**Tugas 1, *Machine Learning***

**Laporan Membangun Sebuah Sistem Klasifikasi Untuk Menentukan Kelas/Label Data Testing Menggunakan Metode *Naïve Bayes***

Oleh:

Enrico Farizky Rustam ( 1301164263 )

IF 40-03 / S1 Informatika / Universitas Telkom

Abstrak: Himpunan data berisi beberapa objek data yang memiliki atribut input dan 1 output yang memiliki 4 kelas/label (0, 1, 2, dan 3) dengan menggunakan Metode *Naïve Bayes* . Metode *Naïve Bayes Classifiers* yaitu salah satu metode klasifikasi teks berdasarkan probabilitas kata kunci dalam membandingkan dokumen latih dan dokumen uji. Keduanya dibandingkan melalui beberapa tahap persamaan, yang akhirnya diperoleh hasil probabilitas tertinggi yang ditetapkan sebagai kategori dokumen baru. Hasil penelitian yang dilakukan oleh peneliti yaitu pengklasifikasian teks otomatis pelaporan dan pengaduan mayarakat dengan menggunakan metode *Naïve Bayes* *Classifiers* menghasilkan rata-rata akurasi yang tinggi, yaitu *recall* 93%, *precission* 90 %,dan *f-measure* 92%.

**Kata Kunci** : *Naïve Bayes*, dokumen latih, dokumen uji, probabilitas

# Pendahuluan

Pengklasifikasi bayes merupakan salah satu pengklasifikasi statistik, dimana pengklasifikasi ini dapat memprediksi probabilitas keanggotaan kelas suatu data tuple yang akan masuk ke dalam kelas tertentu, sesuai dengan perhitungan probabilitas. Pengklasifikasi Bayes didasari oleh teorema bayes yang ditemukan oleh Thomas Bayes pada abad ke-18. Dalam studi pembandingan algoritma klasifikasi telah ditemukan simple bayesian atau yang biasa dikenal dengan *Naïve Bayes Classifier*. *Naïve Bayes Classifier* menunjukkan akurasi dan kecepatan yang tinggi bila diterapkan pada database yang besar. Metode ini sering digunakan dalam menyelesaikan masalah dalam bidang mesin pembelajaran karena metode ini dikenal memiliki tingkat akurasi yang tinggi dengan perhitungan sederhana.

*Naïve Bayes* adalah pengklasifikasian statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu class. *Naïve Bayes* didasarkan pada teorema Bayes yang memiliki kemampuan klasifikasi serupa dengan *decision tree* dan *neural network. Naïve Bayes* terbukti memiliki akurasi

dan kecepatan yang tinggi saat diaplikasikan ke dalam database dengan data yang besar.

Teorema bayes merupakan formula dasar aturan dari *Naïve Bayes Classifier* dengan formula umum sebagai berikut:

***P(H|X)* =**

Dimana X merupakan data tuple hasil pengujian dari suatu set data yang telah ditentukan masuk ke ke dalam kelas tertentu. H merupakan suatu hipotesis yang akan menentukan X masuk ke dalam kelas C. P(H|X) merupakan peluang atau probabilitas X yang merupakan data tuple atau bukti yang diperoleh pada saat observasi masuk ke dalam kelas C, dengan kata lain mencari probabilitas X dimiliki oleh kelas C.

P(H|X) merupakan probabilitas posterior, H dikondisikan pada X. Sebaliknya P(H) merupakan probabilitas prior, atau probabilitas sebelumnya. Kemudian P(X|H) merupakan probabilitas posterior dimana X dikondisikan pada H. Sedangkan P(X) merupakan probabilitas sebelumnya dari X klasifikasi, tetapi membuat batasan antara

setiap klasifikasi menjadi lebih kabur.

# Deskripsi Soal Masalah

Masalah dari soal ini yaitu membangun sebuah kondisi, dimana file yang diberikan yaitu file TrainsetTugas1ML.xlsx berupa himpunan data berisi 160 objek data yang memiliki 7 atribut input (**age**, **workclass**, **education**, **marital-status**, **occupation**, **relationship**, **hours-per-week**) 1 output (label kelas **income**) yang memiliki 4 kelas/label (0, 1, 2, dan 3). Kemudian, dengan menggunakan metode *Naïve Bayes*,, harus menentukan kelas/label data testing dalam file yang diberikan.

Sistem memasukkan file masukan file TrainsetTugas1ML.xlsx dan membaca data dalam file tersbut dan mengeluarkan *output* berupa file **final.xlsx** berupa satu kolom berisi **40 baris** yang menyatakan kelas/label baris yang bersesuaian.

# Metode Penyelesaian

Algoritma *Naïve Bayes**Classifier* (NBC) merupakan salah satu metoda [pembelajaran mesin](https://id.wikipedia.org/wiki/Pemelajaran_mesin) yang memanfaatkan perhitungan proba-bilitas dan statistik yang dikemukakan oleh ilmuan dari Inggris bernama Thomas Bayes, yaitu memprediksi probabilitas di masa depan berdasarkan pengalaman di masa sebelumnya.

Secara umum pada soal ini, mendefinisikan menghitung data yang ada dan membandingkan antara sebuah nilai yang diberikan termasuk nilai income >50 atau <=50 dari file TrainsetTugas1ML.xlsx . Setelah hasilnya didapat, kemudian dijumlahkan sebagai berikut:

*def perulangan(kolom1, kategori1, kolom2, kategori2):*

*i = 0*

*for j in range(trainrow):*

*if (datatrain[kolom1][j] == kategori1) & (datatrain[kolom2][j] == kategori2):*

*i += 1*

*return i*

*def rumus(kolom1, kategori1, kolom2, kategori2, tipeincome):*

*hasil = perulangan(kolom1, kategori1, kolom2, kategori2)/tipeincome*

*return hasil*

# Parameter & Fungsi Klasifikasi

Pada program *Naïve Bayes**Classifier* ini saya menggunakan 2 data inputan, yaitu data Tes dan data Train. Yang mana data test dan data train memiliki data index berupa (**age**, **workclass**, **education**, **marital-status**, **occupation**, **relationship**, **hours-per-week**) dan kelas. Data Tes yang nilai kelas belum ditentukan dan data Train data kelas sudah ditentukan.



Fungsi Klasifikasi digunakan untuk menentukan kelas income dari data Test yang dicari dengan algoritma perulangan naive bayes, lalu diurutkan berdasarkan id data yang jika >50k maka temasuk data peluangMore dan jika <=50k termasuk data peluangLess.

# Output Program

Output program disimpan dalam bentuk xlsx dengan nama file final.xlsx dan dengan indeks sebagai urutan satu kolom berisi 40 baris yang menyatakan kelas income baris yang bersesuaian.

# Sumber

# <https://id.wikipedia.org/wiki/Naive_Bayes_classifier>

1. Septian,Yuda. “DATA MINING MENGGUNAKAN ALGORITMA NAÏVE BAYES UNTUK KLASIFIKASI KELULUSAN MAHASISWA UNIVERSITAS DIAN NUSWANTORO”. Jurusan Sistem Informasi, Fakultas Ilmu Komputer.