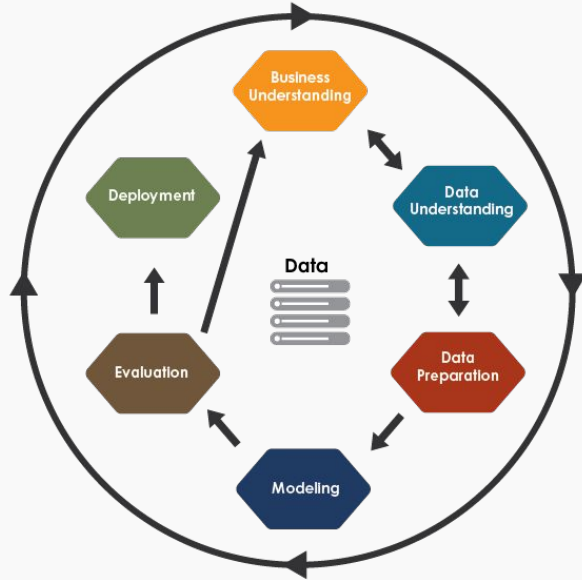


SIC Batch 5

Week 5 - Intro to Machine Learning for IoT

Machine Learning Cycle



Source: datascience-pm.com

Business understanding – What does the business need?

Data understanding – What data do we have / need? Is it clean?

Data preparation – How do we organize the data for modeling?

Modeling – What modeling techniques should we apply?

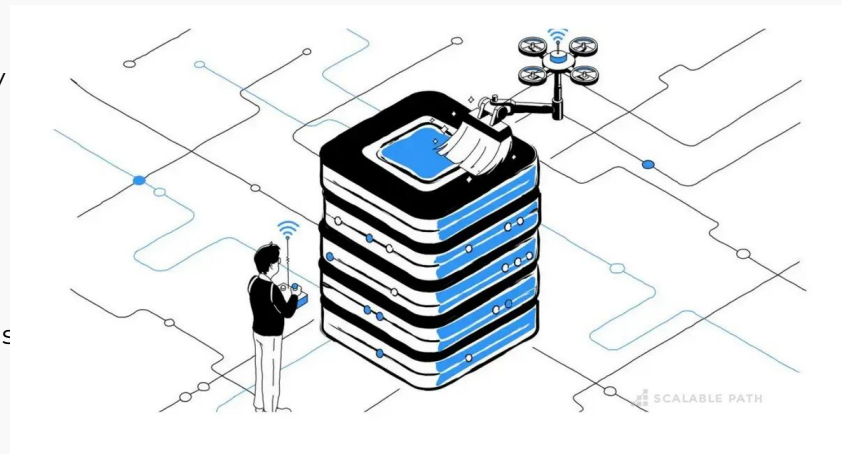
Evaluation – Which model best meets the business objectives?

Deployment – How do stakeholders access the results?

Data Preparation

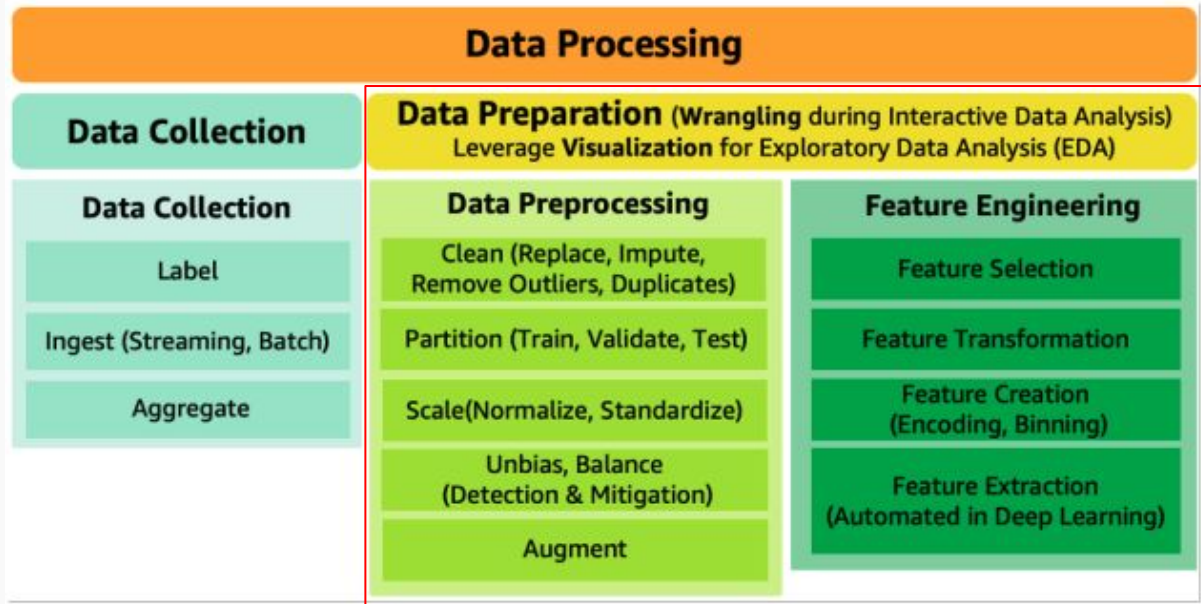
Building good machine learning models requires high-quality training data. This data needs to be prepared through preprocessing and feature engineering to optimize learning and generalization.

Exploratory Data Analysis (EDA) is crucial to uncover hidden patterns in the data, often not readily apparent in tables. Visualization tools and data wrangling tools can be used for fast and interactive analysis to understand the data better. Leveraging no-code/low-code automation and visual features can further improve productivity and reduce costs associated with data analysis for model building.



Source: scalablepath.com

Data Processing



Source: aws

Data Preprocessing

Clean (Replace, Impute,
Remove Outliers, Duplicates)

Partition (Train, Validate, Test)

Scale(Normalize, Standardize)

Unbias, Balance
(Detection & Mitigation)

Augment

Source: aws

Cleaning:

Remove outliers and duplicates & Replace missing or inaccurate data (minimizing bias).

Partitioning:

Split data into training, validation, and testing sets (prevent overfitting and ensure accurate evaluation).

Scaling:

Normalize or standardize numeric data for consistent feature importance and handling of outliers.

Unbiasing and Balancing:

Detect and mitigate biases in data or algorithms to avoid unfair predictions across different groups.

Augmentation:

Artificially increase data volume by creating new data from existing data to improve model performance and reduce overfitting.

Feature Engineering

Feature Selection

Feature Transformation

Feature Creation
(Encoding, Binning)

Feature Extraction
(Automated in Deep Learning)

Source: aws

Feature Creation:

Build new features from existing data to improve predictions (e.g., one-hot encoding, binning, splitting, calculated features).

Feature Transformation & Imputation:

Address missing or invalid features (e.g., forming Cartesian products, non-linear transformations, creating domain-specific features).

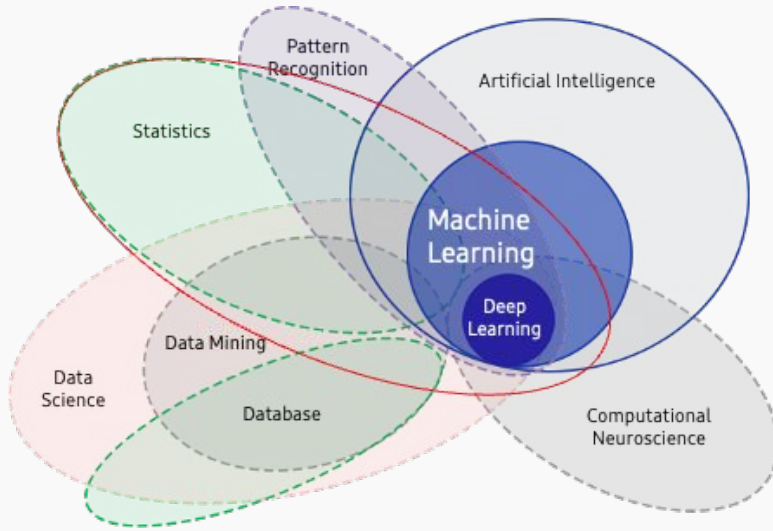
Feature Extraction (Dimensionality Reduction):

Reduce data size using techniques like PCA, ICA, LDA to save memory and processing power while preserving key information.

Feature Selection:

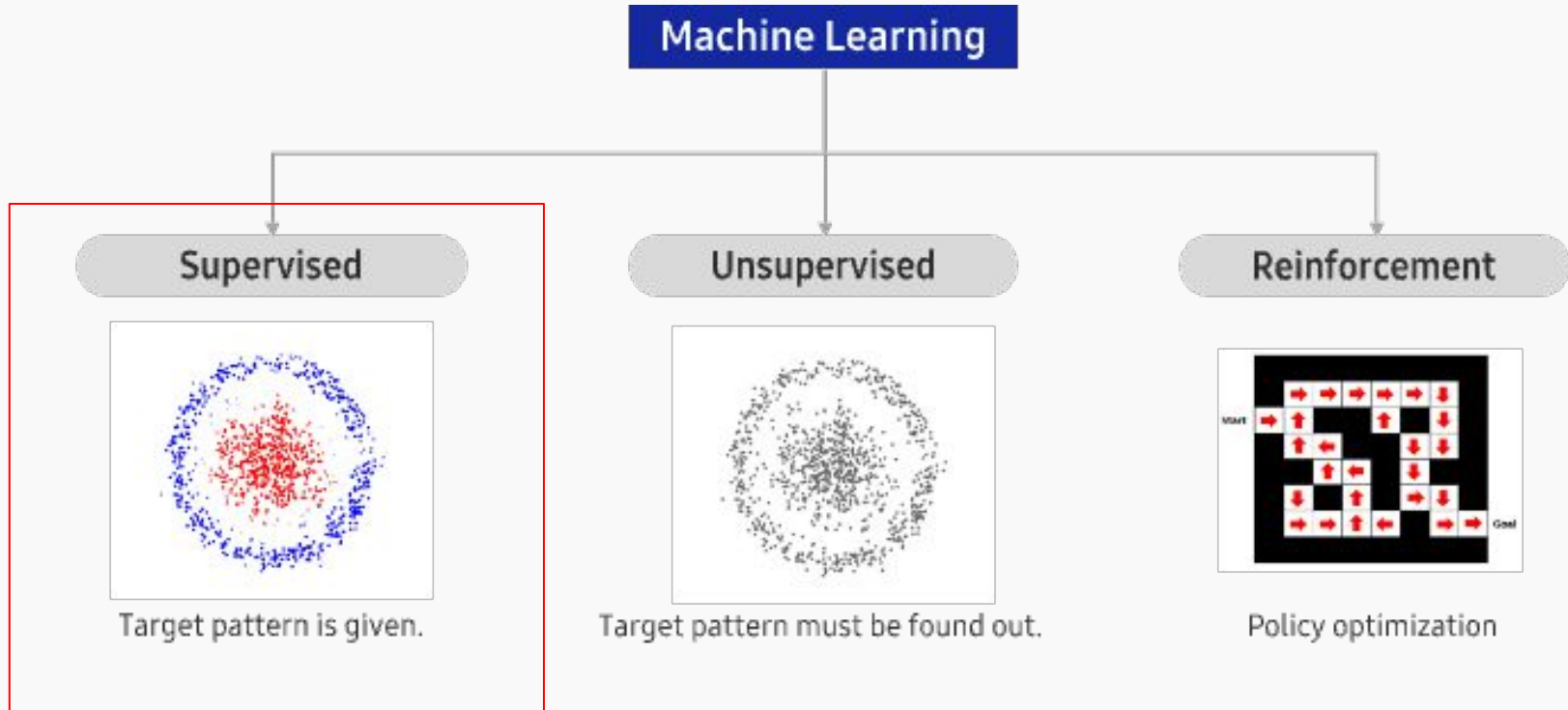
Choose the most relevant features for model training based on importance scores and correlations, reducing complexity and potentially improving model performance.

What is Machine Learning

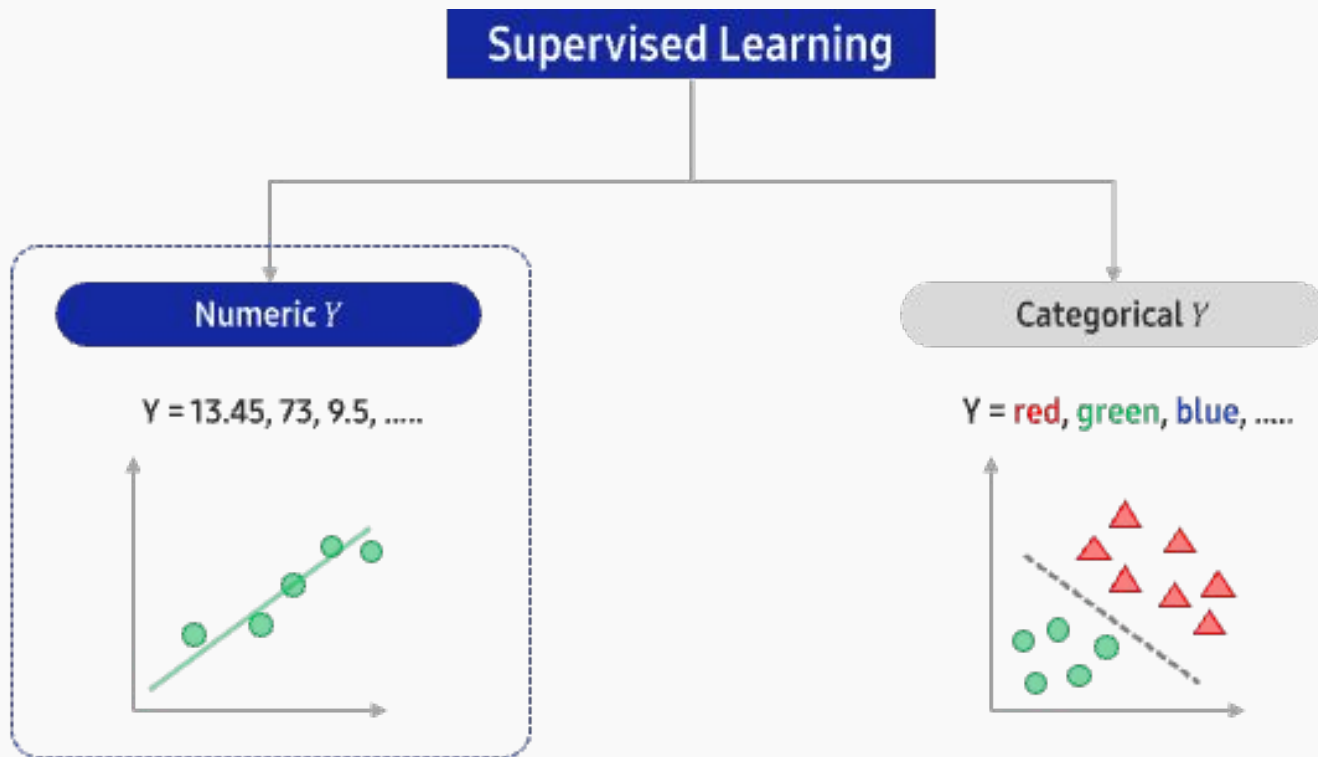


“Machine learning is the science of getting computers to act without being explicitly programmed.” — Stanford University

Types of Machine Learning



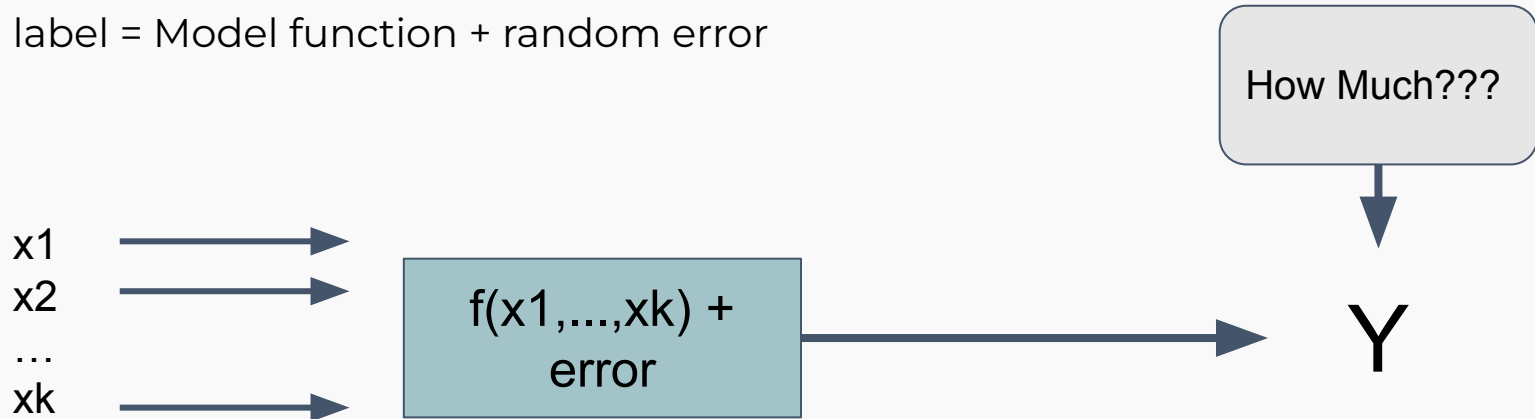
Type of Supervised Learning



Regression

Regression is a supervised machine learning technique which is used to predict continuous values. The ultimate goal of the regression algorithm is to plot a best-fit line or a curve between the data.

label = Model function + random error



Linear Regression :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon$$

Regression House Price Example

Num. of Bed	Num. of Room	...	Garage	Pool	House Price
4	10		yes	no	1000M
2	4		yes	no	500M
3	6		no	yes	120M
...

Houses with known price

We are interested to predict house with unknown price using the available feature

or

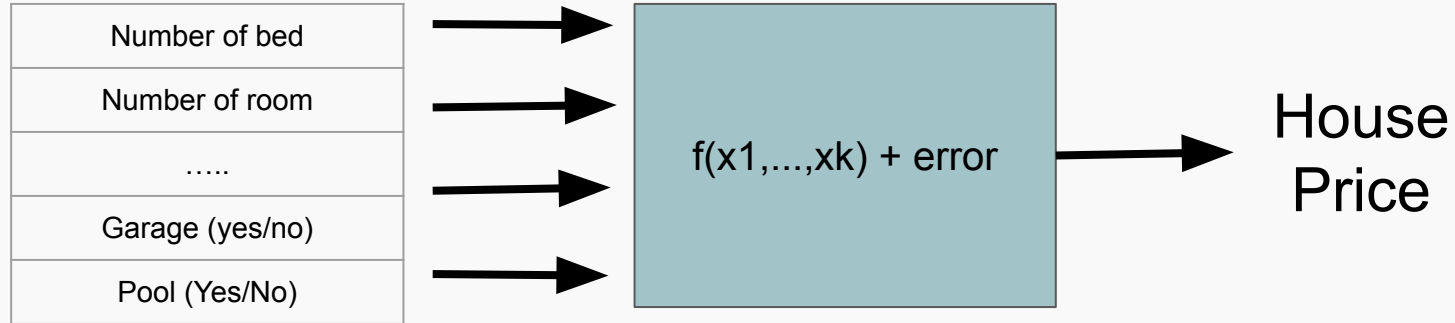
We are interested in analyzing the house price based on its characteristic

Houses characteristics

Num. of Bed	Num. of Room	...	Garage	Pool	House Price
4	7		yes	yes	???
2	5		no	no	???
...

Houses with unknown price

Regression House Price Example



Purpose :
Minimize Overpricing or Underpricing
Phenomenon

Value :
Pricing Strategy

Regression Use Case

Sales
Forecasting

Customer
Satisfaction

Price
Estimation

Employment
Income

Car CO2
Emission

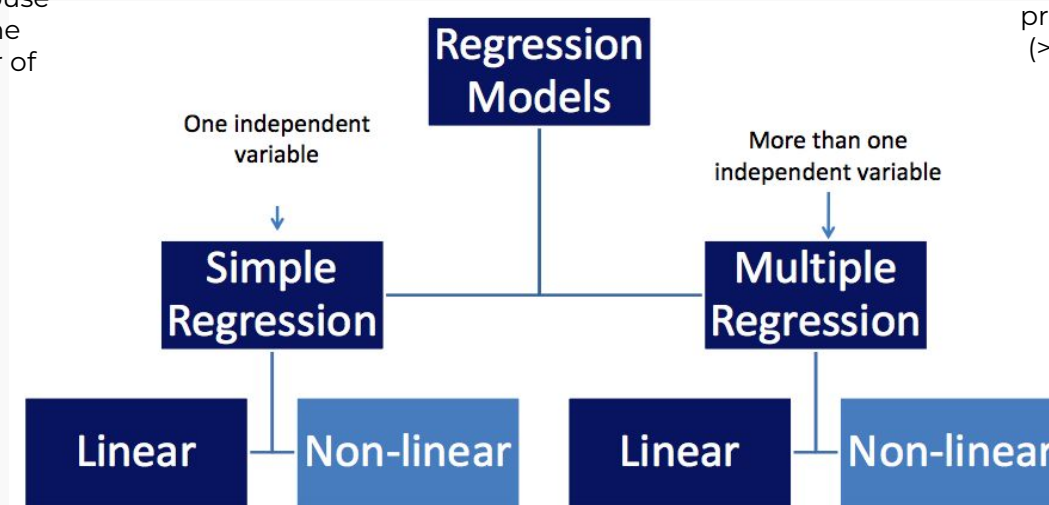
Types of Regression

SIMPLE:

Predict or analyze house price using only one features. ex number of room

MULTIPLE:

Predict or analyze house price using many features (>2). ex number of room, number of bed, etc



Linear Regression

Simple Linear Regression

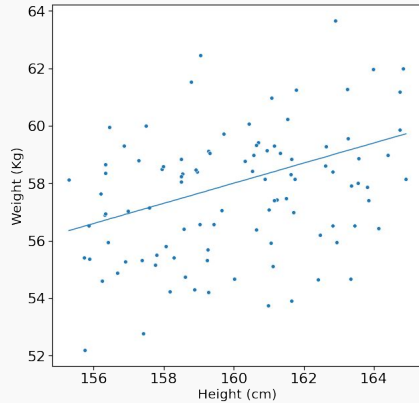
$$Y = \beta_0 + \beta_1 X + \epsilon$$

Multiple Linear Regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \epsilon$$

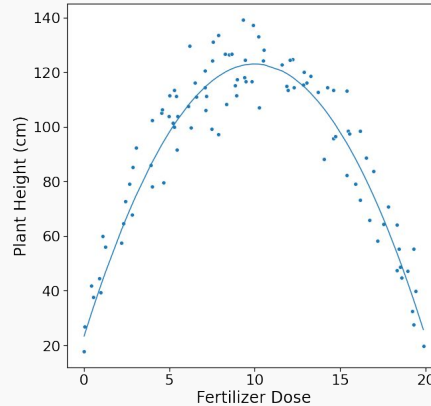
Linear vs Non-Linear Relationship

Ex. height and weight



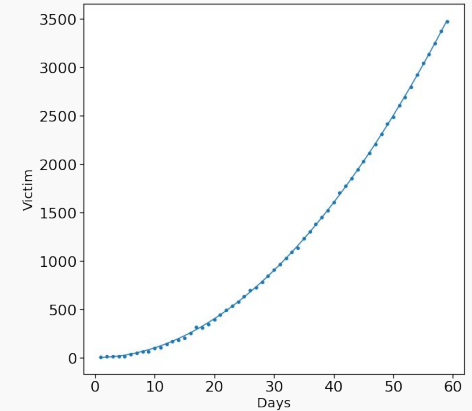
Linear : $y = 2 + 0.35x$

Ex. fertilizer dose and plant height



Non Linear and Non
Monotone

Ex. daily case of COVID-19



Non Linear and
Monotone

Non-Linear Equation Examples

Multiplikatif

$$Y = \beta_0 x^{\beta_1} \varepsilon$$

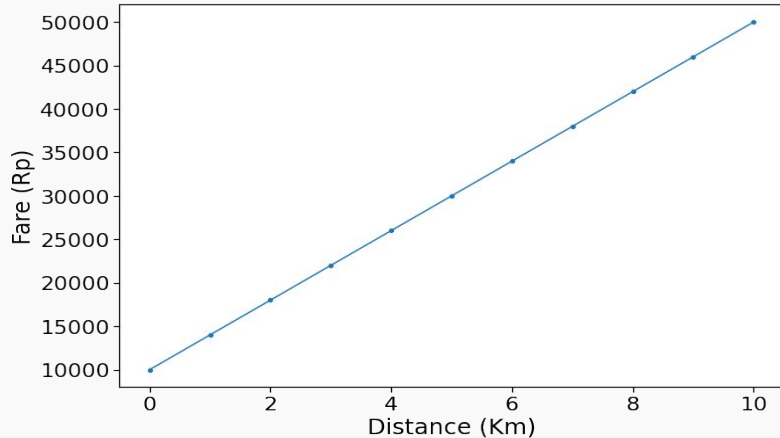
Exponential

$$Y = \beta_0 e^{\beta_1 x} \varepsilon$$

Reciprocal

$$\frac{1}{\beta_0 + \beta_1 x + \varepsilon}$$

Linear Equation : Taxi Fare (Y) vs Distance (X)



General Linear Equation:

$$Y = a + bx$$

Taxi Fare Linear Equation:

$$Y = 10000 + 4000x$$

Interpretation :

- Intercept $a = 10000$: This is interpreted as door open rates, when the customer get out of the taxi and the taxi has not been moving at all ($x = 0$ Km) the customer must pay Rp. 10,000
- Slope $b = 4000$: For each 1 km the fare will increase Rp. 4,000

Simple Linear Regression Model

- Only one independent variable
- Linear in parameters: linear equation is formed between dependent variable and regression parameters

The diagram illustrates the Simple Linear Regression Model equation, $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, set against a dark blue background. The equation is written in white text. The term ε_i is circled in red. Four yellow arrows point from descriptive labels to the corresponding parts of the equation: from 'Population Y-Intercept' to β_0 , from 'Population Slope' to β_1 , from 'Random Error' to ε_i , and from 'Dependent (Response)' to Y_i . Additionally, the label 'Independent (Explanatory)' is positioned below X_i .

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Population Y-Intercept

Population Slope

Random Error

Dependent (Response)

Independent (Explanatory)

Regression Model Performance

In order to get accurate prediction,

We can measure a model performance using :

- MSE
- RMSE
- R2

$$MSE = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n}}$$

Residuals = Real - Prediction

$$R^2 = 1 - \frac{SSE}{SST}$$

MSE & RMSE

- MSE and RMSE measure how accurate the prediction result
- we want MSE and RMSE as small as possible
- MSE is the variance of residuals while RMSE is the standard deviation
- MSE measure the spread of the residuals.

$$MSE = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n}}$$

MSE and RMSE Example

Floor Area (m2)	House Price (IDR) in Millions	Predicted House Price (IDR) in Millions	Residuals
1400	245	252.0	-7.0
1600	312	274.0	38.0
1700	279	285.0	-6.0
1875	308	304.0	4.0
1100	199	219.0	-20.0
1550	219	268.0	-49.0
2350	405	356.0	49.0
2450	324	367.0	-43.0
1425	319	255.0	64.0
1700	255	285.0	-30.0

MSE = 1359.2

RMSE = 36.867

$$MSE = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n}$$

$$MSE = \frac{(-7)^2 + 38^2 + \dots + (-30)^2}{10}$$

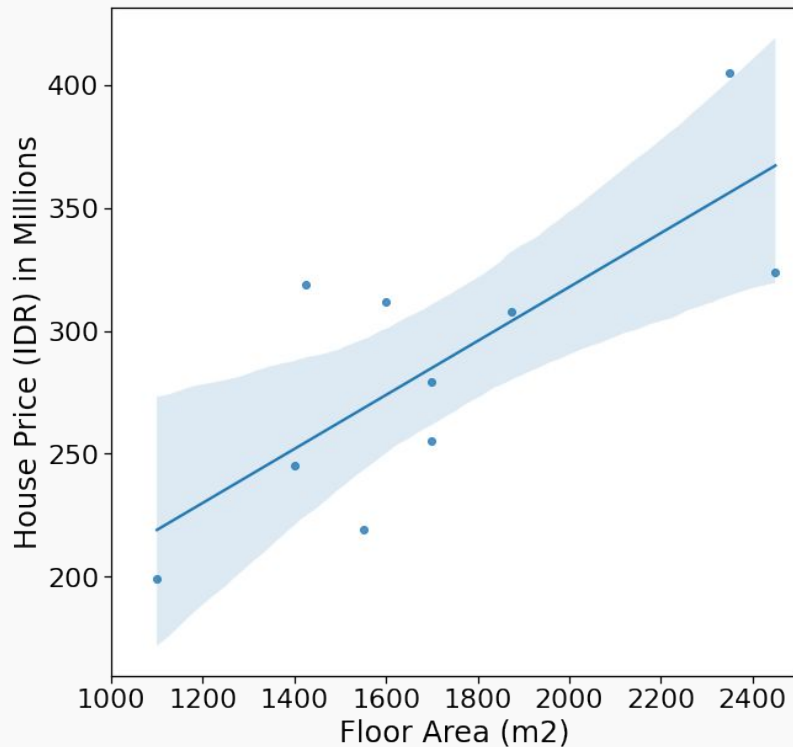
$$RMSE = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n}}$$

$$RMSE = \sqrt{\frac{(-7)^2 + 38^2 + \dots + (-30)^2}{10}}$$

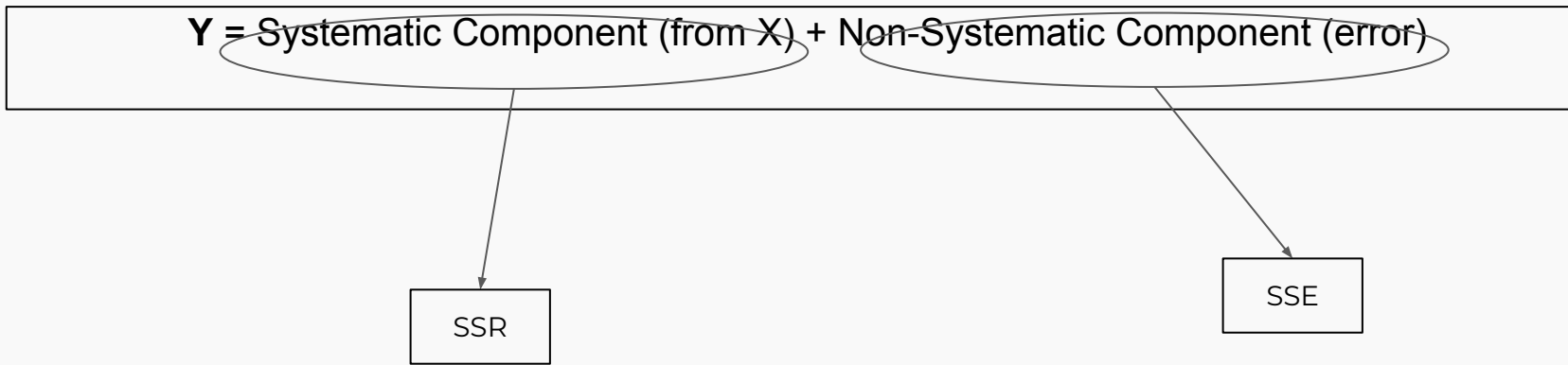
R-Square

The goodness of fit of regression equation can be measured using Coefficient Determination (R-Square)

- Coefficient Determination measure how well the regression line fits the data
- Coefficient Determination lies below 1. it's also standardize version of MSE.
- The closer to 1 the better the regression line in representing data.
- interpretation : Percentage of the variation that can be explained by the regression equation



R-Square



$$SST = SSR + SSE$$

- SST : Sum Square Total
- SSR : Sum Square Regression
- SSE : Sum Square Error

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

R-Square Example

$$SSE = MSE * n = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

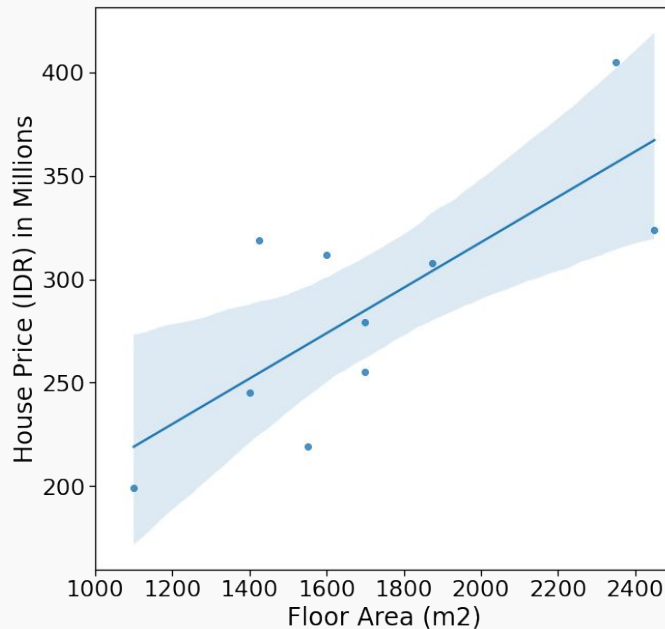
$$SST = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

$$R^2 = 1 - \frac{1359.2 * 10}{(245 - 286.5)^2 + (312 - 286.5)^2 + \dots + (255 - 286.5)^2}$$


MSE = 1359.2

R-Square = 58.30%



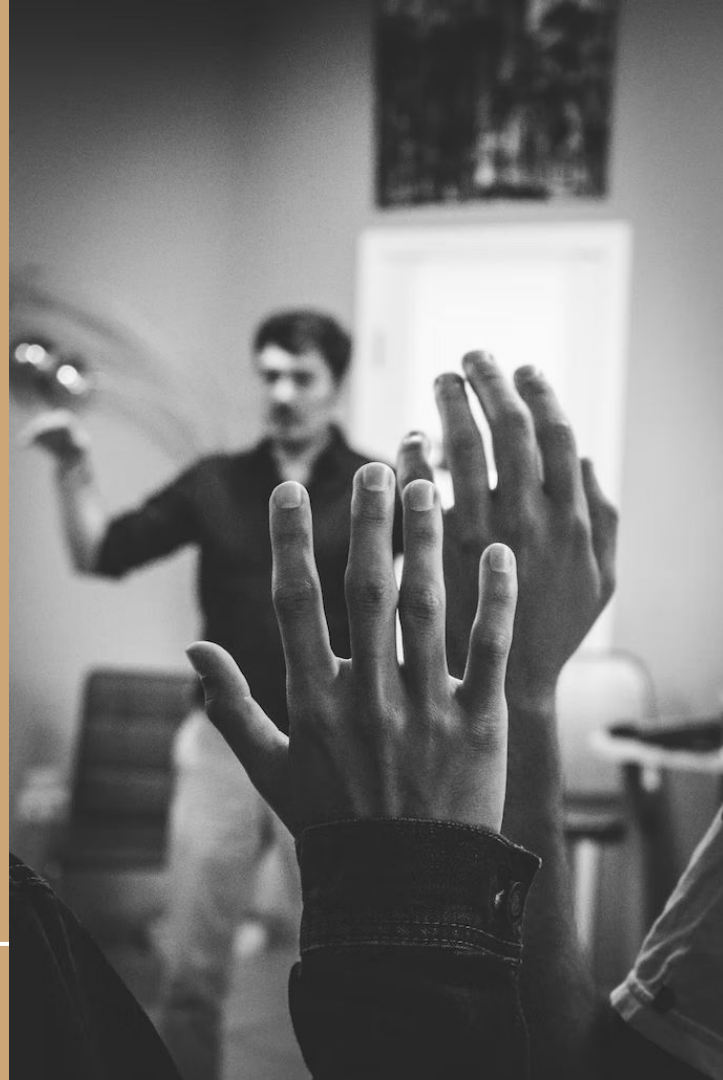
Simple Linear Regression

Variable x and y has Linear relationship	Assumption of the world
$y = \beta_0 + \beta_1 x + \varepsilon,$ Minimize SSE	Fitting a model
Is x really related to y ? Is β_1 statistically significant?	Validating the model
Predict y for a given x .	Using a model



Hands-On

[Collab Link](#)



Use Case : House Price

Harga Rumah (Rp.juta) (y)	Luas Lantai (m2) (x)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

- $Y \rightarrow$ House Price (IDR in millions)
 $x \rightarrow$ floor area (m2)
- We want to know how floor area can affect house price ?
- we want to know whether the effect of floor area to house price is significant ?
- How accurate if we use floor area only to predict house price using simple linear regression ?

[Sample Collab Link](#)

Multiple Linear Regression

- Several independent variables may influence the change in dependent variable we are trying to study
- Linear in parameters: linear equation is formed between response variable and regression parameters

The diagram shows the Multiple Linear Regression equation:
$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + \varepsilon_i$$
 with the following labels and arrows:

- Population Y-intercept** points to β_0 .
- Population slopes** points to the β coefficients ($\beta_1, \beta_2, \dots, \beta_k$).
- Random error** points to ε_i .
- Dependent (response) variable** points to Y_i .
- Independent (explanatory) variables** points to the X variables ($X_{1i}, X_{2i}, \dots, X_{ki}$).

Plant Height (Y) vs Fertilizer Dose and Temperature

Model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon$$

Equation:

$$Y = 90 + 2x_1 + 0.3x_2$$

Y = Plant Height

x₁ = Fertilizer Dose (range 0-10 Liter)

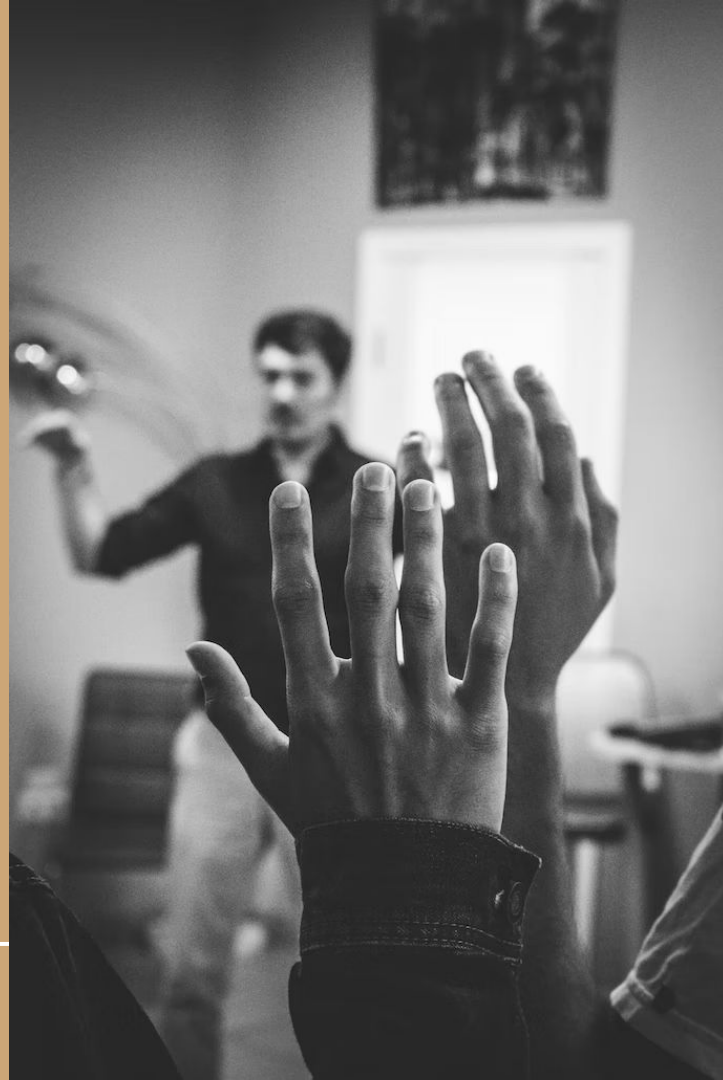
x₂ = Temperatur (C) (range 30 C - 35 C)

Interpretation :

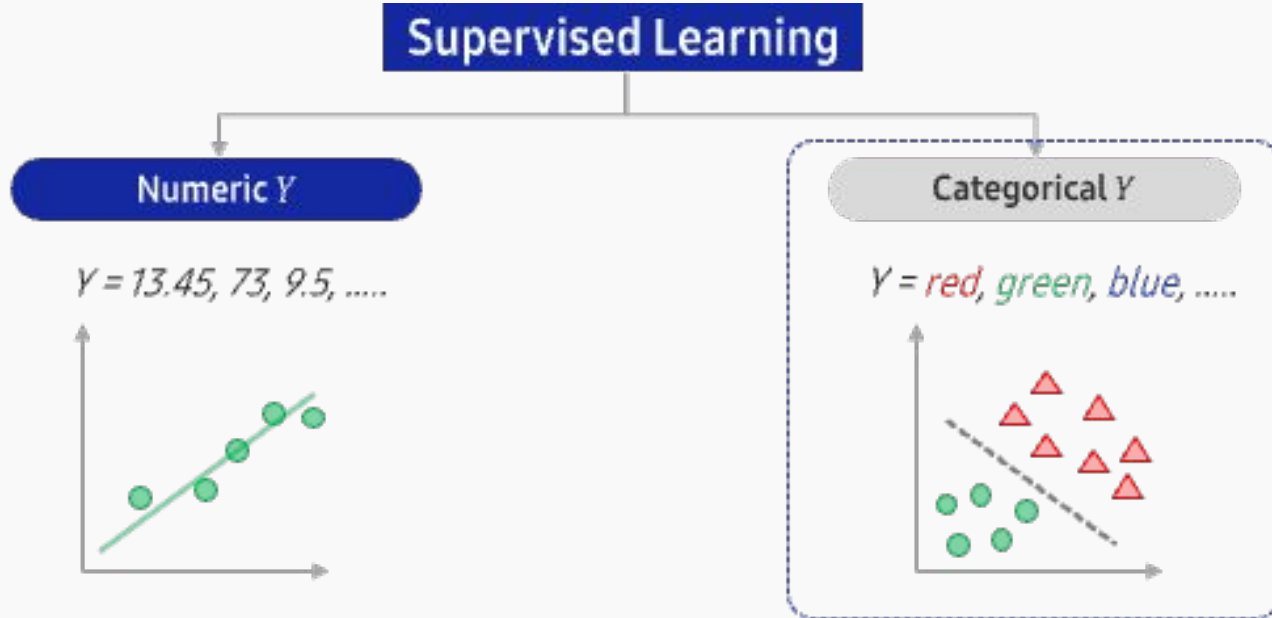
- B₀ = 90 : When we don't give any dose of fertilizer to the plant and the temperature is 30 C the plant will grow **about** 90 + (0.3*30) cm = 90.9 cm
- B₁ = 2 : When fertilizer dose increase 1 Liter the plant height will increase **about** 2 cm
*This interpretation is only recommended when we give dose between 0 and 10 Liter and no changes in another variable (Temperature)
- B₂ = 0.3 : When temperature increase 1 C the plant height will increase **about** 0.3 cm
*This interpretation is only recommended when we give dose between 30 C and 35 C and no changes in another variable (Fertilizer Dose)

Hands-On

[Collab Link](#)



Type of Supervised Learning



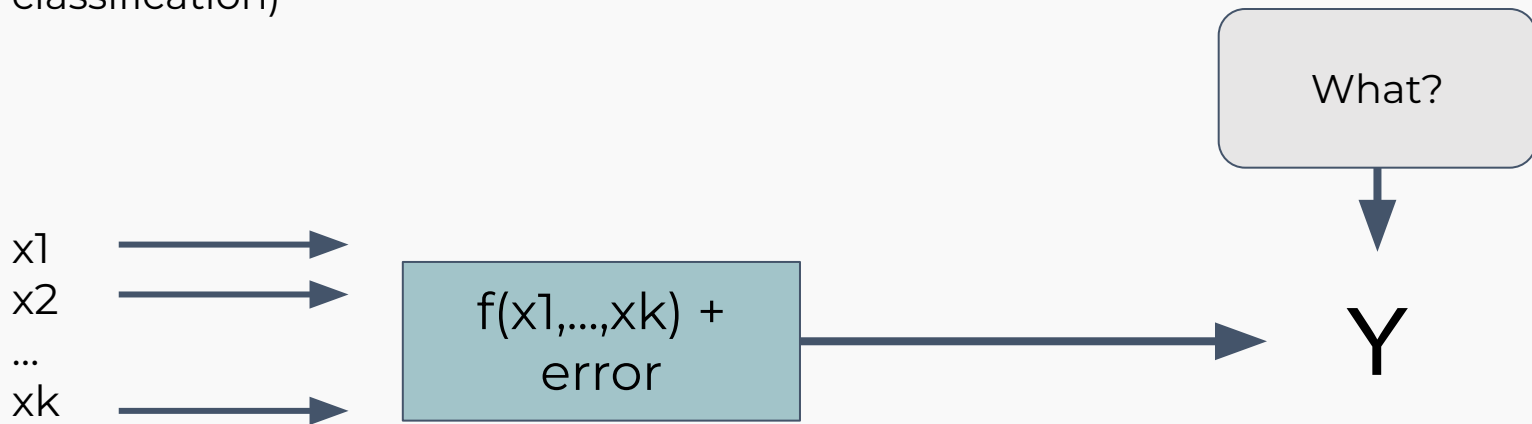
Classficiations

Response variable = Model function + random error

$$Y = f(x_1, x_2, \dots, x_k) + e$$

Y categorical - 2 categories (binary classification)

Y categorical - more than 2 categories (multiclass classification)



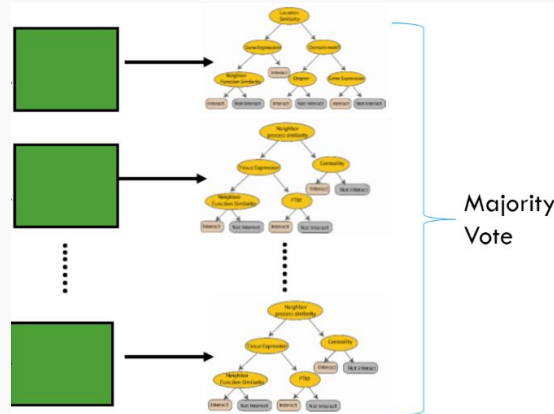
Classification Method

Logistic Regression:

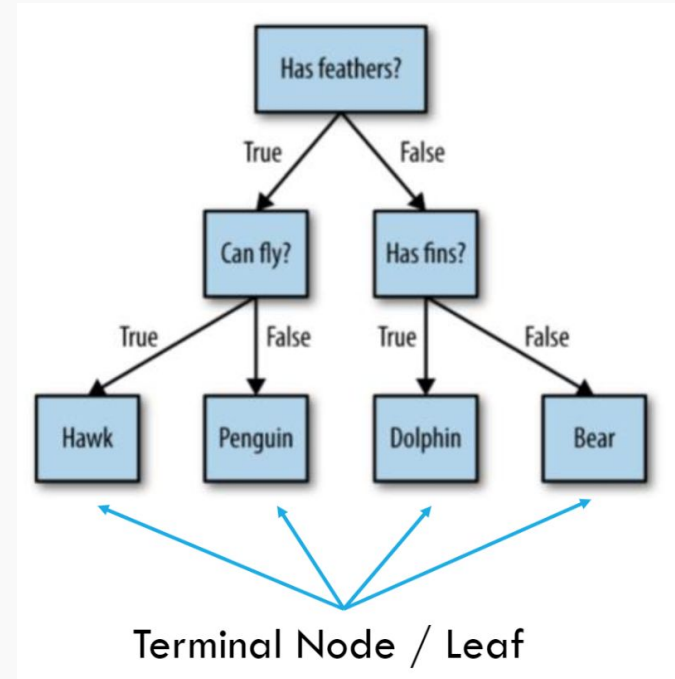
$$P(Y = 1) = \frac{\exp(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)}$$

Ensemble Method:

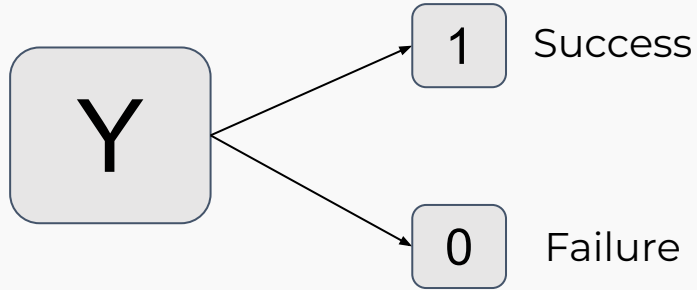
Note:
Other models:
Discriminant Analysis,
K-Nearest Neighbour
(KNN),
Support Vector
Machine (SVM),
Ensemble – Bagging,
Random Forest,
Boosting, etc



Decision Tree Classifier:



Binary Logistic Regression



Remember that
binary logistic
regression model
the success
rate/probability

Has more interest in
success event

Case	1	0
Credit scoring	Bad	Good
Churn Analysis	Turn Over	Stay
Propensity	Buy	Not Buy

What is Binary Logistic Regression

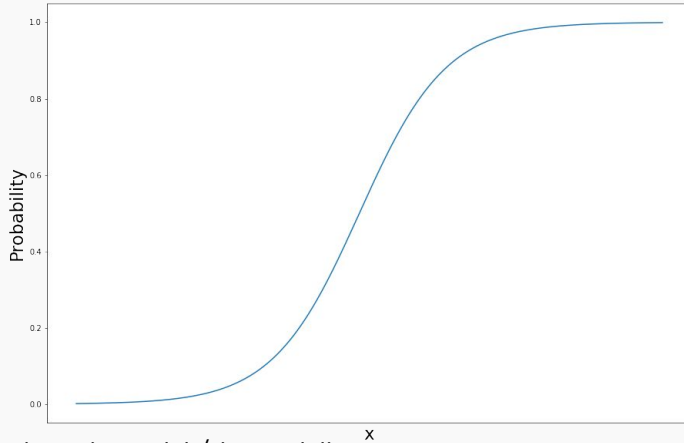
$$P(Y = 1) = \frac{\exp(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)}$$

* $\exp(B_0 + B_1 x_1 + \dots + B_k x_k)$ is approximately equal to $2.71^{(B_0 + B_1 x_1 + \dots + B_k x_k)}$

- Probability to success $P(Y = 1)$ and Probability to fail $P(Y = 0) = 1 - P(Y = 1)$
- Another notation Success (+) Failed (-)
- odds = $\exp(B_0 + B_1 x_1 + \dots + B_k x_k)$, ratio between probability to success and probability to fail
- $B_0 B_1 B_2 \dots B_k$, Regression Parameter
- $x_1 x_2 \dots x_k$, Features/Independent Variable

Sigmoid Curve

$b > 0$, success rate increase
when X increase

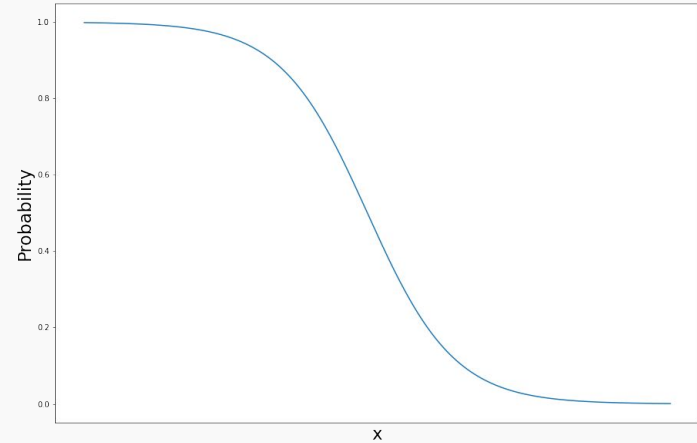


$$P(Y = 1) = \text{odd} / (1 + \text{odd}),$$

with

- $0 < P(Y = 1) < 1$
- Y = dependent variable, success ($Y = 1$) failure ($Y = 0$)
- $\text{odd} = \exp(a + bx)$
- x = independent variable

$b < 0$, success rate decrease
when X increase



Kenapa pakai Logistic Regression ?

- Instead of linear regression, logistic regression is more suitable when the response variable are categorical
 - Linear regression was designed for numerical variable
 - Linear regression can gives meaningless out-of-range prediction
 - Linear regression gives wrong p-value when Y categorical due to violation in normality assumption and equal variance assumption (homoscedasticity)
- Logistic Regression has high interpretability, remember that purpose of the modeling is not always about prediction.

Odds Ratio

- Odds-ratio is used to interpret logistic regression
- Odds-ratio indicate how likely a successful event is to occur in one condition compared to other conditions

The example you provided uses odds ratios to compare loan defaults between men and women. Let's break it down:

Odds-Ratio = $\text{Odd}(\text{Male}) / \text{Odd}(\text{Female}) = 3$

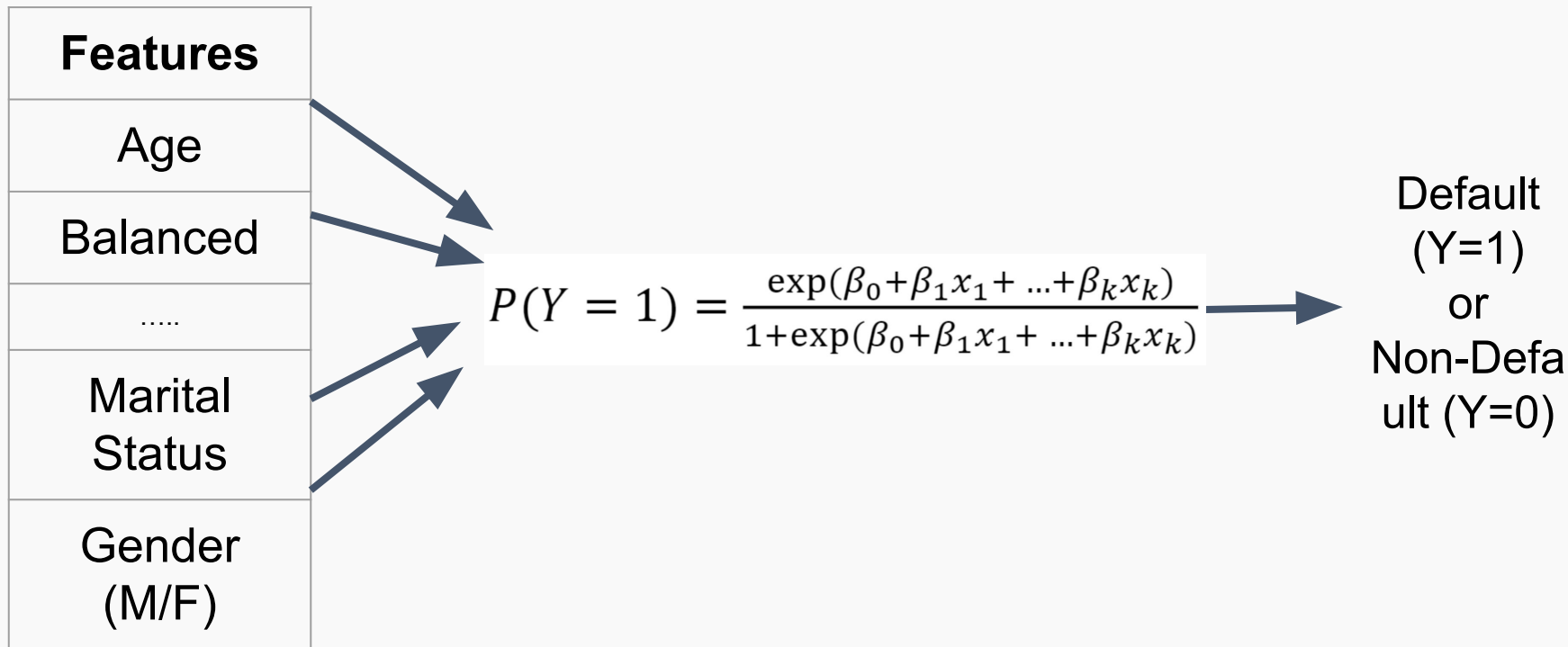
- This means the odds of a man defaulting on a loan are 3 times greater than the odds of a woman defaulting.
- We don't know the exact probabilities of default for men or women, but the odds ratio tells us the relative difference.

Visualizing Odds Ratios

Imagine we have 100 people applying for loans:

- Women: If 20 women default (out of 50), the odds of a woman defaulting are 20 (successes) / 40 (failures) = 1:2.
- Men: If 30 men default (out of 50), the odds of a man defaulting are 30 (successes) / 20 (failures) = 3:2.
- Odds-Ratio = $\text{Odd}(\text{Male}) / \text{Odd}(\text{Female}) = 3/2 \div 1/2 = 3$
- The odds ratio of 3 confirms this: men are 3 times more likely to default than women (based on this example).

Example: Credit Scoring



Example: Credit Scoring

Problem

How to predict **default risk of the new applicant** so we can **allocate loan efficiently** and **increase profit** from loan ?

Data

- What is being predicted ? default risk of the new applicant
- What is needed in prediction ? Demographical, Transaction behaviour, income, ect

ML
Objective

Maximize (profit - potential revenue lost)

Action

Do not allocate loan to a customer when the risk is too high, higher than 50%

Value

Profit Increase

Measuring Classification Performance

Confusion Matrix

- ▶ In the classification machine learning method, the most common method for evaluating the analysis model's result is the metric calculation, including classification accuracy, using a confusion matrix.
- ▶ The confusion matrix refers to the matrix that makes a crosstable of the predicted classification category from the analysis model and the actual classified category of data.

		Predicted categorical value	
		Y	N
Actual categorical value	Y	O (TP: True Positive)	X (FN: False Negative)
	N	X (FP: False Positive)	O (TN: True Negative)

Metric	Calculation Formula	Definition
accuracy	$(TP+TN) / (TP+TN+FP+FN)$	Ratio of accurate prediction of actual classification category (Ratio of TP and TN from the entire prediction)
error rate	$(FP+FN) / (TP+TN+FP+FN)$	Ratio of inaccurate prediction of actual classification category (Identical to 1-accuracy)
sensitivity = TP Rate	$(TP) / (TP+FN)$	Ratio of accurate prediction to 'positive' from actual 'positive' categories (True Positive – also referred to as Recall, Hit Ratio, and TP Rate)
specificity	$(TN) / (TN+FP)$	Ratio of accurate prediction to 'negative' from actual 'negative' categories (True Negative)
FP Rate	$(FP) / (TN+FP)$	Ratio of inaccurate prediction to 'positive' from actual 'negative' categories = 1-specificity
precision	$(TP) / (TP+FP)$	Ratio of actual 'True Positive' from the ratio of predicted 'positive'
F-Measure (F1-Score)	$\frac{2 * (\text{precision}) * (\text{recall})}{(\text{precision} + \text{recall})} = \frac{(2 * TP)}{(2 * TP + FP + FN)}$	Ranged between 0~1. It is the harmonic mean between precision and sensitivity (recall). If both precision and sensitivity are high, f-Measure also tends to have a larger value.
Kappa Statistic	$\{Pr(a) - Pr(e)\} / (1 - Pr(e))$	The value after eliminating coincidental agreement between predicted and actual values of the model (Ranged between 0~1. When the value is closer to 1, the predicted and actual values of the model accurately coincide. The values do not coincide when the value gets closer to 0.)

Measuring Classification Performance

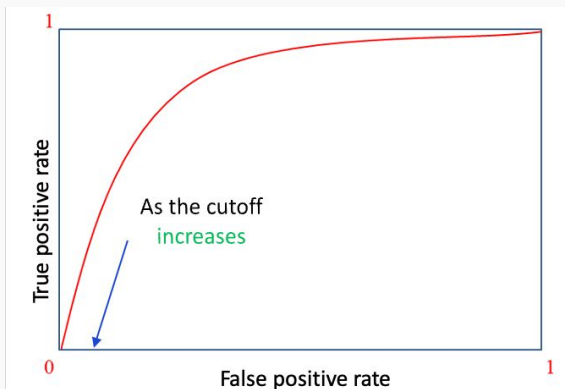
No	Prediction	Actual
1	1	1
2	1	0
3	0	1
..
499	0	0
500	0	1

Prediction	Actual	
	0	1
0	120	23
1	27	330

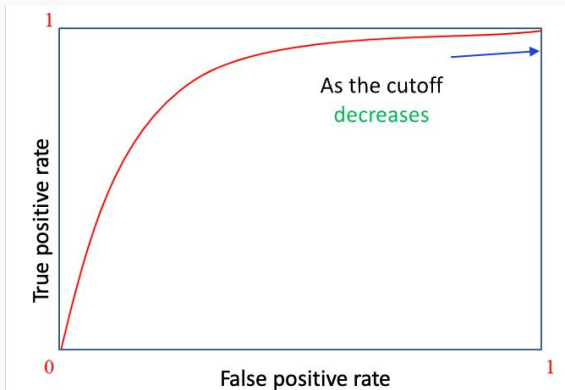
Accuracy Of Prediction = $(120+330)/500 \times 100\%$
 = 90.0%

Our model will correctly predict 9 of 10 People

ROC Curve



As the cutoff increases (closer to 1)	
Performance Metric	Increase/ Decrease
True Positive Rate (Sensitivity)	↓
Specificity	↑
False Positive Rate (1-Specificity)	↓
Precision	↑



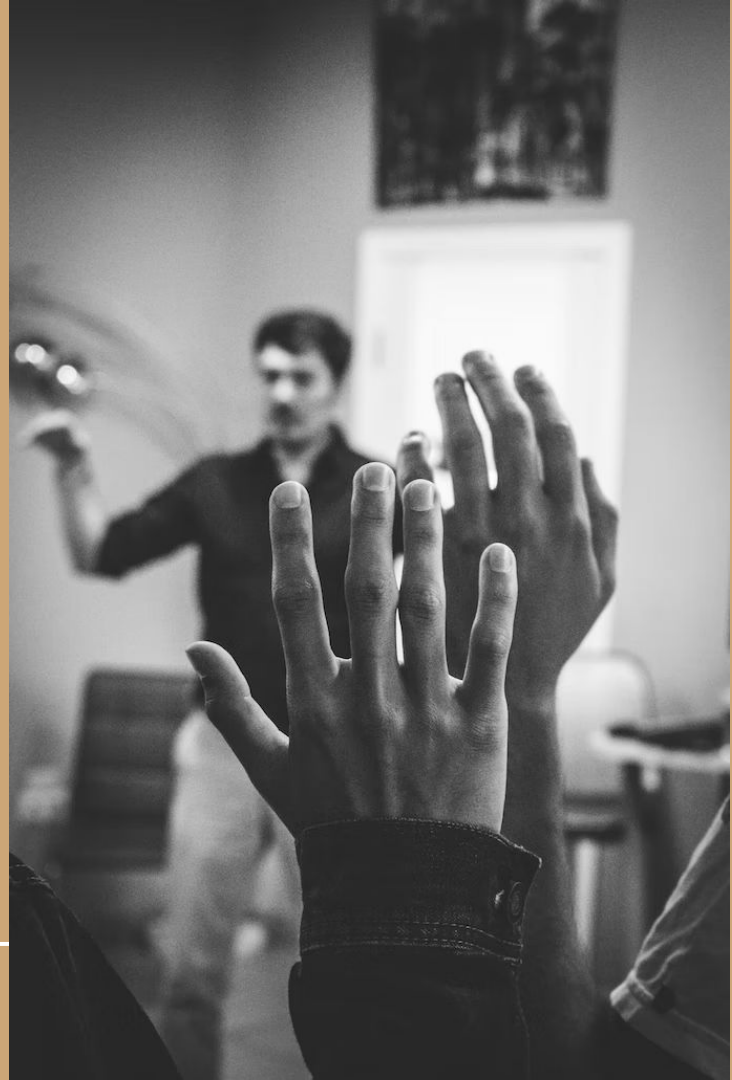
As the cutoff decreases (closer to 0)	
Performance metric	Increase/ Decrease
True Positive Rate (Sensitivity)	↑
Specificity	↓
False Positive Rate (1-Specificity)	↑
Precision	↓

► ROC curve is a parametric plot with respect to the *cutoff* probability.

► AUC stands for Area Under the Curve.

► AUC closer to 1 means good overall performance.

Challenges



Challenge

Develop a logistic regression model to predict loan default risk based on borrower characteristics in the bankloan.csv dataset.

Data:

Source: bankloan.csv

Target Variable: default (indicates loan default)

Tasks:

- Data Exploration and Preprocessing: Explore the bankloan.csv data, handle missing values (if any), and prepare it for modeling.
- Model Building: Construct a logistic regression model using default as the target variable and the specified features as predictors.
- Model Interpretation: Analyze the coefficients and interpret the model's results to understand how each feature influences loan default risk.
- Model Validation: Split the data into training and testing sets (80%/20% split suggested). Train the model on the training data and evaluate its performance on the unseen testing data using accuracy as the metric.

ANOVA F-Test (Simultant) for Multiple Linear

Hypothesis:

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$$

H_A : Not all β values are zero

Test Statistics : F-Statistics

Rejection Criteria:

P-value $\leq \alpha$ (two-sided)

- F-test check for overall significance of multiple regression model.
- F-test checks if there is a statistically significant relationship between Y (dependent variable) and any of the independent variables

T-Test (Partial)

Hypothesis:

$H_0 : \beta_i = 0$

$H_a : \beta_i \neq 0$ (two sided)

$\beta_i > 0$ or $\beta_i < 0$ (one sided)

Rejection Criteria:

$P\text{-value} \leq \alpha$ (two-sided)

$P\text{-value}/2 \leq \alpha$ (one-sided)

Test Statistics : t-Student

$$t = \frac{\hat{\beta}_i}{S_e(\hat{\beta}_i)}$$

- T-test checks if there is a statistically significant relationship between Y (dependent variable) and each of the independent variables

Goodness Of Fit Model : Adjusted R-Square

$$R_A^2 = 1 - \frac{SSE/(n - k - 1)}{SST/(n - 1)}$$

R_A^2 = Adjusted R - Square

n = number of observations

k = number of explanatory variables

- SST (Total Sum of Squares):
- SSE (Sum of Squares Error):
- SSR (Sum of Squares Regression):