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

Kuis 1

1. Jelaskan apa yang dimaksud dengan data mining dan fungsinya.

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.

Secara garis besar, data mining memiliki dua fungsi utama yaitu predictive dan descriptive.
Berikut penjelasannya:

1) Fungsi predictive/prediksi.
   Menggunakan beberapa variabel untuk memprediksi nilai yang belum diketahui (unknown) atau nilai selanjutnya (future) dari variable lain. Dengan kata lain, Fungsi prediksi merupakan sebuah fungsi bagaimana sebuah proses nantinya akan menemukan pola tertentu dari suatu data. Pola-pola tersebut dapat diketahui dari berbagai variabel-variabel yang ada pada data. Ketika sudah menemukan pola, maka pola tersebut bisa digunakan untuk memprediksi variabel lain yang belum diketahui nilai maupun jenisnya.

   Berikut beberapa contoh fungsi prediksi:
   - **Classification** = Menemukan fungsi atau model yang membedakan kelas data. Fungsi tersebut dapat berbentuk aturan if-else, descision tree, formula matematika, atau neural network. Tujuannya untuk memperkirakan kelas dari suatu objek yang labelnya tidak diketahui. Contoh: Deteksi pemakaian kartu kredit secara ilegal
   - **Regression =** menentukan hubungan sebab-akibat antara satu variabel dengan variabel yang lain dengan mengklasifikasikan target data numerik. Contoh = prediksi nilai penjualan di masa mendatang berdasarkan trend penjualan di tahun-tahun sebelumnya.
   - **Time Series** = sekuens data yang nilainya berubah setiap interval waktu tertentu. Dapat dipresentasikan dala bentuk grafik atau kurva yang menunjukkan fungsi variabel data terhadap satuan waktu. Contoh = Prediksi dalam pasar saham.
   - **Deviation Detection** = Sebuah teknik untuk mengidentifikasi outlier yang mengekspresikan sebuah deviasi dari ekspektasi yang sudah diketahui sebelumnya

2) Fungsi Descriptive/Deskripsi
   Menemukan pola pendeskripsian data yang dapat diinterpretasikan oleh manusia. Dengan kata lain, fungsi deskripsi dalam data mining adalah sebuah fungsi untuk memahami lebih jauh tentang data yang diamati. Dengan melakukan sebuah proses diharap bisa mengetahui perilaku dari sebuah data tersebut. Data tersebut itulah yang nantinya dapat digunakan untuk mengetahui karakteristik dari data yang dimaksud.

Berikut beberapa contoh fungsi deskripsi:

- **Clustering** = Mengidentifikasi kelompok alami dari data berdasarkan kemiripan atribut. Fungsi ini disebut juga dengan Segmentation. Contoh: web-document clustering.
- **Association Rule** = Disebut juga Market Base Analysis. Fungsi ini menganalisa tabel transaksi penjualan dan mengidentifikasi produksi produk yang seringkali dibeli bersamaan oleh customer. Contoh: Pengelolaan rak di supermarket.
- **Sequence Analysis** = Digunakan untuk mencari pola pada serangkaian kejadian yang disebut dengan Sequence. Contoh: rangkaian klik pada sebuah website berita.

2. Menurut anda, apakah pre-processing data harus (wajib) dilakukan sebelum mengolah data? Jelaskan.

Menurut saya, data pre-processing wajib dan sangat penting dilakukan sebelum mengolah data. Pertama-tama, mari 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.

Lalu 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.

3. Bagaimana menentukan kelayakan suatu dataset untuk dapat diproses ke tahap selanjutnya?

Pada dasarmya kualitas atau kelayakan data dapat diukur dari akurasi, kelengkapan, konsistensi, ketepatan, kepercayaan, nilai tambah, penafsiran, dan kemudahan akses. Untuk mendapatkan seluruh kualitas tersebut, kita bisa melakukan pembersihan data, integrasi data, transformasi data, reduksi data, dan diskritisasi data.

- **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

4. Apa yang anda ketahui tentang klasifikasi dan bagaimana mengukur kinerja sebuah metode klasifikasi? Jelaskan.

Klasifikasi adalah suatu proses pengelompokan data didasarkan pada ciri-ciri tertentu ke dalam kelas-kelas yang telah ditentukan. Klasifikasi juga dapat didefinisikan sebagai suatu pekerjaan menilai objek data untuk memasukkannya ke dalam kelas tertentu dari sejumlah kelas yang tersedia.

Klasifikasi memiliki 2 pekerjaan utama yaitu:

1) Pembangunan model sebagai prototipe untuk disimpan sebagai memori
2) Penggunaan model tersebut untuk melakukan pengenalan/klasifikasi/prediksi pada suatu objek data lain agar diketahui di kelas mana objek data tersebut

Model klasifikasi dapat dibangun berdasarkan pengetahuan seorang pakar. Namun karena himpunan data yang biasanya sangat besar, model klasifikasi lebih sering dibangun menggunakan Machine Learning. Machine Learning secara otomatis terhadap suatu himpunan data mampu menghasilkan model klasifikasi yang memetakan objek data x (input) ke salah satu kelas y yang telah didefinisikan sebelumnya. Model dalam klasifikasi mempunyai arti yang sama dengan kotak hitam, yaitu ada suatu input, berupa himpunan training set, yang berlabel (memiliki atribut kelas), kemudian mampu melakukan pemikiran terhadap masukan tersebut dan memberikan jawaban sebagai output yang berupa sebuah model klasifikasi.

Untuk mengukur kinerja sebuah metode klasifikasi, kita bisa gunakan **confusion matrix**. Confusion matrix adalah pengukuran performa untuk masalah klasifikasi machine learning dimana keluaran dapat berupa dua kelas atau lebih. Confusion Matrix adalah tabel dengan 4 kombinasi berbeda dari nilai aktual dan nilai prediksi. Ada empat istilah yang merupakan representasi hasil proses klasifikasi pada confusion matrix yaitu True Positive (TP), True Negative(TN), False Positive(FP), dan False Negative(FN).

Berikut tampilan tabel confusion matrix secara sederhana:

	KELAS PREDIKSI		
KELAS AKTUAL	KELAS	True	False
	True	TP	FP
	False	FN	TN

Dari tabel tersebut, ada 4 macam kejadian yang bisa terjadi yaitu:
TP (True Positive) = Jumlah data aktual yang sebenarnya **True** diprediksi **True**
TN (True Negative) = Jumlah data aktual yang sebenarnya **False** diprediksi **False**
FP (False Positive) = Jumlah data aktual yang sebenarnya **True** diprediksi **False**
FN (False Negative) = Jumlah data aktual yang sebenarnya **Fasle** diprediksi **True**

Berikut beberapa perhitungan yang bisa dilakukan dalam confusion matrix

- *Accuracy* menggambarkan seberapa akurat model dalam mengklasifikasikan dengan benar. *Accuracy* = (TP+TN) / (TP+FP+FN+TN)
- *Precision* menggambarkan akurasi antara data yang diminta dengan hasil prediksi yang diberikan oleh model. *Precision* = (TP) / (TP + FP)
- *Recall* atau sensitivity menggambarkan keberhasilan model dalam menemukan kembali sebuah informasi. *Recall* = TP / (TP + FN)

- *F-1 Score* menggambarkan perbandingan rata-rata precision dan recall yang dibobotkan. *F-1 Score* = (2 * Recall * Precision) / (Recall + Precision)
- *Specificity* menggambarkan kebenaran memprediksi negatif dibandingkan dengan keseluruhan data negatif. *Specificity = (TN)/ (TN + FP)*
- *Gmean* = akar(recall * spesifisitas)

5. Berikan sebuah contoh kasus klasifikasi dan sebutkan atribut/fitur dan kelasnya (contoh selain yang ada di materi).

Kasus:

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

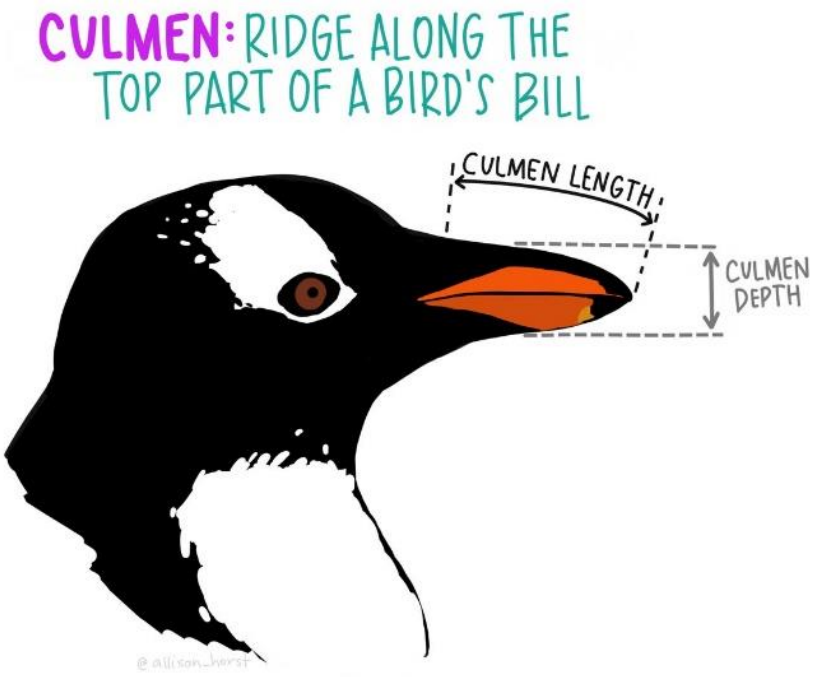
Dataset ini cukup unik karena mempunyai beberapa class/label sehingga dataset ini bisa kita sebut dengan multi-class dataset. Class atau label adalah atribut diskrit yang nilainya ingin kita prediksi berdasarkan nilai atribut lainnya. Dan menurut definisi tersebut, ada beberapa class yang bisa kita prediksi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

untuk keterangan lebih jelasnya, bisa dilihat di gambar di bawah ini:



Terima Kasih