What is a Machine Learning Model?

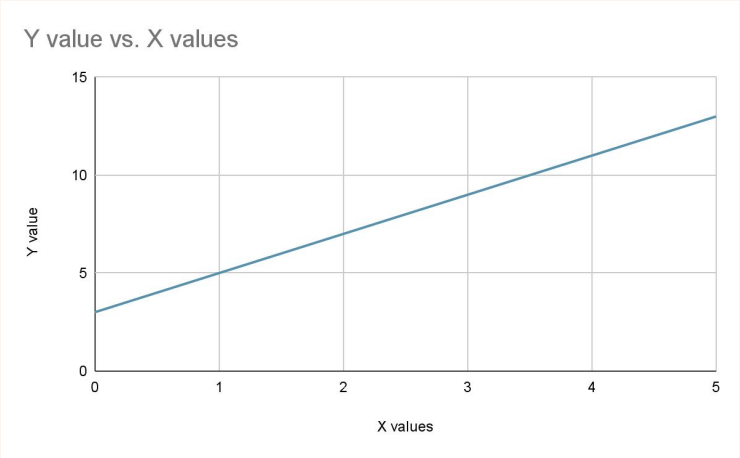
Machine Learning

Machine learning Model

Data

Machine Learning Model

x	1	2	3	4	5
y	5	7	9	11	13



$Y = mX+c$

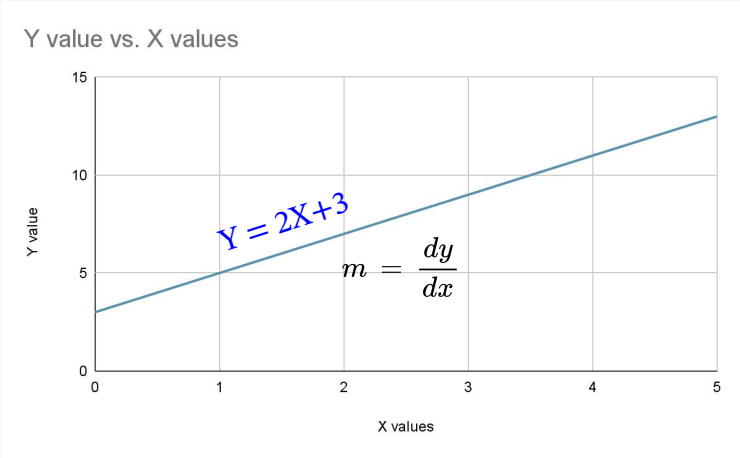
$X \rightarrow$ X value

$Y \rightarrow$ Y value

$m \rightarrow$ Slope

$c \rightarrow$ Intercept

Machine Learning Model



Inference: The above Line equation is a function that relates X and Y.

For a given value of X, we can find the corresponding value of y

Equation of a straight line: **Y = mX+c**

Find the Values of m and c:

Point p1(2,7)

Point p1(3,9)

Slope, m = $\frac{y_2 - y_1}{x_2 - x_1} = \frac{9-7}{3-2} = 2$

m = 2

Intercept, c:

Point (4,11)

Y = 2X+c ; 11 = 2(4) + c

c = 3

Machine learning Model

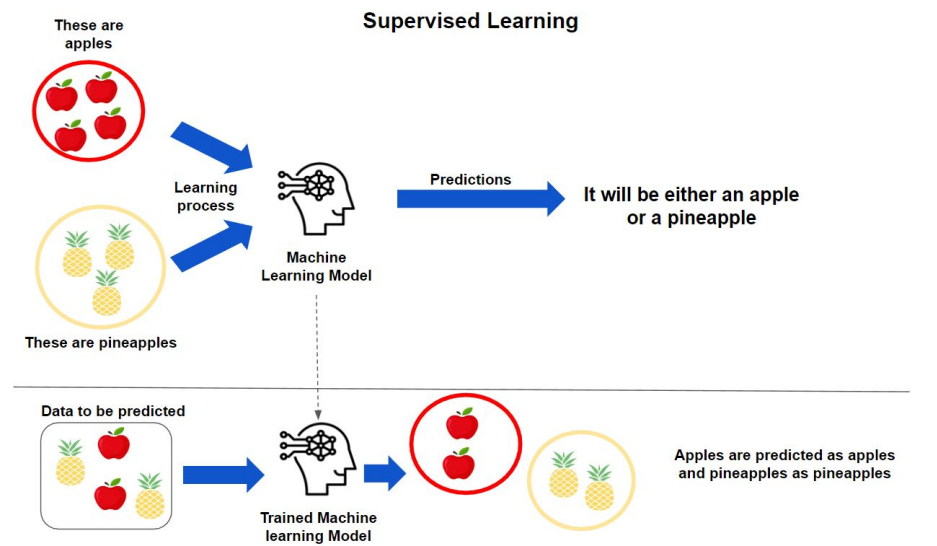
A Machine Learning Model is a function that tries to find the relationship between the **features** and the **target variable**.

It looks for patterns in the data, learns from it, and trains itself accordingly. Based on this learning, the model makes **predictions** and recognizes **patterns**.

We cannot have a Linear relationship between the variables all the time.

Supervised Learning

In **Supervised learning,** the Machine Learning algorithm learns form **Labelled Data**



Types of Supervised Learning

Supervised learning

01

Classification

Classification is about predicting a class or discrete values

Eg: male or female; Ture or False

02

Regression

Regression is about predicting a quantity or continuous values

Eg: salary; age; price

Supervised Learning Algorithms

Classification:

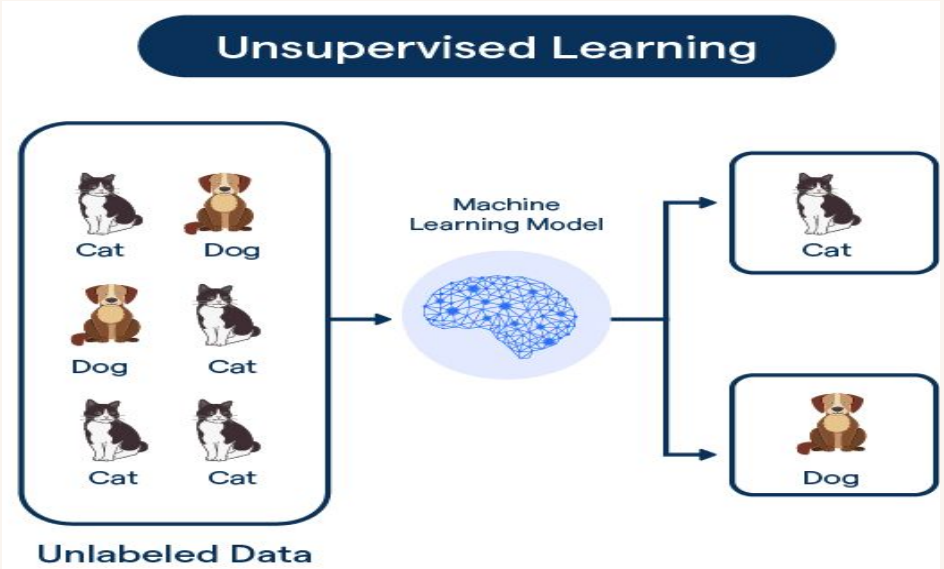
1. Logistic Regression
2. Support Vector Machine Classifier
3. Decision Tree
4. K-Nearest Neighbors
5. Random Forest
6. Naïve Bayes Classifier

Regression:

1. Linear Regression
2. Lasso Regression
3. Polynomial Regression
4. Support Vector Machine Regressor
5. Random Forest Regressor
6. Bayesian Linear Regressor

Unsupervised Learning

In **Unsupervised learning,** the Machine Learning algorithm learns form **Unlabelled Data**



Types of Unsupervised Learning

Unsupervised learning

01

Clustering

Clustering is an unsupervised task which involves grouping the similar data points

02

Association

Association is an unsupervised task that is used to find important relationship between data points

Unsupervised Learning Algorithms

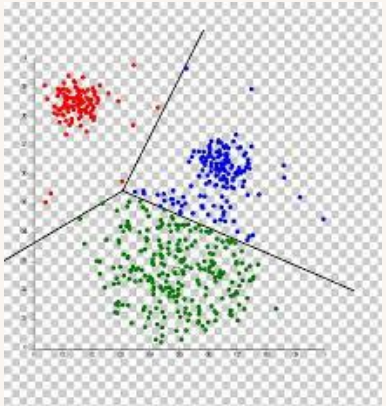
1. K-Means Clustering
2. Hierarchical Clustering
3. Principal Component Analysis (PCA)
4. Apriori
5. Eclat

How to choose the right Machine learning Model? (Model Selection)

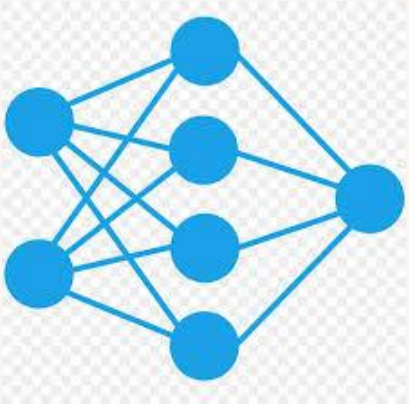
Model Selection

Model Selection in Machine Learning is the process of choosing the best suited model for a particular problem. Selecting a model depends on various factors such as the **dataset, task, and nature of the model,** among others.

Logistic Regression



K-Means Clustering



Neural Network

Model Selection

Models can be selected based on:

1) **Type of data available:**
   - **Images & Videos –** CNN
   - **Text data or Speech data –** RNN
   - **Numerical data –** SVM, Logistic Regression, Decision trees, etc.

2) **Based on the task we need to carry out:**
   - **Classification tasks –** SVM, Logistic Regression, Decision trees, etc.
   - **Regression tasks –** Linear Regression, Random Forest, Polynomial Regression,etc.
   - **Clustering tasks –** K-Means clustering, Hierarchical Clustering.

Cross Validation

- ➔ Accuracy score for SVM = 84.4%
- ➔ Accuracy score for Logistic Regression = 88%

Cross validation Implementation:

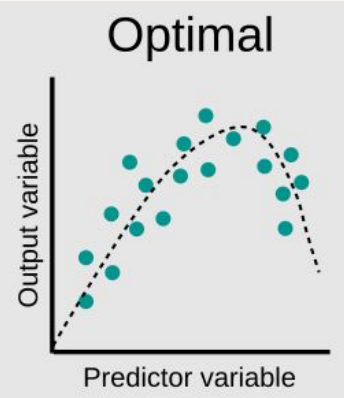
```
from sklearn import datasets, linear_model
from sklearn.model_selection import cross_val_score
diabetes = datasets.load_diabetes()
X = diabetes.data[:150]
y = diabetes.target[:150]
lasso = linear_model.Lasso()
print(cross_val_score(lasso, X, y, cv=3))
```

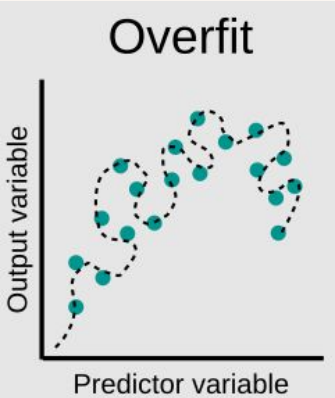
```
[0.3315057  0.08022103 0.03531816]
```

Overfitting & Underfitting in Machine learning

Overfitting

Overfitting refers to a model that models the training data too well. Overfitting happens when a model learns detail and noise in the dataset to the extent that it negatively impacts the performance of the model





Sign that the model has Overfitted: High Training Data Accuracy & very Low Test Data Accuracy

Overfitting

Causes for Overfitting:

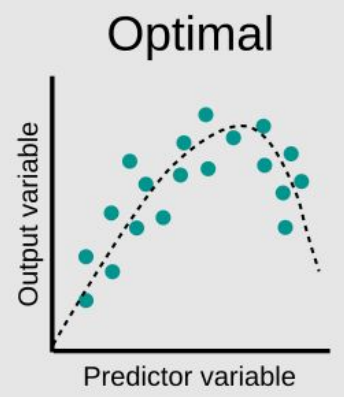
1. Less Data
2. Increased Complexity of the model
3. More number of layers in Neural Network

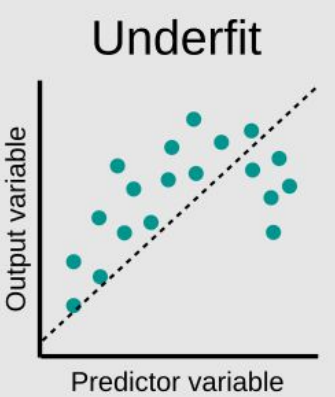
Preventing Overfitting by:

1. Using more data
2. Reduce the number of layers in the Neural network
3. Early Stopping
4. Bias – Variance Tradeoff
5. Use Dropouts

Underfitting

Underfitting happens when the model **does not learn enough** from the data. Underfitting occurs when a machine learning model cannot capture the underlying trend of the data





Sign that the model has Underfitting: Very low training data Accuracy

Underfitting

Causes for Underfitting:

1. Choosing a wrong model
2. Less complexity of the model
3. Less variance but high bias

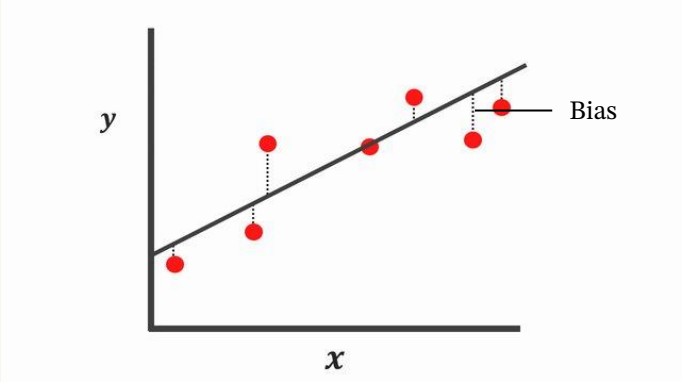
Preventing Underfitting by:

1. Choosing the correct model appropriate for the problem
2. Increasing the complexing of the model
3. More number of parameters to the model

Bias -Variance Tradeoff In Machine learning

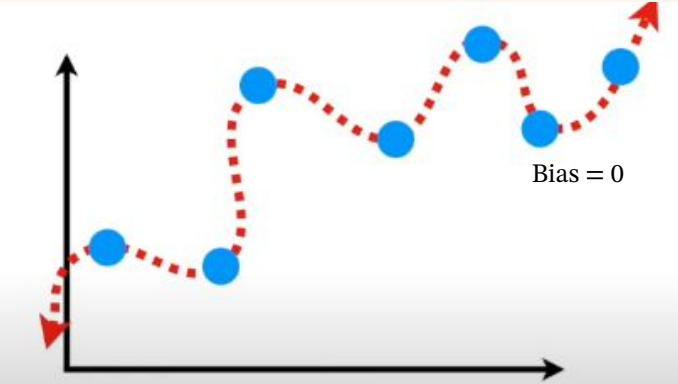
Bias - Variance Tradeoff

Bias: Bias is the difference between the average prediction of our model and the correct value which we are trying to predict.



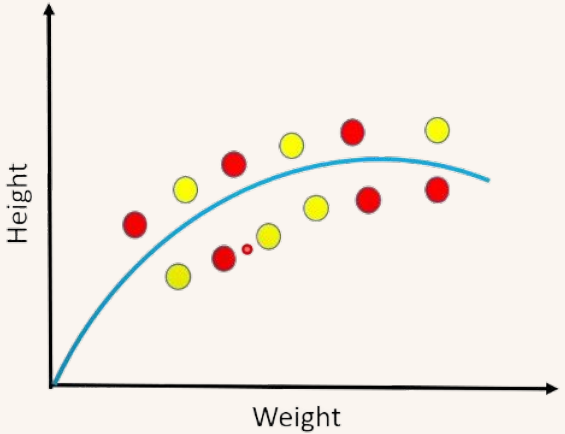
Bias - Variance Tradeoff

Variance: Variance is the amount that the estimate of the target function will change if different training data was used.



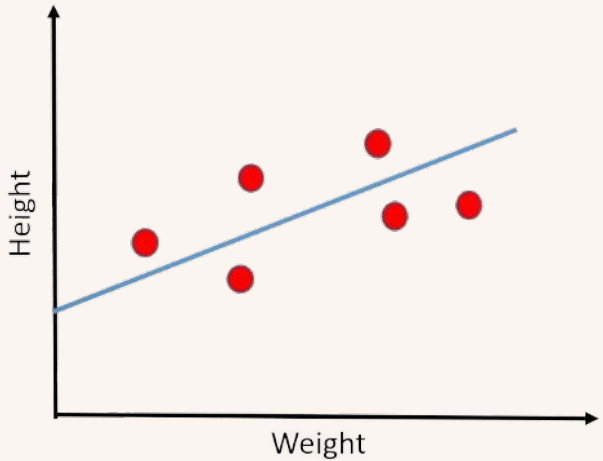
Bias - Variance Tradeoff

Problem statement: Identify an appropriate model to predict the Height of a person, When their weight is given.

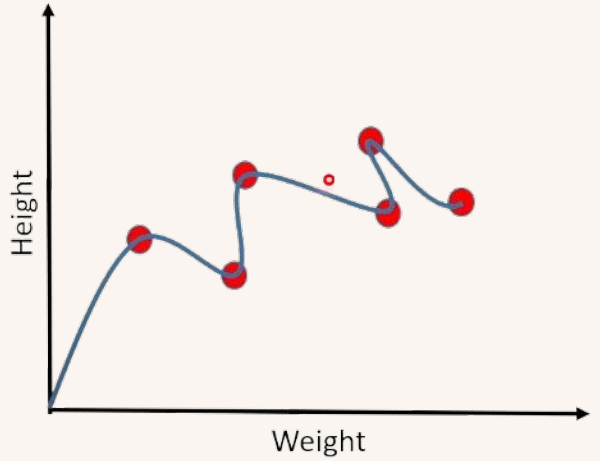


Underfitting & Overfitting

(Plot on training data)



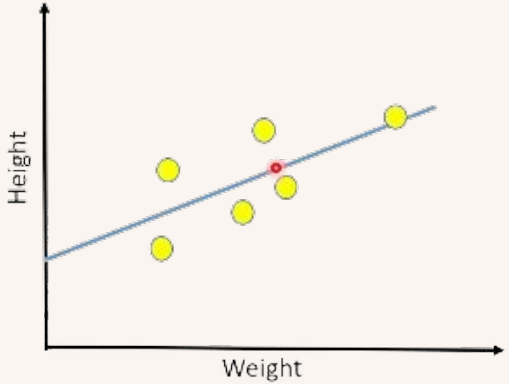
(i) Underfitting



(i) Overfitting

Bias - Variance Tradeoff

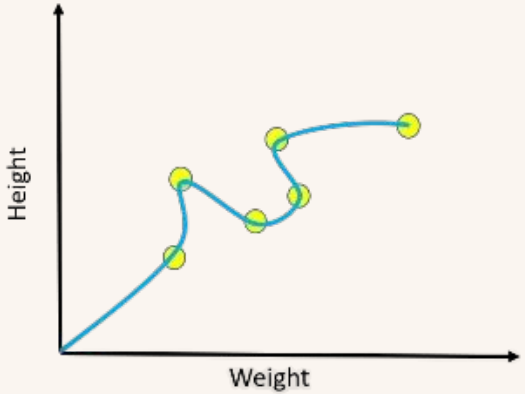
(Testing with different data)



(i) Underfitting

Inference: a. High Bias

b. Low Variance

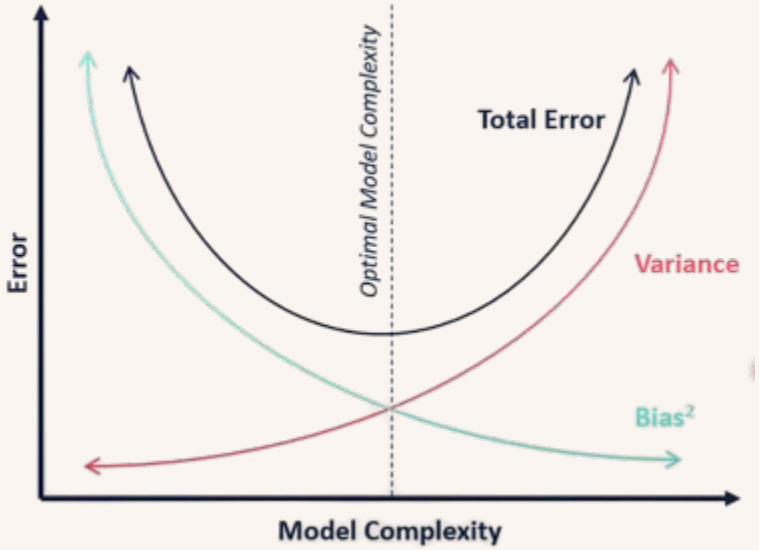


(i) Overfitting

Inference: a. Low Bias

b. High Variance

Bias - Variance Tradeoff



Bias - Variance Tradeoff

Techniques to have better Bias – Variance Tradeoff:

1. **Good Model Selection**
2. **Regularization**
3. **Dimensionality Reduction**
4. **Ensemble methods**

Loss Function in Machine Learning

Loss Function

Loss function measures how far an estimated value is from its true value.

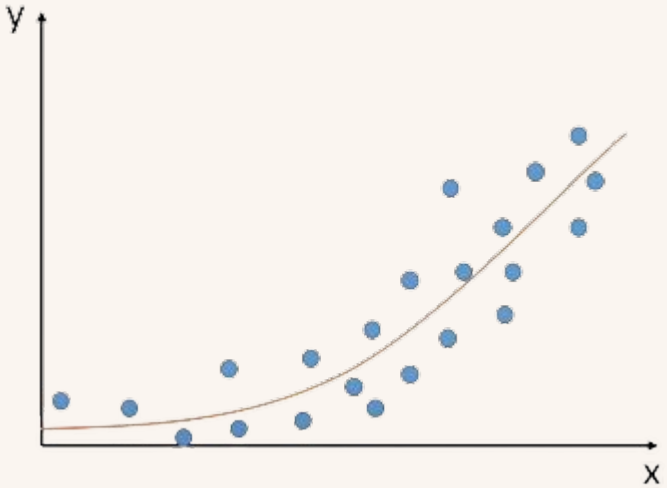
It is helpful to determine which model performs better & which parameters are better.

$$Loss = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i)$$

Types of Loss Function:

- Cross Entropy Loss
- Squared Error Loss
- KL Divergence

Loss Function



$$y = 0.0000003x^3 + 0.0002x^2 + 0.01x + 0.025$$

Degree 3 Polynomial

Loss Function

x	y	y_1	y_2	y_3
0.30	0.35	0.38	0.39	0.41
0.45	0.48	0.45	0.47	0.56
0.50	0.55	0.59	0.58	0.63
0.55	0.63	0.65	0.69	0.70
0.66	0.72	0.75	0.78	0.78

$$Loss = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i)$$

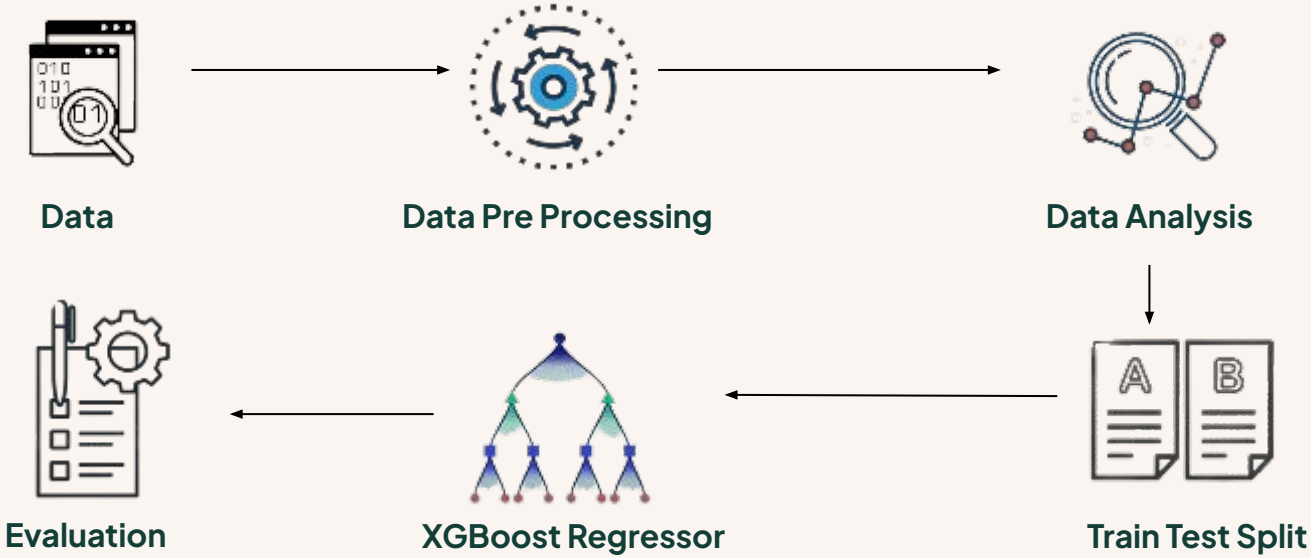
$$loss_1 = \left[(0.35 - 0.38)^2 + (0.48 - 0.45)^2 + (0.55 - 0.59)^2 + (0.63 - 0.65)^2 + (0.72 - 0.75)^2\right]/5$$

$$loss_1 = 0.173$$

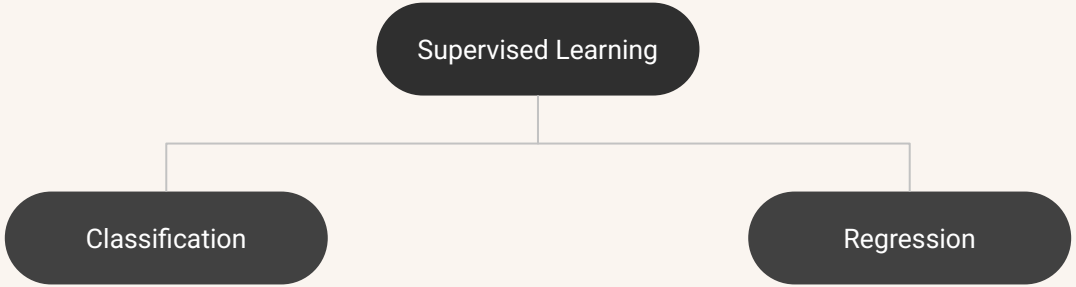
Low Loss Value → High Accuracy

Model Evaluation in Machine Learning

Workflow of a ML project



Workflow of a ML project



Classification is about predicting a class or discrete values

Eg: male or female; Ture or False

Evaluation metric for classification: **Accuracy Score**

Regression is about predicting a quantity or continuous values

Eg: salary; age; price

Evaluation metric for Regression: **Mean Absolute Error**

Accuracy Score

In Classification, **Accuracy Score** is the ratio of **number of correct predictions** to the **total number of input data points.**

$$AccuracyScore = \frac{Number\ of\ correct\ predictions}{Total\ Number\ of\ data\ points} * 100\%$$

Number of correct predictions = 128
Total Number of data points = 150

Accuracy Score = 85.3%c

```
from sklearn.metrics import accuracy_score
```

Mean Squared Error

Mean Squared Error measures the average of the squares of the errors, that is, the average squared difference between the estimated values and the actual value.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i)^2$$

Actual Value (Y, 140 mg/dL)

Predicted Value (Y = 160 mg/dL)

```
from sklearn.metrics import mean_squared_error
```