

Nama : Gigas Taufan Arvyanto

NIM : 1301164211

Kelas : IF-40-12

## TUGAS 1 MACHINE LEARNING

### NAÏVE BAYES

#### 1. Identifikasi Masalah

Diberikan dua buah file yaitu TrainsetTugas1ML.csv dan TestsetTugas1ML.csv yang masing-masing terdapat 160 data dan 40 data. Maka dengan menggunakan naïve bayes akan diprediksi kelas income dari 40 data yang terdapat pada TestsetTugas1ML.csv.

#### 2. Desain Program

Naïve Bayes adalah metode yang digunakan untuk membangun sebuah klasifikasi data. Klasifikasi ini memanfaatkan perhitungan probabilitas dan statistik. Naïve bayes membentuk tabel probabilitas sebagai dasar proses klasifikasi income. Output dari program adalah hasil klasifikasi kelas income dari data test.

Berikut merupakan penerapan yang dilakukan penulis terhadap permasalahan yang diberikan:

##### a. Menghitung Probabilitas Untuk Setiap Kategori Trainset

Perhitungan probabilitas dilakukan untuk data di dalam Trainset. Untuk setiap kategori dilakukan perhitungan probabilitas. Terdapat 7 kategori di dalam Trainset, yakni Age, Workclass, Education, Marital-status, Occupation, Relationship, Hours-per-week, dan Income. Untuk kategori Age, Workclass, Education, Marital-status, Occupation, Relationship, Hours-per-week memiliki 3 kelas. Sementara untuk kategori Income memiliki 2 kelas.

Berikut kategori dan kelas pada Trainset:

Age	Workclass	Education	Marital-status
Young	Private	HS-grad	Married-civ-spouse
Adult	Local-gov	Bachelors	Never-married
Old	Self-emp-not-inc	Some-college	Divorced

Occupation	Relationship	Hours-Per-Week	Income
Craft-repair	Husband	Normal	>50K
Prof-specialty	Not-in-family	Low	<=50K
Exec-managerial	Own-child	Many	

Penghitungan menggunakan kategori Income untuk menjadi patokan. Setiap kelas dari 7 kategori selain Income dibagi menjadi 2 subkelas, yakni untuk subkelas >50K dan subkelas <=50K. Dua subkelas ini akan digunakan untuk membandingkan hasil probabilitas dan menentukan kelas data test.

**b. Menghitung Probabilitas Untuk Setiap Kategori Testset**

Setelah mendapat probabilitas untuk setiap kategori dan kelas di dalamnya, kemudian melakukan perhitungan terhadap Testset. Setiap data pada Testset dicari kategori dan kelasnya.

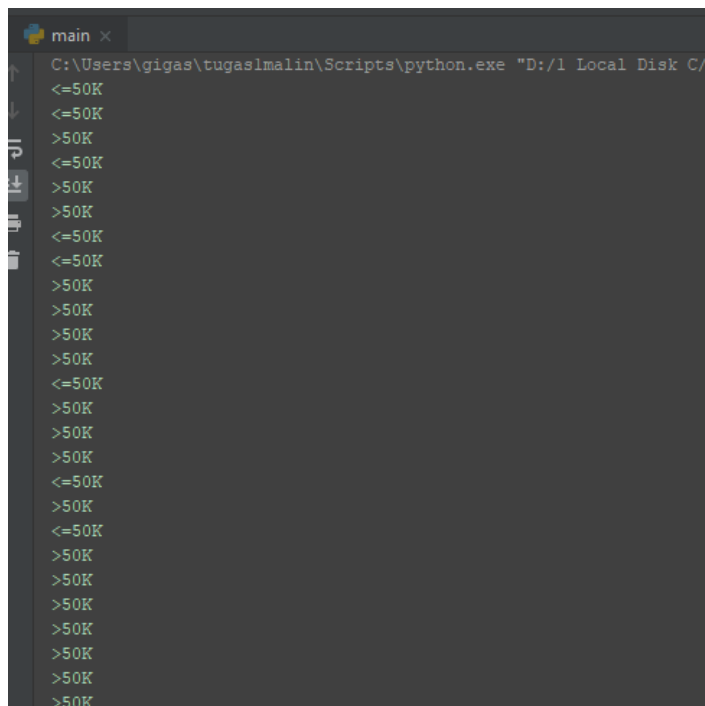
Jika data pada testset masuk ke dalam kategori dan kelas tertentu maka akan dihitung probabilitas apakah data tersebut masuk ke dalam kelas >50K atau <=50K pada kategori Income.

**c. Penentuan Kelas Pada Kategori Income**

Untuk menentukan kelas apa data tersebut adalah dengan mengalikan probabilitas dari setiap kelas dari tiap kategori dan subkelas tadi. Hasil dari perkalian tadi kemudian dibandingkan antara subkelas >50K dan subkelas <=50K. Hasil yang lebih tinggi menjadikan subkelas tersebut sebagai kelas sebenarnya dari data tersebut.

**3. Screenshot Output Program**

Dari penerapan yang dibuat dalam program didapatkan data kelas income untuk tiap baris pada data test. Berikut merupakan screenshot hasil output dari program:



```
main x
C:\Users\gigas\tugas\malin\Scripts\python.exe "D:/l Local Disk C/
<=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
```