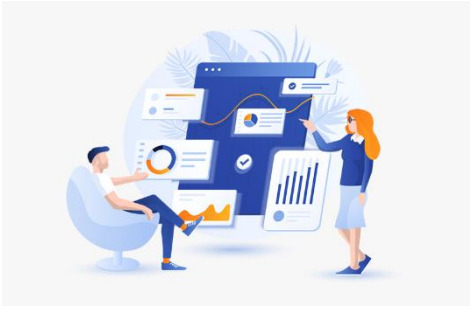
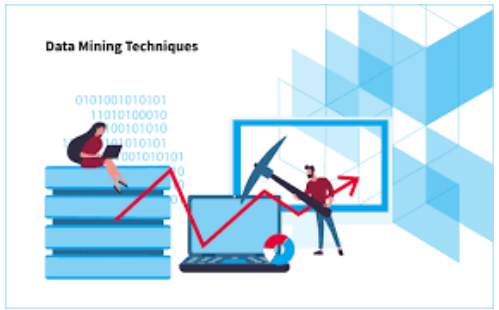
DATA MINING

09 – Hierarchical Clustering

Oleh: Leny Tritanto N., M.Kom.

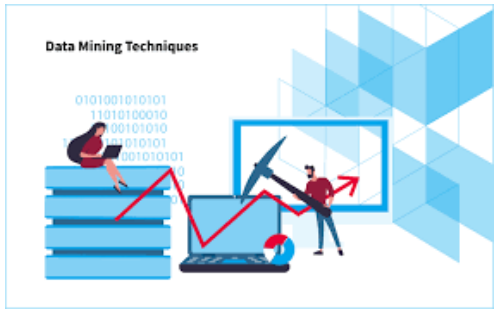




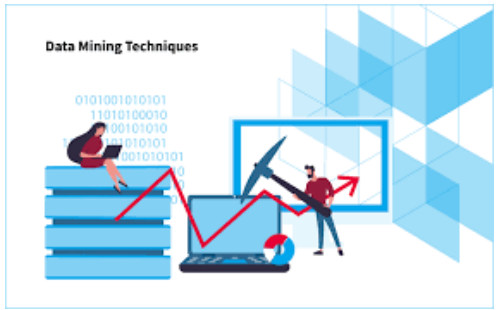
Algoritma Supervised dan UnsupervisedLearning:

Supervised Learning	Unsupervised Learning
▪ Linear Regression	✓ K-Means
✓ Decision Tree and Random Forest	➡ Hierarchical Clustering
✓ Naive Bayes Classifier	▪ DBSCAN
✓ Nearest Neighbour Classifier (KNN)	▪ Association Rule
✓ Artificial Neural Network	▪ Apriori Algorithm
✓ Support Vector Machine (SVM)	

UNBIN



Apa dan bagaimana menerapkan Hierarchical Clustering?

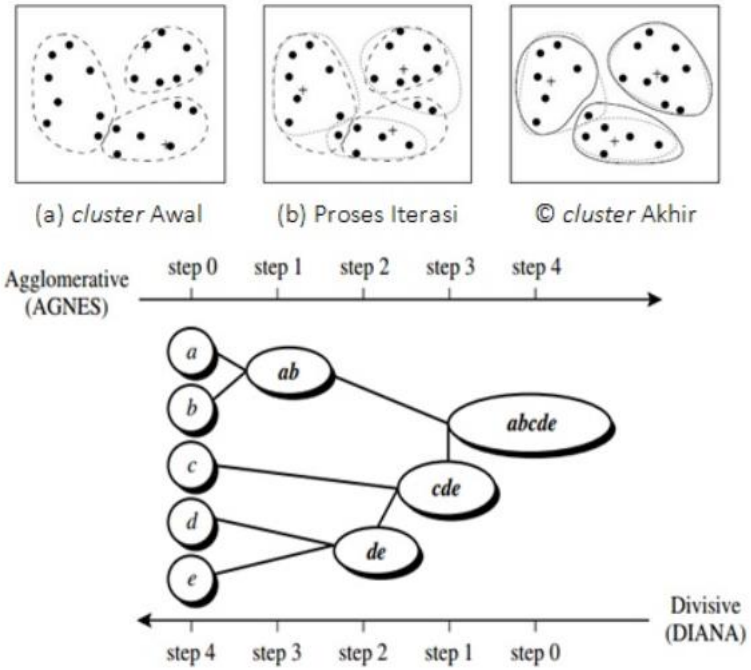


Introduction to Hierarchical Clustering

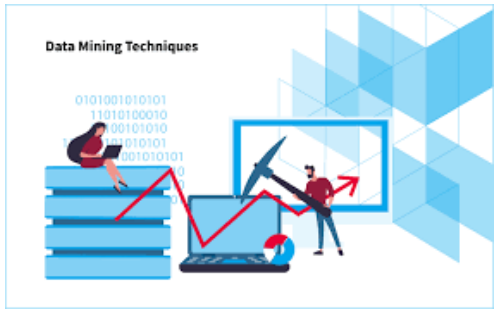
Clustering merupakan salah satu metode Unsupervised Learning yang bertujuan untuk melakukan pengelompokan data berdasasrkan kemiripan/jarak antar data. Clustering memiliki karakteristik dimana anggota dalam satu cluster memiliki kemiripan yang sama atau jarak yang sangat dekat, sementara anggota antar cluster memiliki kemiripan yang sangat berbeda atau jarak yang sangat jauh. Menurut (Tan et al., 2006) dalam bukunya yang berjudul Introduction to Data Mining, metode clustering dibagi menjadi dua jenis, yaitu **Hierarchical Clustering** dan **Partitional Clustering**.

Partitional Clustering umumnya bertujuan untuk mengelompokkan data menjadi beberapa cluster yang lebih kecil. Pada prosesnya, setiap cluster akan memiliki titik pusat cluster (centroid) dan mencoba menghitung setiap data yang paling dekat dengan centroid tersebut. Metode dalam partitional clustering diantaranya *k-means, fuzzy k-means,* dan lain-lain.

Sedangkan dalam **Hierarchical Clustering**, pengelompokan data dilakukan dengan membuat suatu bagan hirarki **(dendrogram)** dengan tujuan menunjukkan kemiripan antar data. Setiap data yang mirip akan memiliki hubungan hirarki yang dekat dan membentuk cluster data. Bagan hirarki akan terus terbentuk hingga seluruh data terhubung dalam bagan hirarki tersebut. Cluster dapat dihasilkan dengan memotong bagan hirarki pada level tertentu. Beberapa metode dalam hierarchical clustering yaitu ***single linkage, complete linkage, average linkage, dan ward's minimum variance.***



Gambaran proses klasterisasi data

Hierachical Clustering Approach

Secara umum, hierarchical clustering dibagi menjadi dua jenis yaitu agglomerative dan divisive3. Kedua metode ini dibedakan berdasarkan pendekatan dalam melakukan pengelompokkan data hingga membentuk dendrogram, menggunakan bottom-up atau top-down manner.

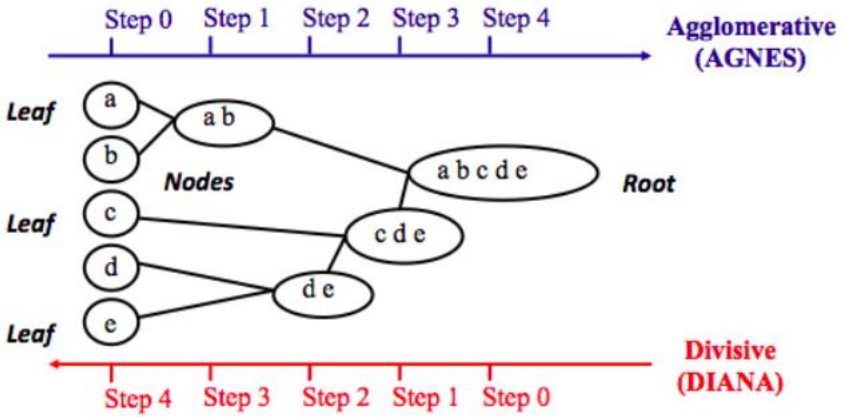
Agglomerative Clustering

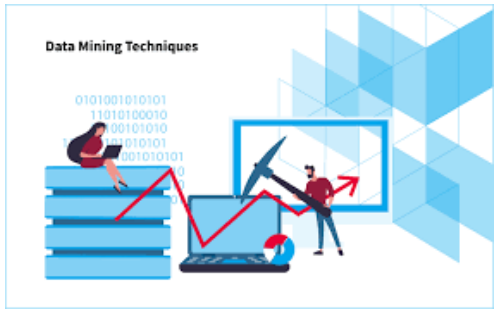
Agglomerative clustering biasa disebut juga sebagai agglomerative nesting (AGNES) dimana cara kerja dalam melakukan pengelompokan data menggunakan bottom-up manner. Prosesnya dimulai dengan menganggap setiap data sebagai satu cluster kecil (leaf) yang hanya memiliki satu anggota saja, lalu pada tahap selanjutnya dua cluster yang memiliki kemiripan akan dikelompokkan menjadi satu cluster yang lebih besar (nodes). Proses ini akan dilakukan terus menerus hingga semua data **menjadi satu cluster besar (root).**

Divisive hierarchical clustering

Divisive hierarchical clustering biasa disebut juga sebagai divisive analysis (DIANA) di mana cara kerja dalam melakukan pengelompokan data menggunakan top-down manner. Prosesnya dimulai dengan menganggap satu set data sebagai satu cluster besar (root), lalu dalam setiap iterasinya setiap data yang memiliki karakteristik yang berbeda akan dipecah menjadi dua cluster yang lebih kecil (nodes) dan proses akan terus berjalan hingga setiap data menjadi satu cluster kecil (leaf) yang hanya memiliki satu anggota saja.

Berikut adalah ilustrasi mengenai bagaimana agglomerative dan divisive clustering bekerja.





(DIS) SIMILARITY MEASURE

Terdapat beragam metode penghitungan (dis)similarity. Pemilihan metode (dis)similarity akan menentukan bagaimana kemiripan antar data dihitung. Itulah mengapa pemilihan metode (dis)similarity menjadi salah satu hal penting dalam pembuatan hierarchical clustering.
Metode penghitungan (dis)similarity yang sering digunakan adalah euclidean distance dan manhattan distance, namun bisa saja menggunakan pengukuran jarak yang lain, bergantung pada data yang sedang kita analisis. Berikut ini formula dalam perhitungan (dis)similarity dari kedua metode tersebut:

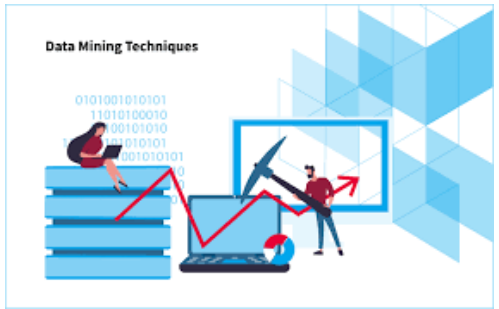
Euclidean distance

$$d_{xy} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Manhattan distance

$$d_{xy} = \sum_{i=1}^{n}|(x_i - y_i)|$$

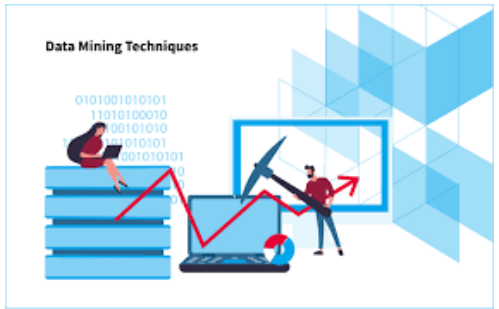
Selanjutnya, nilai (dis)similarity antar data ini akan dibentuk menjadi distance matrix. Kemudian, distance matrix tersebut akan diolah untuk penyusunan dendrogram.



LINKAGE METHOD

Dalam hierarchical clustering, selain menghitung (dis)similarity antar data, diperlukan juga cara untuk menghitung (dis)similarity antar cluster sehingga dapat terbentuk dendrogram dari cluster-cluster yang dekat. Proses penggabungan cluster-cluster kecil menjadi satu dendrogram utuh dilakukan melalui beberapa pendekatan Linkage Method. Berikut ini linkage method yang sering digunakan pada agglomerative approach:

1. **Complete Linkage / Maximum Linkage**
2. **Single Linkage / Minimum Linkage**
3. **Average Linkage**
4. **Centroid Linkage**
5. **Ward's minimum Variance**

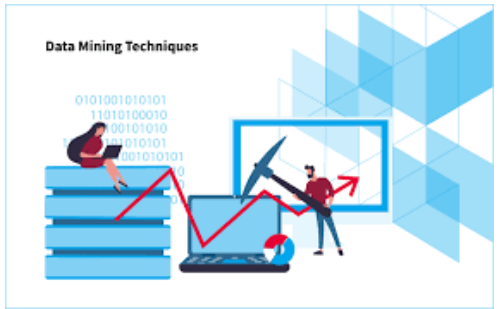
COMPLETE / MAXIMUM LINKAGE

Pengukuran (dis)similarity atau jarak antar cluster dilakukan dengan mengukur terlebih dahulu jarak antar tiap observasi dari cluster yang berbeda (pairwise distances). Jarak paling tinggi (maximum distance) akan menjadi ukuran (dis)similarity antar cluster. Kemudian, dendrogram akan terbentuk dari cluster-cluster yang memiliki (dis)similarity paling kecil. Hal ini membuat dendrogram yang terbentuk menjadi lebih terpisah antar clusternya (terbentuk cluster yang “compact”).
Berikut formula jarak antar cluster menggunakan complete linkage:

$$d_{12} = \max_{ij} d(X_i, Y_j)$$

di mana:

- $X_1, X_2, \ldots, X_k$: observasi pada cluster 1
- $Y_1, Y_2, \ldots, Y_k$: observasi pada cluster 2
- $d(X, Y)$: jarak antara data pada cluster 1 dengan data pada cluster 2

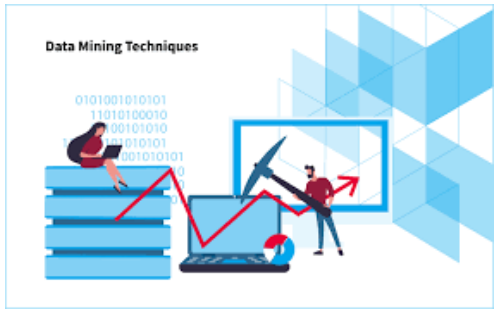



SINGLE LINKAGE

Pengukuran (dis)similarity atau jarak antar cluster dilakukan dengan mengukur terlebih dahulu jarak antar tiap observasi dari cluster yang berbeda pairwise distances. Jarak paling kecil (minimum distance) akan menjadi ukuran (dis)similarity antar cluster. Dendrogram akan terbentuk dari cluster-cluster yang memiliki (dis)similarity paling kecil. Hal ini membuat dendrogram yang terbentuk menjadi lebih “loose” atau berdekatan antar clusternya.
Berikut formula jarak antar cluster menggunakan single linkage:

$$d_{12} = \min_{ij} d(X_i, Y_j)$$

UNBIN



AVERAGE LINKAGE

Pengukuran (dis)similarity atau jarak antar cluster dilakukan dengan mengukur terlebih dahulu jarak antar tiap observasi dari cluster yang berbeda pairwise distances. Kemudian, dihitung rata-rata jarak dari pairwise distance tersebut dan nilai tersebut akan menjadi ukuran (dis)similarity antar cluster. Dendrogram akan terbentuk dari cluster-cluster yang memiliki (dis)similarity paling kecil. Umumnya metode ini akan menghasilkan cluster yang tidak terlalu “loose” maupun “compact”.

Berikut formula jarak antar cluster menggunakan average linkage:

$$d_{12} = \frac{1}{kl} \sum_{i=1}^{k} \sum_{j=1}^{l} d(X_i, Y_j)$$