PENURUNAN DATA

STQD6414 PERLOMBONGAN DATA

Prof. Madya Dr. Nurulkamal Masseran

Jabatan Sains Matematik

Universiti Kebangsaan Malaysia

PENGENALAN:

- Set data yang besar akan menjadikan analisis perlombongan data kurang efisien.
- Malah, saintis data juga mungkin akan mudah terkeliru dalam menjalankan analisis.

- Teknik penurunan data boleh digunakan untuk mengurangkan dimensi atau amaun cerapan bagi suatu set data.
- Namun, prinsipnya ialah data yang diturunkan masih mampu memberikan maklumat yang hampir sama dengan data asal.

- Teknik penurunan data terbahagi kepada:

i) Penurunan Dimensi Data.

ii) Penurunan Numerositi Data.

PENURUNAN DIMENSI DATA:

- Dimensi yang besar dalam data menjadikan kecekapan al-khwarizmi dalam kaedah perlombongan kurang efisien.

- Malah, saiz bagi simpanan data juga mungkin tidak memadai untuk menyimpan data yang terlalu besar.

- Data dengan dimensi besar boleh dikecilkan dimensi menerusi kaedah:

i) Mengeluarkan Atribut.

ii) Analisis Komponen Utama.

iii) Analisis Faktor.

iv) Dan lain-lain.

MENGELUARKAN ATRIBUT:

- Mengeluarkan atribut tertentu merupakan kaedah mengurangkan dimensi data yang paling mudah
- Ini boleh dijalankan dengan mengeluarkan atribut yang bersifat seperti berikut:

i) Atribut yang hampir sama sifat:

- Jika didapati terdapat atribut-atribut yang merupakan duplikasi bagi atribut lain, maka maklumat yang sama boleh diperolehi dalam atribut tersebut.
- **Contoh:** harga pembelian sesuatu produk dan jumlah cukai jualan yang dibayar.

ii) Atribut yang tidak relevan:

- Sesuatu atribut hanya memberikan maklumat berguna jika ianya diperlukan untuk mencapai objektif analisis.
- **Contoh:** atribut ID pelajar adalah tidak relevan untuk menganalisis prestasi CGPA pelajar.

MENGELUARKAN ATRIBUT:

iii) Atribut yang tidak signifikan:

- Sesuatu atribut yang didapati tidak singnifikan secara statistik dalam analisis yang dijalankan.

Contoh:

- Atribut yang tidak singnifikan yang dikesan menerusi analisis model regresi.
- Iaitu, apabila hubungan p/ubah sambutan Y dan p/ubah regressor X_j boleh diterangkan secara linear:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

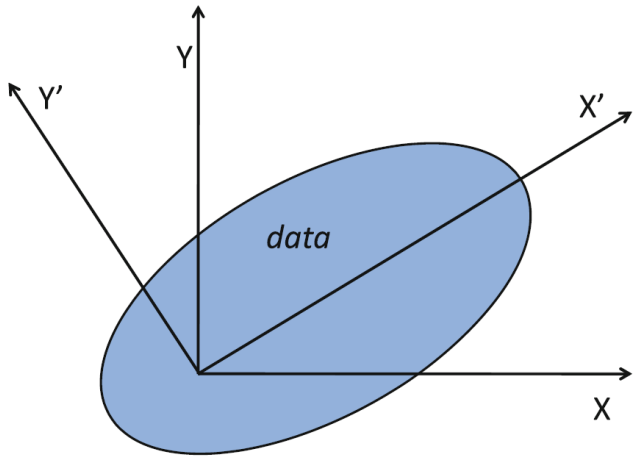
- Kita berminat untuk menguji sama ada $H_0: \beta_j = 0$ vs $H_a: \beta_j \neq 0$.
- Atribut X_j yang mana didapati parameter $\beta_j = 0$ secara signifikan, boleh dikeluarkan daripada data.

ANALISIS KOMPONEN UTAMA (PCA):

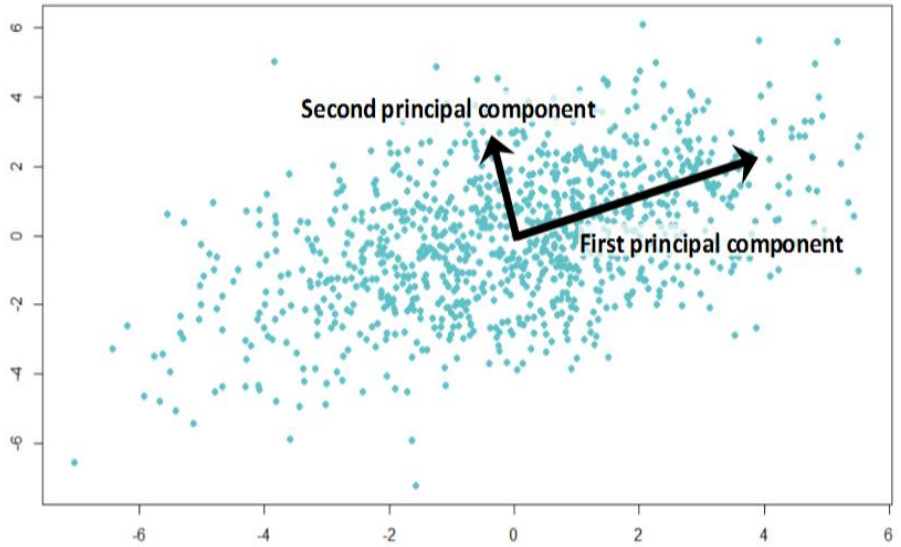
- Idea asas PCA ialah untuk mendapatkan suatu set p/ubah yang lebih kecil daripada set p/ubah asal yang dapat menggambarkan kebanyakan varians data asal menerusi penjelmaan linear.
- Iaitu, set vektor *k* yang bersifat ortogonal yang boleh mewakili data sebenar, dengan $k \leq p$ (*p* ialah dimensi data asal).
- Set atribut baru ini ditunjukkan dalam susunan sumbangan varians yang semakin kecil.
- Pembolehubah pertama PCA, ialah komponen utama pertama yang mengandungi varians terbesar terhadap set data asal.
- Pembolehubah kedua PCA, ialah komponen utama kedua yang mengandungi varians kedua terbesar terhadap set data asal.
- Dan seterusnya.

ANALISIS KOMPONEN UTAMA (PCA):

- Prosedur yang umum adalah untuk menyimpan beberapa komponen utama yang mengandungi 80% atau lebih varians set data asal.
- PCA berguna apabila terdapat set data yang mempunyai banyak atribut dan korelasi antara beberapa atribut adalah agak tinggi (saling berkait).
- Maklumat berkaitan set komponen utama ditunjukkan menerusi nilai-nilai eigen dan vektor-vektor eigen yang diperolehi daripada matriks korelasi.



.1 PCA. X' and Y' are the first two principal components obtained



PROSEDUR PCA:

i) Skalakan data input dengan mempiawaikan julat bagi setiap atribut (skor-*z*).

ii) Dapatkan *k*-set vektor ortogonal berdasarkan data yang telah di piawaikan.

iii) Komponen utama disusun secara sumbangan menurun berdasarkan maklumat nilai eigen. Komponen utama berfungsi sebagai set paksi-paksi baru untuk data yang diselaraskan mengikut varians data asal.

iv) Pengurangan dimensi data dibuat dengan membuang komponen yang memberikan sumbangan varians yang rendah:

- Hanya komponen utama yang menerangkan sumbangan varians yang tinggi dikekalkan sebagai set p/ubah baharu.

- Namun, analisis PCA hanya boleh dijalankan jika semua atribut adalah berangka.

PROSEDUR PCA:

- Bagi vektor rawak $\mathbf{X= (X_1, X_2, \ldots, X_p)}$.
- Matriks kovarians/korelasi var(**X**):

$$\mathrm{var}(\mathbf{X}) = \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{pmatrix}$$

- PCA boleh diperolehi menerusi hubungan linear berikut:

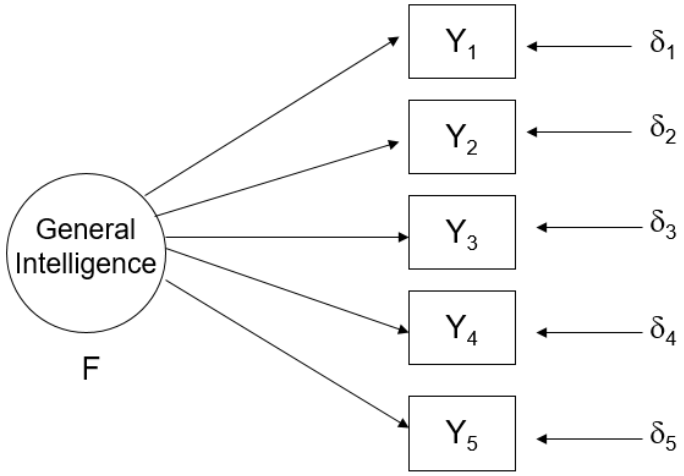
$$\begin{aligned} Y_1 &= e_{11}X_1 + e_{12}X_2 + \cdots + e_{1p}X_p \\ Y_2 &= e_{21}X_1 + e_{22}X_2 + \cdots + e_{2p}X_p \\ &\vdots \\ Y_p &= e_{p1}X_1 + e_{p2}X_2 + \cdots + e_{pp}X_p \end{aligned}$$

- Dengan $\mathbf{Y= (Y_1, Y_2, \ldots, Y_p)}$ set pemboleh ubah PCA.
- $\mathbf{e_i}$ ialah set vektor eigen bagi matriks kovarians (atau matriks korelasi).
- Komponen utama $\mathbf{Y_1}$ mempunyai varians terbesar bagi data, diikuti oleh $\mathbf{Y_2}$ dan seterusnya. Ini ditunjukkan menerusi nilai eigen:

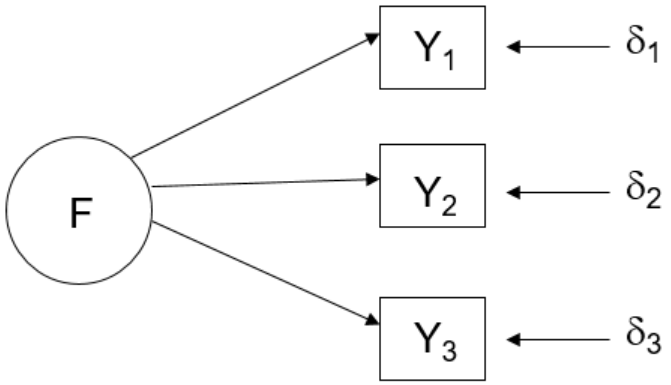
$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$$

ANALISIS FAKTOR (FA):

- Analisis Faktor merupakan teknik penurunan data yang digunakan untuk menerangkan kovarians antara p/ubah tercerap dalam bentuk p/ubah tak tercerap (pendam) yang berdimensi lebih kecil.
- Analisis faktor bertujuan untuk mencari faktor-faktor pendam (*hidden factors*) dalam p/ubah data asal.
- Dalam analisis faktor, kita beranggapan bahawa terdapat suatu set faktor pendam (tidak tercerap) F_j , $j = 1, \ldots, k$; yang boleh diterbitkan daripada data asal.
- Analisis faktor mencirikan sifat kebergantungan antara atribut-atribut data asal melalui faktor-faktor yang berdimensi lebih kecil.



CONTOH: MODEL SATU-FAKTOR:



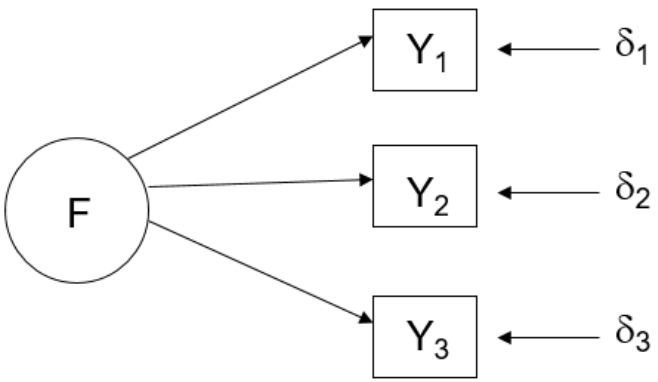
$$Y_1 = \lambda_1 F + \delta_1$$

$$Y_2 = \lambda_2 F + \delta_2$$

$$Y_3 = \lambda_3 F + \delta_3$$

- *F* ialah faktor pendam yang tidak dicerap.
- Y_1, Y_2, Y_3 ialah p/ubah bagi data cerapan.
- δ_i ralat yang mewakili variasi dalam Y_i yang tidak dapat diterangkan oleh faktor *F*.
- Y_i diterangkan menerusi hubungan linear bagi faktor *F* dan ralat δ_i.

ANDAIAN DALAM MODEL SATU-FAKTOR:



$$Y_1 = \lambda_1 F + \delta_1$$

$$Y_2 = \lambda_2 F + \delta_2$$

$$Y_3 = \lambda_3 F + \delta_3$$

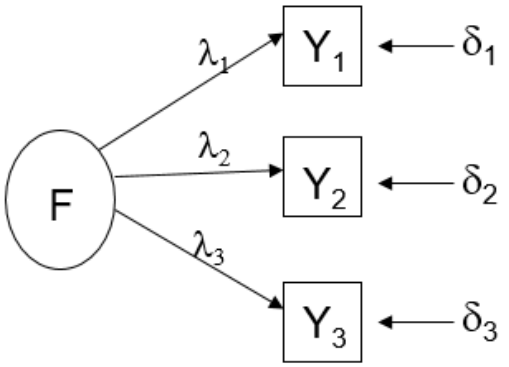
- F ialah faktor penyebab bagi Y_1, Y_2, Y_3.
- F adalah tak bersandar terhadap δ_j, iaitu $cov(F, \delta_j)=0$
- δ_i dan δ_j adalah tak bersandar bagi $i \neq j$, iaitu $cov(\delta_i, \delta_j)=0$
- Ketakbersandaran bersyarat: P/ubah Y_i dan Y_j adalah tak bersandar antara satu sama lain, diberi faktor F, iaitu $cov(Y_i, Y_j | F) = 0$.

TAFSIRAN MODEL SATU-FAKTOR :

- Bagi p/ubah Y_1, Y_2, Y_3 yang telah dipiawaikan: $var(Y_i)=var(F)=1$

- Pemberatan faktor *(Factor loadings)*:

  $\lambda_i = corr(Y_i, F)$



- Komunaliti *(Communality)* bagi p/ubah Y_i:

  $h_i^2 = \lambda_i^2 = [corr(Y_i,F)]^2$

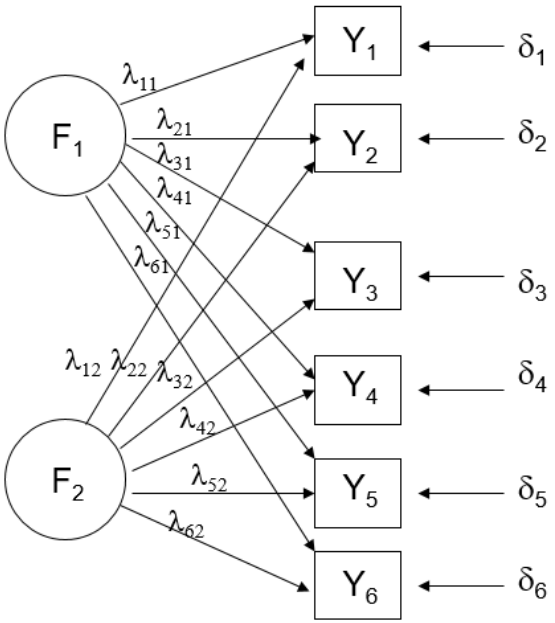
  =% varians bagi Y_i yang diterangkan oleh faktor *F*.

$$Y_1 = \lambda_1 F + \delta_1$$
$$Y_2 = \lambda_2 F + \delta_2$$
$$Y_3 = \lambda_3 F + \delta_3$$

- Keunikan *(Uniqueness)* bagi Y_i:

  $1-h_i^2$ = varians reja bagi Y_i.

CONTOH: MODEL DUA-FAKTOR



$$Y_1 = \lambda_{11}F_1 + \lambda_{12}F_2 + \delta_1$$
$$Y_2 = \lambda_{21}F_1 + \lambda_{22}F_2 + \delta_2$$
$$Y_3 = \lambda_{31}F_1 + \lambda_{32}F_2 + \delta_3$$
$$Y_4 = \lambda_{41}F_1 + \lambda_{42}F_2 + \delta_4$$
$$Y_5 = \lambda_{51}F_1 + \lambda_{52}F_2 + \delta_5$$
$$Y_6 = \lambda_{61}F_1 + \lambda_{62}F_2 + \delta_6$$

- Faktor F_1 and F_2 merupakan faktor sepunya kerana kedua-dua faktor ini berkongsi lebih dari dua p/ubah $Y_1, Y_2, Y_3, Y_4, Y_5, Y_6$ yang sama dalam setiap faktor.
- Model dengan *m*-Faktor dan *n*-p/ubah membawa kepada spesifikasi model yang lebih kompleks.

PUTARAN FAKTOR:

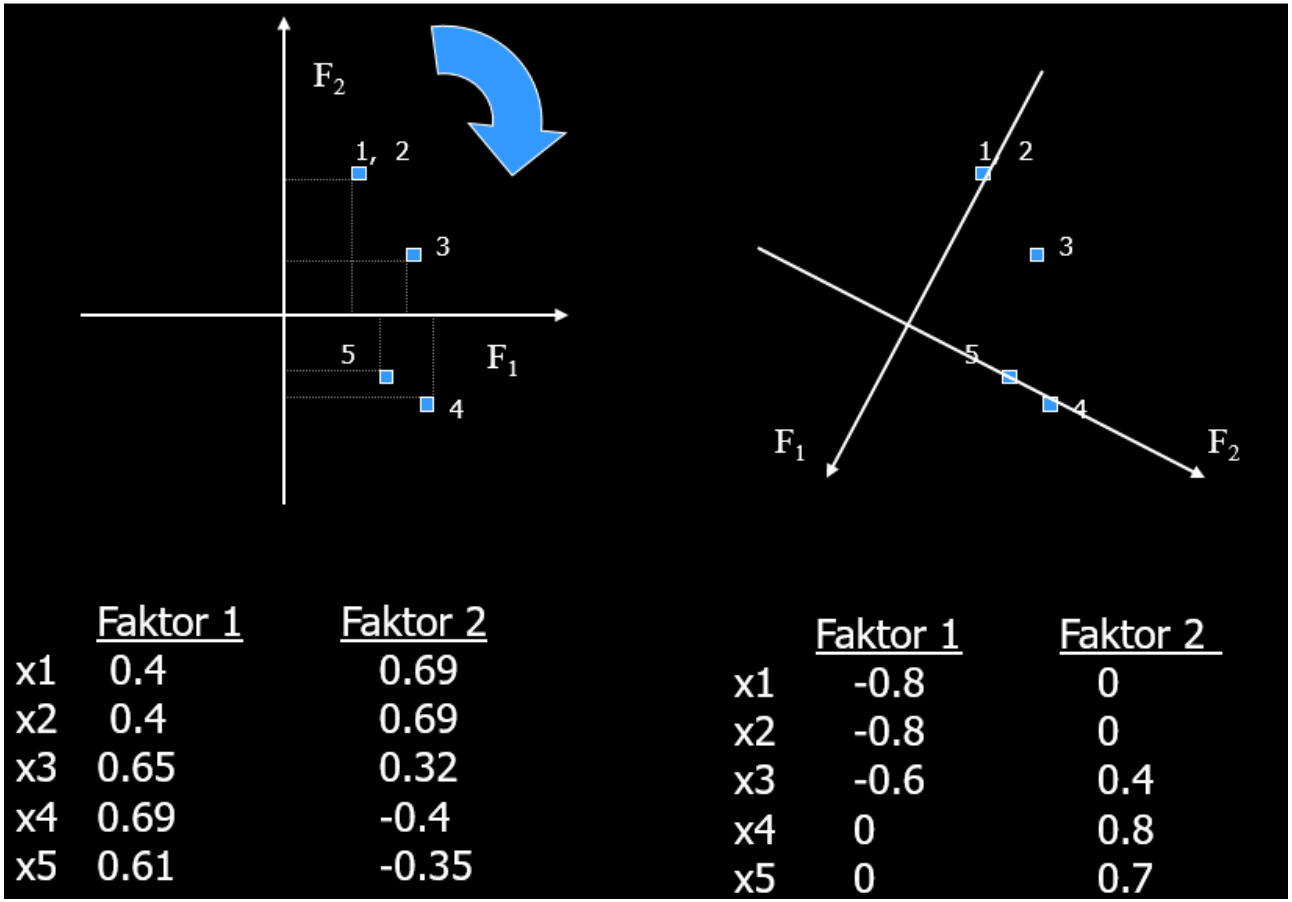
- Bertujuan untuk mendapatkan struktur yang lebih ringkas dan menjadikan faktor-faktor lebih mudah ditafsir.
- Putaran faktor tidak mempengaruhi penyuaian model faktor.
- Bilangan faktor dan komunaliti bagi *Y* yang diperolehi dari penyuaian model adalah tidak berubah selepas putaran faktor dijalankan.
- Putaran faktor dibuat dengan mentakrifkan semula faktor-faktor yang diperolehi sedemikian hingga:

i) Nilai sebahagian faktor cenderung untuk meningkat dengan secara signifikan menghampiri -1 atau 1.

ii) Nilai sebahagian faktor yang lain akan cenderung untuk menurun dengan sangat rendah menghampiri 0.

- Ini menjadikan pengaruh setiap faktor terhadap p/ubah asal lebih ketara dan lebih mudah ditafsir.

CONTOH PUTARAN FAKTOR:



	Faktor 1	Faktor 2
x1	0.4	0.69
x2	0.4	0.69
x3	0.65	0.32
x4	0.69	-0.4
x5	0.61	-0.35

	Faktor 1	Faktor 2
x1	-0.8	0
x2	-0.8	0
x3	-0.6	0.4
x4	0	0.8
x5	0	0.7

PENURUNAN NUMEROSITI DATA:

- Penurunan numerositi data boleh dibuat dengan mengantikan data asal dalam bentuk alternatif lain:

i. Model Berparameter:

contoh:

- Model regresi,
- Model log-linear
- Taburan kebarangkalian, dan lain-lain.

ii. Model Tak Berparameter

contoh:

- Histogram
- Pensampelan semula.
- Pengkelompokan, dan lain-lain.

MODEL BERPARAMETER:

- Model statistik terbaik yang diperoleh dari penyuaian data akan digunakan sebagai perwakilan terhadap data.
- Hanya parameter-parameter model dan maklumat-maklumat penting akan disimpan.
- Data simulasi yang dijana daripada model adalah menghampiri data sebenar.
- **Contoh Model Statistik:**

i) Model Regresi Linear Berganda:

- Data *Y* ialah berangka dan perlu menghampiri taburan Normal.

ii)Model Regresi Log-linear:

Data taburan Y ialah diskrit multi-dimensi.

iii) Model Taburan Kebarangkalian:

- Taburan Univariate: Normal, Poisson, Weibull, Gamma, Pareto dan banyak lagi.
- Taburan Multivariat.

ii) Dan pelbagai model Statistik lain.

MODEL TAK BERPARAMETER:

i) Histogram/pendisketan:

- Data diumpukkan kepada selang tertentu.
- Data disimpan dalam bentuk purata data selang.

ii) Pengkelompokan:

- Data dipartisikan kepada beberapa set kelompok berdasarkan ciri kesamaan yang wujud antara data.
- Data dalam kelompok yang sama mempunyai ciri yang hampir sama dan variasi yang kecil.
- Data antara kelompok yang berbeza tidak mempunyai ciri sama dan variasi yang besar.

iii) Pensampelan semula (butstrap):

- Beberapa sampel cerapan diambil dari set penuh data asal.
- Sampel-sampel cerapan tersebut dianggap mewakili set data asal

JENIS-JENIS PENSAMPELAN:

i)Pensampelan Rawak Lengkap: Setiap item dalam data set mempunyai kebarangkalian yang sama untuk dipilih.

ii) Pensampelan tanpa penggantian: Sebaik sahaja item data dipilih, ia akan dikeluarkan daripada set data asal.

iii) Pensampelan dengan penggantian: Item data yang terpilih, dimasukkan semula dalam set data asal. Ada kemungkinan untuk dipilih semula.

iv) Pensampelan Berstrata: Data dipartisikan kepada kumpulan (strata) tertentu berdasarkan sifat data. Seterusnya sampel rawak diambil daripada setiap strata.

- Lain-lain persampelan termasuklah Pensampelan Berkelompok, Pensampelan Sistematik, Pensampelan Multi-aras (*Multistage*) dan banyak lagi.

TOPIK SETERUSNYA:

Perlombongan Aturan Sekutuan