Laporan Analisis Sistem Klasifikasi untuk Menentukan Label/Kelas Data Test Menggunakan Metode Naïve Bayes

Fauzan Firdaus – 1301164317 – IF 40-04

Diketahui sebuah data train sebanyak 160 data dan data test 70 data. Di dalamnya terdapat 7 atribut. Dengan data yang telah dimiliki pada data train, dicari label/kelas dari masing-masing data yang terdapat di data test. Bahasa yang digunakan untuk sistem ini adalah **Matlab**, dengan versi **R2018A**.

A. Naïve Bayes

Sedikit bahasan mengenai naïve bayes, naïve bayes adalah algoritma / metode yang ada pada machine learning yang digunakan untuk klasifikasi data dengan memanfaatkan perhitungan probabilitas dan statistik.

Dalam program ini, naïve bayes digunakan untuk memecahkan permasalahan klasifikasi pada data test berdasarkan acuan pada data train.

B. Strategi Penyelesaian Masalah

Pertama-tama, yang harus dilakukan adalah memuat / load data dari data train dan data test. Setelah itu, inisialisasi variabel yang akan dibutuhkan pada program, seperti variabel untuk index perulangan, variabel peluang, dan yang lainnya seperti yang sudah didefinisikan di program pada baris 4 sampai 55.

Dalam program ini, terdapat 2 perulangan inti. Yaitu perulangan pertama sebanyak 160 perulangan dengan tujuan untuk mencari dan menghitung data pada train set, dan perulangan kedua sebanyak 40 perulangan dengan tujuan untuk mencari label/kelas income pada test set.

Pada perulangan pertama/train, yang harus dilakukan adalah mencari tau jumlah dari income yang bernilai ‘>50K’ ada berapa, begitu pula dengan income ‘<=50K’.

Terdapat 3 kategori yang berbeda pada masing-masing 7 atribut tersebut, seperti pada atribut age, kategorinya adalah ‘young’, ‘adult’, dan ‘old’. Maka dari itu, yang harus dilakukan selanjutnya adalah menghitung masing-masing jumlah atribut berdasarkan incomenya. Dari jumlah tersebutlah akhirnya bisa didapatkan peluangnya.

Setelah semua atribut dicari jumlahnya (berdasarkan income) maka berhentilah perulangan tersebut. Yang dilakukan selanjutnya adalah menghitung peluang dari masing masing kategori income, yaitu jumlah per kategori income terhadap jumlah data yang ada.

Selanjutnya, perulangan kedua untuk test set lah yang akan dilakukan. Pada perulangan ini, intinya adalah untuk mencari nilai peluang berdasarkan kategori atribut yang ada dan menentukan label di setiap data test nya.

Peluang tersebut didapatkan dari perkalian semua peluang yang didapatkan dari masing-masing atribut dari atribut *age* sampai atribut *hours-per-week*. Untuk masing-masing peluangnya terbagi menjadi 2, yaitu berdasarkan income yang ‘<=50K’ dan ‘>50K’. Jadi income pada setiap data test, didapatkan berdasarkan perbandingan jumlah peluangnya. Jika peluang yang ‘<=50K’ pada sebuah data test lebih besar nilainya daripada peluang yang ‘>50K’, maka label/kelas pada index data test tersebut adalah ‘<=50K’, begitupun sebaliknya.

C. Hasil atau Output Program

Program akan menghasilkan 1 file excel berekstensi ‘.xls’ dengan nama filenya ‘TebakanTugas1ML’. Dalam file tersebut, menghasilkan 40 data yang berisi jawaban berupa tebakan income dari setiap data yang ada pada data test. Data disajikan dalam horizontal dengan indikatornya adalah *‘income-i’*, yang dimana i adalah

index dari setiap data test (1 sampai 40). Berikut adalah *screenshoot* dari hasil tebakannya:

income1	income2	income3	income4	income5
<=50K	<=50K	>50K	<=50K	>50K

income6	income7	income8	income9	income10
>50K	<=50K	<=50K	>50K	>50K

income11	income12	income13	income14	income15
>50K	>50K	<=50K	>50K	>50K

income16	income17	income18	income19	income20
>50K	<=50K	>50K	<=50K	>50K

income21	income22	income23	income24	income25
>50K	>50K	>50K	>50K	>50K

income26	income27	income28	income29	income30
>50K	>50K	>50K	>50K	<=50K

income31	income32	income33	income34	income35
<=50K	<=50K	>50K	>50K	<=50K

income36	income37	income38	income39	income40
>50K	<=50K	>50K	>50K	>50K