**Terapan Data Mining**

**Klasifikasi Metode Naïve Bayes**

Klasifikasi adalah salah satu teknik data mining untuk mengelompokan data berdasarkan kelas atau label yang telah ditentukan. Selain metode Decision Tree, metode naïve bayes juga sering digunakan untuk mengelompokan data. Naïve bayes adalah metode klasifikasi dengan konsep probabilitas atau peluang. Metode ini menghitung sekumpulan peluang dengan menjumlahkan frekuensi dan kombinasi nilai dari dataset yang diberikan.

Terdapat dua asumsi utama pada metode naïve bayes:

* Semua variabel memiliki prioritas yang sama pentingnya.
* Semua variabel bersifat bebas secara statistik (nilai suatu variabel tidak terkait dengan nilai variabel lain).

Asumsi ini kebanyakan tidak selalu benar, namun dalam praktiknya walau asumsi tidak terpenuhi, metode naïve bayes tetap memberikan hasil yang baik. Karena asumsi metode naïve beyes adalah semua variabel bersifat bebas (saling independen/dapat berdiri sendiri).

Rumus teorema naïve bayes:

Dimana:

X : Data dengan kelas yang belum diketahui.

H : Hipotesis data, merupakan suatu kelas yang spesifik

P (H | X) : Probabilitas hipotesis H berdasar kondisi X (Probabilitas posterior)

P (X | H) : Probabilitas X berdasarkna kondisi hipotesis H

P(H) : Probabilitas hipotesis H (Probabilitas prior)

P(X) : Probabilitas X

Contoh studi kasus naïve bayes:

klasifikasi variabel untuk menentukan apakah dapat berpergian atau tidak. Outlook, temperature, humidity, wind adalah variabel bebas. Travel adalah variabel terikat (label).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| No | Outlook | Temperature | Humidity | Wind | Travel |
| 1 | Sunny | Hot | High | False | No |
| 2 | Sunny | Hot | High | True | No |
| 3 | Overcast | Hot | High | False | Yes |
| 4 | Rainy | Mild | High | False | Yes |
| 5 | Rainy | Cool | Normal | False | Yes |
| 6 | Rainy | Cool | Normal | True | No |
| 7 | Overcast | Cool | Normal | True | Yes |
| 8 | Sunny | Mild | High | False | No |
| 9 | Sunny | Cool | Normal | False | Yes |
| 10 | Rainy | Mild | Normal | False | Yes |
| 11 | Sunny | Mild | Normal | True | Yes |
| 12 | Overcast | Mild | High | True | Yes |
| 13 | Overcast | Hot | Normal | False | Yes |
| 14 | Rainy | Mild | High | True | No |

Kemudian dari dataset tersebut akan digunakan untuk menjawab persoalan:

Menggunakan naïve bayes, tentukan label travel (yes or no) jika diketahui

* outlook = sunny AND temperature = cool AND humidity = high AND windy = true

Langkah pertama, Hitung semua kejadian dalam tabel frekuensi sebagai berikut.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Travel | |  | Travel | |  | Travel | |  | Travel | |  | Travel |
| Outlook | Yes | No | Temperature | Yes | No | Humidity | Yes | No | Windy | Yes | No |  |
| Sunny | 2 | 3 | hot | 2 | 2 | high | 3 | 4 | false | 6 | 2 | Yes | 9 |
| overcast | 4 | 0 | mild | 4 | 2 | normal | 6 | 1 | true | 3 | 3 | No | 5 |
| Rainy | 3 | 2 | cool | 3 | 1 |
| TOTAL | 9 | 5 | TOTAL | 9 | 5 | TOTAL | 9 | 5 | TOTAL | 9 | 5 | TOTAL | 14 |

Langkah kedua, ubah setiap frekuensi menjadi bentuk peluang.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Travel | |  | Travel | |  | Travel | |  | Travel | |  | Travel |
| Outlook | Yes | No | Temperature | Yes | No | Humidity | Yes | No | Windy | Yes | No |  |
| Sunny | 0.22 | 0.60 | hot | 0.22 | 0.40 | high | 0.33 | 0.80 | false | 0.67 | 0.40 | Yes | 0.64 |
| overcast | 0.44 | 0.00 | mild | 0.44 | 0.40 | normal | 0.67 | 0.20 | true | 0.33 | 0.60 | No | 0.36 |
| Rainy | 0.33 | 0.40 | cool | 0.33 | 0.20 |

Langkah ketiga, sesuai dengan persoalan.

Hitung outlook = sunny AND temperature = cool AND humidity = high AND windy = true ketika travel = Yes (Warna kuning)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Travel | |  | Travel | |  | Travel | |  | Travel | |  | Travel |
| Outlook | Yes | No | Temperature | Yes | No | Humidity | Yes | No | Windy | Yes | No |  |
| Sunny | 0.22 | 0.60 | hot | 0.22 | 0.40 | high | 0.33 | 0.80 | false | 0.67 | 0.40 | Yes | 0.64 |
| overcast | 0.44 | 0.00 | mild | 0.44 | 0.40 | normal | 0.67 | 0.20 | true | 0.33 | 0.60 | No | 0.36 |
| Rainy | 0.33 | 0.40 | cool | 0.33 | 0.20 |

Sehingga travel = yes adalah 0.22 x 0.33 x 0.33 x 0.33 x 0.64 = 0.0053

Hitung outlook = sunny AND temperature = cool AND humidity = high AND windy = true ketika travel = No (warna merah)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Travel | |  | Travel | |  | Travel | |  | Travel | |  | Travel |
| Outlook | Yes | No | Temperature | Yes | No | Humidity | Yes | No | Windy | Yes | No |  |
| Sunny | 0.22 | 0.60 | hot | 0.22 | 0.40 | high | 0.33 | 0.80 | False | 0.67 | 0.40 | Yes | 0.64 |
| overcast | 0.44 | 0.00 | mild | 0.44 | 0.40 | normal | 0.67 | 0.20 | True | 0.33 | 0.60 | No | 0.36 |
| Rainy | 0.33 | 0.40 | cool | 0.33 | 0.20 |

Sehingga travel = yes adalah 0.60 x 0.22 x 0.80 x 0.60 x 0.36 = 0.0206

Langkah 4.

Masukan nilai yang ada dilangkah 3 kedalam rumus naïve bayes

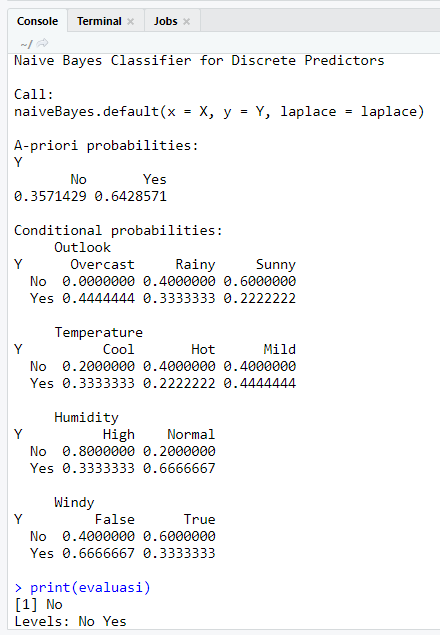
Peluang untuk Travel = yes adalah

Peluang untuk Travel = No adalah

Karena peluang untuk travel = no lebih besar dari peluang travel = yes, maka untuk persoalan outlook = sunny AND temperature = cool AND humidity = high AND windy = true dikelompokan ke dalam travel = no.

Implementasi Bahasa R

|  |
| --- |
| # library for naive bayes classification  library**(**e1071**)**  # load dataset, assumption dataset was done training  data\_latih **=** read.csv**(**"dataset\_travel.csv"**)**  # see the characteristics of training data  summary**(**data\_latih**)**  str**(**data\_latih**)** print**(**data\_latih**)**  # create test data. For evaluation naive bayes model  data\_uji **<-** data.frame**(**  Outlook **=** as.factor**(**"Sunny"**)**,  Temperature **=** as.factor**(**"Cool"**)**,  Humidity **=** as.factor**(**"High"**)**,  Windy **=** as.factor**(**"True"**)**  **)**  # see the characteristics of test data  summary**(**data\_uji**)**  str**(**data\_uji**)**  print**(**data\_uji**)**  # create naive bayes model based on training data  model **<-** naiveBayes**(**data\_latih**$**Travel **~** ., data**=**data\_latih**)**  print**(**model**)**  # evaluation naive bayes model use test data  evaluasi **<-** predict**(**model, data\_uji**)**  print**(**evaluasi**)** |



Gambar 1. Hasil Naïve Bayes