

```
children: [  
  icon(icon, color: color  
  container(  
margin: const EdgeIns  
child:  
  label  
  style
```

 Google Developer Groups  
On Campus • Telkom University Bandung


# Supervised Learning



[hasnatf](https://www.linkedin.com/in/hasnatf)



Scan to download  
this presentation



Hasnat Ferdiananda  
IT practitioner, ex GITS

# Today's Topic

## Supervised Learning

- Overview
- Regression
- Classification

## Evaluation Metrics

- Overview
- Regression Metrics
- Classification Metrics

```
lookup.KeyValue  
f.constant(['em  
=tf.constant([G  
.lookup.StaticV  
_buckets=5)
```

# Overview - Machine Learning Process



## Data Pre-Processing

- Import the data
- Clean the data
- Split into training & test sets
- Feature Scaling



## Modelling

- Build the model
- Train the model
- Make predictions



## Evaluation

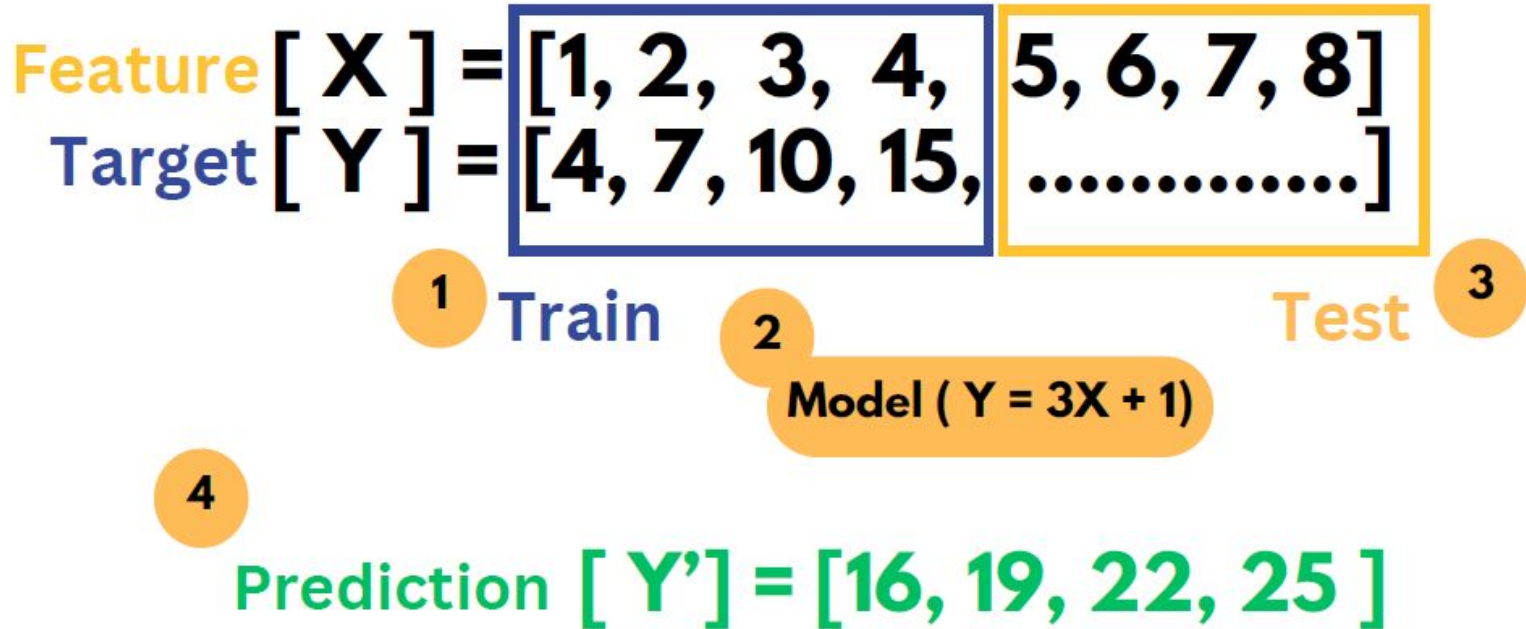
- Calculate performance metrics
- Make a verdict

# Overview - Modeling

What is a model? It's simply a specification of a mathematical (or probabilistic) relationship that exists between different variables.

The business model is probably based on simple mathematical relationships: profit is revenue minus expenses, revenue is units sold times average price, and so on.

# Overview - How ML Works



# Overview - How ML Works

$[X] =$	[1, 2, 3, 4, 5,	6, 7, 8]
$[Y] =$	[4, 7, 10, 15, 16,	19, 22, 25]
		Kunci Jawaban
	Train Set	Test Set

# Overview - How ML Works

## Model Selection

Model selection ibarat kamu memilih orang yang hendak kamu suruh untuk mencari pola atau rumus dari data yang kamu punya



Setiap orang memiliki cara berpikir yang berbeda, jadi usahakan pilih yang paling tepat dan pilih lebih dari satu supaya bisa dibandingkan

## Training Model

Train Set

$[X] = [1, 2, 3, 4, 5, ]$   
 $[Y] = [4, 7, 10, 15, 16, ]$



Setiap orang mempelajari pola dan rumus dari data train yang diberikan dan menetapkan rumus tersebut untuk mengerjakan test set

# Overview - How ML Works

## Testing Model

$[X] = [6, 7, 8]$

$[Y'] = [20, 21, 23]$   $[Y'] = [19, 22, 25]$   $[Y'] = [18, 22, 24]$



Setiap orang menjawab atau memprediksi nilai Y yang seharusnya muncul dengan pola dan rumus yang telah mereka tetapkan sebelumnya

## Evaluation

$[Y] = [19, 22, 25]$

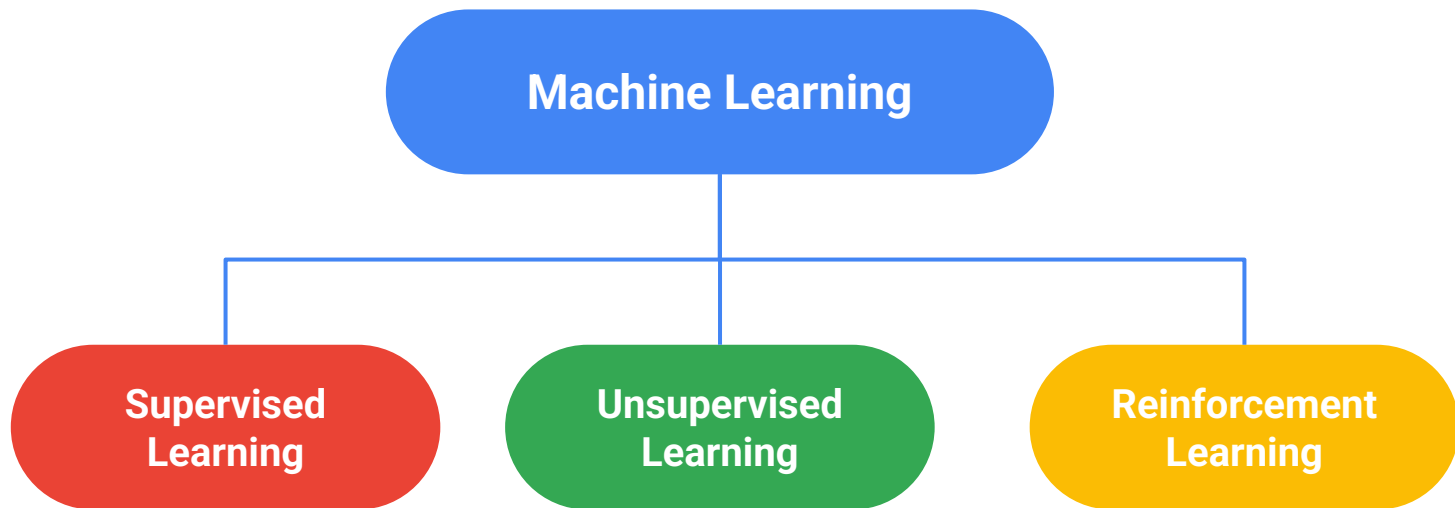
$[Y'] = [20, 21, 23]$   $[Y'] = [19, 22, 25]$   $[Y'] = [18, 22, 24]$



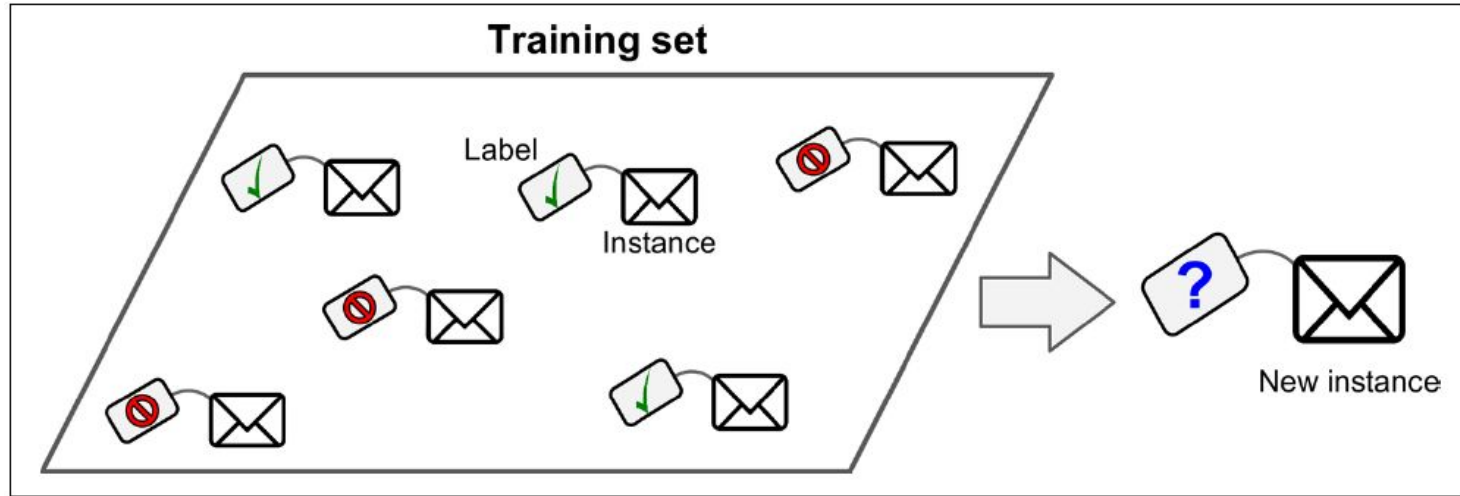
Kamu membandingkan jawaban tiap orang dengan kunci jawaban yang kamu miliki dan melakukan penilaian seberapa baik mereka menjawab



# Overview - ML Type of Supervision



# Overview - Supervised Learning



# Overview - Library

Python libraries are collections of pre-written code and functions that extend the capabilities of the Python programming language.



Used to analyze data



Visualization utility



Machine learning



Used for working with  
arrays



Visualization utility



Deep learning

# Overview - Scikit-learn

This module is designed to help with data processing and training data for machine learning or data science applications.

Scikit-learn is an open source machine learning library that supports supervised and unsupervised learning. It also provides various tools for model fitting, data preprocessing, model selection, model evaluation, and many other utilities.

# Overview - ML Algorithms



# Regression - Definition

Dalam konteks Machine Learning, regresi adalah metode yang digunakan untuk memprediksi nilai kontinu dari suatu variabel, seperti harga, pendapatan, atau suhu. Teknik ini melibatkan analisis hubungan antara variabel dependen (variabel yang ingin diprediksi) dan satu atau lebih variabel independen (variabel yang digunakan untuk memprediksi).

# Regression - Simple Linear Regression

This model involves a linear relationship between a dependent variable and one independent variable.

In simple linear regression, we try to find the best-fitting line to the data, which is called the regression line.

$$\hat{y} = b_0 + b_1 X_1$$

The diagram illustrates the components of the simple linear regression equation  $\hat{y} = b_0 + b_1 X_1$ . Each term is underlined and connected by a vertical line to its label below. The dependent variable  $\hat{y}$  is underlined in grey and labeled 'Dependent variable'. The y-intercept  $b_0$  is underlined in blue and labeled 'y-intercept (constant)'. The slope coefficient  $b_1$  is underlined in yellow and labeled 'Slope coefficient'. The independent variable  $X_1$  is underlined in grey and labeled 'Independent variable'.

Dependent variable

y-intercept (constant)

Slope coefficient

Independent variable

# Regression - Simple Linear Regression



~

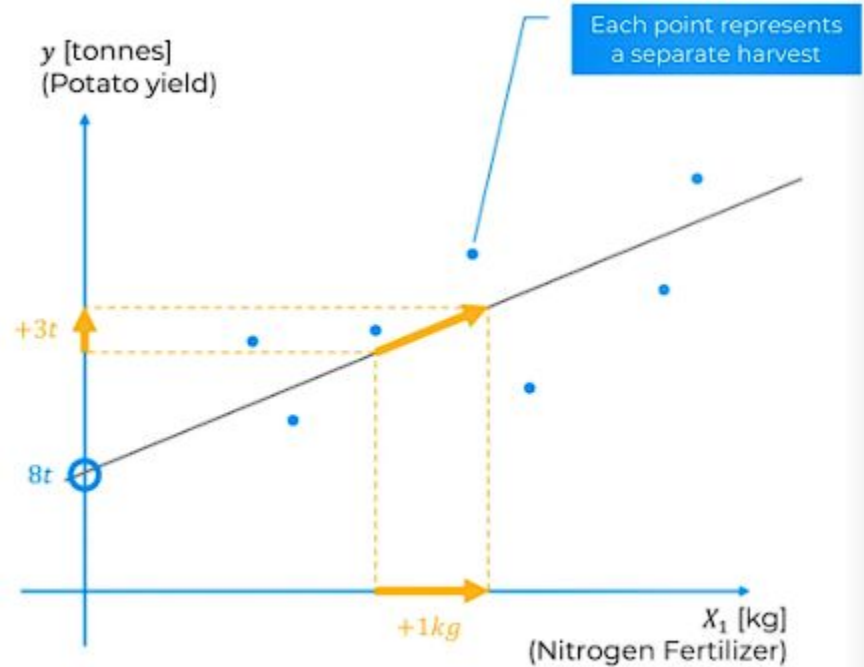


$$\hat{y} = b_0 + b_1 X_1$$

$$\text{Potatoes}[t] = b_0 + b_1 \times \text{Fertilizer}[kg]$$

$$b_0 = 8[t]$$

$$b_1 = 3\left[\frac{t}{kg}\right]$$

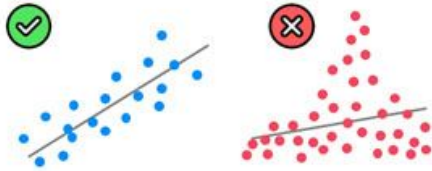




# Regression - Assumptions of Linear Regression

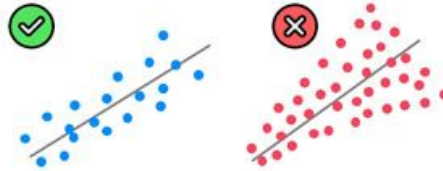
## 1. Linearity

(Linear relationship between Y and each X)



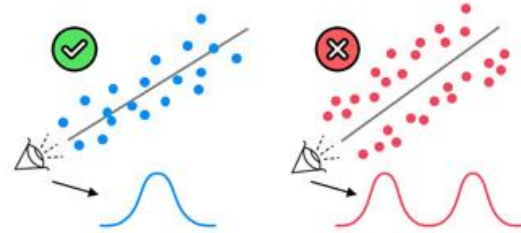
## 2. Homoscedasticity

(Equal variance)



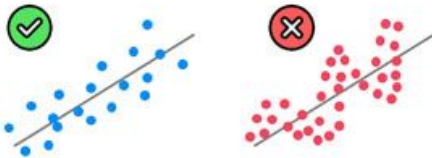
## 3. Multivariate Normality

(Normality of error distribution)



## 4. Independence

(of observations. Includes "no autocorrelation")



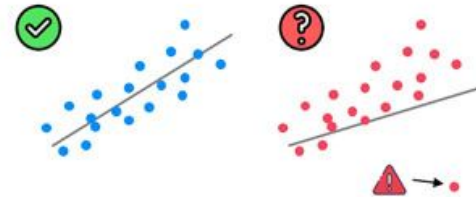
## 5. Lack of Multicollinearity

(Predictors are not correlated with each other)



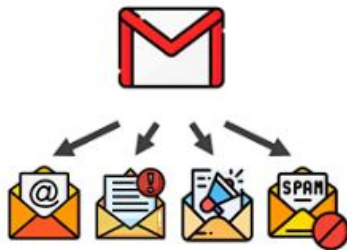
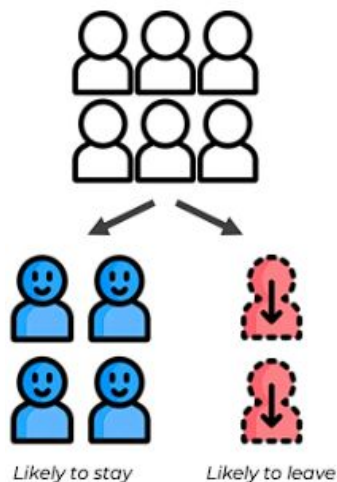
## 6. The Outlier Check

(This is not an assumption, but an "extra")



# Classification - Definition

*Classification: a Machine Learning technique to identify the category of new observations based on training data.*

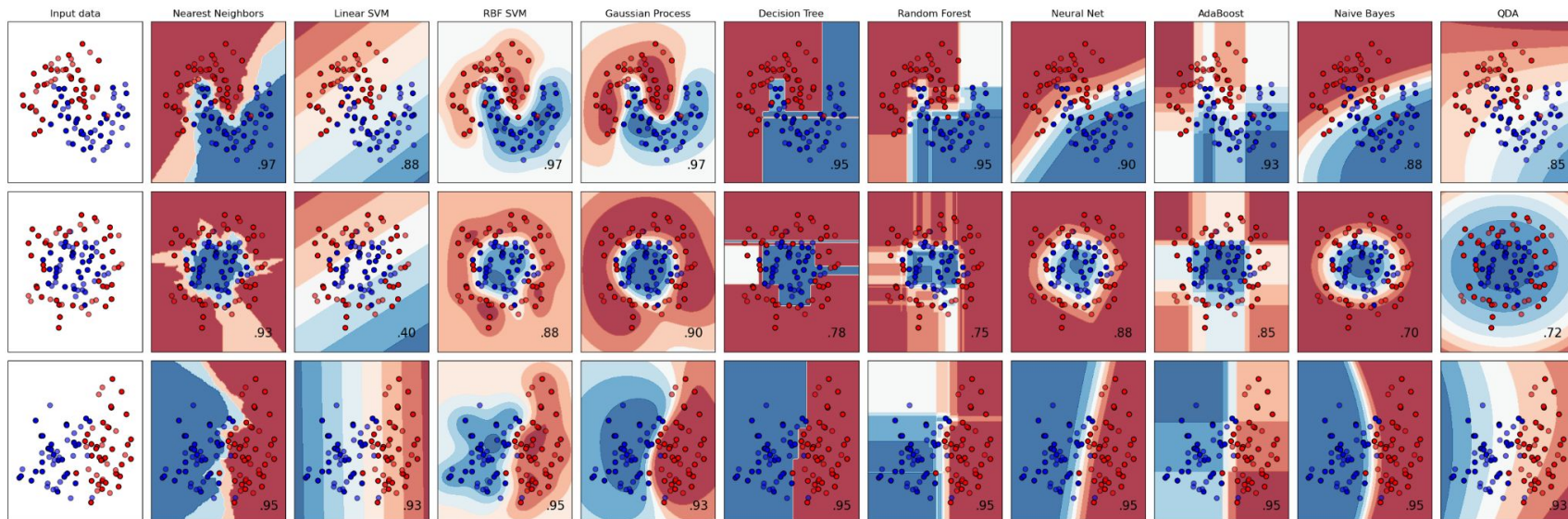


Identifying which category an object belongs to.

**Applications:** Spam detection, image recognition.

**Algorithms:** Gradient boosting, nearest neighbors, random forests, logistic regression, and more...

# Classification - Classifier Comparison

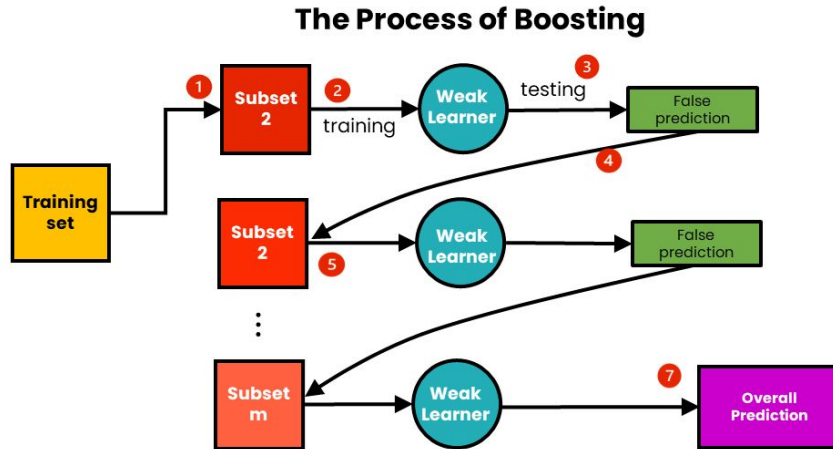


The point of this example is to illustrate the nature of decision boundaries of different classifiers.

# Classification - Ensembling

Ensembling is a technique in machine learning that combines multiple models to produce a prediction that is more robust and accurate than any individual model. The idea is that by aggregating the predictions from multiple models, the errors made by individual models can be compensated for by other models, thus improving the overall accuracy and stability of predictions.

# Classification - Boosting



Boosting is an ensembling technique that trains models sequentially. Each new model attempts to correct the errors of the previous models. Individual models within boosting are typically weak learners.

Ex: GradientBoosting

```
import xgboost

xgb_reg = xgboost.XGBRegressor()
xgb_reg.fit(X_train, y_train)
y_pred = xgb_reg.predict(X_val)
```

# Evaluation Metrics

```
lookup.KeyValue  
f.constant(['em  
=tf.constant([G  
lookup.StaticV  
_buckets=5)
```



# Overview

**Metric** dalam konteks machine learning adalah ukuran yang digunakan untuk mengevaluasi kinerja model. Metrik membantu kita memahami seberapa baik model kita dalam membuat prediksi dan seberapa akurat atau efektif hasil yang diberikan oleh model tersebut.

## Mengapa Metric Penting?

- **Evaluasi Model:** Metric memberi kita cara objektif untuk menilai kualitas model. Tanpa metric, kita tidak tahu seberapa baik model kita dibandingkan dengan model lain atau dengan hasil yang diharapkan.
- **Perbandingan Model:** Dengan metric, kita bisa membandingkan berbagai model atau algoritma untuk melihat mana yang memberikan hasil terbaik.
- **Optimasi Model:** Metric membantu kita memahami area mana dari model yang perlu diperbaiki.

## SUPERVISED

### Regression   Classification

- MAE
  - MSE
  - RMSE
  - $R^2$
- Confusion Matrix

## UNSUPERVISED

- Silhouette Score
- Davies-Bouldin Index
- Elbow Method

# Regression Metrics

Metrics yang dipakai untuk mengukur seberapa jauh nilai prediksi dengan data aktual

## MAE

Mean Absolute Error

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

Level : Kurang Terasa

MAE menghitung rata-rata nilai absolut dari kesalahan. Ini memberikan ukuran kesalahan yang lebih stabil karena tidak mengkuadratkan kesalahan, sehingga setiap kesalahan besar memiliki dampak yang sama pada MAE seperti kesalahan kecil. MAE lebih "rata" dan tidak terlalu dipengaruhi oleh outlier dibandingkan dengan MSE dan RMSE.

## RMSE

Root Mean Squared Error

$$\text{RMSE} = \sqrt{\text{MSE}}$$

Level : Menengah

RMSE adalah akar kuadrat dari MSE, sehingga nilainya kembali dalam satuan yang sama dengan data asli. RMSE memberikan gambaran yang lebih jelas tentang seberapa besar kesalahan rata-rata, tetapi masih mempertimbangkan dampak kesalahan besar karena MSE juga sensitif terhadap outlier.

## MSE

Mean Squared Error

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

Level : Paling Terasa

MSE mengkuadratkan kesalahan, sehingga kesalahan yang lebih besar memberikan kontribusi yang jauh lebih besar pada nilai MSE. Dengan kata lain, kesalahan besar dihukum lebih berat, dan ini membuat MSE sangat sensitif terhadap outlier (data yang sangat berbeda dari yang lain).





# Regression Metrics

## R<sup>2</sup> Score

### Residual Sum of Squares (SSE):

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

di mana  $y_i$  adalah nilai aktual dan  $\bar{y}$  adalah rata-rata nilai aktual.

### Total Sum of Squares (SST):

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

di mana  $\hat{y}_i$  adalah nilai yang diprediksi oleh model.

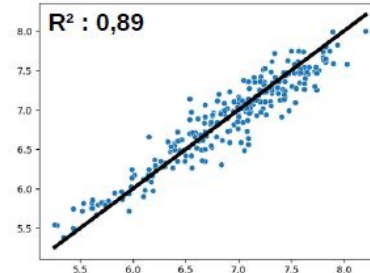
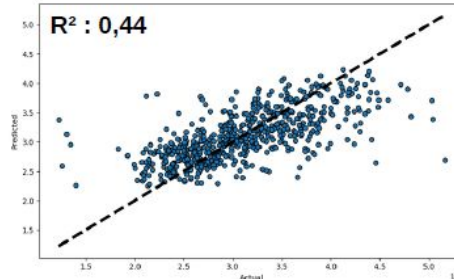
### R-Squared Score

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

R-squared ( $R^2$ ), juga dikenal sebagai koefisien determinasi, adalah ukuran statistik yang digunakan untuk mengevaluasi seberapa baik model regresi menjelaskan variasi dalam data. Ini memberikan indikasi seberapa baik model fit dengan data yang sebenarnya.

- **Tidak Menunjukkan Kualitas Model Secara Keseluruhan:** R-squared tidak memberikan informasi tentang seberapa baik model fit dengan data. Model dengan R-squared tinggi bisa jadi overfitting, terutama jika ada banyak variabel.
- **Sensitif Terhadap Outlier:** Outlier dapat mempengaruhi R-squared, membuatnya tampak lebih baik atau lebih buruk dari yang sebenarnya.
- **Tidak Berlaku untuk Semua Jenis Model:** R-squared paling sering digunakan dalam regresi linear. Untuk model yang lebih kompleks atau non-linear, seperti regresi logistik, R-squared mungkin tidak memberikan informasi yang berarti.

R-Squared biasanya dibuktikan menggunakan Scatter Plot dengan membandingkan data aktual dengan prediksi



# Classification Metrics - Confusion Matrix

		Truth		
		1	0	
Prediction	1	TP	FP	+
	0	FN	TN	-

**Analogi** : Bayangkan kamu lagi suka sama seseorang, terus kamu berpikir apakah dia juga suka kamu atau nggak? terus bagaimana kenyataannya? dari semua pemikiran dan kenyataan itu kita pecah jadi 4 kondisi

## Positive Thinking

- **True Positive** : Kamu berpikir dia suka kamu dan ternyata kenyataannya dia memang suka kamu
- **False Positive** : Kamu berpikir dia suka kamu dan ternyata kenyataannya dia gak suka kamu

## Negative Thinking

- **False Negative** : Kamu berpikir dia gak suka kamu tetapi ternyata kenyataannya dia suka kamu
- **True Negative** : Kamu berpikir dia gak suka kamu dan memang kenyataannya dia gak suka kamu

# Classification Metrics

## Accuracy

		Truth	
		Sakit	Sehat
Pred	Sakit	10	20
	Sehat	30	80

$$\frac{10 + 80}{10 + 80 + 20 + 30} = 0,64$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Accuracy: Proporsi prediksi yang benar dibandingkan dengan total prediksi.

Accuracy bukan segalanya dalam menilai sebuah model klasifikasi

## Recall (Pengukur sensitivitas)

		Truth	
		Sakit	Sehat
Pred	Sakit	10	20
	Sehat	30	80

$$\frac{10}{10 + 30} = 0,25$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

Recall : Proporsi nilai positif yang benar-benar terdeteksi oleh model

Pada beberapa kasus, model yang memiliki nilai recall yang kecil dapat membuat keputusan atau prediksi yang diambil berbahaya

## Precision

		Truth	
		Bersalah	Tidak Bersalah
Pred	Bersalah	10	40
	Tidak Bersalah	30	80

$$\frac{10}{10 + 40} = 0,20$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

Precision : proporsi prediksi positif yang benar dibandingkan dengan total prediksi positif.

Pada beberapa kasus, model yang memiliki nilai Precision yang kecil dapat membuat keputusan atau prediksi yang diambil berbahaya

## F1 Score

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

F1 Score: Harmonik rata-rata antara precision dan recall

Untuk mengetahui seberapa seimbang antara Precision dan Recall pada sebuah model

# Hands-on



Google Colab

Dataset: <https://www.kaggle.com/datasets/yeanzc/telco-customer-churn-ibm-dataset/data>

# Task:

Predicting customer churn on Kaggle

<https://www.kaggle.com/t/eae59a87b88945f1a17384afd23e1eac>



Google Developer Groups

```
er(  
ll(32),  
  
/*1*/  
child: Column(  
  crossAxisAlignment: CrossAxisAlignment.  
  children: [  
    /*2*/  
    Container(  
      padding: const EdgeInsets.  
      child: const Text(  
        'Oeschinen Lake Campg  
        style: TextStyle(  
          fontWeight: FontWeig  
      ),  
    ),  
  ],  
),  
)
```

# The end of this session, Any question guys?

Let's connect and collaborate

Email: [ferdianandahasnat@gmail.com](mailto:ferdianandahasnat@gmail.com)

Personal Page: <https://hasnat.vercel.app/>

LinkedIn: <https://www.linkedin.com/in/hasnatf/>

Instagram: [@hasnat5](https://www.instagram.com/hasnat5)



Scan to download this  
presentation

```
/*1*/  
child: Column(  
  crossAxisAlignment: CrossAxisAlignment.  
  children: [  
    /*2*/  
    Container(  
      padding: const EdgeInsets.  
      child: const Text(  
        'Oeschinen Lake Campg  
        style: TextStyle(  
          fontWeight: FontWeig  
      ),  
    ),  
  ],  
),
```