seri tulisan data mining

.: Pohon Klasifikasi - Bagian 1 :.

Gambaran Umum dan Algoritma Dasar yang Perlu Diketahui

Bagus Sartono

bagusco@gmail.com

July 20, 2015

Abstract

Tulisan ini memamaparkan beberapa hal dasar dan umum terkait dengan pohon klasifikasi yang perlu diketahui oleh mereka yang ingin mempelajari penggunaan dan pengembangan teknik ini. Untuk diketahui, tulisan ini hanya ditujukan untuk sarana berbagi pengetahuan dan tidak disiapkan sebagai rujukan ilmiah. Sebagian besar isi dalam tulisan ini diilhami oleh beberapa chapter dari buku Tom Mithcell (Machine Learning, McGraw-Hill, 1997), yang kemudian diramu berdasarkan pengalaman penulis menerapkan pohon klasifikasi dalam berbagai kasus dan pengalaman menyampaikan materi ini di perkuliahan maupun berbagai pelatihan. Pembaca yang ingin memperoleh rujukan teoritis dapat menelusurinya pada buku dan artikel jurnal yang relevan. Kritik, saran dan pertanyaan terhadap materi tulisan ini dapat disampaikan melalui email pada alamat bagusco@gmail.com. Penulis sangat mengapresiasi berbagai masukan tersebut. Akhirnya, selamat membaca.

1 Pengantar

Bagi pembaca yang belum terbiasa dengan nama di atas, *Pohon Klasifikasi* adalah terjemahan dari *Classification Tree* yang dalam beberapa publikasi dan software disebut sebagai *Decision Tree*. Pohon Klasifikasi merupakan analisis yang menghasilkan aturan jika-maka, dengan bentuk umum "jika karakteristiknya begini dan begitu, maka objek tersebut tergolong dalam kelas tertentu". Karena aturan tersebut dapat digambarkan dalam bentuk yang menyerupai pohon, maka dikenal dengan istilah pohon klasifikasi. Bentuk yang hampir sama juga ditemukan pada diskusi mengenai pengambilan keputusan sehingga beberapa orang juga menggukanan istilah pohon keputusan.

Perhatikan kembali bentuk umum dari aturan jika-maka yang dihasilkan: "jika karakteristiknya begini dan begitu, maka objek tersebut tergolong dalam kelas tertentu". Andaikan

aturan tersebut telah diperoleh, maka kita dapat memanfaatkannya untuk mengkelaskan atau memasukkan suatu objek yang karakteristiknya diketahui ke dalam kelompok yang sesuai. Dengan dasar itulah maka dalam penerapannya pohon klasifikasi banyak digunakan untuk melakukan pengkelasan seperti dalam kasus berikut:

- Persetujuan Aplikasi Kredit

  Dalam rangka menjalankan sistem manajemen resikonya, bank dan lembaga pembiayaan melakukan seleksi terhadap aplikasi pengajuan kredit yang mereka terima. Tidak semua orang yang mengajukan kredit akan disetujui dan dibiayai, karena ada orang-orang tertentu yang dianggap "layak" dan beberapa yang lain dianggap "tidak layak". Kelayakan ini dinilai dari potensi kemampuan bayar seseorang. Analisis klasifikasi digunakan untuk mengkelaskan apakah seseorang dapat disebut layak (yang dalam istilah pembiayaan disebut sebagai Good) ataukah sebaliknya (disebut juga Bad) berdasarkan karakteristiknya seperti usia, banyaknya tanggungan, tingkat penghasilan, kepemilikan rumah, dan lain-lain. Proses penilaian ini sering dikenal sebagai Approval Credit Scoring.

- Penentuan Target Direct Marketing

  Saat ini banyak perusahaan yang melakukan pemasaran langsung (Direct Marketing) melalui berbagai media. Jika beberapa tahun yang lalu mereka menggunakan surat dan telepon, akhir-akhir ini banyak yang menggunakan pesan singkat atau SMS. Pesan singkat itu biasanya berisi ajakan untuk melakukan transaksi atau pembelian tertentu. Setiap pengiriman ajakan tersebut baik menggunakan surat, telepon maupun SMS, semuanya memerlukan biaya. Proses marketing akan lebih efektif jika hanya orang-orang tertentu yang diyakini memiliki peluang besar untuk mengikuti ajakan melakukan pembelian atau traksasi saja yang dijadikan target pengiriman pesan. Dengan demikian diperlukan tahapan untuk mengidentifikasi siapa-siapa saja dari daftar yang dimiliki yang tergolong dalam kelas orang-orang yang prospektif. Dengan kata lain setiap orang dengan karakteristik demografi dan yang lainnya akan ditentukan masuk ke kelas prospektif atau tidak. Di dunia marketing, ini dikenal sebagai Propensity Model.

- Penentuan Segmen Konsumen

Pohon klasifikasi tentu saja bukan satu-satunya tool yang dapat digunakan dalam analisis klasifikasi. Pendekatan lain yang juga dapat digunakan antara lain adalah regresi logistik, analisis diskriminan, Bayesian classifier, k-nearest neighbor, jaringan syarat tiruan (artificial neural network), dan support vector machine. Output visual yang berupa pohon menjadikan pohon klasifikasi banyak disukai orang selain aspek kesederhanaannya.

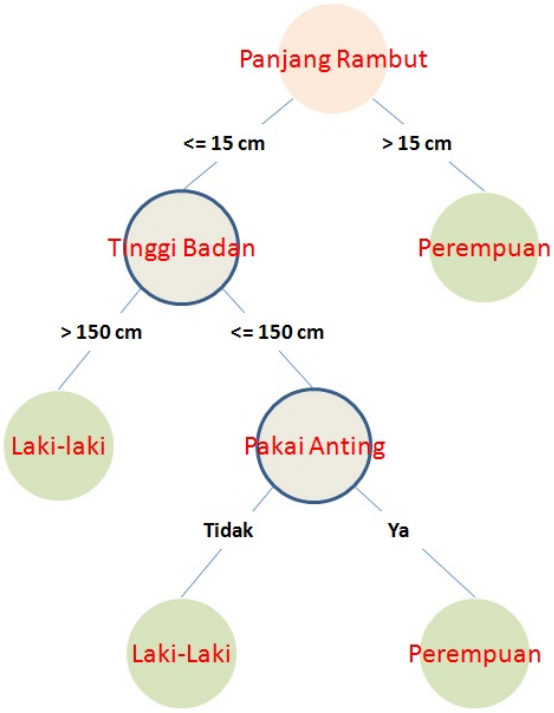

Figure 1: Tampilan dasar pohon klasifikasi

2 Tampilan dari Pohon Klasifikasi

Tampilan visual dari pohon klasifikasi mirip dengan pohon dalam posisi yang terbalik, dimana simpul akar (*root node*) berada di atas dan pohonnya tumbuh ke bawah. Semula semua data berada pada simpul akar dan selanjutnya bercabang menjadi dua atau lebih simpul dengan aturan pencabangan tertentu. Simpul-simpul baru berisi pengamatan yang lebih sedikit, dan kemudian masing-masing dapat bercabang kembali menjadi simpul-simpul yang baru. Simpul akhir yang tidak mengalami pencabangan biasanya dikenal sebagai simpul daun (*leaf node*), yang pada sebagian literatur disebut dengan istilah simpul akhir (*terminal node*). Sedangkan simpul-simpul yang bercabang dikenal sebagai simpul antara (*intermediate node*). Simpul antara dikenal juga sebagai simpul keputusan (*decision node*) karena simpul ini menentukan bagaimana pengamatan tertentu masuk ke simpul-simpul di bawahnya.

Gambar 1 menampilkan struktur dasar dari sebuah pohon klasifikasi yang digunakan untuk mengkelaskan apakah seseorang berjenis kelamin laki-laki ataukah perempuan. Beberapa variabel digunakan untuk menentukan prediksi jenis kelamin itu yaitu antara lain panjang rambut, penggunaan anting dan tinggi badan.

Sebelum kita lihat lebih jauh pohon klasifikasi pada Gambar 1 tersebut, kita diskusikan dulu jenis-jenis simpul yang ada. Simpul yang berwarna merah muda yang posisinya paling atas merupakan simpul akar, simpul yang berwarna hijau adalah simpul daun, serta simpul lain yang ada di bagian tengah adalah simpul antara. Terdapat empat simpul daun dan dua

simpul antara pada pohon tersebut.

Bagaimana penggunaan pohon pada Gambar 1 ini? Pertama untuk menentukan seseorang berjenis kelamin laki-laki atau perempuan, dilihat terlebih dahulu panjang rambutnya. Jika panjang rambutnya lebih dari 15 cm, maka orang tersebut akan diduga berjenis kelamin perempuan. Namun, jika panjang rambutnya tidak lebih dari 15 cm maka belum ada keputusan atau dugaan akhir. Orang-orang yang panjang rambutnya tidak lebih dari 15 cm akan diperiksa data tinggi badannya. Jika tingginya melebihi 150 cm maka diprediksi orang tersebut berjenis kelamin laki-laki, namun jika tidak lebih dari 150 cm maka digunakan data apakah dia menggunakan anting telinga. Seandainya orang tersebut tidak menggunakan maka diduga dia adalah laki-laki dan jika dia menggunakan anting maka diduga dia adalah perempuan.

Merujuk pada penjelasan dari paragraf di atas maka kita dapat menyusun beberapa aturan jika-maka sebagai berikut:

- jika $Panjang\ Rambut > 15$ maka diduga $Perempuan$
- jika $Panjang\ Rambut \leq 15$ dan $Tinggi\ Badan > 150$ maka diduga $Laki - Laki$
- jika $PanjangRambut \leq 15$, $TinggiBadan \leq 150$ dan $PakaiAnting = Tidak$ maka diduga $Laki - Laki$
- jika $Panjang\ Rambut \leq 15$, $Tinggi\ Badan \leq 150$ dan $Pakai\ Anting = Ya$ maka diduga $Perempuan$

Dengan demikian, setiap pohon klasifikasi dapat diubah dalam bentuk aturan jika maka yang dapat digunakan dengan mudah untuk melakukan proses pengkelasan dalam database yang dimiliki.

3 Algoritma Dasar

Perhatikan kembali Gambar 1 yang menampilkan pohon klasifikasi untuk memprediksi jenis kelamin seseorang apakah tergolong dalam laki-laki ataukah perempuan. Beberapa variabel digunakan untuk menentukan prediksi jenis kelamin itu yaitu antara lain panjang rambut, penggunaan anting dan tinggi badan. Dalam pembicaraan di bidang data mining dan pemodelan statistika, jenis kelamin yang akan diprediksi disebut sebagai *variabel respon* atau *variabel target*, sedangkan variabel lain yang digunakan untuk memprediksi disebut sebagai *variabel prediktor* atau *variabel input*.

Berdasarkan penjelasan dalam ilustrasi, kiranya jelas bahwa pohon klasifikasi cocok digunakan pada kasus dimana variabel target merupakan variabel yang bersifat kategorik. Sementara itu, jenis dari variabel prediktornya dapat bersifat numerik maupun kategorik. Hanya saja

pada algoritma-algoritma yang dikembangkan diawal, variabel prediktornya pun mesti bersifat kategorik. Pada algoritma yang lebih baru, proses pengkategorian variabel prediktor dilakukan di dalam proses penyusunan pohon klasifikasinya.

Pada bagian ini akan dipaparkan algoritma ID3 (*Iterative Dichotomiser* 3) yang dikembangkan oleh Quinlan (1983), dan selanjutnya nanti akan didiskusikan beberapa pengembangan algoritma yang disusun oleh berbagai penulis pada artikel lain.

Ide dasar dari penyusunan pohon klasifikasi adalah membagi-bagi gugus data yang semula berada pada satu simpul besar yaity simpul akar menjadi simpul-simpul turunannya yang lebih sedikit anggotanya dan bersifat lebih homogen. Pertanyaan pertama yang harus dijawab dalam penyusunan pohon klasifikasi adalah variabel prediktor mana yang digunakan untuk memisahkan gugus data tersebut? Jika variabel itu telah ditemukan maka akan dimiliki anak gugus - anak gugus atau simpul-simpul baru yang lebih kecil ukurannya. Selanjutnya proses yang sama dilakukan untuk masing-masing simpul yang terbentuk, dan berlanjut sampai setiap simpul tidak perlu atau tidak bisa lagi dibagi/dipisah.

Dalam penjelasan awal ini kita akan membatasi bahwa semua variabel prediktor bersifat kategorik dan variabel target hanya memiliki dua jenis nilai. Untuk menjawab variabel mana yang dapat dijadikan pemisah terbaik maka diperlukan ukuran yang menggambarkan kebaikan pemisahan tersebut. Algoritma ID3 menggunakan sifat statistik yang disebut dengan *information gain* untuk menentukan variabel terbaik. Di setiap tahapan pemisahan, algoritma ini menghitung *information gain* dari setiap variabel prediktor dan selanjutnya variabel prediktor dengan nilai terbaik yang dipilih.

Apa itu *information gain*? Ide dari ukuran ini adalah mencari variabel yang mampu membuat simpul hasil pemisahaan sehomogen mungkin. Dalam bidang informasi, kehomogenan data diukur menggunakan sebuah nilai yang disebut entropy yang merupakan ukuran ketidakhomogenan kumpulan data. Andaikan suatu kumpulan data D yang terdiri atas dua jenis nilai saja yaitu 1 dan 2 dan proporsi banyaknya data bernilai 1 adalah p, maka entropy dari kumpulan data tersebut adalah

$$Entropy(D) = -p \log_2 (p) - (1-p) \log_2 (1-p) \tag{1}$$

Sebagai tambahan, didefinisikan juga bahwa untuk $p = 0$ atau $p = 1$ maka $Entropy(D) = 0$.

Berdasarkan definisi entropy di atas maka suatu gugus data yang seluruhnya bernilai 1 atau seluruhnya bernilai 2 maka nilai entropy-nya adalah nol. Selanjutnya jika banyaknya amatan dengan nilai 1 persis sama dengan yang bernilai 2 atau dengan kata lain $p = 0.5$ maka nilai entropy-nya adalah 1 dan itu adalah nilai entropy yang paling besar.

Sebagai ilustrasi, andaika terdapat sekumpulan data berisi 20 amatan yang terdiri atas 8

amatan bernilai 1 dan 12 amatan lainnya bernilai 2. Maka entropi dari gugus data tersebut adalah

$$-0.4\log_2(0.4) - 0.6\log_2(0.6) = 0.971. \tag{2}$$

Sementara itu, terdapat gugus data lain yang juga beukuran 20 amatan namun hanya memiliki 2 amatan bernilai 1 sedangkan 18 amatan lainnya bernilai 2. Entropi dari gugus data kedua ini adalah

$$-0.1\log_2(0.1) - 0.9\log_2(0.9) = 0.469. \tag{3}$$

Gugus data pertama merupakan gugus yang bersifat lebih heterogen dibandingkan gugus kedua. Sifat lebih heterogen ini ditunjukkan dengan entropi yang lebih besar. Sekali lagi ingin ditegaskan bahwa entropi dengan demikian dapat digunakan sebagai ukuran seberapa homogen atau heterogen gugus data yang dimiliki. Entropi yang semakin tinggi akan dimiliki oleh data yang lebih heterogen, dan sebaliknya data yang homogen akan memiliki entropi yang lebih kecil.

Karena entropi dapat digunakan sebagai ukuran kehogenan data dengan demikian jika seandainya suatu gugus data dibagi menjadi anak gugus - anak gugus baru yang diharapkan bersifat lebih homogen maka penurunan nilai entropi dari gugus data awal menjadi anak gugus dapat dijadikan ukuran kebaikan pemisahan. Inilah yang disebut sebagai information gain. Andaikan D adalah suatu gugus data dan V adalah suatu variabel prediktor kategorik yang memiliki k buah nilai yaitu $v_1, v_2, ..., v_k$. Jika gugus D dipisah-pisah berdasarkan nilai dari variabel V maka akan ada k buah anak gugus data yaitu $D_{v_1}, D_{v_2}, ..., D_{v_k}$, dengan D_{v_i} adalah anak gugus yang beranggotakan amatan-amatan yang berpadanan dengan amatan yang memiliki nilai $V = v_i$. *Information gain* dari pemisahan ini dinotasikan $Gain(D, V)$ dan didefinisikan sebagai

$$Gain(D, V) = Entropy(D) - \sum_{i=1}^{k} \frac{|D_{v_i}|}{|D|} Entropy(D_{v_i}) \tag{4}$$

dengan $|D|$ dan $|D_{v_i}|$ masing-masing adalah banyaknya amatan di D dan D_{v_i}. Tampak bahwa sesungguhnya nilai information gain adalah selisih antara entropi dari gugus asal dengan rata-rata terboboti dari entropi-entropi anak gugusnya. Jika anak gugus bersifat lebih homogen maka mereka akan memiliki entropi yang rendah sehingga nilai $Gain(D, V)$ akan bernilai besar. Sebaliknya jika anak gugus tidak bersifat lebih homogen dari gugus asalnya maka nilai $Gain(D, V)$ akan rendah.

Untuk memperjelas proses perhitungan nilai *information gain* ini, perhatikan gugus data yang disajikan pada Tabel 1. Andaikan Jenis Kelamin adalah variabel target sedangkan Tinggi Badan dan Menggunakan Anting adalah dua buah variabel prediktor.

Table 1: Data ilustrasi

Nomor Nomor	Jenis Kelamin	Menggunakan Anting	Tinggi Badan	Panjang Rambut
1	Perempuan	Ya	Sedang	Panjang
2	Laki-Laki	Tidak	Tinggi	Pendek
3	Laki-Laki	Tidak	Pendek	Pendek
4	Laki-Laki	Tidak	Tinggi	Pendek
5	Perempuan	Tidak	Sedang	Panjang
6	Perempuan	Ya	Pendek	Pendek
7	Perempuan	Tidak	Pendek	Panjang
8	Laki-Laki	Tidak	Sedang	Pendek
9	Laki-Laki	Tidak	Tinggi	Pendek
10	Laki-Laki	Tidak	Sedang	Pendek
11	Laki-Laki	Tidak	Sedang	Pendek
12	Laki-Laki	Ya	Sedang	Pendek
13	Perempuan	Ya	Sedang	Panjang
14	Laki-Laki	Tidak	Pendek	Panjang
15	Laki-Laki	Tidak	Sedang	Pendek
16	Laki-Laki	Tidak	Sedang	Pendek
17	Laki-Laki	Tidak	Tinggi	Pendek
18	Laki-Laki	Tidak	Pendek	Pendek
19	Laki-Laki	Tidak	Sedang	Pendek
20	Perempuan	Ya	Tinggi	Panjang

Variabel Tinggi Badan memiliki tiga buah nilai yaitu {Tinggi, Sedang, Pendek }, sementara variabel Menggunakan Anting memiliki dua nilai berbeda yaitu {Ya, Tidak}. Kita akan gunakan data tersebut untuk ilustrasi penghitungan nilai *information gain* jika pemisahan dilakukan menggunakan masing-masing dari kedua variabel prediktor.

Pertama kita lihat dulu seandainya pemisahnya adalah Menggunakan Anting (MA). Gugus data awal berisi 20 amatan dengan 6 Perempuan dan 14 Laki-Laki. Pemisahan berdasarkan variabel Menggunakan Anting akan menghasilkan dua bagian dengan bagian pertama adalah amatan yang $Menggunakan\,Anting = Tidak$" yang berisi 15 amatan terdiri atas 2 Perempuan dan 13 Laki-Laki, sedangkan bagian kedua merupakan kelompok dengan $MenggunakanAnting = Ya$ yang berisi 5 amatan terdiri atas 4 Perempuan dan 1 Laki-Laki. Entropi dari masing-masing gugus dan anak gugus adalah sebagai berikut:

$$
\begin{aligned}
Entropy(D) &= -\frac{6}{20}\log_2(\frac{6}{20}) - \frac{14}{20}\log_2(\frac{14}{20}) &= 0.8813\\
Entropy(D_{MA=Tidak}) &= -\frac{2}{15}\log_2(\frac{2}{15}) - \frac{13}{15}\log_2(\frac{13}{15}) &= 0.5665\\
Entropy(D_{MA=Ya}) &= -\frac{4}{5}\log_2(\frac{4}{5}) - \frac{1}{5}\log_2(\frac{1}{5}) &= 0.7219
\end{aligned}
$$

Selanjutnya, *information gain* dari pemisahan ini adalah

$$Gain(D, MA) = Entropy(D) - \frac{15}{20}Entropy(D_{MA=Tidak}) - \frac{5}{20}Entropy(D_{MA=Ya}) = 0.276 \quad (5)$$

Sementara itu, jika pemisahan dilakukan menggunakan variabel Tinggi Badan (TB), kita akan memiliki nilai-nilai entropi dan gain sebagai berikut:

$$
\begin{aligned}
Entropy(D) &= -\frac{6}{20}\log_2(\frac{6}{20}) - \frac{14}{20}\log_2(\frac{14}{20}) &= 0.8813\\
Entropy(D_{TB=Tinggi}) &= -\frac{1}{5}\log_2(\frac{1}{5}) - \frac{4}{5}\log_2(\frac{4}{5}) &= 0.7219\\
Entropy(D_{TB=Sedang}) &= -\frac{3}{10}\log_2(\frac{3}{10}) - \frac{7}{10}\log_2(\frac{7}{10}) &= 0.8813\\
Entropy(D_{TB=Pendek}) &= -\frac{2}{5}\log_2(\frac{2}{5}) - \frac{3}{5}\log_2(\frac{3}{5}) &= 0.9705
\end{aligned}
$$

dan

$$
\begin{aligned}
Gain(D, T) = Entropy(D) - \frac{5}{20}Entropy(D_{TB=Tinggi}) - \frac{10}{20}Entropy(D_{TB=Sedang})\\
-\frac{5}{20}Entropy(D_{TB=Pendek}) = 0.017
\end{aligned} \quad (6)
$$

Berdasarkan hasil di atas, terlihat bahwa variabel $MenggunakanAnting$ memberikan pemisahan yang lebih baik dibandingkan *Tinggi Badan* yang diindikasikan dengan *information gain* yang lebih tinggi yaitu sebesar 0.276 dibandingkan 0.017. Dengan demikian jika hanya ada dua variabel ini yang digunakan sebagai prediktor maka variabel *Menggunakan Anting* akan dipilih sebagai variabel pemisah.

4 Ilustrasi Sederhana Penerapan Algoritma ID3

Pada bagian ini akan dipaparkan ilustrasi penerapan algoritma ID3 menggunakan kasus sederhana dengan data yang ditampilkan pada Tabel 1. Variabel targetnya adalah *Jenis Kelamin* sedangkan variabel prediktor tersedia tiga buah yaitu *Menggunakan Anting* dengan dua nilai $\{Ya, Tidak\}$, $TinggiBadan$ yang memiliki tiga nilai $\{Tinggi, Sedang, Pendek\}$, dan $PanjangRambut$ yang memiliki dua nilai $\{Panjang, Pendek\}$.

Langkah pertama adalah menempatkan 20 data dalam simpul akar, dan menghitung nilai entropi dari simpul tersebut. Seperti yang telah dihitung sebelumnya, entropi dari gugus 20

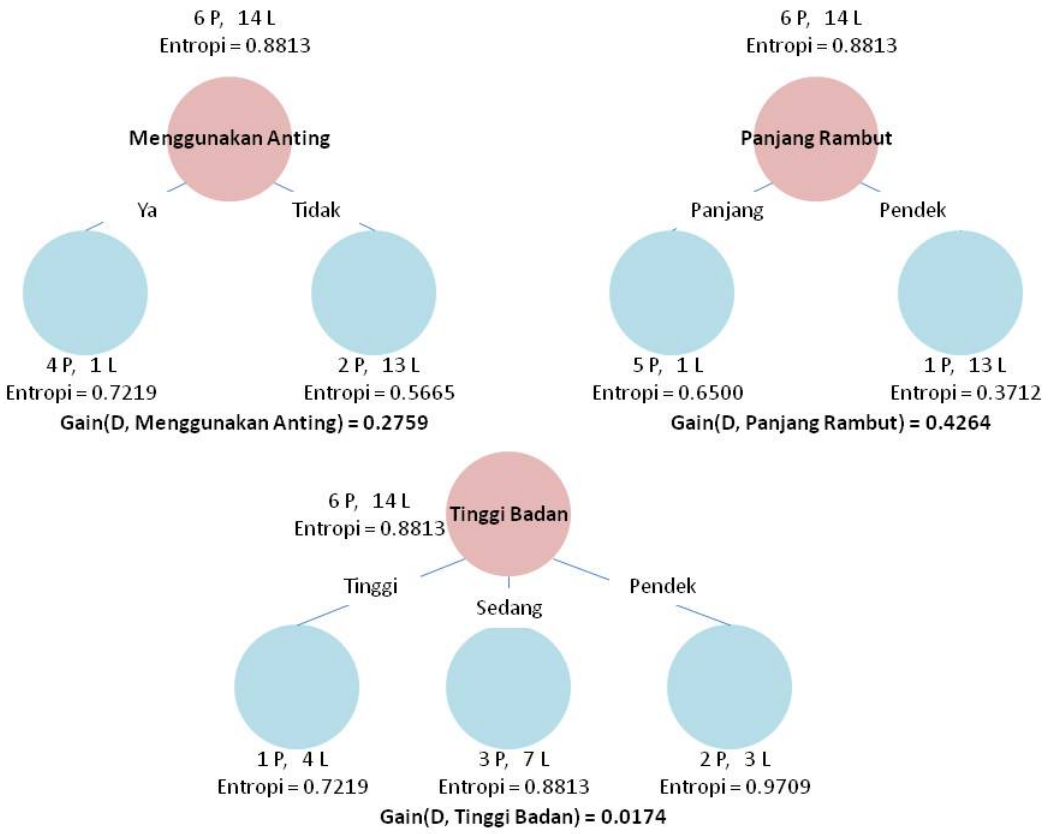

Figure 2: Penghitungan information gain untuk memperoleh variabel pemisah terbaik bagi simpul akar

data ini adalah 0.8813. Selanjutnya dicari variabel terbaik yang memisahkan dengan terlebih dahulu menghitung *information gain* hasil pemisahan dari masing-masing variabel prediktor. Gambar 2 menyajikan ringkasan penghitungannya. Terlihat bahwa pemisahan menggunakan $Panjang\,Rambut$ menghasilkan gain yang paling besar. Dengan demikian, pemisahan pertama menghasilkan dua simpul baru yang kemudian kita sebut S1 dan S2, dengan S1 adalah simpul yang berpadanan dengan amatan-amatan $Panjang\,Rambut = Panjang$ dan S2 untuk amatan dengan $Panjang\ Rambut = Pendek$.

Selanjutnya dilakukan proses yang sama terhadap simpul akar, namun sekarang dilakukan untuk masing-masing simpul S1 dan simpul S2. Kita lakukan dulu pada simpul S1. Perhatikan bahwa pada simpul S1 amatan yang terhimpun seluruhnya memiliki nilai $Panjang\ Rambut = Panjang$ sehingga penentuan variabel pemisah hanya dilakukan pada variabel $MenggunakanAnting$ dan $Tinggi\,Badan$. Gambar 3 menyajikan rangkuman proses tersebut, dimana pemisahan terbaik terjadi saat menggunakan variabel $Tinggi\ Badan$.

Proses yang sama dilakukan terhadap simpul S2 yang ringkasannya disajikan pada Gambar 4. Pada tahap ini, pemisahan menggunakan variabel $Menggunakan\,Anting$ menghasilkan gain sebesar 0.2284, sedangkan pemisahan menggunakan variabel $Tinggi\,Badan$ menghasilkan gain sebesar 0.1744. Dengan demikian, pemisahan terbaik terhadap simpul S2 adalah menggunakan

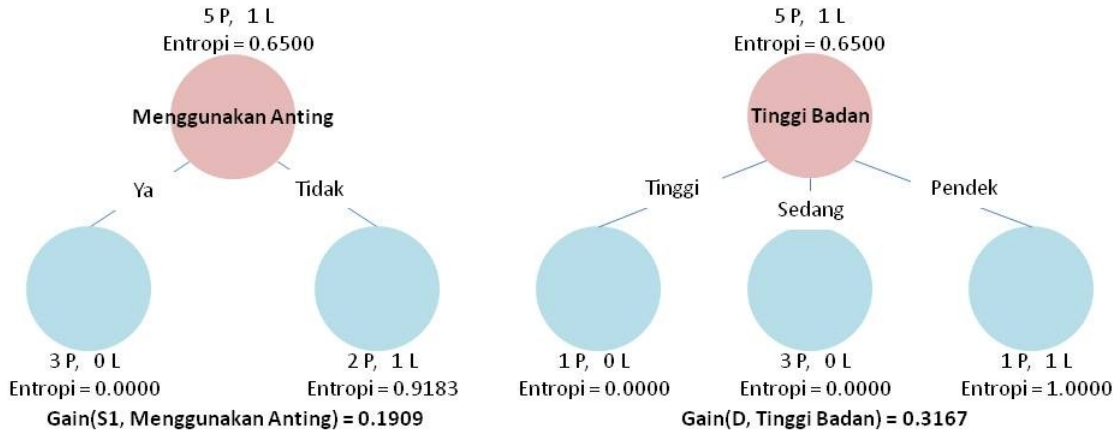

Figure 3: Penghitungan information gain untuk memperoleh variabel pemisah terbaik pada simpul S1

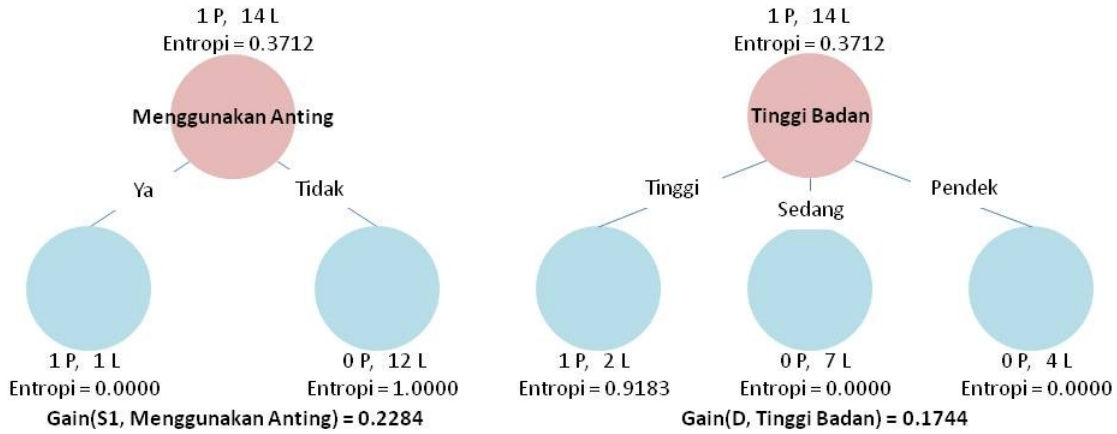

Figure 4: Penghitungan information gain untuk memperoleh variabel pemisah terbaik pada simpul S2

variabel *Menggunakan Anting*.

Merangkum apa yang telah kita kerjakan, semula simpul akar dipisah menjadi dua berdasarkan variabel *Panjang Rambut* menghasilkan dua simpul S1 dan S2. Selanjutnya simpul S1 dipecah menjadi tiga berdasarkan variabel *Tinggi Badan* dan simpul S2 dipecah kembali oleh variabel *Menggunakan Anting* menjadi dua simpul. Gambaran hasil sementara pohon klasifikasi sampai tahap ini diberikan pada Gambar 5.

Perhatikan bahwa pada Gambar 5 yang merupakan pohon klasifikasi yang diperoleh sampai dengan iterasi kedua didapatkan lima buah simpul yang posisinya paling luar. Simpul-simpul tersebut pada gambar diberi nama S3, S4, S5, S6, dan S7. Tiga buah simpul yaitu S3, S4, dan S7 memiliki entropi nol yang berarti bahwa seluruh amatan memiliki nilai variabel target yang sama. Dengan demikian tidak perlu ada proses pemisahan lagi pada ketiga simpul tersebut. Sedangkan pada simpul S5 dan S6 masih dapat dilakukan kembali proses pencarian pemisahan terbaik. Ingat bahwa pada simpul S5, semua amatan merupakan amatan dengan

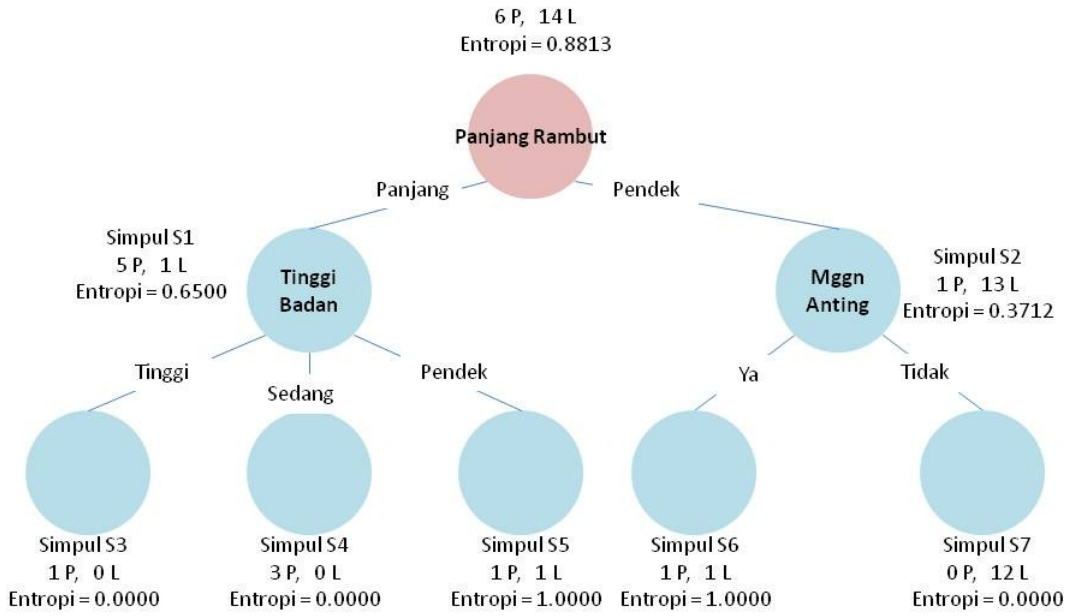

Figure 5: Pohon klasifikasi yang dihasilkan pada iterasi kedua

Panjang Rambut = *Panjang* dan *Tinggi Badan* = *Pendek*, sehingga satu-satunya variabel kandidat pemisah adalah *Menggunakan Anting*. Demikian juga untuk simpul S6, hanya variabel *Tinggi Badan* yang menjadi variabel kandidat pemisah. Karena hanya ada satu variabel untuk masing-masing, proses penghitungan *information gain* tidak lagi dibahas karena tidak ada proses pembandingan nilai tersebut.

Pemeriksaan terhadap data simpul S5 menunjukkan bahwa semua amatan yang memiliki nilai *Panjang Rambut* = *Panjang* dan *Tinggi Badan* = *Pendek* ternyata memiliki nilai yang sama untuk variabel *Menggunakan Anting* yaitu *Tidak*. Dengan demikian tidak ada proses pemisahan lanjuta pada simpul tersebut. Sedangkan untuk simpul S6 dua amatan yang tersisa memiliki tinggi badan yang berbeda dan simpul yang dihasilkan tidak dapat dipecah lebih lanjut. Pohon klasifikasi akhir yang dihasilkan dengan demikian disajikan pada Gambar 6.

Pohon klasifikasi akhir yang diperoleh memiliki enam buah simpul akhir yaitu S3, S4, S5, S8, S9 dan S7 sehingg kita dapat menurunkan 6 (enam) buah aturan jika-maka sebagai berikut:

- S3: jika *Panjang Rambut* = *Panjang* dan *Tinggi Badan* = *Tinggi* maka *Perempuan*
- S4: jika *Panjang Rambut* = *Panjang* dan *Tinggi Badan* = *Sedang* maka *Perempuan*
- S5: jika *Panjang Rambut* = *Panjang* dan *Tinggi Badan* = *Pendek* maka tidak dapat diputuskan jenis kelaminnya karena peluangnya fifty-fifty
- S7: jika *Panjang Rambut* = *Pendek* dan *Menggunakan Anting* = *Tidak* maka *Laki-Laki*
- S8: jika *Panjang Rambut* = *Pendek*, *Menggunakan Anting* = *Ya* dan *Tinggi Badan* = *Sedang* maka *Laki-Laki*

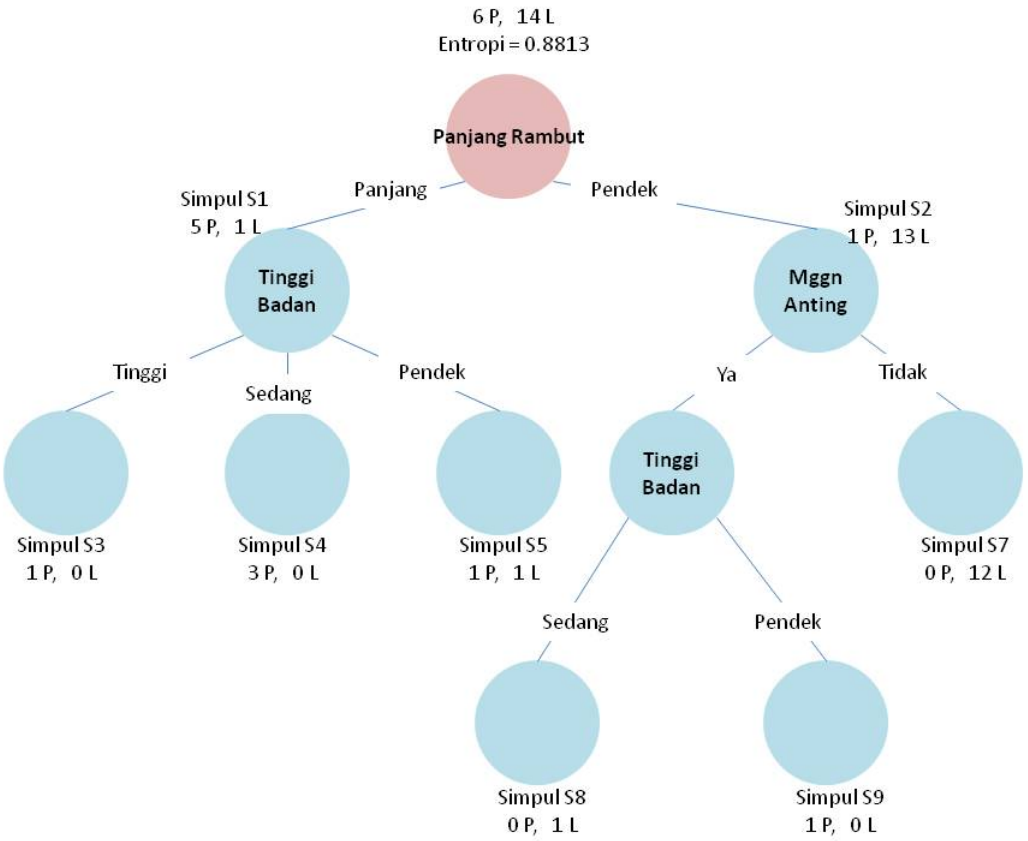

Figure 6: Pohon klasifikasi akhir yang dihasilkan

- S9: jika *Panjang Rambut* = *Pendek*, *Menggunakan Anting* = *Ya* dan *Tinggi Badan* = *Pendek* maka *Laki-Laki*

5 Beberapa Isu Tambahan dalam Algoritma Penyusunan Pohon Klasifikasi

5.1 Kapan iterasi dari algoritma sebaiknya dihentikan?

Pada ilustrasi mengenai klasifikasi jenis kelamin pada bagian sebelumnya, proses pemisah-misahan simpul yang merupakan representasi dari anak gugus data dilakukan secara iteratif dan berhenti jika menemui satu dari dua hal berikut:

- Simpul atau gugus datanya berisi amatan-amatan di dalamnya memiliki nilai dari variabel targetnya yang seluruh sama, misalnya semuanya Laki-Laki atau semuanya Perempuan. Dengan kata lain, jika nilai entropinya nol.
- Simpul atau gugus data berisi amatan-amatan di dalamnya memiliki nilai variabel prediktor yang seluruhnya identik, misalnya seperti yang terjadi pada Simpul S8.

Apabila proses pembentukan pohon klasifikasi dilakukan dengan kriteria penghentian hanya dua hal di atas maka proses pembuatannya disebut sebagai pembentukan pohon secara lengkap atau secara maksimal. Namun demikian dalam prakteknya dapat saja ditambahkan kriteria-kriteria lain sebagai berikut:

- Simpul atau gugus data hanya berisi sedikit amatan saja. Hal ini didasarkan pada pemikiran bahwa jika suatu simpul hanya terdiri atas sedikit amatan maka sudah tidak layak dilakukan inferensia secara statistika untuk penentuan pemisahan terbaik.
- Kedalaman (Depth) dari pohon yang terbentuk sudah cukup memadai. Yang dimaksud dengan kedalaman adalah banyaknya baris simpul yang terbentuk dari pohon. Pohon yang hanya memiliki simpul akar disebut memiliki kedalaman nol. Semakin dalam pohon, umumnya semakin banyak simpul yang terbentuk. Semakin dalam pohon, semakin besar pohon yang dihasilkan. Dengan menentukan sejak awal berapa kedalaman maksimal, maka proses pembentukan pohon dapat dihentikan jika kedalamannya sudah mencapai batas yang ditentukan tersebut.
- Secara statistik sudah tidak signifikan lagi penurunan keheterogenannya.

Penggunaan berbagai kriteria lain ini akan dibahas kemudian ketika membahas algoritma lain selain ID3.

5.2 Menghindari Pohon Klasifikasi yang bersifat Overfitting dan Pemangkasan Pohon

Seperti yang telah disebutkan beberapa kali, apabila pohon klasifikasi telah diperoleh maka pohon tersebut dapat digunakan untuk melakukan prediksi terhadap kelas dari suatu individu yang nilai-nilai variabel prediktornya telah diketahui. Pohon yang baik adalah pohon yang mampu memberikan ketepatan prediksi yang sangat tinggi. Kondisi overfitting adalah kondisi dimana pohon klasifikasi mampu memberikan prediksi yang sangat baik pada data latih (data yang digunakan untuk membangun pohon klasifikasi) namun kemampuan prediksinya jauh menurun pada saat digunakan dengan data lain.

Kondisi overfitting banyak dijumpai pada saat pohon klasifikasinya terlalu dalam atau terlalu kompleks. Membuat pohon dengan kriteria penghentian sampai gugus atau simpul tidak mungkin lagi dipisah karena sudah menjadi simpul yang sangat homogen seringkali bukan menjadi pilihan terbaik. Ini dikarenakan data yang dimiiliki tidak pernah bebas dari *noise* yang dapat mengganggu pembuatan pohon secara umum. Untuk itulah perlu strategi khusus untuk membuat pohon yang lebih sederhana, namun masih memiliki kemampuan prediksi yang memuaskan. Menghentikan proses pemisahan simpul dengan kedalaman tertentu bisa dijadikan

alternatif strategi. Alternatif yang lain adalah melakukan pemangkasan (*pruning*) terhadap pohon yang dihasilkan dan melakukan evaluasi terhadap kemampuan prediksi melalui proses validasi.

Apa itu validasi? Apa itu pemangkasan?

Validasi, secara umum, merupakan proses yang dilakukan untuk meyakinkan kita bahwa model yang diperoleh mampu memberikan kemampuan prediksi yang baik, termasuk pada data-data lain yang tidak digunakan dalam pembuatan model. Yang dimaksud dengan model pada diskusi dalam tulisan ini adalah pohon klasifikasi. Prosedur yang dilakukan secara sederhana dapat diuraikan sebagai berikut. Andaikan kita memiliki suatu gugus data yang akan digunakan dalam membangun pohon klasifikasi. Gugus data tersebut selanjutnya dibagi menjadi dua bagian, yang umumnya dilakukan secara acak. Bagian pertama, dikenal sebagai gugus data *in-sample* adalah gugus data yang digunakan untuk memperoleh pohon klasifikasi. Gugus data ini sering juga disebut gugus data latih atau *training set*. Begitu pohon klasifikasi telah diperoleh selanjutnya pohon tersebut diterapkan untuk memprediksi data pada gugus kedua. Karena pada gugus kedua ini nilai dari variabel target sudah diketahui maka kita dapat mengetahui seberapa baik prediksi yang dihasilkan. Gugus data kedua ini dikenal sebagai gugus data *out-sample* atau gugus data validasi *validation set*.

Pemangkasan pada dasarnya adalah membatalkan proses pemisahan simpul. Andaikan dari gugus data latih dibangun pohon klasifikasi lengkap sehingga setiap simpul tidak lagi dapat dilakukan pemisahan. Pemangkasan pertama dilakukan dengan membatalkan proses pemisahan yang paling akhir. Pemangkasan kedua dilakukan dengan membatalkan proses pemisahan kedua terakhir, dan seterusnya. Pohon-pohon yang diperoleh dari setiap pemangkasan dapat dievaluasi tingkat kemampuan prediksinya. Pohon hasil pemangkasan yang memberikan ketepatan prediksi yang baik

Pola umum dari hubungan antara ukuran pohon dan ketepatan prediksi disajikan pada Gambar 7. Pada data latih atau *in-sample* semakin besar pohon yang berarti semakin banyak simpul maka semakin tinggi ketepatan prediksinya. Namun pada data validasi atau *out-sample* kenaikan ketepatan prediksi hanya terjadi sampai titik tertentu dari ukuran pohon dan kemudian menurun jika pohon tumbuh semakin besar. Pada titik itulah kita dapat menentukan bagaimana sebaiknya pohon itu dipangkas, sehingga pohon akhir yang dihasilkan bukanlah pohon yang sangat besar.

5.3 Bagaimana Mengakomodir Variabel Prediktor yang Bersifat Numerik?

Dalam banyak kasus nyata, variabel prediktor yang terlibat ada yang bersifat kategorik dan ada pula yang numerik. Sejauh ini kita mendiskusikan situasi dimana seluruh prediktor bersi-

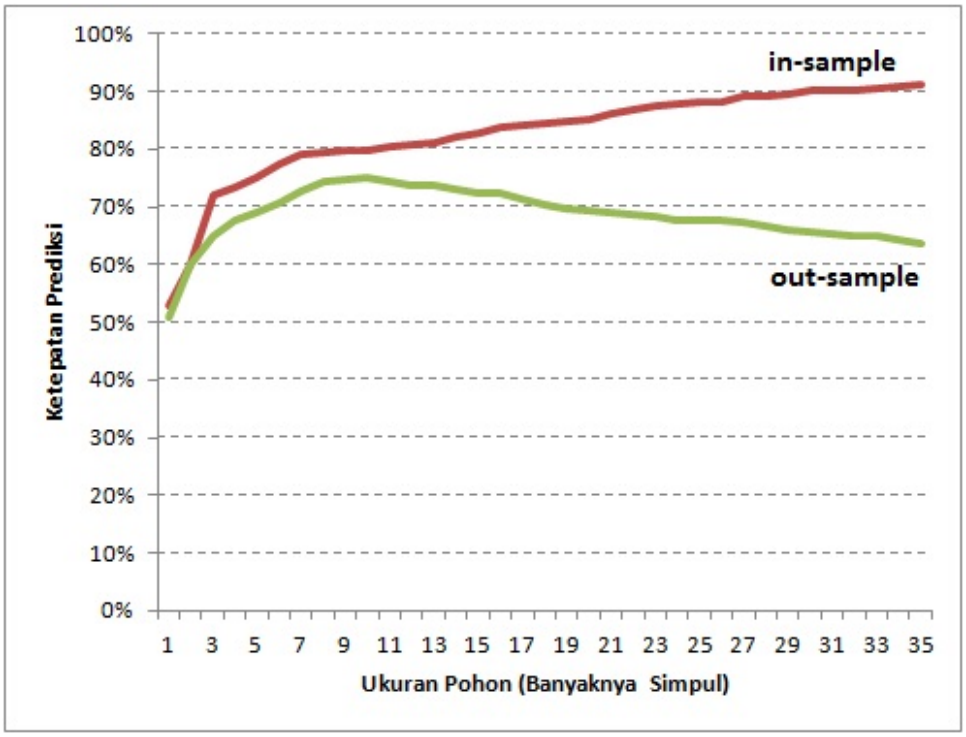

Figure 7: Pola umum hubungan antara ukuran pohon dan ketepatan prediksi

fat kategorik. Pada kasus variabel prediktor bersifat numerik, penerapan algoritma ID3 dapat secara langsung dilakukan dengan cara terlebih dahulu melakukan kategorisasi terhadap variabel numerik tersebut. Secara implisit sebenarnya kita sudah melakukan pada ilustrasi awal yang melibatkan variabel *Tinggi Badan* dan *Panjang Rambut*. Variabel Tinggi Badan yang barangkali data awalnya merupakan bilangan dengan satuan centimeter kemudian dikelompokkan menjadi tiga kelas yaitu Tinggi, Sedang, dan Pendek. Demikian juga dengan variabel Panjang Rambut yang kita kelompokkan menjadi dua kelas yaitu Panjang dan Pendek.

Pertanyaan yang muncul berikutnya adalah bagaimana melakukan kategorisasi tersebut? Berapa kelompok nilai? Pada nilai berapa batas masing-masing kelompok? Dalam beberapa literatur dan software, pengkategorian nilai dari variabel numerik ini dikenal juga sebagai *discretization* dan *binning*.

Pertama mengenai banyaknya kelompok. Secara prinsip tidak ada panduan yang tegas mengenai berapa banyak kelompok yang perlu dibuat. Namun jika mau menerapkan algoritma dasar di atas, penulis menyarankan untuk tidak membuat terlalu banyak kelompok. Dua hingga empat kelompok secara umum akan memadai.

Tentang penentuan batas-batas nilai pengelompokan pun tidak ada patokan baku. Beberapa orang menggunakan cara apriori dengan menentukan batas sesuai dengan keinginan yang merujuk pada konsep tertentu. Beberapa orang lain melakukannya dengan menggunakan informasi data dan mengaitkannya dengan nilai variabel respon/target. Pendekatan yang disebutkan ter-

akhir itu dikenal dengan istilah *supervised binning*. Penentuan batas dilakukan sedemikian rupa sehingga antar kelompok memiliki proporsi nilai tertentu variabel target yang berbeda. Misalnya jika variabel targetnya adalah jenis kelamin seperti pada ilustrasi awal, penentuan batas panjang rambut dilakukan dengan mencari sedemikian rupa sehingga antar kelompok panjang rambut proporsi amatan berjenis kelamin perempuan berbeda-beda. Salah satu teknik yang dapat dilakukan adalah ChiMerge. Teknik ini diawali dengan membagi nilai-nilai prediktor menjadi 10-15 kelompok dengan lebar interval yang kurang lebih seragam. Kemudian untuk setiap interval yang bersebelahan dilakukan pengujian Chi-Square test untuk memeriksa apakah proporsinya sama. Jika uji menyatakan bahwa tidak ada perbedaan proporsi maka kedua interval atau kelompok itu digabungkan (*merged* sehingga pada akhirnya akan diperoleh banyaknya kelompok yang lebih sedikit.

Yang kita diskusikan mengenai variabel prediktor ini adalah bahwa perlu tahapan preprocessing berupa diskretisasi terlebih dahulu terhadap variabel prediktor numerik sebelum diterapkan algoritma ID3 untuk menyusun pohon klasifikasi. Beberapa algoritma pembuatan pohon regresi tidak memerlukan proses diskretisasi secara terpisah. Proses diskretisasi menyatu dengan proses pembentukan pohon, seperti pada algoritma C4.5 dan QUEST yang akan dibahas pada bagian lain nanti.

5.4 Ukuran atau Kriteria Lain dalam Memilih Variabel Prediktor Pemisah Simpul

Salah satu kritik yang sering dikemukakan orang terkait dengan penggunaan *information gain* sebagai ukuran memilih pemisahan terbaik adalah kecenderungan lebih besar terpilihnya variabel prediktor yang memiliki banyak nilai dibandingkan variabel prediktor lain yang hanya memiliki sedikit kemungkinan nilai. Sebut saja misalnya tinggi badan yang dinyatakan dalam bilangan bulat bersatuan centimeter adalah variabel prediktor dalam suatu kasus. Jika variabel ini digunakan langsung dengan setiap nilai tinggi badang dijadikan satu kelompok, maka akan ada puluhan kelompok tinggi badan. Penggunaan variabel ini sebagai pemisah akan memungkinkan memperoleh simpul-simpul baru yang yang berisi sedikit-sedikit amatan (bahkan mungkin hanya berisi satu amatan) sehingga nilai entropinya kecil-kecil dan akibatnya nilai *information gain* menjadi besar. Karena besar, maka variabel tinggi badan akan terpilih sebagai pemisah dan kemudian simpul-simpul yang dihasilkan tidak perlu pemisahan lagi. Pohon yang dihasilkan akan sangat lebar dengan kedalaman hanya 1. Pohon seperti ini akan memiliki ketepatan akurasi yang baik pada data latih namun akan rendah akurasinya pada data-data lain.

Ukuran lain digunakan oleh Quinlan (1986) adalah *Gain Ratio* yang didefinisikan sebagai

$$Gain\ Ratio(D,V) = \frac{Gain(D,V)}{Split\ Information(D,V)} \tag{7}$$

dengan $Gain(D,V)$ adalah nilai *information gain* pemisahan simpul atau gugus data D menggunakan variabel prediktor V, sedangkan $Split\ Information(D,V)$ didefinisikan sebagai

$$Split\ Information(D,V) = -\sum_{i=1}^{k} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|} \tag{8}$$

dengan k adalah banyaknya kategori dari variabel V.

Ukuran lain apa ya? Pada tulisan lain yang mengupas beberapa algoritma selain ID3 akan dipaparkan ukuran-ukuran lain yang digunakan oleh para pengembang teknik ini.

5.5 Variabel Target Memiliki Lebih dari Dua Kelas

Dalam beberapa kasus tentu saja sangat mungkin dihadapi permasalahan yang melibatkan variabel target yang memiliki tiga, empat atau lebih kelas. Algoritma yang telah dipaparkan dapat dengan mudah diadaptasi untuk diterapkan dengan mengganti cara penghitungan entropy dalam bentuk

$$Entropy(D) = \sum_{i=1}^{c} -p_i \log_2 (p_i) \tag{9}$$

dengan c adalah banyaknya kelas dari variabel target.

5.6 Penanganan Data yang Hilang pada Variabel Prediktor

Suatu saat kita akan berhadapan dengan data yang memiliki nilai yang hilang atau tidak lengkap pada beberapa amatan untuk satu atau lebih variabel. Terdapat beberapa pendekatan penanganan yang dilakukan untuk hal tersebut, dan akan disinggung tidak terlalu detail pada tulisan ini.

Pendekatan pertama adalah pendekatan yang disebut sebagai marginalisasi. Pendekatan ini dilakukan dengan cara tidak menyertakan amatan (baris data) atau variabel (kolom data) yang mengandung data hilang. Dengan kata lain menghapus baris atau kolom yang memuat data yang tidak lengkap. Yang lebih banyak dilakukan adalah yang pertama yaitu membuang baris atau amatan karena biasanya baris atau amatan yang dimiliki tersedia cukup banyak. Membuang variabel lebih tidak disukai karena kita akan kehilangan informasi yang berhubungan dengan variabel tersebut.

Pendekatan kedua adalah dengan melakukan proses *imputasi* yaitu mengisi data yang hilang itu dengan nilai tertentu. Untuk variabel prediktor kategorik, kadang-kadang diisi dengan nilai yang paling banyak. Untuk variabel numerik, dapat diisi dengan nilai rata-rata dari amatan

yang lengkap. Imputasi yang lebih rumit dapat dilakukan dengan mencari amatan-amatan yang karakteristiknya mirip dengan amatan yang tidak lengkap dan kemudian mengisi data yang hilang dengan melakukan operasi (baik data terbanyak atau rata-rata) hanya menggunakan subset daya yang mirip tadi.

Pendekatan yang lain yang dapat juga dilakukan, terutama jika data yang hilang frekuensinya agak banyak adalah dengan menjadikan nilai *hilang* sebagai kategori baru. Sehingga misalnya variabel jenis kelamin ada tiga kategori yaitu laki-laki, perempuan dan "hilang".

5.7 Beberapa Algoritma Lain Penyusunan Pohon Klasifikasi

Terdapat beberapa algoritma lain yang dapat ditemui di literatur terkait dengan proses penyusunan pohon klasifikasi. Algoritma-algoritma tersebut antara lain adalah

- CHAID - ChiSquare Automatic Interaction Detection (Kass, G. V. 1980. An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Applied Statistics*, 20, 2, 119-127.)
- CART (Breiman, L., Friedman, J.H., Olshen, R., and Stone, C.J., 1984. *Classification and Regression Tree*. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific California.)
- Algoritma C4.5 (Quinlan, J. R. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers.) yang kemudian berkembang menjadi See5 atau C5.0
- QUEST - Quick, Unbiased, Efficient, Statistical Tree (Loh, W. Y. and Shih, Y. S., 1997. Split selection methods for classification trees. *Statistica Sinica*, Vol. 7, p. 815 - 840.)

Penjelasan mengenai algoritma-algoritma di atas akan diberikan pada tulisan lain.

6 Penilaian Kebaikan Pohon Klasifikasi

Telah dipaparkan bahwa pohon klasifikasi yang sudah terbentuk dapat digunakan untuk melakukan prediksi nilai variabel target dari suatu individu baru jika karakteristik dari individu tersebut diketahui. Karakteristik yang dimaksud adalah karakteristik yang tercermin dalam nilai-nilai variabel prediktor yang digunakan dalam penyusunan pohon klasifikasi.

Sekarang perhatikan Gambar 8 yang merupakan pohon klasifikasi yang serupa dengan ilustrasi awal sebelumnya yang digunakan untuk menduga jenis kelamin seseorang berdasarkan karakterstik Panjang Rambut, Tinggi Badan, dan Menggunakan Anting. Pada pohon tersebut, terdapat enam simpul akhir dengan tiga simpul berakhir pada kesimpulan jenis kelamin perempuan dan tiga simpul yang berakhir dengan kesimpulan jenis kelamin laki-laki.

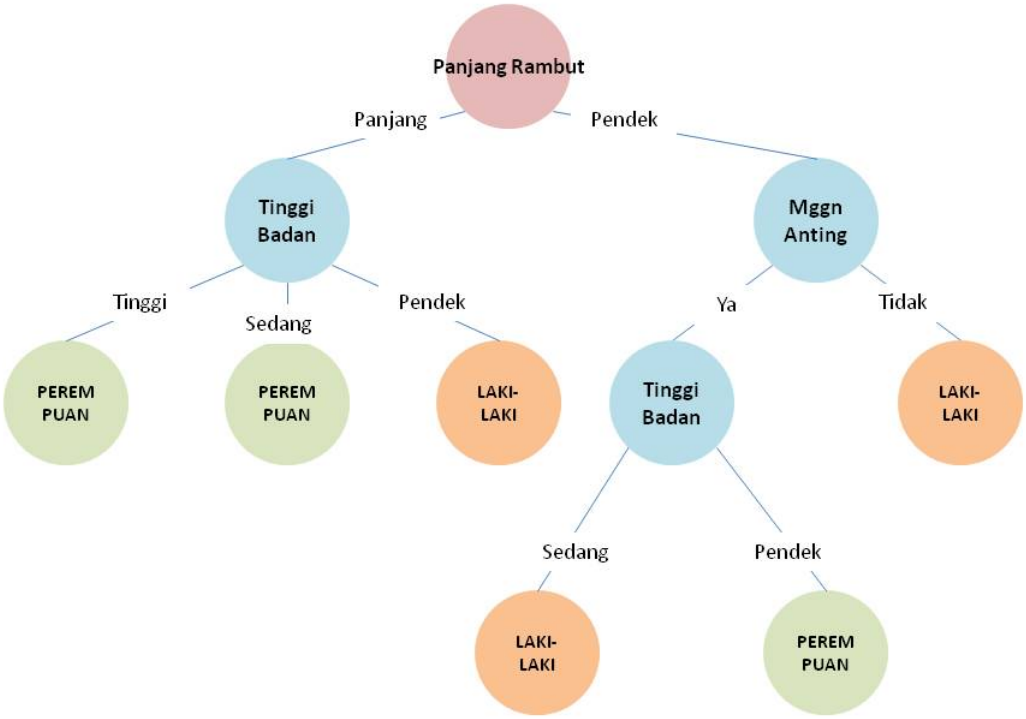

Figure 8: Pohon klasifikasi akhir yang dihasilkan

Pertanyaan yang perlu dijawab setelah kita memiliki suatu pohon klasifikasi adalah: *Seberapa baik pohon klasifikasi tersebut mampu memberikan prediksi yang akurat?* Akurasi prediksi tidak lain adalah kemampuan pohon klasifikasi menghasilkan prediksi sesuai dengan kenyataannya. Dengan demikian perlu dilakukan pemeriksaan dengan cara memprediksi nilai variabel target suatu individu yang sebenarnya diketahui nilai aktualnya.

Bayangkan bahwa pohon klasifikasi pada Gambar 8 diperoleh dengan menggunakan data pada lima kolom pertama dari Tabel 2. Karena itu adalah data yang digunakan untuk menyusun pohon maka tentu saja pada gugus data tersebut setiap individu telah diketahui nilai *Jenis Kelamin* yang sebenarnya dan itu tertera pada kolom kedua dalam tabel. Tiga kolom berikutnya dalam tabel adalah variabel-variabel prediktor yang digunakan. Jika nilai-nilai prediktor kemudian dievaluasi menggunakan pohon yang ada, maka akan diperoleh prediksi seperti yang tertera pada kolom terakhir dalam tabel itu.

Secara umum, pohon klasifikasi dikatakan baik apabila menghasilkan nilai-nilai pada kolom *Prediksi* yang identik dengan nilai-nilai pada kolom *Jenis Kelamin*. Perbedaan antara kedua kolom itu menunjukkan kejadian kesalahan pengklasifikasian atau *misclassification*. Berdasarkan Tabel 2 terlihat ada 3 (tiga) dari 20 amatan yang terjadi salah klasifikasi dan 17 amatan lainnya terklasifikasi dengan benar. Dengan demikian kita dapat menghitung tingkat kesalahan

Table 2: Data ilustrasi dan hasil prediksinya berdasarkan pohon klasifikasi pada Gambar 8

Nomor Nomor	Jenis Kelamin	Menggunakan Anting	Tinggi Badan	Panjang Rambut	Prediksi
1	Perempuan	Ya	Sedang	Panjang	Perempuan
2	Laki-Laki	Tidak	Tinggi	Pendek	Laki-Laki
3	Laki-Laki	Tidak	Pendek	Pendek	Laki-Laki
4	Laki-Laki	Tidak	Tinggi	Pendek	Laki-Laki
5	Perempuan	Tidak	Sedang	Panjang	Perempuan
6	Laki-Laki	Ya	Pendek	Pendek	Perempuan ***
7	Perempuan	Tidak	Pendek	Panjang	Laki-Laki ***
8	Laki-Laki	Tidak	Sedang	Pendek	Laki-Laki
9	Laki-Laki	Tidak	Tinggi	Pendek	Laki-Laki
10	Laki-Laki	Tidak	Sedang	Pendek	Laki-Laki
11	Laki-Laki	Tidak	Sedang	Pendek	Laki-Laki
12	Perempuan	Ya	Sedang	Pendek	Laki-Laki ***
13	Perempuan	Ya	Sedang	Panjang	Perempuan
14	Laki-Laki	Tidak	Pendek	Panjang	Laki-Laki
15	Laki-Laki	Tidak	Sedang	Pendek	Laki-Laki
16	Laki-Laki	Tidak	Sedang	Pendek	Laki-Laki
17	Laki-Laki	Tidak	Tinggi	Pendek	Laki-Laki
18	Laki-Laki	Tidak	Pendek	Pendek	Laki-Laki
19	Laki-Laki	Tidak	Sedang	Pendek	Laki-Laki
20	Perempuan	Ya	Tinggi	Panjang	Perempuan

klasifikasi sebesar $\frac{3}{20}$ atau 15%, serta tingkat ketepatan klasifikasi sebesar $\frac{17}{20}$ atau 85%.

Ukuran paling sederhana yang bisa digunakan untuk menilai kebaikan pohon klasifikasi adalah ukuran yang telah digunakan di atas yaitu **tingkat ketepatan pengklasifikasian** atau *correct classification rate*. Beberapa penulis menyebutnya sebagai *accuracy rate*. Ukuran ini didefinisikan sebagai

$$Correct\,Classification\,Rate = \frac{banyaknya\,amatan\,yang\,tepat\,prediksinya}{banyaknya\,seluruh\,amatan}. \tag{10}$$

Nilai CCR ini berkisar antara 0 dan 1, dan nilai yang semakin besar adalah situasi yang diharapkan. Sedangkan negasi dari ukuran ini dikenal sebagai **tingkat kesalahan klasifikasi** atau *misclassification rate* yaitu

$$Misclassification\,Rate = 1 - Correct\,Classification\,Rate. \tag{11}$$

Perhatikan kembali keberadaan data nilai target yang sebenarnya dengan nilai prediksi. Struktur dari kedua variabel tersebut dapat disusun dalam sebuah tabulasi silang dalam bentuk

sebagai berikut:

Table 3: Tabel klasifikasi berdasarkan ilustrasi pada Tabel 2

	prediksi		Total
Aktual	Laki-Laki	Perempuan	
Laki-laki	13	1	14
Perempuan	2	4	6
Total	15	5	20

Tabulasi silang di atas mendeskripsikan banyaknya amatan dari setiap kondisi. Terdapat 13 amatan yang aktualnya Laki-Laki dan diprediksi sebagai Laki-Laki, terdapat 1 amatan yang aktualnya Laki-Laki namun diprediksi sebagai Perempuan, dan seterusnya. Yang ada pada diagonal utama adalah kondisi dimana nilai prediksi sesuai dengan nilai aktualnya. Sedangkan yang lain adalah kondisi salah klasifikasi. Berdasarkan tabel tersebut kita dapat dengan mudah melihat seberapa besar tingkat ketepatan maupun tingkat kesalahan klasifikasi. Terdapat banyak istilah penamaan bagi tabel itu. Istilah-istilah yang banyak digunakan antara lain adalah *classification table*, *confusion table* dan *coincidence table*. Secara subjektif, penulis lebih menyukai penggunaan istilah tabel klasifikasi atau *classification table*.

Perhatikan baris pertama dari Tabel 3. Baris itu menggambarkan kondisi hasil prediksi terhadap amatan-amatan yang nilai target aktualnya adalah Laki-Laki. Dari baris tersebut dapat terbaca bahwa di dalam data terdapat 14 amatan Laki-Laki, dan 13 diantaranya terprediksi dengan tepat serta 1 amatan salah prediksi. Dengan demikian maka tingkat ketepatan prediksi pada kelompok laki-laki adalah $\frac{13}{14}$ atau 92.86%. Dalam topik peluang, ini adalah peluang bersyarat yaitu

$$P(tepat\ klasifikasi|Y = Laki - Laki) = \frac{13}{14} = 92.86\%.$$

Sementara itu dari baris kedua, kita dapat menghitung tingkat ketepatan lain yaitu tingkat ketepatan prediksi amatan yang Perempuan

$$P(tepat\ klasifikasi|Y = Perempuan) = \frac{4}{6} = 66.67\%.$$

Sampai saat ini kita mengenal tiga buah ukuran ketepatan yaitu: (1) tingkat ketepatan secara keseluruhan sebesar 85%, (2) tingkat ketepatan pada amatan Laki-Laki sebesar 92.86%, dan (3) tingkat ketepatan pada amatan Perempuan sebesar 66.67%. Tentu saja intuisi kita akan mengatakan bahwa pohon klasifikasi yang baik adalah pohon yang menghasilkan ketiga ukuran tadi sebesar-besarnya.

Selanjutnya kita akan coba formulasikan ketiga ukuran di atas dengan lebih formal. Pada kasus nilai variabel target memiliki dua nilai kita dapat kodekan menggunakan 1 dan 0. Bentuk

dari tabel klasifikasi secara umum dapat dituliskan menjadi seperti pada Tabel 4, dengan n_{ij} adalah banyaknya amatan yang aktualnya bernilai i dan diprediksi sebagai j. Notasi titik pada bagian kanan dan bawah tabel merupakan notasi penjumlahan dengan $n_{i*} = \sum_j n_{ij}$ dan $n_{*j} = \sum_i n_{ij}$, serta $n_{**} = \sum_i \sum_j n_{ij}$

Table 4: Tabel klasifikasi umum dengan target berupa variabel dua nilai

	prediksi		Total
Aktual	1	0	
1	n_{11}	n_{10}	n_{1*}
0	n_{01}	n_{00}	n_{1*}
Total	n_{*1}	n_{*0}	n_{**}

Berdasarkan penjelasan di atas, maka banyaknya amatan yang diklasifikasikan dengan tepat adalah sejumlah n_{11} dan n_{00}. Dengan demikian tingkat ketepatan klasifikasi dapat dituliskan sebagai

$$Correct\,Classification\,Rate = \frac{n_{11} + n_{00}}{n_{**}}. \tag{12}$$

Jika kelas target 1 merupakan kelas yang menjadi pusat perhatian sehingga ketepatan akurasi memprediksi amatan yang aktualnya bernilai 1 dianggap lebih penting maka nilai $P(tepat|Y = 1)$ dikenal sebagai *sensitivity* sedangkan $P(tepat|Y = 0)$ dikenal sebagai *specificity*. Selanjutnya kita dapat menuliskan keduanya sebagai

$$Sensitivity = \frac{n_{11}}{n_{11} + n_{10}} = \frac{n_{11}}{n_{1*}}, \tag{13}$$

dan

$$Specificity = \frac{n_{00}}{n_{01} + n_{00}} = \frac{n_{00}}{n_{0*}}. \tag{14}$$

Penilaian yang lebih objektif terhadap kebaikan atau kemampuan prediksi dari suatu pohon klasifikasi adalah dengan melakukan penghitungan nilai *correct rate*, *sensitivity*, dan *specificity* tidak didasarkan pada gugus data yang sama dengan yang digunakan dalam membangun pohon. Penilaian sebaiknya dilakukan dengan melibatkan gugus data lain yang sudah diketahui nilai dari variabel targetnya. Proses demikian ini dikenal sebagai proses **validasi**.

Secara sederhana, proses validasi dilakukan dengan cara sebagai berikut. Andaikan kita memiliki suatu gugus data $\mathcal{G}$. Sebelum melakukan pembuatan pohon, gugus data tersebut dibagi menjadi dua bagian yaitu $\mathcal{T}$ dan $\mathcal{V}$ dengan $\mathcal{G} = \mathcal{T} \cup \mathcal{V}$. Gugus data $\mathcal{T}$ adalah gugus data yang digunakan untuk menyusun pohon klasifikasi, dan selanjutnya pohon klasifikasi diterapkan untuk menduga kelas target dari amatan-amatan pada gugus data $\mathcal{V}$. Berdasarkan nilai aktual dan dugaan kelas pada gugus $\mathcal{V}$ inilah ditentukan ukuran kebaikan prediksi.

Gugus data $\mathcal{T}$ dikenal sebagai gugus data *latih* atau *training set* atau *in-sample*, sedangkan gugus data $\mathcal{V}$ disebut sebgai gugus data *validasi* atau *validation set* atau *out-sample*. Perbandingan banyaknya amatan pada kedua gugus tidak ada aturan yang baku, tetapi dalam praktek sering digunakan pemisahan $70:30$ atau $80:20$.