UNIVERSITAS INDONESIA
FAKULTAS ILMU KOMPUTER
Veritas, Probitas, Iustitia

CSGE603130 – Kecerdasan Artifisial dan Sains Data Dasar
UTS Semester Gasal 2022/2023
24-Oktober-2022, 09:00-11:30 WIB

Pengajar:
Dr. Fariz Darari,
Aruni Yasmin Azizah, M.Comp.Sc.,
Siti Aminah, M.Kom.,
Dr. Dina Chahyati,
Adila Alfa Krisnadhi, Ph.D.

PERNYATAAN KESANGGUPAN MENTAATI TATA TERTIB UJIAN

"Saya telah membaca dan memahami ketentuan tata tertib berikut ini, serta menyatakan bahwa jawaban ujian ini adalah hasil pekerjaan saya sendiri. Saya menyetujui jika melakukan pelanggaran atas ketentuan tersebut, saya bersedia diproses sesuai ketentuan yang berlaku (SK Dekan 103a Tahun 2020) dengan sanksi maksimal **nilai akhir E**."

Nama & Tanda-tangan:	Kelas:	Nomor Pokok Mahasiswa:

TATA TERTIB UJIAN

- Semua alat komunikasi elektronik dalam kondisi non-aktif (dimatikan), dimasukkan ke dalam tas dan diletakkan pada tempat yang telah disediakan.
- Peralatan ujian yang boleh dibawa adalah alat tulis dan yang diperbolehkan sesuai sifat ujian.
- Peserta ujian menempati tempat duduk yang telah ditentukan.
- Peserta ujian menuliskan nama dan NPM pada setiap lembar jawaban ujian.
- Peserta mulai membuka soal dan mengerjakan ketika pengawas mengatakan ujian dimulai dan berhenti bekerja (meletakkan alat tulis) ketika pengawas mengatakan waktu habis.
- Peserta tidak berkomunikasi dalam bentuk apa pun dengan peserta lain selama berada di ruang ujian, termasuk pinjam meminjam alat tulis, serta tidak memberi atau menerima bantuan dari siapapun selama ujian berlangsung.
- Peserta yang meninggalkan ruang ujian dianggap selesai mengerjakan. Jika karena kondisi medis khusus tidak bisa memenuhi ketentuan ini, peserta wajib melaporkan kepada pengawas sebelum ujian dimulai.
- Setelah selesai mengerjakan atau setelah waktu habis, peserta segera meninggalkan berkas soal dan lembar jawaban ujian di meja masing-masing, mengambil tas dan segera keluar tanpa mengganggu peserta lain serta tanpa berkomunikasi dengan peserta lain.
- Jawaban ujian ini tidak akan dinilai jika pernyataan di atas ini tidak ditandatangani.

Informasi Tambahan

- Diperbolehkan membawa catatan (*notes*) sebanyak 8 halaman A4.
- Diperbolehkan menggunakan kalkulator

NPM: Kelas: A | B | C | D | E | F (lingkari salah satu)

Nama:

Ujian Tengah Semester - Kecerdasan Artifisial dan Sains Data Dasar
Semster Gasal 2022/2023 - Senin, 24 Oktober 2022 - Waktu: 150 menit

A. Pilihan Ganda (Total bobot: 30×2 poin $= 60$ poin)

Pilihlah satu dari empat opsi yang tersedia, dengan menyilangi (X) pilihan yang Anda anggap benar. Salinlah semua jawaban Soal PG ke tabel di bawah ini. **Hanya jawaban di tabel ini yang akan dinilai.**

1.	A	B	C	D		16.	A	B	C	D
2.	A	B	C	D		17.	A	B	C	D
3.	A	B	C	D		18.	A	B	C	D
4.	A	B	C	D		19.	A	B	C	D
5.	A	B	C	D		20.	A	B	C	D
6.	A	B	C	D		21.	A	B	C	D
7.	A	B	C	D		22.	A	B	C	D
8.*	A	B	C	D		23.	A	B	C	D
9.	A	B	C	D		24.	A	B	C	D
10.	A	B	C	D		25.	A	B	C	D
11.	A	B	C	D		26.	B	S		
12.	A	B	C	D		27.	B	S		
13.	A	B	C	D		28.	B	S		
14.	A	B	C	D		29.	B	S		
15.	A	B	C	D		30.	B	S		

NPM: Kelas: A | B | C | D | E | F (lingkari salah satu)

Nama:

Soal 1 (2 poin)

Dari beberapa perspektif definisi kecerdasan artifisial, manakah di antara contoh-contoh di bawah ini yang **TIDAK** menunjukkan munculnya sifat cerdas pada mesin?

a. menurut perspektif *thinking humanly*, mesin mampu mengemulasi proses berpikir seorang mahasiswa ketika harus memilih antara belajar di perpustakaan atau bermain game di rumah teman.

b. Menurut perspektif *thinking rationally*, mesin mampu menentukan bahwa cuaca di luar rumah sedang hujan ketika mendapatkan sinyal audio dari atas genting rumah.

c. Menurut perspektif *acting humanly*, mesin mampu berkomunikasi dengan manusia dengan bahasa manusia.

d. Menurut perspektif *acting rationally*, mesin mampu mencegah pintu lift menutup lampu di koridor gedung Fasilkom yang gelap ketika ada orang yang hendak melalui koridor tersebut.

Soal 2 (2 poin)

Definisi kecerdasan artifisial menurut perspektif *acting rationally* bermuara pada pendekatan agen rasional. Manakah pernyataan terkait pendekatan agen rasional yang **TIDAK TEPAT** di antara beberapa pernyataan di bawah ini?

a. Agen rasional yang dapat memilih langkah-langkah terbaik untuk mencapai tujuannya adalah utility-based agent.

b. Agen dapat diwujudkan sebagai sebuah program yang menghasilkan suatu pilihan tindakan terbaik yang dapat dilakukan actuator dengan memperhatikan sederet persepsi yang diterima melalui sensor.

c. Prinsip rasionalitas agen direalisasikan dalam bentuk ukuran kinerja pada keadaan internal agen.

d. Salah satu cara yang berpotensi memperbaiki kinerja agen yang bekerja pada lingkungan yang bersifat partially observable adalah dengan menambah tipe dan jumlah sensor yang dimiliki agen tersebut.

Soal 3 (2 poin)

Manakah di antara pernyataan-pernyataan yang terkait *learning agent* yang menggunakan paradigma pemelajaran mesin berikut yang **TIDAK TEPAT**?

a. Manusialah yang menentukan struktur dasar model agen melalui formulasi representasi model.

b. Algoritma pelatihan model akan menghasilkan output prediksi jawaban dari permasalahan yang dihadapi agen.

c. Proses belajarnya agen diwujudkan melalui pencarian model yang tepat dengan memanfaatkan fungsi evaluasi.

d. Data yang dipakai untuk melatih model agen melalui prosedur optimisasi merupakan sampel dari ruang permasalahan.

NPM: Kelas: A | B | C | D | E | F (lingkari salah satu)

Nama:

Soal 4 (2 poin)
Manakah contoh yang benar untuk problem pemelajaran mesin yang diberikan?

a. Menentukan derajat berbeloknya kemudi pada mobil otonom merupakan problem regresi.

b. Mendeteksi rambu lalu lintas di pinggir jalan pada mobil otonom merupakan problem clustering.

c. Mengelompokkan kendaraan-kendaraan yang melewati pintu tol Cawang berdasarkan tingkat emisi CO2 yang dihasilkannya merupakan problem regresi.

d. Menentukan tingkat akselerasi/percepatan atau deselerasi/perlambatan yang harus dilakukan oleh sebuah mobil otonom sesuai laju kendaraan di depannya merupakan problem klasifikasi.

Soal 5 (2 poin)
Dalam metodologi sains data CRISP-DM, apa yang menyebabkan seorang *data scientist* harus mengulang kembali fase *business understanding* setelah menyelesaikan setelah fase *data understanding*?

a. Hasil EDA menunjukkan indikasi bahwa, walaupun sejumlah data sudah tersedia, jumlahnya tidak memadai untuk pengembangan model analitik yang diinginkan.

b. Upaya pengumpulan data yang sesuai dengan permasalahan bisnis ternyata tidak berhasil dilakukan karena data yang diinginkan sama sekali tidak dapat diakses.

c. Evaluasi model menunjukkan bahwa kinerja model analitik kurang baik karena kurangnya data.

d. Kinerja model setelah di-*deploy* di lapangan kurang memuaskan meskipun evaluasi pada data uji sudah menunjukkan hasil yang baik.

Soal 6 (2 poin)
Diberikan dataset berikut.

A	B	C
5	20	7
15	5	8
20	20	9
20	25	10
10	15	?

Jika dilakukan imputasi data untuk mengisi baris terakhir pada kolom C berdasarkan metode 1-Nearest Neighbors dengan ukuran jarak Euclidean, berapa hasilnya?

a. 7 b. 8 c. 9 d. 10

NPM: Kelas: A | B | C | D | E | F (lingkari salah satu)

Nama:

Soal 7 (2 poin)
Manakah pernyataan yang tepat terkait *pre-processing*?

a. *Binning* hanya dapat diterapkan pada data numerikal, tidak dapat diterapkan pada data kategorikal

b. *Feature extraction* maupun *feature selection* dapat berdampak mengurangi banyaknya kolom dari suatu dataset.

c. Jika suatu kolom mengandung *outlier*, hasil transformasi dengan *Min-Max scaling* akan menghilangkan *outlier*-nya.

d. Pada proyek sains data, tahapan *pre-processing* umumnya memerlukan waktu yang lebih sedikit dibandingkan tahapan *modeling*.

Soal 8 (2 poin)
Mana saja langkah *pre-processing* yang penting dilakukan ketika seseorang melakukan klasifikasi dengan KNN namun tidak harus dilakukan pada klasifikasi dengan Decision Tree **(bisa lebih dari satu jawaban; PILIH SEMUA yang benar)**:

a. Menangani *outlier*

b. Memeriksa konsistensi format penulisan setiap data

c. Membuang fitur yang memiliki korelasi rendah terhadap fitur-fitur lainnya

d. Melakukan normalisasi (*feature scaling*) pada atribut-atribut numerik dengan nilai *range* yang berbeda

Soal 9 (2 poin)
Anda akan melakukan regresi untuk memprediksi nilai inflasi di suatu kota berdasarkan 10 variabel (fitur) yang terkait ekonomi dan keuangan. Kegiatan *pre-processing* berikut perlu untuk Anda lakukan, **KECUALI**:

a. Memeriksa apakah ada nilai Null pada dataset

b. Memeriksa apakah ada *outlier* dan *noise* pada dataset

c. Memeriksa apakah dataset *balance* atau tidak

d. Memeriksa apakah ada variabel yang merupakan turunan dari variabel lainnya (berkorelasi sangat tinggi) dalam dataset tersebut.

NPM: Kelas: A | B | C | D | E | F (lingkari salah satu)

Nama:

Soal 10 (2 poin)
Manakah di bawah ini yang **TIDAK TEPAT** sebagai manfaat dari EDA:

a. Mengetahui apakah data mengandung *outlier*.

b. Mengubah nilai data ke dalam skala 0–1.

c. Mengetahui bentuk persebaran data.

d. Memeriksa nilai hilang (*missing value*) pada data.

Soal 11 (2 poin)
Perhatikan dua pernyataan berikut. Asumsi: setiap dataset terdiri dari satu kolom numerik.

(1) Jika Dataset A memiliki nilai mean yang lebih tinggi daripada Dataset B, dapat dipastikan Dataset A juga memiliki nilai median yang lebih tinggi daripada Dataset B.
(2) Jika Dataset A memiliki nilai maksimum yang lebih tinggi daripada Dataset B, dapat dipastikan Dataset A juga memiliki nilai median yang lebih tinggi daripada Dataset B.

Jawablah sesuai dengan benar/salahnya pernyataan-pernyataan di atas:

a. Pernyataan 1 salah, Pernyataan 2 benar.

b. Pernyataan 1 benar, Pernyataan 2 salah.

c. Pernyataan 1 dan 2 salah.

d. Pernyataan 1 dan 2 benar.

Soal 12 (2 poin)
Perhatikan data berikut: $[5, 6, 3, 1, 4, 2]$. Berapakah nilai IQR-nya?

a. 2 b. 3 c. 5 d. -2

Soal 13 (2 poin)
Perhatikan data berikut: $[100000, 50000, 4000, 300, 20, 1]$. Mana sajakah nilai *outlier*-nya?

a. 1 & 100000 b. 1 c. 100000 d. Tidak ada

Soal 14 (2 poin)
Manakah pernyataan yang **TIDAK TEPAT** di bawah ini mengenai IQR dan *outlier* jika diasumsikan datasetnya hanya memiliki satu kolom numerik?

a. Suatu dataset dapat memiliki beberapa nilai outlier yang berbeda.

b. Outlier pada data dapat disebabkan oleh human error saat penginputan data.

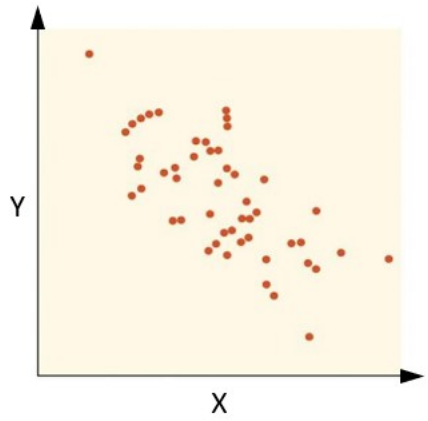
c. Suatu dataset dapat memiliki beberapa nilai IQR yang berbeda.

d. Nilai range pada suatu dataset dapat bernilai sama dengan IQR.

NPM: Kelas: A | B | C | D | E | F (lingkari salah satu)

Nama:

Soal 15 (2 poin)
Perhatikan *scatterplot* berikut.



Berapakah nilai Pearson's correlation coefficient yang paling memungkinkan antara variabel X dan Y?

a. 0 b. -0.99 c. 0.99 d. -0.7

Soal 16 (2 poin)
Manakah pernyataan yang **TIDAK TEPAT** terkait visualisasi data?

a. Scatter plot perlu dibuat terlebih dulu sebelum menghitung korelasi linier dengan Pearson r.

b. Line chart digunakan untuk menunjukkan tren variabel kuantitatif dari waktu ke waktu.

c. Box plot dapat digunakan untuk menunjukkan keberadaan *outlier*.

d. Pie chart dapat digunakan untuk melihat *relationship* antara 2 variabel kuantitatif.

Soal 17 (2 poin)
Pernyataan mana yang benar terkait visualisasi dengan Histogram?

a. Jika ukuran *bin* terlalu besar, ada pola yang tidak tertangkap

b. Semakin besar ukuran bin, semakin banyak jumlah *bin* yang diperlukan untuk merepresentasikan seluruh jangkauan data

c. Permasalahan ukuran *bin* yang tidak cocok dapat diatasi dengan mengunakan Bar chart

d. Histogram cocok digunakan untuk memvisualisasikan akurasi model KNN pada proses klasifikasi dengan berbagai nilai k.

NPM: Kelas: A | B | C | D | E | F (lingkari salah satu)

Nama:

Soal 18 (2 poin)
Manakah pernyataan yang **TIDAK TEPAT** tentang PCA?

a. Vektor-vektor yang membentuk *principle component* saling ortogonal

b. Matriks yang didiagonalisasi pada PCA adalah matriks input yang berukuran $n \times m$ dimana n adalah jumlah data dan m adalah jumlah fitur

c. *Principle component* pertama merupakan vektor eigen yang berkaitan dengan nilai eigen tertinggi

d. PCA melakukan transformasi data dari suatu ruang ke ruang lainnya yang memiliki basis berbeda

Soal 19 (2 poin)
Misalkan principle component suatu dataset adalah $(\sqrt{2}/2, \sqrt{2}/2)$ dan $(\sqrt{2}/2, -\sqrt{2}/2)$. Jika suatu data setelah di-*adjust* berada pada koordinat $(1, 1)$, maka data koordinat barunya setelah ditransformasi PCA adalah:

a. $(\sqrt{2}, 0)$ b. $(\sqrt{2}, -\sqrt{2})$ c. $(\sqrt{2}/2, -\sqrt{2}/2)$ d. $(0, \sqrt{2})$

Soal 20 (2 poin)
Jika $a(x)$ menyatakan *intra-cluster distance* dari data x, $b(x)$ menyatakan *inter-cluster distance* terdekat dari x, dan S menyatakan nilai Silhouette dari suatu *clustering*, maka *clustering* yang baik seharusnya memenuhi kondisi berikut:

a. $b(x) > a(x)$ untuk sebagian besar x

b. $b(x) < 0$ untuk sebagian besar x

c. $S < 0.25$

d. $a(x) > 1$ untuk sebagian besar x

Soal 21 (2 poin)
Perbedaan *k-means clustering* dan *hierarchical clustering* adalah:

a. *k-means supervised, hierarchical unsupervised*

b. Jumlah *cluster* pada *k-means* harus ditentukan di awal, sedangkan pada *hierarchical* dapat ditentukan di akhir tergantung pada nilai *threshold* ketinggian dendogram yang dipilih

c. *K-means* sensitif terhadap *outlier*, sedangkan *hierarchical* tidak

d. Jarak pada *k-means* bisa menggunakan jarak Manhattan, sedangkan pada *hierarchical* harus menggunakan jarak Euclid

NPM: Kelas: A | B | C | D | E | F (lingkari salah satu)

Nama:

Soal 22 (2 poin)

Pernyataan berikut yang **paling tepat** terkait *random forest* adalah:

a. Setiap *tree* dilatih dengan dataset yang diperoleh dengan *sampling* secara acak tanpa pengembalian

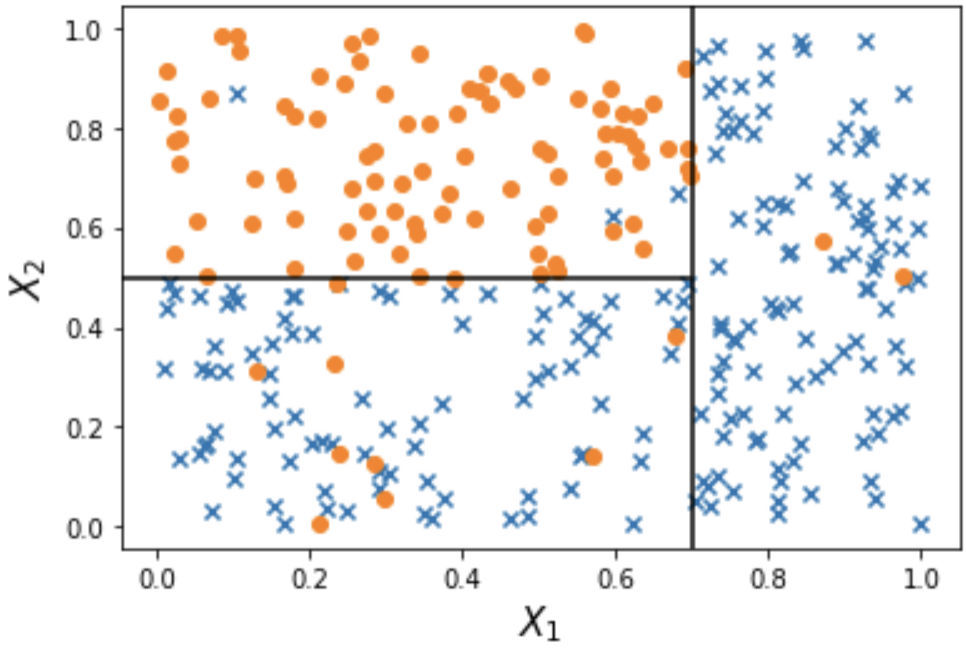
b. *Random forest* digunakan untuk menurunkan *training error* dibandingkan dengan *decision tree*

c. Hanya digunakan untuk klasifikasi yang mana prediksi dilakukan dengan *majority vote*

d. Data tidak perlu dinormalisasi/distandarisasi

Soal 23 (2 poin)

Perhatikan *decision boundary* di bawah ini yang dihasilkan dari sebuah *decision tree*. Tanda "x" dan "o" di-plot tersebut merepresentasikan semua data yang digunakan saat *training*.



Pernyataan berikut yang **KURANG TEPAT** tentang *decision tree* yang dihasilkan adalah:

a. Data dengan $X_1 = 0.5$ dan $X_2 = 0.4$ termasuk kelas "o"

b. Kedalaman *tree* adalah 2

c. Tidak ada *leaf* yang *pure*

d. Atribut pada *root* yang ditanyakan nilainya adalah X_1 dengan threshold 0.7

NPM: Kelas: A | B | C | D | E | F (lingkari salah satu)

Nama:

Soal 24 (2 poin)
Pernyataan berikut yang **paling tepat** tentang evaluasi sebuah model prediksi adalah:

a. Proporsi data yang digunakan untuk *training* dan *testing* secara berturut-turut adalah 20% dan 80%

b. Nilai k pada *k-fold cross validation* menentukan berapa kali skema *training-validation* dilakukan

c. Setiap data pada metode *random subsampling* setidaknya pernah digunakan pada tahap *validation*

d. Metode *hold-out* dapat memperoleh eror generalisasi yang paling akurat dibandingkan dengan metode evaluasi model lainnya

Soal 25 (2 poin)
Berikut ini diberikan confusion matrix sebuah model klasifikasi.

		Predicted: Negative	Predicted: Positive
Actual	Negative	50	10
	Positive	5	100

Pernyataan yang **KURANG TEPAT** mengenai kinerja model tersebut adalah:

a. Dataset tidak seimbang di mana contoh kelas positif lebih banyak

b. Nilai akurasi sama dengan nilai *precision*

c. F1-Score = 0.93

d. Nilai *specificity* sama dengan *recall*

Soal 26 (2 poin)
(B/S) Similaritas Jaccard antara $\{a, b, c, d\}$ dan $\{b, d, e, f\}$ adalah 0.25.

Soal 27 (2 poin)
(B/S) Cara mengklasifikasi suatu data baru menggunakan KNN adalah dengan mencari jarak terdekat ke pusat-pusat (*centroid*) kelas.

Soal 28 (2 poin)
(B/S) Budi hendak melakukan klasifikasi terhadap suatu dataset yang terdiri dari satu juta baris data, dimana 75% diantaranya digunakan pada tahap pelatihan. Budi menggunakan metode CART dan KNN. *Running time* tahap pengujian kedua metode tersebut seharusnya sama saja.

NPM: Kelas: A | B | C | D | E | F (lingkari salah satu)

Nama:

Soal 29 (2 poin)

(B/S) Yang membedakan antara *random forest* dengan teknik *bagging* yang hanya terdiri dari *decision trees* adalah setiap *tree* harus dibangun sampai kedalaman tertentu.

Soal 30 (2 poin)

(B/S) Lain halnya dengan MAE, MSE, dan RMSE, R^2 menghitung proporsi *variance* data yang dapat dijelaskan melalui regresi yang dilakukan sehingga semakin tinggi nilai R^2, maka hasil regresi cenderung menunjukkan prediksi yang lebih baik.

NPM: Kelas: A | B | C | D | E | F (lingkari salah satu)

Nama:

B. Uraian (Total bobot: 5×8 poin $= 40$ poin)

Soal 31 (*Principal Component Analysis* – 8 poin)

a. Jelaskan bagaimana proses reduksi dimensi pada PCA

b. Pada Lab 3 yang lalu, *explained variance* yang didapatkan untuk dataset Iris yang awalnya terdiri dari 4 dimensi adalah $[0.72962445, 0.22850762, 0.03668922, 0.00517871]$. Jika kita hanya mengambil 2 *principle component*, apakah dapat dikatakan bahwa dimensi dataset tersebut direduksi menjadi $(0.72962445 + 0.22850762) \times 100\%$ dari dimensi awalnya? Jelaskan jawaban Anda.

Jawaban:

NPM: Kelas: A | B | C | D | E | F (lingkari salah satu)

Nama:

Soal 32 (*Clustering* – 8 poin)
Diberikan dua pertanyaan yang tidak terkait satu sama lain:

a. Anda melakukan *agglomerative clustering* dari 5 poin:

A (1, 2), B (5, 3), C (5, 7), D (6, 3), E (4, 1).

Dengan menggunakan *rectilinear distance*, *Distance Matrix* yang terbentuk pada tahap inisial sebagai berikut:

	A	B	C	D	E
A	0	5	9	6	4
B		0	4	1	3
C			0	5	7
D				0	4
E					0

Jika dilakukan *clustering* dengan menggunakan *bottom-up, proximity measure* berupa *complete link*, *Distance Matrix* tahap berikutnya adalah

	A	BD	C	E
A	0	6	9	4
BD		0	5	4
C			0	7
E				0

Tugas Anda melanjutkan proses *clustering* satu tahap berikutnya, tunjukkan *Distance Matrix* dan Dendogram yang terbentuk setelah proses tersebut.

b. Suatu dataset Giant Superstore memiliki kolom-kolom: Row ID, Order ID, Order Date, Ship Date, Ship Mode, Customer ID, Customer Name, Segment, Country, City State, Postal Code, Region, ProductID, Category, SubCategory, ProductName, Sales, Quantity, Discount, Profit. Segment disini berisi deskripsi apakah customer merupakan individual customer, corporate ataupun home office. Suatu tim sains data diberikan tugas clustering untuk membantu Giant Superstore tersebut menemukan pola belanja dari setiap segment. Jika pola belanja sudah didapatkan, diharapkan akan membantu Giant Superstore untuk memberikan paket promosi yang sesuai dengan hasil clustering. Langkah-langkah yang dilakukan tim sebagai berikut:

NPM: Kelas: A | B | C | D | E | F (lingkari salah satu)

Nama:

---

(1) Melakukan EDA dan visualisasi

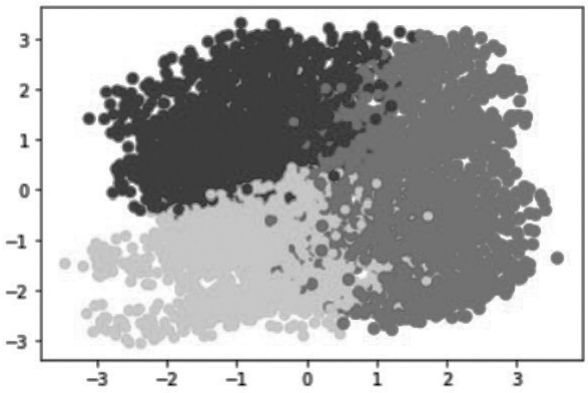
(2) Menangani duplikasi data, nilai Null dan *outlier*

(3) Mengubah atribut kategorikal ke numerik

(4) Melakukan normalisasi data dengan *standard score*

(5) Melakukan PCA hingga mendapat 9 *principal components* (80% *explained variance*)

(6) Melakukan *K-Means Clustering* dan dengan metode Elbow didapatkan hasil akhir berupa 3 *clusters* yang divisualisasikan sebagai berikut:



Terdapat langkah-langkah yang kurang tepat untuk mencapai tujuan *clustering* yang diharapkan. Tunjukkan langlah yang kurang tepat tersebut dan bagaimana perbaikan yang sebaiknya dilakukan.

Jawaban:

NPM: Kelas: A | B | C | D | E | F (lingkari salah satu)

Nama:

Soal 33 (*Model evaluation* – 8 poin)

Pada problem prediksi terjadinya *fraud* pada transaksi kartu kredit sebuah bank, input dari problem adalah suatu transaksi $\boldsymbol{x}$ dan outputnya adalah label y yang mana $y = 1$ ("*fraud*") jika $\boldsymbol{x}$ adalah transaksi *fraud*, dan $y = 0$ ("*legitimate*") jika $\boldsymbol{x}$ adalah transaksi yang legal.

Andaikan sebuah model *machine learning* f dilatih sebagai solusi problem tersebut. Model bersifat probablistik: untuk setiap input transaksi $\boldsymbol{x}$, model akan menghasilkan output $f(\boldsymbol{x})$, $0 \leq f(\boldsymbol{x}) \leq 1$, yang merupakan nilai **peluang** bahwa $\boldsymbol{x}$ tergolong ke dalam kelas *fraud*. Aturan prediksi yang kemudian dipakai adalah :

$$y = \begin{cases} 0 & \text{jika } f(\boldsymbol{x}) \leq \tau \\ 1 & \text{jika sebaliknya} \end{cases}$$

Di sini τ adalah nilai *threshold* prediksi yang dapat kita tentukan sendiri. Contohnya, jika $\tau = 0.4$, maka model akan memprediksi "*fraud*" apabila $f(\boldsymbol{x})$ bernilai lebih dari 0.4.

Pada soal ini, model diujikan pada himpunan data uji yang terdiri dari 4 transaksi *fraud* dan 8 buah transaksi *legal*. Setelah proses pengujian dilakukan diperoleh data hasil sebagai berikut:

Transaksi	Prediksi $f(\boldsymbol{x})$	*Ground truth*
$\boldsymbol{x}_1$	0.9	1
$\boldsymbol{x}_2$	0.8	1
$\boldsymbol{x}_3$	0.4	1
$\boldsymbol{x}_4$	0.4	1
$\boldsymbol{x}_5$	0.6	0
$\boldsymbol{x}_6$	0.6	0
$\boldsymbol{x}_7$	0.6	0
$\boldsymbol{x}_8$	0.2	0
$\boldsymbol{x}_9$	0.2	0
$\boldsymbol{x}_{10}$	0.2	0
$\boldsymbol{x}_{11}$	0.2	0
$\boldsymbol{x}_{12}$	0.2	0

a. Jika *threshold* $\tau = 0.5$, hitung akurasi, *precision*, *recall*, dan F1-*score* dari model f.

b. Hitung nilai *false positive rate* (FPR) dan *true positive rate* (TPR) dari model f untuk **setiap** kemungkinan nilai threshold $\tau \in \{0.85, 0.7, 0.5, 0.3\}$. (Petunjuk: Jadi ada 4 nilai TPR dan 4 nilai FPR yang dihitung, masing-masing satu untuk setiap pilihan threshold).

c. Di samping 4 pasang nilai (FPR, TPR) yang Anda peroleh di bagian (b), dapat dihitung pula bahwa ketika $\tau = 1$ dan $\tau = 0$, Anda juga akan dapatkan 2 pasang (FPR,TPR), yakni (0,0) dan (1,1), sehingga total Anda akan memperoleh total 6 pasang nilai (FPR,TPR). Urutkanlah keenam pasang nilai (FPR,TPR) tersebut dimulai dari FPR terkecil hingga terbesar. Kemudian, gambarlah plot **kurva ROC** dengan menghubungkan pasangan-pasangan (FPR,TPR) tersebut dengan garis takterputus.

d. Hitung luas daerah di bawah kurva ROC yang Anda buat.

NPM: Kelas: A | B | C | D | E | F (lingkari salah satu)

Nama:

Soal 34 (*K Nearest Neighbours* – 8 poin)

a. Jelaskan dengan singkat dan jelas perbedaan utama cara kerja kNN untuk regresi vs. klasifikasi!

b. Diberikan dataset berikut. Jika dilakukan klasifikasi (ke kategori A dan B) pada dua baris data terakhir untuk kolom Z berdasarkan metode 3-Nearest Neighbors dengan ukuran jarak Manhattan (= rectilinear), apa prediksinya? Jelaskan langkah demi langkah!

X	Y	Z
5	8	B
2	6	A
2	1	A
8	8	B
3	5	A
6	7	B
5	7	?
4	6	?

Jawaban:

NPM: Kelas: A | B | C | D | E | F (lingkari salah satu)

Nama:

Soal 35 (*Classification Tree* – 8 poin)
Anda diberikan dataset berisi informasi mengenai planet-planet yang baru ditemukan oleh ASAN, seperti yang ditunjukkan pada tabel di bawah ini.

Ukuran	Orbit	Temperatur	Dapat dihuni?
besar	jauh	205	Tidak
besar	dekat	205	Tidak
kecil	jauh	205	Tidak
besar	dekat	260	Iya
kecil	jauh	260	Iya
kecil	dekat	260	Iya
besar	dekat	380	Iya
kecil	dekat	380	Tidak

Anda diminta untuk membangun sebuah *decision tree* yang dapat mengklasifikasi apakah sebuah planet dapat dihuni atau tidak berdasarkan 3 sifat, yaitu ukuran planet, jarak orbit, dan suhu permukaan planet.

a. Pasangan atribut beserta *threshold* manakah yang paling cocok menjadi *root* dari *decision tree*? Jelaskan jawaban Anda dengan menunjukkan perhitungan yang dilakukan berdasarkan algoritma CART. Asumsikan algoritma hanya melakukan *binary split*, GINI *index* digunakan dalam perhitungan *cost*, dan *midpoint* dari nilai-nilai suatu atribut numerik digunakan sebagai *threshold*.

b. Untuk setiap *child node* yang dihasilkan dari proses (a), jelaskan apa yang dilakukan oleh algoritma CART pada tahap berikutnya.

Jawaban: