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’)** \*

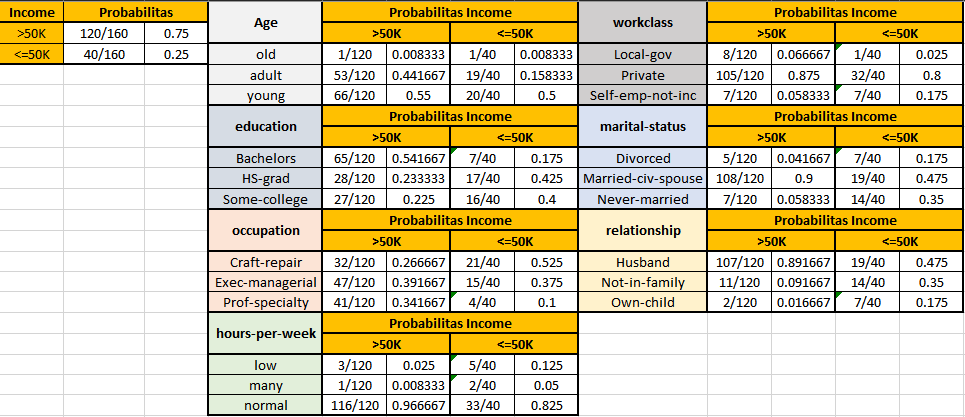
**P(Education=’HS-grad’|Income=’<=50K’)** \*

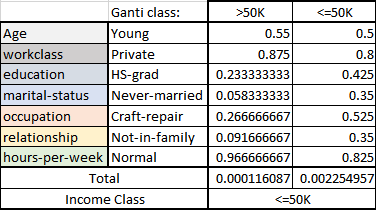
**P(Marital-status=’never-married’|Income=’<=50K’)** \*

**P(occupation=’craft-repair’|Income=’<=50K’)** \*

**P(relationship=’Not-in-family’|Income=’<=50K’)** \*

**P(hours-per-week=’normal’|Income=’<=50K’) \* P(Income=’<=50K’)**

****

****

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**.