K-Means Clustering Program Studi Saintek di Universitas Gadjah Mada Berdasarkan Peminat SNMPTN Tahun 2021 dan Daya Tampung SNMPTN Tahun 2022

Ahmad Habib Hasan Zein[1, a)], Dinda Awanda Ramadhani[2, b)], Mozaya Zakiyah Anisa[3, c)], Mufrih Nur Huda Tri Putra[4, d)], Salsadila Aulia Fauzi[5, e)], Sindhi Mery Handayani[6, f)]

[1-9] Program Studi Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Gadjah Mada

Email :
[a)] ahmad.habib.hasan@mail.ugm.ac.id
[b)] dindaawanda@mail.ugm.ac.id
[c)] mozayazakiyah02@mail.ugm.ac.id
[d)] mufrih.n.h@mail.ugm.ac.id
[e)] salsadila.aulia@mail.ugm.ac.id
[f)] sindhimery3@mail.ugm.ac.id

ABSTRAK

Seleksi Nasional Masuk Perguruan Tinggi Negeri biasa disebut dengan SNMPTN merupakan salah satu bentuk jalur seleksi penerimaan mahasiswa untuk memasuki perguruan tinggi negeri yang dilaksanakan serentak seluruh Indonesia. Salah satu strategi mengikuti SNMPTN yang biasanya dilakukan adalah dengan mengetahui peluang masuk program studi di universitas tersebut. Dengan data daya tampung, peminat, dan banyak siswa/siswi yang diterima untuk setiap program studi kita dapat mengkategorikan program studi apa yang kiranya memiliki peluang besar dan kecil. Tujuan dari Penelitian ini adalah untuk mengategorikan program studi saintek di Universitas Gadjah Mada berdasarkan data peminat SNMPTN pada tahun 2021 dan juga daya tampung tahun 2022 pada program studi tersebut. Kategori yang dihasilkan berdasarkan data siswa peminat SNMPTN 2021 dan daya tampung SNMPTN 2022 dapat menghasilkan informasi mengenai siswa yang berminat untuk memilih jurusan yang diinginkannya agar peluangnya lebih besar. Data yang digunakan dalam penelitian ini adalah data yang diambil dari laman resmi Lembaga Tes Masuk Perguruan Tinggi (LTMPT) dengan cara web scraping. Data yang diperoleh dianalisis menggunakan metode analisis *clustering* dengan metode *K-means Clustering*, untuk menghasilkan beberapa kategori berdasarkan daya tampung 2022 dan peminat SNMPTN 2021. Dengan menggunakan metode ini, selanjutnya diperoleh 3 *cluster* yaitu *cluster* 0, *cluster* 1, dan *cluster* 2. *Cluster* 1 merupakan *cluster* program studi saintek dengan peminat dan daya tampung yang rendah. *Cluster* 0 merupakan *cluster* program studi saintek dengan peminat dan daya tampung sedang. Sedangkan, *cluster* 2 merupakan *cluster* program studi saintek dengan peminat dan daya tampung tinggi.

Kata kunci: K-Means, *Cluster*, Program Studi, Saintek, SNMPTN.

PENDAHULUAN

A. Latar Belakang

Di era modern ini, pendidikan merupakan suatu hal yang penting dan menjadi kebutuhan primer yang patut diperjuangkan. Tidak dapat dipungkiri persaingan memperoleh pendidikan yang terbaik menjadi hal yang wajar sehingga membuat kita harus mempersiapkan kualitas diri kita sebaik mungkin. Semakin berkualitas dan tinggi

jenjang pendidikan seseorang, maka semakin besar pula peluangnya untuk dapat diterima di dunia kerja.

Dekat ini, persaingan memasuki dunia perkuliahan jalur SNMPTN berlangsung cukup ketat. Setiap siswa/siswi yang sesuai kriteria nilai dan kuota yang diberikan sekolah memiliki kesempatan yang sama untuk diterima. Mereka memiliki teknik dan strateginya masing-masing untuk mendapatkan program studi dan universitas yang didambakannya. Salah satu strategi yang biasanya dilakukan adalah dengan mengetahui peluang masuk program studi di universitas tersebut. Dengan data daya tampung, peminat, dan banyak siswa/siswi yang diterima untuk setiap program studi kita dapat mengkategorikan program studi apa yang kiranya memiliki peluang besar dan kecil. Pengelompokkan program studi saintek di salah satu universitas berdasarkan peminat SNMPTN tahun 2021 dan daya tampung SNMPTN tahun 2022 pada analisis ini dilakukan menggunakan analisis clustering, salah satu metodenya adalah *K-Means Clustering*.

Clustering merupakan metode penganalisaan data, yang sering kali dimasukkan sebagai salah satu metode Data Mining, dengan tujuan utama yaitu untuk mengelompokkan objek-objek data berdasarkan karakteristik yang dimiliki. Pengelompokkan objek didasari pada kemiripan ciri-ciri umum antar objek, yang mana objek yang ada pada setiap kelompok atau cluster akan memiliki kemiripan satu sama lain. Analisis cluster biasa digunakan pada berbagai bidang ilmu seperti psikologi, sosiologi, biologi, ekonomi, bisnis, dan lain-lain. Selain digunakan pada berbagai bidang ilmu, dan yang lainnya. Ada beberapa pendekatan yang digunakan dalam mengembangkan metode clustering. Berdasarkan karakteristik dari algoritmanya, terdapat 4 jenis teknik clustering, yaitu *partitioning methods, hierarchical methods, density-based methods*, dan *grid-based method.*

B. Tujuan dan Manfaat

Penelitian ini dilakukan untuk mengategorikan program studi saintek di Universitas Gadjah Mada berdasarkan data peminat SNMPTN pada tahun 2021 dan juga mengkategorikan daya tampung di program studi saintek. Kategori yang dihasilkan berdasarkan data siswa peminat SNMPTN 2021 dan daya tampung SNMPTN 2022 dapat menghasilkan informasi mengenai siswa yang berminat untuk memilih jurusan yang diinginkannya agar peluangnya lebih besar. Setelah dilakukan analisis nanti diharapkan dapat memberikan manfaat bagi pembaca untuk mengetahui tingkat akurasi algoritma K-Means dalam memprediksi ketepatan peminat suatu jurusan di Universitas Gadjah Mada dan daya tampungnya sehingga dapat melihat seberapa ketat persaingan SNMPTN untuk masuk program studi saintek di Universitas Gadjah Mada.

DASAR TEORI

Analisis kluster merupakan metode dalam membagi rangkaian data menjadi beberapa kelompok berdasarkan kesamaan. Terdapat beberapa manfaat *clustering*, yakni untuk eksplorasi data, reduksi data, dan pelapisan data. Eksplorasi data dilakukan untuk memperoleh gambaran dari data; reduksi data untuk mewakili seluruh anggota cluster dengan suatu ringkasan *cluster*; dan hasil analisis *cluster* dapat digunakan sebagai pelapisan atau stratifikasi objek.

Pada analisis kluster, berdasarkan karakteristik algoritmanya dibagi menjadi 4 metode yaitu, *partitioning methods, hierarchical methods, density-based methods*, dan *grid-based method*. Penelitian kali ini menggunakan metode *partitioning*. Pada metode partisi, terdapat 2 jenis algoritma, yaitu *K-means Clustering* dan *Partitioning Around*

Medoids atau biasa disebut dengan *K-Medoids*. Dalam penelitian ini menggunakan jenis algoritma *K-means* sebagai solusi untuk pengklasifikasian karakteristik dari objek. Alasan penggunaan algoritma *K-Means* karena algoritma ini memiliki ketelitian yang cukup tinggi terhadap ukuran objek sehingga relatif lebih terukur dan efisien untuk pengolahan objek dalam jumlah besar. Selain itu, algoritma *K-Means* ini tidak terpengaruh terhadap urutan objek.

K-Means adalah salah satu algoritma dari teknik data mining yang mampu melakukan klasterisasi terhadap data heterogen karena pada dasarnya algoritma pengelompokkan hanya mampu mengenali nilai atribut homogen saja. Algoritma *K-Means Clustering* akan memilih sebanyak k titik awal *centroid* secara acak dan melakukan perhitungan jarak masing-masing observasi ke masing-masing *centroid* tersebut dengan menggunakan *Euclidean Distance*. Dari perhitungan jarak tersebut, akan diambil jarak yang paling dekat dari masing-masing observasi untuk menentukan masuk ke cluster manakah observasi tersebut.

Misal, diberikan sekumpulan objek x = (x_1 ,x_2, …., x_n) maka algoritma *K-Means Cluster Analysis* akan mempartisi x dalam sebanyak k *cluster*, setiap cluster memiliki centroid dari objek-objek dalam cluster tersebut. Awalnya, algoritma *K-Means Cluster Analysis* dipilih secara acak k buah objek sebagai *centroid*, kemudian jarak antara objek dengan *centroid* dihitung dengan menggunakan jarak euclidean, objek ditempatkan dalam *cluster* yang terdekat dihitung dari titik tengah *cluster*. Jika semua objek sudah ditempatkan dalam *cluster* terdekat maka, barulah *centroid* ditetapkan. Penentuan *centroid* dan penempatan objek dalam *cluster* diulangi sampai nilai *centroid* dari semua *cluster* tidak berubah lagi. Secara umum metode *K-Means Cluster Analysis* menggunakan algoritma sebagai berikut :

1. Tentukan k sebagai jumlah *cluster* yang di bentuk.
   Untuk menentukan banyaknya *cluster* k dilakukan dengan beberapa pertimbangan seperti pertimbangan teoritis dan konseptual yang mungkin diusulkan untuk menentukan berapa banyak *cluster*.
2. Bangkitkan k *Centroid* (titik pusat cluster) awal secara random.
   Penentuan centroid awal dilakukan secara random/acak dari objek-objek yang tersedia sebanyak k *cluster*, kemudian untuk menghitung *centroid cluster* ke-i berikutnya, digunakan rumus sebagai berikut :

$$v = \frac{\sum_{i=1}^{n} x_i}{n} \quad ;i = 1,2,3,...n$$

   dimana;
   v : *centroid* pada cluster
   x_i : objek ke-i
   n : banyaknya objek/jumlah objek yang menjadi anggota *cluster*
3. Hitung jarak setiap objek ke masing-masing *centroid* dari masing-masing *cluster*.
   Untuk menghitung jarak antara objek dengan *centroid* penulis menggunakan *Euclidean Distance*.

$$d(x,y) = \|x - y\| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad ;i = 1,2,3,..., n$$

   dimana;
   x_i : objek x ke-i
   y_i : data y ke-i
   n : banyaknya objek

4. Alokasikan masing-masing objek ke dalam *centroid* yang paling terdekat. Untuk melakukan pengalokasian objek kedalam masing-masing *cluster* pada saat iterasi secara umum dapat dilakukan dengan dua cara yaitu dengan hard k-means, dimana secara tegas setiap objek dinyatakan sebagai anggota *cluster* dengan mengukur jarak kedekatan sifatnya terhadap titik pusat *cluster* tersebut, cara lain dapat dilakukan dengan *fuzzy C-Means*.
5. Lakukan iterasi, kemudian tentukan posisi *centroid* baru dengan menggunakan persamaan (1).
6. Ulangi langkah 3 jika posisi *centroid* baru tidak sama

METODOLOGI PENELITIAN

A. Tentang Data

Untuk dapat melakukan pengelompokkan program studi saintek pada SNMPTN 2021 di Universitas Gadjah Mada berdasarkan seberapa ketat persaingan untuk masuk ke program studi yang ada, diperlukan data yang dapat digunakan untuk proses analisis. Pengambilan data diambil dengan cara *web scraping*. *Web scraping* dapat didefinisikan sebagai proses pengambilan data dari sebuah *website.* Untuk itu, data yang digunakan perlu diperoleh dari sumber yang dapat dipercaya.

Pada proses analisis data kali ini, data diambil dari laman resmi Lembaga Tes Masuk Perguruan Tinggi (LTMPT). LTMPT adalah lembaga pendidikan di bawah Kementrian Pendidikan, Budaya, Riset, dan Teknologi (Kemendikbud Ristek) yang bertugas menyelenggarakan tes untuk masuk perguruan tinggi bagi para calon mahasiswa. Adapun LTMPT ini bagian dari serangkaian alur yang harus diikuti calon mahasiswa baru yang ingin masuk perguruan tinggi (PTN). Laman yang digunakan untuk

Dari *web scraping* yang telah dilakukan, diperoleh data SNMPTN 2021 yang terdiri dari nama-nama program studi pada kelompok saintek yang berjumlah 61, masing-masing jenjang, daya tampung 2022, peminat 2021, serta jenis portofolio.

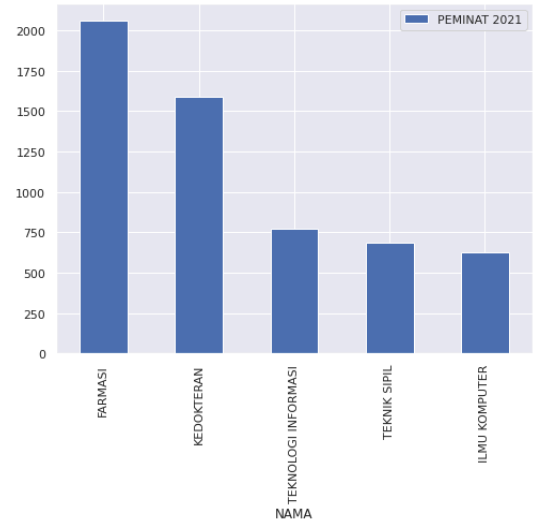
B. Metode yang Digunakan

Setelah memperoleh data dari LTMPT, serta sesuai dengan tujuan dari analisis kali ini yaitu untuk mengetahui pengelompokan program studi saintek pada SNMPTN 2021 di Universitas Gadjah Mada berdasarkan seberapa ketat persaingan untuk masuk ke program studi yang ada, maka analisis yang dilakukan adalah analisis *clustering* dengan metode *K-means clustering*. Hal ini dikarenakan tidak banyaknya *outlier* pada data. Lebih lanjut mengenai analisis data dan pembahasan terhadap analisis data dijelaskan pada bagian Analisis dan Pembahasan.

ANALISIS DAN PEMBAHASAN

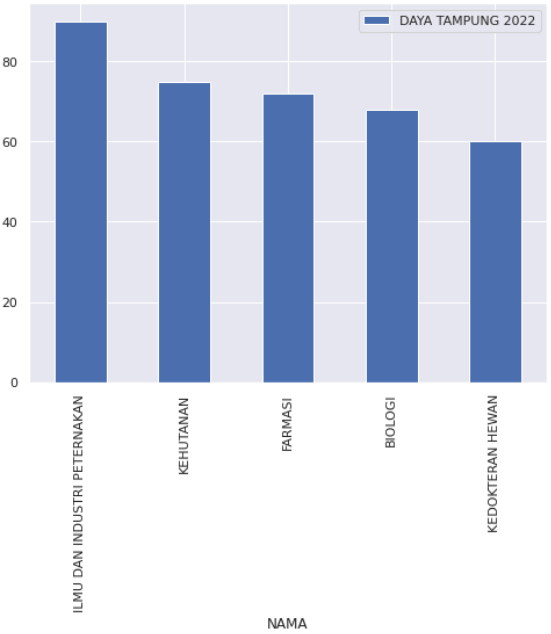
A. Analisis Data Eksploratif

Tahapan pertama yang dilakukan sebelum melakukan analisis adalah eksplorasi data. Hal ini dilakukan bertujuan untuk mengetahui gambaran umum dari data yang akan dianalisis. Eksplorasi data dilakukan dengan membentuk diagram batang untuk kolom peminat dan kolom daya tampung program studi saintek pada SNMPTN 2021. Selain itu, dibentuk juga *scatter plot* antara peminat dan daya tampung untuk melihat pola hubungan di antara kedua kolom tersebut.



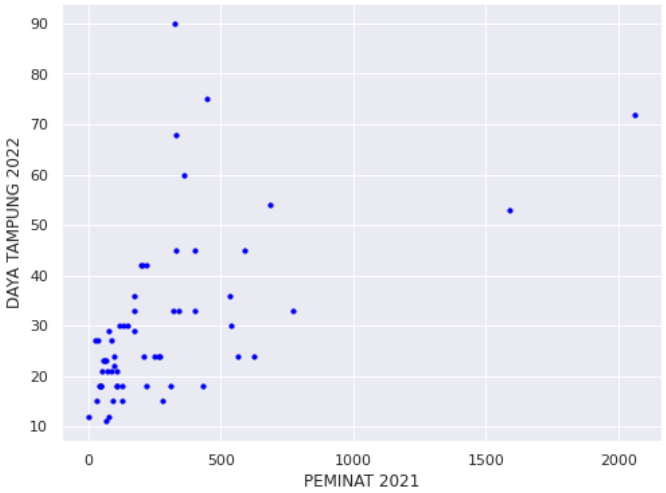
Gambar 4.1 Program Studi Saintek dengan Peminat Jalur SNMPTN Tertinggi Tahun 2021

Berdasarkan diagram batang di atas diperoleh informasi lima program studi saintek dengan peminat tertinggi pada SNMPTN tahun 2021. Lima program studi tersebut secara berurutan adalah farmasi dengan peminat sebanyak 2061 orang, kedokteran sebanyak 1588 orang, teknologi informasi sebanyak 771 orang, teknik sipil sebanyak 684 orang, dan ilmu komputer sebanyak 626 orang.



Gambar 4.2 Program Studi Saintek dengan Daya Tampung Jalur SNMPTN Tertinggi Tahun 2022

Berdasarkan diagram batang di atas diperoleh informasi lima program studi saintek dengan daya tampung tertinggi pada SNMPTN tahun 2021. Lima program studi tersebut secara berurutan adalah ilmu dan industri peternakan sebanyak 90 kursi, kehutanan sebanyak 75 kursi, farmasi sebanyak 72 kursi, biologi sebanyak 68 kursi, dan kedokteran hewan sebanyak 60 kursi.

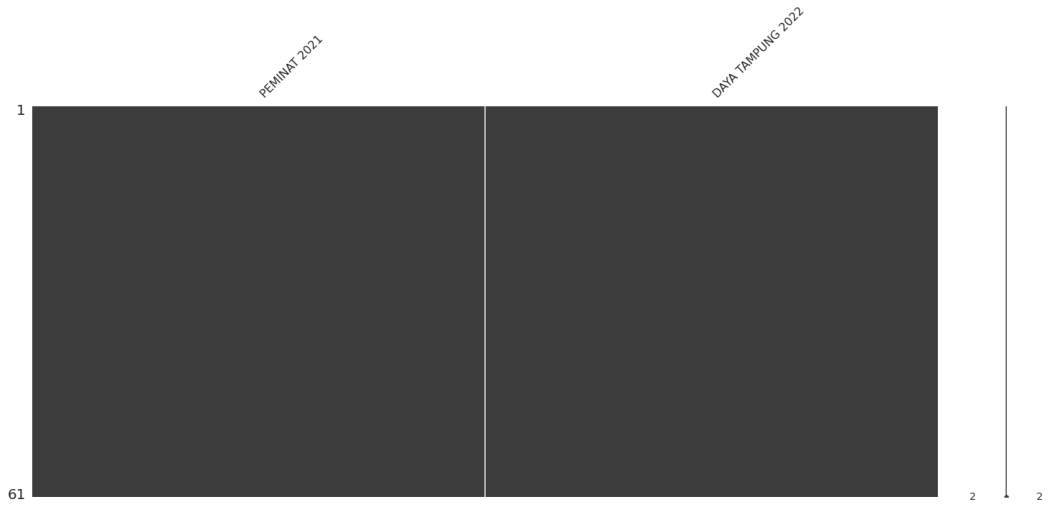


Gambar 4.3 Scatter Plot antara Peminat 2021 dengan Daya Tampung 2022

Berdasarkan scatter plot di atas terlihat bahwa antara peminat dan daya tampung saintek pada SNMPTN 2021 tidak saling berhubungan secara linear. Hal tersebut terlihat dari persebaran titik-titik terletak menumpuk di pojok kiri bawah dan tidak tersebar dari pojok kiri bawah ke pojok kanan atas. Untuk menguatkan argumen tersebut, maka perlu dilakukan uji korelasi antara kedua variabel, yaitu daya tampung 2022 dengan peminat 2021.

B. Pengecekan *Missing Value* dan *Outlier*

Sebelum melakukan analisis, penting untuk melakukan pengecekan missing value dan outlier



Gambar 4.4 *Matrix* Pengecekan Missing Value

Dari plot matrix diatas didapatkan hasil bahwa tidak ada missing value dari variabel peminat 2021 dan daya tampung 2022. Hal ini ditunjukan dengan box hitam yang berdekatan tidak garis penghubung antara 2 variabel tersebut.

Selain melalui plot di atas, kita juga dapat mencari tahu ada tidaknya *missing value* dengan menjalankan perintah `print(" \nCount total NaN at each column in a DataFrame : \n\n",mydata.isnull().sum())`. Dengan menjalankan perintah ini, diperoleh output sebagai berikut:

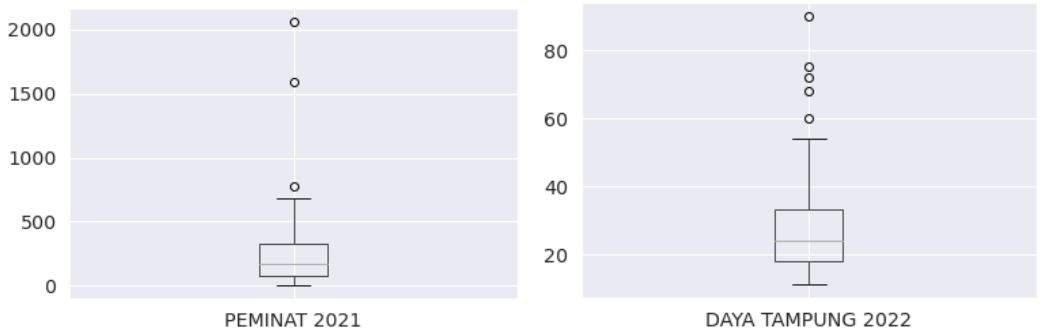
```
Count total NaN at each column in a DataFrame :

 NO                   0
KODE                  0
NAMA                  0
JENJANG               0
DAYA TAMPUNG 2022     0
PEMINAT 2021          0
JENIS PORTOFOLIO      0
dtype: int64
```

Gambar 4.5 Output Pengecekan Missing Value

Dari output diatas didapatkan informasi bahwa untuk variabel NO, KODE, NAMA, JENJANG, DAYA TAMPUNG 2022, PEMINAT 2021, JENIS PORTOFOLIO, tidak memiliki *missing value* atau berarti tidak terdapat *missing value* pada data yang akan kami gunakan untuk penelitian kali ini.

Pengecekan *outlier* dilakukan dengan melihat boxplot dari masing-masing variabel, diperoleh *output* sebagai berikut



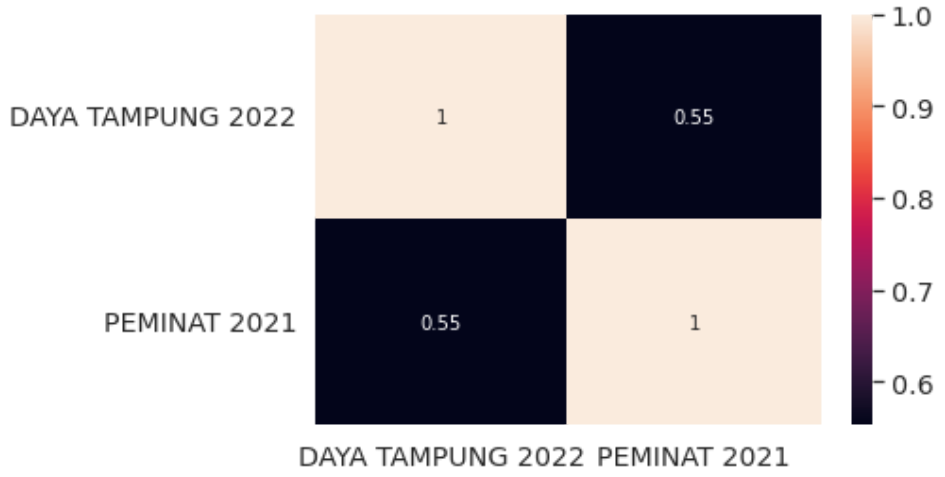
Gambar 4.6 dan 4.7 Boxplot Peminat 2021 dan Daya Tampung 2022

Dari output diatas didapatkan bahwa untuk variable PEMINAT 2021, bentuk boxplot yang menjurai ke atas dimana median atau garis tengah boxplot lebih dekat dengan kuartil 1 dan terdapat 3 outlier pada nilai tinggi (di atas kuartil 3). Dan untuk variabel Daya Tampung 2022, bentuk boxplot yang menjurai ke atas dimana median atau garis tengah boxplot lebih dekat dengan kuartil 1 dan terdapat 4 *outlier* pada nilai tinggi (di atas kuartil 3). Karena jumlah *outlier* yang sedikit maka dapat digunakan *K-Means Clustering* untuk pengujian kali ini.

C. Pengujian Asumsi

Dalam analisis kluster terdapat dua asumsi yang harus dipenuhi, yaitu:

1. Kecukupan Sampel
   Sampel yang dimiliki harus bisa merepresentasikan populasi. Pada penelitian ini data yang digunakan merupakan populasi itu sendiri, maka asumsi ini secara otomatis terpenuhi.
2. No Multikolinieritas
   Uji asumsi no multikolinieritas dapat dilakukan dengan beberapa cara, salah satunya adalah dengan melihat nilai koefisien korelasi antar variabelnya. Nilai koefisien korelasi antara variabel Peminat 2021 dengan Daya Tampung 2022 dapat dilihat pada gambar berikut.



Gambar 4.8 *Heatmap* Korelasi Peminat 2021 dengan Daya Tampung 2022

Dari *heatmap* diatas didapatkan informasi, koefisien korelasi antara variabel PEMINAT 2021 dengan variabel DAYA TAMPUNG 2022 sebesar 0.55. korelasi antara keduanya rendah. Maka dari itu tidak terdapat multikolinearitas dalam variabel yang akan diuji. Selain dengan heatmat diatas kami juga menguji asumsi no multikolinearitas dengan *Variance Inflation Factor* (VIF).

	feature	VIF
0	DAYA TAMPUNG 2022	2.338802
1	PEMINAT 2021	2.338802

Gambar 4.9 *Variance Inflation Factor* setiap variabel

Variance Inflation Factor (VIF) dari variabel PEMINAT 2021 dan variabel DAYA TAMPUNG 2022 sebesar 2.338802. Nilai *Variance Inflation Factor* (VIF) dari masing masing variabel kurang dari 10, multikolinearitas terjadi jika nilai VIF lebih dari 10 (Sarwoko dalam Zaenuddin, 2015). Maka tidak terdapat multikolinearitas antar variabel. Asumsi no multikolinearitas terpenuhi.

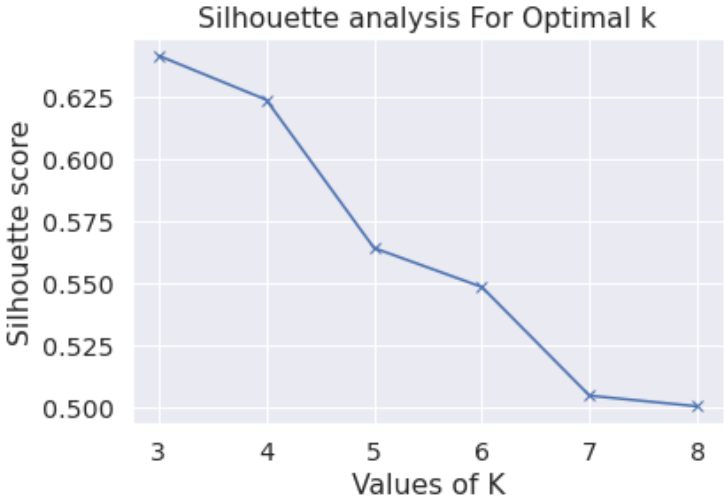
D. Penentuan Jumlah *Cluster*

Penentuan jumlah *cluster* yang paling optimal dapat dilakukan dengan validasi hasil cluster. Pada penelitian ini, validasi dilakukan dengan menggunakan metode validasi Silhouette Index.

```
3
Silhouetter Score: 0.642
4
Silhouetter Score: 0.624
5
Silhouetter Score: 0.564
6
Silhouetter Score: 0.548
7
Silhouetter Score: 0.505
8
Silhouetter Score: 0.500
```

Gambar 4.10 *Silhouette Score* dari Beberapa Jumlah *Cluster*

Dari gambar diatas didapatkan informasi, nilai koefisien *Silhouette* terbesar yang berada pada antara jumlah *cluster* 3 sampai dengan 8 adalah *cluster* 3. Maka *cluster* yang akan digunakan dalam metode *K-Means Clustering* ini adalah 3. Selain itu jumlah *cluster* terbaik dapat diketahui dengan plot *Silhouette* dengan jumlah *cluster*. Dimana semakin tinggi koefisien *Silhouette* semakin baik jumlah *cluster* yang digunakan



Gambar 4.11 *Silhouette Analysis for Optimal k*

Sama halnya seperti Gambar 4.10, pada Gambar 4.11 dapat dilihat nilai-nilai *Silhouette Score* antara jumlah *cluster* sebesar 3 sampai dengan 8. Dari plot yang terbentuk, terlihat bahwa jumlah *cluster* yang memiliki nilai *Silhouette Score* terbesar adalah 3. Oleh karena itu, jumlah *cluster* yang akan dibentuk dalam penelitian ini adalah sebanyak 3 *cluster*.

E. Analisis *Cluster*

Pada penelitian ini, algoritma *k-means* digunakan untuk *clustering*. Algoritma k-means merupakan algoritma yang bertujuan untuk mengelompokkan data menjadi beberapa cluster berdasarkan jarak terdekat. Analisis *cluster* dilakukan dengan jumlah *cluster* sebanyak 3 *cluster*, diperoleh hasil sebagai berikut.

Tabel 4.1 Jumlah Data pada Setiap *Cluster*

Cluster	Jumlah Data
0	18
1	41
2	2

Berdasarkan tabel di atas, diketahui bahwa *cluster* 0 terdiri dari 18 program studi, *cluster* 1 terdiri atas 41 program studi, dan *cluster* 2 terdiri dari 2 program studi. Adapun anggota dari masing-masing *cluster* dapat dilihat pada tabel berikut.

Tabel 4.2 Anggota Tiap-tiap *Cluster*

	Cluster 0	*Cluster 1*	*Cluster 2*
Anggota (Kode Prodi)	3611526, 3611012,	3611534, 3611542,	3611027 dan

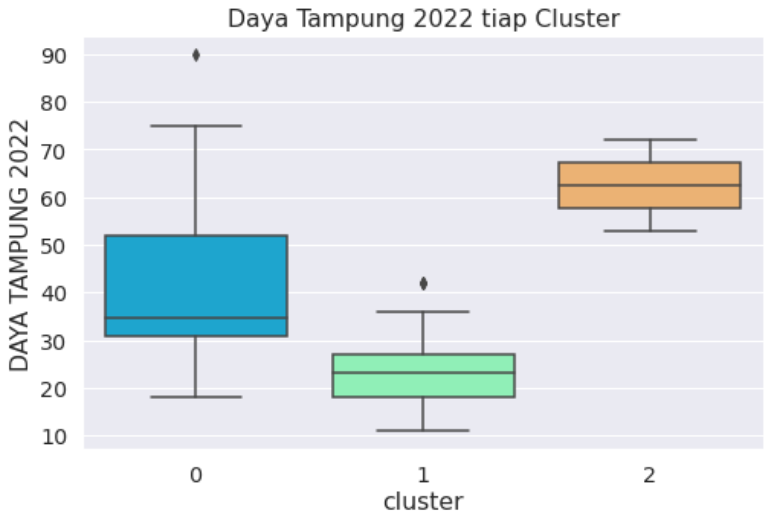
	3611074, 3611082, 3611097, 3611101, 3611186, 3611194, 3611325, 3611333, 3611372, 3611387, 3611395, 3611406, 3611437, 3611453, 3611484, dan 3611492	3611557, 3611565, 3611573, 3611581, 3611596, 3611607, 3611615, 3611623, 3611631, 3611646, 3611654, 3611035. 3611043, 3611051, 3611155, 3611163, 3611171, 3611205, 3611213, 3611221, 3611221, 3611244, 3611252, 3611267, 3611275, 3611283, 3611291, 3611302, 3611317, 3611341, 3611356, 3611364, 3611414, 3611422, 3611445, 3611461, 3611476, 3611503, 3611511, dan 3611662.	3611066

Untuk mengetahui karakteristik dari tiap *cluster* yang terbentuk, akan dibuat visualisasi dengan menggunakan boxplot. Dibentuk dua boxplot yaitu berdasarkan Peminat Tahun 2021 dan Daya Tampung Tahun 2022. Berikut merupakan boxplot untuk tiap *cluster* berdasarkan Peminat Tahun 2021

Gambar 4.12 Boxplot tiap *Cluster* Berdasarkan Peminat 2021

Dari boxplot di atas terlihat bahwa *cluster* 2 merupakan *cluster* program studi dengan peminat pendaftar SNMPTN Tahun 2021 tertinggi. Sedangkan, *cluster* 0 merupakan *cluster* dengan peminat pendaftar SNMPTN Tahun 2021 sedang dan *cluster* 1 merupakan *cluster* dengan peminat pendaftar SNMPTN Tahun 2021 terendah. Dengan demikian dapat disimpulkan bahwa program studi yang masuk *cluster* 2 merupakan program studi yang sangat diminati pada SNMPTN Tahun 2021.

Selain berdasarkan Peminat Tahun 2021, akan dibandingkan pula ketiga *cluster* yang terbentuk berdasarkan variabel Daya Tampung Tahun 2022



Gambar 4.13 Boxplot tiap *Cluster* Berdasarkan Daya Tampung 2022

Jika dibandingkan dengan boxplot pada Gambar 4.12, boxplot pada Gambar 4.13 tidak terlalu jelas perbedaan antar *cluster* yang terbentuk. Hal ini mengindikasikan bahwa perbedaan daya tampung antar *cluster* tidak terlalu berbeda. Meskipun demikian, dapat dilihat bahwa *cluster* 2 merupakan *cluster* yang daya tampungnya relatif lebih tinggi dibandingkan kedua *cluster* lainnya. Sedangkan, *cluster* 1 merupakan *cluster* yang memiliki daya tampung yang lebih rendah dibandingkan dengan kedua *cluster* lainnya.

Selain dilihat satu per satu melalui boxplot, kita juga dapat melihat karakteristik setiap *cluster* melalui *scatter plot* berikut.

Gambar 4.14 Scatter Plot Berdasarkan Peminat 2021 dan Daya Tampung 2022

Berdasarkan Gambar 4.14, dapat diketahui bahwa antar *cluster* terdapat perbedaan peminat SNMPTN Tahun 2021 yang signifikan. Hal ini terlihat dari batas-batas yang jelas antara *cluster* 0, *cluster* 1, dan *cluster* 2. Sedangkan jika berdasarkan daya tampung perbedaan antar *cluster*-nya tidak terlalu jelas. Meskipun

perbedaan daya tampung antar *cluster* tidak terlalu jelas, namun jika diamati daya tampung tiap *cluster* terlihat mengumpul pada suatu interval tertentu.

Terlihat bahwa pada *cluster* 1 data mengumpul pada peminat 0 sampai 250 dan daya tampung 0 sampai 30. Selanjutnya, pada *cluster* 0 terlihat bahwa data mengumpul pada peminat 250 sampai dengan 750 dan daya tampung 20 sampai 50. Sedangkan, pada *cluster* 2 terlihat bahwa data peminat lebih dari 1500 dan daya tampung 50 sampai 75. Oleh karena itu, dapat disimpulkan bahwa *cluster* 1 merupakan *cluster* program studi saintek dengan peminat dan daya tampung yang rendah. *Cluster* 0 merupakan *cluster* program studi saintek dengan peminat dan daya tampung sedang. Sedangkan, *cluster* 2 merupakan *cluster* program studi saintek dengan peminat dan daya tampung tinggi.

PENUTUP

A. Kesimpulan

Dilakukan analisis *cluster* untuk mengelompokkan program studi yang ada di Universitas Gadjah Mada berdasarkan peminat SNMPTN Tahun 2021 dan daya tampung SNMPTN Tahun 2022. Pada penelitian ini *clustering* dilakukan dengan metode K-Means. Dengan menggunakan metode ini, diperoleh 3 *cluster* sebagai berikut.

1. *Cluster* 0 beranggotakan 18 program studi, yaitu program studi dengan kode 3611526, 3611012, 3611074, 3611082, 3611097, 3611101, 3611186, 3611194, 3611325, 3611333, 3611372, 3611387, 3611395, 3611406, 3611437, 3611453, 3611484, dan 3611492. *Cluster* ini merupakan *cluster* program studi saintek dengan peminat dan daya tampung sedang.
2. *Cluster* 1 beranggotakan 41 program studi, yaitu program studi dengan kode 3611534, 3611542, 3611557, 3611565, 3611573, 3611581, 3611596, 3611607, 3611615, 3611623, 3611631, 3611646, 3611654, 3611035. 3611043, 3611051, 3611155, 3611163, 3611171, 3611205, 3611213, 3611221, 3611221, 3611244, 3611252, 3611267, 3611275, 3611283, 3611291, 3611302, 3611317, 3611341, 3611356, 3611364, 3611414, 3611422, 3611445, 3611461, 3611476, 3611503, 3611511, dan 3611662. *Cluster* 1 merupakan *cluster* program studi saintek dengan peminat dan daya tampung yang rendah.
3. *Cluster* 2 beranggotakan 2 program studi, yaitu program studi dengan kode 3611027 dan 3611066. *Cluster* 2 merupakan *cluster* program studi saintek dengan peminat dan daya tampung tinggi.

B. Saran

Berdasarkan hasil analisis yang diperoleh dapat diketahui prodi-prodi dengan peminat serta daya tampung yang rendah, sedang, dan tinggi. Hasil ini dapat dijadikan pertimbangan dalam memilih program studi saat pendaftaran SNMPTN di Universitas Gadjah Mada. Saran yang dapat kami berikan kepada para peserta SNMPTN adalah agar melihat program studi yang hendak dipilih masuk ke dalam *cluster* mana. Dengan demikian, para peserta SNMPTN dapat memperkirakan bagaimana peluang diterima pada program studi tersebut.

DAFTAR PUSTAKA

A. Rencher, 2002. Method of Multivariate Analysis. 2nd ed. New York: John Wiley and Sons, Inc

Agusta Y. K-Means-Penerapan, Permasalahan dan Metode Terkait. Denpasar, Bali: Jurnal Sistem dan Informatika (Februari 2007) Vol. 3: 47-60; 2007.

J. W. Tukey, *Exploratory Data Analysis* (Addison-Wesley, Philippines, 1997).

K. Monika, N. Lal and S. Qamar, "K-mean clustering algorithm approach for data mining of heterogeneous data," in Information and Communication Technology for Sustainable Development, vol. 10, Singapore, Springer, 2018, pp. 61-70.

P. J. Rousseeuw (1987), "*Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis*", *Computational and Applied Mathematics*. 20: 53–65.

Simamora B. Analisis Multivariat Pemasaran. Jakarta: PT. Gramedia Pustaka Utama; 2005.

Zaenuddin, Muhammad. 2015. Isu, Problematika, Dan Dinamika Perekonomian, Dan Kebijakan Publik. Yogyakarta: Deepublish.