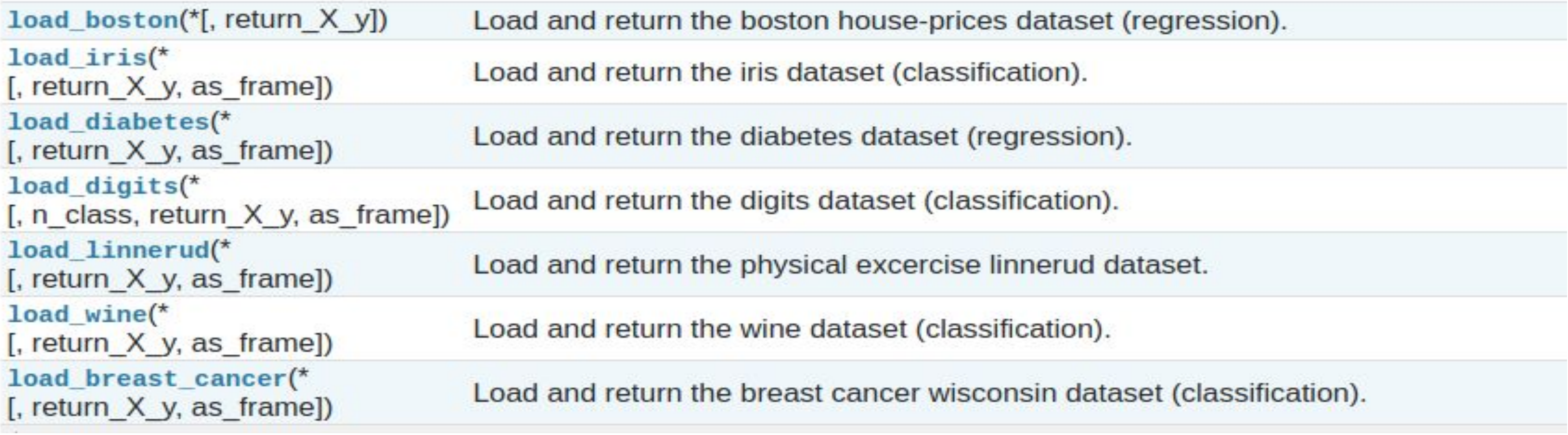
How to use Scikit-Learn Datasets for Machine Learning

scikit
learn

- Scikit-Learn provides clean datasets for you to use when building ML models.
- I mean the type of clean that’s ready to be used to train a ML model.
- The datasets come with the Scikit-learn package itself. Your don’t need to download anything.
- Scikit-Learn provides seven datasets, which they call toy datasets. Don’t be fooled by the word “toy”.
- These datasets are powerful and serve as a strong starting point for learning ML.



load_boston(*[, return_X_y])	Load and return the boston house-prices dataset (regression).
load_iris(* [, return_X_y, as_frame])	Load and return the iris dataset (classification).
load_diabetes(* [, return_X_y, as_frame])	Load and return the diabetes dataset (regression).
load_digits(* [, n_class, return_X_y, as_frame])	Load and return the digits dataset (classification).
load_linnerud(* [, return_X_y, as_frame])	Load and return the physical excercise linnerud dataset.
load_wine(* [, return_X_y, as_frame])	Load and return the wine dataset (classification).
load_breast_cancer(* [, return_X_y, as_frame])	Load and return the breast cancer wisconsin dataset (classification).

Breast Cancer Dataset

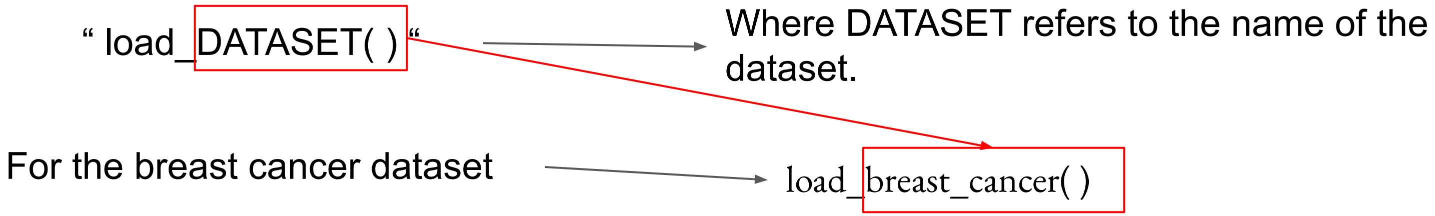
- We will working with the “Breast Cancer Wisconsin” Datasets.
- We will import the data and understand how to read it.
- We’ll build a simple ML model that is able to classify cancer scans either as malignant or benign.

How do I Import the Datasets

- The datasets can be found in sklearn.datasets Lets import the data . We first import datasets which holds all the seven datasets.

```
from sklearn import datasets
```

- Each dataset has a corresponding function used to load the dataset.
- These function follow the same format:



“ load_DATASET( ) “

Where DATASET refers to the name of the dataset.

For the breast cancer dataset

load_breast_cancer( )

- Similarly, for the wine dataset we would use load_wine( ).

- Let’s load the dataset and store it into a variable called data.

```
Data = datasets.load_breast_cancer( )
```

- These load functions don’t return data in the tabular format we may expect.
- They return a Bunch object. Don’t know what a Bunch is ? No worries.

- Think of a Bunch object as Scikit-learn’s fancy name for a dictionary

What’s in our Dictionary (Bunch) ?

- Scikit’s dictionary or Bunch is really powerful.
- Let’s begin this dictionary by looking at its keys.

```
Print ( data.keys( ) )
```

data :- data is all the feature data (the attributes of the scan that help us identify if the tumor is malignant or benign, such as radius, area,etc.) in a NumPy array.

target :- target is the target data (the variable you want to predict, in this case whether the tumor is malignant or benign) in a NumPy array.

- ➢ These two keys are the actual data. The remaining keys (below), server a descriptive purpose , It’s important to note that all of Scikit-Learn datasets are divided into data and target.
- ➢ Data represents the features, which are the variables that help the model learn how to predict.
- ➢ Target includes the actual labels.
- ➢ In our case, the target data is one column classifies the tumor as either o indicating malignant or 1 for benign.

- **feature_name** are the names of the feature variables, in other words names of the target columns(s)
- **target_names** is the name(s) of the target variable(s), in other words name(s) of the target column(s)
- DESCR, short for DESCRIPTION , is a description of the dataset
- **filename** is the path to the actual file of the data in csv format.

```
Print ( data.DESCR )
```