**Ulangan Tengah Semester Big Data 2021/2022**

**(Take Home 3 Hari)**

**Nurhayati, Ph.D**

**Aryajaya Alamsyah, S.Kom.**

**Nomor 1 Pemahaman Dasat**

1. Jelaskan apa yang dimaksud dengan Big Data, Kecerdasan Buatan, Machine Learning, Deep Learning, Data Mining dan Statistik.
2. Bagaimana hubungan antar ke 6 sub bidang tersebut.

**Nomor 2. Pengenalan R dan Visualisasi Data.**

1. Buatlah data dummy sebanyak 100 baris tentang ojek onlie. Jika variabel yang digunakan adalah tanggal, nama pengemudi, jumlah Go Ride, jumlah Go Food, jumlah penghasilan dan jumlah bonus, total penghasilan. Dengan ketentuan:
2. Setiap 10 baris data hanya diisi 1 tanggal saja. Sehingga data berisi 1 jan 2020 sampai 10 jan 2020. Variabel nama berisi 10 nama yang berbeda. Setiap nama berpasangan dengan variabel tanggal. Sehingga contoh data menjadi berikut.

|  |  |
| --- | --- |
| Tanggal | Nama |
| 1 Jan 2020 | Pengemudi 1 |
| ……….. | ……….. |
| 1 Jan 2020 | Pengemudi 10 |
| ……….. | ……….. |
| 10 Jan 2020 | Pengemudi 1 |
| ……….. | ……….. |
| 10 Jan 2020 | Pengemudi 10 |

1. Variabel Go Ride, Go Food diisi dengan angka dengan rentang 5-15 (mewakili jumlah orderan yang didapat perhari).
2. Variabel GoRide dan GoFood diisi menggunakan kaidah distribusi normal.
3. Jika persatu kali order Go Ride seharga 15.000 dan Go Food 30.000
4. Variabel jumlah penghasilan memiliki rumus berikut

jumlah penghasilan = (Jumlah Go Ride\*15.000) + (Jumlah Go Food\*30.000)

1. Variabel jumlah bonus diisi 0 dan 100.000 rb. Jika hanya pengemudi yang mengumpulkan point minimal 25 yang mendapatkan bonus 100.000rb. Rumus menghitung point adalah (Jumlah Go Ride\*1) + (Jumlah Go Food\*2).
2. Variabel total penghasilan diisi dengan menjumlahkan jumlah penghasilan dan jumlah bonus.

Pertanyaanya adalah

1. Siapakah pengemudi yang memiliki total penghasilan terbesar jika diakumulasi selama 10 hari?
2. Siapakah pengemudi yang memiliki total penghasilan terkecil jika diakumulasi selama 10 hari?
3. Berapa rata-rata penghasilan setiap pengumudi ojek onlie jika diakumulasi selama 10 hari?
4. Lalukan visualisasi data time series untuk melihat fluktuasi total penghasilan pengemudi ojek oline.
5. Lakukan visualisasi group bar (jumlah penghasilan, jumlah bonus, total penghasilan) setiap pengemudi ojek online.

**Nomor 3. Pengenalan Analisa Statistik**

1. Jelaskan apa yang dimaksud dengan distribusi normal, seragam dan poisson.
2. Bagaimana contoh data yang mengikuti kaidah distribusi normal, seragam dan poisson (untuk pembuktiannya dilakukan visualisasi histrogram).
3. Perhatikan dataset/ dataset\_iklim\_harian\_mar.xlsx. Hitung secara manual korelasi antara suhu rata-rata dengan kelembaban rata-rata menggunakan metode pearson, kendall, spearman rank. Kemudian hitung juga nilai determinasi dan signifikansinya.
4. Jelaskan mengapa terjadi perbedaan nilai korelasi antar metode tersebut.

**Nomor 4. Pra Proses Data dan Explorasi Data.**

1. Perhatikan Bahan-Ajar-Big-Data/dataset/UTS-EDA-vgsales.csv. Sumber

<https://www.kaggle.com/gregorut/videogamesales>

1. Kemudian buat visualisasi data seperti lampiran tugas7 Gambar 1 sampai Gambar 5

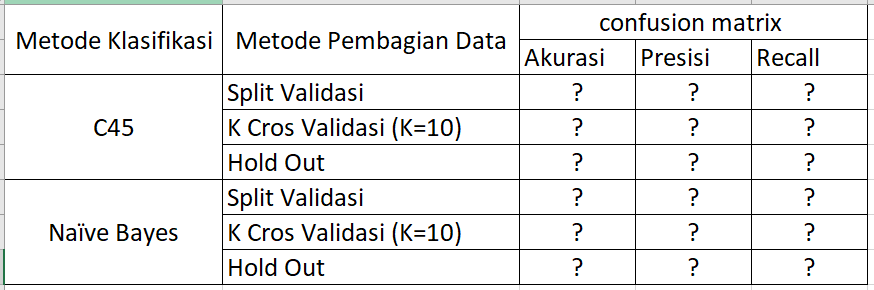
**Nomor 5 Klasifikasi C45 dan Naïve Bayes**

1. Perhatikan Bahan-Ajar-Big-Data/dataset/UTS-Pima-Indians-Diabetes.csv. Sumber

<https://github.com/npradaschnor/Pima-Indians-Diabetes-Dataset/blob/master/diabetes.csv>

1. Lakukan proses normaliasi data menggunakan metode max-min skala -1 sampai 1
2. Pembagian data menggunakan persentase 80% dan 20%
3. Berapa nilai confusion matrix dari percobaan berikut:

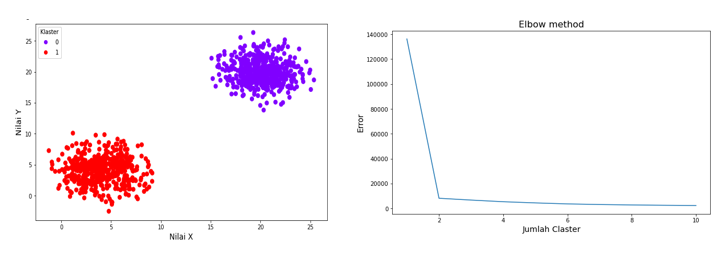
Tabel x. Template Jawaban Nomor 5



1. Buatlah kesimpulan dan analisa hasil dari percobaan dan perlakukan yang telah dilakukan menggunakan metode C45 dan Naïve Bayes

**Nomor 6. Klasterisasi Kmeans**

Perhatikan Bahan-Ajar-Big-Data/dataset/ UTS-Dummy-Kmean.xlsx

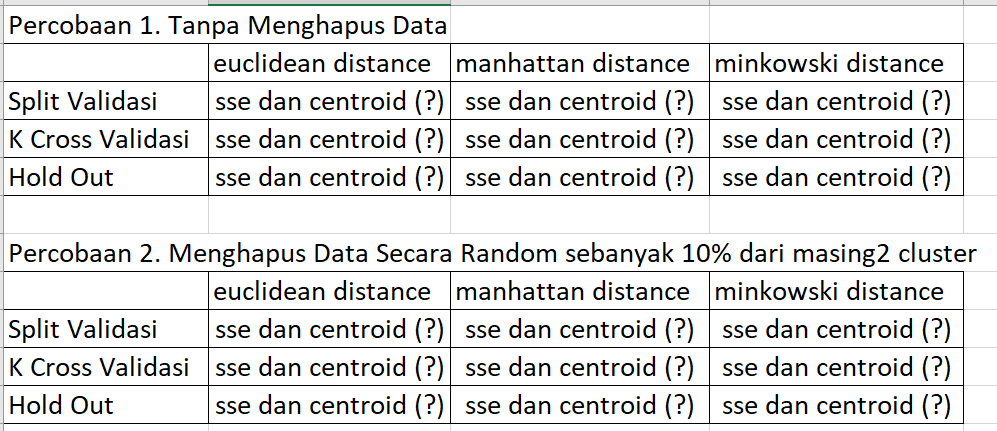


|  |  |
| --- | --- |
| a | b |

Gambar x. Visualisasi Kmean (a) dan Visualisasi Nilai K-Optimal (b)

Berapa nilai Sum Square Eror dan nilai titik centroid dari percobaan berikut:

Tabel x. Template Jawaban Nomor 6



* Pembagian data menggunakan persentase 80% dan 20%
* Makna percobaan 2 adalah membuktikan teori bahwa algoritma k-mean dapat memberikan hasil yang optimal walaupun terdapat nilai pencilan dan nilai kosong.
* Buatlah kesimpulan dan analisa hasil dari percobaan dan perlakukan yang telah dilakukan menggunakan metode kmean.