics, and more...

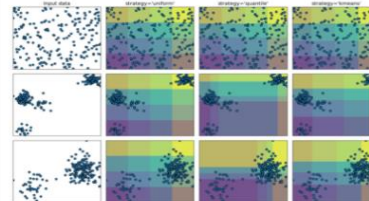


Preprocessing

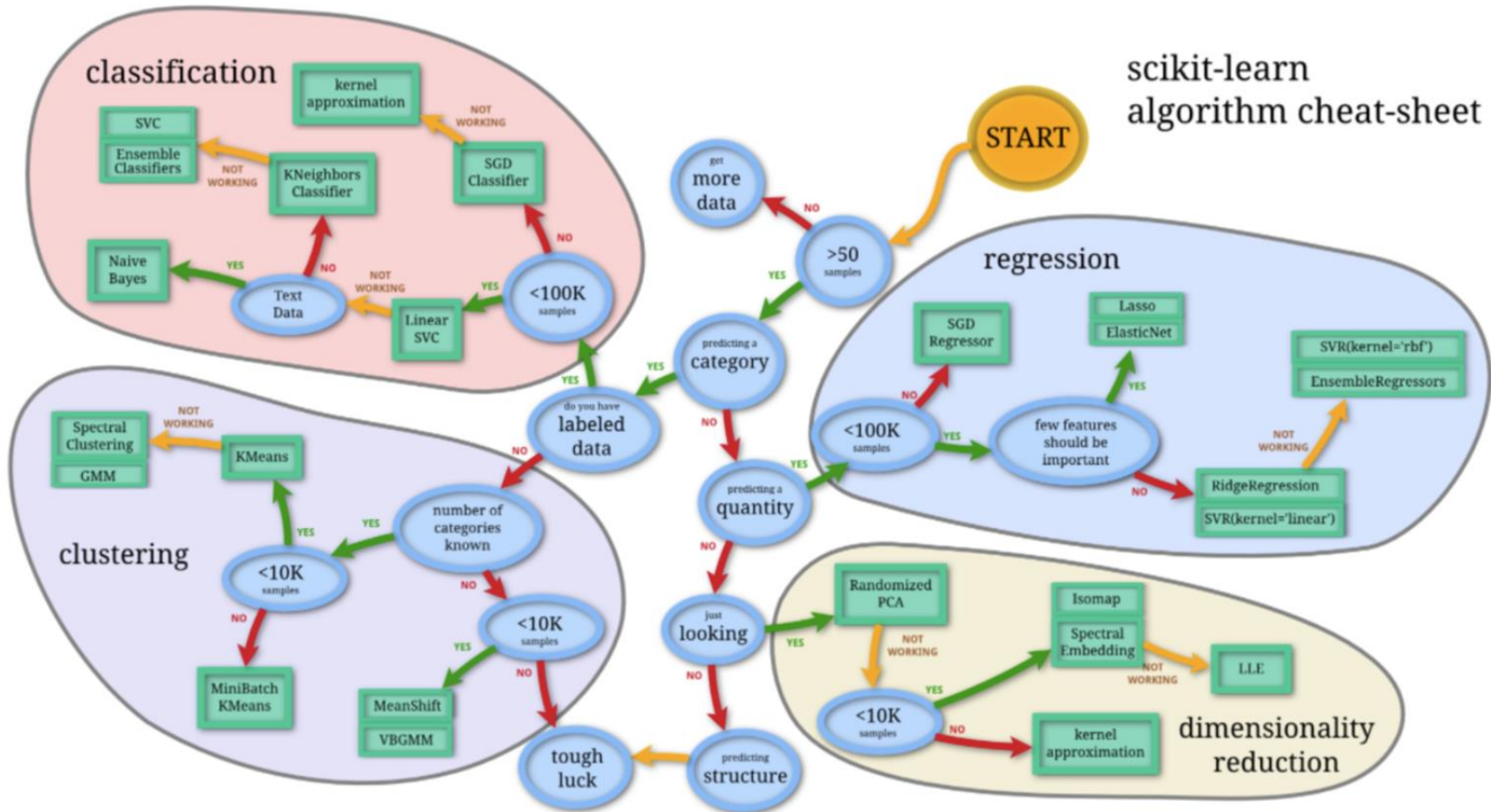
Feature extraction and normalization.

Applications: Transforming input data such as text for use with machine learning algorithms.

Algorithms: preprocessing, feature extraction, and more...



한눈에 보기



scikit-learn 라이브러리 (2/4)

- 전처리(pre-processing) 기능
 - 특성 스케일링(Scaling of features): remove mean and keep unit variance
 - 정규화(Normalization to have unit norm)
 - 이진화(Binarization to turn data into 0 or 1 format)
 - 원핫인코딩 기법(One Hot Encoding for categorical features)
 - 결측데이터 관리(Handling of missing values)
 - 상위단계특성 생성(Generating higher order features)
 - 맞춤 변환기 생성(Build custom transformations)

scikit-learn 라이브러리 (3/4)

- 차원 축소(영어 그대로)
 - Principal Component Analysis (PCA)
 - Singular Value Decomposition (SVD)
 - Factor Analysis
 - Independent Component Analysis
 - Matrix Factorization
 - Latent Dirichlet Allocation (LDA)

scikit-learn 라이브러리 (4/4)

- 모델 선택
 - 교차검증: Provides methods for Cross Validation (CV)
 - 하이퍼 파라미터 튜닝: Library functions for tuning hyper parameters
 - 모델 평가 지표: Model Evaluation mechanisms to measure model performance
 - 모델 평가의 시각화: Plotting methods for visualizing scores to evaluate models

다음 영상에서 배울 내용

- 머신러닝 Mini project 실습
 - 축구데이터 분석
 - 데이터 살펴보기
 - 데이터 정제
 - 데이터 상관관계
 - 데이터 시각화
 - 데이터 모델링

수고하셨습니다