



DIGITAL TALENT SCHOLARSHIP 2019



Program Fresh Graduate Academy Digital Talent Scholarship 2019 | Machine Learning

Classification : Decision Tree

M. Ramli & M. Soleh





Bagian Satu

Apa itu Decision Tree?



Pengantar Deision Tree

- Apa itu *Decision Tree*?
- Bagaimana cara menggunakannya untuk melakukan klasifikasi?
- Bagaimana caranya untuk menumbuhkan Decision Tree kita sendiri?
- Mungkin beberapa pertanyaan tersebut muncul dalam benak kita ketika mendengar kata Decision Tree
- Materi ini akan menjawab semua pertanyaan tersebut.



Studi Kasus Decision Tree

Patient ID	Age	Sex	BP	Cholesterol	Drug
p1	Young	F	High	Normal	Drug A
p2	Young	F	High	High	Drug A
p3	Middle-age	F	Hiigh	Normal	Drug B
p4	Senior	F	Normal	Normal	Drug B
p5	Senior	M	Low	Normal	Drug B
p6	Senior	M	Low	High	Drug A
p7	Middle-age	M	Low	High	Drug B
p8	Young	F	Normal	Normal	Drug A
p9	Young	M	Low	Normal	Drug B
p10	Senior	M	Normal	Normal	Drug B
p11	Young	M	Normal	High	Drug B
p12	Middle-age	F	Normal	High	Drug B
p13	Middle-age	M	High	Normal	Drug B
p14	Senior	F	Normal	High	Drug A
p15	Middle-age	F	Low	Normal	?

- Bayangkan anda sebagai peneliti medis yang sedang melakukan observasi data pasien.
- Data pasien telah terkumpul, dimana pasien-pasien tersebut memiliki penyakit yang sama.
- Selama penanganan, setiap pasien harus menerima salah satu dari dua obat yang tersedia.
 - *Drug A*
 - *Drug B*
- **Tugas anda:** memberikan saran obat kepada pasien yang baru.

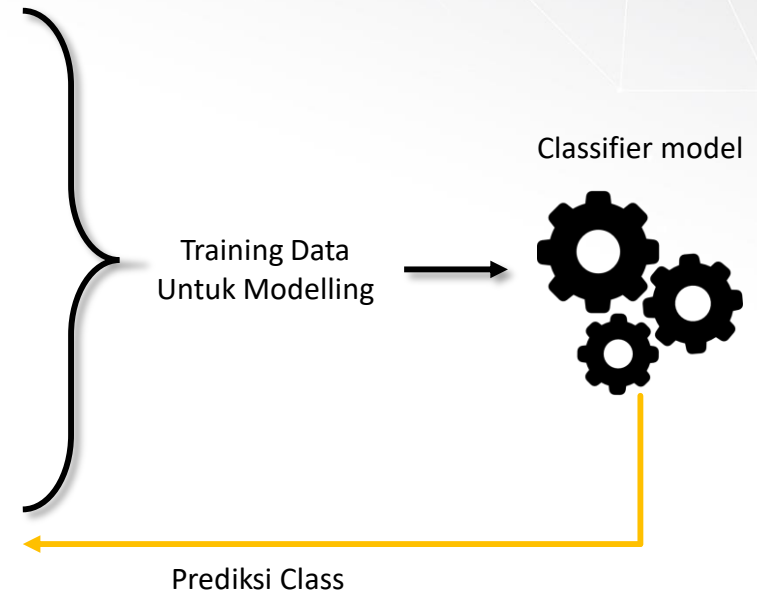
Studi Kasus Decision Tree

Patient ID	Age	Sex	BP	Cholesterol	Drug
p1	Young	F	High	Normal	Drug A
p2	Young	F	High	High	Drug A
p3	Middle-age	F	Hiigh	Normal	Drug B
p4	Senior	F	Normal	Normal	Drug B
p5	Senior	M	Low	Normal	Drug B
p6	Senior	M	Low	High	Drug A
p7	Middle-age	M	Low	High	Drug B
p8	Young	F	Normal	Normal	Drug A
p9	Young	M	Low	Normal	Drug B
p10	Senior	M	Normal	Normal	Drug B
p11	Young	M	Normal	High	Drug B
p12	Middle-age	F	Normal	High	Drug B
p13	Middle-age	M	High	Normal	Drug B
p14	Senior	F	Normal	High	Drug A
p15	Middle-age	F	Low	Normal	?

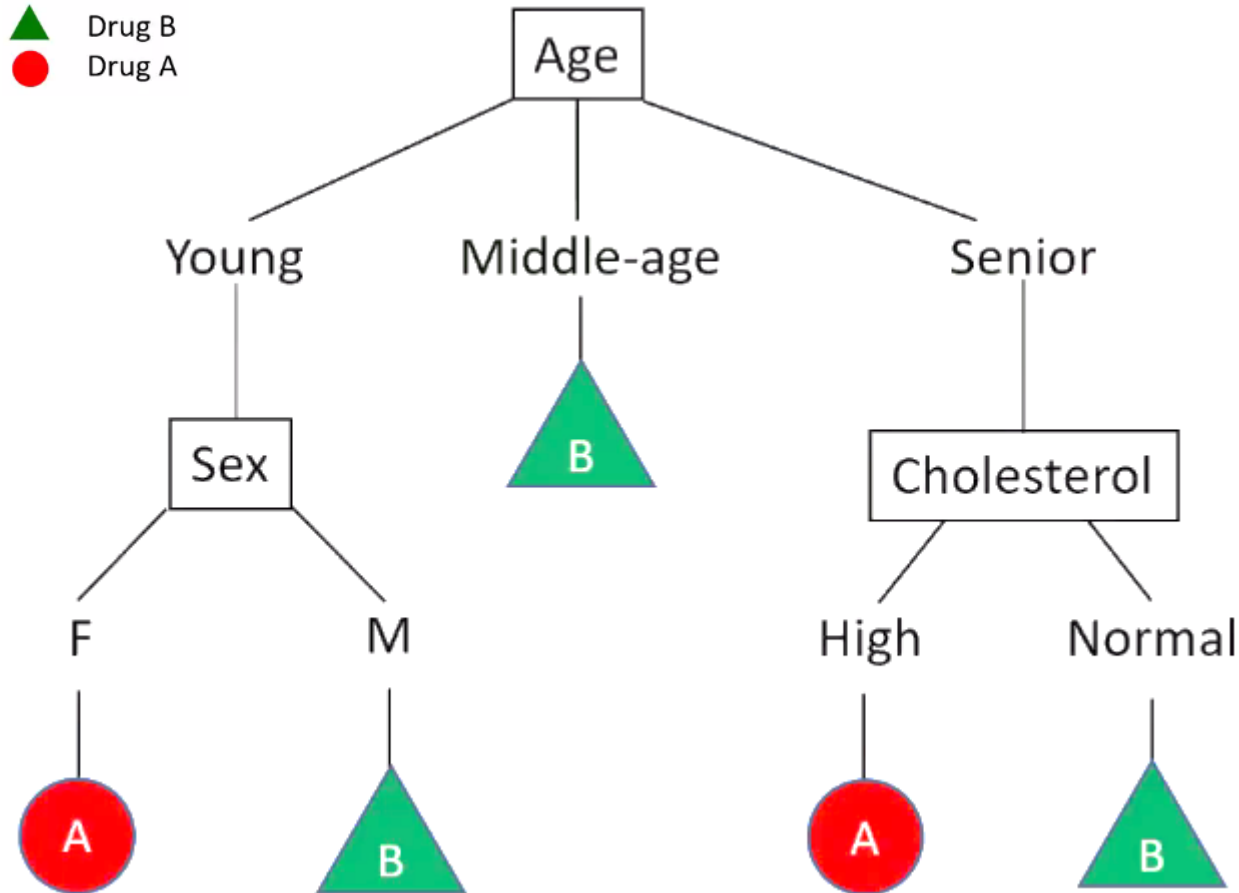
- **Fitur:**
 - Age
 - Sex
 - Blood Pressure (BP)
 - Cholesterol
- **Target/Class:**
 - Drug A atau Drug B

Studi Kasus Decision Tree

Patient ID	Age	Sex	BP	Cholesterol	Drug
p1	Young	F	High	Normal	Drug A
p2	Young	F	High	High	Drug A
p3	Middle-age	F	Hiigh	Normal	Drug B
p4	Senior	F	Normal	Normal	Drug B
p5	Senior	M	Low	Normal	Drug B
p6	Senior	M	Low	High	Drug A
p7	Middle-age	M	Low	High	Drug B
p8	Young	F	Normal	Normal	Drug A
p9	Young	M	Low	Normal	Drug B
p10	Senior	M	Normal	Normal	Drug B
p11	Young	M	Normal	High	Drug B
p12	Middle-age	F	Normal	High	Drug B
p13	Middle-age	M	High	Normal	Drug B
p14	Senior	F	Normal	High	Drug A
p15	Middle-age	F	Low	Normal	?

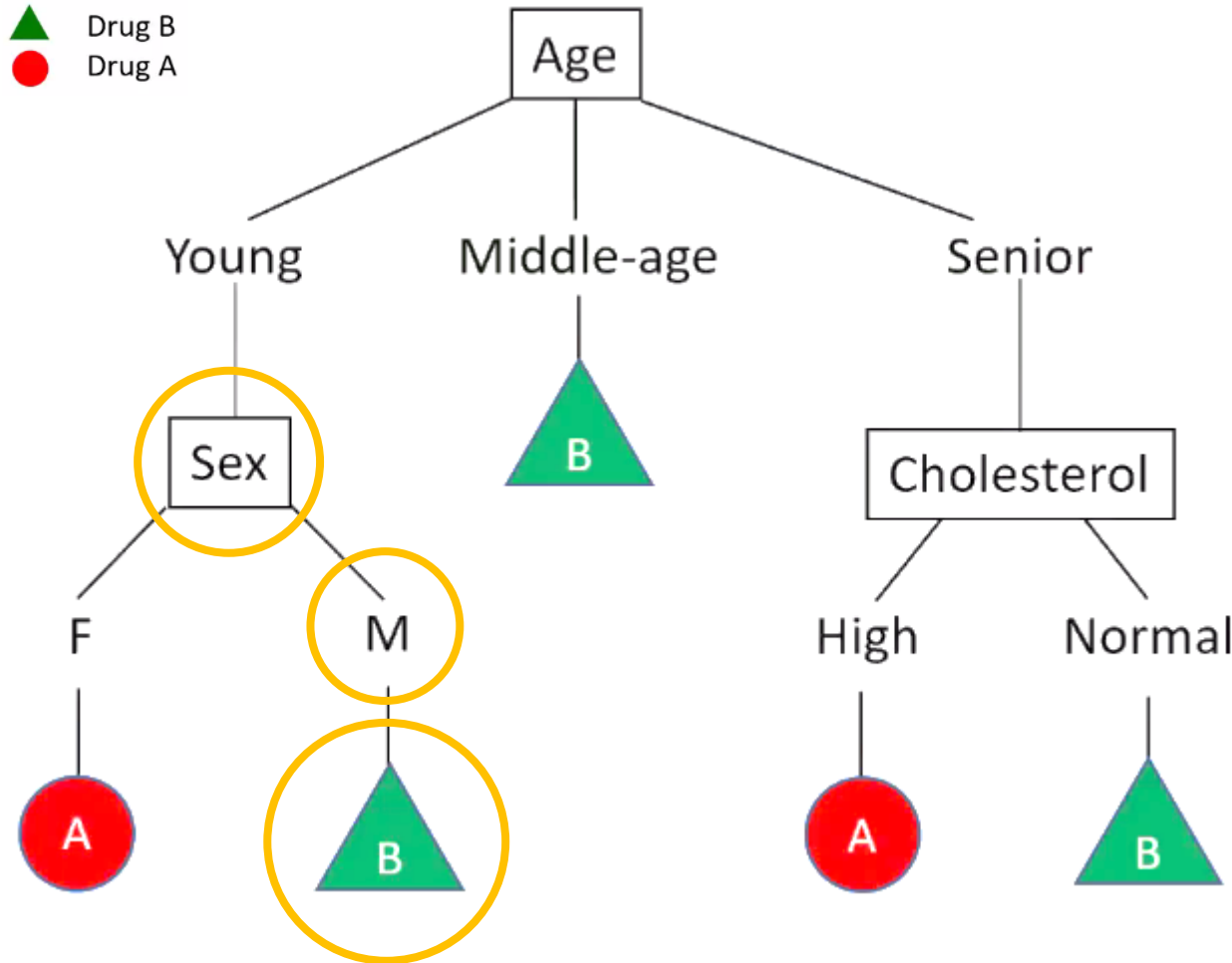


Bentuk Decision Tree



- Kita ingin melakukan klasifikasi pasien baru
- Keputusan obat yang akan diterimanya akan tergantung dari bentuk pohon.
- Tree memiliki **node akar, Age**
 - Age dianggap sebagai variable paling berpengaruh

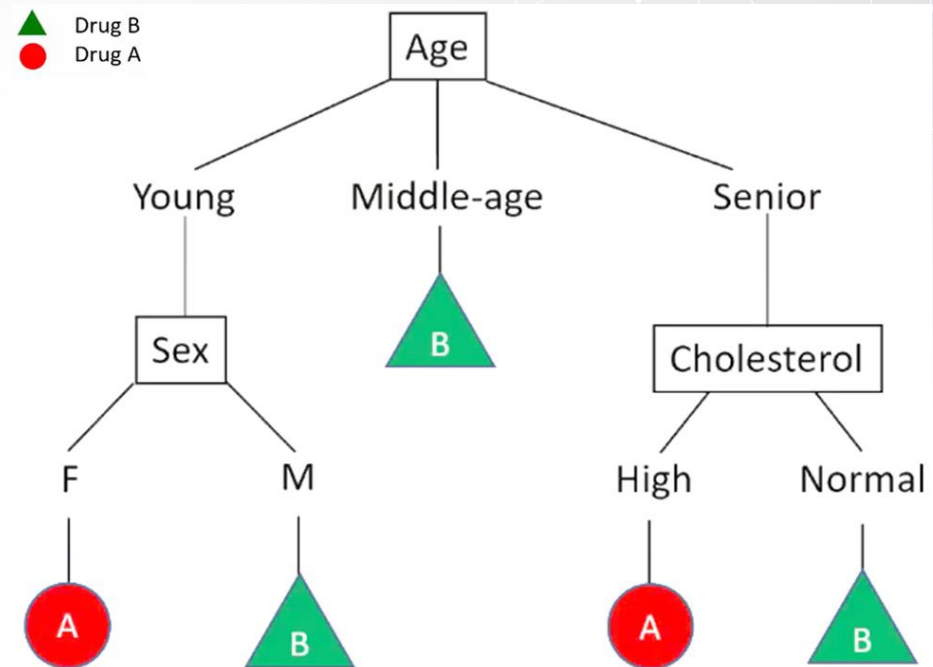
Bentuk Decision Tree



- Setiap *Node* mendeskripsikan sebuah tindakan observasi
 - Apa gender/sex pasien tsb.
- Setiap *Branch* mendeskripsikan hasil observasi
 - Pasien ternyata seorang pria
- Setiap *Leaf* mendeksripsikan hasil klasifikasi.

Bagaimana Bentuk Decision Tree

1. Pilih salah satu atribut dari dataset
2. Hitung nilai signifikansi atribut dalam pemecahan data
 - Nilai signifikansi mendeskripsikan seberapa besar pengaruh atribut tersebut dalam sebaran data
 - Kalkulasi ini akan kita bahas di bagian selanjutnya
3. Pecah data berdasarkan atribut yang memiliki nilai signifikansi terbesar
4. Kembali ke langkah 1



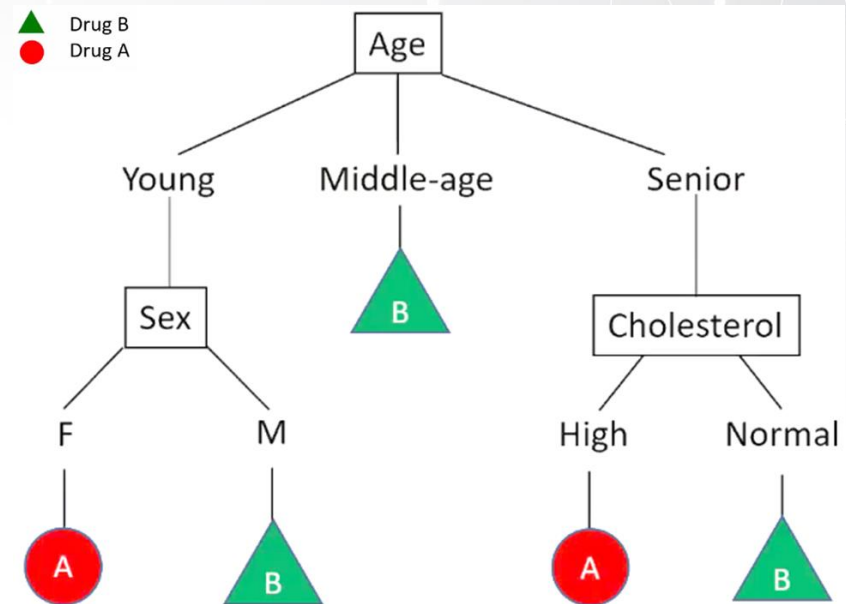


Bagian Dua

Algoritma Membangun Decision Tree

Membangun Decision Tree

Patient ID	Age	Sex	BP	Cholesterol	Drug
p1	Young	F	High	Normal	Drug A
p2	Young	F	High	High	Drug A
p3	Middle-age	F	Hiigh	Normal	Drug B
p4	Senior	F	Normal	Normal	Drug B
p5	Senior	M	Low	Normal	Drug B
p6	Senior	M	Low	High	Drug A
p7	Middle-age	M	Low	High	Drug B
p8	Young	F	Normal	Normal	Drug A
p9	Young	M	Low	Normal	Drug B
p10	Senior	M	Normal	Normal	Drug B
p11	Young	M	Normal	High	Drug B
p12	Middle-age	F	Normal	High	Drug B
p13	Middle-age	M	High	Normal	Drug B
p14	Senior	F	Normal	High	Drug A
p15	Middle-age	F	Low	Normal	?





Membangun Decision Tree

▲ Drug B
● Drug A



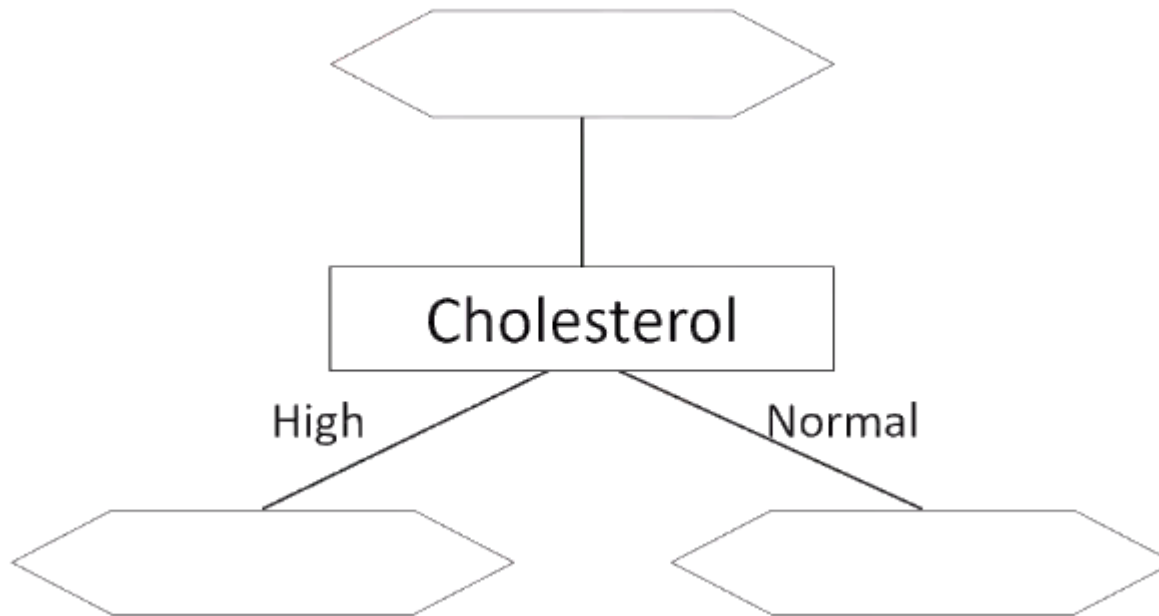
Atribut apa yang paling terbaik memisahkan data?

- Decision tree dibuat dengan menggunakan teknik recursive partitioning untuk klasifikasi data
- Berdasarkan dataset, kita memiliki 14 pasien dengan 7 diklasifikasikan Drug A, 7 lagi diklasifikasikan Drug B.
- Algoritma harus bisa memilih fitur/atribut yang paling baik dalam melakukan klasifikasi



Membangun Decision Tree

▲ Drug B
● Drug A



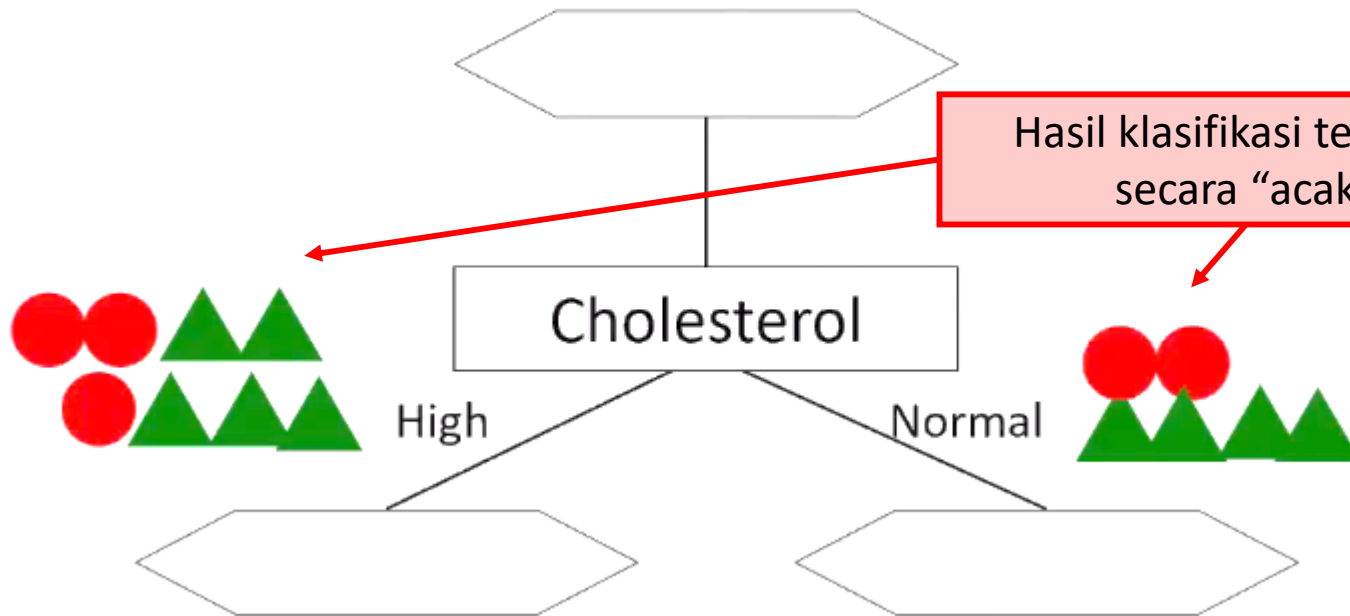
Apakah atribut
cholesterol terbaik?

- Untuk mengetahui mana atribut yang terbaik, yang kita lakukan adalah kita coba satu persatu.
- Pertama kita mulai dari Cholesterol.



Membangun Decision Tree

▲ Drug B
● Drug A



Apakah atribut
cholesterol terbaik?

- Klasifikasi yang dihasilkan jika kita memiliki Cholesterol sebagai atribut pertama adalah:
 - Untuk High: 3 buah Drug A; 4 buah Drug B
 - Untuk Normal: 2 buah Drug A, 4 buah Drug B

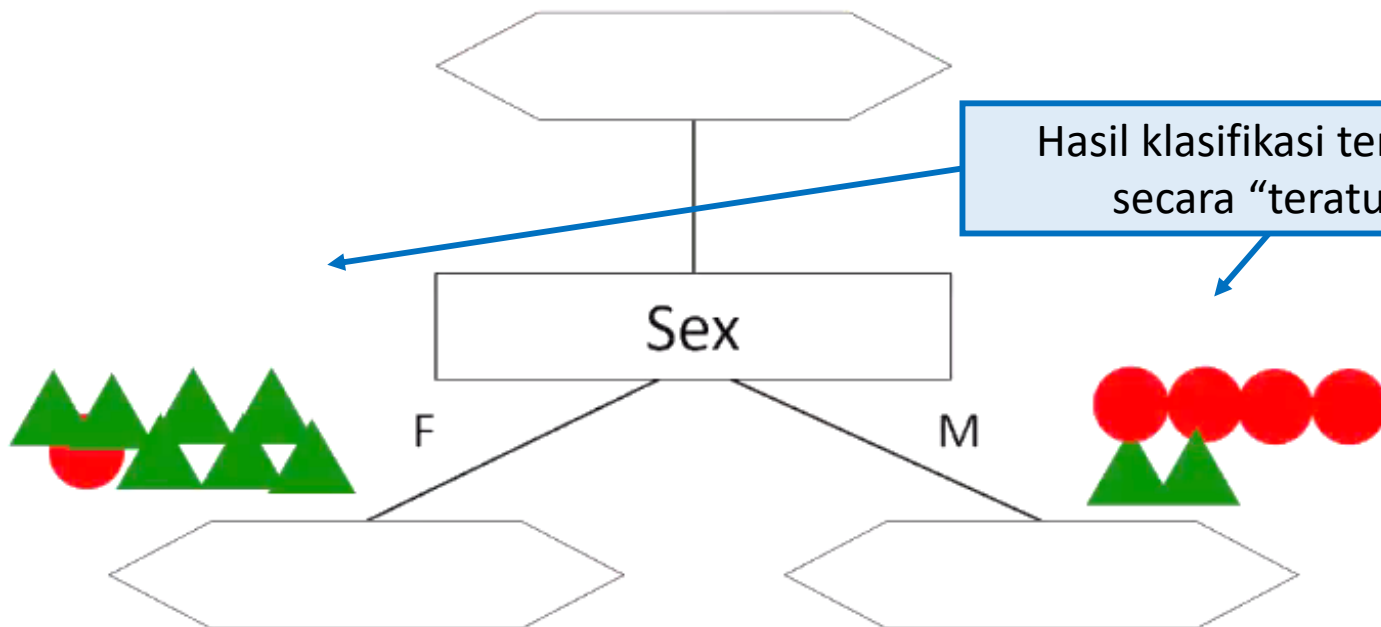


Membangun Decision Tree

Sekarang kita coba atribut yang lain.

Membangun Decision Tree

▲ Drug B
● Drug A



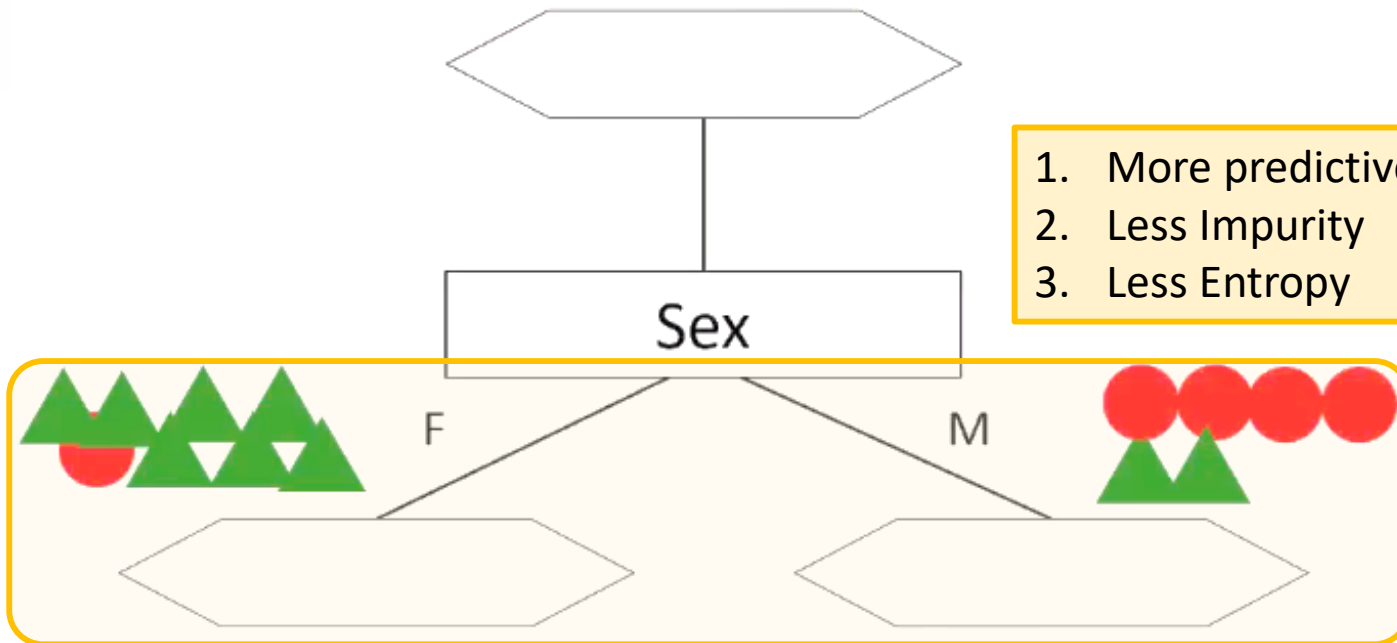
Apakah atribut sex yang terbaik?

- Klasifikasi yang dihasilkan jika kita memiliki Sex sebagai atribut pertama adalah:
 - Untuk Female: 1 buah Drug A; 7 buah Drug B
 - Untuk Male: 4 buah Drug A, 2 buah Drug B



Membangun Decision Tree

▲ Drug B
● Drug A



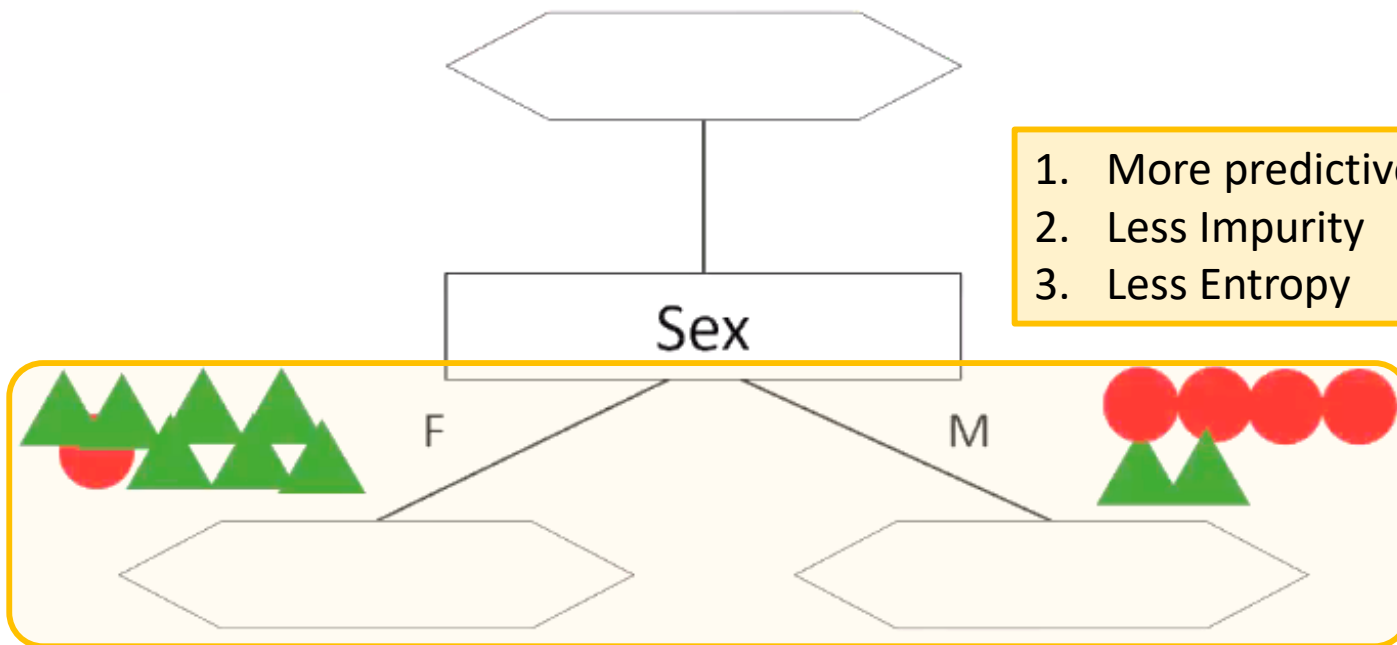
Apakah atribut sex
yang terbaik?

- More Predictive = Lebih dapat diprediksi klasifikasinya
- Less Impurity = Berkurang ketidak-murniannya
- Less Entropy = Berkurang keacakannya



Membangun Decision Tree

▲ Drug B
● Drug A



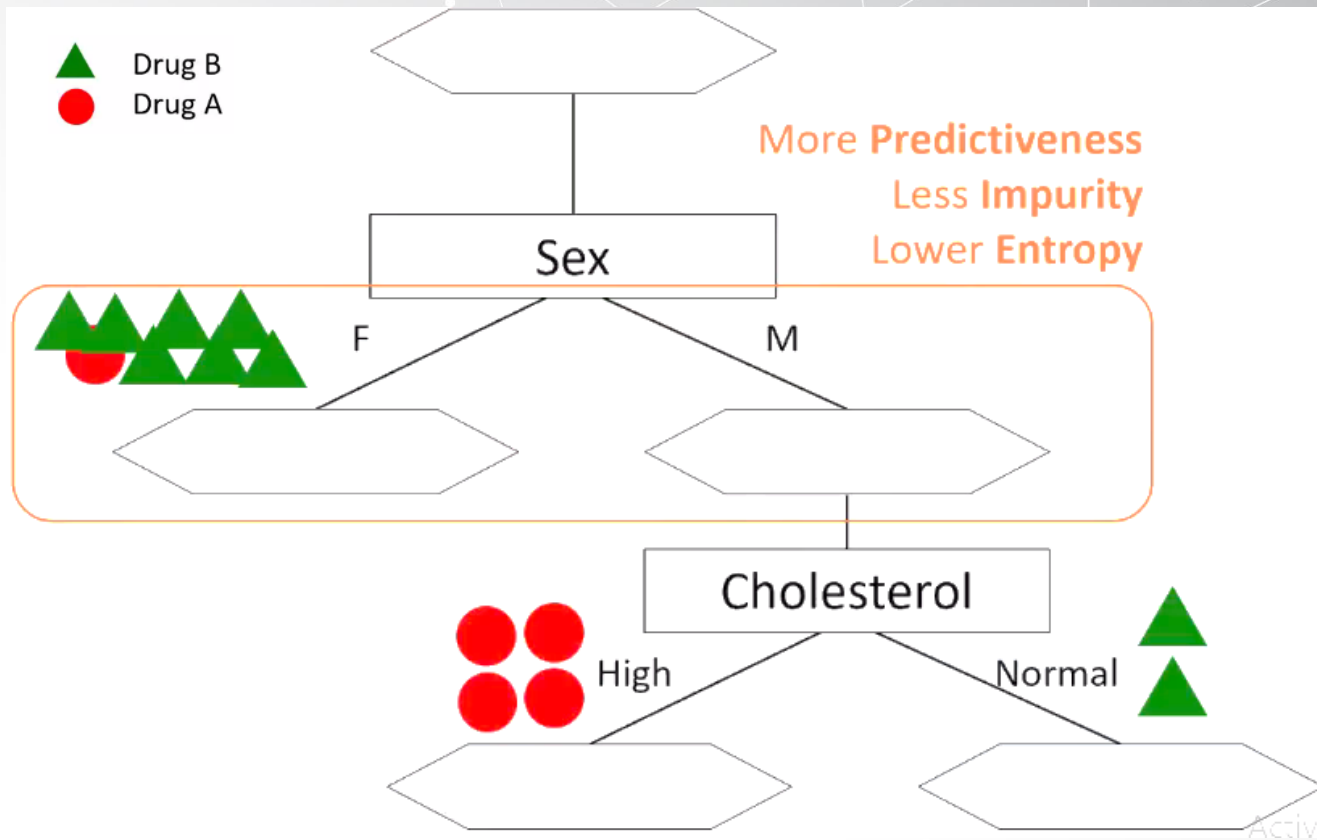
Apakah atribut sex yang terbaik?

- Secara kualitatif, atribut Sex dikatakan memiliki nilai signifikasni lebih banyak dibanding atribut Cholesterol



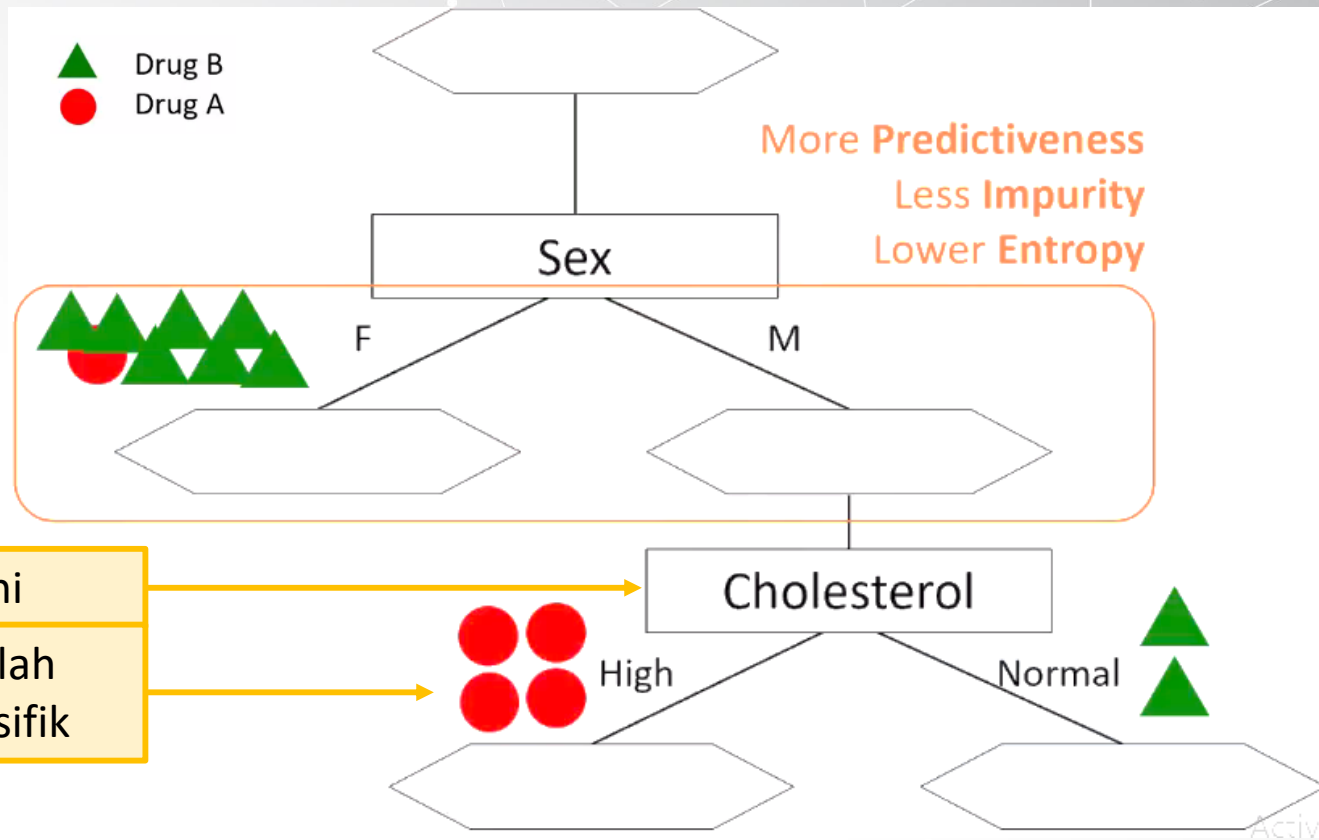
Membangun Decision Tree

Ayo melangkah lebih dalam lagi.



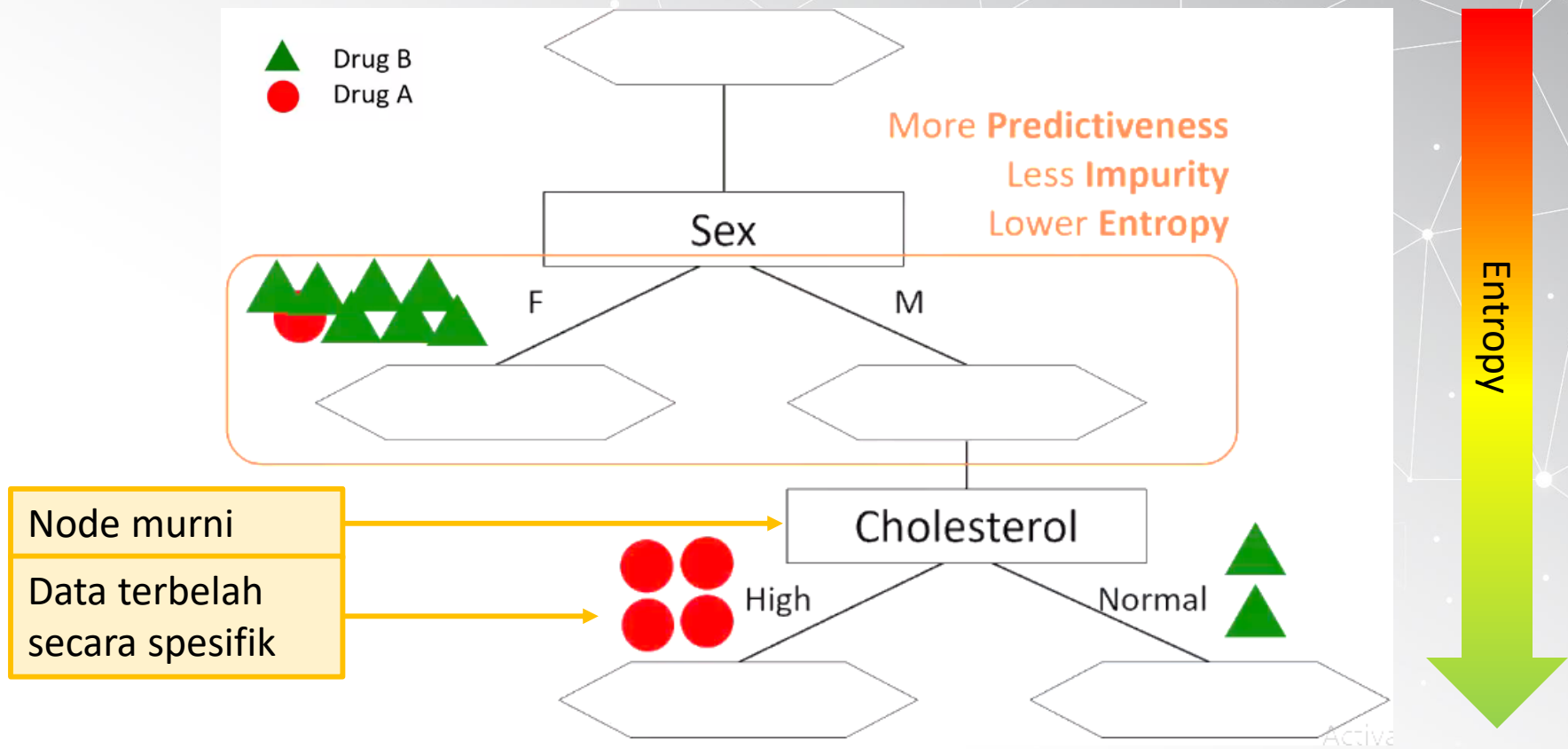
Kemurnian suatu node

- Jika kita tambahkan satu atribut lagi, Cholesterol, klasifikasi menjadi terbagi sangat baik.
- jika seorang pasien merupakan **Pria** dengan Cholesterol **Tinggi**, maka kita bisa sarankan ia menggunakan Drug A dengan tingkat *confident* yang sangat tinggi



Kemurnian suatu node

- Sebuah node dari Decision Tree dikatakan murni atau *pure* jika node tersebut membelah class secara spesifik di 100% kasus.



Kemurnian dan Entropy

- Semakin kebawah, Decision Tree seharusnya semakin kecil Entropynya
- Semakin teratur klasifikasinya.



Bagian Tiga

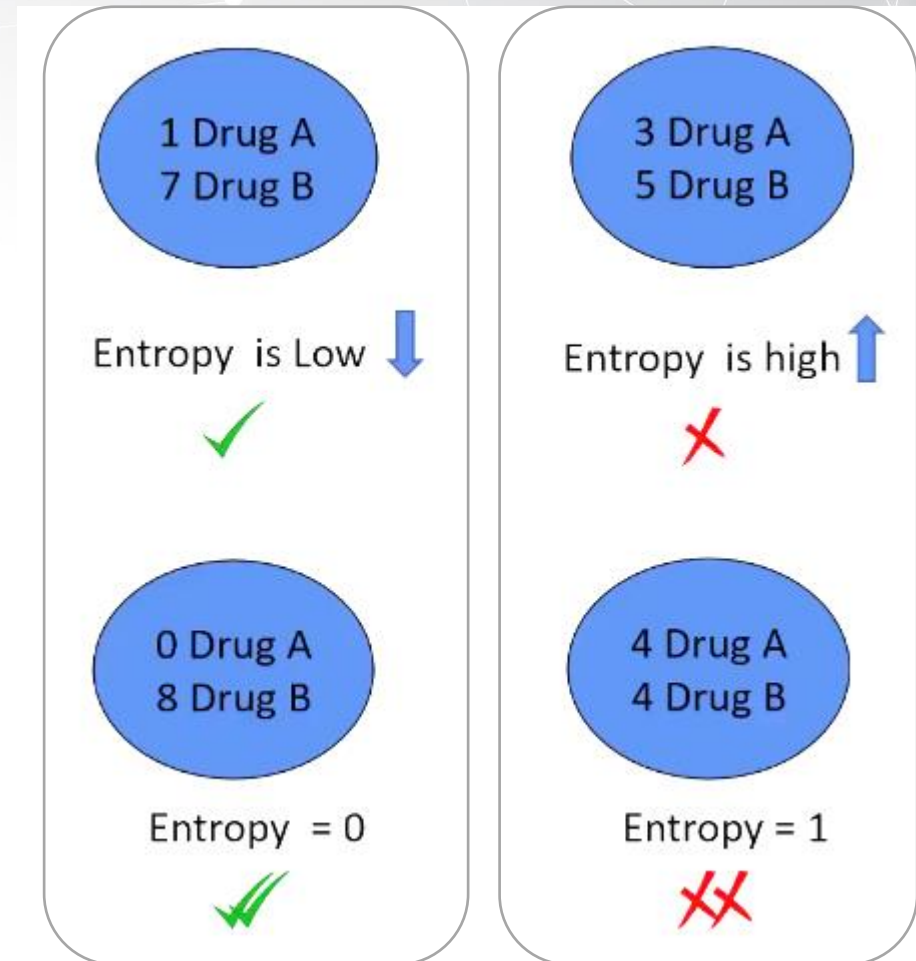
Pemahaman Entropy

Pemahaman Entropy

- **Entropy = Pengukuran tingkat ketidak-aturan**
- **Semakin rendah entropy**, semakin teratur dan seragam distribusi data yang kita punya.
- **Semakin tinggi entropy**, semakin tidak teratur dan acak distribusi data yang kita punya.

$$E = -p(a) \log(p(a)) - p(b) \log(p(b))$$

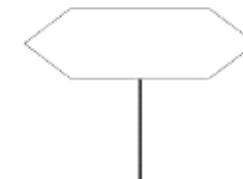
- Dengan $p(\cdot)$ probabilitas atau rasio proporsional Drug A atau Drug B





Pemahaman Entropy

Patient ID	Age	Sex	BP	Cholesterol	Drug
p1	Young	F	High	Normal	Drug A
p2	Young	F	High	High	Drug A
p3	Middle-age	F	Hiigh	Normal	Drug B
p4	Senior	F	Normal	Normal	Drug B
p5	Senior	M	Low	Normal	Drug B
p6	Senior	M	Low	High	Drug A
p7	Middle-age	M	Low	High	Drug B
p8	Young	F	Normal	Normal	Drug A
p9	Young	M	Low	Normal	Drug B
p10	Senior	M	Normal	Normal	Drug B
p11	Young	M	Normal	High	Drug B
p12	Middle-age	F	Normal	High	Drug B
p13	Middle-age	M	High	Normal	Drug B
p14	Senior	F	Normal	High	Drug A



Pemahaman Entropy

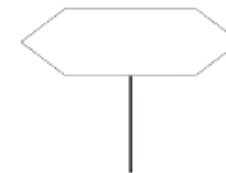
Patient ID	Age	Sex	BP	Cholesterol	Drug
p1	Young	F	High	Normal	Drug A
p2	Young	F	High	High	Drug A
p3	Middle-age	F	Hiigh	Normal	Drug B
p4	Senior	F	Normal	Normal	Drug B
p5	Senior	M	Low	Normal	Drug B
p6	Senior	M	Low	High	Drug A
p7	Middle-age	M	Low	High	Drug B
p8	Young	F	Normal	Normal	Drug A
p9	Young	M	Low	Normal	Drug B
p10	Senior	M	Normal	Normal	Drug B
p11	Young	M	Normal	High	Drug B
p12	Middle-age	F	Normal	High	Drug B
p13	Middle-age	M	High	Normal	Drug B
p14	Senior	F	Normal	High	Drug A

S: [9 B, 5 A]

$$E = -p(B)\log(p(B)) - p(A)\log(p(A))$$

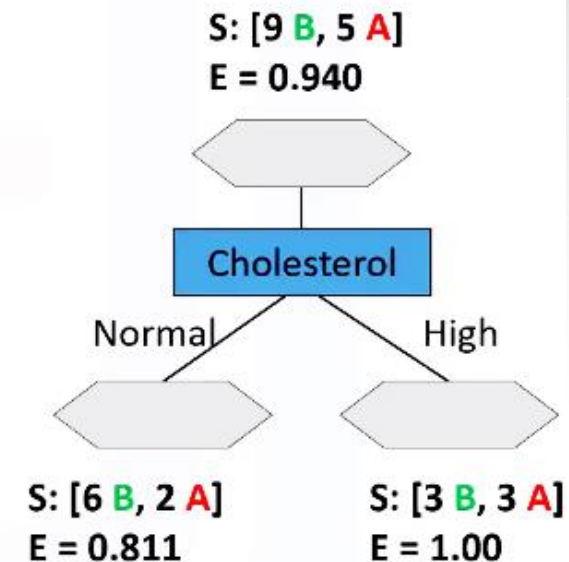
$$E = -(9/14)\log(9/14) - (5/14)\log(5/14)$$

$$E = 0.940$$



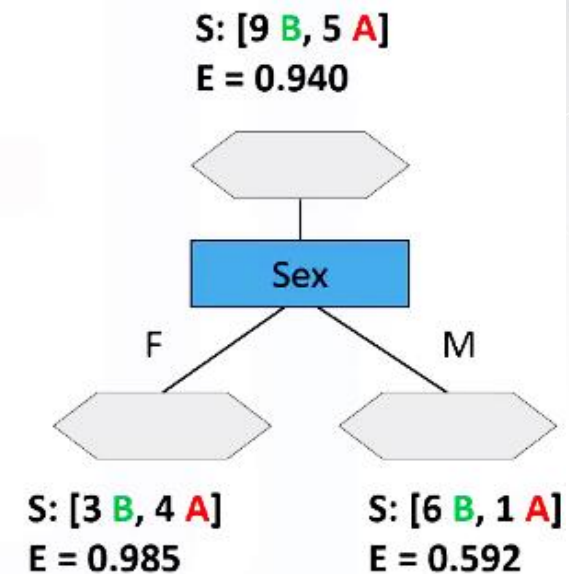
Pemahaman Entropy

Patient ID	Age	Sex	BP	Cholesterol	Drug
p1	Young	F	High	Normal	Drug A
p2	Young	F	High	High	Drug A
p3	Middle-age	F	High	Normal	Drug B
p4	Senior	F	Normal	Normal	Drug B
p5	Senior	M	Low	Normal	Drug B
p6	Senior	M	Low	High	Drug A
p7	Middle-age	M	Low	High	Drug B
p8	Young	F	Normal	Normal	Drug A
p9	Young	M	Low	Normal	Drug B
p10	Senior	M	Normal	Normal	Drug B
p11	Young	M	Normal	High	Drug B
p12	Middle-age	F	Normal	High	Drug B
p13	Middle-age	M	High	Normal	Drug B
p14	Senior	F	Normal	High	Drug A



Pemahaman Entropy

Patient ID	Age	Sex	BP	Cholesterol	Drug
p1	Young	F	High	Normal	Drug A
p2	Young	F	High	High	Drug A
p3	Middle-age	F	High	Normal	Drug B
p4	Senior	F	Normal	Normal	Drug B
p5	Senior	M	Low	Normal	Drug B
p6	Senior	M	Low	High	Drug A
p7	Middle-age	M	Low	High	Drug B
p8	Young	F	Normal	Normal	Drug A
p9	Young	M	Low	Normal	Drug B
p10	Senior	M	Normal	Normal	Drug B
p11	Young	M	Normal	High	Drug B
p12	Middle-age	F	Normal	High	Drug B
p13	Middle-age	M	High	Normal	Drug B
p14	Senior	F	Normal	High	Drug A



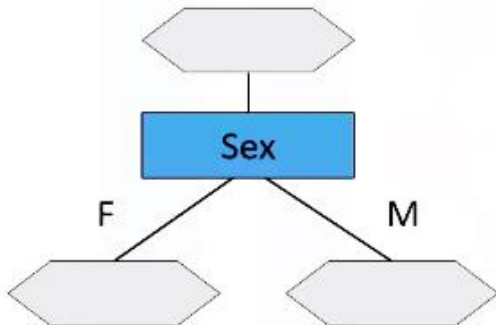


Bagian 4

Menentukan Node yang Terbaik

Mana Node yang Terbaik?

S: [9 B, 5 A]
E = 0.940



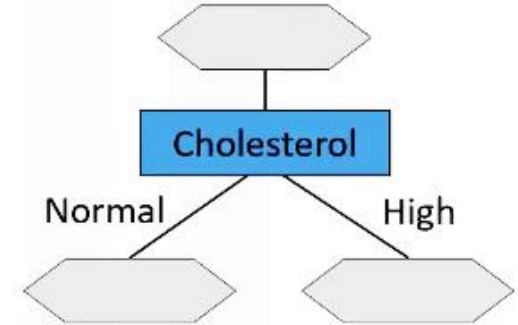
S: [3 B, 4 A]
E = 0.985

S: [6 B, 1 A]
E = 0.592

Patient ID	Age	Sex	BP	Cholesterol	Drug
p1	Young	F	High	Normal	Drug A
p2	Young	F	High	High	Drug A
p3	Middle-age	F	High	Normal	Drug B
p4	Senior	F	Normal	Normal	Drug B
p5	Senior	M	Low	Normal	Drug B
p6	Senior	M	Low	High	Drug A
p7	Middle-age	M	Low	High	Drug B
p8	Young	F	Normal	Normal	Drug A
p9	Young	M	Low	Normal	Drug B
p10	Senior	M	Normal	Normal	Drug B
p11	Young	M	Normal	High	Drug B
p12	Middle-age	F	Normal	High	Drug B
p13	Middle-age	M	High	Normal	Drug B
p14	Senior	F	Normal	High	Drug A

?

S: [9 B, 5 A]
E = 0.940



S: [6 B, 2 A]
E = 0.811

S: [3 B, 3 A]
E = 1.00

Decision Tree dengan Information Gain lebih besar yang kita pilih!

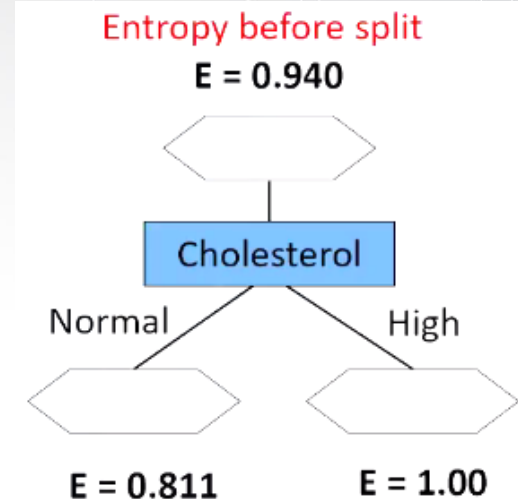


Information Gain

- **Information Gain** adalah sebuah nilai yang menspesifikasikan tingkat informasi yang dimiliki oleh *decision tree*.
- Nilai yang akan meningkatkan tingkat kepastian setelah dibelah.

$$\text{Information Gain} = (\text{Entropy sebelum dibelah}) - (\text{Entropy setelah dibelah})$$

- Membangun Decision Tree adalah tentang menemukan atribut yang memberikan perolehan informasi tertinggi.

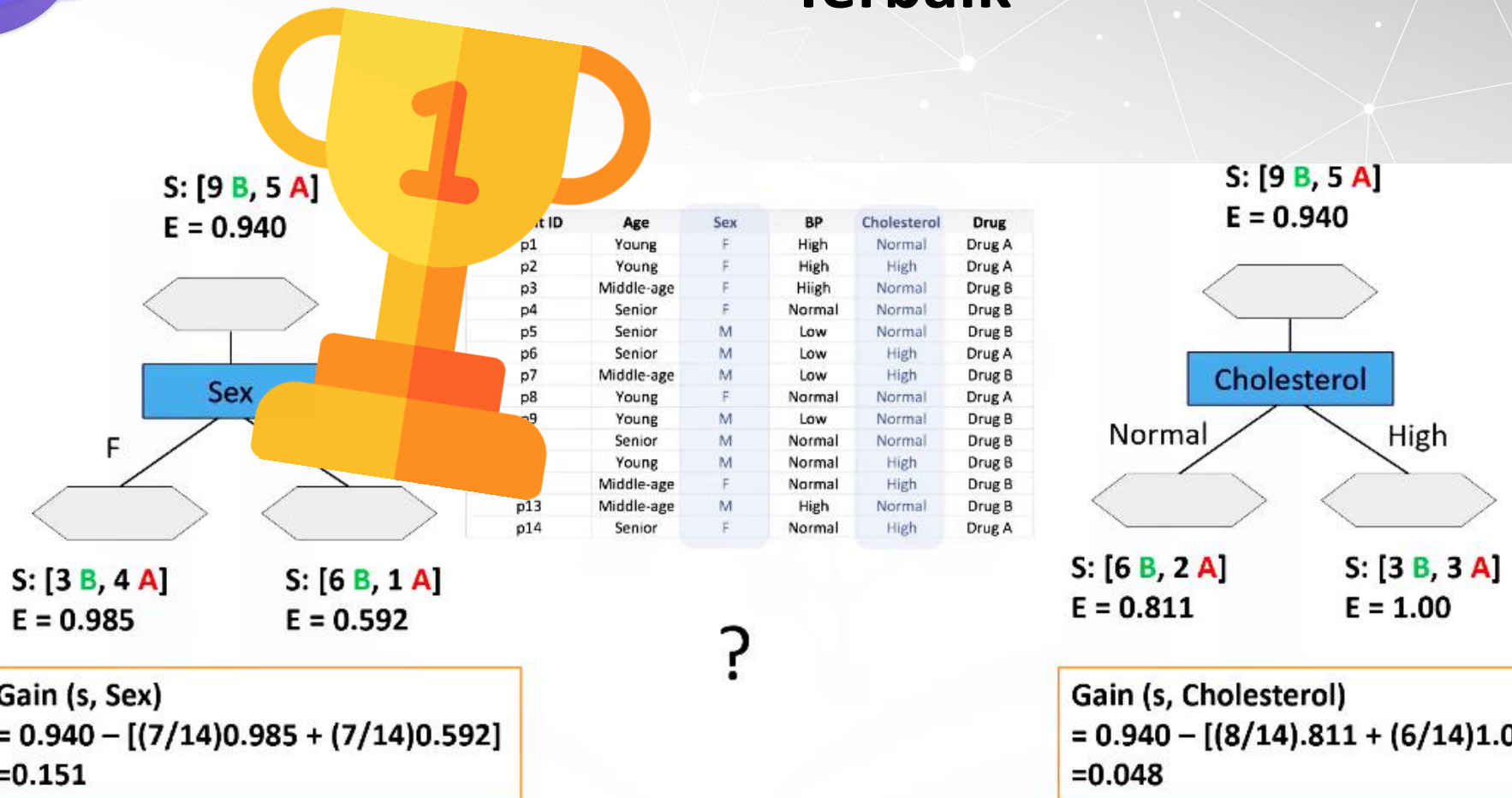


Weighted entropy after split

Weighted Entropy ↓

Information Gain ↑

Information Gain untuk Pemilihan Node Terbaik









Bagian 5

Praktikum Lab

ML0101EN-Clas-Decision-Trees-drug-py-v1.ipynb

IKUTI KAMI



-  digitalent.kominfo
-  digitalent.kominfo
-  DTS_kominfo
-  Digital Talent Scholarship 2019

Pusat Pengembangan Profesi dan Sertifikasi
Badan Penelitian dan Pengembangan SDM
Kementerian Komunikasi dan Informatika
Jl. Medan Merdeka Barat No. 9
(Gd. Belakang Lt. 4 - 5)
Jakarta Pusat, 10110

