PENDAHULUAN

Machine Learning

Data Analysis Process

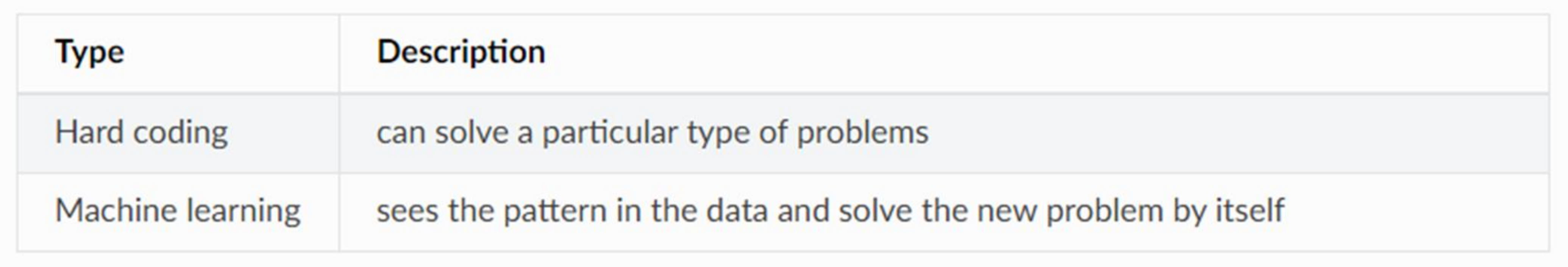
Machine learning algorithms is a part of data analysis process. The data analysis process involves following steps :

- Collecting the data from various sources
- Cleaning and rearranging the data e.g. filling the missing values from the dataset etc.
- Exploring the data e.g. checking the statistical values of the data and visualizing the data using plots etc.
- Modeling the data using correct machine learning algorithms.
- Lastly, check the performance of the newly created model.

- Data analysis requires the knowledge of multiple field e.g. data cleaning using Python or R language.
- Good knowledge of mathematics for measuring the statistical parameter of the data.
- Also, we need to have the knowledge of some specific field on which we want to apply the machine learning algorithm.
- Lastly, we must have the understanding of the machine learning algorithms.

Machine Learning

- In general programming methods, we write the codes to solve the problem; and the code can solve a particular types of problem only. This is known as ‘hard coding’ method
- Machine learning can be defined as the process of extracting knowledge from the data, such that an accurate predication can be made on the future data. In the other words, machine learning algorithms are able to predict the outcomes of the new data based on their training



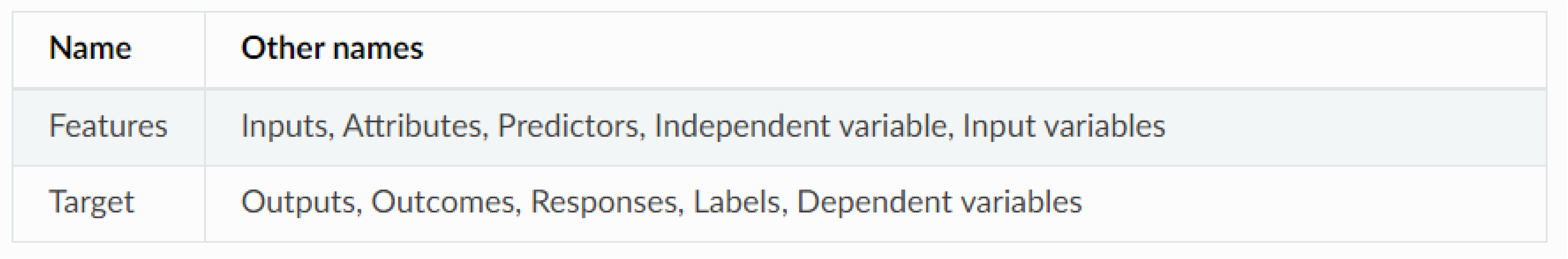
Type	Description
Hard coding	can solve a particular type of problems
Machine learning	sees the pattern in the data and solve the new problem by itself

Data: samples and features

- Samples: Each data has certain number of samples.
- Features: Each sample has some features, e.g if we have samples of lines, then features of this lines can be ‘x’ and ‘y’ coordinates.
- All the features should be identical. For example, all the lines should have only two features i.e. ‘x’ and ‘y’ coordinates. If some lines have third feature as ‘thickness of line’, then we need to append/delete this feature to all the lines.

Target

- Target: There **may be** the certain numbers of possible outputs for the data, which is known as ‘target’. For example, the the points can be on the ‘straight line’ or on the ‘curve line’. Therefore, the possible targets for this case are ‘line’ and ‘curve’.



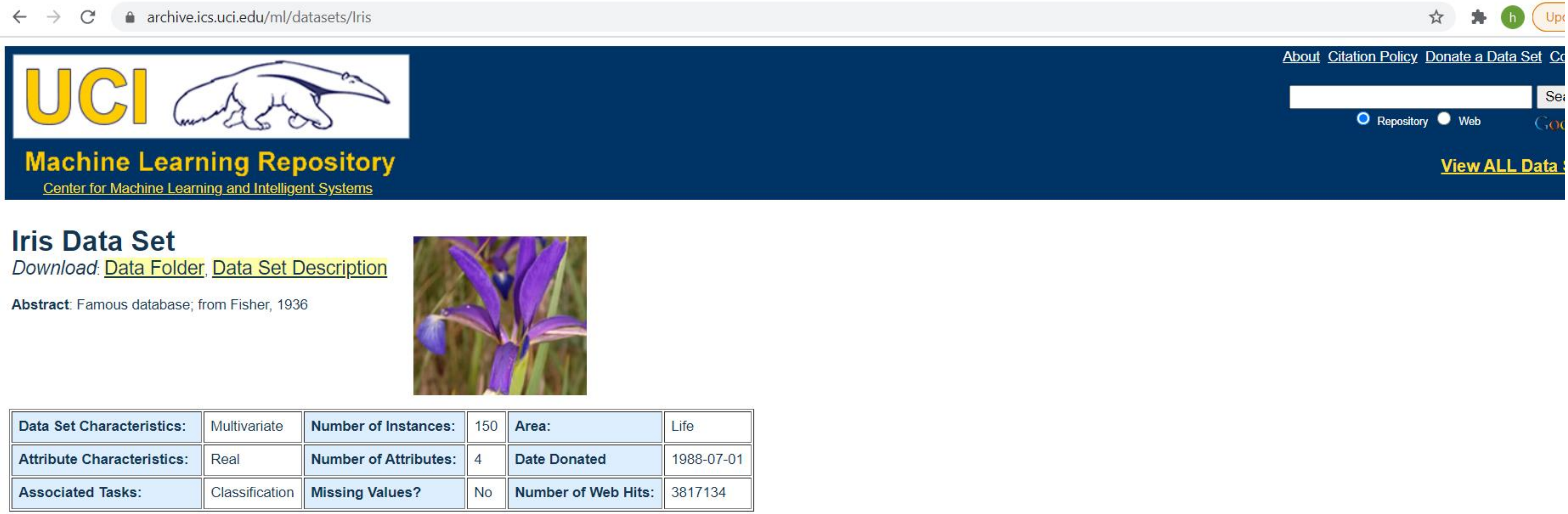
Name	Other names
Features	Inputs, Attributes, Predictors, Independent variable, Input variables
Target	Outputs, Outcomes, Responses, Labels, Dependent variables

Dataset

- Following are the important points about the dataset, which we discussed in this section,
- Datasets have samples of data, which includes some features of the data.
- All the features should be available in every data. If there are missing/extra features in some data, the we need to add/remove those features from the data.
- Also, the dataset may contain the ‘target’ values in it.

Example dataset : UCI dataset, Kaggle

Example dataset Iris



Machine Learning Repository

Center for Machine Learning and Intelligent Systems

Iris Data Set

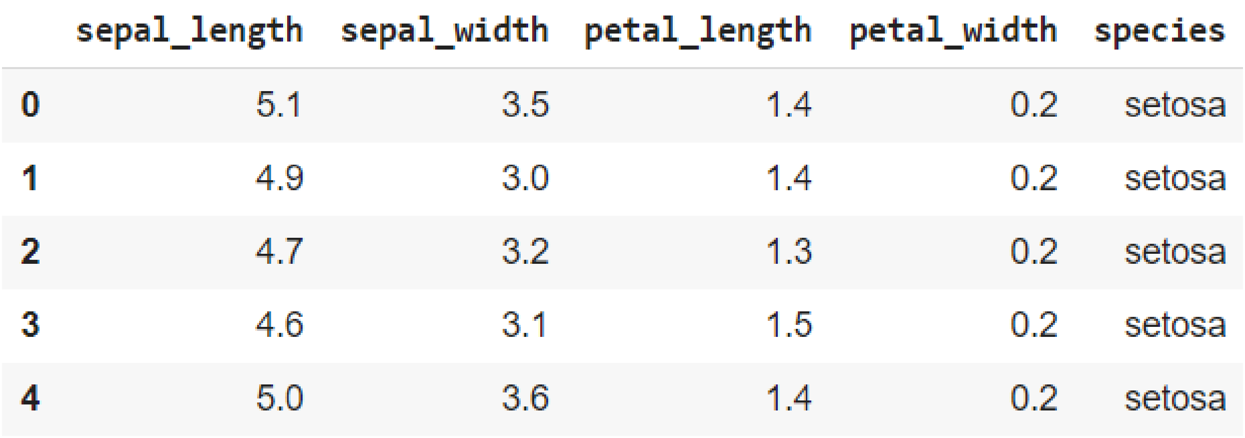
Download: Data Folder, Data Set Description

Abstract: Famous database; from Fisher, 1936

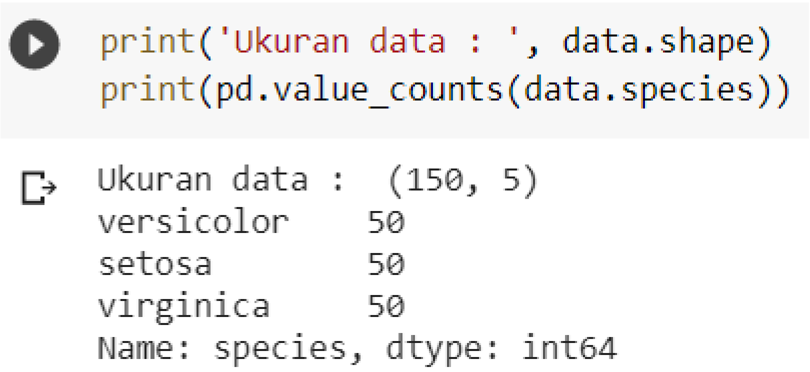
Data Set Characteristics:	Multivariate	**Number of Instances:**	150	**Area:**	Life
Attribute Characteristics:	Real	**Number of Attributes:**	4	**Date Donated**	1988-07-01
Associated Tasks:	Classification	**Missing Values?**	No	**Number of Web Hits:**	3817134

https://archive.ics.uci.edu/ml/index.php

Example data iris

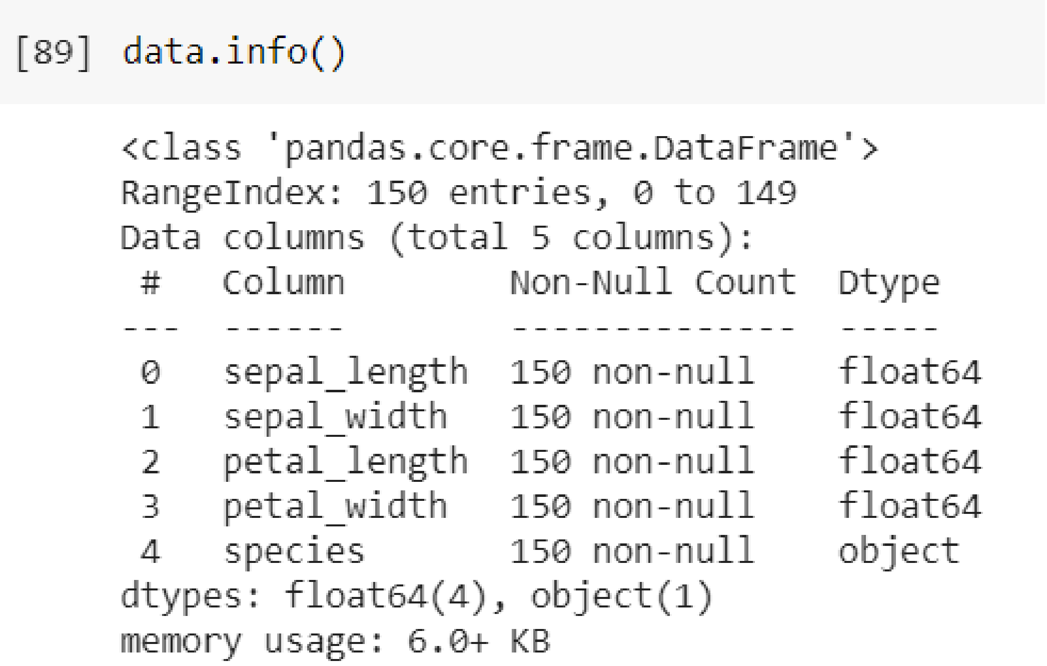


	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa



```
print('Ukuran data : ', data.shape)
print(pd.value_counts(data.species))
```

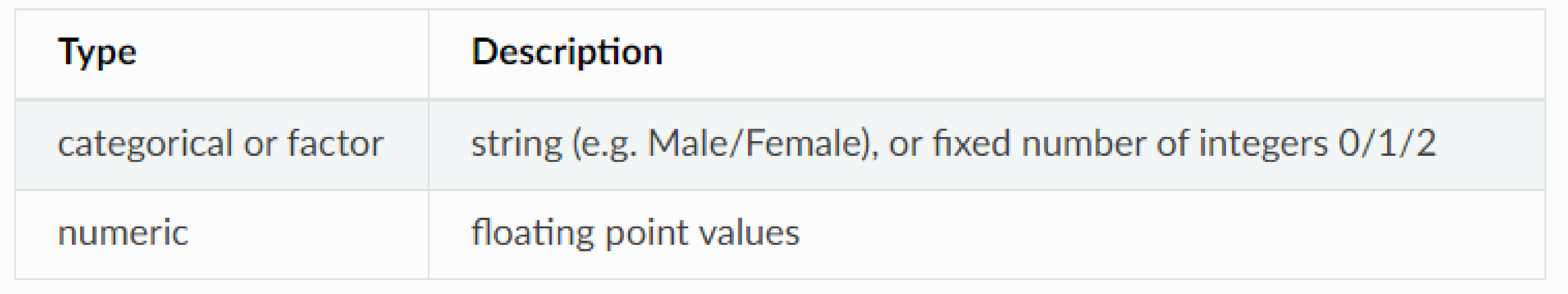
```
Ukuran data :  (150, 5)
versicolor    50
setosa        50
virginica     50
Name: species, dtype: int64
```



```
[89] data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150 entries, 0 to 149
Data columns (total 5 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   sepal_length  150 non-null    float64
 1   sepal_width   150 non-null    float64
 2   petal_length  150 non-null    float64
 3   petal_width   150 non-null    float64
 4   species       150 non-null    object 
dtypes: float64(4), object(1)
memory usage: 6.0+ KB
```

Type of variable



Type	Description
categorical or factor	string (e.g. Male/Female), or fixed number of integers 0/1/2
numeric	floating point values

Types of machine learning

1. Supervised learning

In Supervised Learning, we have a dataset which contains both the input ‘features’ and output ‘target’,

Example where Iris flower dataset has both ‘features’ and ‘target

2. Unsupervised learning

Unsupervised Learning, the dataset contains only ‘features’ and ‘no target’. Here, we need to find the relationship between the various types of data. In the other words, we have to find the labels from the given dataset.

Supervised learning

a. Classification: In classification the targets are discrete i.e. there are fixed number of values of the outputs

Example IRIS there are only three types of flower.

Also, these outputs are represented using strings e.g. (Male/Female) or with fixed number of integers as shown for ‘iris’ dataset in Section 1.3.3 where 0, 1 and 2 are used for three types of flower.

- If the target has only two possible values, then it is known as ‘binary classification’.
- If the target has more than two possible values, then it is known as ‘multiclass classification’.

b. Regression: In regression the targets are continuous e.g. we want the calculate the ‘age of the animal (i.e. target)’ with the help of the ‘fossil dataset (i.e. feature)’. In this case, the problem regression problem as the age is a continuous quantity as it does not have fixed number of values.

Unsupervised learning

Unsupervised learning can be divided into three categories i.e. Clustering, Dimensionality reduction and Anomaly detection

a. Clustering: It is process of reducing the observations. This is acheived by collecting the simialar data in one class.
b. Dimensionality reduction: This is the reduction of higher dimensional data to 2 dimensional or 3 dimensional data, as it is easy to visualize the data in 2 dimensional and 3 dimensional form.
c. Anomaly detection: This is the process of removal of undesired data from the dataset.

Combination Supervised and Unsupervised

- Sometimes these two methods, i.e. supervised and unsupervised learning, are combined. For example the unsupervised learning can be used to find useful features and targets; and then these features can be used by the supervised training method.
- For example, we have a the ‘titanic’ dataset, where we have all the information about the passengers e.g. age, gender, traveling-class and number of people died during accident etc. Here, we need to find the relationship between various types of data e.g. people who are traveling in higher-class must have higher chances of survival etc.

Please note the following points

- Not all the problems can be solved using Machine learning algorithms.
- If a problem can be solved directly, then do not use machine learning algorithms.
- Each machine learning algorithms has it’s own advantages and disadvantages. In the other words, we need to choose the correct machine learning algorithms to solve the problem.
- We need not to be expert in the mathematics behind the machine learning algorithms; but we should be aware of pros and cons of the algorithms.

TUGAS

1. Buka dataset di UCI machine learning https://archive.ics.uci.edu/ml/index.php dan download data.csv atau txt
2. Simpan data di folder tertentu.
3. Identifikasi dataset.

   Data : typical, feature, target, number of sample

Reference

- https://mclguide.readthedocs.io/en/latest/sklearn/sklearn.html