What is scikit-learn?

scikit-learn is a Python module integrating classic machine learning algorithms in the tightly-knit scientific Python world (numpy, scipy, matplotlib). It aims to provide simple and efficient solutions to learning problems, accessible to everybody and reusable in various contexts: machine-learning as a versatile tool for science and engineering.

- Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python.
- Scikit-learn is licensed under a permissive simplified BSD license and is distributed under many Linux distributions, encouraging academic and commercial use.

- The library is built upon the **SciPy** (Scientific Python) that must be installed before you can use scikit-learn.
- This stack that includes:

  NumPy: Base n-dimensional array package

  SciPy: Fundamental library for scientific computing

  Matplotlib: Comprehensive 2D/3D plotting

  IPython: Enhanced interactive console

  Sympy: Symbolic mathematics

  Pandas: Data structures and analysis
- Extensions or modules for SciPy are conventionally named SciKits. As such, the module provides learning algorithms is named scikit-learn.
- The vision for the library is a level of robustness and support required for use in production systems. This means a deep focus on concerns such as easy of use, code quality, collaboration, documentation and and performance.
- Although the interface is Python, c-libraries are leverage for performance such as numpy for arrays and matrix operations, LAPACK, LibSVM and the careful use of cython.

Underlying Technologies

Numpy: the base data structure used for data and model parameters. Input data is presented as numpy arrays, thus integrating seamlessly with other scientific Python libraries.
Numpy's viewbased memory model limits copies, even when binding with compiled code (*Van der Walt et al., 2011*). It also provides basic arithmetic operations.

Scipy: efficient algorithms for linear algebra, sparse matrix representation, special functions and basic statistical functions. Scipy has bindings for many Fortran-based standard numerical packages, such as LAPACK.
This is important for ease of installation and portability, as providing libraries around Fortran code can prove challenging on various platforms.

Cython: a language for combining C in Python. Cython makes it easy to reach the performance of compiled languages with Python-like syntax and high-level operations. It is also used to bind compiled libraries, eliminating the boilerplate code of Python/C extensions.