

*Unsupervised Learning*

*Communication & Presentation  
Skill*

*Evaluation Metrics and Model  
Selection*

*Learning Progress Review Week 14*

By:  
**MARVEL TEAM**

Fikrie | Natalia | Satria

# 1. *Unsupervised Learning*

## Supervised Learning

**Supervised Learning** adalah sebuah pendekatan dimana sudah terdapat data yang di latih, dan terdapat variable yang di targetkan sehingga tujuan dari pendekatan ini adalah mengelompokkan suatu data ke data yang sudah ada.

## Supervised Learning

Algoritma Supervised Learning:

- Decision Tree
- Nearest-Neighbour (KNN)
- Naive Bayes Classifier
- Artificial Neural Network
- Support Vector Machine
- Fuzzy K-Nearest Neighbour

# Unsupervised Learning

**Unsupervised Learning** tidak memiliki data latih, sehingga dari data yang ada kita mengelompokkan data tersebut menjadi 2 bagian atau 3 bagian dan seterusnya.

## Unsupervised Learning

Algoritma Unsupervised Learning:

- K-Means
- Hierarchical Clustering
- DBScan
- Fuzzy C-Means
- Self-Organizing Map

## K-Means Algorithm

**K-Means** Clustering adalah salah satu Unsupervised Learning Algorithm. Tujuan dari algorithm ini adalah untuk menemukan grup dalam data, dengan jumlah grup di wakili oleh variable K. Variabel K sendiri adalah jumlah cluster yang kita inginkan.

## K-Means Algorithm

Proses :

Untuk memproses data algoritma **K-Means** clustering, data dimulai dengan kelompok pertama centroid yang dipilih secara acak, yang digunakan sebagai titik awal untuk setiap cluster, dan kemudian melakukan perhitungan berulang untuk mengoptimalkan posisi centroid.



## K-Means Algorithm

Proses :

Untuk memproses data algoritma **K-Means** clustering, data dimulai dengan kelompok pertama centroid yang dipilih secara acak, yang digunakan sebagai titik awal untuk setiap cluster, dan kemudian melakukan perhitungan berulang untuk mengoptimalkan posisi centroid.

# K-Means Algorithm

Proses ini berhenti atau telah selesai dalam mengoptimalkan cluster Ketika:

- Centroid telah stabil – tidak ada perubahan dalam nilai-nilai mereka karena pengelompokan telah berhasil.
- Jumlah iterasi yang ditentukan telah tercapai.

Hasil dari K-Mean:

- Centroid dari cluster K, yang dapat digunakan untuk memberi label data baru.
- Label untuk data pelatihan (setiap titik data ditugaskan ke satu cluster).

# Hierarchical Clustering

Hierarchical clustering adalah Teknik clustering membentuk hirarki atau berdasarkan tingkatan tertentu sehingga menyerupai struktur pohon. Dengan demikian proses pengelompokannya dilakukan secara bertingkat atau bertahap. Biasanya, metode ini digunakan pada data yang jumlahnya tidak terlalu banyak dan jumlah cluster yang akan di bentuk belum diketahui.

Teknik clustering ini memiliki 2 tipe:

- ❖ Agglomerative
- ❖ Divisive

Density – Based Spatial Clustering Algorithm with Noise (DBSCAN) adalah algoritma pengelompokan yang didasarkan pada kepadatan (density) data.

Konsep kepadatan yang dimaksud dalam DBSCAN adalah banyaknya data (minPts) yang berada dalam radius Eps dari suatu data.

# Comparing Different Clustering Algorithms

	K-Means	Hierarchical Clustering	DBSCAN	Mean Shift
<b>Parameters</b>	Need to specify #clusters ( $k$ ) first	Have to try $k$ values	Doesn't need to specify #clusters, but not intuitive in determining parameters	Doesn't need to specify #clusters
<b>Scalability</b>	Scalable on large dataset	Very heavy computational on large dataset (complexity proportional to square number of observations)	Can handle lots of data	Slow with a lot of data
<b>Cluster Sizes</b>	Tends to find even cluster sizes (that's why sometimes we need to get the large $k$ to get more insights)	Can find uneven cluster sizes	Can find uneven cluster sizes	Can find uneven cluster sizes
<b>Evaluation Metrics</b>	Can use many metrics evaluation	Lot of distance metrics & linkage options	Can use many metrics evaluation	Limited to use Euclidean Distance only
<b>General Use Case/Applications</b>	General purposes	General purpose	Outlier detection	Often used in video analysis

## **2.** *Communication and Presentation Skill*

## *Exploratory & Explanatory*

**Exploratory** adalah bagaimana mencari tahu dan memahami data yang menarik untuk dijelaskan.

*Contoh data :*

- Tren penjualan kopi selama lima tahun terakhir

**Explanatory** adalah menemukan bagian yang menarik dan menjelaskannya secara gamblang.

*Contoh :*

- Big idea worksheet

## *Understand The Context*

Ada beberapa hal yang perlu diperhatikan untuk memahami konteks :

- *Mendefinisikan secara spesifik target audiens*
- Big Idea Worksheet
- *Mengkritisi Ide*
- Storyboarding



## Target Audiens

- Cari masukan/saran mengenai profil audiens dari kolega
- Saat menjelaskan data, gunakan perspektif audiens yang akan dihadapi
- Buat skala prioritas siapa audiens yang paling membutuhkan data tersebut

## *Big Idea Worksheet*

- Definisikan secara spesifik & detil target audiens
- Gambarkan keuntungan & resiko ketika audiens melakukan atau tidak melakukan berdasarkan data yang kita jelaskan
- Deskripsikan & artikulasikan ide/gagasan besar

## Kritisi Ide

- Pikirkan pertanyaan apa yang mungkin muncul
- Apa yang dijelaskan mungkin tidak terpikirkan oleh audiens, maka ajak mereka untuk terus mengembangkan ide/gagasan dasar

## *Storyboarding*

- Penjelasan secara spesifik mengenai pengelolaan ide/gagasan dasar
- Hal ini meliputi bagaimana produk tersebut dibeli, apa yang membedakan produk dengan yang lainnya dan pengelolaan media/website penjualan

## *Choosing an Effective Visual*

Ada beberapa hal yang perlu diperhatikan untuk memilih visualisasi yang efektif :

- *Visualisasi digunakan untuk memudahkan pemahaman data, bukan memperumit pemahaman data*
- *Selalu cantumkan informasi umum di bagian atas data*
- *Posisikan informasi zig-zag karena kecenderungan mata melihat seperti membaca buku, dari kiri atas – kanan atas – kiri bawah – kanan bawah*
- *Posisikan judul dengan baik dan berikan sub-judul sebagai informasi spesifik*

Ada beberapa hal yang perlu diperhatikan dalam membangun narasi:

- *Gunakan judul yang menarik untuk membangun ekspektasi audiens*
- *Lengkapi grafik data dengan deskripsi/penjelasan*
- *Identifikasikan data-data yang menarik/menantang*

## *Presentation Tools*

Pilihlah aplikasi presentasi yang sesuai dengan tujuan presentasi data.



# Practice

### **3. *Evaluation Metrics and Model Selection***

### **Mengapa model *machine learning* harus dievaluasi?**

Evaluasi diperlukan untuk mengetahui mana yang terbaik dari hasil uji coba pada metode dan algoritma *machine learning* yang kita pilih. Dari evaluasi tersebut, kita dapat membandingkan kinerja dari tiap-tiap algoritma.

## Cross Validation

**Cross Validation** digunakan untuk menghindari *overlapping* pada *data testing*. Ini disebut juga dengan *k-fold cross-validation*.

Tahapan *cross-validation* :

1. Bagi data menjadi *k* subset yang berukuran sama
2. Gunakan setiap subset untuk *data testing* dan sisanya untuk *data training*

Seringkali subset dibuat *stratified* (bertingkat) sebelum *cross-validation* dilakukan, karena stratifikasi akan mengurangi variansi dari estimasi.



## *Evaluation Metrics : Regression*

**Tiga (3) metrik evaluasi yang umum untuk model regresi adalah:**

1. *R square / Adjusted R Square*
2. *Mean Square Error (MSE) / Root Mean Square Error (RMSE)*
3. *Mean Absolute Error (MAE)*

## *Evaluation Metrics : Regression*

### **1. R square / Adjusted R Square**

Nilai  $R^2$  mengukur seberapa dekat nilai data yang diketahui dengan garis regresi yang dipasang. Nilai  $R^2$  biasanya berkisar dari 0,0 hingga 1,0. Nilai  $R^2$  yang mendekati 1,0 menunjukkan kesesuaian model yang lebih kuat.

## *Evaluation Metrics : Regression*

### **2. Mean Square Error (MSE) / Root Mean Square Error (RMSE)**

MSE (Mean Square Error) mengacu pada nilai rata-rata dari nilai kesalahan kuadrat dihitung untuk setiap datapoint. Nilai MSE yang mendekati nol menunjukkan kinerja model yang lebih baik.

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

$y_i$  = actual value

$\hat{y}_i$  = predicted value

$n$  = # of observations

## *Evaluation Metrics : Regression*

### **2. Mean Square Error (MSE) / Root Mean Square Error (RMSE)**

RMSE (Root Mean Squared Error) adalah salah satu metrik evaluasi paling populer untuk masalah regresi. RMSE dihitung dengan mengambil akar kuadrat dari MSE. Nilai RMSE yang lebih rendah menunjukkan performa model yang lebih baik.

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

$y_i$  = actual value

$\hat{y}_i$  = predicted value

$n$  = # of observations

## *Evaluation Metrics : Regression*

### **3. Mean Absolute Error (MAE)**

MAE (Mean Absolute Error) mengacu pada nilai rata-rata dari nilai kesalahan mutlak dihitung untuk setiap titik dalam dataset. MAE dihitung menggunakan persamaan di samping. Model yang sempurna menghasilkan MAE nol dan semakin dekat MAE yang diamati ke nol, semakin baik model tersebut cocok dengan datanya.

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

$\hat{y}_i$  = *predicted value*

$y_i$  = *actual value*

$n$  = *# of observations*

## *Evaluation Metrics : Classification*

**Tiga (3) metrik evaluasi yang umum untuk model klasifikasi adalah:**

1. *Confusion Matrix*
2. *Precision*
3. *Recall*

# Evaluation Metrics : Classification

## 1. Confusion Matrix

*Confusion Matrix* sering disebut *error matrix*. Pada dasarnya *confusion matrix* memberikan informasi perbandingan hasil klasifikasi yang dilakukan oleh sistem (model) dengan hasil klasifikasi sebenarnya. Gambar disamping merupakan *confusion matrix* dengan 4 kombinasi nilai prediksi dan nilai aktual yang berbeda.

		Actual Values	
		1 (Positive)	0 (Negative)
Predicted Values	1 (Positive)	<b>TP</b> (True Positive)	<b>FP</b> (False Positive) <small>Type I Error</small>
	0 (Negative)	<b>FN</b> (False Negative) <small>Type II Error</small>	<b>TN</b> (True Negative)

## *Evaluation Metrics : Classification*

### **2. Precision**

*Precision* merupakan pembagian dari jumlah total contoh positif yang diklasifikasikan bernilai benar dengan jumlah total contoh positif yang diprediksi. *High precision* menunjukkan contoh berlabel positif memang positif (False Positif rendah).

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$



## *Evaluation Metrics : Classification*

### **3. Recall**

*Recall* dapat didefinisikan sebagai rasio dari jumlah total contoh positif yang diklasifikasikan bernilai benar dibagi dengan jumlah total contoh positif. *High recall* menunjukkan kelas dikenali dengan baik (False Negatif rendah).

$$\begin{aligned}\text{Recall} &= \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \\ &= \frac{\text{True Positive}}{\text{Total Actual Positive}}\end{aligned}$$

Special Thanks to :



Slide template by SlideCarnival