Data Science using Python

Dr. Kamini Solanki, Associate Professor
Parul Institute of Computer Application - BCA

CHAPTER - 6

Machine Learning with Scikit–Learn

Introduction to Machine Learning

- Machine Learning is a discipline that deals with the study of methods for pattern recognition in data sets undergoing data analysis. In particular, it deals with the development of algorithms that learn from data and make predictions.
- Each methodology is based on building a specific model.
- There are very many methods that belong to the learning machine, each with its unique characteristics, which are specific to the nature of the data and the predictive model that you want to build.
- The choice of which method is to be applied is called **learning problem**.

Why Machine Learning?

- If we stored the data generated in a day on Blu-ray disks and stacked them up, it would be equal to the height of four Eiffel towers.
- Machine learning helps analyze this data easily and quickly.

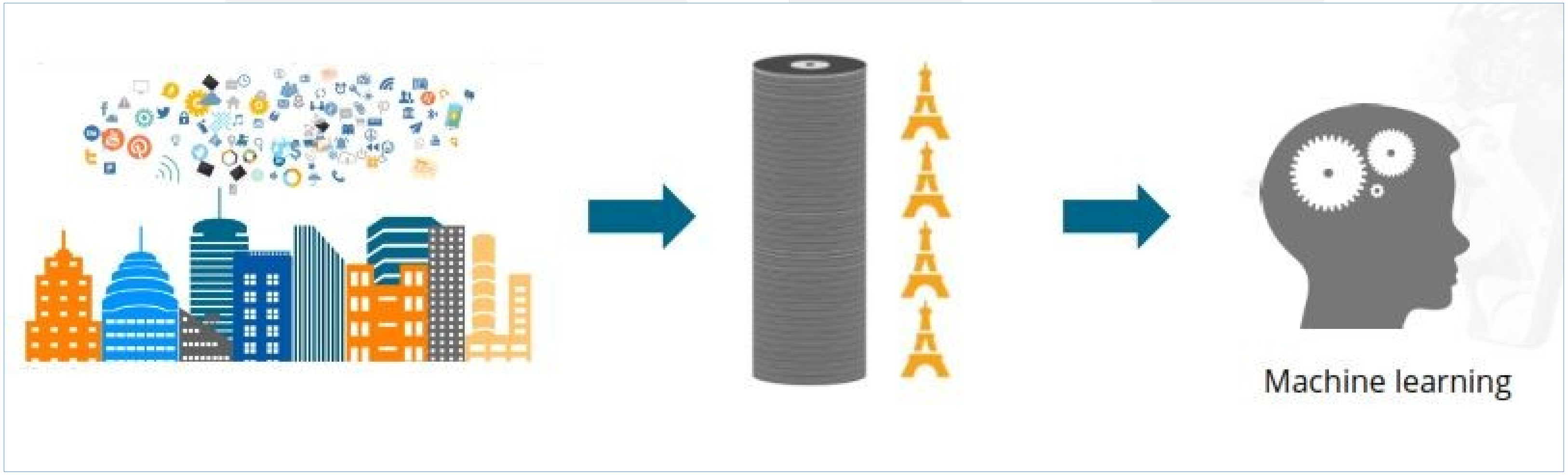


Image source : Simplilearn

Purpose of Machine Learning

- Machine learning is a great tool to analyze data, find hidden data patterns and relationships, and extract information to enable information-driven decisions and provide insights.

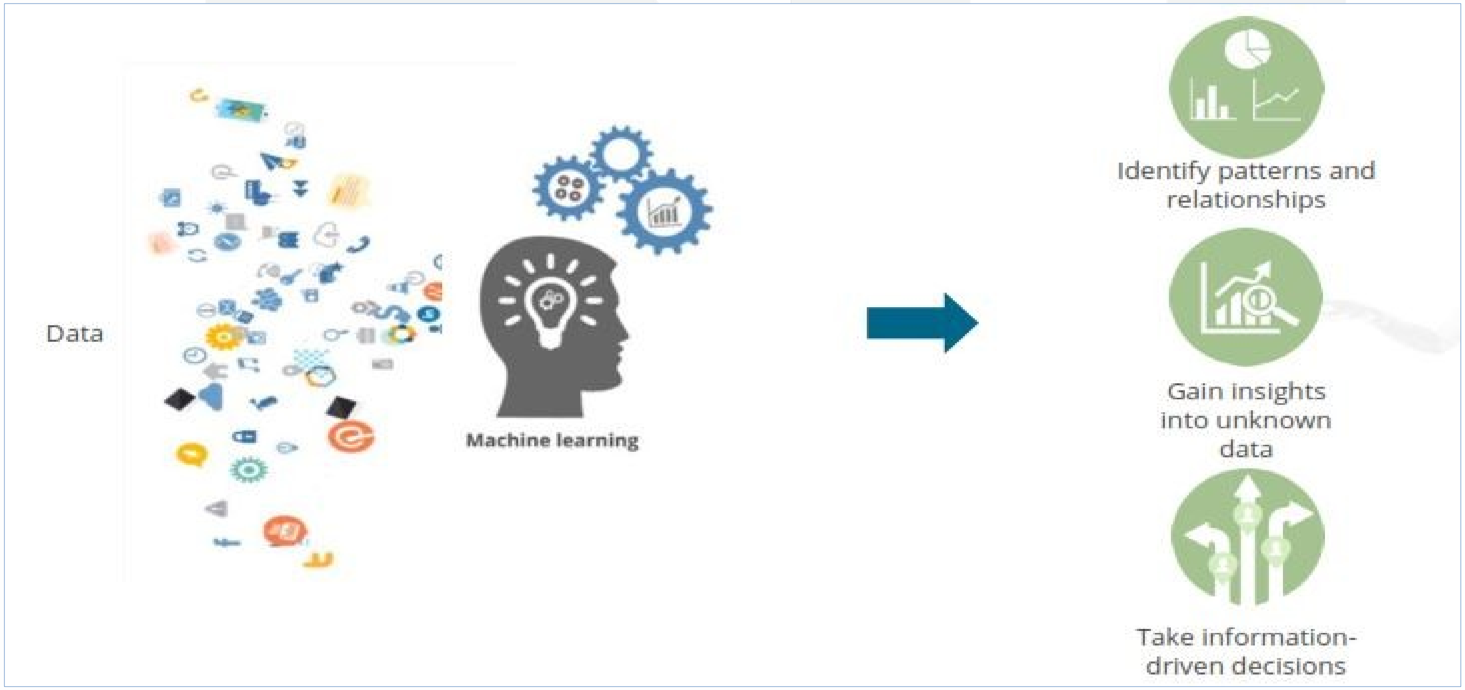

Data
Machine learning
Identify patterns and relationships
Gain insights into unknown data
Take information-driven decisions

Image source : Simplilearn

Machine Learning Terminology

- These are some machine learning terminologies that you will come across in this lesson:

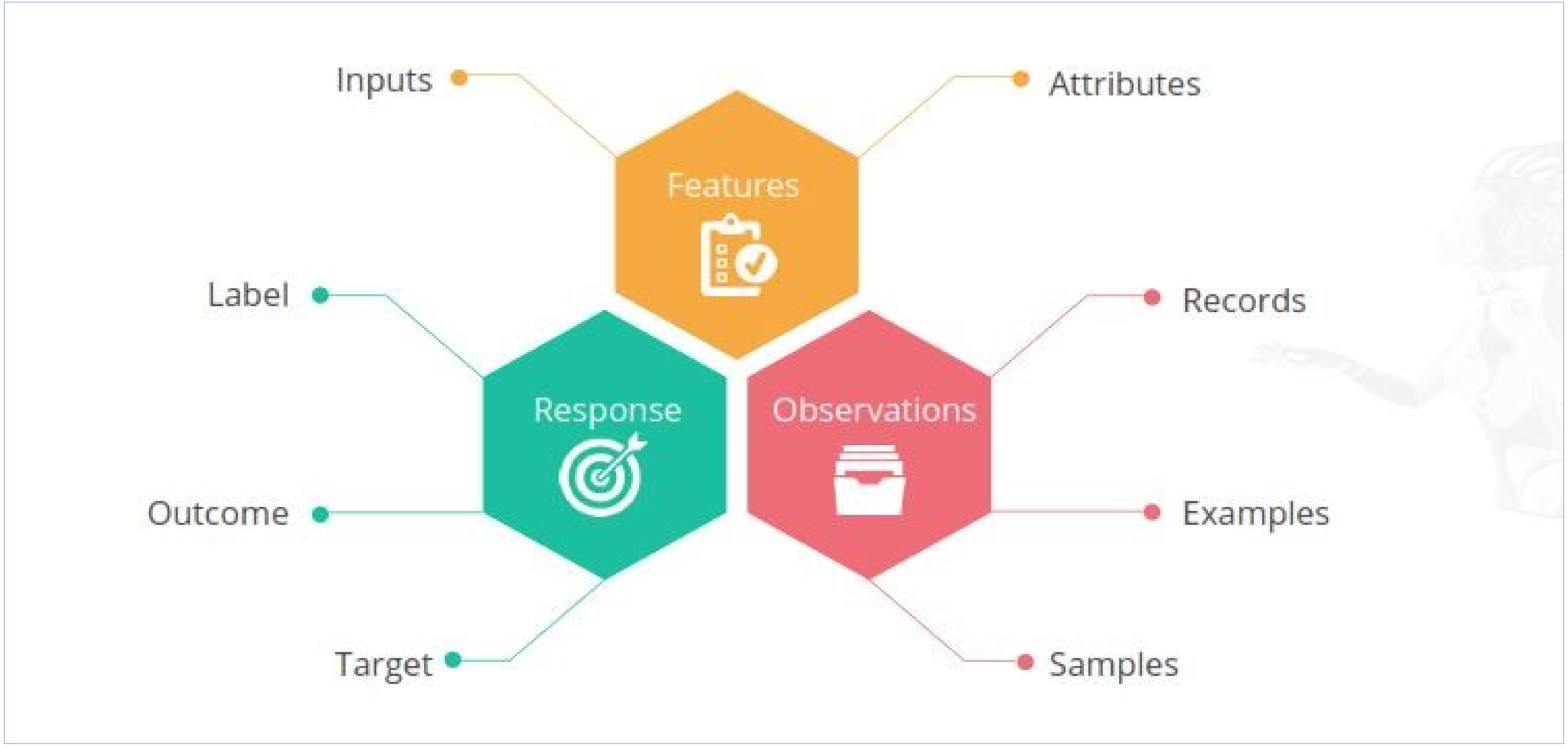


Image source : Simplilearn

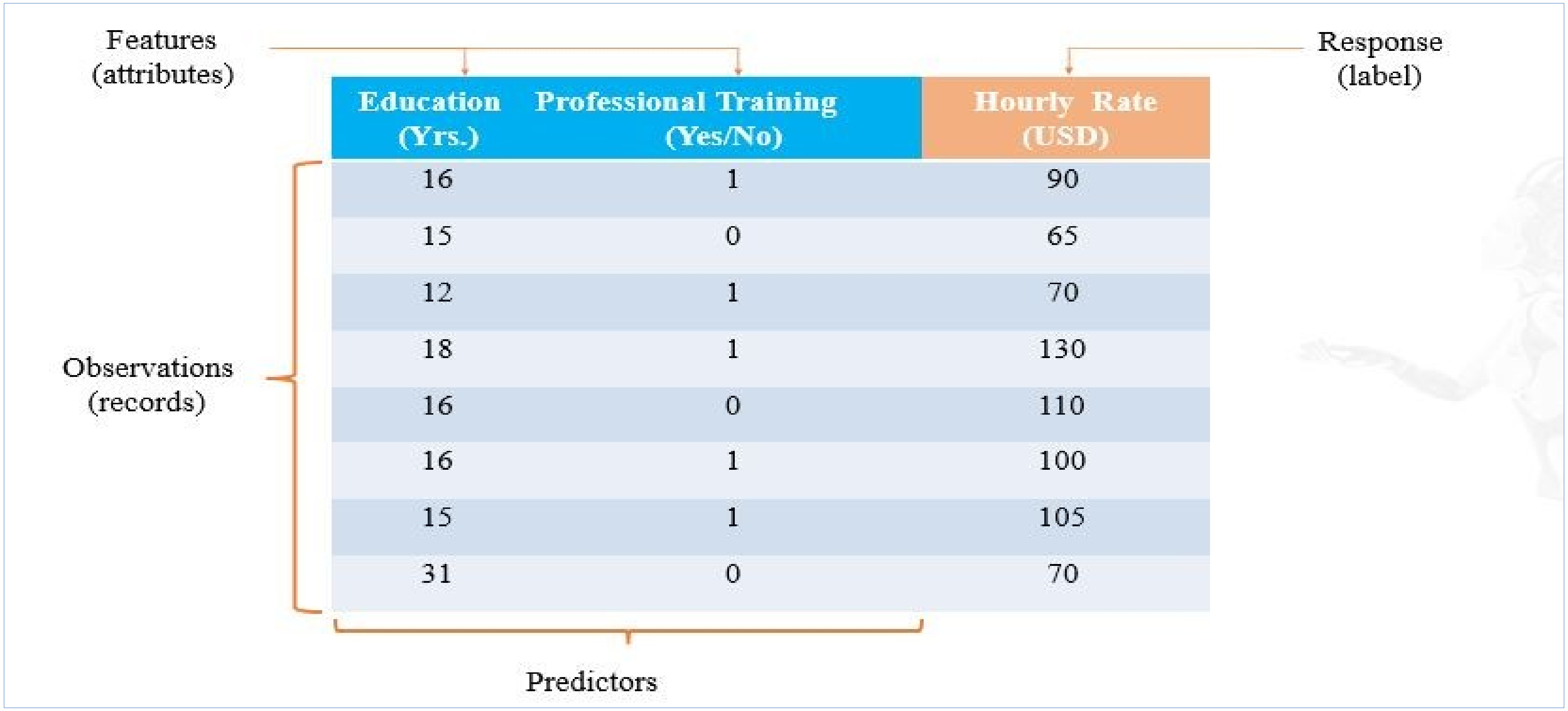
Machine Learning Approach

- The machine learning approach starts with either a problem that you need to solve or a given dataset that you need to analyze.

1. Understand the problem/dataset
2. Extract the features from the dataset
3. Identify the problem type
4. Choose the right model
5. Train and test the model
6. Strive for accuracy

Steps 1 and 2: Understand the Dataset and Extract Its Features

- Let us look at a dataset and understand its features in terms of machine learning.



Features (attributes)

Response (label)

Education (Yrs.)	Professional Training (Yes/No)	Hourly Rate (USD)
16	1	90
15	0	65
12	1	70
18	1	130
16	0	110
16	1	100
15	1	105
31	0	70

Observations (records)

Predictors

Steps 3 and 4: Identify the Problem Type and Learning Model

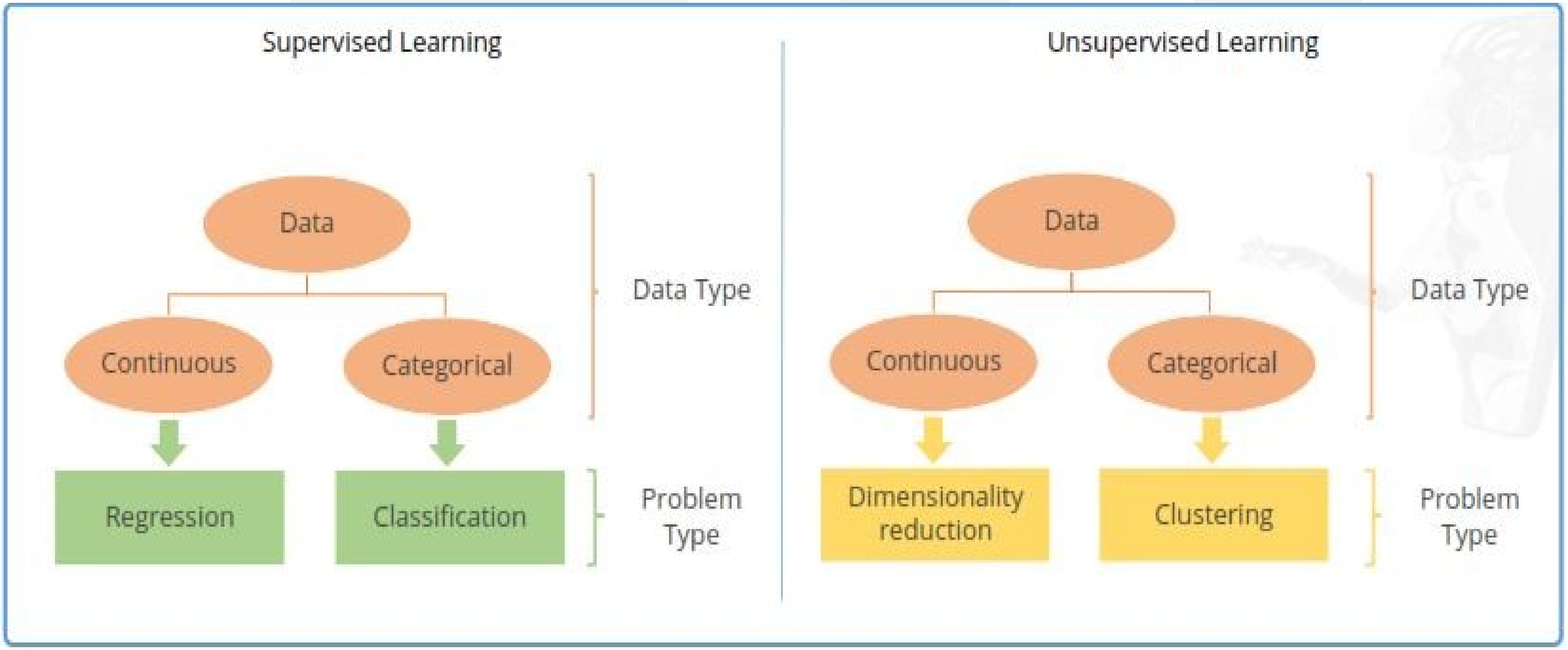
- Machine learning can either be supervised or unsupervised. The problem type should be selected based on the type of learning model.
- Concept
- Problem Types
- Example

Concept

Supervised Learning	Unsupervised Learning
In supervised learning, the dataset used to train a model should have observations, features, and responses. The model is trained to predict the “right” response for a given set of data points.	In unsupervised learning, the response or the outcome of the data is not known.
Supervised learning models are used to predict an outcome.	Unsupervised learning models are used to identify and visualize patterns in data by grouping similar types of data.
The goal of this model is to “generalize” a dataset so that the “general rule” can be applied to new data as well.	The goal of this model is to “represent” data in a way that meaningful information can be extracted.

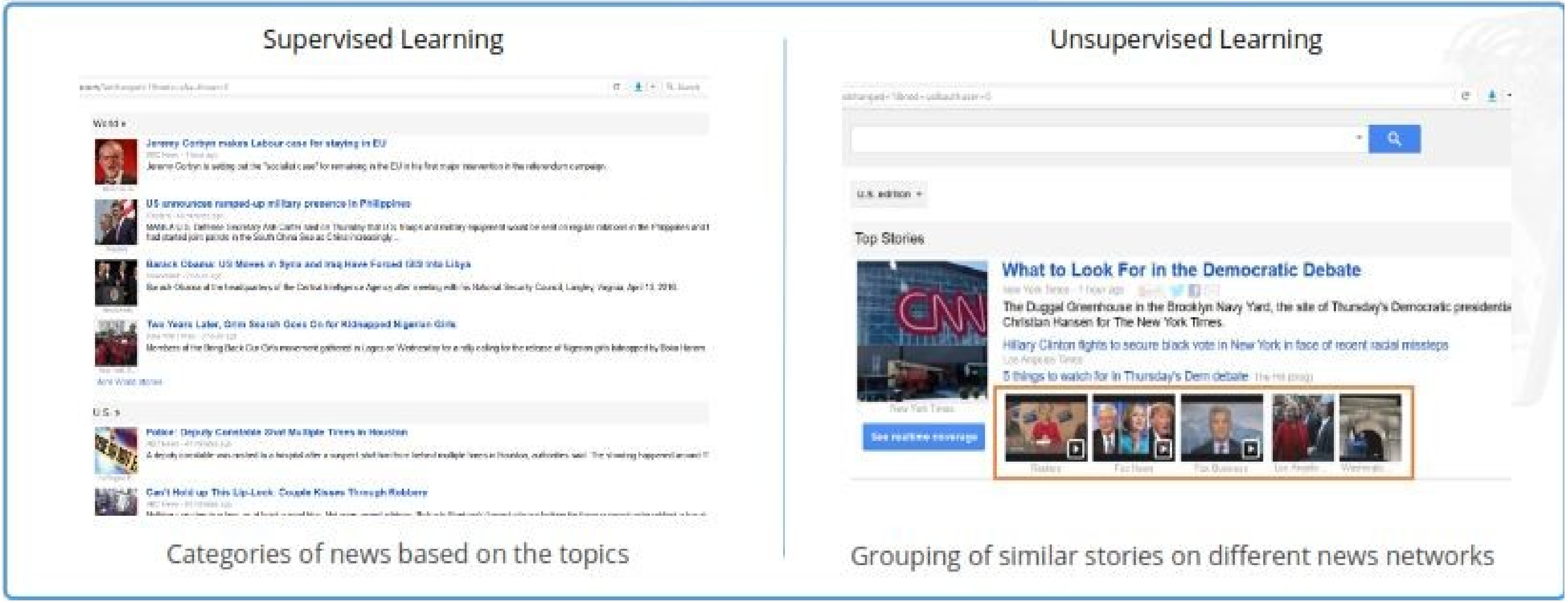
Problem Types

- Data can either be continuous or categorical. Based on whether it is supervised or unsupervised learning, the problem type will differ.



Examples

- Some examples of supervised and unsupervised learning models are:



Supervised Learning

World ▸

Jeremy Corbyn makes Labour case for staying in EU

US announces ramped-up military presence in Philippines

Barack Obama: US Moves in Syria and Iraq Have Forced ISIS Into Libya

Two Years Later, Grim Search Goes On for Kidnapped Nigerian Girls

More World stories

U.S. ▸

Police: Deputy Constable Shot Multiple Times in Houston

Can't Hold up This Lip-Lock: Couple Kisses Through Robbery

Categories of news based on the topics

Unsupervised Learning

U.S. edition ▾

Top Stories

What to Look For in the Democratic Debate

New York Times · 1 hour ago

The Duggal Greenhouse in the Brooklyn Navy Yard, the site of Thursday's Democratic presidentia Christian Hansen for The New York Times.

Hillary Clinton fights to secure black vote in New York in face of recent racial missteps

Los Angeles Times

5 things to watch for in Thursday's Dem debate The Hill (blog)

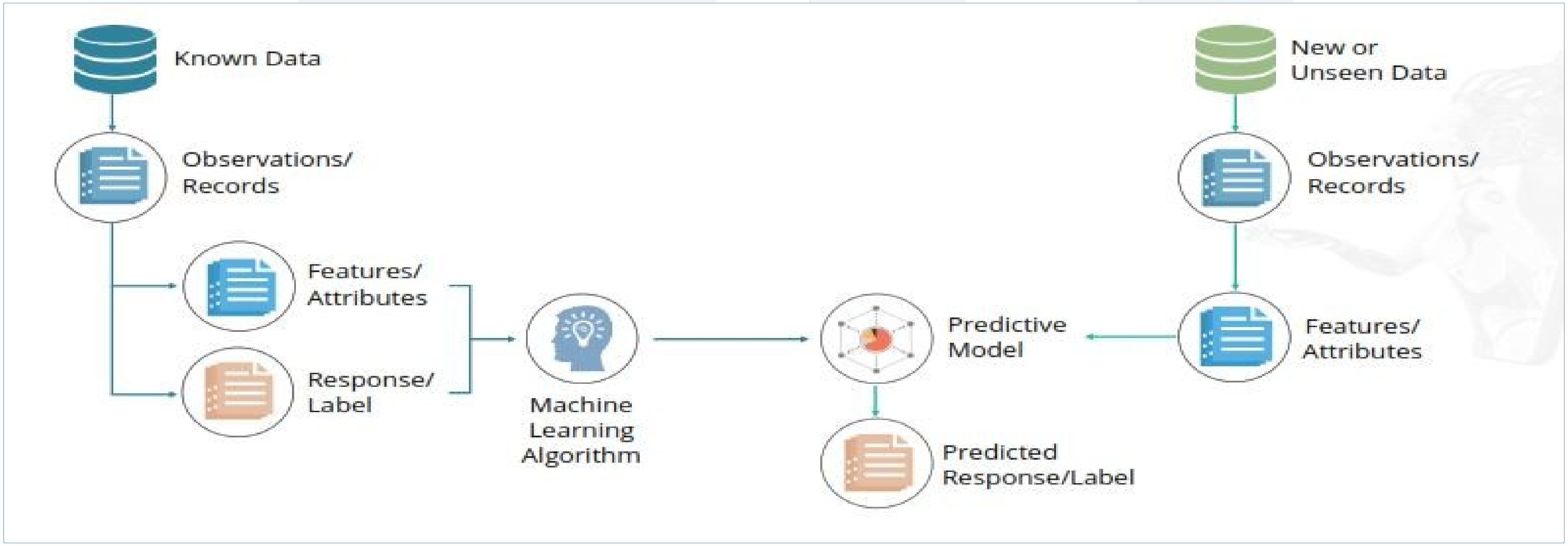
New York Times

See realtime coverage

Grouping of similar stories on different news networks

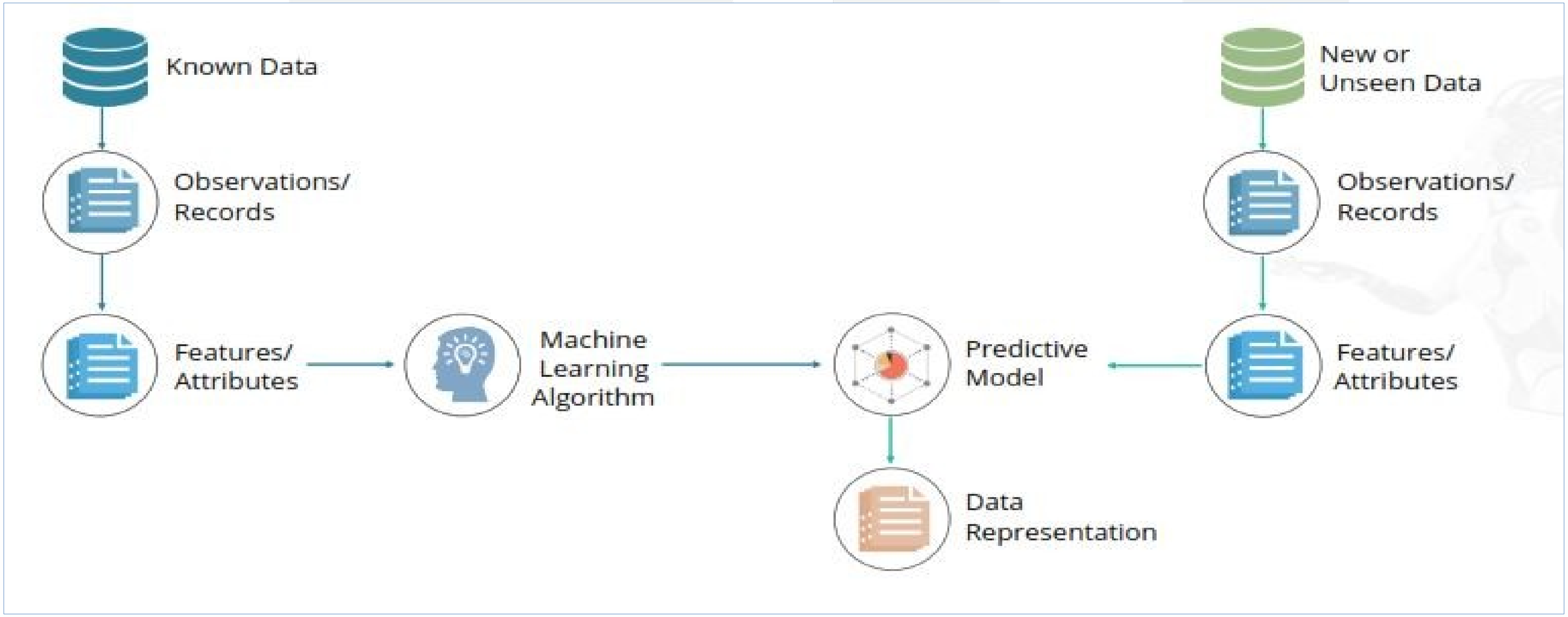
Working of Supervised Learning Model

- In supervised learning, a known dataset with observations, features, and response is used to create and train a machine learning algorithm.
- A predictive model, built on top of this algorithm, is then used to predict the response for a new dataset that has the same features.



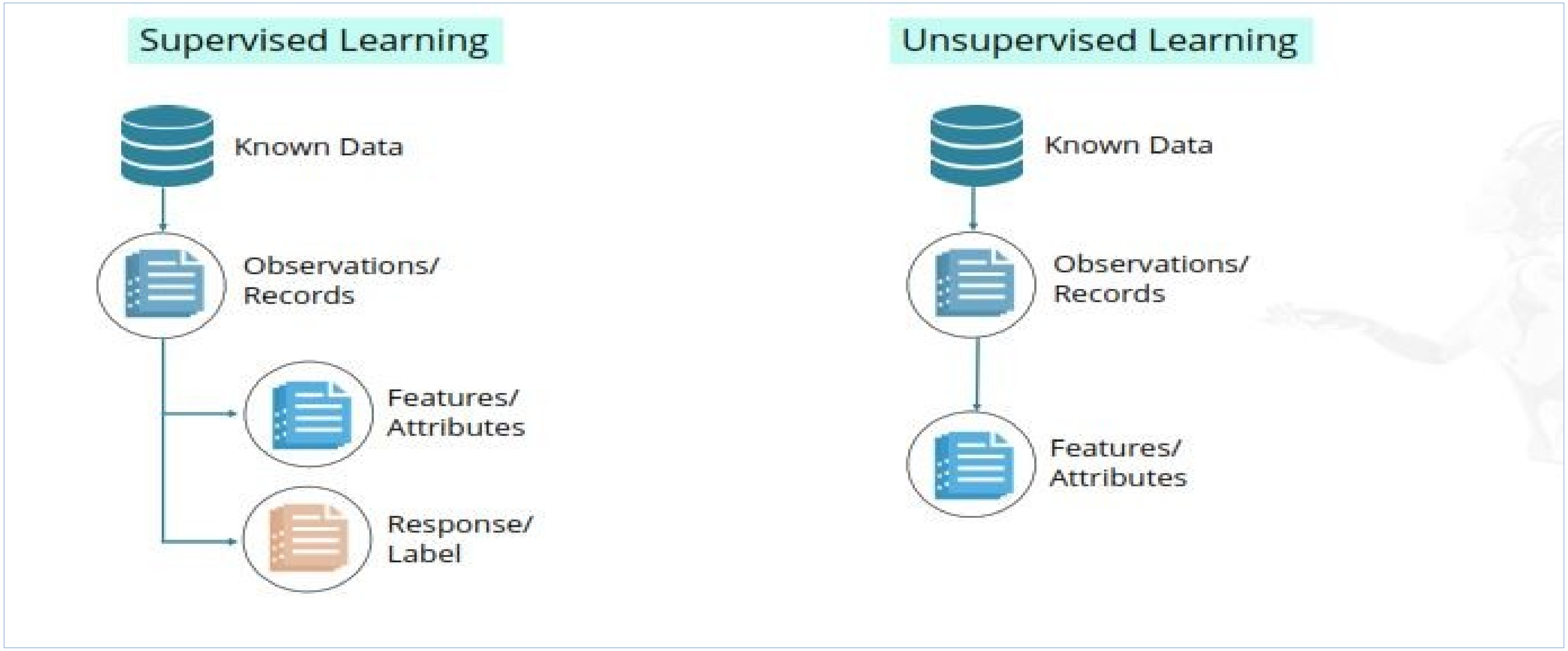
Working of Unsupervised Learning Model

- In unsupervised learning, a known dataset has a set of observations with features. But the response is not known.
- The predictive model uses these features to identify how to classify and represent the data points of new or unseen data.

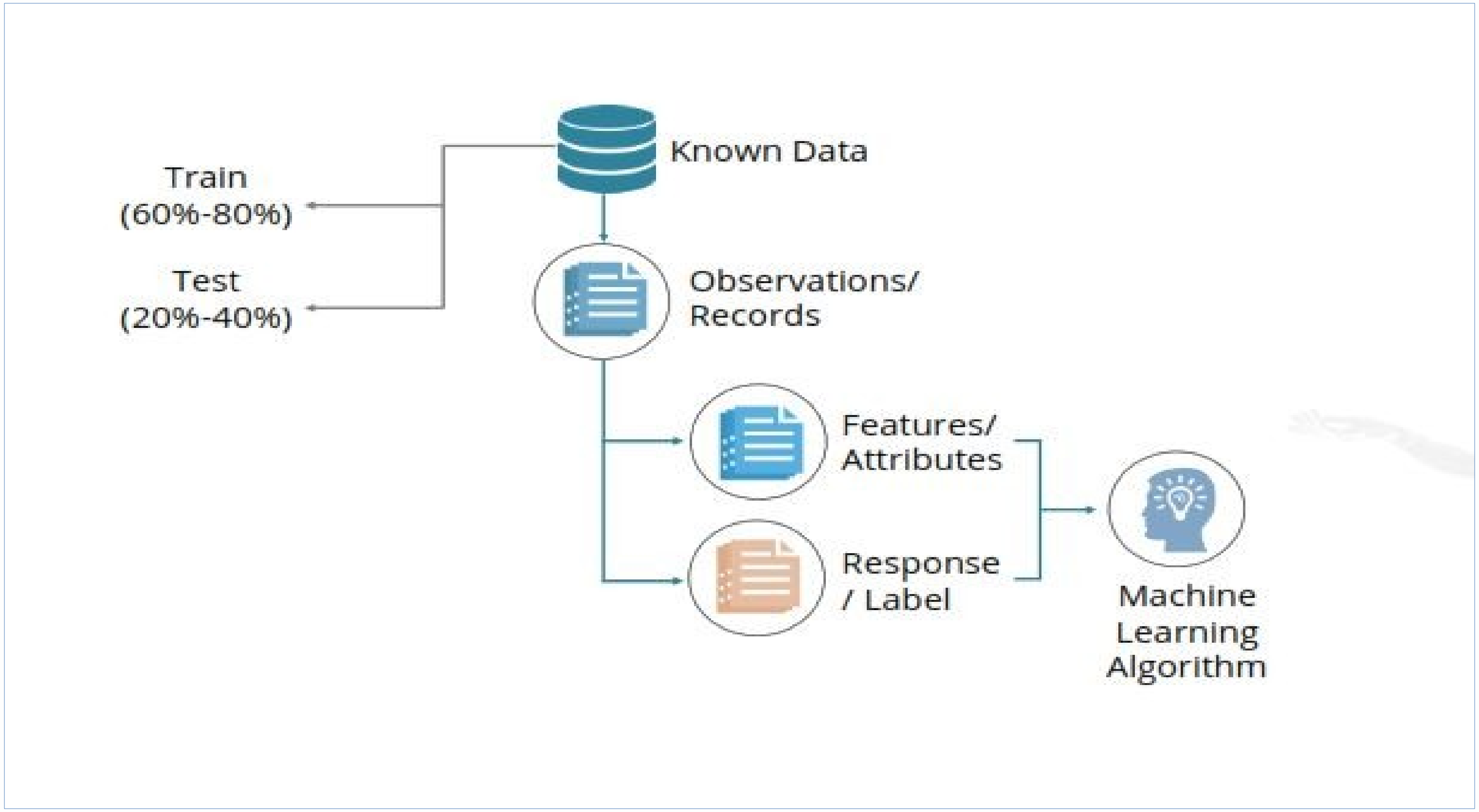


Steps 5 and 6: Train, Test, and Optimize the Model

- To train supervised learning models, data analysts usually divide a known dataset into training and testing sets.



Steps 5 and 6: Train, Test, and Optimize the Model



Scikit-Learn

- Scikit is a powerful and modern machine learning Python library for fully and semi automated data analysis and information extraction.
- Efficient tools to identify and organize problems (Supervised/ Unsupervised)
- Free and open datasets
- Rich set of libraries for learning and predicting
- Model support for every problem type
- Model persistence
- Open source community and vendor support

Scikit-Learn: Problem-Solution Approach

- Scikit-learn helps Data Scientists organize their work through its problem-solution approach.

1. Model Selection
2. Estimator Object
3. Model Training
4. Predictions
5. Model Tuning
6. Accuracy

Scikit-Learn: Problem-Solution Considerations

- While working with a Scikit-Learn dataset or loading your own data to Scikit - Learn, always consider these points:

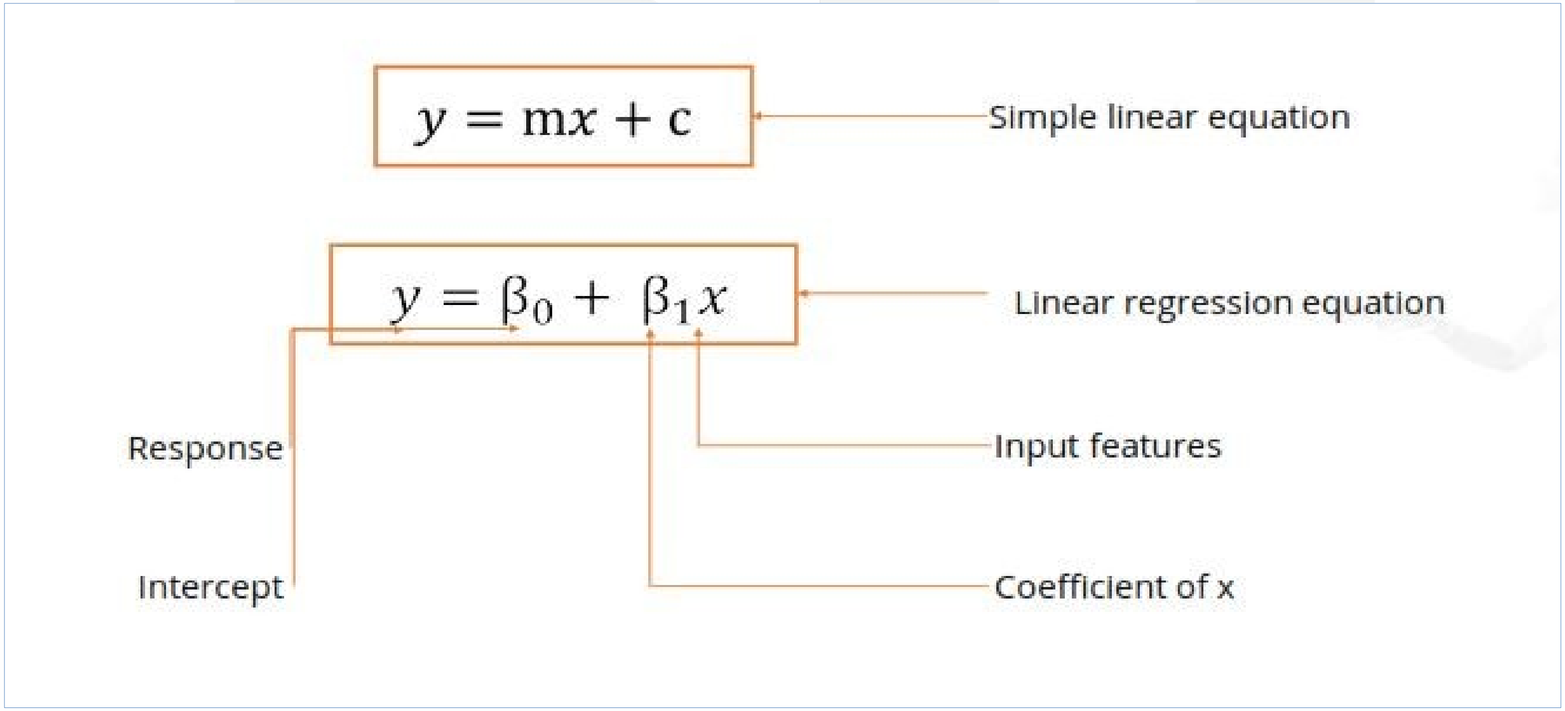
1. Create separate objects for feature and response
2. Ensure that features and response have only numeric values
3. Features and response should be in the form of a NumPy ndarray
4. Since features and response would be in the form of arrays, they would have shapes and sizes
5. Features are always mapped as ***x, and response is mapped as y***

Supervised Learning Models: Linear Regression

- Linear regression is a supervised learning model used to analyze continuous data.
- It is the most basic and widely used technique to predict a value of an attribute
- It is easy to use as the model does not require a lot of tuning
- It runs very fast, which makes it more time-efficient

Supervised Learning Models: Linear Regression

- The linear regression equation is based on the formula for a simple linear equation.

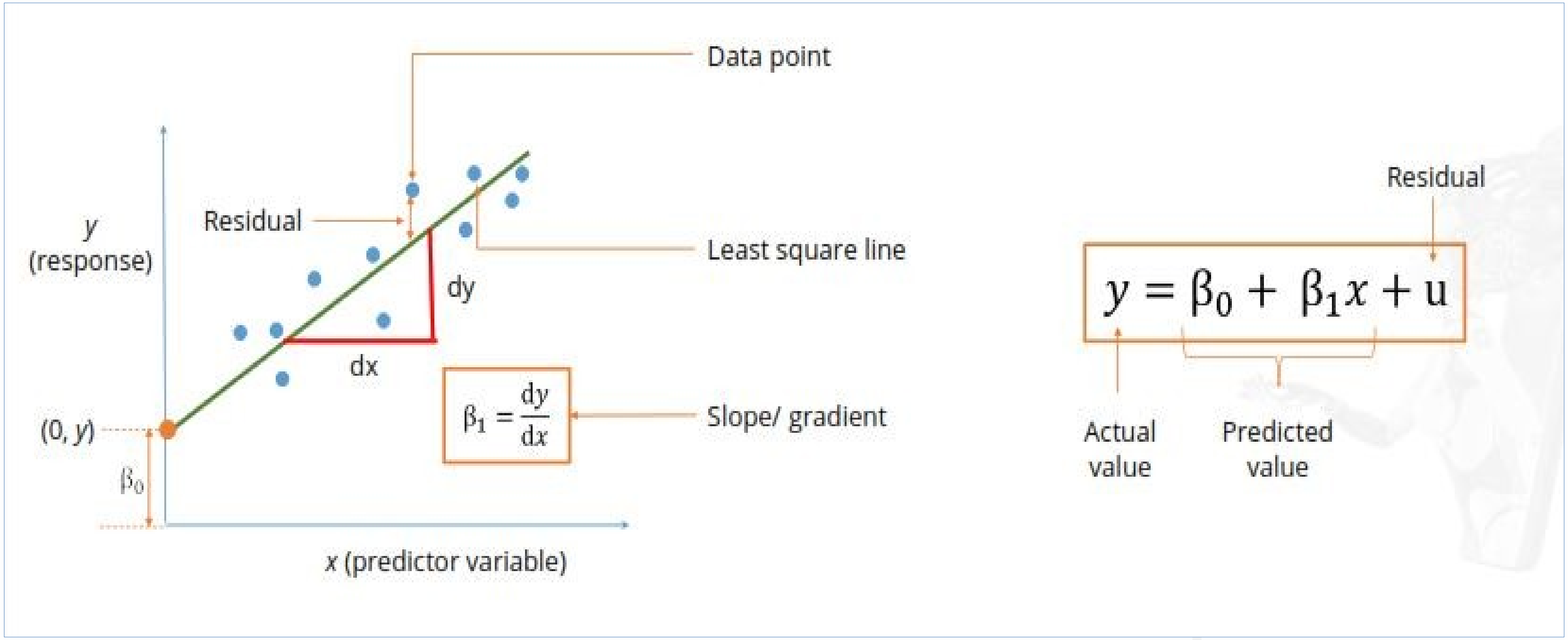


$$y = mx + c$$

$$y = \beta_0 + \beta_1 x$$

Supervised Learning Models: Linear Regression

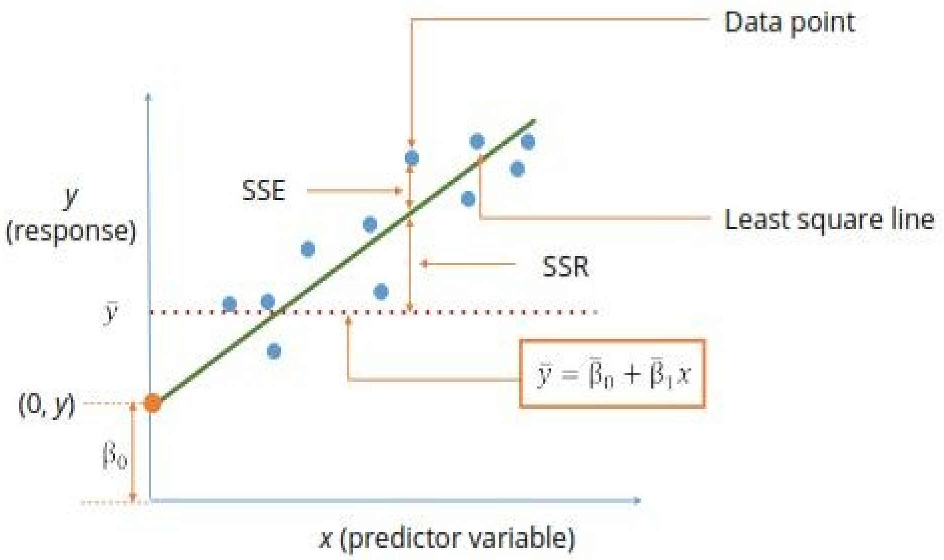
- Linear regression is the most basic technique to predict a value of an attribute.



- The attributes are usually fitted using the “least square” approach.

Supervised Learning Models: Linear Regression

- Smaller the value of SSR or SSE, the more accurate the prediction will be, which would make the model the best fit.



$$y = \beta_0 + \beta_1 x + u$$

$$SSR = \sum(\hat{y}_i - \bar{y})^2$$

Regression of sum of squares

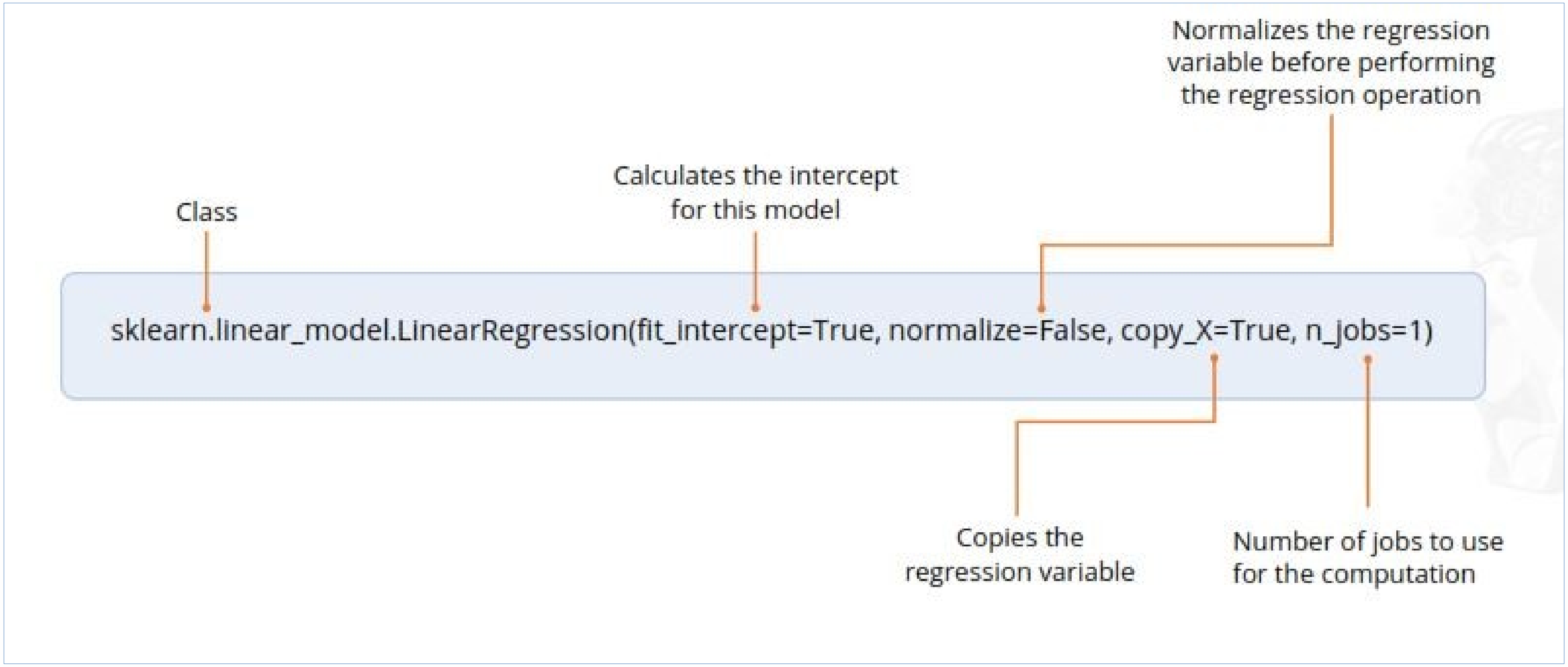
$$SSE = \sum(y_i - \hat{y}_i)^2$$

Error of sum of squares

- The attributes are usually fitted using the “least square” approach.

Supervised Learning Models: Linear Regression

- Let us see how linear regression works in Scikit-Learn.

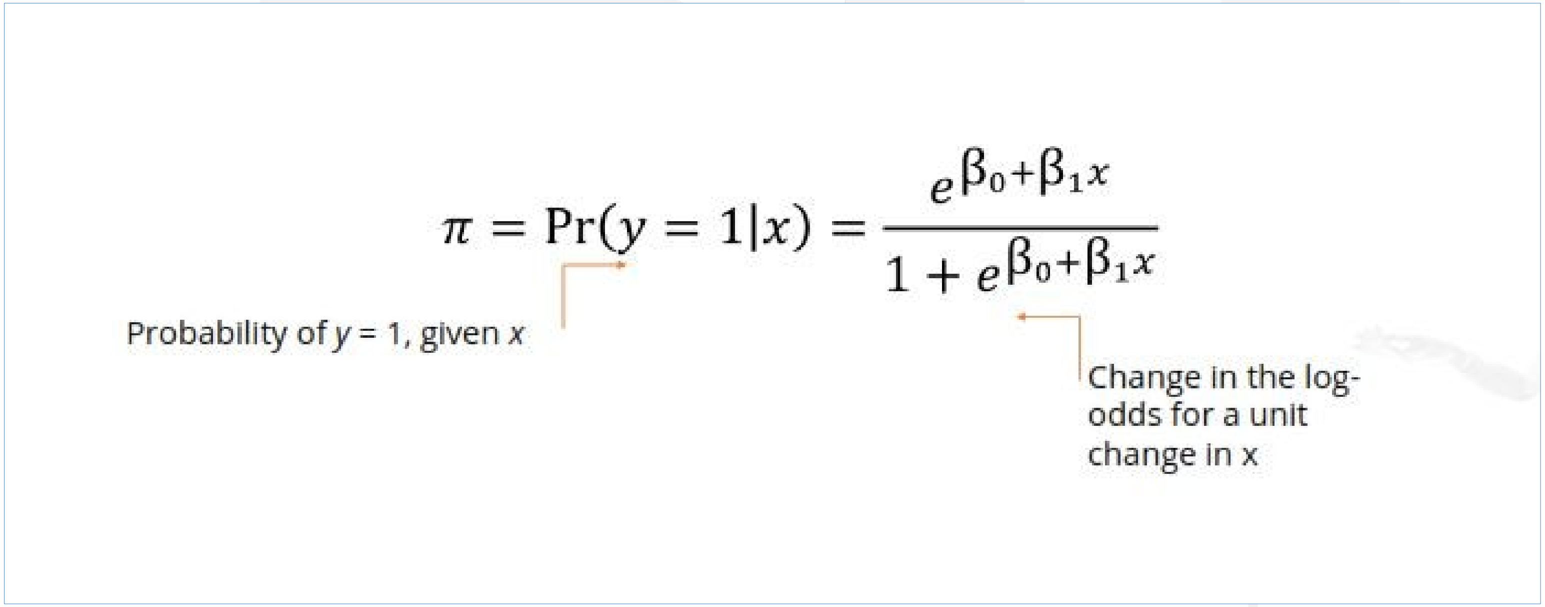


```
sklearn.linear_model.LinearRegression(fit_intercept=True, normalize=False, copy_X=True, n_jobs=1)
```

Normalizes the regression variable before performing the regression operation
Class
Calculates the intercept for this model
Copies the regression variable
Number of jobs to use for the computation

Supervised Learning Models: Logistic Regression

- Logistic regression is a generalization of the linear regression model used for classification problems.



$$\pi = \Pr(y = 1|x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

Probability of *y* = 1, given *x*

Change in the log-odds for a unit change in x

- The purpose of K-NN is to predict the class for each observation.

Supervised Learning Models: Logistic Regression

- Logistic regression is a generalization of the linear regression model used for classification problems.

$$\text{Odds} = \frac{\pi}{1-\pi}$$

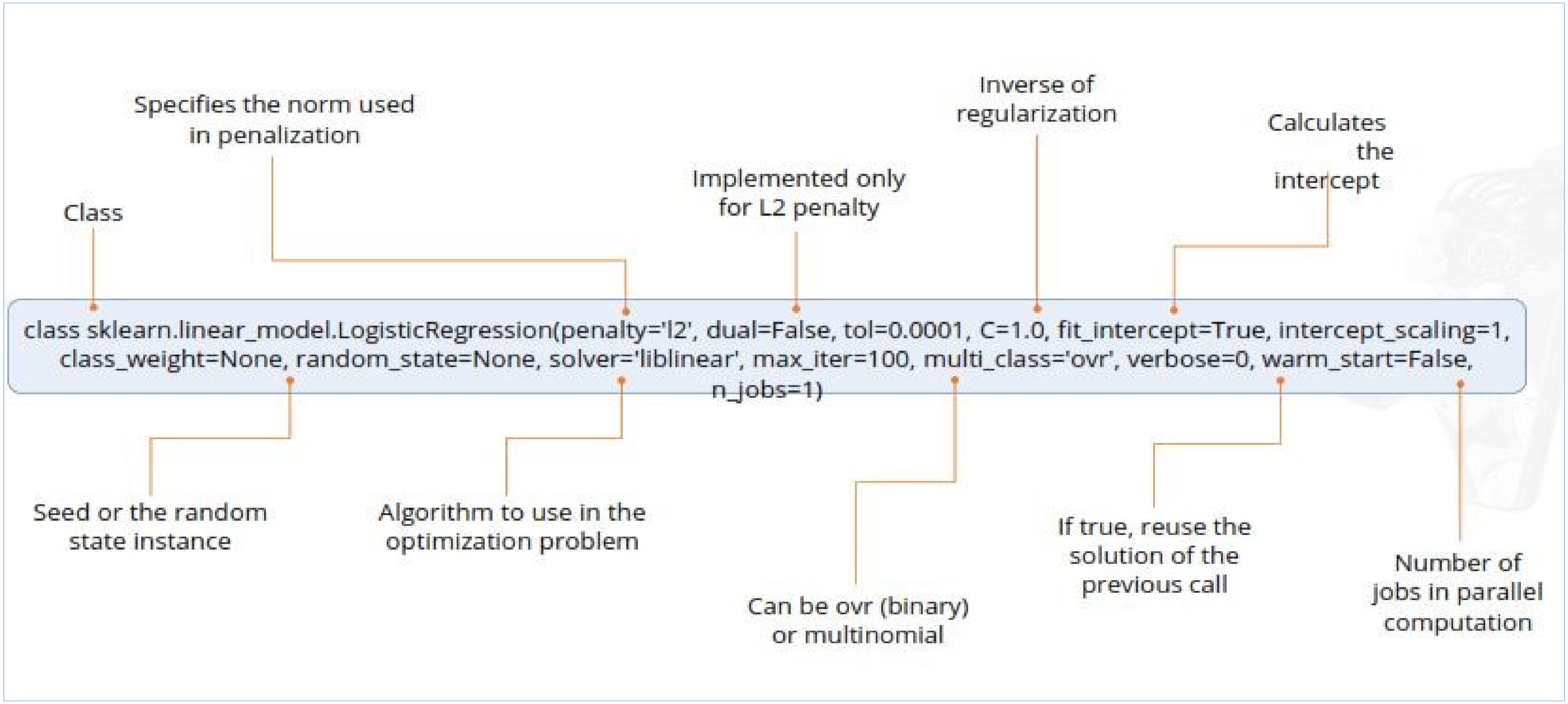
Probability

$$\log\left(\frac{\pi}{1-\pi}\right) = \log\left(e^{\beta_0+\beta_1 x}\right) = \beta_0 + \beta_1 x$$

Logarithm of odds

Linear regression

Supervised Learning Models: Logistic Regression



```
class sklearn.linear_model.LogisticRegression(penalty='l2', dual=False, tol=0.0001, C=1.0, fit_intercept=True, intercept_scaling=1,
class_weight=None, random_state=None, solver='liblinear', max_iter=100, multi_class='ovr', verbose=0, warm_start=False,
n_jobs=1)
```

Class

Specifies the norm used in penalization

Implemented only for L2 penalty

Inverse of regularization

Calculates the intercept

Seed or the random state instance

Algorithm to use in the optimization problem

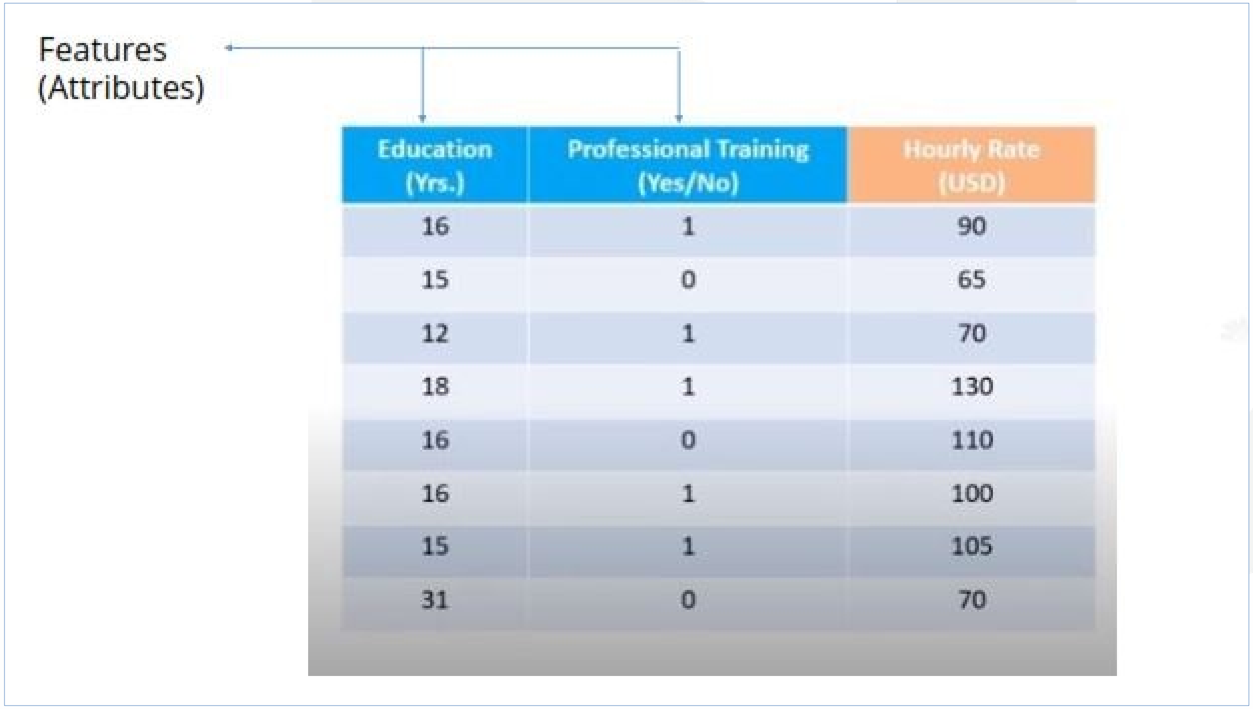
Can be ovr (binary) or multinomial

If true, reuse the solution of the previous call

Number of jobs in parallel computation

Supervised Learning Models: : K-Nearest Neighbors

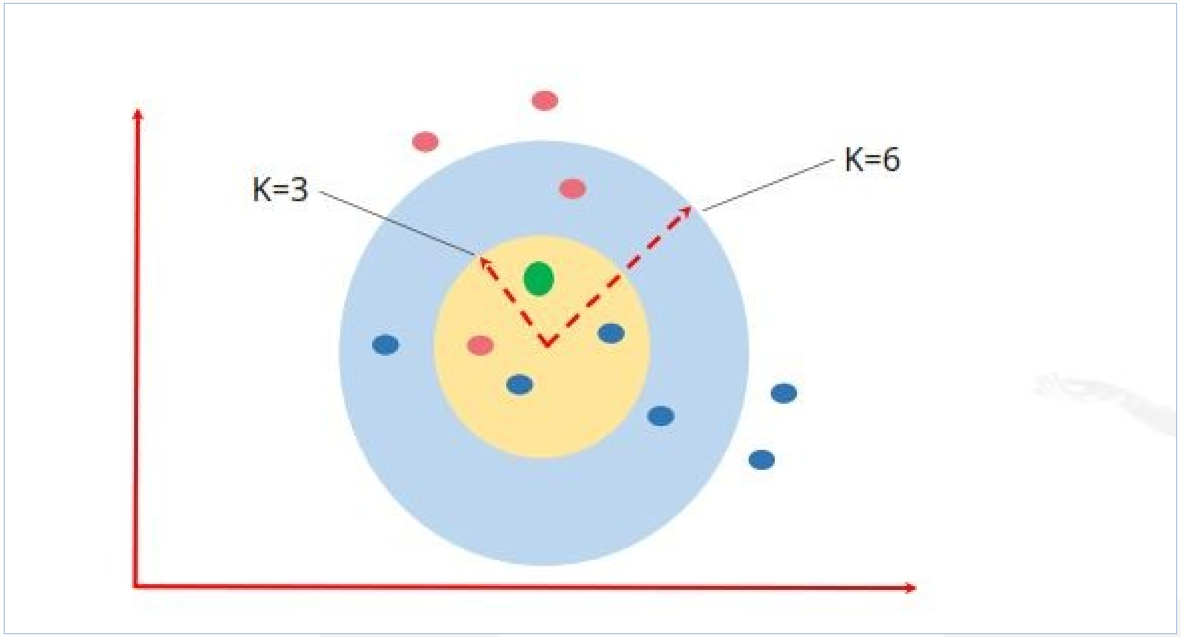
- K-Nearest Neighbors, or K-NN, is one of the simplest machine learning algorithms used for both classification and regression problem types.



Features (Attributes)

Education (Yrs.)	Professional Training (Yes/No)	Hourly Rate (USD)
16	1	90
15	0	65
12	1	70
18	1	130
16	0	110
16	1	100
15	1	105
31	0	70

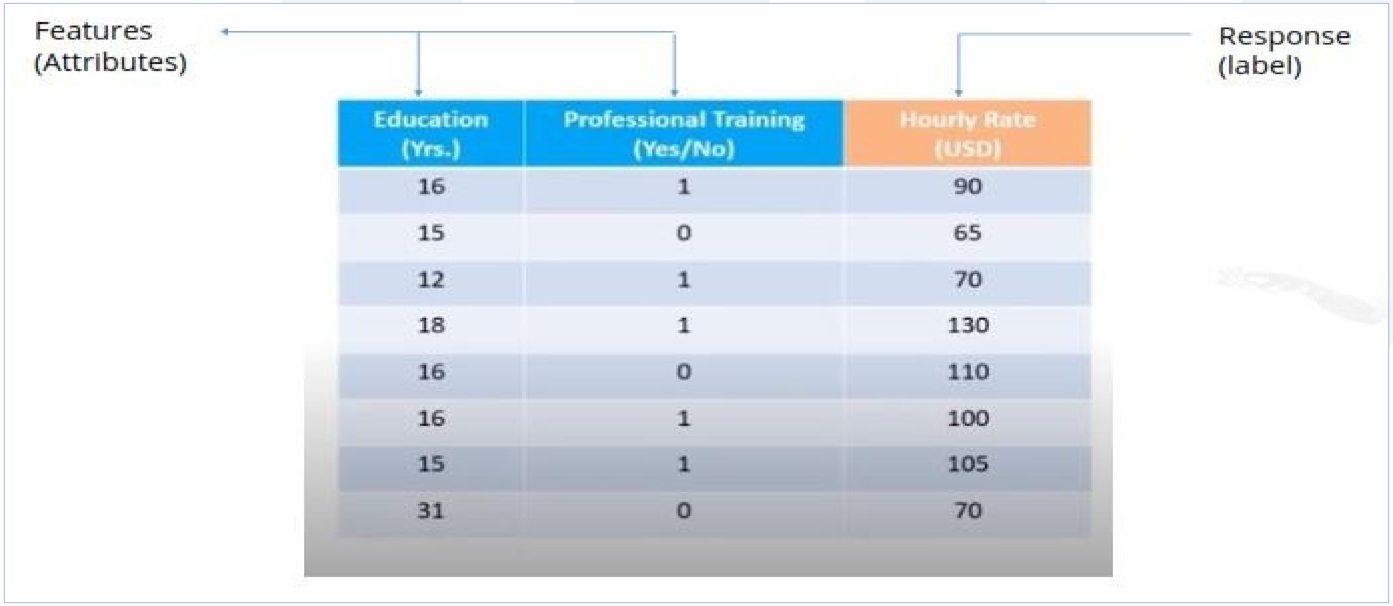
Supervised Learning Models: : K-Nearest Neighbors



- If you are using this method for binary classification, choose an odd number for k to avoid the case of a **tied** distance between two classes.

Supervised Learning Models: : K-Nearest Neighbors

- K-means finds the best centroids by alternatively assigning random centroids to a dataset and selecting mean data points from the resulting clusters to form new centroids. It continues this process iteratively until the model is optimized.



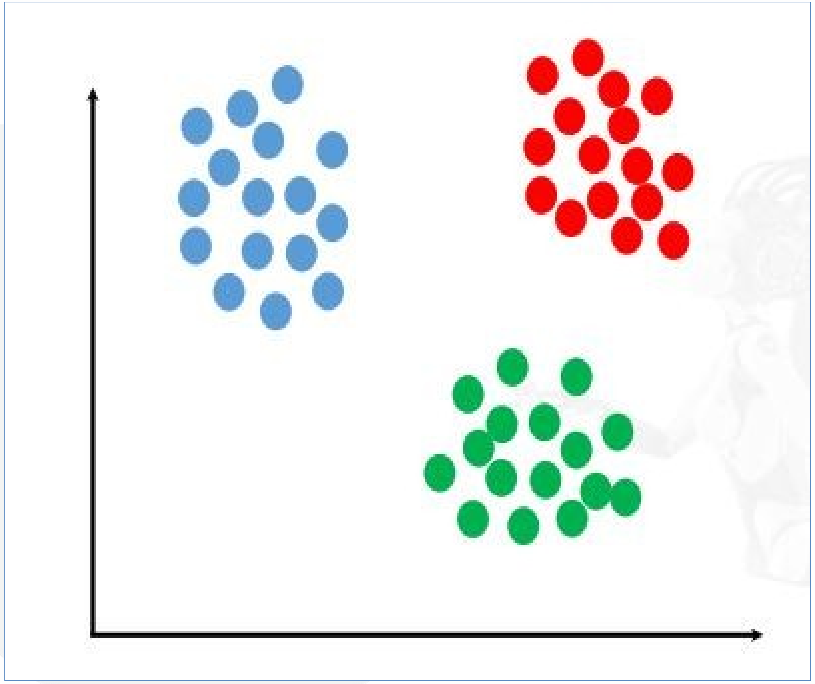
Features (Attributes)

Response (label)

Education (Yrs.)	Professional Training (Yes/No)	Hourly Rate (USD)
16	1	90
15	0	65
12	1	70
18	1	130
16	0	110
16	1	100
15	1	105
31	0	70

Unsupervised Learning Models: Clustering

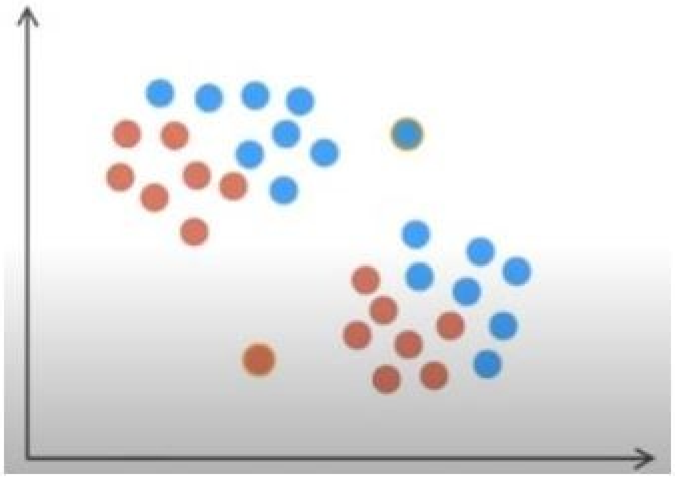
- A cluster is a group of similar data points.
- Clustering is used to:
- Extract the structure of the data
- Identify groups in the data



- Greater similarity between data points results in better clustering.

Unsupervised Learning Models: K-Means Clustering

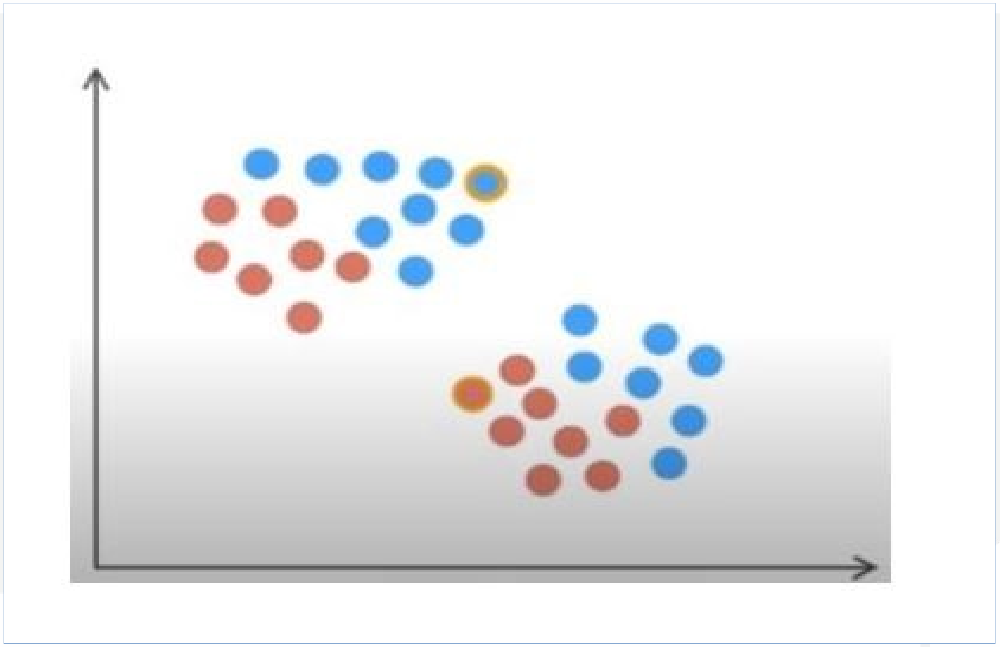
- K-means finds the best centroids by alternatively assigning random centroids to a dataset and selecting mean data points from the resulting clusters to form new centroids. It continues this process iteratively until the model is optimized.



Assign data points to the centroids

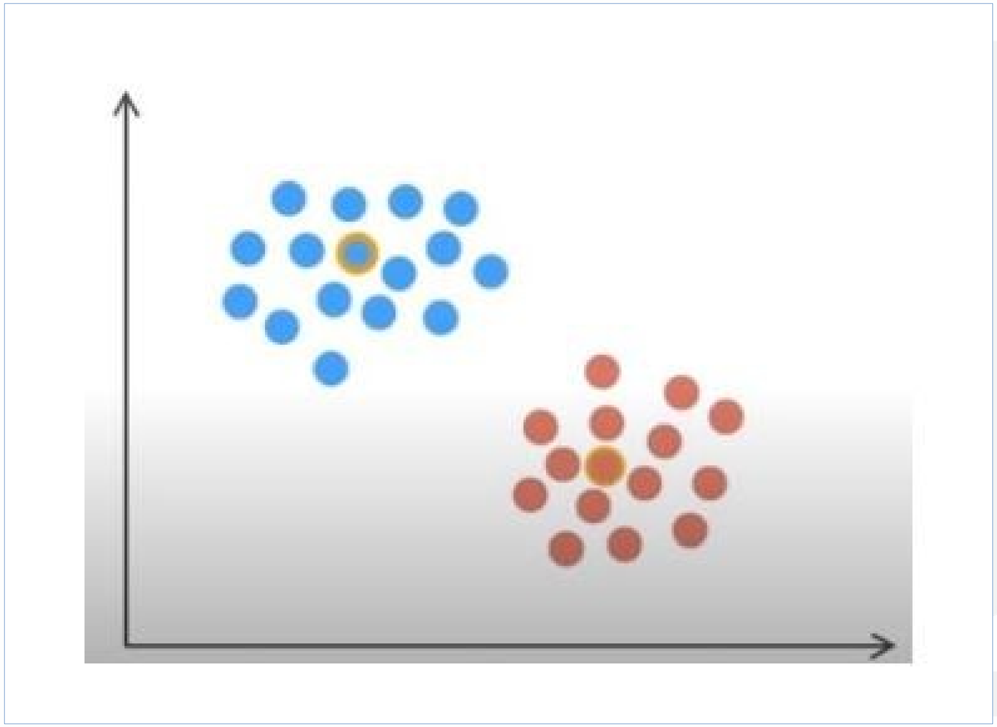
Unsupervised Learning Models: K-Means Clustering

Choose a mean from each cluster as a centroid



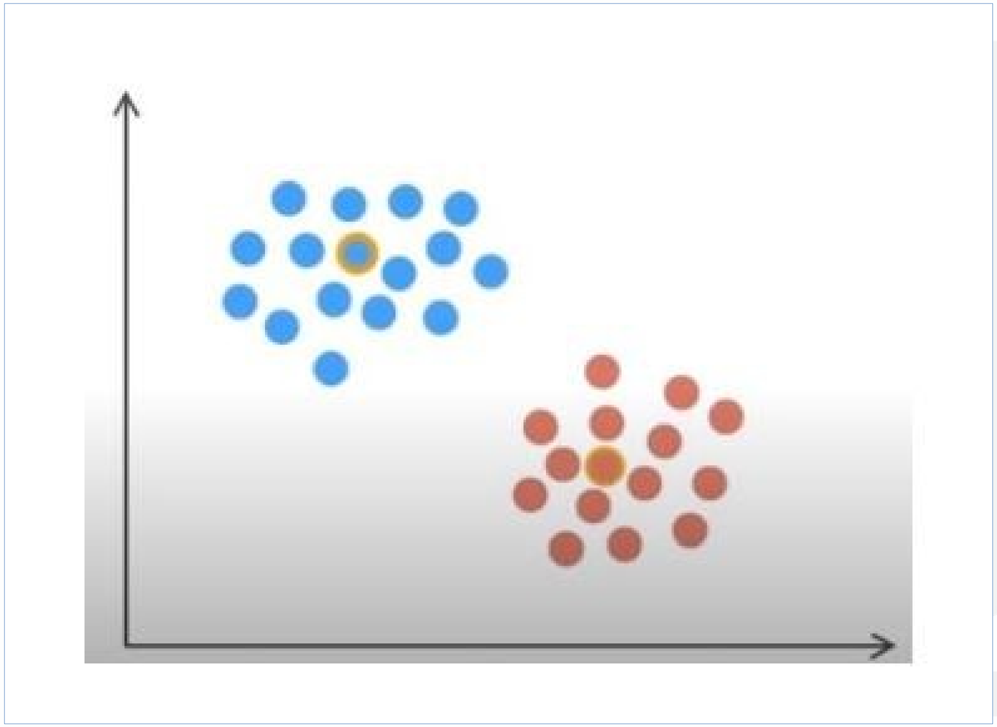
Unsupervised Learning Models: K-Means Clustering

Reassign data points to new centroids



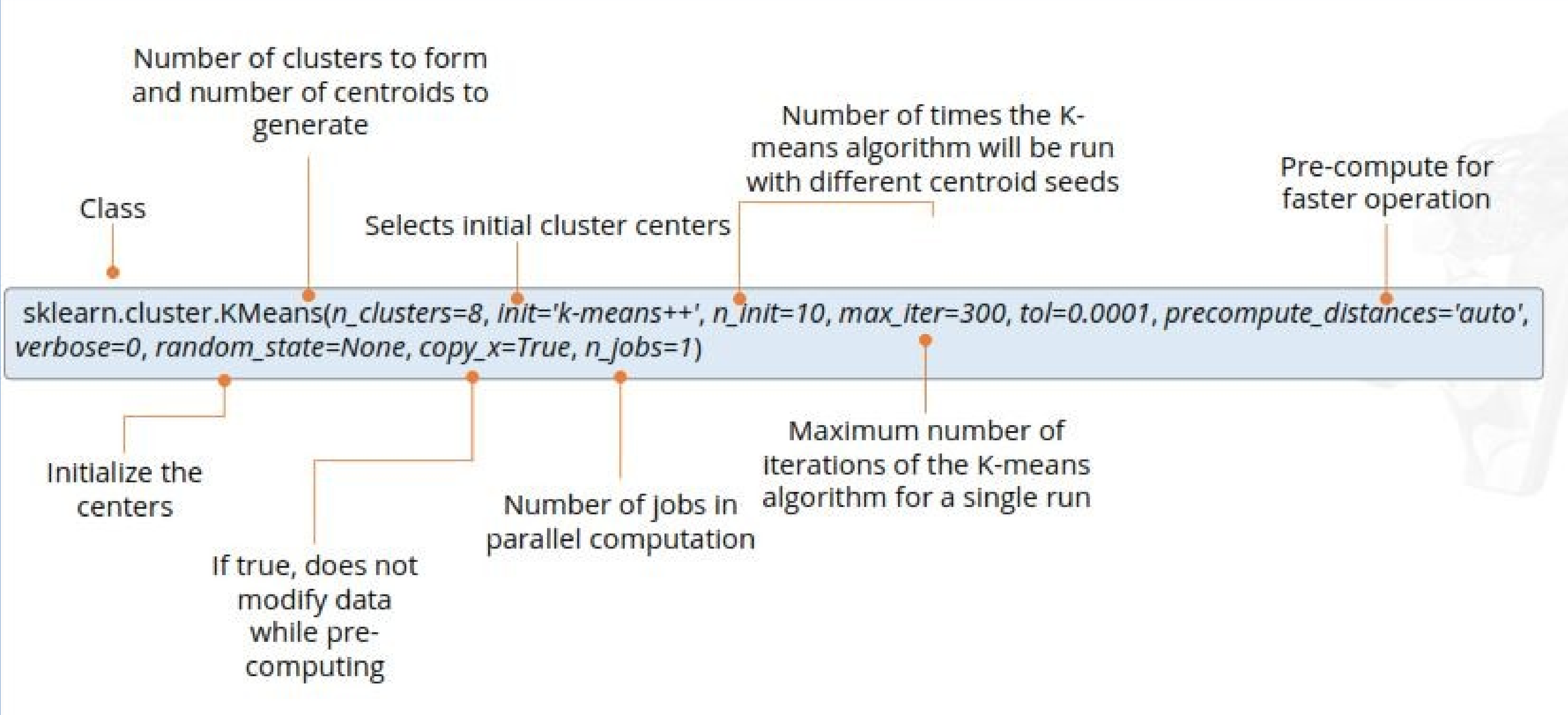
Unsupervised Learning Models: K-Means Clustering

Iterate the process till the model is optimized



Unsupervised Learning Models: K-Means Clustering

- Let us see how the k-means algorithm works in Scikit-Learn.

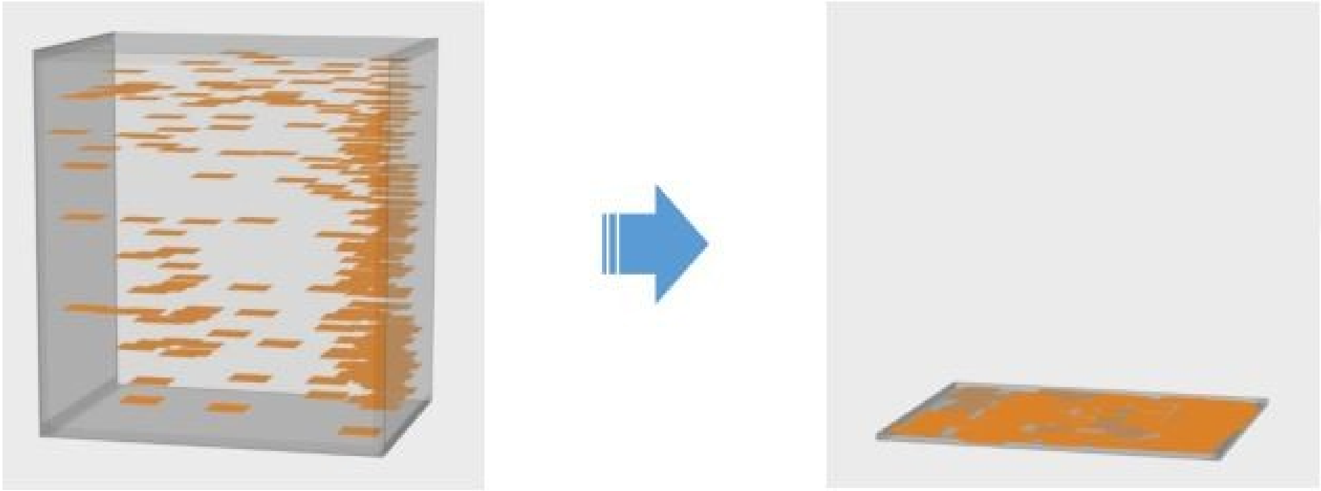


```
sklearn.cluster.KMeans(n_clusters=8, init='k-means++', n_init=10, max_iter=300, tol=0.0001, precompute_distances='auto',
verbose=0, random_state=None, copy_x=True, n_jobs=1)
```

Class
Number of clusters to form and number of centroids to generate
Selects initial cluster centers
Number of times the K-means algorithm will be run with different centroid seeds
Pre-compute for faster operation
Initialize the centers
If true, does not modify data while pre-computing
Number of jobs in parallel computation
Maximum number of iterations of the K-means algorithm for a single run

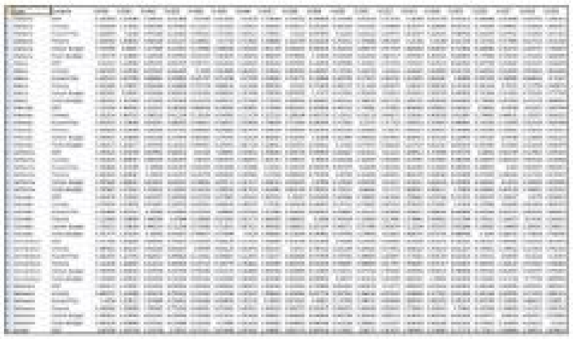
Unsupervised Learning Models: Dimensionality Reduction

- It reduces a high-dimensional dataset into a dataset with fewer dimensions. This makes it easier and faster for the algorithm to analyze the data.



Unsupervised Learning Models: Dimensionality Reduction

- These are some techniques used for dimensionality reduction:
- Drop data columns with missing values
- Drop data columns with low variance
- Drop data columns with high correlations
- Apply statistical functions - PCA

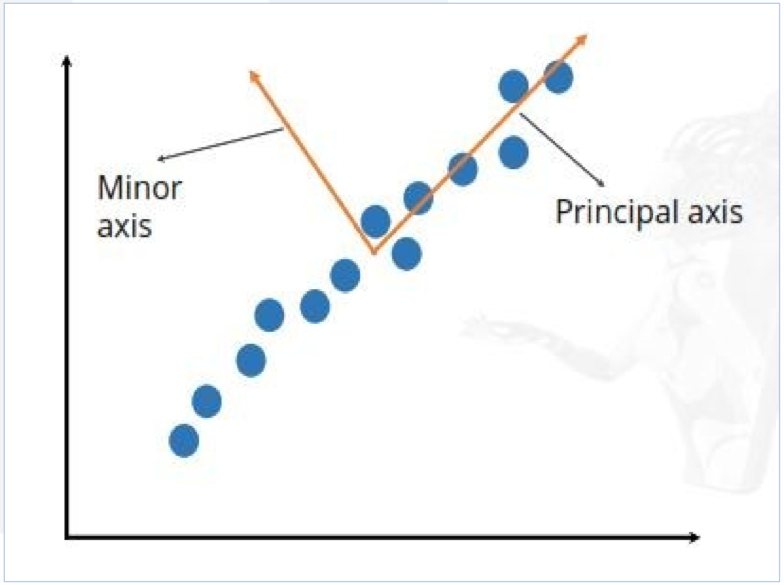


Large dataset
(a few thousand columns and rows)

Unsupervised Learning Models: Principal Component Analysis

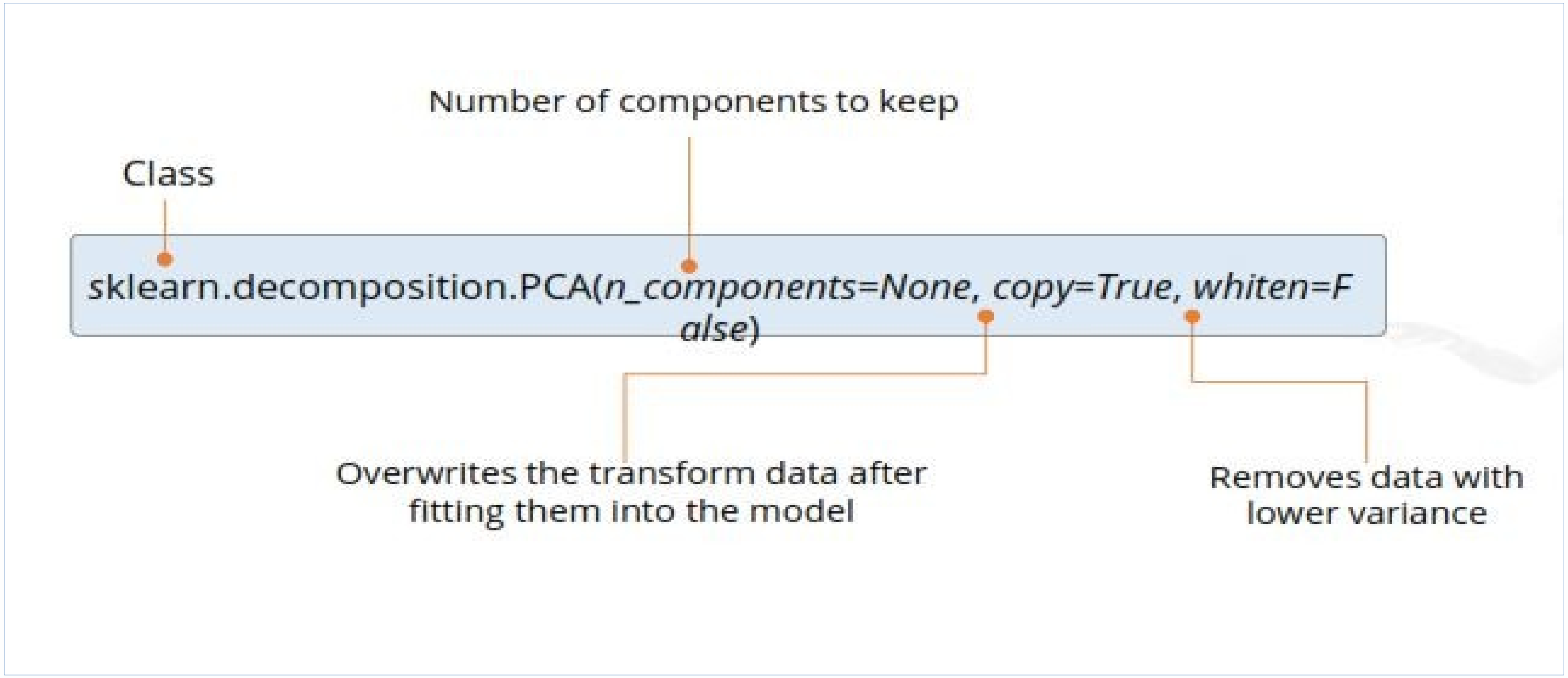
- It is a linear dimensionality reduction method which uses singular value decomposition of the data and keeps only the most significant singular vectors to project the data to a lower dimensional space.

- It is primarily used to compress or reduce the data.
- PCA tries to capture the variance, which helps it pick up interesting features.
- PCA is used to reduce dimensionality in the dataset and to build our feature vector.
- Here, the principal axes in the feature space represents the direction of maximum variance in the data.
- This method is used to capture variance.



Unsupervised Learning Models: Principal Component Analysis

- Let us look at how the PCA algorithm works in Scikit-Learn.



```
sklearn.decomposition.PCA(n_components=None, copy=True, whiten=False)
```

Number of components to keep
Class
Overwrites the transform data after fitting them into the model
Removes data with lower variance

Pipeline

- It simplifies the process where more than one model is required or used.
- All models in the pipeline must be transformers. The last model can either be a transformer or a classifier, regressor, or other such objects.
- Once all the data is fit into the models or estimators, the predict method can be called.
- Estimators are known as ‘model instance’.

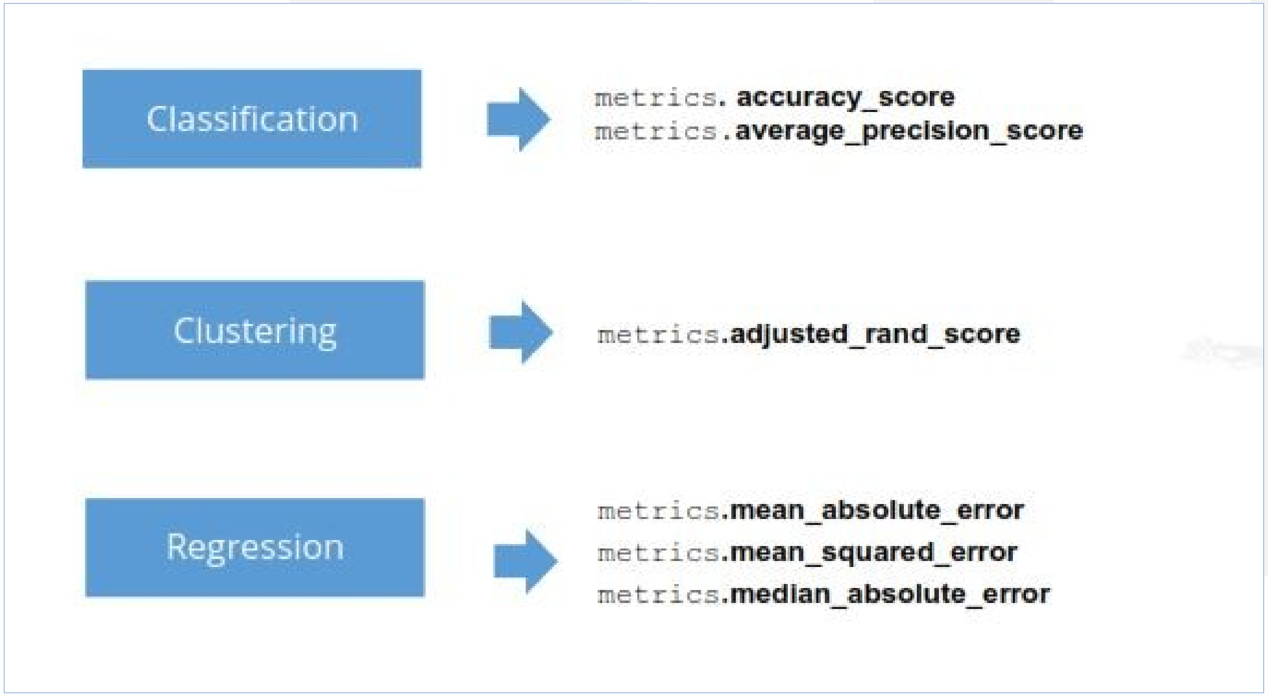
Model Persistence

- You can save your model for future use. This avoids the need to retrain the model.

- This can be saved using the Pickle method.
- It can also be replaced with the joblib of Sci-kit team.
- Both joblib.dump and joblib.load can be used.
- These would be efficient for Big Data.

Model Evaluation: Metric Functions

- You can use the “Metrics” function to evaluate the accuracy of your model’s predictions.



Classification → metrics. accuracy_score
metrics.average_precision_score

Clustering → metrics.adjusted_rand_score

Regression → metrics.mean_absolute_error
metrics.mean_squared_error
metrics.median_absolute_error

DIGITAL LEARNING CONTENT

Parul® University

www.paruluniversity.ac.in