## Algortitma Decision Tree C4.5

Algoritma data mining C4.5 merupakan salah satu algoritma yang digunakan untuk melakukan klasifikasi dan bersifat prediktif. Salah satu proses data mining adalah klasifikasi, proses ini bertujuan untuk menemukan pola yang berharga dari data yang memiliki ukuran relatif besar hingga sangat besar. Algortima C4.5 merupakan pengembangan dari [algortima Iterative Dichotomiser 3 (ID3](https://informatikalogi.com/algoritma-id3/)). Algoritma ID3 merupakan algoritma yang dapat dipergunakan dalam membangun sebuah pohon keputusan (*decision tree*). Algoritma ini ditemukan oleh J. Ross Quinlan (1979), dengan memanfaatkan Teori Informasi atau Information Theory milik Shanon. ID3 sendiri merupakan singkatan dari.

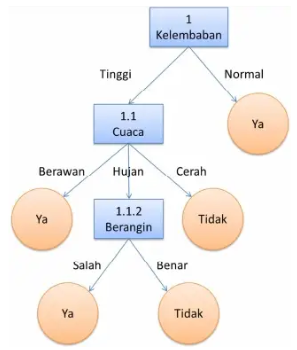
1. **Decision Tree**

Decision tree adalah salah satu metode klasifikasi dengan menggunakan struktur hierarki dalam pengambilan kelas atau keputusan akhir. Proses dari decision tree dimulai dari root node hingga leaf node yang dilakukan secara rekursif. Di mana setiap percabangan menyatakan suatu kondisi yang harus dipenuhi dan pada setiap ujung pohon menyatakan kelas dari suatu data.

Proses dalam decision tree yaitu mengubah data dalam bentuk tabel menjadi model pohon (tree) yang kemudian model pohon tersebut diubah menjadi serangkaian aturan (rule). Tabel 2.2 merupakan data yang telah diketahui atribut dan klasifikasi hasil akhir. Tabel ini akan dirubah menjadi sebuah pohon keputusan seperti gambar 2.4

Tabel 2‑2 Rules Data Keputusan

Salah satu kelemahan algoritma dari decision tree melalui pendekatan proses perubahannya yaitu faktor skalabilitas dimana algoritma tersebut hanya dapat digunakan untuk menangani sampel-sampel yang dapat disimpan secara keseluruhan dan pada waktu yang bersamaan di memori.



Gambar2‑3‑4 Algoritma C4.5 Decision Tree

1. **Arsitektur Decision Tree**

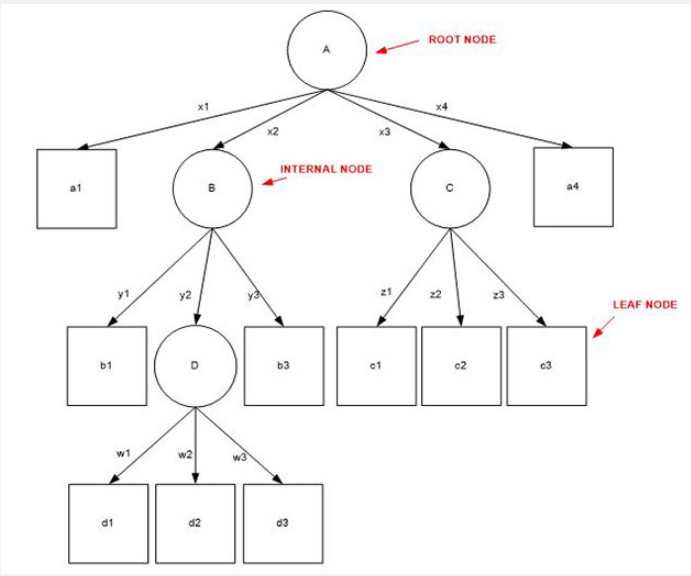
Arsitektur pohon keputusan menyerupai bentuk pohon, dimana pada umumnya sebuah pohon terdapat akar (root), cabang dan daun (leaf). Pada pohon keputusan mempunyai tiga elemen yang sama sebagai berikut :

Root node (node akar) merupakan attribut node yang terletak paling atas dari suatu pohon keputusan.

Internal Node adalah node percabangan yang memiliki satu input dan minimal dua output.

Leaf Node adalah node akhir yang hanya memiliki satu input, dan tidak memiliki output. Pada pohon keputusan setiap leaf node menandai label kelas.

Pada pohon keputusan di setiap percabangan menyatakan kondisi yang harus dipenuhi dan tiap ujung pohon menyatakan nilai kelas data. Gambar 2.5 merupakan bentuk arsitektur pohon keputusan.



Gambar2‑3‑5 Arsitektur Decision Tree

1. **Konstruksi Pohon Keputusan dengan Algoritma ID3**

Algoritma ID3 merupakan algoritma data mining yang dapat mengumpulkan infromasi baru dari kumpulan data yang tersimpan didalam basis data. Konstruksi dari algoritma ini yaitu pohon keputusan yang dapat menghasilkan suatu keputusan. langkah-langkah dalam konstruksi pohon keputusan adalah sebagai berikut :

Langkah 1 : Pohon dimulai dengan sebuah simpul yang mereperesentasikan sampel data pelatihan yaitu dengan membuat simpul akar.

Langkah 2 : Jika semua sampel berada dalam kelas yang sama, maka simpul ini menjadi daun dan dilabeli menjadi kelas. Jika tidak, information gain akan digunakan untuk memilih atribut terbaik dalam memisahkan data sampel menjadi kelas-kelas individu.

Langkah 3 : Cabang akan dibuat untuk setiap nilai pada atribut dan data sampel akan dipartisi lagi.

Langkah 4 : Algoritma ini menggunakan proses rekursif untuk membentuk pohon keputusan pada setiap data partisi. Jika sebuah atribut sduah digunakan disebuah simpul, maka atribut ini tidak akan digunakan lagi di simpul anak-anaknya.

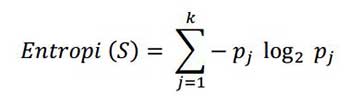
Langkah 5 : Proses ini berhenti jika dicapai kondisi seperti berikut :

* Semua sampel pada simpul berada di dalam satu kelas
* Tidak ada atribut lainnya yang dapat digunakan untuk mempartisi sampel lebih lanjut. Dalam hal ini akan diterapkan suara terbanyak. Hali Ini berarti sebuah simpul dirubah menjadi daun serta melabelinya dengan kelas pada suara terbanyak.

1. **Entropy dan Information Gain**

Algoritma pada metode ini menggunakan konsep dari entropi. Konsep Entropi yang digunakan untuk mengukur “seberapa informatifnya” sebuah node tersebut.

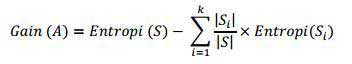
Entropi(S) = 0, jika semua contoh pada S berada dalam kelas yang sama.  
Entroiy(S) = 1, jika jumlah contoh positif dan jumlah contoh negatif dalam S adalah sama.  
0 < Entropi(S) < 1, jika jumlah contoh positif dan negatif dalam S tidak sama.



Gambar 2‑3‑6 Rumus Entropi

Dimana:  
• S adalah himpunan (dataset) kasus  
• k adalah banyaknya partisi S  
• pj adalah probabilitas yang di dapat dari Sum(Ya) dibagi Total Kasus.

Setelah mendapat nilai entropi, pemilihan atribut dilakukan dengan memilih nilai information gain terbesar.



Gambar 2‑3‑7 Rumus Gain

Dimana:  
S = ruang (data) sample yang digunakan untuk training.  
A = atribut.  
|Si| = jumlah sample untuk nilai V.  
|S| = jumlah seluruh sample data.  
Entropi(Si) = entropy untuk sample-sample yang memiliki nilai i