**Pokok Bahasan XIV**

**Analisis Cluster**

**Kode Pokok Bahasan**: TIK.RPL03.001.011.01

**Deskripsi Pokok Bahasan**:

Membahas tentang Association Rule pada R dengan dataset yang diberikan.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No | Elemen Kompetensi | Indikator Kinerja | Jml Jam | Hal |
| 1 | Memahami cara implementasi Analisis Cluster | 1.1 Mampu memahami cara implementasi Analisis Cluster dengan Hierarchical Cluster, Dendogram, k-mean clustering | 1 | 12 |

**TUGAS PENDAHULUAN**

Hal yang harus dilakukan dan acuan yang harus dibaca sebelum praktikum :

1. Menginstal R pada PC masing-masing praktikan.

2. Menginstal R Studio pada PC masing-masing praktikan.

**DAFTAR PERTANYAAN**

1. Berikan penjelasan mengenai apa itu “Analisis Cluster”?

**Jawab : Analisis klaster adalah metode statistik dalam penelitian yang memungkinkan peneliti untuk mengelompokkan atau mengelompokkan sekumpulan objek ke dalam kluster-kluster kecil namun berbeda yang berbeda karakteristiknya dari kluster-kluster lain yang berbeda.**

2. Jelaskan pengertian Hierarchical Cluster, Dendogram, k-mean clustering?

**Jawab :**

Hierarchical Cluster = metode pengelompokan data yang dimulai dengan setiap satu pengamatan sebagai clusternya sendiri kemudian terus mengelompokkan pengamatan ke dalam kelompok yang semakin besar.

Dendogram = sejenis diagram treelike yang digunakan secara hirarkis kekelompokan. Ini merupakan daftar semua peserta pada satu akhir dan kemudian mengarahkan cabang keluar dari peserta tersebut yang serupa dan menghubungkan mereka dengan simpul yang mewakili sebuah cluster.

k-mean clustering = Pengklasteran k rata-rata adalah algoritme untuk membagi n pengamatan menjadi k kelompok sedemikian hingga tiap pengamatan termasuk ke dalam kelompok dengan rata-rata terdekat. Hasilnya adalah pembagian pengamatan ke dalam sel-sel Voronoi. Pengklasteran k rata-rata meminimalkan ragam dalam klaster.

**TEORI SINGKAT**

Analisis cluster sering juga disebut analisis gerombol. Analisis cluster adalah analisis statistika yang bertujuan untuk mengelompokkan data sedemikian sehingga data yang berada dalam kelompok yang sama mempunyai sifat yang relatif homogen daripada data yang berada dalam kelompok yang berbeda.

Ditinjau dari hal-hal yang dikelompokkan, analisis cluster dibagi menjadi dua macam, yaitu pengelompokan observasi dan pengelompokan variabel. Dalam pembahasan ini, pengelompokan yang dilakukan adalah pengelompokan observasi.

**LAB SETUP**

Hal yang harus disiapkan dan dilakukan oleh praktikan untuk menjalankan praktikum modul ini.

1. Menginstall library yang dibutuhkan untuk mengerjakan modul.

2. Menjalankan R Studio.

**ELEMEN KOMPETENSI I**

**Deskripsi:**

Memahami cara implementasi Analisis Cluster

**Kompetensi Dasar**:

1. Mampu memahami cara implementasi Analisis Cluster dengan Hierarchical Cluster, Dendogram, k-mean clustering.

**Latihan 1.1.1**

**Penjelasan Singkat :**

Pada latihan ini anda akan diminta untuk menerapkan Analisi Cluster pada data yang diberikan.

**Langkah-Langkah Praktikum:**

Dataset 1 :

|  |  |  |
| --- | --- | --- |
| observation | Income | Education |
| s1 | 5 | 5 |
| s2 | 6 | 6 |
| s3 | 15 | 14 |
| s4 | 16 | 15 |
| s5 | 25 | 20 |
| s6 | 30 | 19 |

Hierarchical Cluster :

|  |
| --- |
| library("readxl")  #my\_data <- read\_excel(file.choose(), sheet = "",range = "")  my\_data <- read\_excel("E:/Cluster analysis/clusterPage127.xlsx",na = "-")  print(my\_data)  str(my\_data)  d.euc <- dist(my\_data)  d.sqeuc <- d.euc^2  cluster<- hclust(d = d.sqeuc, method = "centroid")  library("factoextra")  fviz\_dend(cluster, cex = 0.6)    #if(!require(devtools)) install.packages("devtools")   #devtools::install\_github("kassambara/factoextra") |

Output :

|  |
| --- |
|  |

Output Dendogram :

|  |
| --- |
|  |

Hierarchical Cluster in Python :

|  |
| --- |
| import pandas as pd  import numpy as np  import matplotlib.pyplot as plt  df = pd.read\_excel("data1.xlsx")  df.head()  data= df.iloc[:,[1,2]]  data.head()  from sklearn.preprocessing import normalize  data\_scaled = normalize(data)  data\_scaled = pd.DataFrame(data\_scaled, columns=data.columns)  data\_scaled.head()  import scipy.cluster.hierarchy as shc  plt.figure(figsize=(10,7))  plt.title("Dendrograms")  dend = shc.dendrogram(shc.linkage(data\_scaled, method = 'ward')) |

Output :

|  |
| --- |
|  |

Output Dendogram :

|  |
| --- |
|  |

Script k-mean clustering

> klaster <- kmeans(dataku[, 2:3], 3, nstart = 20)

> klaster

> table(klaster$cluster, dataku$X)

Output :

|  |
| --- |
|  |

Script k-mean clustering

|  |
| --- |
| import matplotlib.pyplot as plt  from sklearn.cluster import KMeans  import pandas as pd  df2 = pd.read\_excel("data1.xlsx")  df2.head()  data2= df.iloc[:,[1,2]]  data2.head()  kmeans =KMeans(n\_clusters=3).fit(data2)  centroids = kmeans.cluster\_centers\_  print(centroids)  plt.scatter(df.iloc[:,1], df.iloc[:,2], c = kmeans.labels\_.astype(float), s  =50, alpha = 0.5)  plt.scatter(centroids[:,0], centroids[:,1], c='red', s=50)  plt.show |

Output :

|  |
| --- |
|  |

Dataset 2 :

|  |  |  |  |
| --- | --- | --- | --- |
| Supplier | DAR | DSR | Quality |
| S1 | 96.81 | 73.85 | 100 |
| S2 | 99.64 | 65.79 | 100 |
| S3 | 96.5 | 71.63 | 62.86 |
| S4 | 99.349 | 79.38 | 96.86 |
| S5 | 100 | 88.24 | 100 |
| S6 | 71.4 | 60.7 | 71.4 |
| S7 | 99.827 | 84.54 | 79.43 |
| S8 | 99.9 | 82.98 | 88 |
| S9 | 99.058 | 95.16 | 98 |
| S10 | 98.1 | 90.77 | 81.43 |
| S11 | 99.574 | 79.66 | 95.14 |
| S12 | 99.606 | 80.4 | 71.1 |

Script k-mean clustering

|  |
| --- |
| dataku=read.delim("clipboard") data\_latih <-dataku[,c(2,3,4)]  data\_latih\_matrix <- as.matrix(scale(data\_latih)) library(factoextra) library(NbClust)  lnb <- NbClust(data\_latih, distance = "euclidean", min.nc = 2,max.nc = 8, method =  "complete", index ="all")  km.res=kmeans(data\_latih, 3, nstart=25) km.res library(dplyr)  fviz\_cluster(km.res, data = data\_latih, geom = "point",stand = FALSE, frame.type =  "norm") fviz\_cluster(km.res, data = data\_latih) |

Output :

|  |
| --- |
|  |

Script k-mean clustering in Python

|  |
| --- |
| import matplotlib.pyplot as plt  from sklearn.cluster import KMeans  import pandas as pd  df2 = pd.read\_excel("data2.xlsx")  df2.head()  data2= df.iloc[:,[1,2,3]]  data2.head()  kmeans =KMeans(n\_clusters=3).fit(data2)  centroids = kmeans.cluster\_centers\_  print(centroids)  plt.scatter(df.iloc[:,1], df.iloc[:,2], c = kmeans.labels\_.astype(float), s  =50, alpha = 0.5)  plt.scatter(centroids[:,0], centroids[:,1], c='red', s=50)  plt.show |

Output :

|  |
| --- |
|  |

Script hierarchical clustering

|  |
| --- |
| dataku2=read.delim("clipboard") df=scale(dataku2[,2:4]) res.dist <- dist(df, method = "euclidean") res.hc <- hclust(d = res.dist, method = "ward.D2") library("factoextra") fviz\_dend(res.hc, cex = 0.5)  fviz\_dend(res.hc, k = 4, cex = 0.5, k\_colors = c("#2E9FDF", "#00AFBB", "#E7B800", "#F  C4E07"), color\_labels\_by\_k = TRUE, rect = TRUE) |

Output :

|  |
| --- |
|  |

Script hierarchical clustering in Python

|  |
| --- |
| import pandas as pd  import numpy as np  import matplotlib.pyplot as plt  df = pd.read\_excel("data2.xlsx")  df.head()  data= df.iloc[:,[1,2,3]]  data.head()  from sklearn.preprocessing import normalize  data\_scaled = normalize(data)  data\_scaled = pd.DataFrame(data\_scaled, columns=data.columns)  data\_scaled.head()  import scipy.cluster.hierarchy as shc  plt.figure(figsize=(10,7))  plt.title("Dendrograms")  dend = shc.dendrogram(shc.linkage(data\_scaled, method = 'ward')) |

Output :

|  |
| --- |
|  |

Interpretasi :

|  |
| --- |
| Dendogram menggambarkan bagaimanaa dari data-data yang disediakan bisa digabungkan menjadi suatu dataset yang lebih mudah dibaca, semisal ada data yang berhubungan langsung digambarkan dari mulai data primer sampai data sekunder |

Tugas :

Diberikan data harga beberapa komoditas berbagai pasar di Jakarta. Lakukan analisis cluster menggunakan metode hierarchical clustering dan k-means serta interpretasikan  maknanya.

|  |  |  |  |
| --- | --- | --- | --- |
| **Pasar** | Beras | Jeruk | Minyak |
| Pasar Senen Blok III - VI | 13000 | 18000 | 13000 |
| Pasar Jembatan Merah | 11000 | 25000 | 11500 |
| Pasar Sunter Podomoro | 12000 | 30000 | 12000 |
| Pasar Rawa Badak | 12000 | 23000 | 12000 |
| Pasar Grogol | 11000 | 25000 | 11000 |

|  |  |  |  |
| --- | --- | --- | --- |
| Pasar Minggu | 13125 | 25000 | 10000 |
| Pasar Mayestik | 12000 | 20000 | 13000 |
| Pasar Pramuka | 12000 | 25000 | 14000 |
| Pasar Kramat Jati | 11700 | 25000 | 12000 |
| Pasar Jatinegara | 11000 | 22000 | 12000 |
| Pasar Perumnas Klender | 10900 | 20000 | 13000 |
| Pasar Pulo Gadung | 10700 | 22000 | 12000 |
| Pasar Pal Meriam | 10000 | 20000 | 12000 |
| Pasar Ciplak | 11200 | 17000 | 12000 |
| Pasar Cijantung | 11000 | 23000 | 12000 |
| Pasar Cibubur | 11500 | 20000 | 12500 |
| Pasar Ujung Menteng | 11500 | 30000 | 11000 |
| Pasar Tanah Abang Blok A-G | 11000 | 25000 | 12000 |
| Pasar Petojo Ilir | 11000 | 23000 | 12000 |
| Pasar Gondangdia | 13000 | 20000 | 12000 |
| Pasar Paseban | 12000 | 20000 | 11000 |
| Pasar Cempaka Putih | 12000 | 19000 | 12000 |
| Pasar Johar Baru | 12000 | 25000 | 13000 |
| Pasar Baru Metro Atom | 11500 | 20000 | 11000 |
| Pasar Kebayoran Lama | 12500 | 23000 | 13000 |
| Pasar Cipete | 10500 | 25000 | 12500 |
| Pasar Pondok Labu | 12000 | 25000 | 13000 |
| Pasar Lenteng Agung | 11500 | 28000 | 11500 |
| Pasar Mampang Prapatan | 13000 | 17000 | 12000 |
| Pasar Tebet Barat | 12000 | 20000 | 11000 |
| Pasar Rumput | 13000 | 20000 | 13000 |
| Pasar Tomang Barat | 13000 | 24000 | 12000 |
| Pasar Pos Pengumben | 11400 | 22000 | 12000 |
| Pasar Pal Merah | 10800 | 25000 | 12000 |
| Pasar Jembatan Lima | 10000 | 35000 | 12500 |
| Pasar Kelapa Gading | 12000 | 25000 | 12500 |
| Pasar Pademangan Timur | 12000 | 23000 | 12000 |
| Pasar Kalibaru | 12000 | 20000 | 11000 |
| Pasar Koja Baru | 12000 | 18000 | 12500 |

K-means Clustering

|  |
| --- |
| library(factoextra)  library(NbClust)  library(dplyr)  dataku=read.delim("clipboard")  dataku  data\_latih <-dataku[,c(2,3,4)]  data\_latih\_matrix <- as.matrix(scale(data\_latih))  lnb <- NbClust(data\_latih, distance = "euclidean", min.nc = 2,max.nc = 8, method = "complete", index ="all")  km.res=kmeans(data\_latih, 3, nstart=25)  km.res  fviz\_cluster(km.res, data = data\_latih, geom = "point",stand = FALSE, frame.type = "norm")  fviz\_cluster(km.res, data = data\_latih) |
|  |

Hierarchical Clustering

|  |
| --- |
| df=scale(dataku[,2:4])  res.dist <- dist(df, method = "euclidean")  res.hc <- hclust(d = res.dist, method = "ward.D2")  library("factoextra")  fviz\_dend(res.hc, cex = 0.5)  fviz\_dend(res.hc, k = 4, cex = 0.5, k\_colors = c("#2E9FDF", "#00AFBB", "#E7B800", "#FC4E07"), color\_labels\_by\_k = TRUE, rect = TRUE) |
|  |

Intepretasi

|  |
| --- |
| Interpretasi dari data clustering diatas adalah dari sleuruh pasar yang ada bisa dilihat bahwa harga mayoritas mengikuti harga pasar yang ada, mulai dari beras, jagung, dan minyak semuanya dari pasar ke pasar masih mempunyai harga pasaran pada umumnya |

**CEK LIST**

|  |  |  |  |
| --- | --- | --- | --- |
| Elemen Kompetensi | No Latihan | Penyelesaian | |
| Selesai | Tidak selesai |
| 1 | 1.1.1 |  |  |
| 1.1.2 |  |  |

**FORM UMPAN BALIK**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Elemen Kompetensi** | **Tingkat Kesulitan** | | | **Tingkat Ketertarikan** | | | **Waktu Penyelesaian dalam menit** | |
| Memahami cara implementasi Analisa Cluster |  |  |  |  |  |  | |  | |
|  |  | Sangat Mudah |  |  | Tidak Tertarik | | 30 menit | |
|  |  |  |  |  |  | |  | |
|  |  | Mudah |  |  | Cukup Tertarik | |  | |
|  |  |  |  |  |  | |  | |
|  |  | * Biasa |  |  | * Tertarik | |  | |