**Laporan Tugas Pembelajaran Mesin Naive Bayes**

**Nama : Badrus Shoolehk Al Ar Fanny**

**NIM : 1301164131**

**Kelas : IF-40-04**

1. Deskripsi

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

Berdasarkan deskripsi permasalah di atas dapat disimpulkan bahwa:

1. Diberikan 160 Data yang memiliki 8 kategori *id, age, workclass, education, marital-status, occupation, relationship, dan hours-per-week*
2. **Proses Validasi tidak perlu dilakukan**, gunakan 160 datatrain untuk meng-kategorikan 40 datatest.
3. Pada atribut input **age**, terdapat 3 tipe, yaitu : **old, adult, young**
4. Pada atribut input **workclass**, terdapat 3 tipe, yaitu : **Local-gov, Private, Self-emp-not-inc**
5. Pada atribut input **education**, terdapat 3 tipe, yaitu : **Bachelors, HS-grad, Some-college**
6. Pada atribut input **marital-status**, terdapat 3 tipe, yaitu : **Divorced, Married-civ-spouse, Never-married**
7. Pada atribut input **occupation**, terdapat 3 tipe, yaitu : **Craft-repair, Exec-managerial, Prof-specialty**
8. Pada atribut input **relationship**, terdapat 3 tipe, yaitu : **Husband, Not-in-family, Own-child**
9. Pada atribut input **hours-per-week**, terdapat 3 tipe, yaitu : **low, many, normal**
10. Diberikan 40 Data yang tidak diketahui kelas *income*-nya
11. Menggunakan metode **Naive Bayes** buatlah sebuah model untuk memprediksi kelas dari suatu data
12. Kelas dari tiap data adalah ‘>50K’ atau ‘<=50K’

2. Strategi Pemodelan dan Analisis

2.1. Pre-processing

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

2.2. Pemodelan Naive Bayes

Pertama-tama akan dihitung probabilitas *Prior* yaitu probabilitas dari *dependent variable* dalam hal ini label pada fitur *income*. Dengan menghitung banyaknya kemunculan kategori >50K dan <=50K lalu dibagi dengan jumlah data.

Hasil dari tahap ini adalah sebagai berikut.



Selanjutnya akan dihitung probabilitas *Likelihood* untuk seluruh kategori pada tiap fitur. *Likelihood* adalah probabilitas suatu kejadian apabila kejadian tertentu dipenuhi. Pada kasus yang dapat kita hitung adalah probabilitas income diantara >50K dan <=50L terhadap kategori-kategori pada *independent variable*. Contohnya adalah probabilitas income >50K apabila nilai dari fitur *age* adalah *young*.

Algoritma yang digunakan adalah awalnya akan dipisahkan data yang memiliki label >50K dan <=50K. Lalu akan dihitung kemunculan pada tiap kategori dan dibagikan dengan total jumlah data pada tiap label.

Berikut tabel dari probabilitas *Likelihood*.

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’)

Setelah didapatkan *prior* dan *likelihood* maka dapat diprediksi kelas dari suatu data. Prediksi dapat dibuat dengan membandingkan antara P(‘>50K’|X) dengan P(‘<=50K’|X), probabilitas yang lebih besar lah yang akan menjadi kelas dari data tersebut.

2.3. Analisis Hasil Prediksi

Setelah model untuk memprediksi data dibangun, maka kita dapat memprediksi seluruh data pada testset dengan melakukan perhitungan *posterior*. Berikut adalah hasil prediksi beberapa testset.

