Bab 2

Landasan teori

2.1 data mining

Data mining adalah suatu proses kegiatan menemukan pola yang menarik dari data dalam jumlah besar yang disimpan dalam penyimpanan database atau penyimpanan dengan menggunakan teknik pengenalan pola **seperti teknik statistic atau matematika**.

data mining merupakan sebuah proses dari input yang memiliki beberapa tahapan pengerjaan sehingga menghasilkan sebuah output. Tahap-tahap tersebut bersifat interaktif, pemakai terlibat langsung dengan perataraan knowledge base.

GAMBAR

Tahap-tahap data mining sebagai berikut:

1. Pembersihan data (data cleaning)

Pembersihan data merupakan proses menghilangkan noise dan data yang tidak konsisten atau data tidak relevan.

1. Integrasi data (data integration)

Integrasi data merupakan proses menggabungkan data dari berbagai database atau sumber data kedalam satu database baru.

1. Seleksi data (data selection)

Seleksi data merupakan proses pemilihan data pada database yang sering tidak dipakai oleh karena itu hanya data yang sesuai untuk di analisis yang akan diambil dari database.

1. Transformasi data (data transformation)

Transformasi data merupakan data diubah atau digabung kedalam format yang sesuai untuk di proses dalam data mining.

1. Proses mining (data mining)

Proses data mining merupakan proses utama saat metode diterapkan untuk menemukan pengetahuan berharga dan tersembunyi dari data.

1. Evaluasi pola (pattern evaluation)

Evaluasi pola merupakan tahap untuk mengidentifikasi pola-pola menarik kedalam based yang ditemukan sehingga menghasilkan pengetahuan yang jelas dipahami.

1. Presentasi pengetahuan (knowledge presentation)

Presentasi pengetahuan merupakan proses mempresentasikan pengetahuan hasil dari penambangan data dengan menggunakan teknik visualisasi untuk membuat data dipahami oleh pengguna.

2.2 naïve bayes

Klasifikasi bayes sederhana yang lebih dikenal sebagai naïve Bayesian classifier dapat di asumsikan bahwa efek dari suatu nilai atribut sebuah kelas yang diberikan adalah bebas dari atribut-atribut lain. Asumsi ini disebut class conditional independence yang dibuat untuk memudahkan perhitungan-perhitungan, pengertian ini dianggap “naïve”, dalam bahasa lebih sederhana naïve itu mengasumsikan bahwa kemunculan suatu term kata dalam suatu kalimat tidak dipengaruhi kemungkinan kata-kata yang lain dalam kalimat padahal kenyataannya bahwa kemungkinan kata dalam kalimat sangat dipengaruhi kemungkinan keberadaan kata-kata dalam kalimat. (Sulaksono & Darsono, 2015)

Secara umum, teorema bayes dapat ditulisakn dalam bentuk persamaan berikut :

P(H|X) =

Keterangan :

X : data dengan class yang belum diketahui.

H : Hipotesis data merupakan suatu class spesifik.

P(H|X) : Probabilitas hipotesis H berdasarkan kondisi X.

P(X|H) : Probabilitas X berdasarkan kondisi pada hipotesis H.

P(H) : Probabilitas hipotesis H.

P(X) : Probabilitas X

Salah satu hal berpotensi untuk menambah nilai akurasi dari naïve bayes classifier adalah dengan nilai korelasi atribut terhadap kelas. Perhitungan korelasi nilai atribut terhadap kelas akan menjadi dasar ketepatan dari klasifikasi yang tidak hanya probabilitas namun juga seberapa besar korelasi atribut dengan kelas.

2.3 correlated naïve bayes

Metode correlated naïve bayes classifier merupakan sebuah pengembangan dari metode naïve bayes. Pada metode correlated naïve bayes classifier memperhitungkan nilai korelasi (R-Square) antara variabel bebas (X) terhadap variabel terikat (Y). penambahan parameter korelasi digunakan untuk mengukur tinggi rendahnya derajat hubungan antara variabel bebas (X) terhadap variabel terikat (Y). (Marzuki et al., 2018)

Rumus algoritma correlated naïve bayes untuk klasifikasi:

P(Y|X) =

Keterangan :

X : Data dengan kelas yang belum diketahui.

Y : Hipotesis data X merupakan suatu kelas spesifik.

P(X|Y) : Probabilitas hipotesis Y berdasarkan kondisi Y.

P(Y) : Probabilitas hipotesis Y (prior probability)

P(X) : Probabilitas dari X.

R() : R-Square setiap atribut dari data X berdasarkan kondisi hipotesis Y.

: Probabilitas setiap atribut dari data X berdasarkan kondisi hipotesis Y

: R-Square setiap atribut dari data X berdasarkan kondisi hipotesis Y.

: Bilangan laplacian

Berikut merupakan persamaan untuk menentukan perhitungan korelasi atribut :

Keterangan :

R : R-Square fitur antar kelas.

r : Nilai Korelasi antar fitur kelas.

n : Total data pada dataset.

: Total perkalian variabel X dengan variabel Y.

: Total variabel X.

: Total variabel Y.

: Total variabel X yang dikuadratkan.

: Total variabel Y yang dikuadratkan.

: Kuadrat dari total variabel X.

: Kuadrat dari total variabel Y.

Nilai (r) memiliki ketentuan dari nilai koefisien korelasi yakni -1 <= r <= 1. Interprestasi koefisien korelasi nilai (r) ditunjukkan pada tabel :

|  |  |
| --- | --- |
| Interval Koefisien | Tingkat Hubungan |
| 0 – 0.199 | Sangat Rendah |
| 0.20 – 0.299 | Rendah |
| 0.40 – 0.599 | Cukup |
| 0.60 – 0.799 | Kuat |
| 0.80 – 1 | Sangat Kuat |

2.4 crossvalidation

c

Bab 3

Metodologi penelitian

3.1 Data

Data yang digunakan dalam penelitian ini adalah data jantung yang diperoleh dari situs kaggle.com yang berjudul “heart”. Data dalam penelitian ini berjumlah 918 record data yang terdiri dari 12 atribut.

Penjelasan masing-masing dari tiap atribut:

Daftar pustaka

Azis, H., Fattah, F., & Putri, P. (2020). *Performa Klasifikasi K-NN dan Cross-validation pada Data Pasien Pengidap Penyakit Jantung*. *12*(2), 81–86.

Marzuki, J. I., Mataram, K., & Bar, N. T. (2018). KOMPARASI AKURASI METODE CORRELATED NAIVE BAYES CLASSIFIER DAN NAIVE BAYES CLASSIFIER UNTUK DIAGNOSIS PENYAKIT DIABETES Hairani , Gibran Satya Nugraha , Mokhammad Nurkholis Abdillah , Muhammad Innuddin InfoTekJar ( Jurnal Nasional Informatika dan Teknolog. *InfoTekJar (Jurnal Nasional Informatika Dan Teknologi Jaringan)*, *3*(1), 6–11.

Sulaksono, J., & Darsono. (2015). Sistem pakar penentuan penyakit gagal jantung menggunakan metode naive bayes classifier. *Seminar Nasional Teknologi Informasi Dan Multimedia 2015*, 6–8.