

기계학습

ML프로그래밍을 위한 라이브러리

Numpy 라이브러리

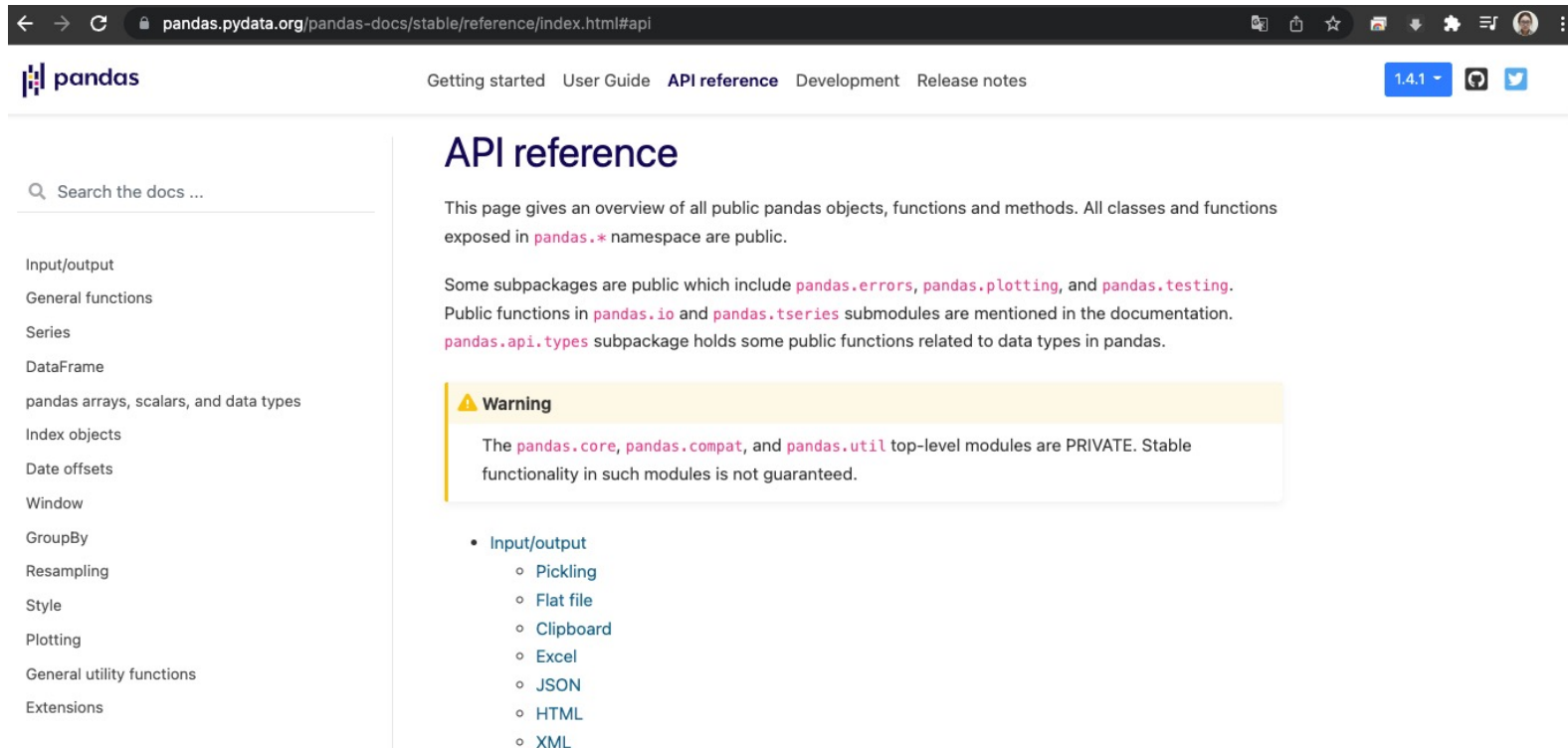
- 넘파이(numpy)
- 판다스(Pandas)
- 맷플랏립(Matplotlib)

Numpy 라이브러리

- 넘파이(Numpy)란?
 - Numerical Python으로 수치계산을 위해 만들어진 파이썬 라이브러리
 - 넘파이 배열(ndarray)이라는 자료구조를 사용함
 - 넘파이 배열이란 다차원 배열과 행렬을 지원하고 벡터, 행렬 등의 연산을 쉽고 빠르게 수행
- 넘파이 라이브러리 불러오기
 - `import numpy as np`
 - as 뒤에 numpy라 해도 되지만 간결성을 위해 관례적으로 np를 사용함
- 넘파이 실습
 - <https://www.kaggle.com/yukyungchoi/2022-ml-numpy-cheatsheet>

판다스 라이브러리

- 판다스(Pandas)란?
 - 파이썬을 이용한 데이터 처리/분석 작업의 필수 라이브러리
 - 판다스 공식 문서
 - <https://pandas.pydata.org/pandas-docs/stable/>



The screenshot shows the pandas API reference page. The browser address bar displays `pandas.pydata.org/pandas-docs/stable/reference/index.html#api`. The page header includes the pandas logo, navigation links (Getting started, User Guide, API reference, Development, Release notes), and a version dropdown set to 1.4.1. A search bar on the left is labeled "Search the docs ...". A sidebar on the left lists various pandas topics: Input/output, General functions, Series, DataFrame, pandas arrays, scalars, and data types, Index objects, Date offsets, Window, GroupBy, Resampling, Style, Plotting, General utility functions, and Extensions. The main content area is titled "API reference" and contains an overview of public pandas objects, functions, and methods. It mentions that all classes and functions are public, except for those in the `pandas.*` namespace. It also lists some subpackages: `pandas.errors`, `pandas.plotting`, and `pandas.testing`. Public functions in `pandas.io` and `pandas.tseries` submodules are mentioned. The `pandas.api.types` subpackage holds some public functions related to data types. A yellow warning box states that the `pandas.core`, `pandas.compat`, and `pandas.util` top-level modules are PRIVATE and their stable functionality is not guaranteed. A bulleted list under "Input/output" includes: Pickling, Flat file, Clipboard, Excel, JSON, HTML, and XML.

← → ↻ 🔒 pandas.pydata.org/pandas-docs/stable/reference/index.html#api

pandas Getting started User Guide **API reference** Development Release notes 1.4.1

API reference

This page gives an overview of all public pandas objects, functions and methods. All classes and functions exposed in `pandas.*` namespace are public.

Some subpackages are public which include `pandas.errors`, `pandas.plotting`, and `pandas.testing`. Public functions in `pandas.io` and `pandas.tseries` submodules are mentioned in the documentation. `pandas.api.types` subpackage holds some public functions related to data types in pandas.

Warning

The `pandas.core`, `pandas.compat`, and `pandas.util` top-level modules are PRIVATE. Stable functionality in such modules is not guaranteed.

- Input/output
 - Pickling
 - Flat file
 - Clipboard
 - Excel
 - JSON
 - HTML
 - XML

판다스 라이브러리

- 판다스 라이브러리 불러오기
 - `Import pandas as pd`
- 판다스 데이터 구조
 - 자료구조 3요소: 시리즈 (Series), 데이터프레임 (DataFrame), 패널 (Panel)
 - 데이터프레임이 가장 많이 사용됨
 - 시리즈 란?
 - 1차원 배열의 값에 각 값에 대응하는 인덱스를 부여할 수 있는 구조
 - 데이터프레임 이란?
 - 행과 열을 가지는 자료구조로, 2차원 리스트를 매개변수로 전달
- 판다스 데이터프레임 실습
 - <https://www.kaggle.com/yukyungchoi/2022-ml-pandas-cheatsheet>
- 판다스 프로파일링
 - <https://www.kaggle.com/yukyungchoi/2022-ml-pandas-profiling>

판다스 라이브러리

- (실습) 4주차 실습 과제1과 실습 과제2에 사용되는 데이터를 판다스 데이터프레임으로 읽고 프로파일링 해보기
 - <https://www.kaggle.com/c/2022-ml-w4p1>
 - <https://www.kaggle.com/t/e4d47e37ea3b41879d6b4670bc9f06b1>

[기계학습][4주차][실습과제1] KNN을 이용하여 재배환경 별 작물 종류 예측 문제를 해결하라

- 캐글 리더보드: <https://www.kaggle.com/c/2022-ml-w4p1>
- 과제 제출 (1) : 캐글리더보드에 답안 제출하여 베이스라인 넘기 후 캐글 노트북 담당 조교에게 공유
- 과제 제출 (2) : KNN의 하이퍼파라미터 변경에 따른 성능결과 분석 리포트 A4 한장 이내로 제출
- 과제 제출 기한: 2022년 04월 03일 오후 11시 59분
- 제출할 곳: admin@rcv.sejong.ac.kr
 - 이메일 제목 : [기계학습][4주차][실습과제1] 재배환경별 작물종류 예측 (학번_이름)

<https://www.kaggle.com/yukyungchoi/2022-ml-w4p1-profiling>

[기계학습][4주차][실습과제2] KNN을 이용하여 자동차 가격 예측 문제를 해결하라

- 캐글 리더보드: <https://www.kaggle.com/t/e4d47e37ea3b41879d6b4670bc9f06b1>
- 과제 제출 (1) : 캐글리더보드에 답안 제출하여 베이스라인 넘기 후 캐글 노트북 담당 조교에게 공유
- 과제 제출 (2) : KNN의 하이퍼파라미터 변경에 따른 성능결과 분석 리포트 A4 한장 이내로 제출
- 과제 제출 기한: 2022년 04월 03일 오후 11시 59분
- 제출할 곳: admin@rcv.sejong.ac.kr
 - 이메일 제목 : [기계학습][4주차][실습과제2] 자동차 가격예측 (학번_이름)

<https://www.kaggle.com/yukyungchoi/2022-ml-w4p2-profiling>

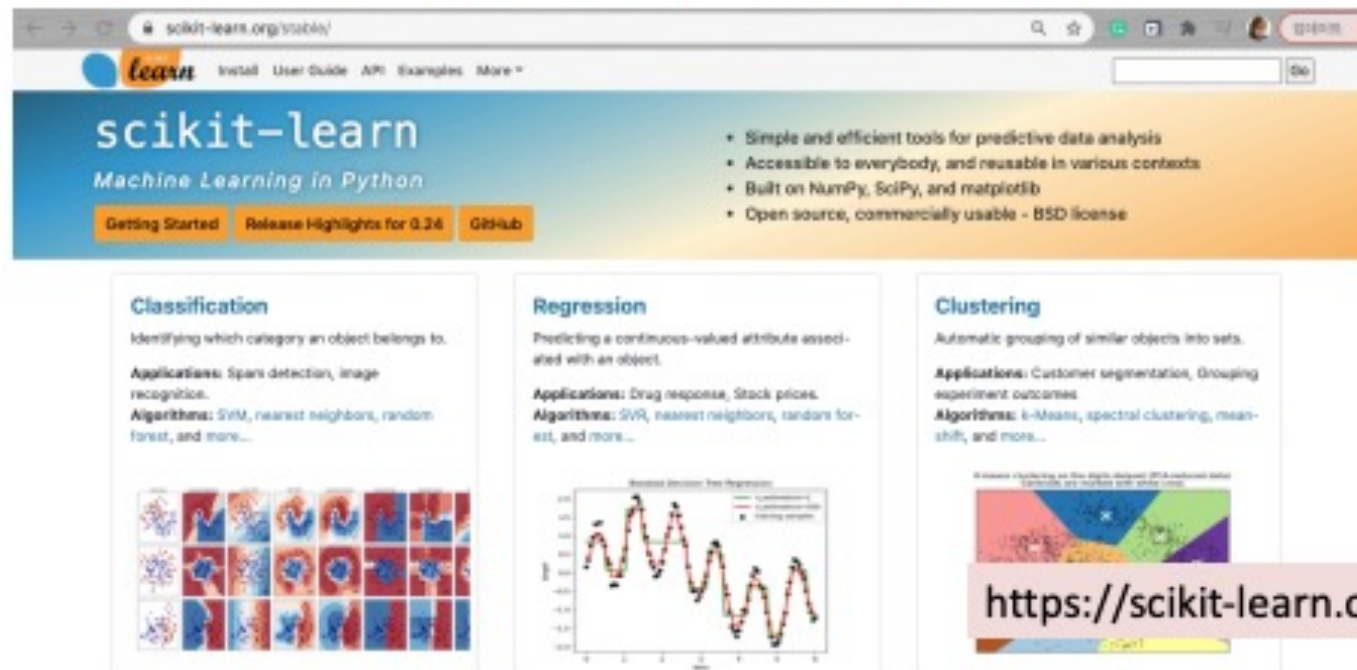
Matplotlib 라이브러리

- Matplotlib이란?
 - 맷플롯립(Matplotlib)은 데이터를 차트나 플롯으로 시각화하는 패키지임
 - 데이터 분석에서 Matplotlib은 데이터 분석 이전에 데이터 이해를 위한 시각화나, 데이터 분석 후에 결과를 시각화하기 위해서 사용됨
 - Matplotlib을 사용할 때 주로 서브패키지인 pyplot을 사용하며, pyplot은 MATLAB의 인터페이스와 유사하게 작동할 수 있도록 MATLAB을 사용하는 사용자층이 쉽게 matplotlib으로 옮겨오도록 돕고 있음
- Matplotlib 실습
 - <https://www.kaggle.com/yukyungchoi/2022-ml-matplotlib-cheatsheet>
- 다른 시각화 툴
 - Matplotlib으로 간단한 차트나 그래프를 그리는 것은 쉬운 일이나 예쁘게 다듬고 커스터마이징 하기에는 부적합함
 - 추천할 만한 시각화 툴
 - seaborn, plotly, plotnine

Scikit-Learn 라이브러리 (별도 영상)

■ 기계 학습을 위한 라이브러리 #1: Scikit-Learn

- 다양한 머신러닝 알고리즘을 구현한 파이썬 라이브러리
- 심플하고 일관성 있는 API, 유용한 온라인 문서, 풍부한 예제
- 머신러닝을 위한 쉽고 효율적인 개발 라이브러리 제공
- 다양한 머신러닝 관련 알고리즘과 개발을 위한 프레임워크와 API제공
- 많은 사람들이 사용하며 다양한 환경에서 검증된 라이브러리



The screenshot shows the Scikit-Learn website at <https://scikit-learn.org/stable/>. The header includes the Scikit-Learn logo, navigation links (Install, User Guide, API, Examples, More...), and a search bar. The main banner features the text "scikit-learn Machine Learning in Python" and a list of key features: Simple and efficient tools for predictive data analysis, Accessible to everybody, and reusable in various contexts, Built on NumPy, SciPy, and matplotlib, and Open source, commercially usable - BSD license. Below the banner, three main categories are highlighted: Classification (Identifying which category an object belongs to, Applications: Spam detection, image recognition, Algorithms: SVM, nearest neighbors, random forest, and more...), Regression (Predicting a continuous-valued attribute associated with an object, Applications: Drug response, Stock prices, Algorithms: SVR, nearest neighbors, random forest, and more...), and Clustering (Automatic grouping of similar objects into sets, Applications: Customer segmentation, Grouping experiment outcomes, Algorithms: k-Means, spectral clustering, mean-shift, and more...). Each category includes a representative image: a grid of handwritten digits for Classification, a line plot for Regression, and a scatter plot for Clustering.

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