

***CLUSTERING DATA DARI NILAI KELULUSAN***  
**UTBK JURUSAN IPA**



**Dosen Pengampu :**

I Made Sunia Raharja, S.Kom., M.Cs.

**Disusun Oleh :**

Ni Made Nadya Dewindra Wirata

2205551021

**PROGRAM STUDI TEKNOLOGI INFORMASI**  
**FAKULTAS TEKNIK**  
**UNIVERSITAS UDAYANA**  
**2023**

## LATAR BELAKANG

Kelulusan nilai UTBK (Ujian Tulis Berbasis Komputer) jurusan IPA berkaitan dengan penentuan kelulusan calon mahasiswa dan penerimaan mereka ke perguruan tinggi. Nilai UTBK jurusan IPA menjadi salah satu faktor penting yang dipertimbangkan dalam proses seleksi dan penentuan kelulusan. Pada umumnya, universitas atau perguruan tinggi menetapkan ambang batas nilai UTBK jurusan IPA yang harus dicapai oleh calon mahasiswa agar dapat diterima. Ambang batas ini dapat bervariasi antara perguruan tinggi yang satu dengan yang lain, tergantung pada kebijakan dan kebutuhan dari masing-masing perguruan tinggi.

Selain sebagai persyaratan kelulusan, nilai UTBK jurusan IPA juga dapat digunakan sebagai penilaian kompetensi calon mahasiswa dalam bidang ilmu alam. Hasil UTBK jurusan IPA dapat memberikan gambaran mengenai kemampuan calon mahasiswa dalam memahami dan menerapkan konsep-konsep ilmu pengetahuan alam seperti Matematika, Fisika, Kimia, dan Biologi. Namun, penting untuk dicatat bahwa nilai UTBK jurusan IPA bukan satu-satunya faktor penentu dalam kelulusan. Perguruan tinggi juga mempertimbangkan faktor lain seperti nilai rapor, tes kemampuan tertulis, tes wawancara, dan portofolio akademik atau non-akademik. Semua faktor ini bersama-sama digunakan untuk menilai kemampuan dan potensi calon mahasiswa serta memastikan kesesuaian mereka dengan program studi yang dituju.

Dengan menganalisis nilai kelulusan UTBK jurusan IPA bahwa dapat dilihat tujuan dari *clustering* data nilai kelulusan UTBK Jurusan IPA untuk mengetahui pengelompokan atau pengklasteran nilai kelulusan UTBK Jurusan IPA. Dari hasil ujian tersebut dapat dilihat juga satuan baku nilai, minimal nilai, dan maksimal nilai menjadi pertimbangan utama dalam menentukan kelulusan calon mahasiswa dan penerimaan mereka ke perguruan tinggi. Namun, keputusan akhir tetap bergantung pada kebijakan dan kriteria seleksi dari masing-masing perguruan tinggi.

## **TUJUAN DARI CLUSTERING**

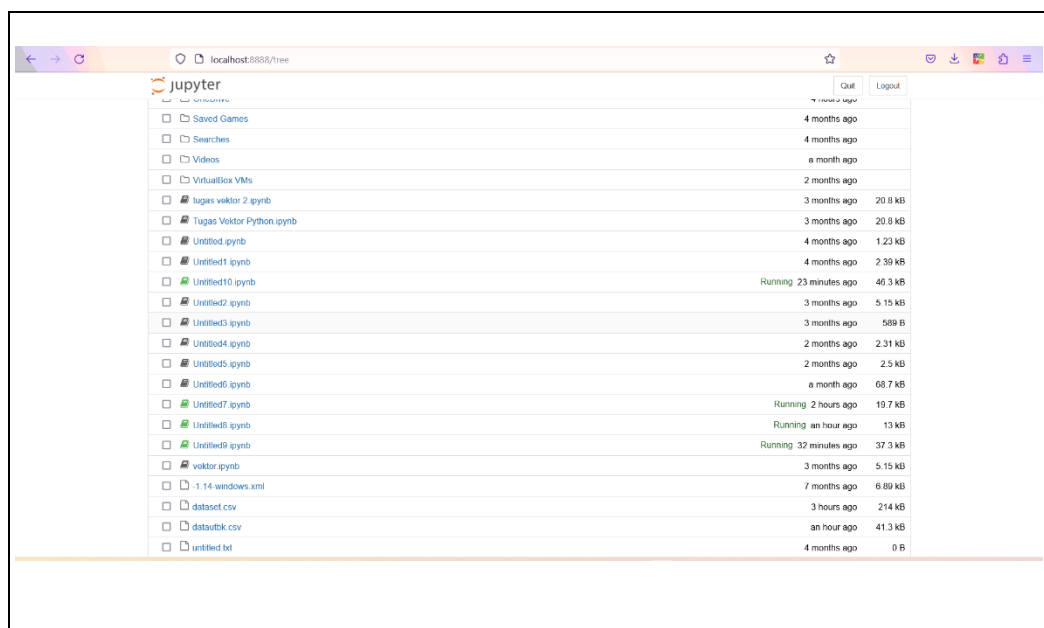
Tujuan dari clustering data kelulusan nilai UTBK (Ujian Tulis Berbasis Komputer) jurusan IPA adalah untuk mengelompokkan siswa berdasarkan kesamaan karakteristik atau pola pada nilai-nilai UTBK mereka. Clustering merupakan teknik analisis data yang digunakan untuk mengidentifikasi kelompok-kelompok yang serupa dalam kumpulan data, tanpa memiliki informasi sebelumnya tentang kelompok-kelompok tersebut. Dalam konteks ini, clustering dapat membantu mengungkap pola-pola atau grup-grup siswa yang memiliki performa serupa dalam UTBK jurusan IPA. Jadi Dapat dianalisis tujuan dari clustering data nilai kelulusan UTBK Jurusan IPA ini untuk mengetahui pengelompokan atau pengklasteran nilai kelulusan UTBK Jurusan IPA.

## ANALISIS DATA DAN PROSES CLUSTERING

Proses *clustering* data dari Nilai Kelulusan UTBK Jurusan IPA menggunakan bahasa pemrograman *Python* pada aplikasi *Jupyter Notebook* dengan metode algoritma *K-Means Clustering*. Dengan mengumpulkan data dari nilai UTBK Jurusan IPA calon mahasiswa yang melamar masuk ke perguruan tinggi. Kemudian data yang digunakan dalam studi kasus ini adalah data S.BAKU (Simpangan Baku Nilai) dan data MAX (Maksimum Nilai). Berikut merupakan proses atau tahapan yang dilakukan dalam aplikasi *Python Jupyter Notebook*.

### 1. Mengupload Dataset di Jupyter Notebook

Pertama-pertama membuka aplikasi pemrograman bahasa *Java* yang akan digunakan yakni aplikasi *Jupyter Notebook*. Kemudian mengupload *file datasheet* pada aplikasi *Jupyter Notebook*.



Gambar 1 Upload Dataset

Gambar di atas merupakan gambar tampilan awal atau *homepage* dari aplikasi *Jupyter Notebook*. Dan gambar diatas adalah proses *upload dataset* yaitu *datautbk*, dimana proses ini akan memudahkan dalam membaca *file dataset* saat proses *clustering*.

```
In [1]: #Tugas Ujian Akhir Semester 2
        #Nama : Ni Made Nadya Dewindra Wirata
        #NIM : 2205551021
        #Kelas : Aljabar Linear (F)

        #Menampilkan Data Clustering dari Nilai Kelulusan UTBK Jurusan IPA

In [2]: #Clustering adalah teknik dalam analisis data yang digunakan untuk mengelompokkan objek atau data ke dalam kelompok-kelompok
        #yang memiliki kesamaan tertentu. Dalam proses clustering, algoritma akan mencari pola-pola atau kesamaan antarobjek berdasarkan
        #atribut-atribut yang dimiliki. Objek yang memiliki atribut atau karakteristik yang serupa akan dikelompokkan ke dalam kelompok
        #yang sama, sedangkan objek yang memiliki atribut yang berbeda akan dikelompokkan ke dalam kelompok yang berbeda.
        #Tujuan utama dari clustering adalah mengidentifikasi pola atau struktur tersembunyi dalam data tanpa adanya label kelas
        #sebelumnya.
```

**Gambar 2** Membuat Project Identitas

Gambar di atas merupakan tampilan penulisan identitas yang berisi nama, nim, dan kelas di ikuti dengan penjelasan dan tujuan dari *clustering* pada *Jupyter Notebook*.

## 2. Import Library

Dalam *Python library* merupakan kumpulan modul terkait yang berisi kumpulan kode yang dapat digunakan berulang kali dalam program yang berbeda. Adanya *library* membuat pemrograman *python* menjadi lebih sederhana dan nyaman bagi programmer karena tidak perlu menulis kode yang sama berulang kali untuk program yang berbeda. *Library python* memainkan peran yang sangat vital dalam bidang pembelajaran mesin, data *science*, visualisasi data, aplikasi manipulasi gambar dan data, dan masih banyak lagi. Adapun *library* yang dimasukkan atau di-*import* pada studi kasus ini adalah sebanyak 4 jenis *library* seperti :

### 1. Pandas

Pandas (*Python for Data Analysis*) adalah *library Python* yang digunakan untuk melakukan manipulasi dan analisis data. Ini menyediakan struktur data yang efisien dan mudah digunakan, terutama dalam bentuk objek *Dataframe*. *Dataframe* adalah struktur data berbentuk tabel dengan baris dan kolom yang mirip dengan tabel dalam *database* atau *spreadsheet*. Pandas memadukan *library NumPy* yang memiliki kemampuan manipulasi data yang fleksibel dengan *database* relasional

### 2. Numpy

NumPy (*Numerical Python*) adalah *library Python* yang digunakan untuk melakukan komputasi numerik dan ilmiah. NumPy menyediakan objek array multidimensi yang efisien, bersama dengan kumpulan fungsi matematika yang

kuat, untuk memanipulasi dan mengoperasikan data numerik. Keunggulan NumPy array dibandingkan dengan *list* pada Python adalah konsumsi *memory* yang lebih kecil serta *runtime* yang lebih cepat. NumPy juga memudahkan kita pada Aljabar Linear, terutama operasi pada Vector (1-d *array*) dan Matrix (2-d *array*).

### 3. Kmeans

K-means adalah salah satu algoritma dalam analisis klustering (*clustering*) yang digunakan untuk mengelompokkan data ke dalam kelompok-kelompok yang serupa berdasarkan kemiripan fitur. Tujuan utama dari algoritma K-Means adalah untuk mencari pusat kelompok (*centroid*) yang optimal sehingga jarak antara titik data dalam kelompok tersebut dengan *centroid* minimal. Algoritma ini didesain untuk memungkinkan kita mengelompokkan data ke dalam group yang berbeda dengan cara yang lebih mudah berdasarkan variabel tertentu tanpa perlu melakukan proses training. Hal ini karena *k-means clustering* merupakan algoritma *unsupervised learning* berbasis *centroid*, yang dimana setiap cluster diasosiasikan dengan *centroid*. Tujuan utama dari algoritma ini adalah untuk meminimalkan jumlah jarak antara titik data dan cluster yang sesuai.

### 4. Matplotlib.pyplot

Matplotlib.pyplot adalah modul dalam library Matplotlib yang menyediakan fungsi dan metode untuk membuat plot grafik. Awalnya matplotlib dirancang untuk menghasilkan *plot* grafik yang sesuai pada publikasi jurnal atau artikel ilmiah. Matplotlib dapat digunakan dalam skrip Python, Python dan IPython *shell*, server aplikasi web, dan beberapa toolkit *graphical user interface* (GUI) lainnya. Fungsi ini berbasis pada Matplotlib, sebuah library yang kuat dan fleksibel untuk visualisasi data. Library ini menyediakan banyak fungsi dan metode untuk mengontrol tampilan grafik, termasuk sumbu, label, judul, warna, dan gaya garis.

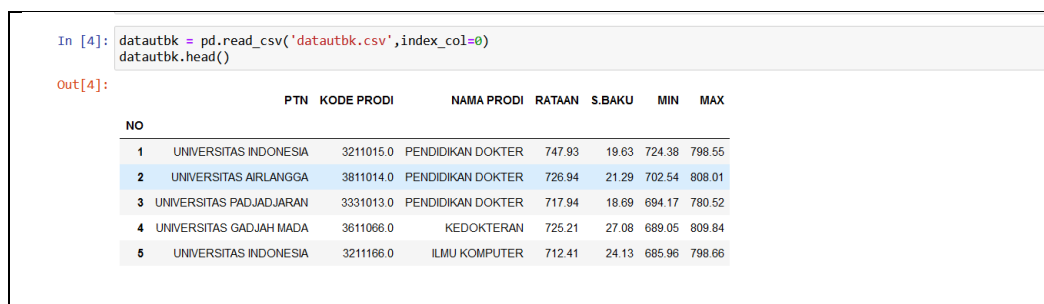
```
In [3]: #Import Library
import pandas as pd
import numpy as np
from sklearn.cluster import KMeans
from matplotlib import pyplot as plt
```

**Gambar 3** Import Library

Gambar di atas merupakan langkah yang dilakukan dalam proses *import library* ke dalam *Jupyter Notebook*. Pada langkah ini dilakukan *import library* berupa *library pandas, numpy, sklearn.cluster*, dan juga *matplotlib*.

### 3. Membaca File CSV ke dalam Dataset

Setelah *import library* berhasil selanjutnya adalah membaca *file dataset* yang telah di *import* kemudian membacanya dengan menggunakan sintaks seperti gambar berikut.



```
In [4]: dataautbk = pd.read_csv('dataautbk.csv', index_col=0)
dataautbk.head()
```

Out[4]:

	PTN	KODE PRODI	NAMA PRODI	RATAAN	S.BAKU	MIN	MAX
NO							
1	UNIVERSITAS INDONESIA	3211015.0	PENDIDIKAN DOKTER	747.93	19.63	724.38	798.55
2	UNIVERSITAS AIRLANGGA	3811014.0	PENDIDIKAN DOKTER	726.94	21.29	702.54	808.01
3	UNIVERSITAS PADJADJARAN	3331013.0	PENDIDIKAN DOKTER	717.94	18.69	694.17	780.52
4	UNIVERSITAS GADJAH MADA	3611066.0	KEDOKTERAN	725.21	27.08	689.05	809.84
5	UNIVERSITAS INDONESIA	3211166.0	ILMU KOMPUTER	712.41	24.13	685.96	798.66

**Gambar 4** Membaca File Dataset

Gambar diatas merupakan proses pembacaan *file dataset*. Dimana dalam proses tersebut pertama adalah memberikan nama *file* yaitu *dataautbk* yang nantinya akan digunakan untuk menyimpan data – data yang ada pada *file .csv* yang akan dibaca. Kemudian itu, menggunakan *method read\_csv()* yang digunakan untuk membaca *file dataset* CSV. Kemudian, menuliskan nama file *dataautbk* didalam kurung *method*, diikuti *method index\_col=0* yang digunakan untuk menentukan kolom yang harus dijadikan index dalam sebuah *data frame*. Lalu, *method data.head()* digunakan untuk menampilkan 5 data pertama.

### 4. Pemilihan Atribut Clustering

Selanjutnya Pemilihan Atribut Clustering dimana, pemilihan atribut ini yang nantinya akan digunakan serta ditampilkan pada proses *clustering*. Berikut adalah sintaks yang digunakan untuk melakukan proses ini.

```

In [7]: x = dataautbk[['S.BAKU', 'MIN']]
In [8]: x = x.values
In [9]: x = np.nan_to_num(x)

```

**Gambar 5** Pemilihan Atribut Clusterig

Gambar diatas merupakan gambar *method* yang digunakan untuk memilih atribut *clustering*. Dimana atribut yang akan di *clustering* adalah kolom file dataset S.BAKU menggunakan *method* `data[['S.BAKU', 'MIN']]` yang akan disimpan dalam variabel x. Kemudian, nilai pada kolom dataframe tersebut akan diubah menjadi nilai dalam *array* numpy menggunakan *method* `x.values` yang nantinya akan memudahkan proses *clustering*. Setelah itu menambahkan *method* `x = np.nan.to_num(x)` untuk Menggantikan nilai NaN dan *infinity* dengan nol.

## 5. Membuat Objek KMeans

Tahapan Pada langkah ini akan membuat sebuah objek K-Means dengan para meter tertentu yang kemudian model K-Means tersebut akan dilatih. Berikut adalah *method* yang digunakan untuk membuat objek K-Means ini.

```

In [10]: #Membuat Clustering menggunakan metode K-Means Clustering
kmeans = KMeans(n_clusters=3, random_state=42)
kmeans.fit(x)

Out[10]: KMeans(n_clusters=3, random_state=42)

```

**Gambar 6** Mebuat Objek K-Means

Gambar diatas adalah membuat objek KMeans `kmeans= KMeans(n_clusters=3, random_state=42)` KMeans adalah kelas dalam library scikit-learn yang digunakan untuk melakukan clustering dengan metode K-Means. `n_clusters=3` menentukan jumlah cluster yang ingin dibentuk. `random_state=42` digunakan untuk mengatur seed yang menentukan inisialisasi titik-titik pusat cluster secara acak. Kemudian melakukan *clustering* `kmeans.fit(x)`, `fit(x)` adalah metode dari objek `kmeans` yang digunakan untuk melakukan *clustering* pada data x. x adalah matriks data yang akan digunakan untuk *clustering*. Data ini harus dalam format *array* atau matriks dengan dimensi  $(n\_samples, n\_features)$ , di mana *n\_samples* adalah jumlah data dan *n\_features* adalah jumlah fitur dalam setiap data.



Dalam konteks ini, kode tersebut menciptakan objek `kmeans` dengan 3 cluster dan melakukan clustering pada data `x` dengan menggunakan metode K-Means. Hasil clustering akan tersimpan dalam objek `kmeans` untuk digunakan selanjutnya.

```
In [11]: #Untuk memperoleh label klaster
labels = kmeans.labels_

In [12]: #Untuk menambahkan label dengan nama Cluster ke dalam data
datautbk['Cluster'] = labels

In [13]: #Untuk menampilkan data hasil clustering
print(datautbk[['S.BAKU', 'MIN', 'Cluster']])
```

	S.BAKU	MIN	Cluster
NO			
1	19.63	724.38	0
2	21.29	702.54	0
3	18.69	694.17	0
4	27.08	689.05	0
5	24.13	685.96	0
...	...	...	...
496	16.02	582.23	2
497	13.04	582.14	2
498	17.46	582.05	2
499	16.76	581.97	2
500	13.54	581.92	2

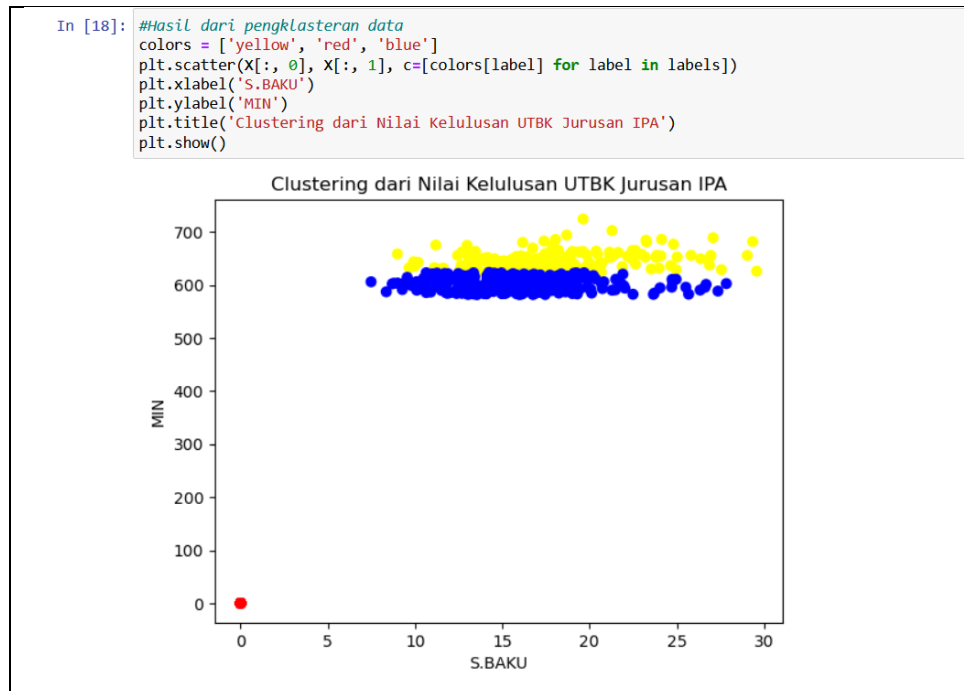
[500 rows x 3 columns]

**Gambar 7** Label Pada Dataset

Gambar diatas merupakan memberikan labels *cluster* yang dimana menggunakan *method* `labels = kmeans.labels` digunakan untuk memperoleh label klaster dari hasil clustering yang telah dilakukan menggunakan metode K-Means. Kemudian, diikuti dengan *method* `data['Cluster'] = labels` yang dimana digunakan untuk menambahkan kolom baru bernama '*Cluster*' ke dalam *data frame* diikuti dengan *method* `print(data[['S.BAKU', 'MIN', 'Cluster']])` digunakan untuk menampilkan data hasil clustering dari kolom S.BAKU, MIN, dan *Cluster* dalam dataframe `datautbk`.

## 6. Proses dan Hasil Clustering

Tahapan ini merupakan tahapan akhir yang dilakukan untuk mendapatkan data clustering dari studi kasus diatas dengan menggunakan metode *K-Means Clustering* dan penggunaan sintaks seperti gambar dibawah ini



**Gambar 8** Proses dan Hasil *Clustering*

Gambar diatas merupakan gambar dari proses dan hasil *clustering* Kelulusan Nilai UTBK Jurusan IPA. Dimana `Color` berfungsi untuk menentukan skema warna yang akan digunakan untuk memberikan warna pada *cluster*. Kemudian menggunakan *method* `plt.scatter` yang berfungsi untuk membuat *plot scatter* dengan sumbu X yang diambil dari kolom pertama dari variabel X dan sumbu Y yang diambil dari kolom kedua dari variabel Y. Kemudian, *parameter* `c=labels` digunakan untuk menentukan label *cluster* sebagai input pewarnaan dan *parameter*. Dilanjutkan dengan *method* `plt.xlabel()` dan `plt.ylabel` yang berfungsi untuk memberikan label pada sumbu x dan sumbu y. lalu, *method* `plt.title()` berfungsi untuk memberikan judul pada *plot*. Dan terakhir adalah *method* `plt.show()` yang digunakan untuk menampilkan *plot* yang telah dibuat.

Adapun Deskripsi dari hasil *clustering* diatas yaitu hasil nilai MIN yang tertinggi berada di rentang nilai 600 sampai 700 dan untuk nilai S.Baku nilai berada pada rentang nilai 10 sampai 30. Kemudian Dapat dilihat dari klaster diatas dimana kelompok klaster berwarna kuning menunjukkan data nilai kelulusan UTBK dengan nilai yang sangat baik. Kemudian kelompok klaster berwarna biru menunjukkan data nilai kelulusan UTBK yang baik dan yang terakhir adalah klaster yang berwarna merah menunjukkan nilai UTBK yang buruk.

## KESIMPULAN

Dapat ditarik kesimpulan dari data *clustering* diatas adalah teknik pengelompokan data yang bertujuan untuk mengidentifikasi pola atau kelompok yang ada dalam data. Dalam konteks data nilai kelulusan UTBK jurusan IPA, clustering dapat membantu mengelompokkan calon mahasiswa berdasarkan kesamaan atau perbedaan nilai mereka. Setelah itu pada data diatas juga terdapat *K-Means*. *K-Means* adalah salah satu metode clustering yang populer dan sederhana. Metode ini bekerja dengan mengelompokkan data ke dalam  $k$  klaster, di mana  $k$  merupakan jumlah klaster yang telah ditentukan sebelumnya. *K-Means* dapat membantu mengelompokkan calon mahasiswa berdasarkan nilai-nilai mereka. Dalam contoh kode yang diberikan, k-means clustering dengan 3 klaster digunakan untuk mengelompokkan data nilai kelulusan UTBK jurusan IPA menjadi 3 kelompok. Maka dari itu clustering dan penggunaan metode K-Means pada data nilai kelulusan UTBK jurusan IPA dapat membantu dalam mengelompokkan calon mahasiswa berdasarkan kesamaan atau perbedaan nilai mereka.