Tugas

MAKALAH PROSES DATA MINING

Disusun Oleh:

Nim: 2000019

Nama: Muhammad Ajriel Rahayu

Universitas: Universitas Pendidikan Indonesia

UNIVERSITAS NURUL JADID PAITON PROBOLINGGO

FAKULTAS TEKNIK

DAFTAR ISI

BAB I

PENDAHULUAN

1.1 Latar belakang Masalah

Pada jaman sekarang ini jumlah data di dunia sudah sulit terhitung banyaknya, hal ini dikarenakan semakin pesatnya manusia membuat data-data baru, apalagi dengan berbagai teknologi yang turut serta memudahkan untuk menghasilkan data-data yang sangat banyak, data-data yang dihasilkan tersebut ternyata memiliki banyak keuntungan bagi kita semua, apalagi jika data-data tersebut di teliti dan diolah kembali oleh kita semua, akan ada banyak sekali informasi baru yang baru kita ketahui dari data-data tersebut. Salah satu metode untuk mengolah kumpulan data tersebut adalah dengan menggunakan metode data mining, yaitu metode pengolahan data untuk mencari informasi atau pengetahuan yang menarik di dalam sebuah kumpulan data, untuk menggunakan metode tersebut tentu saja kita harus mengetahui proses apa saja yang dilalui pada metode data mining ini supaya hasilnya akan terasa baik.

1.2 Rumusan Masalah

Apa saja proses yang ada pada metode data mining?

1.3 Tujuan

Mendeskripsikan proses apa saja yang dilakukan pada pengolahan data melalui metode data mining.

BAB II

PEMBAHASAN

Untuk melakukan pengolahan metode data mining yang benar, diperlukan proses-proses terstruktur yang dilakukan, yaitu:

2.1 Pembersihan Data

Langkah ini merupakan langkah awal pada metode data mining. Pembersihan data dilakukan untuk membuang data yang mempunyai missing value pada record dan juga menghilangkan noise. Jika pada data terdapat nilai yang kosong dapat dibersihkan dengan cara-cara berikut:

a. Mengabaikan tuple
b. Mengisi manual atribut yang kosong
c. Mengisi atribut kosong dengan nilai rata-rata

2.2 Integrasi Data

Proses ini dilakukan dengan menggabungkan data dari berbagai sumber database ke dalam suatu database baru, proses ini bertujuan untuk meningkatkan akurasi dan juga kecepatan proses data mining.

2.3 Reduksi Data

Proses ini dilakukan dengan mengurangi dimensi atau atribut pada datasets, proses ini bertujuan untuk mengoptimalkan atribut yang berpengaruh saja pada proses data mining sehingga dapat menjaga keakuratan proses data mining. Reduksi data dilakukan dengan beberapa metode yaitu:

a. Metode Information Gain

Metode Information Gain adalah metode yang menggunakan teknik scoring untuk pembobotan sebuah fitur dengan menggunakan maksimal entropy. Fitur yang dipilih adalah fitur dengan nilai Information Gain yang lebih besar atau sama dengan nilai threshold tertentu. Persamaan matematis nya sebagai berikut:

- Tanpa missing value

$IG(S, A) = Entropy(S) - \sum c \in values(A)\ |Sv|\ Entropy(Sv)$

- Dengan missing value

$Entropy(S) = \sum c - pi log2pi$

Dimana:

S : himpunan kasus

A : atribut

$|Si|$: jumlah kasus pada partisi ke-i

$|S|$: jumlah kasus dalam S

c : jumlah partisi S

Pi : proporsi dari Si terhadap S.

b. *Particle Swarm Optimization* (PSO)

Particle Swarm Optimization (PSO) adalah sebuah teknik stochastic optimization berdasarkan populasi (ikan, lebah, burung dll), dikemukakan oleh Russell C. Eberhart dan James Kennedy di tahun 1995 yang terinspirasi oleh perilaku sosial dari pergerakan burung atau ikan (Hu, 2005). Proses dasar dari algoritma PSO yaitu sebagai berikut.

1. Inisialisasi : secara acak menghasilkan partikel awal

2. Fitness: ukuran fitness setiap partikel dalam populasi

3. Update: menghitung kecepatan setiap partikel

4. Konstruksi: untuk setiap partikel, bergerak ke posisi berikutnya

5. Penghentian: hentikan algoritma jika kriteria penghentian dipenuhi, dan kembali ke langkah 2 (fitness). Iterasi dihentikan jika jumlah iterasi mencapai jumlah maksimum yang telah ditentukan dan grafik konvergen

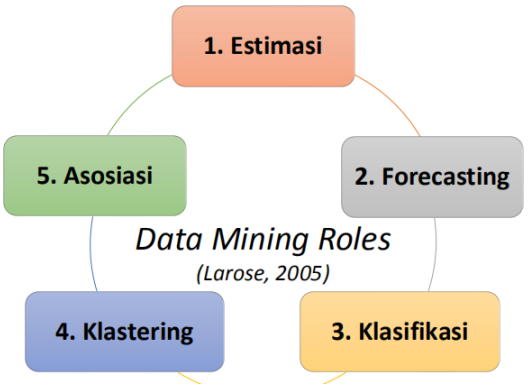
2.4 Data Transformasi

Pada proses ini dilakukan pengubahan data menjadi bentuk atau format yang sesuai untuk pemprosesan data mining, langkah ini dilakukan karena pada beberapa metode pada data mining membutuhkan format data yang khusus sebelum bisa diaplikasikan. Ada beberapa transformasi yang bisa dilakukan diantaranya

a. Diskritisasi : mengubah atribut kontinyu menjadi atribut kategoris
b. *Entropy-Based Discretization*: merupakan supervised, dengan menggunakan teknik split topdown

2.5 Teknik Data Mining

Proses ini merupakan proses yang paling utama dalam data mining, karena pada proses ini kita bisa menemukan suatu pengetahuan tersembunyi dari data, ada beberapa cara dalam melakukan teknik data diantaranya yaitu klasifikasi, klastering, asosiasi dan juga regresi



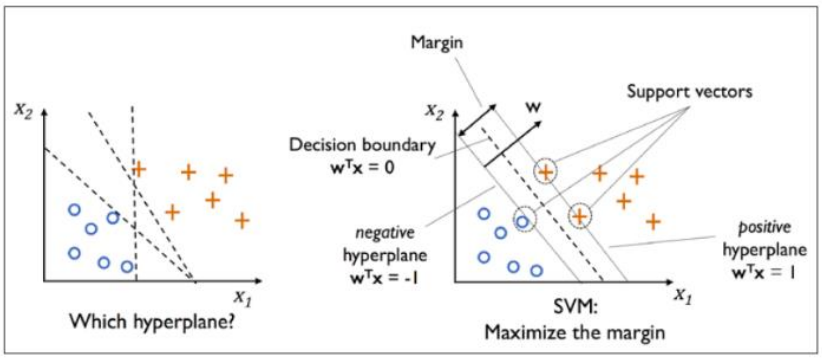
a. Klasifikasi

Metode ini melakukan pengelompokan obyek berdasarkan kelompok yang sudah ada. Klasifikasi ini memerlukan data pelatihan yang sudah diberi label kelompok/kelas. Prediksi pengelompokan dilakukan dengan membangun model terlebih dahulu melalui proses pelatihan menggunakan dara yang sudah kita siapkan, setelah model berbentuk dari proses pelatihan, data baru bisa dikelompokkan menggunakan model tersebut.

Ada beberapa metode klasifikasi diantaranya *Support Vector Machine, Multilayer Perceptron, Naive bayes, ID3, Ensemble Methode*, dll.

1. *Support Vector Machine* (SVM)

Konsep SVM dapat dijelaskan secara sederhana sebagai usaha mencari hyperplane terbaik yang berfungsi sebagai pemisah dua buah class pada input. Hyperplane pemisah terbaik antara kedua class dapat ditemukan dengan mengukur margin hyperplane tersebut dan mencari titik maksimal nya. Margin adalah jarak antara hyperplane tersebut dengan pattern terdekat dari masing-masing class. Pattern yang paling dekat ini disebut sebagai support vector



Langkah yang dilakukan pada proses ini yaitu diawali dengan pendefinisian persamaan yang dituliskan dengan

w. X + b = 0

jika b dianggap suatu bobot tambahan *w*0, maka persamaan suatu hyperplane pemisah dapat ditulis ulang seperti pada persamaan berikut.

$w_0 + w_1x_1 + w_2x_2 = 0$

Setelah persamaan dapat didefinisikan, nilai *x*1 dan *x*2 dapat dimasukkan ke dalam persamaan untuk mencari bobot *w*1, *w*2, dan *w*0 atau b

2. *Multilayer Percepton*(MLP)

Multilayer Perceptron atau yang biasa disingkat MLP adalah salah satu jenis dari feed-forward neural network dengan satu atau lebih hidden layer. Pada umumnya MLP terdiri dari layer input yang merupakan kumpulan neuron untuk memasukkan data; minimal satu hidden layer sebagai neuron komputasi dan layer output sebagai neuron penampung hasil komputasi.

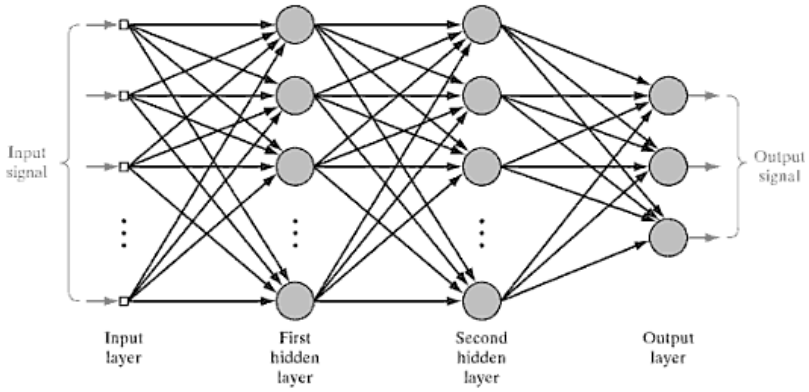


FIGURE 4.1 Architectural graph of a multilayer perceptron with two hidden layers.

3. Naïve Bayes

Naïve Bayes Classifier merupakan sebuah metoda klasifikasi yang berakar pada teorema Bayes . Bayes merupakan teknik prediksi berbasis probalistik sederhana yang berdasar pada penerapan teorema bayes atau aturan bayes dengan asumsi independensi (ketidaktergantungan) yang kuat (naïve). Persamaan naïve bayes dapat diformulasikan sebagai berikut:

Likelihood Class Prior Probability

$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

Posterior Probability Predictor Prior Probability

$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

b. *Clustering* (Klasterisasi)

Mengelompokkan obyek ke dalam beberapa kelompok bedasarkan kemiripan antar obyek, dimana dalam satu klaster harus berisi obyek yang saling mirip dan antar klaster obyek saling tidak mirip. Klustering ini tidak memerlukan data pelatihan yang sudah di beri label, klustering ini mempunyai beberapa metode diantaranya, *K-Medoids, K-Means, Fuzzy C-Means, Self-Organizing Map (SOM)*, dll.

c. Regresi/estimasi

Regresi pada dasarnya mirip dengan klasifikasi, yakni memerlukan data pelatihan yang sudah diberi label. Bedanya output klasifikasi adalah nilai diskrit, sedangkan output dari regresi adalah nilai kontinyu. Regresi ini mencari hubungan antara atribut predictor dan atribut dependen, dimana atribut dependen nya juga berupa nilai kontinyu. Metode yang sering digunakan *adalah linear regression, logistic regression, support vector regression*, dll.

2.6 Pengetahuan (Knowledge)

Setelah proses data mining selesai tentu saja hasil dari temuan informasi baru atau juga pengetahuan harus bisa dikomunikasikan kepada pengguna Pengetahuan dari data mining adalah bagaimana cara mengatur pembuatan keputusan tertentu atau aksi selanjutnya yang akan dilakukan sesuai dari hasil analisis yang didapat. Memvisualisasikan temuan bisa menjadi cara untuk mengkomunikasikan data ke pengguna.

2.7 Evaluasi

Evaluasi adalah proses penilaian pola-pola menarik atau model prediksi apakah memenuhi hipotesa awal atau belum. Apabila hasil yang didapatkan tidak sesuai dengan hipotesa maka dapat diambil beberapa pilihan alternatif untuk memperbaiki proses data mining dengan cara mencoba metode data mining lain yang sesuai atau menerima hasil ini sebagai suatu hasil di luar dugaan yang mungkin bermanfaat.

BAB III

PENUTUP

3.1 Kesimpulan

Tahapan proses pengolahan data melalui metode data mining dimulai dengan membersihkan data, mengintegrasi data, mereduksi data, transformasi data, kemudian melakukan pemodelan dengan teknik data mining yang digolongkan menjadi klasifikasi, klasterisasi, regresi dan asosiasi, setelah pemodelan selesai, dilanjutkan dengan mempresentasikan pengetahuan yang didapat melalui visualisasi, dan diakhiri dengan evaluasi model, jika model sudah sesuai maka proses selesai dan jika model dirasa kurang tepat maka dilakukan proses pemodelan ulang.

DAFTAR PUSTAKA

Widianto, M.H. 2019. *Algoritma Naïve Bayes.* Diakses dari : https://binus.ac.id/bandung/2019/12/algoritma-naive-bayes/

Umam, A. & Santosa, B. 2018. *Data Mining dan Big Data Analitics.* Yogyakarta : Media Pustaka

Faid, M dkk. 2021. *Modul Pembelajaran Data Mining With Python.* Diakses dari : https://elearning.unuja.ac.id/pluginfile.php/40718/mod_resource/content/1/Modul%20Data%20Mining%20Fix.pdf

Wibawa, M.S & Maysanjaya, D . 2018. *Multi Layer Percepton dan Principal Component Alanysis Untuk Diagnosa Kanker Payudara*