NAMA : RADEN ISNAWAN ARGI ARYASATYA
NIM : 195410257
KELAS : IF-5
MATKUL : DATA MINING

UJIAN TENGAH SEMESTER

1. **Jelaskan apa yang anda ketahui tentang data mining dan bidang ilmu terkait data mining?**
   Jawab:
   Data mining adalah sebuah kegiatan atau sebuah proses untuk melakukan ekstraksi untuk mendapatkan informasi penting yang sifatnya implisit dan sebelumnya tidak diketahui dari suatu data. Data mining juga bisa diartikan dengan suatu kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola, dan hubungan dalam set data berukuran besar. Selain itu, data mining bisa juga disebut sebagai penambangan atau penemuan informasi baru dengan mencari pola atau aturan tertentu dari sejumlah data yang sangat besar.

   Data mining memiliki beberapa karakteristik yaitu:
   - **Pertama,** data mining berhubungan dengan penemuan sesuatu yang tersembunyi dan pola data tertentu yang tidak diketahui sebelumnya.
   - **Kedua,** data mining biasanya menggunakan data yang sangat besar. Data yang besar digunakan untuk membuat hasil yang lebih otentik atau bisa lebih dipercaya.
   - **Ketiga,** data mining berguna untuk membuat keputusan yang kritis, terutama dalam strategi.

   Berikut bidang-bidang ilmu yang terkait dengan data mining:
   a. <u>Statistika</u>
      Statistika adalah sebuah ilmu yang mempelajari bagaimana cara merencanakan, mengumpulkan, menganalisis, lalu menginterpretasikan, dan akhirnya mempresentasikan data. Statistika berkaitan dengan data mining karena statistika merupakan bagian inti dari data mining yang mencakup seluruh proses analisis data. Statistika membantu dalam mengidentifikasi pola yang bisa membantu mengidentifikasi gangguan pada data, memberikan teori untuk memperkirakan probabilitas atau prediksi, dan lain-lain.

   b. <u>Machine Learning</u>
      Machine learning (ML) adalah mesin yang dikembangkan untuk bisa belajar dengan sendirinya tanpa arahan dari penggunanya. Machine learning merupakan cabang dari AI yang berfokus pada penggunaan data dan algoritma untuk meniru cara manusia belajar dan secara bertahap dapat meningkatkan akurasinya. Machine Learning berkaitan dengan Data Mining karena data mining diimplementasikan pada dataset tertentu untuk menemukan suatu pola unik antar item dalam dataset, dan untuk mengambil data dan memprediksi hasil dari dataset tersebut, Data Mining menggunakan teknik yang dibuat oleh Machine Learning.

   c. <u>Artificial Intelligence</u>
      Artificial Intelligence atau juga bisa disebut dengan Kecerdasan Buatan adalah suatu simulasi dari kecerdasan yang dimiliki oleh manusia yang dimodelkan di dalam mesin dan diprogram agar bisa berpikir seperti halnya manusia. AI berkaitan dengan data mining karena pada proses data mining akan memerlukan beberapa teknik dan algoritma AI untuk memeriksa data dan mendapatkan beberapa hasil penting dari dataset. Teknik ini termasuk dalam teknik Machine Learning dan dibagi menjadi supervised learning dan unsupervised learning.

d. Sistem Basis Data

Sistem basis data adalah sistem yang terdiri atas kumpulan tabel data yang saling berhubungan dan kumpulan program yang memungkinkan beberapa pemakai atau program lain untuk mengakses dan memanipulasi tabel tabel data tersebut. Basis data berhubungan dengan data mining karena data mining mengambil dan menemukan data-data dari sebuah dataset di dalam sebuah basis data atau database.

e. Visualisasi

Visualisasi adalah proses merepresentasikan data secara grafis dan sekaligus berinteraksi dengan representasi untuk mendapatkan wawasan atau pengetahuan tentang data. Visualisasi adalah representasi grafis dari informasi dan data kuantitatif dengan menggunakan elemen visual seperti grafik, bagan, dan map. Visualisasi data mengubah kumpulan data besar dan kecil menjadi sebuah bentuk visual yang mudah dipahami dan diproses oleh manusia. Tool visualisasi data menyediakan cara yang dapat diakses untuk memahami outlier, pola, dan tren dalam data. Visualisasi berkaitan dengan data mining karena visualisasi merupakan cara untuk mempresentasikan hasil pengumpulan atau pengolahan data pada data mining.

2. Tahap/proses/bagian mana yang paling penting dalam data mining? Jelaskan dan urutkan dari yang paling penting.

Jawab:

Proses paling penting dalam data mining tentunya adalah proses pertama yaitu data cleaning karena data mining tidak bisa dimulai jika data yang diseleksi tidak dibersihkan terlebih dahulu. Dan jika memaksakan untuk memulai proses data mining tanpa data cleaning, maka akan terjadi gangguan pada tahap-tahap selanjutnya yang bisa berakibat cukup buruk terutama jika data yang akan dipresentasikan atau divisualisasikan cukup penting dan krusial.

Tahapan data mining:

1) Data cleaning

Adalah proses menghapus data yang tidak akurat, tidak lengkap, salah, atau "kotor" dari source data yang digunakan. Data dibersihkan dengan mengembalikan data yang hilang atau menghapus data yang “kotor”.

2) Data integration

Mengumpulkan dan menggabungkan berbagai macam sumber data menjadi satu sumber yang cocok untuk proses analisis dan manipulasi data. Data integration mengintegrasikan semua data yang tersedia dengan meningkatkan kecepatan dan akurasi data mining.

3) Data reduction

Merupakan proses “slicing and dicing” data untuk mendapatkan informasi yang paling relevan dari keseluruhan data yang besar, tanpa mengganggu integritas keseluruhan dari sumber data atau sampel yang akan diambil.

4) Data transformation

Data yang telah melewati proses data cleaning, data reduction, dan juga telah dioptimalkan selanjutnya ditransformasikan dalam tahap ini. Data disempurnakan dengan lebih teliti, menghaluskan outlier, meringkas kumpulan data jika, mengganti nilai data raw dengan ranges seperti discretization.

5) Data mining

Data mining adalah sebuah kegiatan atau sebuah proses untuk melakukan ekstraksi untuk mendapatkan informasi penting yang sifatnya implisit dan sebelumnya tidak diketahui dari suatu data. Teknik, metode, atau algoritma dalam data mining sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses data mining secara keseluruhan.

6) Pattern Evaluation

pola yang ditemukan selama penambangan data dianalisis dan diubah menjadi informasi berguna yang dapat dipahami oleh para end-user. Di tahap ini juga mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya.

7) Knowledge Representation

mengubah informasi yang telah diproses ke dalam format multimedia berbentuk visualisasi untuk tinjauan, analisis, dan presentasi lebih lanjut. Pengetahuan atau wawasan yang diperoleh selama evaluasi pola dapat digunakan untuk membuat perkiraan penjualan, penyesuaian rantai pasokan, jadwal produksi baru, dll tergantung dengan kasusnya.

3. Sebutkan dan jelaskan contoh implementasi data mining dalam kehidupan sehari-hari.

Jawab:

a) Pendidikan

Di dunia pendidikan, ada yang disebut dengan Educational Data Mining (EDM). EDM merupakan sebuah metode pengembangan yang dapat menemukan knowledge atau informasi dari data yang berasal dari lingkungan pendidikan. EDM digunakan sebagai prediksi perilaku belajar siswa, prediksi hasil pembelajaran siswa, mempelajari dampak dukungan pendidikan, dan memajukan pengetahuan ilmiah pada pembelajaran.

b) Kesehatan

Dengan data mining, tenaga kerja IT di rumah sakit bisa menggunakan data dan analisis untuk mengidentifikasi praktik terbaik untuk meningkatkan kualitas perawatan pada pasien, menghitung volume pasien dalam setiap periode tertentu, sekaligus mengurangi biaya perawatan dan operasional rumah sakit. Hal itu bisa dilakukan dengan menggunakan pendekatan data mining seperti database multi dimensi, machine learning, soft computing, visualisasi data dan statistik.

c) Olahraga

Data mining digunakan sebagai sarana untuk mencatat jumlah statistik pemain atau tim pada statistik pertandingan sepak bola, basket, tenis, voli, dll. Data mining di bidang olahraga diterapkan pada IBM Advanced Scout untuk analisis statistik permainan NBA. Sehingga dapat mencapai keunggulan dalam bersaing utuk tim Miami Heat dan New York Knicks.

d) Keuangan

Financial Crimes Enforcement Network yang berada di Amerika Serikat memakai data mining untuk mengumpulkan berbagai subyek seperti rekening bank, properti hingga berbagai transaksi keuangan lainnya. Tujuannya adalah untuk mendeteksi transaksi keuangan yang mencurigakan.

e) Manufaktur

Data mining bisa sangat berguna untuk menemukan pola dalam proses manufaktur yang kompleks. Data mining dapat digunakan dalam perancangan tingkat sistem untuk mengekstrak hubungan antara arsitektur produk, portofolio produk, dan data kebutuhan pelanggan. Data mining juga bisa digunakan untuk memprediksi perkembangan produk seperti span time, cost, dan dependencies.

4. **Mengapa perlu dilakukan pra-proses data? Jelaskan alasan anda dan berikan contoh.**

Jawab:

Menurut saya, data pre-processing wajib dan sangat penting dilakukan sebelum mengolah data. Pertama-tama, kita definisikan dahulu apa itu data pre-processing. Pre-processing adalah proses mengubah data mentah menjadi data dengan format yang lebih dapat dipahami. Teknik ini adalah teknik data mining untuk mengubah data mentah yang dikumpulkan dari berbagai sumber menjadi informasi yang lebih bersih dan lebih cocok untuk diproses lebih lanjut. Kenapa pre-processing data wajib dilakukan? Karena pada pemrosesan data, kita bisa mejumpai banyak kesalahan, redundansi, missing values dan inkonsistensi yang mempengaruhi integritas sebuah data set, dan kita harus memperbaiki semua masalah itu untuk hasil akhir yang lebih akurat. Jadi, jika diambil kesimpulan dari penjelasan saya tentang data definisi pre-processing di atas, kita bisa menyimpulkan bahwa sebelum menggunakan data untuk tujuan yang kita inginkan, kita harus mengatur kumpulan data tersebut supaya bisa mengatasi kesalahan, redundansi, missing values, dll supaya data menjadi sebersih mungkin dengan data pre-processing.

Contoh data pre-processing:

- **Pembersihan data** adalah mengisi nilai-nilai yang hilang, menghaluskan noisy data, mengenali atau menghilangkan outlier, dan memecahkan ketak-konsistenan.
- **Integrasi data** adalah proses menggabungkan beberapa sumber menjadi satu kumpulan data. Seperti Integrasi banyak database, banyak kubus data, atau banyak file
- **Transformasi data** adalah perubahan yang dibuat dalam format atau struktur data
- **Reduksi data** adalah mendapatkan representasi yang direduksi dalam volume tetapi menghasilkan hasil analitikal yang sama atau mirip
- **Diskritisasi data** adalah bagian dari reduksi data tetapi dengan kepentingan khusus, terutama data numerik

5. **Berikan contoh implementasi klasifikasi (selain yang ada di slide materi) dan jelaskan.**

Jawab:

Saya menggunakan sebuah dataset yang strukturnya mirip dengan iris dataset, yaitu penguin dataset, atau lengkapnya bisa disebut dengan “Palmer Archipelago (Antarctica) penguin dataset”

	species	island	culmen_length	culmen_depth	flipper_length	body_mass	sex
1	Adelie	Torgersen	39.1	18.7	181	3750	MALE
2	Adelie	Torgersen	39.5	17.4	186	3800	FEMALE
3	Adelie	Torgersen	40.3	18	195	3250	FEMALE
4	Adelie	Torgersen	NaN	NaN	NaN	NaN	NaN
5	Adelie	Torgersen	36.7	19.3	193	3450	FEMALE
154	Chinstrap	Dream	46.5	17.9	192	3500	FEMALE
155	Chinstrap	Dream	50	19.5	196	3900	MALE
156	Chinstrap	Dream	51.3	19.2	193	3650	MALE
157	Chinstrap	Dream	45.4	18.7	188	3525	FEMALE

158	Chinstrap	Dream	52.7	19.8	197	3725	MALE
222	Gentoo	Biscoe	46.1	13.2	211	4500	FEMALE
223	Gentoo	Biscoe	50	16.3	230	5700	MALE
224	Gentoo	Biscoe	48.7	14.1	210	4450	FEMALE
226	Gentoo	Biscoe	47.6	14.5	215	5400	MALE
227	Gentoo	Biscoe	46.5	13.5	210	4550	MALE

Dataset ini cukup unik karena mempunyai beberapa class/label sehingga dataset ini bisa kita sebut dengan multi-class dataset. Class atau label adalah atribut diskrit yang nilainya ingin kita prediksi berdasarkan nilai atribut lainnya. Dan menurut definisi tersebut, ada beberapa class yang bisa kita prediksi menggunakan atribut/fitur yang ada dalam tabel tersebut. Sehingga, bisa disimpulkan bahwa:

Class: Species, island, sex
Attribute: culmen_length, culmen_depth, flipper_length, body_mass

Keterangan class dan atribut:

- **species:** penguin species (Adelie, Chinstrap, Gentoo)
- **island:** pulau
- **culmen_length:** panjang paruh atas (mm)
- **culmen_depth:** tinggi/kedalaman paruh (mm)
- **flipper_length:** panjang sayap
- **sex:** jenis kelamin
- **body_mass:** berat badan (gram)

6. Perhatikan data berikut ini :

Temperature udara (^{0}C)	Kecepatan angin (km/jam)	klasifikasi
10	0	Dingin
25	0	Panas
15	5	Dingin
20	3	Panas
18	7	Dingin
20	10	Dingin
22	5	Panas
24	6	Panas

Terdapat data klasifikasi suhu udara yang terdiri dari 2 kelas yaitu Panas dan Dingin. Kedua kelas ini dapat diukur berdasarkan 2 variabel yaitu temperature dalam derajat Celsius dan kecepatan angin dalam km/h. Bagaimana hasil klasifikasi saat temperature udara 16 C dan kecepatan angin 3km/jam, jika nilai k=3? Tampilkan hasil perhitungan anda dalam tabel (seperti contoh).

Jawab:

Pertama, tentukan parameter K yaitu K = 3

Kedua, hitung jarak antara data baru dengan semua data training. Kita menggunakan Euclidean Distance. Gunakan X untuk temperatur udara, dan Y untuk kecepatan angin.

X	Y	Euclidean Distance (16, 3)
10	0	$\sqrt{(10-16)^2} + \sqrt{(0-3)^2} = \sqrt{(-6)^2 + (-3)^2} = \sqrt{45} = 6.7$
25	0	$\sqrt{(25-16)^2} + \sqrt{(0-3)^2} = \sqrt{(9)^2 + (-3)^2} = \sqrt{90} = 9.4$
15	5	$\sqrt{(15-16)^2} + \sqrt{(5-3)^2} = \sqrt{(-1)^2 + (2)^2} = \sqrt{5} = 2.2$
20	3	$\sqrt{(20-16)^2} + \sqrt{(3-3)^2} = \sqrt{(4)^2 + (0)^2} = \sqrt{16} = 4$
18	7	$\sqrt{(18-16)^2} + \sqrt{(7-3)^2} = \sqrt{(2)^2 + (4)^2} = \sqrt{20} = 4.4$
20	10	$\sqrt{(20-16)^2} + \sqrt{(10-3)^2} = \sqrt{(4)^2 + (7)^2} = \sqrt{65} = 8.1$
22	5	$\sqrt{(22-16)^2} + \sqrt{(5-3)^2} = \sqrt{(6)^2 + (2)^2} = \sqrt{40} = 6.3$
24	6	$\sqrt{(24-16)^2} + \sqrt{(6-3)^2} = \sqrt{(8)^2 + (3)^2} = \sqrt{73} = 8.5$

Ketiga, urutkan jarak dari data baru dengan data training dan menentukan tetangga terdekat berdasarkan jarak minimum K.

X	Y	Euclidean Distance (16, 3)	Urutan jarak	Termasuk 3-NN ?
10	0	6.7	5	Tidak (K>3)
25	0	9.4	8	Tidak (K>3)
15	5	2.2	1	Ya (K<3)
20	3	4	2	Tidak (K>3)
18	7	4.4	3	Tidak (K>3)
20	10	8.1	6	Tidak (K>3)
22	5	6.3	4	Tidak (K>3)
24	6	8.5	7	Tidak (K>3)

Keempat, tentukan kategori dari tetangga terdekat.

X	Y	Euclidean Distance (16, 3)	Urutan Jarak	Termasuk 3-NN ?	Kategori KKN
10	0	6.7	5	Tidak (K>3)	-
25	0	9.4	8	Tidak (K>3)	-
15	5	2.2	1	Ya (K<3)	Dingin
20	3	4	2	Tidak (K>3)	-
18	7	4.4	3	Tidak (K>3)	-
20	10	8.1	6	Tidak (K>3)	-
22	5	6.3	4	Tidak (K>3)	-
24	6	8.5	7	Tidak (K>3)	-

Kelima, gunakan kategori mayoritas yang sederhana dari tetangga yang terdekat tersebut sebagai nilai prediksi data yang baru. Kita ubah kembali X menjadi temperatur udara dan Y menjadi kecepatan angin.

Temperature udara (^{0}C)	**Kecepatan angin (km/jam)**	**Klasifikasi**
10	0	Dingin
25	0	Panas
15	5	Dingin
20	3	Panas
18	7	Dingin
20	10	Dingin
22	5	Panas
24	6	Panas
10	0	Dingin
16	**3**	**Dingin**

Data yang kita miliki adalah pada baris 3 yaitu Dingin. Dan tidak ada baris lain yang memiliki kategori Dingin maupun Panas karena tidak termasuk 3-NN. Sehingga tinggal menyisakan baris 3 (temperature 15 dan kecepatan angin 5) yang memiliki klasifikasi dingin. Sehingga bisa kita simpulkan bahwa klasifikasi data baru yaitu X=16 dan Y=3 adalah **Dingin**.

Terima Kasih