**SCIKIT-LEARN**

**Features**

Supervised learning:-

* Linear regression
* Support vector machine (SVM)
* Decision tree etc.

Unsupervised learning:-

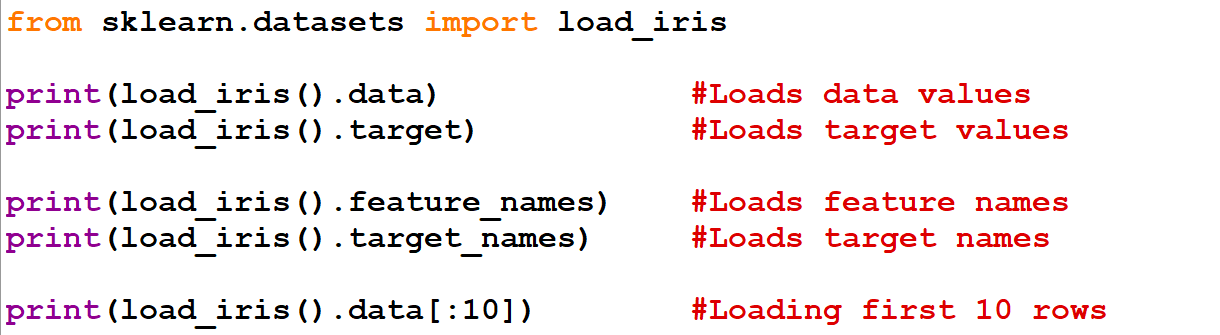
* Clustering
* Factor analysis
* Principal component analysis
* Neural networks etc.

**Basic Terminologies**

* **Cross validation:** Method used to check **accuracy** of **supervised** **models**.
* **Dimensionality reduction:** Reducing data to **lower** **dimensions** for ease of visualization.
* **Ensemble methods:** Combines the predictions of multiple supervised models.
* **Feature extraction:** Attribute **conversion** from text to image or vice-versa.
* **Feature selection:** Picking up useful attributes to feed our model.

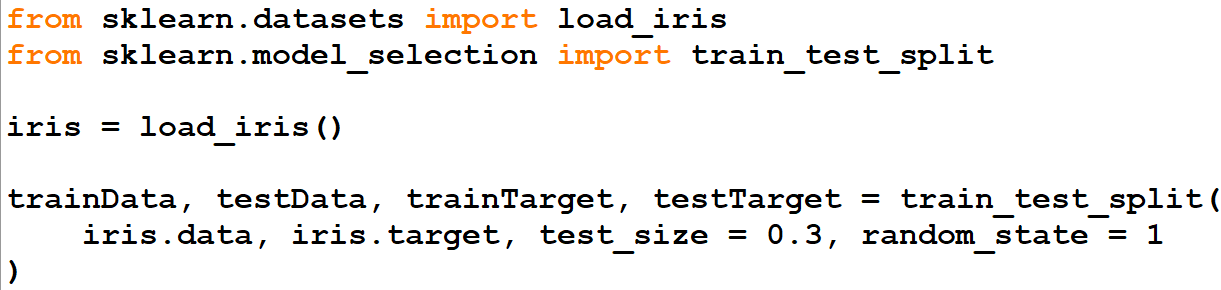
**Dataset Loading**

* **Features:** Variables of a given data also known as **predictors** or **input**.
* **Feature matrix:** Matrix showing **collection** of all features.
* **Response:** Output variable also known as **target** or **label**.
* **Response vector:** Refers to the **response column** in our matrix.



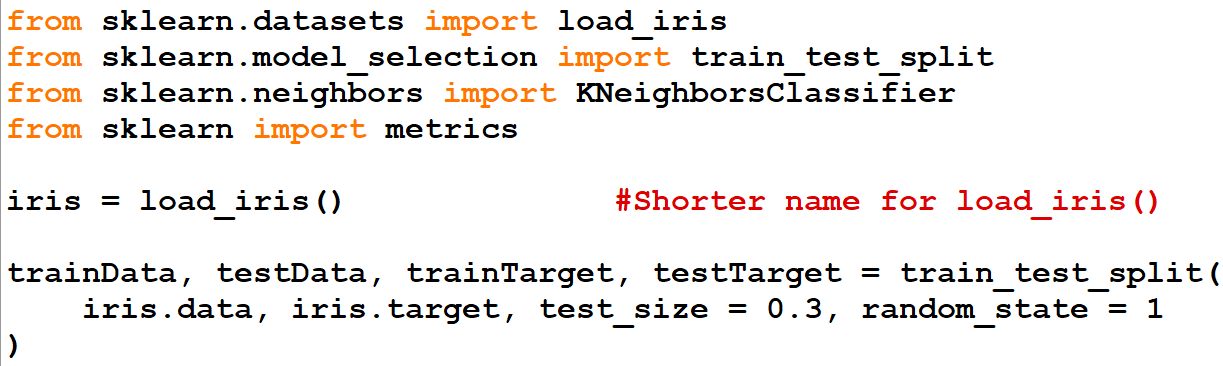
* Data above are represented in form of **matrices**.

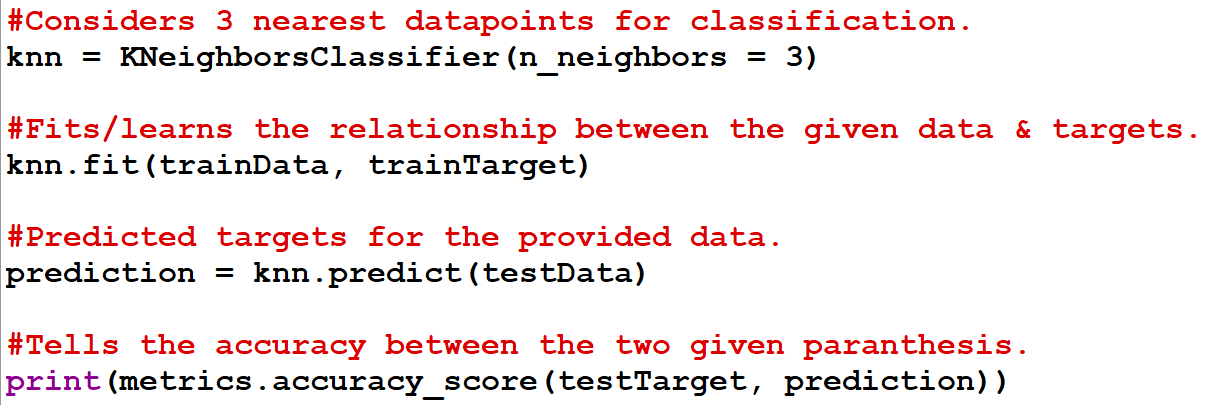
**Splitting the Dataset**



* This way, we kill two birds with same stone.
* We are splitting the data into **training** & **testing** part.
* Also, we are assigning values to **4** variables for **training data**, **testing data**, **training target** & **testing target**.
* **1st argument** are **data** from datasets.
* **2nd argument** are **targets** from datasets.
* **3rd argument** is **0.3**, meaning **30%** of data will be used for **testing**.
* **4th argument** is the **random seed** & with an integer value it ensures that same result is produced when compiled each time.

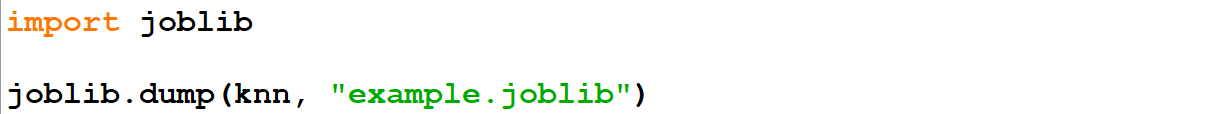
**KNN Example Code**





**Saving a Model**

Saving:-



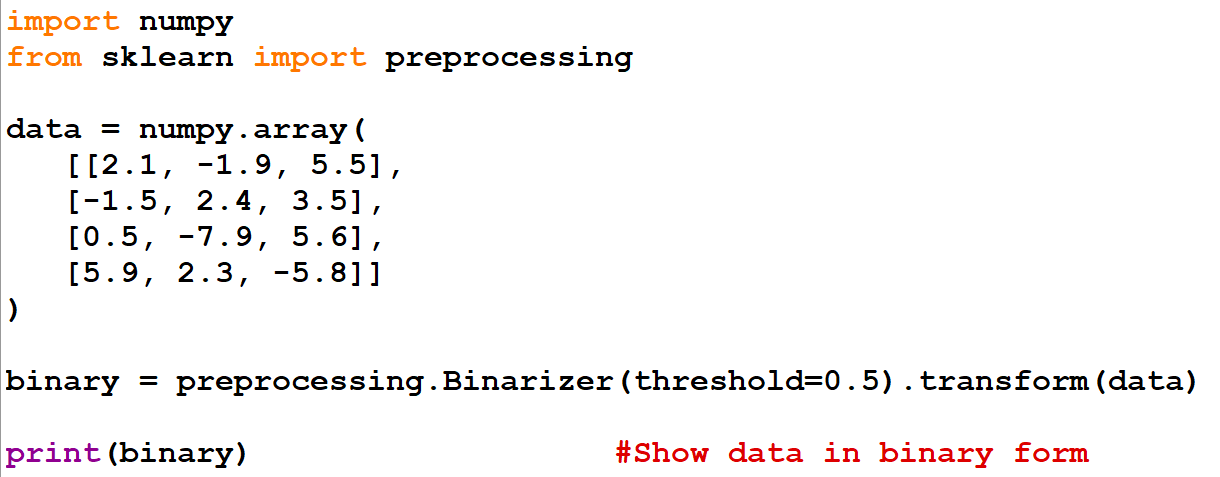
Loading:-



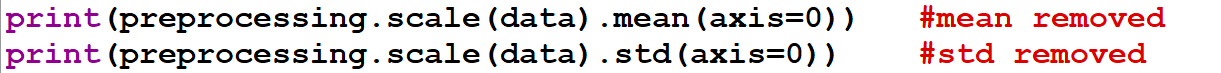
**Preprocessing**

* Cleansing of **raw data** into **meaningful** form.
* Preprocessing techniques include **binarization**, **mean removal** & **normalization** etc.
* **Binarization:** Values above a certain **threshold** value will be set to **1**.
* **Mean removal:** **Subtracting** **mean** of data from each value, to **centre** them around **0**.
* **Normalization: Modifying** feature vectors in such a certain way.

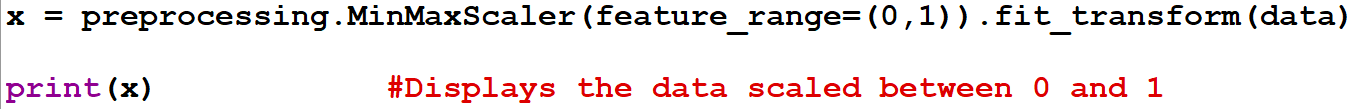
**Binarization**



**Mean Removal**

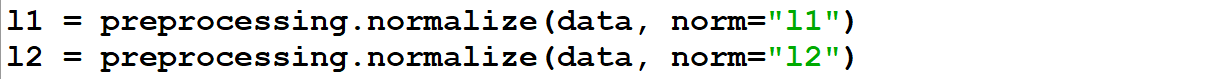


**Scaling**



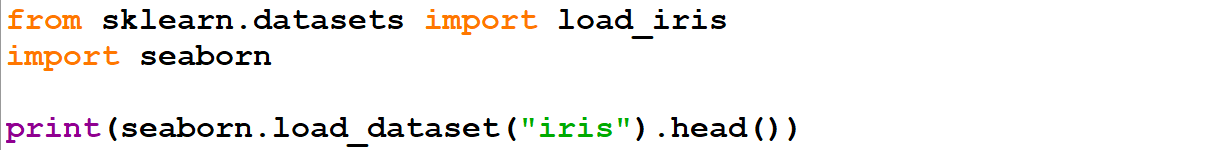
**Normalization**

* **L1 normalization:** Normalizing the matrix in such a way that the **sum of values** in each row goes up to **1**.
* **L2 normalization:** Normalizing the matrix in such a way that the **sum of squares of values** in each row goes up to **1**.



**Data Representation**

Representation in form of table:-

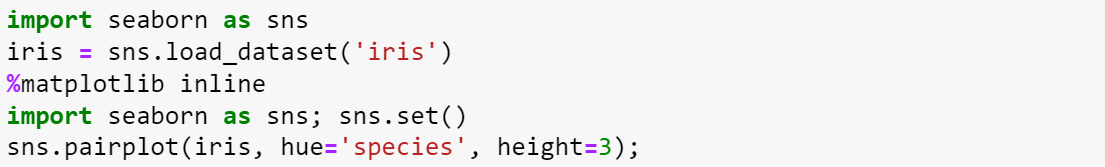


* **Rows** are referred as **sample** & **columns** as **features** here.

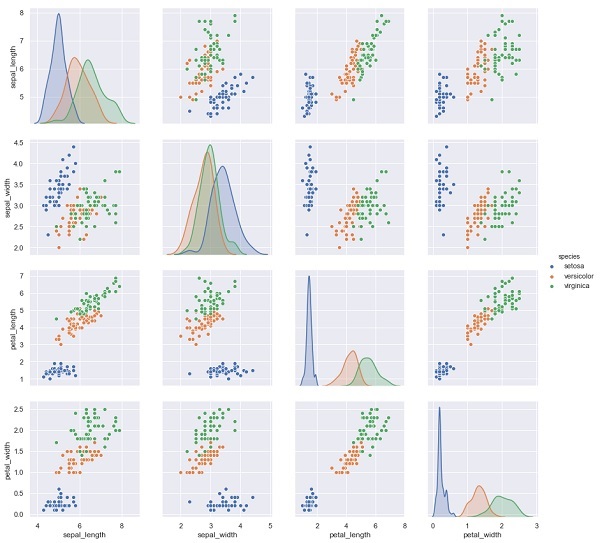
Representation in form of **matrix**:-



More attractive representation:-



***\*Specific to "Jupyter Notebook" only\****



**Estimator API**

* An API used by **Scikit-learn** & by all **ML algorithms** in **Scikit-learn**.
* It is that **object** that learns by fitting the data.

Uses of estimator API:-

* **Estimation** & **decoding** of a model.
* **Mapping** **non-rectangular** data representation into **rectangular** data (array).
* **Distinction** between different samples.

Linear Regression Model

* Hyperparameter: Parameter to be set before fitting the model to data.