Laporan Tugas 1 Machine Learning – Naïve Bayes

Diberikan sebuah Trainset berupa himpunan data berisi 160 objek data yang memiliki 7 atribut input (**age, workclass, education, marital-status, occupation, relationship, hours-per-week**) dan 1 output (label kelas **income**) yang memiliki ~~4 kelas/label (0, 1, 2, dan 3)~~ 2 kelas ('>50K' dan '<=50K'). Bangunlah sebuah sistem klasifikasi menggunakan metode **Naïve Bayes** untuk menentukan kelas/label data testing dalam Testset. Sistem membaca masukan file TrainsetTugas1ML.xlsx dan TestsetTugas1ML.xlsx dan mengeluarkan *output* berupa file **TebakanTugas1ML.xlsx** berupa satu kolom berisi **40 baris** yang menyatakan kelas/label baris yang bersesuaian pada file TestsetTugas1ML.xlsx

Analisa :

- Terdapat 160 data didalam TrainsetTugas1ML.xlsx
- Terdapat 120 data dengan label '>50K' dan 40 data dengan label '<=50K' pada TrainsetTugas1ML.xlsx
- Terdapat 40 data didalam TestsetTugas1ML.xlsx
- Tidak terdapat data yang membutuhkan *Laplacian Correction* dalam kasus yang diberikan, sehingga untuk meminimalisir waktu pengerjaan *Laplacian Correction* tidak di implementasikan didalam *Source Code.*
- **Proses Validasi tidak perlu dilakukan**, gunakan 160 datatrain untuk meng-kategorikan 40 datatest.
- Pada atribut input **age**, terdapat 3 tipe, yaitu : **old, adult, young**
- Pada atribut input **workclass**, terdapat 3 tipe, yaitu : **Local-gov, Private, Self-emp-not-inc**
- Pada atribut input **education**, terdapat 3 tipe, yaitu : **Bachelors, HS-grad, Some-college**
- Pada atribut input **marital-status**, terdapat 3 tipe, yaitu : **Divorced, Married-civ-spouse, Never-married**
- Pada atribut input **occupation**, terdapat 3 tipe, yaitu : **Craft-repair, Exec-managerial, Prof-specialty**
- Pada atribut input **relationship**, terdapat 3 tipe, yaitu : **Husband, Not-in-family, Own-child**
- Pada atribut input **hours-per-week**, terdapat 3 tipe, yaitu : **low, many, normal**
- **ID** tidak dimasukkan kedalam tuple, karena tidak termasuk atribut input.

Strategi Penyelesaian :

Mengambil contoh dari data pertama pada TestsetTugas1ML.xlsx, data sebagai berikut:

Age	Workclass	Education	Marital-status	Occupation	Relationship	Hours-per-week
Young	Private	HS-Grad	Never-married	Craft-repair	Not-in-family	Normal

Langkah pertama adalah untuk menemukan nilai dari probabilitas label, **P(Income='>50K')** dan **P(Income='<=50K')**, didapatkan :

P(Income='>50K') = 120/160 = 0.75
P(Income='<=50K') = 40/160 = 0.25

Lalu lakukan perhitungan probabilitas untuk semua atribut input (**age, workclass, education, marital-status, occupation, relationship, hours-per-week**). Setelah didapatkan semua probabilitas atribut input untuk **P(Income='>50K')** dan **P(Income='<=50K')**, lakukan cari **P(X|Income='>50K')** dan **P(X|Income='<=50K')**, didapatkan dua hasil, dan hasil yang terbesar adalah kelas label pada data tersebut, seperti contoh berikut:

P(X|Income='>50K') =

P(Age='Young'|Income='>50K') * P(workclass='Private'|Income='>50K') * P(Education='HS-grad'|Income='>50K') * P(Marital-status='never-married'|Income='>50K') * P(occupation='craft-repair'|Income='>50K') * P(relationship='Not-in-family'|Income='>50K') * P(hours-per-week='normal'|Income='>50K') * P(Income='>50K')

P(X|Income='<=50K') =

P(Age='Young'|Income='<=50K') * P(workclass='Private'|Income='<=50K') * P(Education='HS-grad'|Income='<=50K') * P(Marital-status='never-married'|Income='<=50K') * P(occupation='craft-repair'|Income='<=50K') * P(relationship='Not-in-family'|Income='<=50K') * P(hours-per-week='normal'|Income='<=50K') * P(Income='<=50K')

Income	Probabilitas	
>50K	120/160	0.75
<=50K	40/160	0.25

Age	Probabilitas Income			
	>50K		<=50K	
old	1/120	0.008333	1/40	0.008333
adult	53/120	0.441667	19/40	0.158333
young	66/120	0.55	20/40	0.5

education	Probabilitas Income			
	>50K		<=50K	
Bachelors	65/120	0.541667	7/40	0.175
HS-grad	28/120	0.233333	17/40	0.425
Some-college	27/120	0.225	16/40	0.4

occupation	Probabilitas Income			
	>50K		<=50K	
Craft-repair	32/120	0.266667	21/40	0.525
Exec-managerial	47/120	0.391667	15/40	0.375
Prof-specialty	41/120	0.341667	4/40	0.1

hours-per-week	Probabilitas Income			
	>50K		<=50K	
low	3/120	0.025	5/40	0.125
many	1/120	0.008333	2/40	0.05
normal	116/120	0.966667	33/40	0.825

workclass	Probabilitas Income			
	>50K		<=50K	
Local-gov	8/120	0.066667	1/40	0.025
Private	105/120	0.875	32/40	0.8
Self-emp-not-inc	7/120	0.058333	7/40	0.175

marital-status	Probabilitas Income			
	>50K		<=50K	
Divorced	5/120	0.041667	7/40	0.175
Married-civ-spouse	108/120	0.9	19/40	0.475
Never-married	7/120	0.058333	14/40	0.35

relationship	Probabilitas Income			
	>50K		<=50K	
Husband	107/120	0.891667	19/40	0.475
Not-in-family	11/120	0.091667	14/40	0.35
Own-child	2/120	0.016667	7/40	0.175

	Ganti class:	>50K	<=50K
Age	Young	0.55	0.5
workclass	Private	0.875	0.8
education	HS-grad	0.233333333	0.425
marital-status	Never-married	0.058333333	0.35
occupation	Craft-repair	0.266666667	0.525
relationship	Not-in-family	0.091666667	0.35
hours-per-week	Normal	0.966666667	0.825
Total		0.000116087	0.002254957
Income Class		<=50K	

Maka Kelas label Income dari contoh data yang diberikan adalah <=50K, karena setelah dilakukan perhitungan menggunakan naïve bayes, income label >50K bernilai 0.000116087, dan income label <=50K bernilai 0.002254957, **sehingga <=50K dijadikan label dari data karena memiliki nilai yang lebih besar**.