SNS COLLEGE OF TECHNOLOGY

Coimbatore-35
An Autonomous Institution

Accredited by NBA – AICTE and Accredited by NAAC – UGC with 'A++' Grade
Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

23AMB201 - MACHINE LEARNING

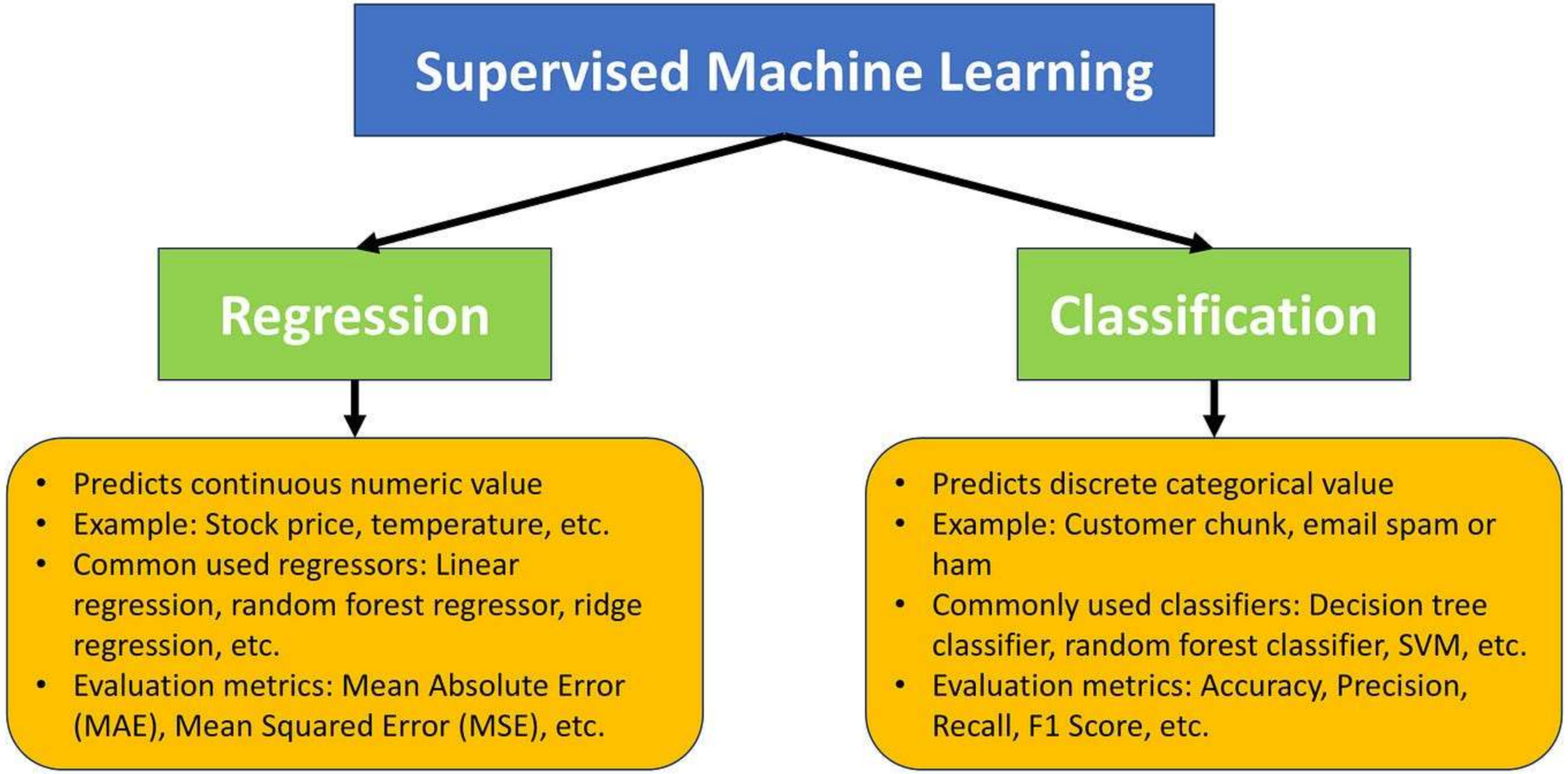
II YEAR IV SEM

UNIT II – SUPERVISED LEARNING ALGORITHMS

TOPIC 2,3 – Classification and Regression – Linear Regression

Supervised Learning

1. Supervised Learning is a branch of Artificial Intelligence that focuses on training models to make predictions or decision based on labeled training data.



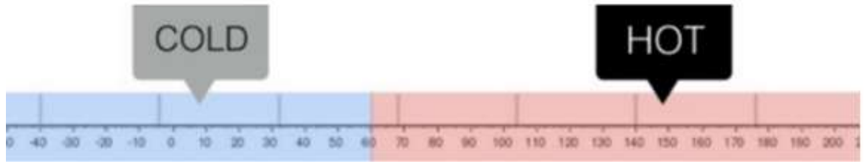
Regression

What will be the temperature tomorrow?

Predictors				Target
Outlook	Temp.	Humidity	Windy	Hours Played
Rainy	Hot	High	False	25
Rainy	Hot	High	True	30
Overcast	Hot	High	False	48
Sunny	Mild	High	False	46
Sunny	Cool	Normal	False	52
Sunny	Cool	Normal	True	23
Overcast	Cool	Normal	True	43
Rainy	Mild	High	False	35
Rainy	Cool	Normal	False	38

Classification

Will play or don't play



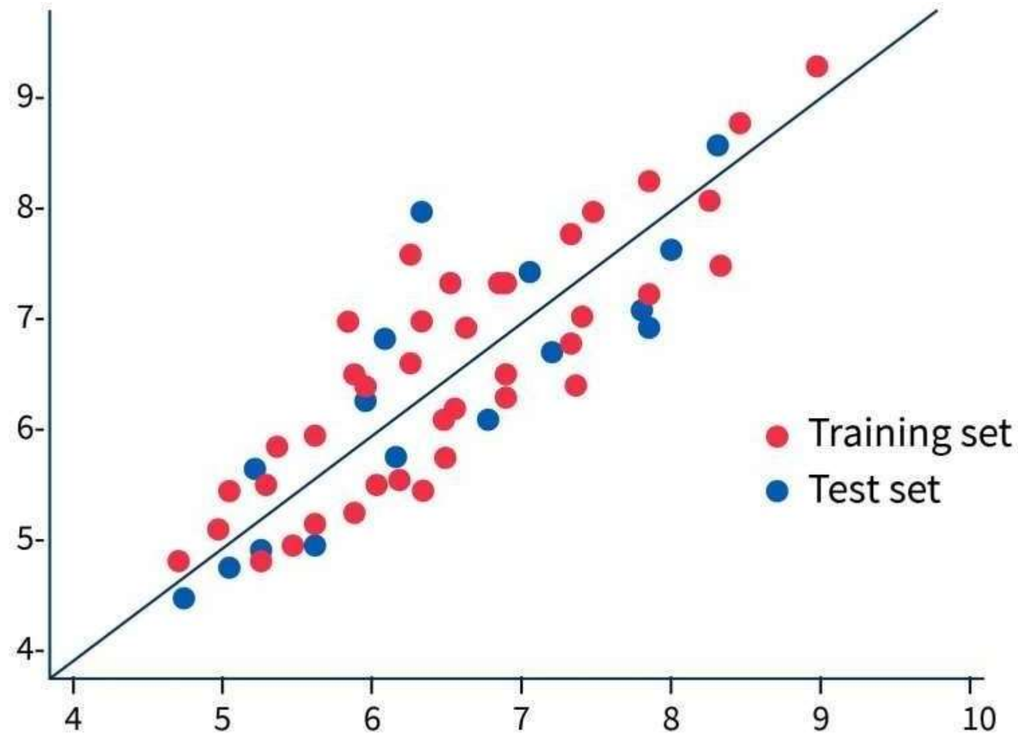
Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	false	Don't Play
Sunny	Hot	High	true	Don't Play
Overcast	Hot	High	false	Play
Rain	Mild	High	false	Play
Rain	Cool	Normal	false	Play
Rain	Cool	Normal	true	Don't Play

What is Linear Regression?

Linear regression is **an algorithm that provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events**.

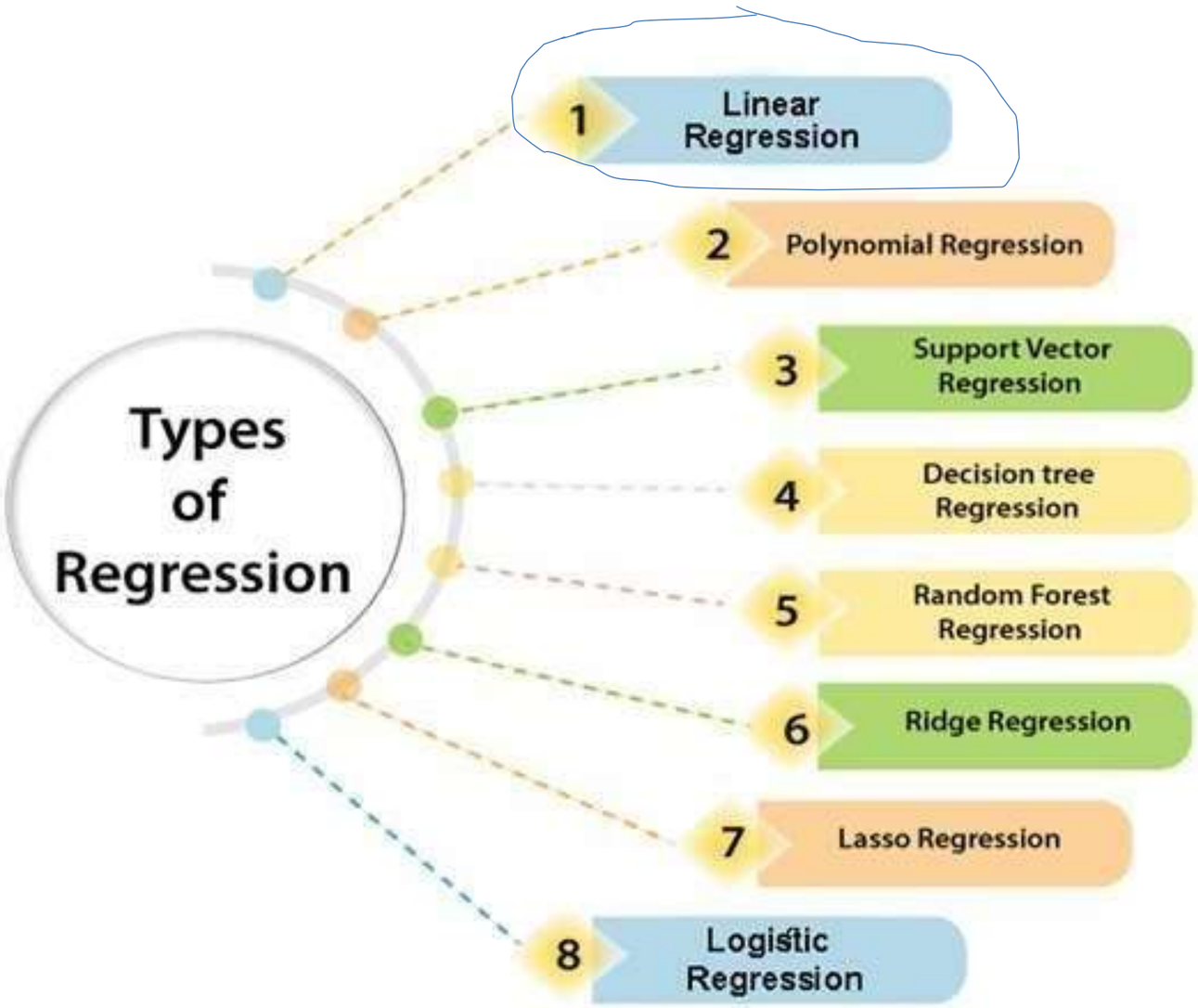
	A	B
1	Exp	Sal
2	8	10
3	10	13
4	12	16
5		

Exp =20 Salary=?



Regression

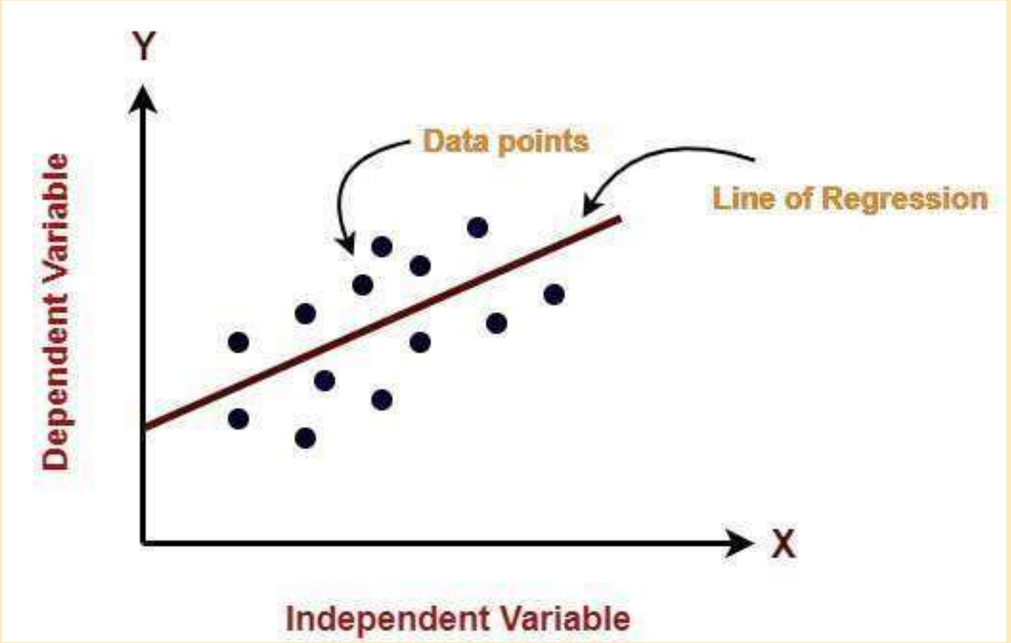
1. Regression analysis is a statistical method to model the relationship between a dependent (target) and independent (predictor) variables with one or more independent variables.
2. Regression analysis helps us to understand how the value of the dependent variable is changing corresponding to an independent variable when other independent variables are held fixed.
3. It predicts continuous/real values such as temperature, age, salary, price, etc.



Univariate Linear Regression

1. Prediction of rain using temperature and other factors
2. Determining Market trends
3. Prediction of road accidents due to rash driving.

Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), hence called linear regression.



- Below is the mathematical equation for Linear regression:

$$Y = aX + b$$

Here, Y = dependent variables (target variables),
X= Independent variables (predictor variables),
a and b are the linear coefficients

Terminologies Related to the Regression Analysis:

Dependent Variable: The main factor in Regression analysis which we want to predict or understand is called the dependent variable. It is also called **target variable**.

Independent Variable: The factors which affect the dependent variables or which are used to predict the values of the dependent variables are called independent variable, also called as a **predictor**.

Outliers: Outlier is an observation which contains either very low value or very high value in comparison to other observed values. An outlier may hamper the result, so it should be avoided.

Underfitting and Overfitting: If our algorithm works well with the training dataset but not well with test dataset, then such problem is called **Overfitting**. And if our algorithm does not perform well even with training dataset, then such problem is called **underfitting**.

Case Example: Employee Salary prediction

Imagine you're studying the relationship between **years of employee experience ("x")** and **salary ("y")** using linear regression.

1. The **intercept** would represent the **predicted salary of an employee with 0 years of experience** (e.g., a baseline salary of $30,000).
2. The **slope** would indicate **how much the salary increases for each additional year of experience** (e.g., a slope of $2,000 means each additional year of experience adds $2,000 to the predicted salary).

Example Equation:

Predicted Salary = 30,000 + 2,000 * Years of Experience

= 30,000 + 2,000 * 21= 72,000

Example: Predicting Employee Salary

Experience in Year (X)	Salary in K(Y)
8	10
10	13
12	16

How much will a worker with 20 years of experience be paid?

Example

X	Y	Mean X	Mean Y	Deviation X	Deviation Y	Product of Deviation	Sum of Product of Deviation	Square of Deviation for X
8	10	(8+10+12)/3 =10	(10+13+16)/3 =13	(8-10) = -2	(10-13) = -3	6	12	4
10	13			(10-10) = 0	(13-13) = 0	0		0
12	16			(12-10) = 2	(16-13) = 3	6		4

m=Sum of Product of Deviations/Sum of Square of Deviations X

m=12/8

m=1.5

b=Mean of Y-(m*Mean of X)

b=13-(1.5*10) b=-2

y=mx+b

y=1.5*20-2

y=28

The Salary for an employee with twenty years of experience is: = 28K

1. Implement Linear regression using SciKit Learn [May use appropriate library]

NumPy is a *Python* library used for working with arrays. It also has functions for working in domain of linear algebra, fourier transform, and matrices.

Matplotlib is **data visualization and graphical plotting library** for Python and its numerical extension NumPy

Pandas is **an open source Python package that is most widely used for data science/data analysis and machine learning tasks**.

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
```

Program

The iloc() function in python is defined in the Pandas module that **helps us to select a specific row or column from the data set**.

```
dataset = pd.read_csv('emp.csv')
x = dataset.iloc[:,:-1].values      #YoE
y = dataset.iloc[:, :1].values      #Salary
```

Program

The train_test_split function of the sklearn. model_selection package in Python **splits arrays or matrices into random subsets for train and test data**

```
# Splitting the dataset into the Training set and Test set
from sklearn.model_selection import train_test_split

x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=1/3)
```

Program

linear_model is a class of the sklearn module if contain different functions for **performing machine learning with linear models**.

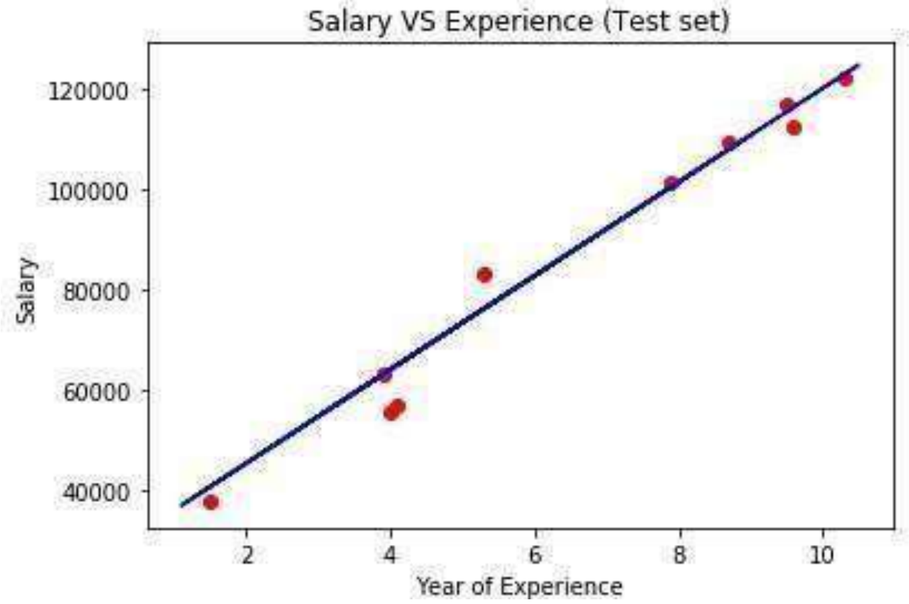
```
# Fitting Simple Linear Regression to the Training set
from sklearn.linear_model import LinearRegression
regressor = LinearRegression()
regressor.fit(x_train, y_train)
```

Program

```
# Visualizing the Training set results
viz_train = plt
viz_train.scatter(X_train, y_train, color='red')
viz_train.plot(X_train, regressor.predict(X_train), color='blue')
viz_train.title('Salary VS Experience (Training set)')
viz_train.xlabel('Year of Experience')
viz_train.ylabel('Salary')
viz_train.show()
```

Program

```
# Visualizing the Training set results
viz_test = plt
viz_test.scatter(x_test, y_test, color='red')
viz_test.plot(x_test, regressor.predict(x_test), color='blue')
viz_test.title('Salary Vs Experience (Test set)')
viz_test.xlabel('Year of Experience')
viz_test.ylabel('Salary')
viz_test.show()
```



Program

```
y_pred = regressor.predict(20)
print(y_pred)
```

The value of y_pred with **X = 5** (5 Years Experience) is 28

Colab

```
[35] import numpy as np
     import matplotlib.pyplot as plt
     import pandas as pd
```

```
df=pd.read_csv('Employee.csv')
df
```

	Exp	Sal
0	8	10
1	10	13
2	12	16

```
x=df.iloc[:,:-1]
y=df.iloc[:,1]
y
```

	Sal
0	10
1	13
2	16

dtype: int64

Colab

```
[45] from sklearn.model_selection import train_test_split
     x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=1/3,random_state=0)
     y_train
```

	Sal
1	13
0	10

dtype: int64

```
from sklearn.linear_model import LinearRegression
reg=LinearRegression()
reg.fit(x_train, y_train)
```

```
LinearRegression()
```

Colab

0s

Generated code may be subject to a license | ShrilathaKS/ML-lab

```
viz_train=plt
viz_train.scatter(x_train,y_train,color='red')
viz_train.plot(x_train,reg.predict(x_train),color='blue')
viz_train.title('Salary vs Experience(Training set)')
viz_train.xlabel('Years of Experience')
viz_train.ylabel('Salary')
viz_train.show()
```

Colab

0s

```
viz_test=plt
viz_test.scatter(x_test,y_test,color='red')
viz_test.plot(x_test,reg.predict(x_test),color='blue')
viz_test.title('Salary vs Experience(Test set)')
viz_test.xlabel('Years of Experience')
viz_test.ylabel('Salary')
viz_test.show()
```

Colab

```
y_pred=reg.predict([[20]])
print("The salary for Employee with 20 year of experience:",y_pred)
```

```
The salary for Employee with 20 year of experience: [28.]
```

Program

In conclusion, with Simple Linear Regression, we have to do 5 steps as per below:

1. Importing the dataset.
2. Splitting dataset into training set and testing set (2 dimensions of X and y per each set). Normally, the testing set should be 5% to 30% of dataset.
3. Visualize the training set and testing set to double check (you can bypass this step if you want).
4. Initializing the regression model and fitting it using training set (both X and y).
5. Let's predict!!

References

Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin, ―Learning from Data, AML Book Publishers, 2012.

P. Flach, ―Machine Learning: The art and science of algorithms that make sense of data‖, Cambridge University Press, 2012.

W3school.com