filkom.ub.id

DIGITAL TALENT SCHOLARSHIP 2019

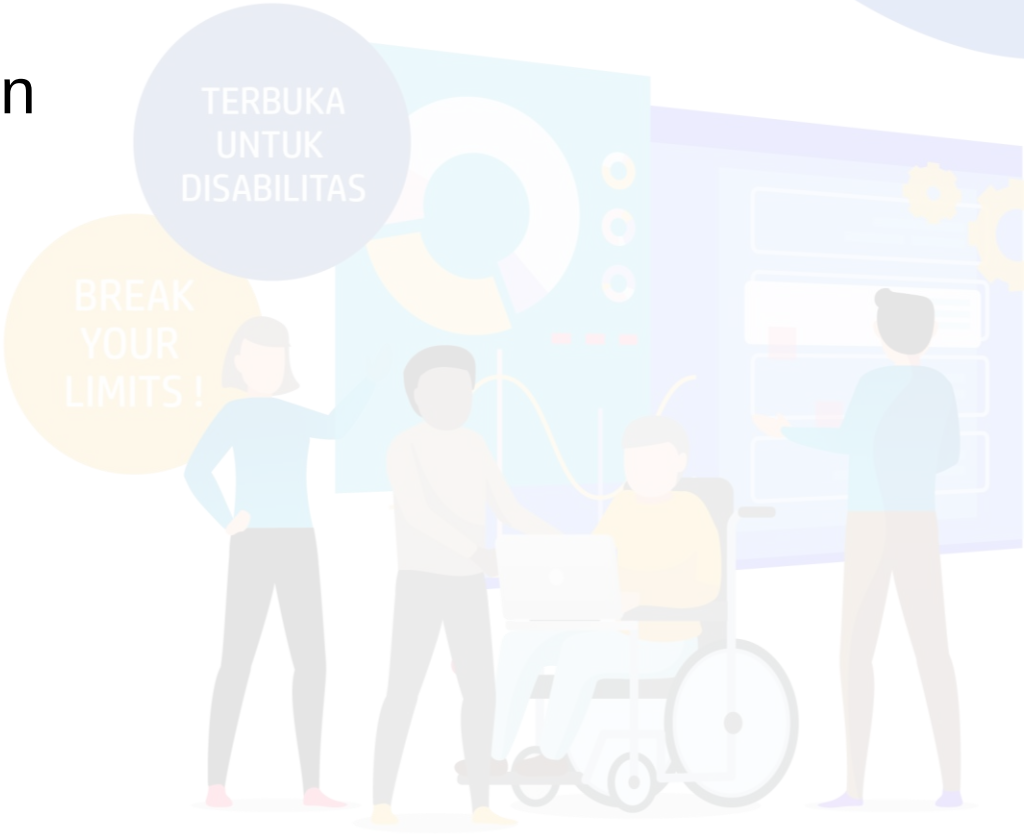
Big Data Analytics


# Association Rule, Clustering & Classification

Oleh: Imam Cholissodin | imamcs@ub.ac.id, Putra Pandu Adikara, Sufia Adha Putri
Asisten: Guedho, Sukma, Anshori, Aang dan Gusti
**Fakultas Ilmu Komputer (Filkom) Universitas Brawijaya (UB)**

# Pokok Pembahasan

- Association Rule
- Clustering & Classification
- Tugas



# Lecture's Objective

- Mempelajari Association Rule (Market Basket Analysis, dll.)
- Mempelajari pengelompokan obyek atau beberapa varibel berdasarkan kecenderungan yang ada pada data secara supervised dan unspervised.
- Peserta dapat memahami metode Association Rule, clustering dan classification dan menerapkan pada kasus yang tepat

# Association Rule

- Mencari suatu kaidah keterhubungan dari data
- Diusulkan oleh Agrawal, Imielinski, and Swami (1993)
- Contoh: Dalam suatu supermarket kita ingin mengetahui seberapa jauh orang yang membeli celana juga membeli sabuk?
- Input
  - Adanya sejumlah transaksi
  - Setiap transaksi memuat kumpulan item
- Problem
  - Bagaimana caranya menemukan association rule yang memenuhi minimum support dan minimum confidence yang kita berikan

# Association Rule: Manfaat

- Dapat digunakan untuk Market Basket Analysis (menganalisa kebiasaan customer dengan mencari asosiasi dan korelasi dari data transaksi)
  - Sebagai saran penempatan barang dalam supermarket
  - Sebagai saran produk apa yang dipakai dalam promosi

Database transaksi menyimpan data transaksi. Data transaksi bisa juga disimpan dalam suatu bentuk lain dari suatu database m x n.

# Association Rule: Definisi umum

- Itemset: himpunan dari item-item yang muncul bersama-sama
- Kaidah asosiasi: peluang bahwa item-item tertentu hadir bersama-sama.
- Support dari suatu itemset X (supp(X) ) adalah rasio dari jumlah transaksi dimana itemset muncul dengan total jumlah transaksi
- Konfidence (keyakinan) dari kaidah X→Y, ditulis conf(X → Y) adalah
  - conf(X→Y) = supp(X∪Y) / supp(X)
  - Konfindence bisa juga didefinisikan dalam terminologi peluang bersyarat

$$\text{conf}(X \rightarrow Y) = P(Y|X) = P(X \cap Y) / P(X)$$

- Lift adalah ratio dari nilai pengamatan *support* yang diharapkan jika dua aturan tersebut independen

$$\text{lift}(X \Rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X) \times \text{supp}(Y)}$$

# Association Rule: Contoh

| Transaksi | A | B | C | D |
|---|---|---|---|---|
| T1 | 1 | 0 | 1 | 14 |
| T2 | 0 | 0 | 6 | 0 |
| T3 | 1 | 0 | 2 | 4 |
| T4 | 0 | 0 | 4 | 0 |
| T5 | 0 | 0 | 3 | 1 |
| T6 | 0 | 0 | 1 | 13 |
| T7 | 0 | 0 | 8 | 0 |
| T8 | 4 | 0 | 0 | 7 |
| T9 | 0 | 1 | 1 | 10 |
| T10 | 0 | 0 | 0 | 18 |

Jumlah transaksi |D| = 10

Kemunculan item A pada transaksi (|Ta|) sebanyak 3 kali yaitu pada T1, T3, T8.

Supp(A)=|Ta|/|D| = 3/10 = 0.3.

|Tcd| sebanyak 5 kali, yaitu pada T1, T3, T5, T6, T9.
Supp(CD)=|Tcd|/|D| = 5/10 = 0.5.

- Frequent itemset adalah itemset yang mempunyai support >= minimum support yang diberikan oleh user.

# Association Rule: Contoh

| Itemset | Sp |
|---|---|
| A | 0.3 |
| B | 0.1 |
| C | 0.8 |
| D | 0.7 |
| AB | 0 |
| AC | 0.2 |
| AD | 0.3 |
| BC | 0.1 |
| BD | 0.1 |
| CD | 0.5 |
| ABC | 0 |
| ABD | 0 |
| ACD | 0.2 |
| BCD | 0.1 |
| ABCD | 0 |

Jika minsupport diberikan oleh user sebagai threshold adalah 0.2, maka frequent itemset adalah semua itemset yang support-nya >= 0.2, yakni

A, C, D, AC, AD, CD, ACD

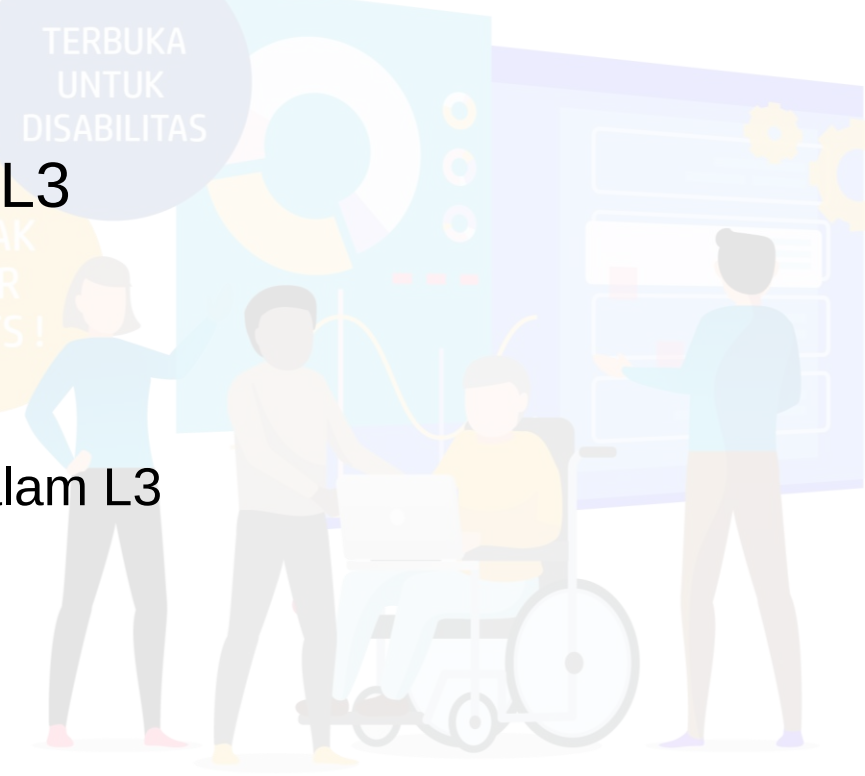
Dari frequent itemset bisa dibangun kaidah asosiasi sbb:

A → C  C → A
A → D  D → A
C → D  D → C
A,C → D  A,D → C  C,D → A

- Misal hitung berkut Conf(A→C) = supp(A,C) / supp(A)

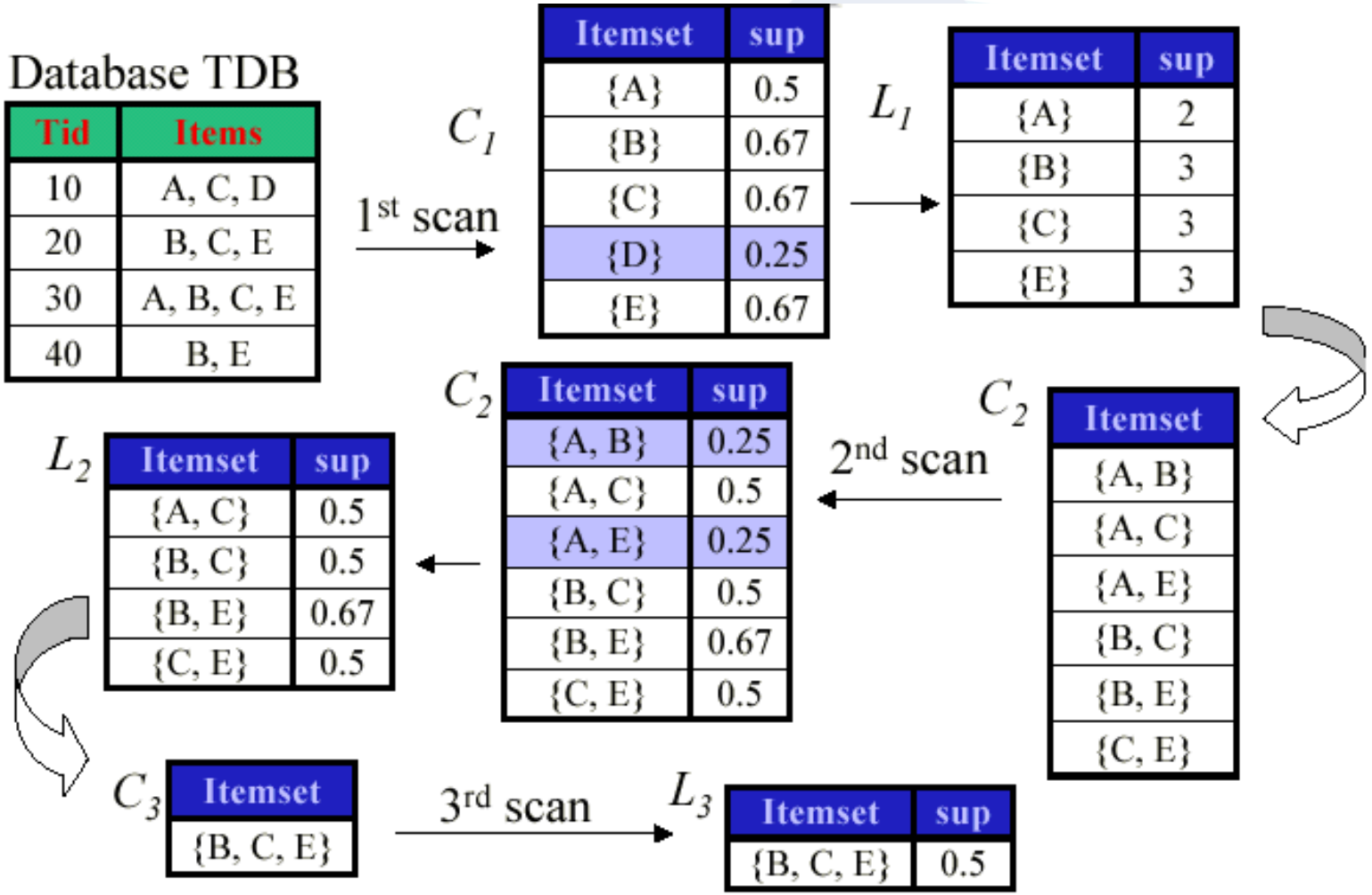
# Association Rule: Apriori

- Prinsip apriori : Subset apapun dari suatu frequent itemset harus frequent
- L3={abc, abd, acd, ace, bcd}
- Penggabungan sendiri : L3*L *L3
  - abcd dari abc dan abd
  - acde dari acd dan ace
- Pemangkasan Pemangkasan:
  - acde dibuang sebab ade tidak dalam L3
- C4={abcd}



# Association Rule: Apriori

- Contoh apriori dengan minimum support 50%



# Association Rule: Contoh 2

- Suatu supermarkat mempunyai sejumlah transaksi seperti dalam tabel

| | |
|---|---|
| T1 | {roti, selai, mentega} |
| T2 | {roti, mentega} |
| T3 | {roti, susu, mentega} |
| T4 | {coklat, roti} |
| T5 | {coklat, susu} |

a. Buatlah association rule dari data tersebut dengan cara menghitung support
b. Pakailah metode apriori dengan minimum support=0.3 dan **confidence=0.8**

Jawab.

| Itemset | Sp |
|---|---|
| {roti} | 0.8 |
| {selai} | 0.2 |
| {mentega} | 0.6 |
| {susu} | 0.4 |
| {coklat} | 0.4 |

⇨

| Itemset | Sp |
|---|---|
| {roti,mentega} | 0.6 |
| {roti,susu} | 0.2 |
| {roti,coklat} | 0.2 |
| {mentega,susu} | 0.2 |
| {mentega,coklat} | 0 |
| {susu,coklat} | 0.2 |

Conf(roti→mentega) =
Supp({roti,mentega})/Supp({roti})
= 0.6 / 0.8 = 0.75
= (75%)

Conf(mentega→roti) =
Supp({mentega,roti})/Supp({mentega})
= 0.6 / 0.6 = 1
**= (100%)**

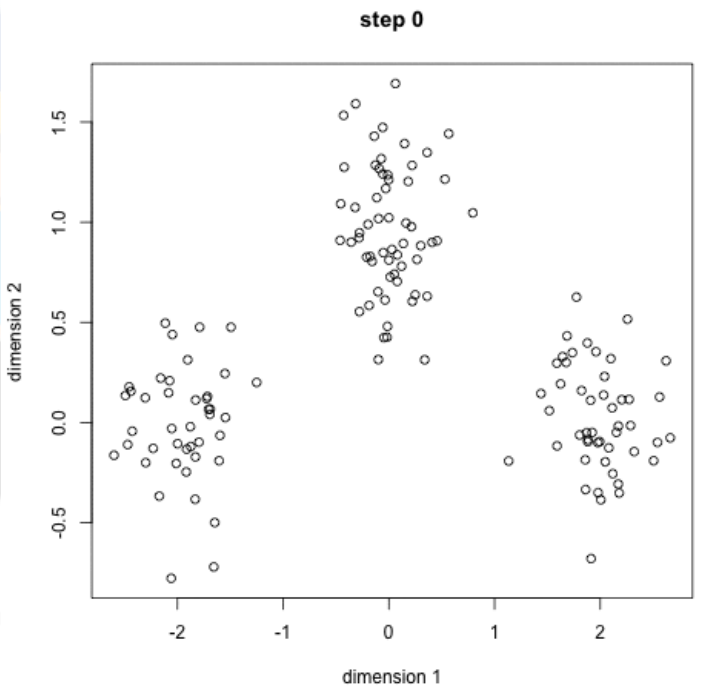
# *Clustering*

- ***Clustering*** adalah teknik *machine learning* berupa algoritma pengelompokkan objek-objek data berjumlah *N* menjadi kelompok-kelompok data tertentu (*cluster).*
- Objek data yang berada dalam satu kelompok / *cluster* harus memiliki kemiripan.
- Semakin banyak data yang diperoleh = Semakin akurat hasil yang didapatkan.
- *Clustering* merupakan salah satu jenis dari algoritma ***unsupervised learning***, algoritma yang bertujuan untuk mempelajari dan menemukan pola dari suatu input yang diberikan tanpa menggunakan label.
- Dengan penggunaan *supervised learning*, maka beberapa hal berikut ini dapat dilakukan:

  1. *Search*: Membandingkan antar dokumen, gambar atau suara untuk menampilkan *item* serupa.
  2. Deteksi anomali: Mendeteksi perilaku yang tidak biasa yang biasanya berhubungan dengan hal-hal yang ingin dicegah atau dideteksi, seperti contoh penipuan.

# K-Means *Clustering*

- Tentukan jumlah cluster
- Alokasikan data ke dalam cluster secara random
- Hitung centroid/rata-rata dari data yang ada di masing-masing cluster
- Alokasikan masing-masing data ke centroid/rata-rata terdekat
- Kembali ke Step 3, apabila masih ada data yang berpindah cluster atau apabila perubahan nilai centroid, ada yang di atas nilai threshold yang ditentukan atau apabila perubahan nilai pada objective function yang digunakan di atas nilai threshold yang ditentukan



# *Classification*

- Classification adalah teknik machine learning berupa algoritma pengklasifikasian objek-objek data ke dalam kelompok kelas yang telah ada.
- Pada classification, tidak akan ada pembentukan kelompok kelas baru.
- Classification merupakan salah satu jenis dari algoritma supervised learning, algoritma yang mempelajari korelasi antara sekumpulan input-output yang diinginkan dalam jumlah yang cukup besar menggunakan label.

# Algoritma *KNN*

1. **Nearest centroid**
   a) Menghitung centroid untuk setiap kelas
   b) Menghitung jarak antara test sample dan setiap kelas centroid
   c) Memprediksi kelas dengan metode centroid terdekat
2. **K-nearest neighbor**
   a) Dalam hal ini setiap data baru akan dibandingkan dengan data training.
   b) Lalu 3 data training terdekat (misalkan kita ambil k = 3) dengan data baru akan diambil.
   c) Misalkan ketiga data tersebut masuk ke dalam kelompok 1, 2 dan 1, maka data baru tersebut dimasukan ke dalam kelompok 1 (seperti voting, karena suara yang terbanyak adalah 1, maka keputusannya adalah 1).
   d) Menggunakan prediksi dengan majority vote dengan jumlah yang ganjil.

# Latihan langsung di Kelas Ke-1 & Pembahasan
# Link kode “http://bit.ly/2YBUU47”

*Silahkan dicoba dijalankan dengan Jupyter notebook yang Anda buat sebelumnya di Ubuntu 16.04 atau dengan SageMaker notebook (JupyterLab) yang baru Anda buat hari ini.*

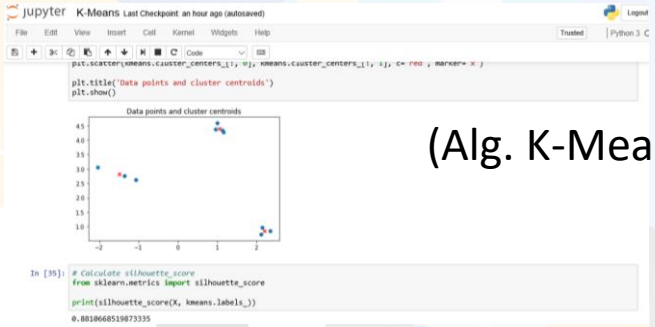
## Lab-Sesi28-1

Suatu supermarkat mempunyai sejumlah transaksi seperti dalam tabel

| | |
|---|---|
| T1 | {roti, selai, mentega} |
| T2 | {roti, mentega} |
| T3 | {roti, susu, mentega} |
| T4 | {coklat, roti, susu, mentega} |
| T5 | {coklat, susu} |

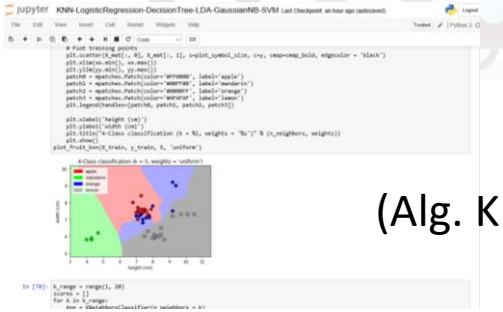
a. Buatlah association rule dari data tersebut dengan cara menghitung support
b. Pakailah metode apriori dengan minimum support=0.3 dan **confidence=0.8**

## Lab-Sesi28/29-2/1




(Alg. K-Means)

## Lab-Sesi28/29-3/2




(Alg. KNN, SVM & Others)

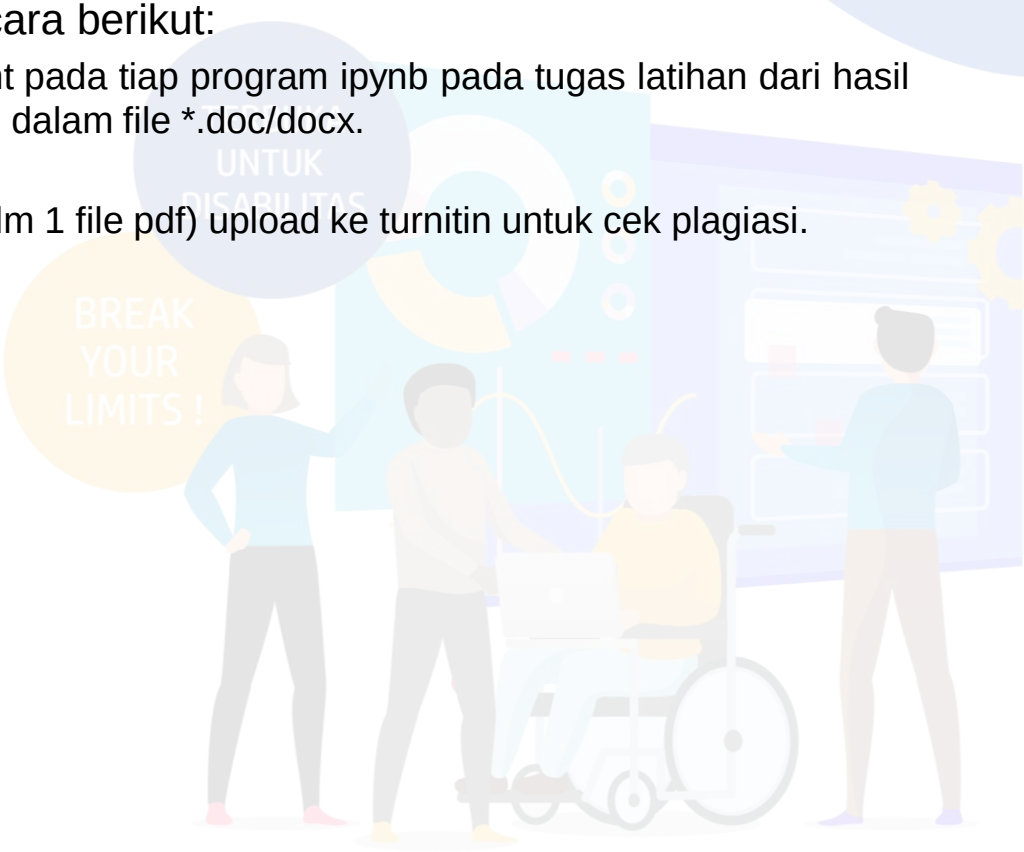
# Latihan langsung di Kelas Ke-2 & Pembahasan

- Tidak ada

# Tugas Individu

1. Buatlah rangkuman materi dengan cara berikut:
   - Menambahkan penjelasan/comment pada tiap program ipynb pada tugas latihan dari hasil “Latihan langsung di Kelas Ke-1” ke dalam file *.doc/docx.

*semua bentuk tugas tersebut (merger dlm 1 file pdf) upload ke turnitin untuk cek plagiasi.



filkom.ub.id

# DIGITAL TALENT SCHOLARSHIP 2019

## Big Data Analytics


# Terimakasih

Oleh: Imam Cholissodin | imamcs@ub.ac.id, Putra Pandu Adikara, Sufia Adha Putri
Asisten: Guedho, Sukma, Anshori, Aang dan Gusti
**Fakultas Ilmu Komputer (Filkom) Universitas Brawijaya (UB)**