

Nama : Muhammad Alfisar Rachman
NIM : 1301164036
Kelas : IF-40-08

Problem :

Melakukan klasifikasi terhadap 40 data apakah 40 data tersebut termasuk kedalam kelas/label >50K atau <=50K , dimana sudah diberikan 160 data yang sudah di ketahui labelnya dan memiliki 7 atribut yaitu **age, workclass, education, marital-status, occupation, relationship, dan hours-per-week**, untuk 40 data yang akan diklasifikasikan belum di ketahui labelnya. Dalam tugas ini pencarian klasifikasi terhadap 40 data tersebut diharuskan menggunakan dan mengimplementasikan metode *Naive Bayes*.

Strategi Penyelesaian Masalah :

Strategi yang digunakan yaitu dengan mengimplementasikan *Naive Bayes* kedalam sebuah program sehingga program tersebut dapat menentukan label dari 40 data tersebut. Dalam program yang saya buat, saya mengambil seluruh data train dan seluruh data test untuk mencari label pada data test, cara untuk mendapatkan class/label yaitu dengan cara menghitung probabilitas kelas dengan cara mencari seluruh label >50K dan <=50K yang berada di data train setelah itu jumlah label >50K akan program saya bagikan dengan seluruh jumlah label >50K ditambah seluruh jumlah label <=50K dan untuk label <=50K juga sama caranya dengan cara mencari probabilitas class/label >50K. Setelah mendapatkan probabilitas class / label program saya akan mencari probabilitas setiap atribut terhadap class/label tersebut yang berada di data train, misalnya mencari probabilitas atribut young terhadap >50K yaitu menghitung jumlah atribut young yang memiliki label >50K dan akan di bagi dengan jumlah label >50K dan untuk mencari probabilitas atribut terhadap label <=50K sama langkahnya dengan probabilitas atribut terhadap label >50K. Setelah itu maka program akan mencari besar probabilitas class/label terhadap atribut di setiap baris data, cara untuk mendapatkannya yaitu dengan cara setiap 1 baris data test akan di kalikan semua atributnya terhadap class/label dan hasil dari perkalian tersebut akan di kalikan dengan probabilitas class/label tersebut. Setelah itu program akan sampai di step final yaitu membandingkan probabilitas class/label >50K terhadap atribut dengan probabilitas class/label <=50K terhadap atribut dan akan dipilih probabilitas yang paling besar. Berikut rumus yang di pakai untuk mendapatkan semua probabilitas tersebut:

$P(C_i)$ = jumlah class tersebut / jumlah seluruh data.

$P(X_1, X_2, \dots, X_n | C_i) = P(X_1 | C_i) * P(X_2 | C_i) * \dots * P(X_n | C_i)$.

$P(C_i | X_1, X_2, \dots, X_n) = P(X_1, X_2, \dots, X_n | C_i) * P(C_i)$.

Class/Label = Perbandingan antara $P(C_i | X_1, X_2, \dots, X_n)$ terhadap $P(C_i | X_1, X_2, \dots, X_n)$ dan diambil yang paling besar.

Nama : Muhammad Alfisar Rachman
NIM : 1301164036
Kelas : IF-40-08

Analisis :

Berdasarkan penjelasan yang saya jabarkan maka dapat diketahui bahwa pemakaian rumus untuk menghitung $P(C_i)$, $P(X_1, X_2, \dots, X_n | C_i)$, dan $P(C_i | X_1, X_2, \dots, X_n)$ ternyata sangat berperan penting terhadap penghasilan class/label yang akan dipilih, Jika terjadi kesalahan dalam proses perhitungan salah satu atau semua probabilitas maka pengambilan keputusan terhadap class/label yang akan dipilih cenderung akan salah, maka pemakaian rumus tersebut sangat penting terhadap hasil class/label yang akan di pilih.

Hasil Percobaan :

Akurasi:

```
mylinux@mylinux:~/Perkuliahan/semester 6/Malen/Machine-Learning$ cd "/  
nv "PYTHONIOENCODING=UTF-8" "PYTHONUNBUFFERED=1" /usr/bin/python3 /hom  
nFiles/ptvsd_launcher.py --default --client --host localhost --port 34  
g/akurasi2.py"  
83.125 %
```

Hasil Running program

<=50K	<=50K
<=50K	>50K
>50K	<=50K
<=50K	>50K
>50K	>50K
>50K	>50K
<=50K	>50K
<=50K	
>50K	
>50K	
>50K	
>50K	
<=50K	
>50K	
>50K	
>50K	
<=50K	
>50K	
<=50K	
>50K	
>50K	
>50K	
>50K	
>50K	
>50K	
>50K	
>50K	
>50K	
<=50K	
<=50K	
<=50K	
>50K	
>50K	
<=50K	
>50K	