Nama : Indira Septianita Larasati

NIM : H071191023

Decision tree adalah model prediksi menggunakan struktur pohon atau struktur berhirarki. *Decision tree* digunakan untuk pengenalan pola termasuk pola pengenalan pola secara statistik.

Tahapan algoritma klasifikasi yakni:

1. Konstruksi model yang dimana menguraikan suatu himpunan kelas yang telah ditentukan sebelumnya.
2. Penggunaan model digunakan untuk mengklasifikasi tuple data yang label kelasnya belum diketahui.

Proses Klasifikasi:

Data dibagi dua jenis yakni data training dan data testing. Data training digunakan membangun model. Data training dilakukan fitting/training menggunakan algoritma tertentu yang dimana proses fitting akan menghasilkan model. Sedangkan untuk data testing digunakan sebagai evaluasi model yang digunakan.

Entropy(S) adalah total bit yang diprediksi dibutuhkan untuk mengekstrak suatu kelas dari banyaknya data acak dalam suatu ruang sampel. Entropy dikatakan sebagai kebutuhan bit jika menyatakan suatu kelas. Semakin kecil nilai entropy maka semakin baik digunakan dalam mengekstraksi suatu data. Panjang kode dalam menyatakan informasi optimal yakni menggunakan

log_2.p bits untuk messages yang mempunyai probabilitas p sehingga banyaknya bit yang diperkirakan untuk mengekstraksi S ke dalam kelas adalah: $-p_+ log_2\, p_+ - p_- log_2\, p_-$

Uji coba chi square digunakan untuk menguji :

1. Uji x^2 untuk mengetahui ada tidaknya hubungan antara dua variabel (*independency test)*
2. Uji x^2 untuk mengetahui homogenitas antar bagian kelompok (*homogeneity test)*
3. Uji x^2 untuk mengetahui bentuk distribusi (*goodness of a bit)*

Adapun rumus yang digunakan:

$$X^2 = \sum_{i=1}^{k} \frac{(Oi - Ei)^2}{Ei}$$

Ket:

k : Banyaknya kategori dalam sel 1,2, …. k

O : frekuensi amatan

i : frekuensi observasi untuk kategori ke – i

Ei : frekuensi observasi untuk kategori ke -i dalam menghitung frekuensi ekspektasi dengan perbandingan nilai H_0 yang dimana derajat kebebasan db = (k-1).

Information grain mengukur seberapa relevan / berpengaruh sebuah feature terhadap hasil pengukuran. Penggunaan teknik ini dapat mereduksi dimensi feature dengan cara mengukur reduksi Entropy sebelum dan sesudah pemisahan. Information grain pada atribut A dihitung menggunakan berdasarkan :

$$Info_A\,(S) = \sum_{j=1}^{k} \frac{|S_j|}{|S|}\, Info(S_j)$$

$$Gain(A) = Info(S) - Info_A\,(S)$$