

Nama : Akmal Zuhdy Prasetya

NIM : H071191035

Kelas : Pengantar Data Mining

Resume Pertemuan 2

A. Apa itu Data Mining

Data mining adalah proses mengekstraksi dan menemukan pola dalam kumpulan data besar yang melibatkan metode machine learning, statistik, dan sistem basis data. Data mining adalah subbidang ilmu komputer dan statistik dengan tujuan keseluruhan untuk mengekstrak informasi dari kumpulan data dan mengubah informasi menjadi struktur yang dapat dipahami untuk penggunaan lebih lanjut. Data mining adalah langkah analisis dari proses "Knowledge Discovery in Databases" atau KDD. Selain langkah analisis mentah, data mining juga melibatkan aspek database dan manajemen data, pra-pemrosesan data, pertimbangan model dan inferensi, metrik ketertarikan, pertimbangan kompleksitas, pasca-pemrosesan struktur yang ditemukan, visualisasi, dan pembaruan online. Istilah "data mining" sebenarnya merupakan miskonsepsi, karena tujuannya adalah ekstraksi pola dan pengetahuan dari sejumlah besar data, bukan ekstraksi data itu sendiri.

B. Mengapa Data Mining Penting

Data mining penting karena dengannya kita dimungkinkan untuk:

- Menyaring semua kebisingan (noise) yang kacau dan berulang dalam data
- Memahami apa yang relevan dan memanfaatkan informasi itu dengan baik untuk menilai kemungkinan hasil
- Mempercepat langkah membuat keputusan berdasarkan informasi

C. Proses Data Mining

Proses data mining melibatkan sejumlah langkah dari pengumpulan data hingga visualisasi untuk mengekstrak informasi berharga dari kumpulan data besar. Seperti disebutkan di atas, teknik data mining digunakan untuk menghasilkan deskripsi dan prediksi tentang kumpulan data target. Ilmuwan data menggambarkan data melalui pengamatan mereka terhadap pola, asosiasi, dan korelasi. Data mining biasanya terdiri dari empat langkah utama yaitu menetapkan tujuan, pengumpulan dan persiapan data, menerapkan algoritma data mining, dan mengevaluasi hasil.

1. Menetapkan Tujuan

Ini dapat menjadi bagian tersulit dari proses data mining, dan banyak organisasi menghabiskan terlalu sedikit waktu untuk langkah penting ini. Ilmuwan data dan pemangku kepentingan bisnis perlu bekerja sama untuk mendefinisikan masalah bisnis, yang membantu menginformasikan pertanyaan dan parameter data untuk proyek tertentu. Analisis mungkin juga perlu melakukan penelitian tambahan untuk memahami konteks bisnis dengan tepat.

2. Persiapan Data

Setelah ruang lingkup masalah ditentukan, lebih mudah bagi ilmuwan data untuk mengidentifikasi kumpulan data mana yang akan membantu menjawab pertanyaan terkait bisnis. Setelah mereka mengumpulkan data yang relevan, data akan dibersihkan, menghilangkan kebisingan, seperti duplikat, nilai yang hilang, dan outlier. Bergantung pada kumpulan data, langkah tambahan dapat diambil untuk mengurangi jumlah dimensi karena terlalu banyak fitur dapat memperlambat komputasi berikutnya.

3. Pembuatan Model & Penambangan Pola

Bergantung pada jenis analisis, ilmuwan data dapat menyelidiki hubungan data yang menarik, seperti pola sekuensial, aturan asosiasi, atau korelasi. Sementara pola frekuensi tinggi memiliki aplikasi yang lebih luas, terkadang penyimpangan dalam data bisa lebih menarik, menyoroti area potensi penipuan.

4. Evaluasi Hasil & Implementasi Pengetahuan

Setelah data dikumpulkan, hasilnya perlu dievaluasi dan diinterpretasikan. Ketika menyelesaikan hasil, mereka harus valid, baru, berguna, dan dapat dimengerti. Ketika kriteria ini terpenuhi, organisasi dapat menggunakan pengetahuan ini untuk menerapkan strategi baru, mencapai tujuan yang diinginkan.

D. Teknik Data Mining

Data mining bekerja dengan menggunakan berbagai algoritma dan teknik untuk mengubah volume data yang besar menjadi informasi yang berguna. Berikut adalah beberapa yang paling umum:

1. Association Rules

Aturan asosiasi adalah metode berbasis aturan untuk menemukan hubungan antara variabel dalam kumpulan data yang diberikan. Metode ini sering digunakan untuk analisis keranjang pasar, memungkinkan perusahaan untuk lebih memahami hubungan antara produk yang berbeda. Memahami kebiasaan konsumsi pelanggan memungkinkan bisnis mengembangkan strategi penjualan silang dan mesin rekomendasi yang lebih baik.

2. Neural Networks

Dimanfaatkan terutama untuk algoritma pembelajaran mendalam, neural networks memproses data pelatihan dengan meniru interkoneksi otak manusia melalui lapisan node. Setiap node terdiri dari input, bobot, bias (atau ambang batas), dan output. Jika nilai output itu melebihi ambang batas yang diberikan, ia "membakar" atau mengaktifkan node, meneruskan data ke lapisan berikutnya dalam jaringan. Neural networks mempelajari fungsi pemetaan ini melalui pembelajaran yang diawasi, menyesuaikan berdasarkan fungsi kerugian melalui proses penurunan gradien. Ketika fungsi biaya berada pada atau mendekati nol, kita dapat yakin dengan akurasi model untuk menghasilkan jawaban yang benar.

3. Decision Tree

Teknik data mining ini menggunakan metode klasifikasi atau regresi untuk mengklasifikasikan atau memprediksi hasil potensial berdasarkan serangkaian keputusan. Seperti namanya, ini menggunakan visualisasi seperti pohon untuk mewakili hasil potensial dari keputusan ini.

4. K-Nearest Neighbor (KNN)

Juga dikenal sebagai algoritma KNN, adalah algoritma non-parametrik yang mengklasifikasikan titik data berdasarkan kedekatan dan asosiasinya dengan data lain yang tersedia. Algoritma ini mengasumsikan bahwa titik data yang serupa dapat ditemukan berdekatan satu sama lain. Akibatnya, ia berusaha menghitung jarak antara titik data, biasanya melalui jarak Euclidean, dan kemudian menetapkan kategori berdasarkan kategori atau rata-rata yang paling sering.