**BAB 2**

**Konsep Learning**

# POKOK BAHASAN :

* Konsep Learning
* Algoritma Find-S

# TUJUAN PEMBELAJARAN :

Setelah mempelajari materi dalam bab ini, mahasiswa diharapkan mampu:

* Mengenal dan memahami konsep pembelajaran pada mesin
* Memahami cara mengambil sebuah mesin mengambil kesimpulan
* Membuat rancangan sebuah mesin yang dapat belajar dari data contoh yang diberikan dana dapat menganalisa sebuah permasalahan

**2.1. Konsep Learning**

Learning mempunyai arti belajar, yaitu menambah pengetahuan atau memahami dan menganalisa sesuatu dengan cara mengikuti aturan-aturan yang ada atau pengalaman yang dimiliki. Pada Konsep Learning disini yang dimaksud dengan learning adalah sebuah studi yang mempelajari cara untuk memprogram sebuah komputer untuk dapat berbelajar.

Manusia selama hidupnya selalu melakukan learning. Hal ini terjadi tanpa disadari dan alamiah. Manusia belajar melalui pengalaman yang dia alami sehari-hari. Dari pengalaman tersebut, manusia akan mendapatkan knowledge.Untuk mendapatkan knowledge dapat melalui berbagai cara. Cara yang paling sederhana adalah rote learning atau menyimpan informasi yang sudah dikalkulasi. Cara lainnya adalah dengan mendapatkan pengetahuan dari orang lain yang sudah ahli. Manusia juga dapat belajar melalui pengalaman pemecahan masalah yang ia lakukan. Setelah berhasil mengatasi sebuah masalah, manusia akan mengingat struktur dan cara mengatasi masalah tersebut. Apabila manusia mengalami sebuah masalah yang hampir serupa, maka manusia dapat mengatasi masalah tersebut secara lebih efisien.

Permasalahan akan muncul manakala manusia membutuhkan sebuah alat bantu yang dapat belajar dari pengalaman yang diberikan. Maka dikembangkan sebuah progam yang belajar. Untuk membuat program/mesin yang dapat belajar cara yang paling mudah adalah dengan meniru cara menusia belajar. Manusia belajar dengan cara menambah pengetahuan, membuat kesimpulan dan ketika ada permasalahan baru akan melalukukan analisa berdasarkan kesimpulan yang sudah diambil.

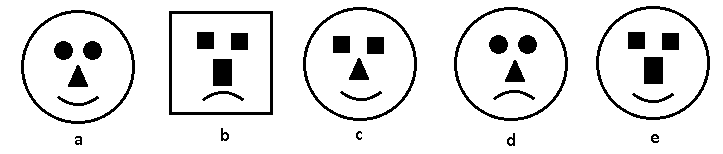
Konsep Pembelajaran

Jika meniru cara belajar yang dilakaukan oleh manusia maka proses pembelajaran pada mesin dapat dibagi menjadi beberapa tahap yaitu:

1. Proses pebelajaran dari contoh yang diberikan .
2. Proses pengurutan hipotesa

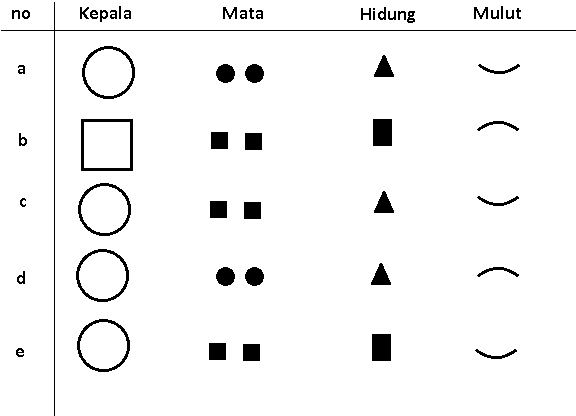
Berikut adalah kasus proses pembelajaran dari contoh.

Perhatikan gambar berikut:



**Gambar 2.1** Contoh Data Pembelajaran

Dari gambar tersebut diatas maka dapat dilakukan pembelajaran dengan target apakah gambar tersebut tersenyum atau tidak. Untuk merepersentasikan pengetahuan tersebut makan diamati berdasarkan bentuk kepala, bentuk mata, bentuk hidung dan bentuk mulut. Repesentasi dari pengetahuan dari hasil belajar tersebut dapat dibuat dalam bentuk table berikut:



**Gambar 2.2** Analisa Dari Gambar Data Pembelajaran

Jika gambar a, c dan e mengambarkan raut muka tersenyum maka dari pembelajaran table diatas dapat direpresentasikan dalam bentuk table sebagai berikut:

**Tabel 2.1.** Tabel Hasil Analisa Data Keseluruhan

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| no | Kepala | Mata | Hidung | Mulut | Senyum |
| a | lingkaran | bulat | Segi tiga | melengkung keatas | ya |
| b | bujursangkar | kotak | Segi empat | melengkung kebawah | tidak |
| c | lingkaran | kotak | Segi tiga | melengkung keatas | ya |
| d | lingkaran | bulat | Segi tiga | melengkung kebawah | tidak |
| e | lingkaran | kotak | Segi empat | melengkung keatas | ya |

Dari tabel diatas maka didapat data bahwa untuk senyum yang bernilai ‘ya’ adalah sebagai berikut:

**Tabel 2.2.** Tabel Hasil Analisa Data Senyum =”ya”

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| no | Kepala | Mata | Hidung | Mulut | senyum |
| a | lingkaran | bulat | segitiga | melengkung keatas | ya |
| c | lingkaran | kotak | segitiga | melengkung keatas | ya |
| e | lingkaran | kotak | segiempat | melengkung keatas | ya |

Maka hipotesa yang dapat dari tabel diatas Senyum adalah ya jika bentuk kepala lingkaran, bentuk mata terserah, bentuk Hidung terserah dan bentuk mulut melengkug keatas. Atau dapat ditulis

(Kepala=lingkaran,mata=\*, hidung=\*,Mulut=melengkung keatas)

atau aturan untuk senyum adalah

(lingkaran,\*, \*,melengkung keatas)

Sedangkan untuk aturan Senyum = tidak disimpulkan dengan cara sebagai berikut:

**Tabel 2.3**. Hasil Analisa Data Senyum =”Tidak”

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| no | Kepala | Mata | Hidung | Mulut | senyum |
| b | bujursankar | kotak | segiempat | melengkung kebawah | tidak |
| d | lingkaran | bulat | segitiga | melengkung kebawah | tidak |

Dari tabel diatas bisa dibaca bahwa untuk senyum yang bernilai tidak(tidak tersenyum) dapat di simpulkan dengan hipotersa adalah bentuk kepala terserah, bentuk mata terserah, bentuk hidung terserah, bentuk mulut melengkung ke bawah. Atau dapat ditulis menjadi

(Kepala=\*,mata=\*, hidung=\*,Mulut=melengkung kebawah)

atau aturan untuk tidak senyum adalah

(\*,\*, \*,melengkung kebawah)

Contoh diatas dapat juga di terapkan dipada data yang berbentuk katagorikal, data katagorikal adalah data yang sudah dalam bentuk pengelompokan, misalnya tua, muda besar kecil, bukan dalam bentuk angka.

Dengan cara membuat tabel kemudian dilanjutkan dengan pengambilan hipotesa maka proses tersebut dapat diterap pada perhitungan menggunakan computer. Atau dengan kata lain proses pembelajaran tersebut dapat dikalukan oleh mesin.

Tetapi proses pembelajan tersebut membutuhkan data training yang konsisten dan tidak bias. Yang dimaksud dengan data konsisten adalah atribut dari data tersebut mempunyai nilai yang sama dalam satu keputusan. sedangkan data bias adalah jika ada dua atau lebih data mempunyai data sama persis (identik) tetapi keputusannya berbeda.

**Tabel 2.4.** Data Tidak Konsisten

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| no | Kepala | Mata | Hidung | Mulut | senyum |
| b | bujursankar | kotak | segiempat | melengkung kebawah | tidak |
| d | lingkaran | bulat | segitiga | melengkung kebawah | tidak |

Contoh diatas menunjukkan bahwa atribut Mulut memiliki data yang konsisten terhadap keputusan tidak senyum.

**Tabel 2.5.** Data Bias

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| no | Kepala | Mata | Hidung | Mulut | senyum |
| - | bujursankar | kotak | segiempat | melengkung kebawah | ya |
| - | bujursankar | kotak | segiempat | melengkung kebawah | tidak |

Kedua atribut dari data diatas masing-masing nilai atributnya sama tetapi keputusannya berbeda data yang demikian disebut dengan data yang bias.

Pada penerapan di dunia nyata data yang tidak bias dan konsisten itu sulit ditemukan. sehingga perlu pengembangan cara pembelajaran yang lain.

**2.2 Algoritma Find-S**

Find-S adalah suatu metode paling sederhana yang dapat digunakan untuk mendapatkan suatu hipotesa berdasarkan data. Sehingga Find-S ini digukanan untuk member inspirasi bahwa manusia dapat membuat sebuah mesin bisa menarik sebuah kesimpulan atau hipotesa. Sehingga Find-S ini sesuai jika digunakan untuk penengalan konsep pembelajaran.

Untuk membuat sebuah hipotesa Find-S mencari kesamaan nilai attribut dari data contoh yang diberikan. Sehingga hipotesa sangat terpengaruh oleh data contoh yang diberikan, hal ini menjadikan Find-S mempunyai kelemahan yaitu data yang digunakan harus bersifat konsisten dan tidak bias. Padahal terlalu sulit untuk dapat memperoleh data semacam ini pada persoalan nyata.

Berikut adalah contoh dari Find-S.

**Tabel 2.6.** Data Cuaca

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Day | Cuaca | Temperatur | Kecepatan Angin | Berolah-raga |
| D1 | Cerah | Normal | Pelan | Ya |
| D2 | Cerah | Normal | Pelan | Ya |
| D3 | Hujan | Tinggi | Pelan | Tidak |
| D4 | Cerah | Normal | Kencang | Ya |
| D5 | Hujan | Tinggi | Kencang | Tidak |
| D6 | Cerah | Normal | Pelan | Ya |

Untuk memperoleh hipotesa dengan Find-S dari data training di atas, langkah pertama yang harus dilakukan adalah memecah data berdasarkan keputusannya, sehingga akan diperoleh 2 data: pertama untuk keputusan=ya dan kedua untuk keputusan=tidak , seperti dibawah ini

**Tabel 2.7.** Data Cuaca Berolahraga= Ya

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Day | Cuaca | Temperatur | Kecepatan Angin | Berolah-raga |
| D1 | Cerah | Normal | Pelan | Ya |
| D2 | Cerah | Normal | Pelan | Ya |
| D4 | Cerah | Normal | Kencang | Ya |
| D6 | Cerah | Normal | Pelan | Ya |

**Tabel 2.8.** Data Cuaca Berolahraga= Tidak

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Day | Cuaca | Temperatur | Kecepatan Angin | Berolah-raga |
| D3 | Hujan | Tinggi | Pelan | Tidak |
| D5 | Hujan | Tinggi | Kencang | Tidak |

Dari hasil pemisahan tersebut, terlihat bahwa data training tersebut Konsisten dan tidak bias. Langkah berikutnya adalah membuat hipotesa untuk masing-masing keputusan. Pembuatan hipotesa ini dilakukan dengan mengambil data pertama sebagai hipotesa awal yang dianggap sebagai hipotesa spesifik, dan diteruskan hingga data terakhir dengan memperhatikan kesamaan sampai didapatkan hipotesa umum.

Langkah ini digunakan untuk menentukan hipotesa dari keputusan=ya seperti pada table 2.7

1. Hipotesa awal disamakan dengan data pertama:

H(Cerah, Normal, Pelan) = Ya

1. Data kedua, tidak ada perubahan karena semua nilai instancenya sama:

H(Cerah, Normal, Pelan) = Ya

1. Data ketiga, ada perubahan di attribut kec angin, sehingga :

H(Cerah, Normal, \*) = Ya

1. Data keenam, ada perubahan di attribut kec angin, sehingga :

H(Cerah, Normal, \*) = Ya

Sehingga Hipotesa ya adalah

**H(Cerah, Normal, \*) = Ya**

Langkah ini digunakan untuk menentukan hipotesa dari keputusan=tidak dengan memperhatikan table 2.8

1. Hipotesa awal disamakan dengan data pertama (D3):

H(Hujan, Tinggi, Pelan) = Tidak

1. Data kelima, ada perubahan di kec.angin

H(Hujan, Tinggi, \*) = Tidak

Sehingga Hipotesa tidak adalah

**H(Hujan, Tinggi, \*) = Tidak**

jika ada kondisi Bila cuaca cerah, apakah akan berolahraga? jawabnya dibandingkan dengan kedua hipotersa tersebut yaitu

**(Cerah,\*,\*)? dibandingkan dengan**

**H(Cerah, Normal, \*) = Ya**

**H(Hujan, Tinggi, \*) = Tidak**

Jawabnya adalah YA

**3.5. Latihan Soal**

1. Buatlah hipotesa dari mahasiswa yang dapat biasiswa dan yang tidak dapat biasiswa dari table dibawah

**Tabel 2.9.** Data Biasiswa

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Pintar | Rajin | Sopan | pendapatan ortu | Biasiswa |
| iya | iya | tidak | rendah | dapat |
| iya | tidak | iya | rendah | dapat |
| tidak | tidak | tidak | rendah | tidak |
| tidak | tidak | tidak | tinggi | tidak |
| iya | iya | iya | rendah | dapat |
| tidak | iya | tidak | rendah | tidak |
| iya | tidak | tidak | rendah | dapat |
| tidak | iya | tidak | tinggi | tidak |

1. Buatlah hipotesa dari jenis ikan koi dan ikan koi dari tabel dibawah

**Tabel 2.10.** Data Jenis Ikan

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| gerakan | panjang | lebar | warna | jenis ikan |
| cepat | kecil | kecil | cerah | koi |
| lambat | kecil | kecil | cerah | koi |
| lambat | kecil | lebar | cerah | koi |
| cepat | panjang | kecil | cerah | lele |
| lambat | panjang | kecil | cerah | lele |
| cepat | panjang | kecil | buram | lele |
| lambat | panjang | kecil | buram | lele |
| cepat | kecil | kecil | cerah | koi |
| lambat | kecil | kecil | cerah | koi |
| lambat | panjang | kecil | cerah | lele |
| cepat | panjang | kecil | buram | lele |
| cepat | kecil | lebar | cerah | koi |
| lambat | kecil | lebar | cerah | koi |
| cepat | kecil | lebar | cerah | koi |
| lambat | panjang | kecil | buram | lele |
| cepat | panjang | kecil | cerah | lele |

**BAB 3**

**Metode Naïve Bayes**

# POKOK BAHASAN :

* Probabilitas Bersyarat
* Teorima Bayes
* HMAP Metode Naive Bayes

# TUJUAN PEMBELAJARAN :

Setelah mempelajari materi dalam bab ini, mahasiswa diharapkan mampu:

* Mengenal dan memahami Metode Bayes
* Memahami cara kerja dari teori bayes
* Membuat rancangan sebuah program yang dapat menyelesaikan permasalahan dengan menggunakan metode Naive Bayes

Metode Find-S tidak dapat digunakan untuk data yang tidak konsisten dan data yang bias, sehingga untuk bentuk data semacam ini salah satu metode sederhana yang dapat digunakan adalah metode bayes. Metode Bayes ini merupakan metode yang baik di dalam mesin pembelajaran berdasarkan data training, dengan menggunakan probabilitas bersyarat sebagai dasarnya.

**3.1 Probabilitas Bersyarat**

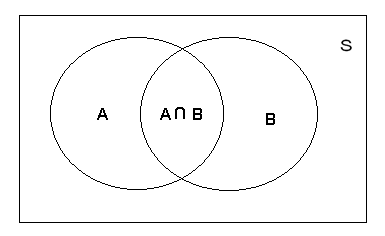
Ditentukan set B dan set A. Probabilitas terjadinya A sama dengan syarat bahwa B sudah terjadi atau akan terjadi. Ditulis :



dimana P(B) > 0. Dengan kata lain kejadian B merupakan syarat terjadinya kejadian A.

Jika yang menjadi syarat adalah kejadian A maka dapat ditulis sebagai berikut :





**Gambar 3.1** Kemungkinan A dan B adalah irisan A dan B

Misalkan dari 100 orang mahasiswa menunjukkan 20 orang mahasiswa menyukai keduanya, 30 orang mahasiswa menyukai bulu tangkis tapi tidak menyukai bola volley, 40 orang mahasiswa menyukai bola volley tapi tidak menyukai bulu tangkis, dan 10 orang mahasiswa tidak menyukai kuduanya.

Dari data ini dapat disusun bentuk distribusi bersama sebagai berikut:

**Tabel 3.1.** Tabel probabilitas

|  |  |  |  |
| --- | --- | --- | --- |
| Suka | Suka bola volley (B) | | P(A) |
| bulu tangkis (A) | Ya | Tidak |
| Ya | 0.2 | 0.3 | **0.5** |
| Tidak | 0.4 | 0.1 | **0.5** |
| P(B) | **0.6** | **0.4** | 1 |

Kejadian(A,B) = {(ya,ya), (ya,tidak), (tidak,ya), (tidak,tidak)}

p(A=ya,B=ya) = 0.2

p(A=ya,B=tidak) = 0.3

p(A=tidak,B=ya) = 0.4

p(A=tidak,B=tidak) = 0.1

Probabilitas seseorang menyukai bulutangkis bila diketahui dia menyukai bola volley adalah:



Probabilitas seseorang tidak menyukai bolavolley bila diketahui dia menyukai bola bulutangkis adalah:



Dari tabel tersebut dapat di simpulkan bahwa 0.2 suka keduanya, 0.3 suka bulu tangkis saja, 0.4 suka bola voley saja dan 0.1 tidak suka keduaya. Data tersebut dinamakan *Distribusi Bersama*. Sedangkan nilai 0.5 suka bulu tangkis dan 0.5 tidak suka bulu tangkis serta 0.6 suka bola voley dan 0.4 tidak suka bola voley adalah *Distribusi Marginal.*

**3.2  Teorima Bayes**

Teori Bayes atau lebih dikenal dengan kaidah Bayes, memainkan peranan yang sangat penting dalam penerapan probabilitas bersyarat. Teori ini pertama kali dikembangkan oleh Thomas bayes (1702-1763). Kaidah Bayes merupakan kaidah yang memperbaiki atau merevisi suatu probabilitas dengan cara memanfaatkan informasi tambahan. Maksudnya, dari probabilitas awal (prior probability) yang belum diperbaiki dengan rumuskan berdasarkan informasi yang tersedia saat ini, kemudian dibentuklah probabilitas berikutnya (posterior probability).

Teori lain mengatakan Teori Bayes adalah kesimpulan statistik yang membuktikan atau pengamatan yang digunakan untuk memperbarui atau menarik kesimpulan yang baru suatu probabilitas yang mungkin benar. Teori Bayes berasal dari kebiasaan menggunakan rumus bayes untuk memproses suatu kesimpulan/dugaan.

Keadaan Posteriror (Probabilitas Xk di dalam Y) dapat dihitung dari keadaan prior (Probabilitas Y di dalam Xk dibagi dengan jumlah dari semua probabilitas Y di dalam semua Xi)

**Gambar 3.2** Kemungkinan Xk didalam Y



**3.3 HMAP (*Hypothesis Maximum Appropri Probability*)**

HMAP (Hypothesis Maximum Appropri Probability) menyatakan hipotesa yang diambil berdasarkan nilai probabilitas berdasarkan kondisi prior yang diketahui. HMAP adalah model penyederhanaan dari metode bayes yang disebut dengan Naive Bayes. HMAP inilah yang digunakan di dalam macine learning sebagai metode untuk mendapatkan hipotesis untuk suatu keputusan.

* + 1. *Contoh permasalahan 1 HMAP atau naïve Bayes*

Diketahui hasil survey yang dilakukan sebuah lembaga kesehatan menyatakan bahwa 30% penduduk di dunia menderita sakit paru-paru. Dari 90% penduduk yang sakit paru-paru ini 60% adalah perokok, dan dari penduduk yang tidak menderita sakit paru-paru 20% perokok.

*Fakta ini bisa didefinisikan dengan: X=sakit paru-paru dan Y=perokok.*

*Maka : P(X) = 0.9*

*P(~X) = 0.1*

*P(Y|X) = 0.6 🡪 P(~Y|X) = 0.4*

*P(Y|~X) = 0.2 🡪 P(~Y|~X) = 0.8*

Dengan metode bayes dapat dihitung:

*P({Y}|X) = P(Y|X).P(X) = (0.6) . (0.9) = 0.54*

*P({Y}|~X) = P(Y|~X) P(~X) = (0.2).(0.1) = 0.02*

Bila diketahui seseorang merokok, maka dia menderita sakit paru-paru karana P({Y}|X) lebih besar dari P({Y}|~X). HMAP diartikan mencari probabilitas terbesar dari semua instance pada attribut target atau semua kemungkinan keputusan. Pada persoalan keputusan adalah sakit paru-paru atau tidak.

* + 1. *Contoh permasalahan 2 HMAP atau naïve Bayes Dari Data Training*

Bila diketahui kebiasaan olah raga seseorang dan sudah ditabelkan dalam bentuk sebagai berikut:

**Tabel 3.2.** Tabel olahraga

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| # | Cuaca | Temperatur | Kecepatan Angin | Berolah-raga |
| 1 | Cerah | Normal | Pelan | Ya |
| 2 | Cerah | Normal | Pelan | Ya |
| 3 | Hujan | Tinggi | Pelan | Tidak |
| 4 | Cerah | Normal | Kencang | Ya |
| 5 | Hujan | Tinggi | Kencang | Tidak |
| 6 | Cerah | Normal | Pelan | Ya |

*Asumsi:*

*Y = berolahraga,*

*X1 = cuaca,*

*X2 = temperatur,*

*X3 = kecepatan angin*

Fakta menunjukkan:

*P(Y=ya) = 4/6 🡪 P(Y=tidak) = 2/6*

Apakah bila cuaca cerah dan kecepatan angin kencang, orang akan berolahraga?

Fakta: P(X1=cerah|Y=ya) = 1, P(X1=cerah|Y=tidak) = 0

P(X3=kencang|Y=ya) = 1/4 , P(X3=kencang|Y=tidak) = 1/2

HMAP dari keadaan ini dapat dihitung dengan:

P( X1=cerah,X3=kencang | Y=ya )

= { P(X1=cerah|Y=ya).P(X3=kencang|Y=ya) } . P(Y=ya)

= { (1) . (1/4) } . (4/6) = 1/6

P( X1=cerah,X3=kencang | Y=tidak )

= { P(X1=cerah|Y=tidak).P(X3=kencang|Y=tidak) } . P(Y=tidak)

= { (0) . (1/2) } . (2/6) = 0

Keputusannya adalah berolahraga = YA

* + 1. *Contoh 2 HMAP atau naïve Bayes Dari Data Training*

Jika diketahui dari data kelulusan mahasiswa disebuh perguruan tinggi adalah sebagai berikut:

**Tabel 3.3.** Tabel kelulusan mahasiswa

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **NO** | **JENIS KELAMIN** | **STATUS MAHASISWA** | **STATUS PRENIKAHAN** | **IPK Semester 1-6** | **STATUS KELULUSAN** |
| 1 | LAKI - LAKI | MAHASISWA | BELUM | 3.17 | TEPAT |
| 2 | LAKI - LAKI | BEKERJA | BELUM | 3.30 | TEPAT |
| 3 | PEREMPUAN | MAHASISWA | BELUM | 3.01 | TEPAT |
| 4 | PEREMPUAN | MAHASISWA | MENIKAH | 3.25 | TEPAT |
| 5 | LAKI - LAKI | BEKERJA | MENIKAH | 3.20 | TEPAT |
| 6 | LAKI - LAKI | BEKERJA | MENIKAH | 2.50 | TERLAMBAT |
| 7 | PEREMPUAN | BEKERJA | MENIKAH | 3.00 | TERLAMBAT |
| 8 | PEREMPUAN | BEKERJA | BELUM | 2.70 | TERLAMBAT |
| 9 | LAKI - LAKI | BEKERJA | BELUM | 2.40 | TERLAMBAT |
| 10 | PEREMPUAN | MAHASISWA | MENIKAH | 2.50 | TERLAMBAT |
| 11 | PEREMPUAN | MAHASISWA | BELUM | 2.50 | TERLAMBAT |
| 12 | PEREMPUAN | MAHASISWA | BELUM | 3.50 | TEPAT |
| 13 | LAKI - LAKI | BEKERJA | MENIKAH | 3.30 | TEPAT |
| 14 | LAKI - LAKI | MAHASISWA | MENIKAH | 3.25 | TEPAT |
| 15 | LAKI - LAKI | MAHASISWA | BELUM | 2.30 | TERLAMBAT |

Jika seorang mahasiswa dengan data sebagai berikut, buatlah prediksi kelulusannya!

**Tabel 3.4.** Tabel Testing

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| KELAMIN | STATUS | PRENIKAHAN | IPK | KETERANGAN |
| LAKI - LAKI | MAHASISWA | BELUM | 2.70 | ??? |

Jawab:

Tahap 1 menghitung jumlah class/label

* P(Y= TEPAT)  = 8/15   ‘ jumlah data “TEPAT” pada komom ‘STATUS KELULUSAN’ dibagi jumlah data
* P(Y= TERLAMBAT) = 7/15 ‘ jumlah data “TERLAMBAT” pada komom ‘STATUS KELULUSAN’ dibagi jumlah data

Tahap 2 menghitung jumlah kasus yang sama dengan class yang sama

* P(JENIS KELAMIN = LAKI - LAKI | Y= TEPAT) = 5/8 ‘jumlah data jenis kelamin “laki-laki” dengan keterangan “TEPAT” dibagi jumlah data TEPAT
* P(JENIS KELAMIN = LAKI - LAKI | Y= TERLAMBAT) = 3/7 ‘jumlah data jenis kelamin “laki-laki” dengan keterangan “TERLAMBAT” dibagi jumlah data TERLAMBAT
* P(STATUS MAHASISWA = MAHASISWA | Y= TEPAT) = 5/8 ‘jumlah data dengan status mahasiswa  dengan keterangan “TEPAT” dibagi jumlah data TEPAT
* P(STATUS MAHASISWA = MAHASISWA | Y= TERLAMBAT) = 3/7 ‘jumlah data dengan status mahasiswa  dengan keterangan “TERLAMBAT” dibagi jumlah data TERLAMBAT
* P(STATUS PRENIKAHAN = BELUM| Y= TEPAT) = 4/8 ‘jumlah data dengan status pernikahan “Belum menikah” dan  keterangan “TEPAT” dibagi jumlah data TEPAT
* P(STATUS PRENIKAHAN = BELUM| Y= TERLAMBAT) = 4/7 ‘jumlah data dengan status pernikahan “Belum menikah” dan keterangan “TERLAMBAT”dibagi jumlah data TERLAMBAT
* P(IPK = 2.70| Y= TEPAT) = 0/8 ‘jumlah data IPK “2.70” dengan keterangan “TEPAT” dibagi jumlah data TEPAT
* P(IPK = 2.70| Y= TERLAMBAT) = 1/7 ‘jumlah data IPK “2.70” dengan keterangan “TERLAMBAT” dibagi jumlah data TERLAMBAT

Tahap 3 kalikan semua hasil variable TEPAT & TERLAMBAT

P (KELAMIN=LAKI – LAKI), (STATUS MHS=MAHASISWA), (PRENIKAHAN = BELUM), (IPK= 2.70 ) |TEPAT)

= {P(P(KELAMIN =LAKI-LAKI|Y=TEPAT). P(STATUS MHS = MAHASISWA | Y= TEPAT) . P(PRENIKAHAN =BELUM|Y=TEPAT). P(IPK = 2.70| Y= TEPAT)

=    5/8    .      5/8 .   4/8   .   0/8   . 8/15

= 0

P (KELAMIN=LAKI – LAKI), (STATUS MHS=MAHASISWA), (PRENIKAHAN = BELUM), (IPK= 2.70 ) |TERLAMBAT)

= {P(P(KELAMIN =LAKI-LAKI|Y= TERLAMBAT). P(STATUS MHS = MAHASISWA | Y= TERLAMBAT) . P(PRENIKAHAN = BELUM|Y= TERLAMBAT). P(IPK = 2.70| Y= TERLAMBAT)

= 3/7  .  3/7    .  4/7 .  1/7 . 7/15

= 0,0069

Tahap 4 Bandingkan hasil class TEPAT & TERLAMBAT

* Karena hasil (P|TERLAMBAT) lebih besar dari (P|TERLAMBAT) maka keputusanya adalah “TERLAMBAT”

**Tabel 3.5.** Tabel Testing

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| KELAMIN | STATUS | PRENIKAHAN | IPK | KETERANGAN |
| LAKI - LAKI | MAHASISWA | BELUM | 2.70 | TERLAMBAT |

**3.4. Rangkuman**

Teori Bayesian  mempunyai beberapa kelebihan, yaitu:

1.  Mudah untuk dipahami.

2.  Hanya memerlukan pengkodean yang sederhana.

3.   Lebih cepat dalam penghitungan.

Kekurangan dari Teori probabilitas Bayes:

1. Metode Bayes hanya bisa digunakan untuk persoalan klasifikasi dengan supervised learning dan data-data kategorikal.
2. Metode Bayes memerlukan pengetahuan awal untuk dapat mengambil suatu keputusan. Tingkat keberhasilan metode ini sangat tergantung pada pengetahuan awal yang diberikan.

**3.5. Latihan Soal**

1. Diketahui hasil survey yang dilakukan sebuah lembaga kesehatan menyatakan bahwa 30% penduduk di dunia menderita sakit paru-paru. Dari 90% penduduk yang sakit paru-paru ini 60% adalah perokok, dan dari penduduk yang tidak menderita sakit paru-paru 20% perokok. Dengan menggunakan HMAP (Hypothesis Maximum Appropri Probability), hitunglah kemungkinan seseorang perokok mengidap penyakit paru-paru, dan kemungkinan orang tersebut tidak sakit paru-paru!
2. Misalkan populasi penduduk di suatu desa dibagi menurut jenis kelamin dan status pekerjaan sebagai berikut:

**Tabel 3.6.** Tabel data penduduk

|  |  |  |  |
| --- | --- | --- | --- |
| Jenis kelamin | Bekerja | Menganggur | Jumlah |
| Laki-laki | 460 | 40 | 500 |
| Wanita | 140 | 260 | 400 |
| Jumlah | 600 | 300 | 900 |

Bila diambil seorang dari mereka secara acak maka berapakah peluang yang terpilih adalah status yang bekerja dan berjenis kelamin laki-laki. Serta berapa peulang yang terpilih adalah status yang bekerja dan berjenis kelamin perempuan!

1. Dibawah ini adalah data Ikan Lele dan Ikan Gurami .

**Tabel 3.6.** Tabel data jenis ikan

|  |  |  |  |
| --- | --- | --- | --- |
| **Panjang** | **Lebar** | **Berat** | **Jenis Ikan** |
| Panjang | kecil | ringan | Ikan Lele |
| Panjang | sedang | Berat | Ikan Lele |
| Panjang | Lebar | Berat | Gurami |
| Sedang | kecil | Berat | Ikan Lele |
| Sedang | sedang | ringan | Ikan Lele |
| Sedang | Lebar | Berat | Gurami |
| Kecil | kecil | ringan | Gurami |
| Kecil | sedang | ringan | Gurami |
| Kecil | Lebar | Berat | Gurami |

Jika diketahui ikan “Panjang Lebar Berat” maka termasuk ikan apakah? hitunglah menggunakan Naïve Bayes!

1. Diketahui data orang yang membeli Laptop sebagai berikut

**Tabel 3.7.** Tabel data penjualan laptop

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Umur | Gaji | Status | Hutang | Beli Laptop |
| <=30 | tinggi | single | punya | tidak |
| <=30 | tinggi | single | tidak | tidak |
| 30…40 | tinggi | single | punya | iya |
| >40 | sedang | single | punya | iya |
| >40 | rendah | menikah | punya | iya |
| >40 | rendah | menikah | tidak | tidak |
| 31…40 | rendah | menikah | tidak | iya |
| <=30 | sedang | single | punya | tidak |
| <=30 | rendah | menikah | punya | iya |
| >40 | sedang | menikah | punya | iya |
| <=30 | sedang | menikah | tidak | iya |
| 31…40 | sedang | single | tidak | iya |
| 31…40 | tinggi | menikah | punya | iya |
| >40 | sedang | single | tidak | tidak |

Jika ada orang dengan data X =(Umur<=30, Gaji=Sedang, Status=menikah, Hutang= punya) hitunglah kemungkinan dia beli laptop dan kemungkinan dia tidak beli laptop!

**BAB 4**

**K-Nearest Neighbor (KNN)**

# POKOK BAHASAN :

* Prinsip K-NN
* Contoh K-NN

# TUJUAN PEMBELAJARAN :

Setelah mempelajari materi dalam bab ini, mahasiswa diharapkan mampu:

* Mengenal dan memahami algoritma K-NN
* Dapat menyelesaikan pemasalahan dengan menggunakan K-NN
* Membuat rancangan sebuah program yang dapat menyelesaikan permasalahan dengan menggunakan metode K-NN
  1. **Prinsip K-NN**

*K-Nearest Neighbor* (KNN) adalah suatu metode yang menggunakan algoritma *supervised* dimana hasil dari query instance yang baru diklasifikan berdasarkan mayoritas dari kategori pada KNN. Tujuan dari algoritma ini adalah mengklasifikasikan obyek baru bedasarkan atribut dan training sample. Classifier tidak menggunakan model apapun untuk dicocokkan dan hanya berdasarkan pada memori. Diberikan titik query, akan ditemukan sejumlah k obyek atau (titik training) yang paling dekat dengan titik query. Klasifikasi menggunakan voting terbanyak diantara klasifikasi dari k obyek.. algoritma KNN menggunakan klasifikasi ketetanggaan sebagai nilai prediksi dari query instance yang baru.

Algoritma metode KNN sangatlah sederhana, bekerja berdasarkan jarak terpendek dari query instance ke training sample untuk menentukan KNN-nya. Training sample diproyeksikan ke ruang berdimensi banyak, dimana masing-masing dimensi merepresentasikan fitur dari data. Ruang ini dibagi menjadi bagian-bagian berdasarkan klasifikasi training sample. Sebuah titik pada ruang ini ditandai kelas tertentu  jika kelas itu merupakan klasifikasi yang paling banyak ditemui pada k buah tetangga terdekat dari titik tersebut. Dekat atau jauhnya tetangga biasanya dihitung berdasarkan *Euclidean Distance* yang direpresentasikan sebagai berikut :

Nama lain dari NN

* lazy algorithm
* memory-based
* instance-based
* exemplar-based
* case-based
* experience-based

Algoritma KNN sangat sederhana. Algoritma ini bekerja berdasarkan jarak minimum dari data baru terhadap K tetangga terdekat yang telah ditetapkan. Setelah diperoleh K tetangga terdekat, prediksi kelas dari data baru akan ditentukan berdasarkan mayoritas K tetangga terdekat.

KNN dapat dibagi menjadi 2 jenis bersadarkan tetangga yang digunakan sebagai acuan perhitungan yaitu :

* 1-NN yaitu pengklasifikasikan dilakukan terhadap 1 data tetangga yang memiliki label terdekat
* k-NN yaitu pengklasifikasikan dilakukan terhadap k data tetangga yang memiliki label terdekat dengan K harus lebih besar 1 dan ganjil

Algoritma KNN:

1. Tentukan parameter K = jumlah tetangga terdekat
2. hitung jarak antara data baru dengan semua data training
3. urutkan jarak tersebut dan tetapkan tetangga terdekat berdasarkan jarak minimum ke-K
4. periksa kelas dari tetangga terdekat
5. gunakan mayoritas sederhana dari kelas tetangga terdekat sebagai nilai prediksi data baru

Contoh kasus:  
Pengenalan untuk menentukan seseorang itu mempunyai hipertensi atau tidak

**Tabel 4.1.** Tabel data hipertensi

|  |  |  |
| --- | --- | --- |
| Umur | Kegemukan | Hipertensi |
| muda | gemuk | Tidak |
| muda | sangat gemuk | Tidak |
| paruh baya | gemuk | Tidak |
| paruh baya | terlalu gemuk | Ya |
| tua | terlalu gemuk | Ya |

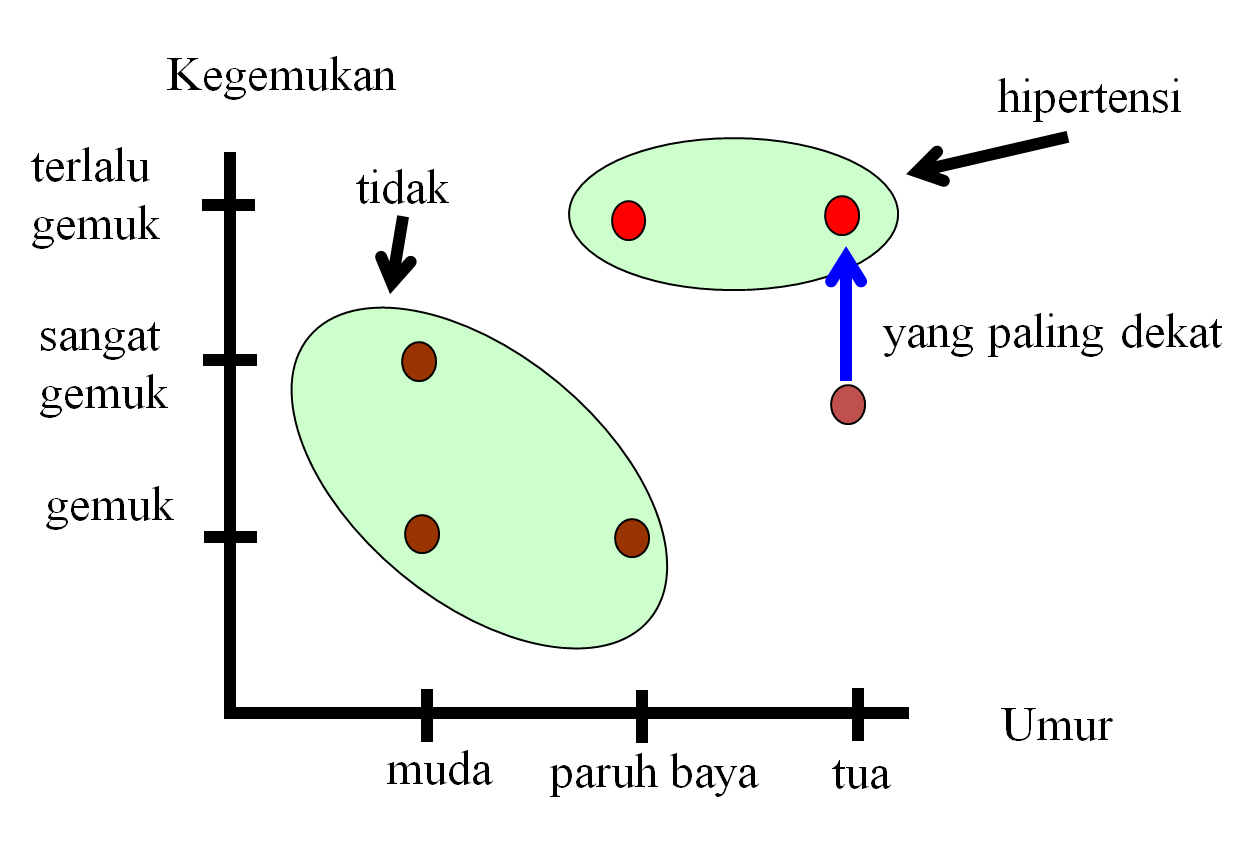
Jika data uji sebagai berikut maka termasuk hipertensi atau tidak?

**Tabel 4.2.** Tabel uji data hipertensi

|  |  |  |
| --- | --- | --- |
| Umur | Kegemukan | Hipertensi |
| tua | sangat gemuk | ? |

Penyelesaian dengan 1-NN

1. Gambarkan dalam bentuk grafik sebagai berikut



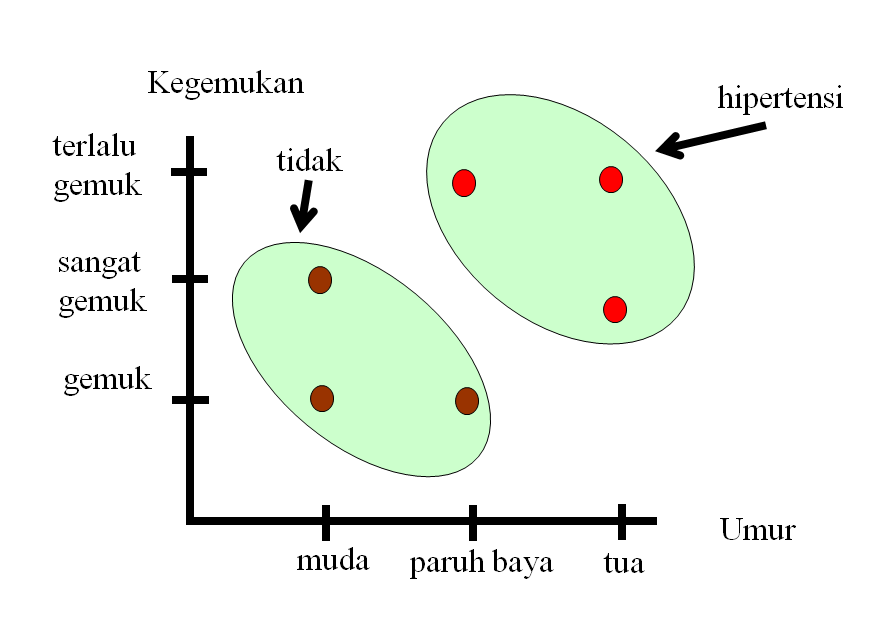
**Gambar 4.1.** Hasil grafik dari tabel data hipertensi

1. Hitunglah jarak titik yang diujikan dengan menggunakan Euclidean Distance, dengan asumsi nilai sebagai berikut:

**Tabel 4.3.** tabel konversi katagorikal ke interger

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| muda | 1 |  | gemuk | 1 |
| paruh baya | 2 |  | sangat gemuk | 2 |
| tua | 3 |  | terlalu gemuk | 3 |

1. Cari data terdekat

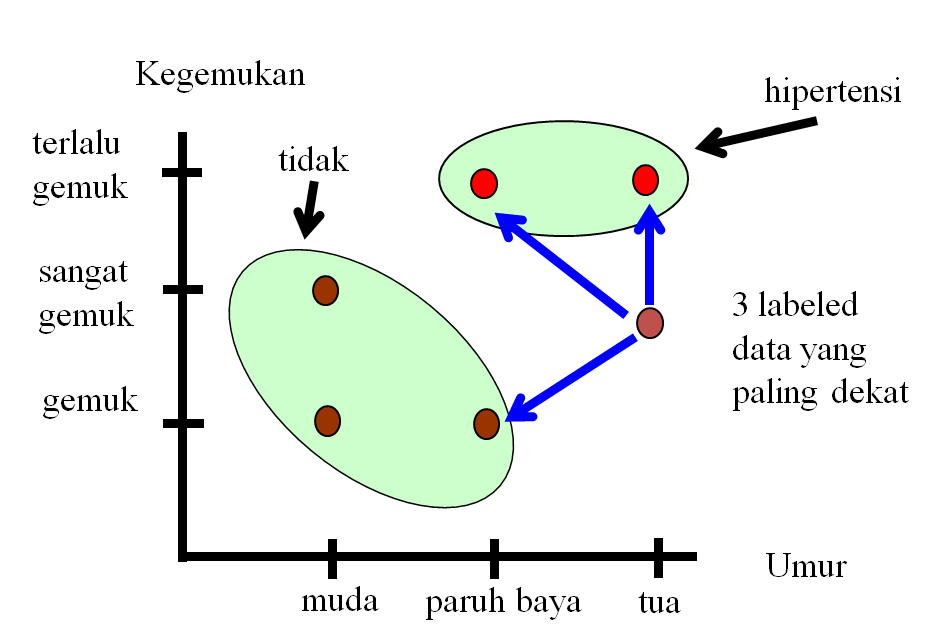


**Gambar 4.2.** Data baru masuk didalam kelas hipertensi

1. Berilah label data uji sesui dengan label data terdekat

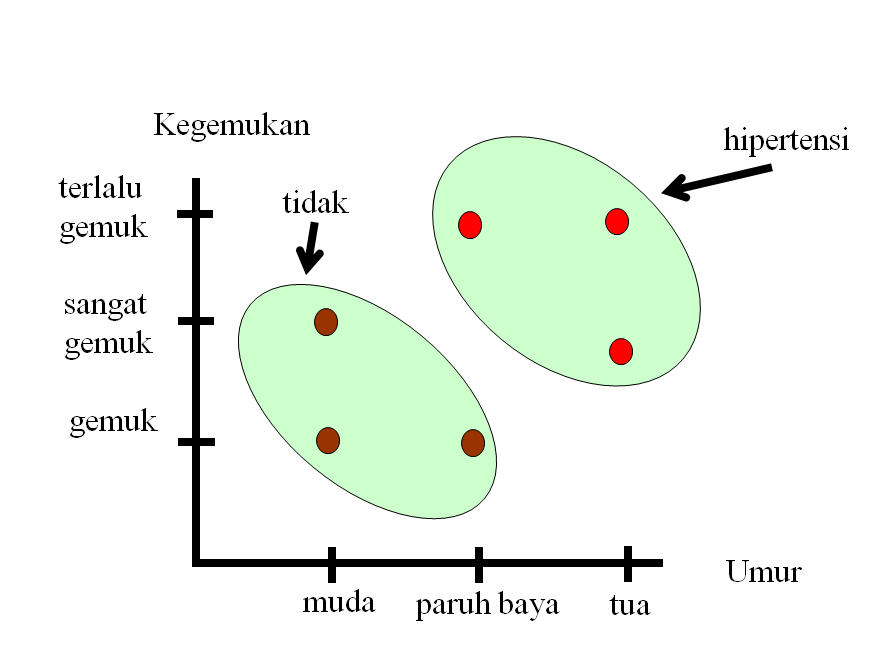
Algoritma k-NN

1. Tentukan k misalkan 3
2. Hitung jarak antara data baru ke setiap data
3. Tentukan k atau 3 data yang memilki label dan mempunyai jarak yang paling dekat



**Gambar 4.3.** Data baru dibandingkan dengan 3 data terdekat

1. Klasifikasikan data baru ke dalam data yang memilili label mayoritas



**Gambar 4.4.** Data baru termasuk kedalam kelas hipertensi

* 1. **Contoh Permasalahan 1**

Diberikan data training berikut dibawah ini, terdiri dari 2 atribut dengan skala kuantitatif yaitu X1 dan X2 serta 2 kelas yaitu baik dan buruk. Jika terdapat data baru dengan nilai X1=3 dan X2=7, tentukan kelasnya

data training:**Tabel 4.4.** data latih K-NN

|  |  |  |
| --- | --- | --- |
| X1 | X2 | Y |
| 7 | 7 | Buruk |
| 7 | 4 | Buruk |
| 3 | 4 | Baik |
| 1 | 4 | Baik |

1. Tentukan parameter K = jumlah tetangga terdekat

Misalkan ditetapkan K = 3

1. Hitung jarak antara data baru dengan semua data training

**Tabel 4.5.** Jarak data latih dengan data uji

|  |  |  |
| --- | --- | --- |
| X1 | X2 | Kuadrar jarak dengan data baru(3,7) |
| 7 | 7 | (7-3)2+(7-7)2=16 |
| 7 | 4 | (7-3)2+(4-7)2=25 |
| 3 | 4 | (3-3)2+(4-7)2=9 |
| 1 | 4 | (1-3)2+(4-7)2=13 |

1. Urutkan jarak tersebut dan tetapkan tetangga terdekat berdasarkan jarak minimum ke-K

**Tabel 4.6.** Mencari 3 data terdekat

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| X1 | X2 | Kuadrar jarak dengan data baru(3,7) | Peringkat Jarak minimum | Termasuk 3 tetangga terdekat |
| 7 | 7 | (7-3)2+(7-7)2=16 | 3 | Ya |
| 7 | 4 | (7-3)2+(4-7)2=25 | 4 | Tidak |
| 3 | 4 | (3-3)2+(4-7)2=9 | 1 | Ya |
| 1 | 4 | (1-3)2+(4-7)2=13 | 2 | Ya |

1. Periksa kelas dari tetangga terdekat

**Tabel 4.7.** Memberi label berdasarkan label terbanyak dari 3 data terdekat

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| X1 | X2 | Kuadrar jarak dengan data baru(3,7) | Peringkat Jarak minimum | Termasuk 3 tetangga terdekat | Y= kelas tetangga terdekat |
| 7 | 7 | (7-3)2+(7-7)2=16 | 3 | Ya | Buruk |
| 7 | 4 | (7-3)2+(4-7)2=25 | 4 | Tidak | - |
| 3 | 4 | (3-3)2+(4-7)2=9 | 1 | Ya | Baik |
| 1 | 4 | (1-3)2+(4-7)2=13 | 2 | Ya | Baik |

1. gunakan mayoritas sederhana dari kelas tetangga terdekat sebagai nilai prediksi data baru

Hasil pada no 4 menunjukkan bahwa dari 3 tetangga terdekat, terdapat 2 kelas Baik dan 1 kelas Buruk, maka disimpulkan bahwa data baru termasuk ke dalam ***kelas Baik.***

* 1. **Contoh Permasalahan 2**

Jika diketahui data seperti dalam gambar berikut ini, rubahlah dalam bentuk tabel untuk mempermudah perhitungan K-NN



**Gambar 4.5.** Data training ó’dan ‘+’ data, data testing ⏺

Data uji adalah data (3,4), fitur X=3, Y=4. Akan dilakukan prediksi, masuk dalam kelas yang manakah seharusnya jika mengunakan jarak Euclidean .

**Tabel 4.7.** Data dalam bentuk tabel

|  |  |  |  |
| --- | --- | --- | --- |
| Data | X | Y | Kelas |
| 1 | 1 | 1 | 0 |
| 2 | 2 | 1 | 0 |
| 3 | 3 | 1 | 0 |
| 4 | 3 | 2 | 0 |
| 5 | 7 | 2 | 1 |
| 6 | 1 | 3 | 0 |
| 7 | 2 | 3 | 0 |
| 8 | 5 | 3 | 1 |
| 9 | 4 | 4 | 1 |
| 10 | 6 | 4 | 1 |
| 11 | 1 | 5 | 0 |
| 12 | 6 | 5 | 1 |
| 13 | 1 | 6 | 0 |
| 14 | 4 | 6 | 1 |
| 15 | 5 | 6 | 1 |
| 16 | 2 | 7 | 1 |
| 17 | 4 | 7 | 1 |

Hitung jarak data uji (3,4) ke 17 data latih

**Tabel 4.7.** Jarak ke data latih dengan 1,3,7 NN

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Nomor data | x | y | Kelas asli | Jarak data uji ke data latih | 1-NN | 3-NN | 7-NN |
| 1 | 1 | 1 | 0 | 3.6055 | 0 | 0 | 0 |
| 2 | 2 | 1 | 0 | 3.1622 | 0 | 0 | 0 |
| 3 | 3 | 1 | 0 | 3 | 0 | 0 | 0 |
| 4 | 3 | 2 | 0 | 2 | 0 | 1 | 1 |
| 5 | 7 | 2 | 1 | 4.4721 | 0 | 0 | 0 |
| 6 | 1 | 3 | 0 | 2.236 | 0 | 0 | 1 |
| 7 | 2 | 3 | 0 | 1.4142 | 0 | 1 | 1 |
| 8 | 5 | 3 | 1 | 2.236 | 0 | 0 | 1 |
| 9 | 4 | 4 | 1 | 1 | 1 | 1 | 1 |
| 10 | 6 | 4 | 1 | 3 | 0 | 0 | 0 |
| 11 | 1 | 5 | 0 | 2.236 | 0 | 0 | 1 |
| 12 | 6 | 5 | 1 | 3.1622 | 0 | 0 | 0 |
| 13 | 1 | 6 | 0 | 2.8284 | 0 | 0 | 0 |
| 14 | 4 | 6 | 1 | 2.236 | 0 | 0 | 1 |
| 15 | 5 | 6 | 1 | 2.8284 | 0 | 0 | 0 |
| 16 | 2 | 7 | 1 | 3.1622 | 0 | 0 | 0 |
| 17 | 4 | 7 | 1 | 3.1622 | 0 | 0 | 0 |

Prediksi dengan K-NN

Untuk K=1

Data latih yang terdekat adalah data nomor 9 (4,4) dengan kelas 1, maka data uji (3,4) diprediksi masuk *kelas 1*.

Untuk K=3

Data latih yang terdekat adalah data nomor 9 (4,4) dengan kelas 1, data nomor 7 (2,3) dan data nomor 4 (3,2) dengan kelas 0, karena kelas 1 berjumlah 1 dan kelas 0 berjumlah 2 (*lebih banyak kelas 0 daripada kelas 1*) maka data uji (3,4) diprediksi masuk *kelas 0.*

Untuk K=7

Data latih yang terdekat adalah data nomor 8 (5,3), 9 (4,4), 14 (4,6) dengan kelas 1, data nomor 4 (3,2), 6 (1,3), 7 (2,3), dan 11 (1,5) dengan kelas 0, karena kelas 1 berjumlah 3 dan kelas 0 berjumlah 4 (*lebih banyak kelas 0 daripada kelas 1*) maka data uji (3,4) diprediksi masuk *kelas 0*.

**4.4. Rangkuman**

Algoritma K-NN dihitung berdasarkan tetangga terdekat dengan menggunakan jarak ecludian. Algoritma ini memiliki kelebihan sebagai berkikut:

* Analytically tractable
* Implementasi sangat sederhana
* Tingkat error > bayesian, < 2xbayesian
* Memungkinkan parallel implementation
* Robust terhadap data training yang memiliki noise (terutama jika digunakan invers kuadrat jarak terboboti sebagai “jarak”)
* Efektif jika data training berukuran besar

Sedangan kelemahan dari algoritma ini adalah:

* Butuh memori besar
* Komputasi besar

**4.5. Latihan Soal**

1. Dibawah ini adalah data Sapi dan Kambing .

**Tabel 4.7.** Jenis Hewan

|  |  |  |
| --- | --- | --- |
| Berat | Tinggi | Jenis Hewan |
| Berat | Pendek | Kambing |
| Berat | sedang | Sapi |
| Berat | Tinggi | Sapi |
| Sedang | Pendek | Kambing |
| Sedang | sedang | Kambing |
| Sedang | Tinggi | Kambing |
| Ringan | Tinggi | Sapi |

Jika terdapat data baru yang Beratnya Ringan dan Tingginya sedang, maka data tersebut kemungkinan Kambing atau Sapi?. Pakailah metode 1-NN dan k-NN (k=3) untuk menyelesaikan persoalan diatas!

1. Terdapat beberapa data yang berasal dari survey questioner tentang klasifikasi kualitas kertas tissue apakah baik atau jelek, dengan objek training menggunakan dua attribute yaitu daya tahan terhadap asam dan kekuatan.

**Tabel 4.7.** Jenis klasifikasi tisue

|  |  |  |
| --- | --- | --- |
| X1= Daya tahan asam(detik) | X2= Kekuatan (Kg/m2) | Y=Klasifikasi |
| 8 | 4 | Baik |
| 4 | 5 | Jelek |
| 4 | 6 | Jelek |
| 7 | 7 | Baik |
| 5 | 6 | Jelek |
| 6 | 5 | Baik |

Akan diproduksi kembali kertas tisu dengan attribute X1=7 dan X2=4 tanpa harus mengeluarkan biaya untuk melakukan survey, maka dapat diklasifikasikan kertas tise tersebut termasuk yang baik atau jelek.