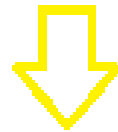


# DATA MINING

## ( 3 SKS )

## METODE DATA MINING



- Data mining model dibuat berdasarkan salah satu dari dua jenis pembelajaran: supervised dan unsupervised
- Fungsi **Supervised**: untuk memprediksi suatu nilai
- Fungsi **Unsupervised**: untuk mencari struktur intrinsik, relasi dalam suatu data yang tidak memerlukan class atau label sebelum dilakukan proses pembelajaran
- Contoh dari algoritma **Unsupervised** → K-means clustering, Apriori association rules
- Contoh dari algoritma **Supervised** → NaiveBayes untuk klasifikasi

# *Supervised dan Unsupervised Learning*

| <i>Supervised Learning</i>  | <i>Unsupervised Learning</i>   |
|---|--|
| Teknik ini melibatkan fase <b>pelatihan</b> : <b>data pelatihan historis</b> yang karakter-karakternya dipetakan ke hasil-hasil yang telah diketahui dan diolah dalam algoritma Data Mining | Teknik ini bergantung pada <b>penggunaan algoritma</b> yang mendeteksi semua pola, seperti <i>Associations</i> dan <i>Sequences</i> (dari kriteria penting-spesifik) dari data masukan |
| Melatih algoritma mengenali variabel dan nilai kunci yang nantinya akan digunakan sebagai dasar membuat perkiraan ketika ada data baru  | Pendekatan ini mengarah pada pembuatan banyak aturan ( <i>rules</i> ) yang mengkarakteristikkan penemuan <i>Associations</i> , <i>Clusters</i> dan <i>Segments</i> .                   |

# Perbedaan *Supervised* dan *Unsupervised learning*:

| <i>Supervised Learning</i>  | <i>Unsupervised Learning</i>  |
|---|---|
| <b>KEGUNAAN</b>   |   |
| Mengumpulkan/memproduksi <i>output data</i> dari pengalaman yang sudah pernah terjadi   | Digunakan untuk menemukan seluruh pola yang tidak dikenal dalam data  |
| Mirip memori manusia: Bisa mengingat nama ketika sudah pernah berkenalan atau bertemu   | Contoh penerapan yang sering digunakan dalam kehidupan sehari-hari adalah prediksi waktu pada peta digital (rute-waktu tempuh)                                  |
| <b>PROSES KERJA</b>   |   |
| Mendapatkan variabel data <i>input</i> dan <i>output</i>  | Hanya mendapatkan data <i>input</i>   |
| Dapat mengumpulkan atau memproduksi <i>output data</i> dari pengalaman masa lalu  | <b>Tidak menghasilkan <i>output data</i></b> (tidak dirancang untuk “belajar” dari pengalaman masa lalu)  |
| <b>PROSES BELAJAR</b>   |   |
| Algoritma komputer melakukan pembelajaran secara <i>offline</i> sebelum menghadapi data → Komputer “dibekali” sejumlah materi tertentu agar nanti dapat mengenali data dengan mudah | algoritma komputer mempelajari data secara <i>real-time</i> → Ketika komputer berhadapan dengan data, pada saat itu juga, komputer baru belajar mengenali data. |

# *Kasus pada Supervised dan Unsupervised learning:*

| <i>Supervised Learning</i>  | <i>Unsupervised Learning</i>  |
|---|---|
| <b>Banyak diterapkan dalam kasus:</b>   |   |
| <ul style="list-style-type: none"><li>• <b>Object recognition:</b> Algoritma supervised learning dapat digunakan untuk menemukan, mengisolasi, dan mengkategorikan objek dari video atau gambar, menjadikannya berguna ketika diterapkan pada berbagai teknik komputer vision dan analisis citra.</li><li>• <b>Predictive analysis:</b> Algoritma supervised learning sangat populer digunakan untuk keperluan ini. Menggunakan data-data kejadian masa lalu, teknik supervised learning digunakan untuk memprediksi kondisi atau trend di masa depan.</li><li>• <b>Sentiment analysis:</b> Dengan algoritma supervised learning, dapat mengekstrak dan mengklasifikasikan informasi penting dari data termasuk mendeteksi “emosi” manusia. Proses ini sangat berguna, misal untuk mengetahui persepsi konsumen terhadap produk tertentu, melalui sentiment analysis pada kolom review produk di sebuah toko online</li></ul> | <ul style="list-style-type: none"><li>• <b>Mesin Rekomendasi.</b> Menggunakan data sebelumnya, unsupervised learning dapat membantu menemukan tren data yang dapat digunakan untuk memberikan rekomendasi produk, sehingga konsumen dapat tertarik untuk melakukan pembelian kembali.</li><li>• <b>Segmentasi pasar/ konsumen.</b> Unsupervised learning dapat digunakan untuk membantu mendefinisikan persona konsumen. Proses ini membuat lebih mudah untuk memahami ciri-ciri umum dan kebiasaan pembelian oleh konsumen. Melalui proses ini, penyedia produk/ jasa dapat melakukan evaluasi strategi pemasaran yang tepat, seperti kapan waktu pemberian diskon yang paling menguntungkan.</li><li>• <b>Deteksi anomali.</b> Unsupervised learning dapat digunakan untuk menyisir data dalam jumlah besar dan menemukan titik data yang “berbeda” atau “aneh” dalam kumpulan data. Deteksi anomali ini dapat bermanfaat untuk menemukan kemungkinan adanya kesalahan manusia atau kerusakan alat rekam yang menyebabkan datanya jauh berbeda dengan yang lain</li></ul> |

## *Algoritma pada Supervised dan Unsupervised learning:*

| <i>Supervised Learning</i>  | <i>Unsupervised Learning</i>  |
|---|---|
| <p><b>Beberapa algoritma yang termasuk dalam supervised learning:</b></p> <ul style="list-style-type: none"><li>• Linear regression</li><li>• Decision Tree dan Random Forest</li><li>• Naive Bayes Classifier</li><li>• Nearest Neighbor Classifier</li><li>• Artificial Neural Network</li><li>• Support Vector Machine</li></ul> | <p><b>Beberapa contoh algoritma unsupervised learning:</b></p> <ul style="list-style-type: none"><li>• K-means clustering</li><li>• Hierarchical clustering</li><li>• DBSCAN</li><li>• Association Rule</li><li>• Apriori algorithm</li></ul> |

## KELEBIHAN

### *Supervised Learning*

Kelebihan/ keunggulan:

- Output hasil model sesuai dengan input di dataset training, sehingga kita bisa mengatur kelas output (misal, jumlah kelas) dengan mudah
- Hasil yang dihasilkan lebih akurat dan andal dibandingkan dengan hasil yang dihasilkan unsupervised learning
- Secara logika lebih mudah dipahami dan dijelaskan prosesnya
- Sangat berguna untuk melakukan analisis prediksi



Universitas Binaniaga Indonesia

### *Unsupervised Learning*

Kelebihan/ keunggulan:

- Tidak perlu label data, sehingga menghindari proses melabel data membutuhkan effort waktu dan tenaga yang tinggi,
- Proses pelabelan dapat dilakukan saat data sudah terklaster, sehingga proses labeling lebih cepat
- Sangat bermanfaat jika digunakan untuk memahami data yang masih raw (exploratori analysis) atau mencari pola di dalam data.
- Tidak diperlukan atau sedikit diperlukan pengetahuan sebelumnya tentang data kita
- Peluang kesalahan manusia diminimalkan.
- Relatif mudah dan cepat untuk dilaksanakan.
- Melalui dimensionality reduction, data yang kompleks dapat direduksi dimensionalitasnya.

## KELEMAHAN

### *Supervised Learning*

Kelemahan/ kekurangan/ tantangan:

- Memerlukan tingkat keahlian tertentu untuk menyusun model secara akurat.
- Proses training bisa sangat memakan waktu.
- Hasilnya sangat bergantung pada dataset training.
- Kesalahan dalam proses sampling akan sangat berpengaruh pada hasil model
- Dataset input dengan label yang salah akan memberikan hasil yang salah, meskipun model memiliki akurasi tinggi
- Tidak dapat mengelompokkan atau mengklasifikasikan data sendiri.
- Tidak dapat memberikan informasi yang belum diketahui dari data

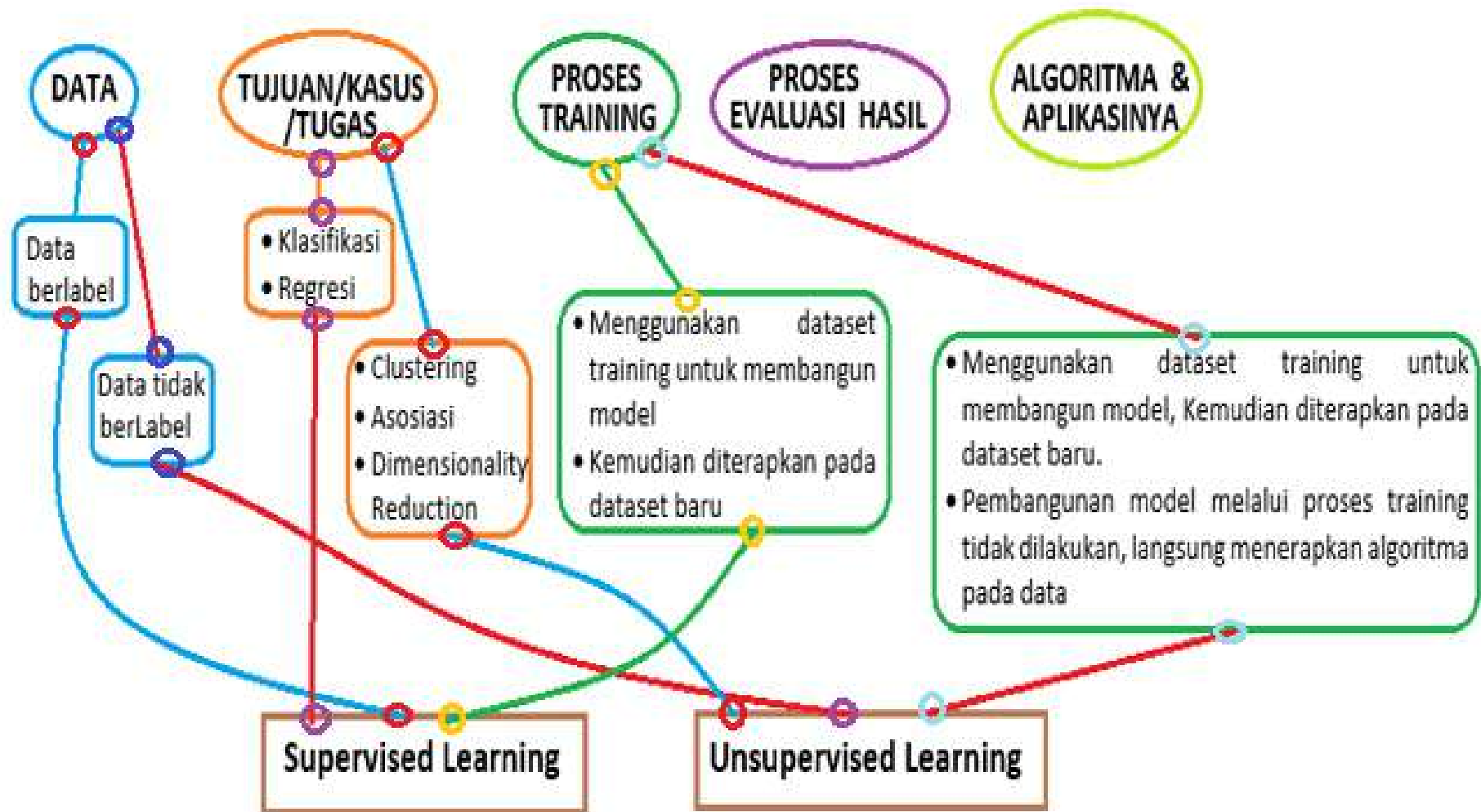
### *Unsupervised Learning*

Kelemahan/ kekurangan/ tantangan:

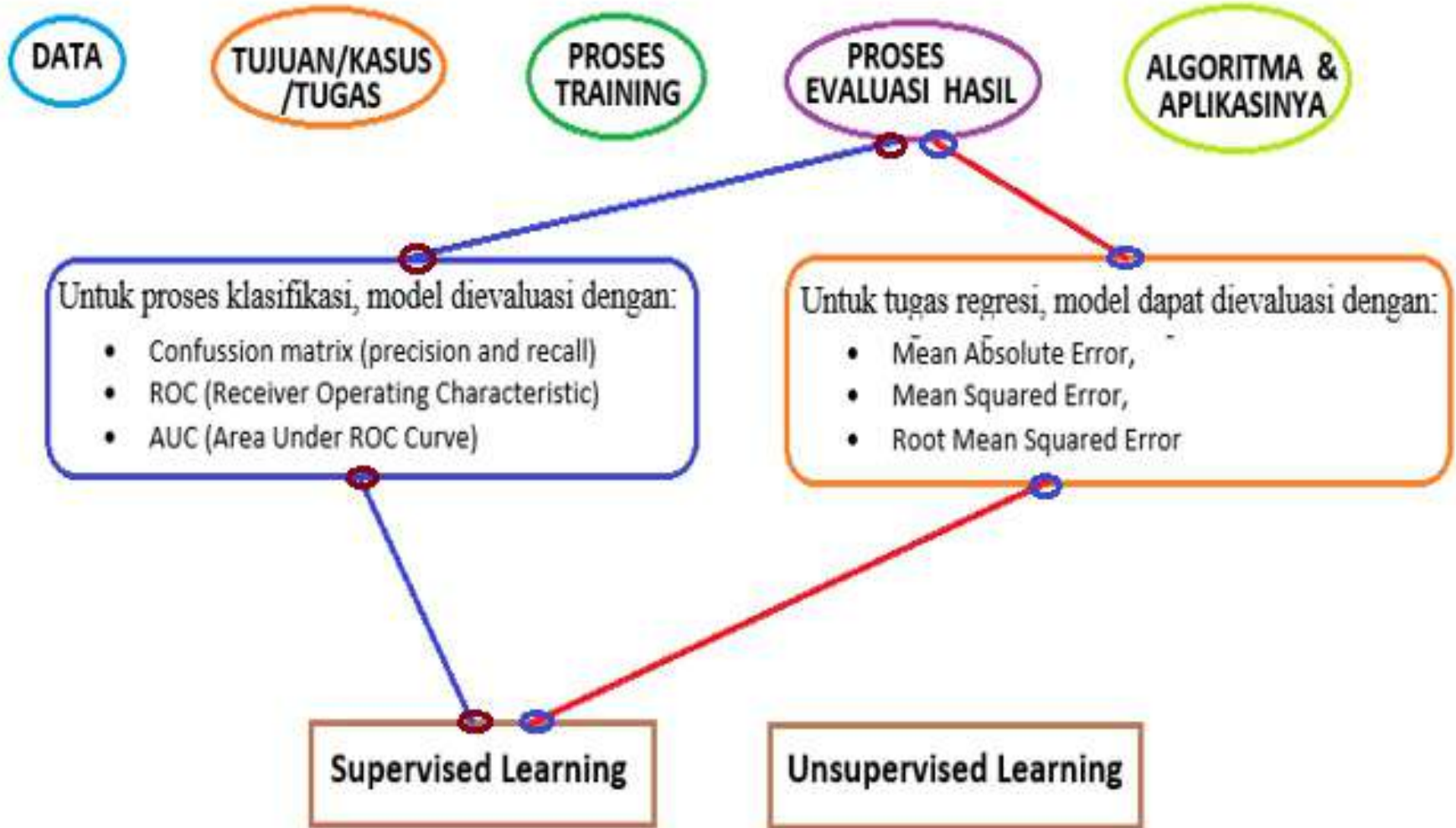
- Kompleksitas komputasi karena volume data pelatihan yang tinggi
- Risiko hasil yang tidak akurat lebih tinggi
- Intervensi manusia dibutuhkan untuk memvalidasi variabel keluaran
- Kurangnya transparansi dalam proses pengelompokan data



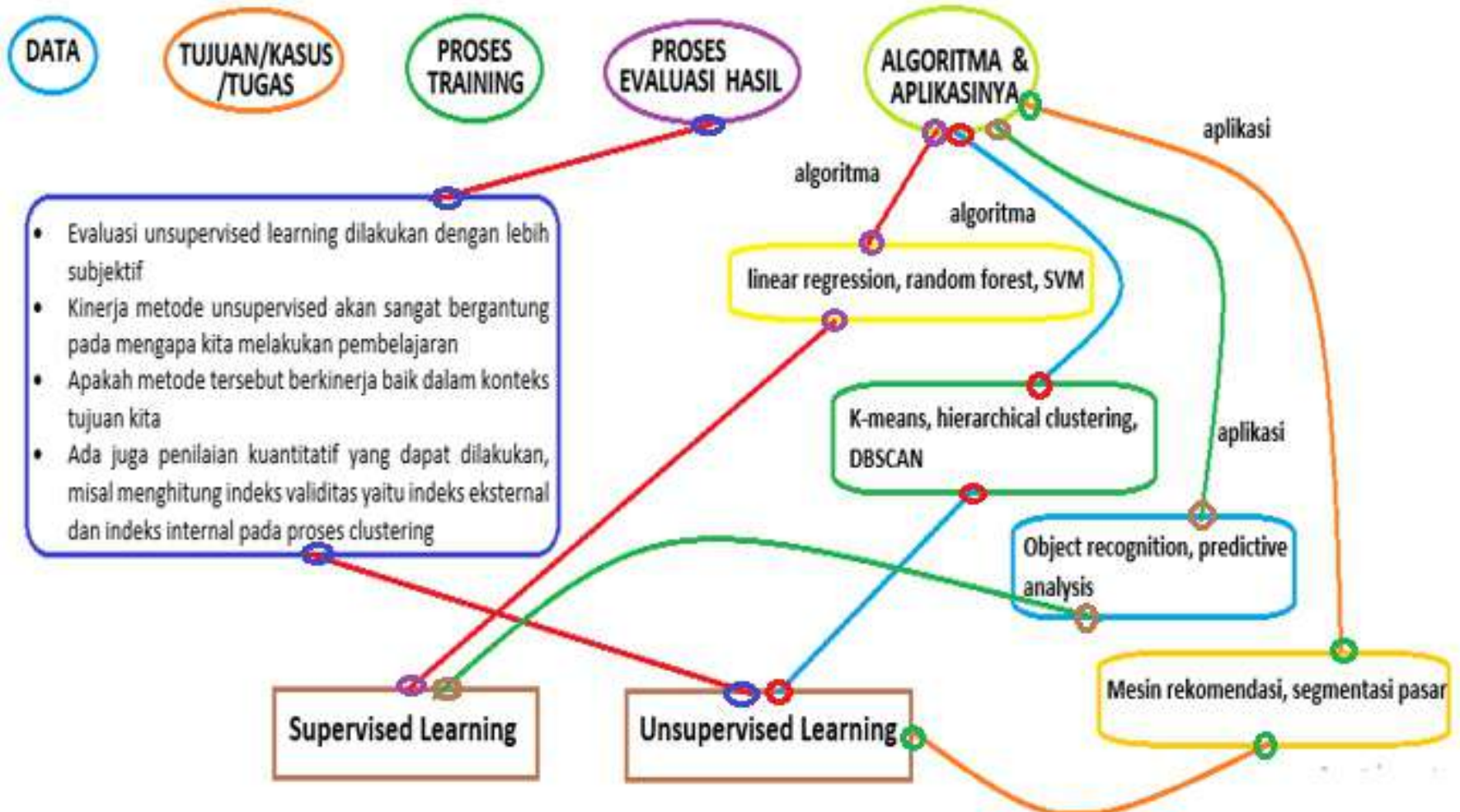
# Kapan menggunakan *Supervised / Unsupervised learning??*



# Kapan menggunakan *Supervised / Unsupervised learning??*



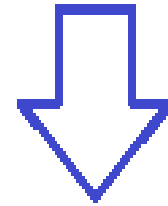
# Kapan menggunakan *Supervised / Unsupervised learning*??



# Rangkuman perbedaan Supervised & Unsupervised

| Aspek pembeda         | Supervised                              | Unsupervised                                   |
|-----------------------|---|--|
| Label data            | Data berlabel                           | Data tanpa label                               |
| Tujuan                | Klasifikasi, prediksi, regresi          | Klastering, asosiasi, dimensionality reduction |
| Proses training model | Ada                                     | Tidak ada                                      |
| Proses evaluasi       | Menggunakan test data                   | Dievaluasi secara subjektif,                   |
| Contoh algoritma      | linear regression, random forest, SVM   | K-means, hierarchical clustering, DBSCAN       |
| Contoh aplikasi       | Object recognition, predictive analysis | Mesin rekomendasi, segmentasi pasar            |

Klasifikasi Metode data mining berdasarkan fungsi yang dilakukan/ berdasarkan jenis aplikasi yang menggunakannya:



- **Klasifikasi (supervised)**
- **Clustering (unsupervised)**
- **Association Rules (unsupervised)**
- **Attribute Importance (supervised)**

## KLASIFIKASI (SUPERVISED)

- Pada klasifikasi, kita memiliki sejumlah kasus (sampel data) dan ingin **memprediksi beberapa class** yang ada pada sampel data tersebut
- Tiap instan data berisi banyak atribut, dimana masing-masing atribut memiliki satu dari beberapa kemungkinan nilai
- Hanya satu atribut diantara banyak atribut tersebut yang disebut dengan **atribut target**, sedangkan atribut yang lain disebut sebagai **atribut prediktor**
- Tiap kemungkinan nilai yang dimiliki oleh atribut target menunjukkan class yang diprediksi berdasarkan nilai-nilai dari atribut prediktor
- Klasifikasi digunakan untuk segmentasi customer, pemodelan bisnis, analisa kartu kredit, dan banyak aplikasi yang lain
- Sebagai contoh, perusahaan kartu kredit ingin memprediksi customer berdasarkan tipe pembayaran

## CLUSTERING (UNSUPERVISED)

- **Clustering** adalah teknik yang berguna untuk mengeksplorasi data
- Digunakan pada saat banyak kasus dan tidak memiliki pengelompokan secara alami
- Dalam hal ini algoritma data mining dapat digunakan untuk mencari pengelompokan yang ada pada data
- **Analisa Clustering** mengidentifikasi cluster yang ada pada data
- **Cluster** adalah kumpulan obyek data yang mirip satu sama lain
- Metode clustering yang bagus menghasilkan cluster yang berkualitas untuk memastikan kesamaan pada data-data yang ada dalam satu cluster
- Clustering model berbeda dari model prediktif dikarenakan pada clustering tidak perlu ada atribut target
- Clustering yang diorganisasi ke dalam struktur hirarkikal akan mendefinisikan taksonomi dari data.
- Dalam ODM (Oracle9 i Data Mining), suatu cluster dikarakterisasi oleh centroid, attribute histograms, dan clustering model hierarchical tree
- ODM membentuk hierarchical clustering dengan menggunakan versi perbaikan dari algoritma k-means dan O-Cluster

## ASSOCIATION RULES (UNSUPERVISED)

- **Fungsi Association Rules** seringkali disebut dengan "market basket analysis", yang digunakan untuk menemukan relasi atau korelasi diantara himpunan item2
- Fungsi ini paling banyak digunakan untuk menganalisa data dalam rangka keperluan strategi pemasaran, desain katalog, dan proses pembuatan keputusan bisnis
- **Tipe association rule** bisa dinyatakan sebagai misal : "70% dari orang-orang yang membeli mie, juice dan saus akan membeli juga roti tawar".
- Aturan asosiasi mengcapture item atau kejadian dalam data berukuran besar yang berisi data transaksi
- Dengan kemajuan teknologi, data penjualan dapat disimpan dalam jumlah besar yang disebut dengan "basket data."
- Aturan asosiasi yang didefinisikan pada basket data, digunakan untuk keperluan promosi, desain katalog, segmentasi customer dan target pemasaran.




## ATTRIBUTE IMPORTANCE (SUPERVISED)

- *Attribute Importance*, disebut juga dengan *feature selection*, menyediakan solusi otomatis untuk meningkatkan kecepatan dan akurasi dari model klasifikasi yang dibangun pada table data yang memiliki jumlah atribut yang sangat banyak
- *Attribute Importance* meranking atribut prediktif dengan melakukan eliminasi nilai yang *redundant*, *tidak relevant* atau *tidak informative* dan mengidentifikasi atribut predictor yang banyak paling berpengaruh dalam pengambilan keputusan.
- Dengan menggunakan atribut yang lebih sedikit akan mereduksi waktu untuk membangun suatu model, juga dapat meningkatkan akurasi dari kemampuan prediksi
- Jika terlalu banyak atribut yang dilibatkan maka akan banyak pula noise yang terlibat yang akan berpengaruh terhadap model karena dapat menurunkan performansi dan akurasi

## *Supervised Learning*

**Beberapa algoritma yang termasuk dalam supervised learning:**

- Linear regression
-  **Decision Tree** dan Random Forest
- Naive Bayes Classifier
- Nearest Neighbor Classifier
- Artificial Neural Network
- Support Vector Machine

# DECISION TREE

# DECISION TREE

terdiri dari tiga elemen

## Node akar

Bagian diagram paling atas yang mewakili tujuan akhir atau keputusan besar yang akan dibuat

## Ranting

Bagian cabang yang berasal dari akar dan mewakili opsi yang berbeda atau rangkaian tindakan yang tersedia saat membuat keputusan tertentu

## Simpul daun

Bagian ini berada pada ujung cabang yang mewakili kemungkinan hasil untuk setiap

### Daun Persegi

Keputusan lain yang harus dibuat

### Daun Lingkaran

hasil yang tidak diketahui / kemungkinan perubahan peristiwa

## MANFAAT (4)

### Menawarkan kejelasan

- Memperjelas pilihan, risiko, tujuan dan keuntungan untuk setiap pilihan
- Dapat memetakan berbagai kemungkinan
- Dapat menentukan tindakan yang menghasilkan sukses tertinggi.

### Efisiensi

- Menyajikan informasi yang mudah dipahami  
Dengan cepat dapat melakukan analisis
- Membantu Pengambilan keputusan dengan efisien

### Kompatibilitas

Dapat dikolaborasikan dengan metodologi manajemen proyek lain

### Menghindari bias

- Menerima pendapat orang lain → sangat ber-resiko, karena dipengaruhi oleh pendapat pribadi (bias)
- *Decision Tree* memberikan pandangan yang seimbang karena didasarkan pada perhitungan risiko dan imbalan

## Beberapa Contoh penerapan *Decision Tree* dalam perusahaan:

### Mencari calon klien

Untuk menemukan calon klien, dapat digunakan data historis

### Menilai peluang atau prospek pertumbuhan

Data historis penjualan dapat digunakan untuk menghasilkan keputusan perubahan strategi bisnis yang dapat membantu pertumbuhan bisnis

### Menganalisa *credit scoring*

Dalam industri perbankan, *decision tree* juga dapat digunakan untuk memprediksi risiko kemungkinan peminjam gagal membayar pinjaman, dengan menerapkan pembuatan model prediktif menggunakan data masa lalu Klien

### Manajemen strategis

Dapat membantu dalam menentukan strategi yang tepat yang akan membantu perusahaan mencapai tujuan

Bidang lain di mana *decision tree* dapat diterapkan termasuk teknik, pendidikan, hukum, bisnis, kesehatan, keuangan, dll

## Langkah-langkah membuat *Decision Tree*:

1. Mulailah dengan menempatkan objektif atau keputusan besar di bagian atas  
→ menjadi bagian **akar** dari seluruh diagram.
2. Gambar **garis panah** untuk setiap tindakan yang mungkin dilakukan (dari akarnya)
  - Jika tindakan yang diambil melibatkan perhitungan nilai atau biaya, tuliskan biaya pada setiap tindakan serta kemungkinan untuk sukses
3. Pasang simpul daun di ujung cabang untuk menggambarkan apa hasil dari setiap pilihan tindakan
  - Jika keputusan lain harus dibuat, gambarlah simpul daun persegi.
  - Jika hasilnya tidak pasti gambarlah simpul daun melingkar (bulat)
4. Tentukan peluang keberhasilan setiap titik keputusan
  - Saat membuat *decision tree* penting untuk melakukan riset dengan melihat data atau evaluasi dari proyek sebelumnya sehingga dapat memprediksi kemungkinan sukses secara akurat
5. Evaluasi risiko vs *reward*.
  - Menghitung nilai yang diharapkan dari setiap keputusan di *decision tree* membantu kamu meminimalkan risiko dan meningkatkan kemungkinan mencapai hasil yang menguntungkan

## Tip membuat *decision tree* yang efektif:

- **Kode warna pohon.** Beri kode warna pada cabang dan simpul Anda untuk mengidentifikasi hasil dengan mudah. Penggunaan skema warna untuk membuatnya menarik secara visual
- **Gunakan simbol diagram alur.** Jika tujuan membuat *decision tree* untuk dibagikan dengan tim atau manajer, simbol diagram alur standar memastikan pohon mudah dipahami oleh banyak pembaca
- **Buat simbol dengan ukuran yang sama.** Ini akan membantu memberikan nilai yang sama pada masing-masing dan membuat pohon lebih mudah dibaca.
- **Gunakan template.** Ada banyak template online yang dapat digunakan untuk membuat pohon terlihat sederhana. Beberapa juga memiliki fungsi matematika jika menggunakan pohon untuk menangani data dan statistik.
- **Ketahui kapan harus menggunakan *decision tree*.** Pohon keputusan bekerja paling baik ketika memiliki tujuan khusus dan perlu melihat hasil untuk setiap pilihan yang dapat dibuat.



## Kelebihan & Kekurangan *Decision Tree*

### Kelebihan:

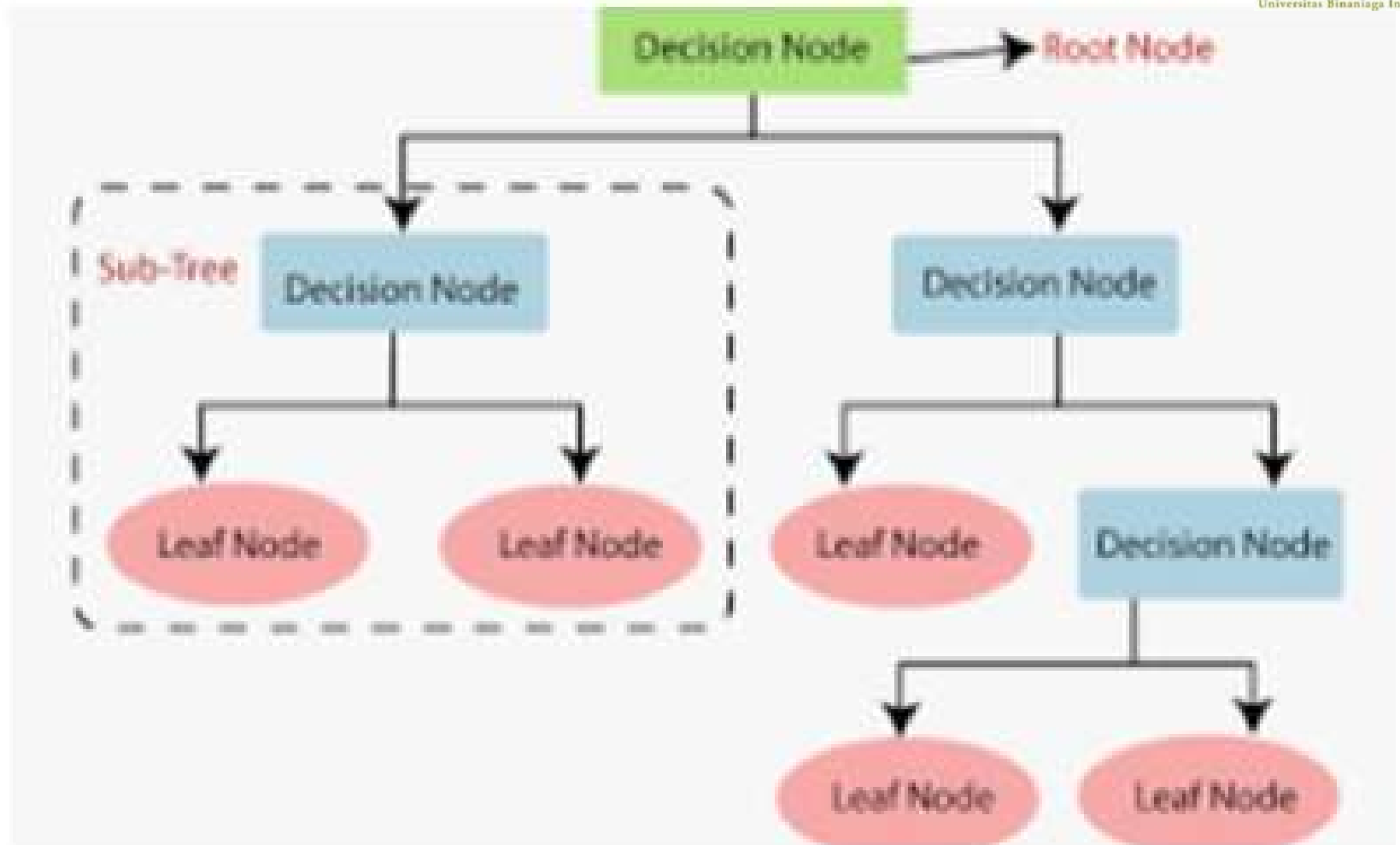
- Mudah dibaca dan ditafsirkan
- Mudah disiapkan
- Lebih sedikit pembersihan data yang diperlukan

### Kekurangan:

- Sifat tidak stabil
- Kurang efektif dalam memprediksi hasil dari variabel kontinu

## DECISION TREE:

Format Umum

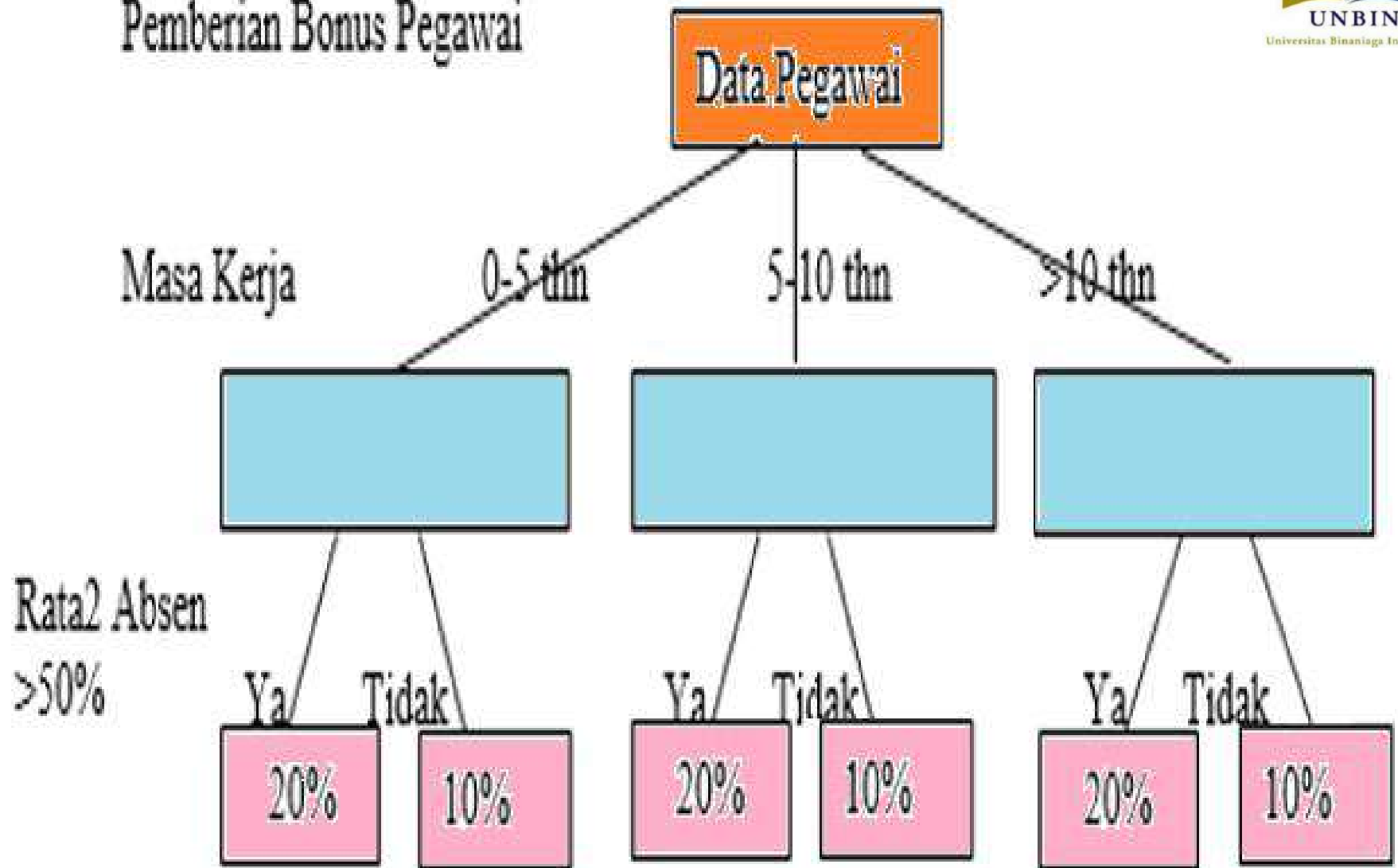


## Beberapa Contoh Decision Tree:



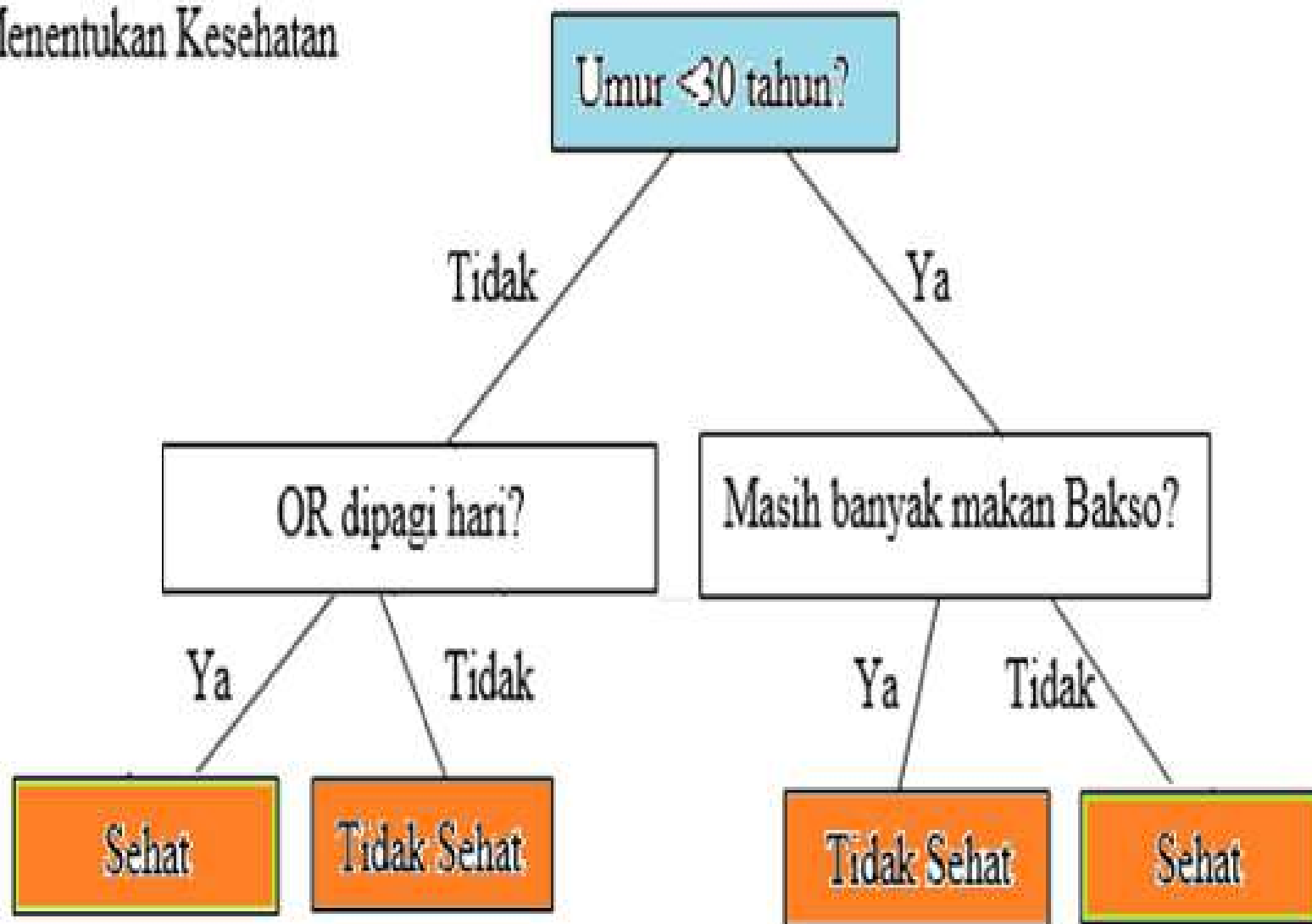
# DECISION TREE:

## Pemberian Bonus Pegawai



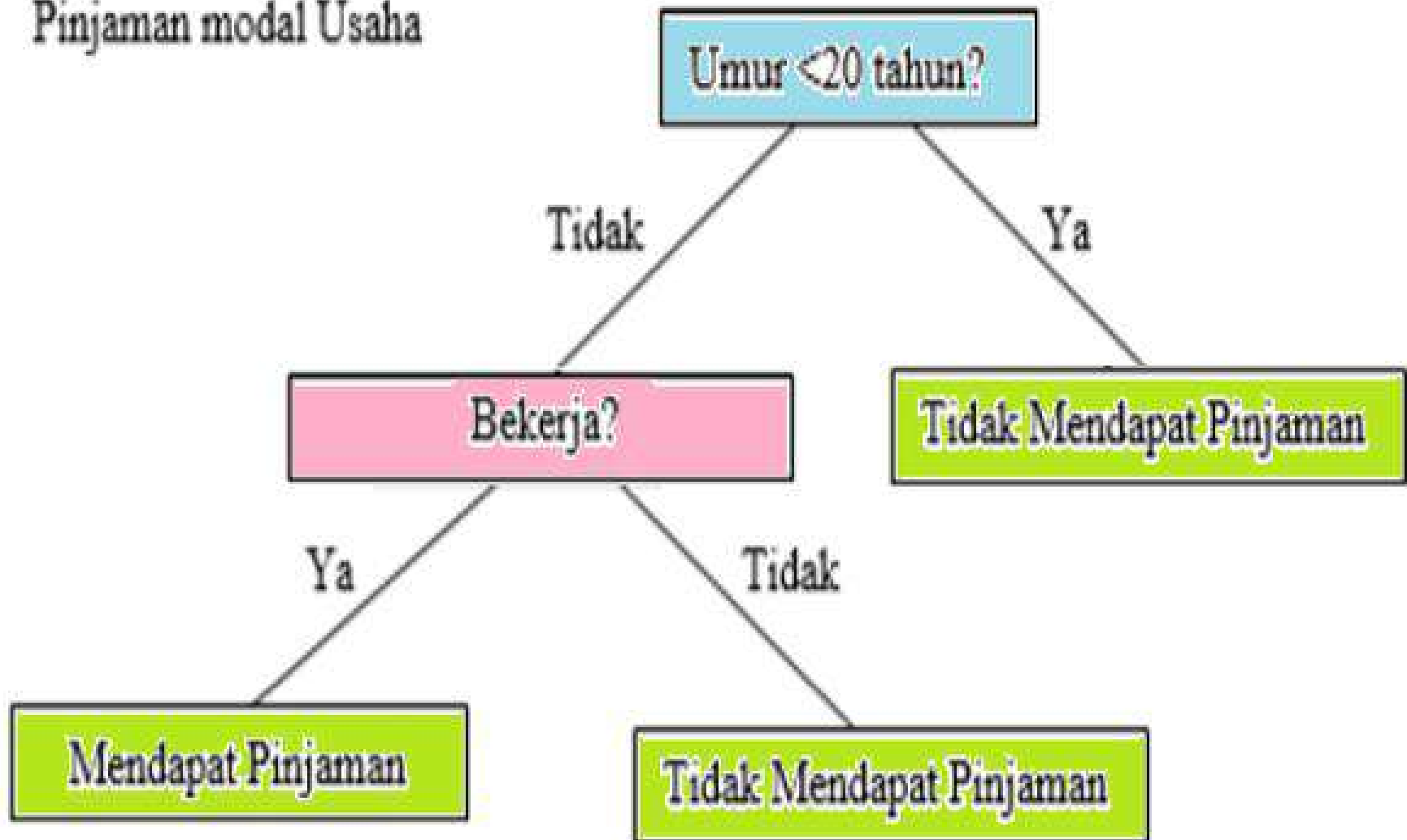
## DECISION TREE:

Menentukan Kesehatan



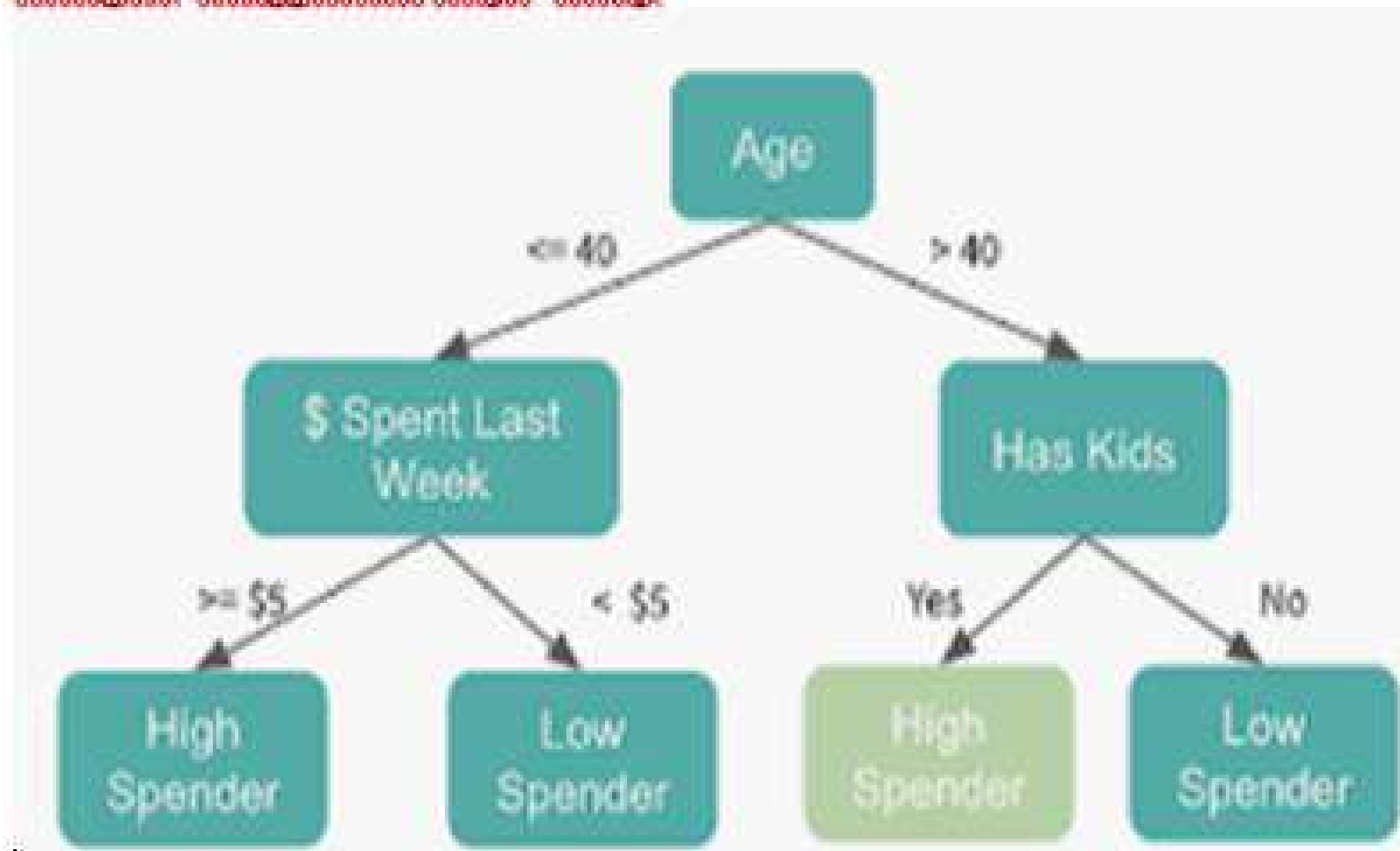
## DECISION TREE:

Pinjaman modal Usaha



## DECISION TREE:

Kategori Penggunaan biaya hidup



## **Latihan :**

Buatlah 3 Diagram Decision Tree yang menceritakan kondisi disekitar tempat tinggalmu !!

**Kumpulkan paling lambat Senin, 26/09/2022 pukul 21.00**