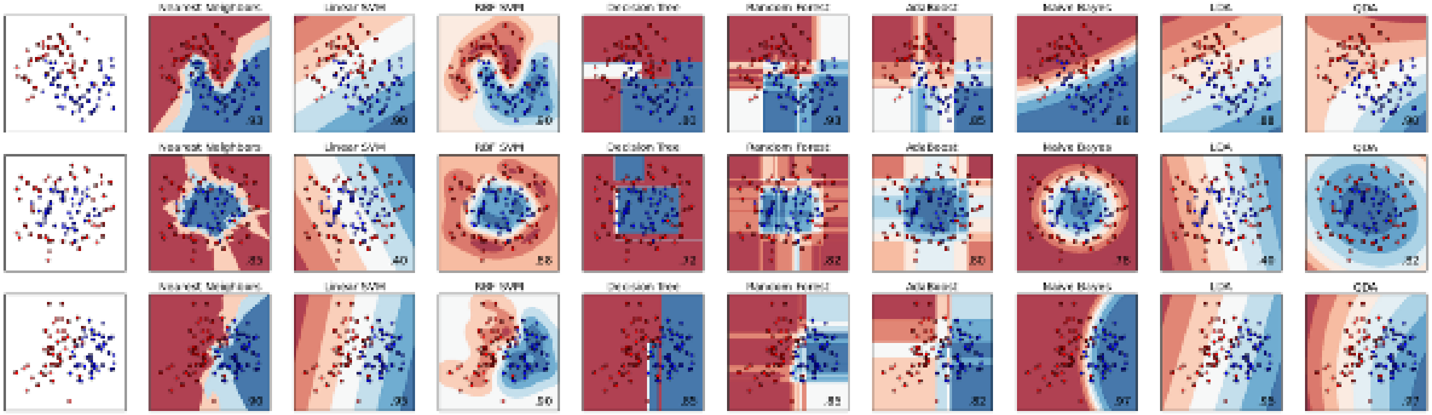
An Introduction to Scikit Learn: The Gold Standard of Python Machine Learning

George Seif Follow
Dec 26, 2018 · 3 min read ★



A comparison Scikit Learn's many Machine Learning models

Machine Learning Gold

If you're going to do Machine Learning in Python, Scikit Learn is the gold standard. Scikit-learn provides a wide selection of supervised and unsupervised learning algorithms. Best of all, it's by far the easiest and cleanest ML library.

Scikit learn was created with a software engineering mindset. It's core API design revolves around being easy to use, yet powerful, and still maintaining flexibility for research endeavours. This robustness makes it perfect for use in any end-to-end ML project, from the research phase right down to production deployments.

What Scikit Learn has to Offer

Scikit Learn is built on top of several common data and math Python libraries. Such a design makes it super easy to integrate between them all. You can pass numpy arrays and pandas data frames directly to the ML algoirthms of Scikit! It uses the following libraries:

- **NumPy**: For any work with matrices, especially math operations
- **SciPy**: Scientific and technical computing
- **Matplotlib**: Data visualisation
- **IPython**: Interactive console for Python
- **Sympy**: Symbolic mathematics
- **Pandas**: Data handling, manipulation, and analysis

Scikit Learn is focused on Machine Learning, e.g *data modelling*. It *is not* concerned with the loading, handling, manipulating, and visualising of data. Thus, it is natural and common practice to use the above libraries, especially NumPy, for those extra steps; they are made for each other!

Scikit's robust set of algorithm offerings includes:

- **Regression:** Fitting linear and non-linear models
- **Clustering:** Unsupervised classification
- **Decision Trees:** Tree induction and pruning for both classification and regression tasks
- **Neural Networks:** End-to-end training for both classification and regression. Layers can be easily defined in a tuple
- **SVMs:** for learning decision boundaries
- **Naive Bayes**: Direct probabilistic modelling

Even beyond that, it has some very convenient and advanced functions not commonly offered by other libraries:

- **Ensemble Methods:** Boosting, Bagging, Random Forest, Model voting and averaging

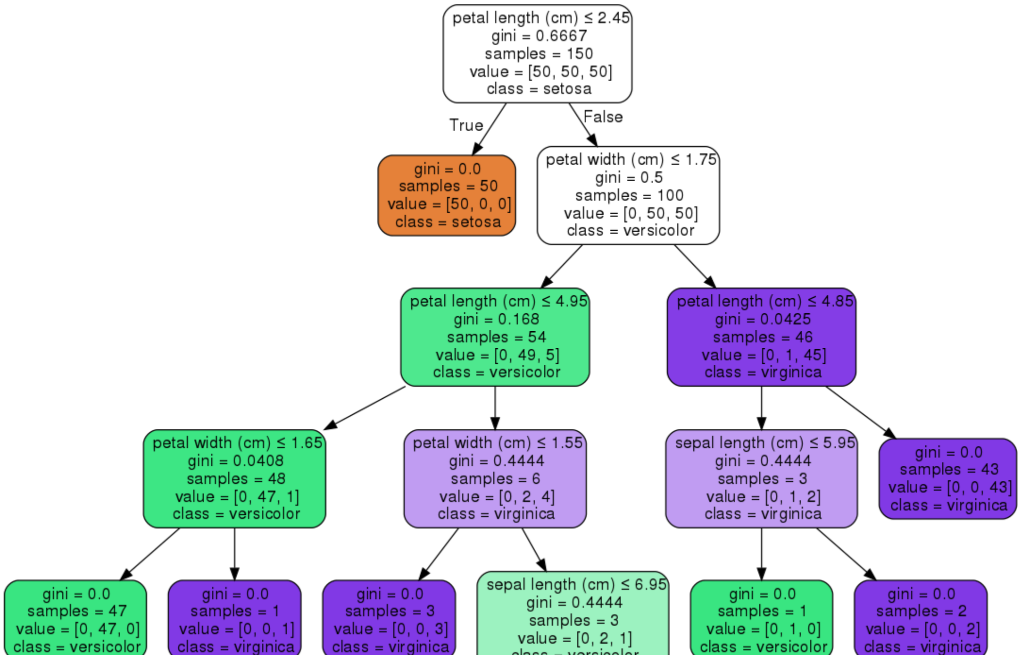
- **Feature Manipulation**: Dimensionality reduction, feature selection, feature analysis
- **Outlier Detection:** For detecting outliers and rejecting noise
- **Model selection and validation:** Cross-validation, Hyperparamter tuning, and metrics

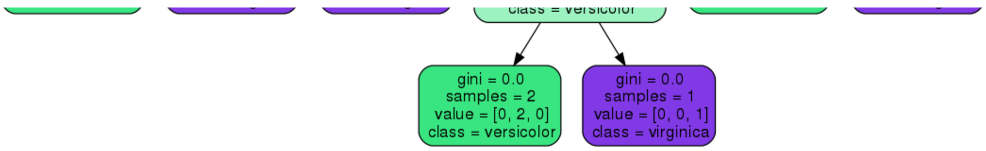
A Taste Test

To give you a taste of just how easy it is to train and test an ML model using Scikit Learn, here's an example of how to do just that for a Decision Tree Classifier!

Decision trees for both classification and regression are super easy to use in Scikit Learn with a built in class. We'll first load in our dataset which actually comes built into the library. Then we'll initialise our decision tree for classification, also a built in class. Running training is then a simple one-liner! The `.fit(X, Y)` function trains the model where *X* is the numpy array of inputs and *Y* is the corresponding numpy array of outputs

Scikit Learn also allows us to visualise our tree using the graphviz library. It comes with a few options that will help in visualising the decision nodes and splits that the model learned which is super useful for understanding how it all works. Below we will colour the nodes based on the feature names and display the class and feature information of each node.





Beyond that, Scikit Learn's documentation is exquisite! Each of the algorithm parameters are explained clearly and are intuitively named. Moreover, they also offer tutorials with example code on how to train and apply the model, its pros and cons, and practical application tips!

. . .

Like to learn?

Follow me on twitter where I post all about the latest and greatest AI, Technology, and Science! Connect with me on LinkedIn too!

Recommended Reading

Want to learn more about Machine Learning? The ***Hands-On Machine Learning*** book is the best resource out there for learning how to do *real* Machine Learning with Python!

And just a heads up, I support this blog with Amazon affiliate links to great books, because sharing great books helps everyone! As an Amazon Associate I earn from qualifying purchases.

Machine Learning Data Science Artificial Intelligence Technology Innovation