

# Analisis Peserta *Online Test* dengan Pembentukan *Cluster* dengan Algoritma *K-Means* dan Analisis Tabulasi

Kevin Chandra Gunawan<sup>1</sup>

---

**Abstract:** PT XYZ is one of the leading tobacco industries in Indonesia. PT XYZ manages more than 25,000 permanent employees both at the center and at subsidiaries. The company management has an international internship program to find quality human resources from an early age and as an employer branding container. The internship program is promoted to all universities in Indonesia with an initial selection stage using an online test. The company wants to know the grouping formed based on the first stage test data, namely the Situational Judgment Test (SJT) and Numerical. Grouping is done by using k-means clustering. Cluster results are used to classify the origin of student university. The relationship between the university and the cluster formed will be tested using tabulation and chi-square analysis. The results of grouping there are 4 clusters, there are clusters with both values having a high average. Another cluster with both low average values and the other with a combination of the two. The grouping shows that there are 4 universities that have a greater percentage than others in cluster 2 which have high average scores. There are 4 universities included in cluster 1 or groups with both low test scores.

**Keywords:** k-means; cluster; tabulation; chi-square

---

## Pendahuluan

PT XYZ merupakan salah satu industri tembakau terkemuka di Indonesia dan berada pada taraf internasional baik dari segi manajemen maupun pemasaran produk yang dimiliki. Tim manajemen pada PT XYZ senantiasa menerapkan praktek global terbaik dan sistem manajemen berkelas dunia dalam mengelola lebih dari 25.000 karyawan. Manajemen membentuk sistem dengan memiliki program magang (INTERN) yang disediakan untuk mencari sumber daya manusia dan diharapkan mampu dipersiapkan sejak dini sekaligus sebagai wadah *employer branding*. Program INTERN telah berjalan dari 2014–2020 dan memiliki peserta dari *background* universitas yang berbeda, dengan semester yaitu 4 atau 6, dan IPK lebih dari sama dengan 3,2. Manajemen perusahaan dalam melakukan promosi program INTERN memiliki universitas yang menjadi *targeted university*. Terdapat 10 universitas yang menjadi target promosi setiap tahunnya. Proses seleksi INTERN memiliki 5 tahap yang harus dipenuhi. Tahap pertama adalah pendaftaran, kedua tes *online*, ketiga wawancara

melalui video *assessment center*, lalu rekrutmen. Setiap universitas yang mengikuti program INTERN memiliki kemampuan yang berbeda dalam mengikuti program. Hal ini dapat dilihat pada hasil di tahap rekrutmen dan juga tes *online*. Fokus masalah akan diarahkan pada tahap tes *online* yang masih memiliki persebaran peserta dalam jumlah yang besar. Tahap tes *online* dilakukan dengan dua jenis tes, yaitu *Numerical* dan *Situational Judgment Test (SJT)*. Lulus tahap tes *online* akan dilihat melalui nilai pada dua jenis tes yang ada. Peserta akan dinyatakan lulus tahap tes *online* ketika kedua nilai tes lebih dari atau sama dengan 30. Perbedaan ini menyebabkan perlu evaluasi dalam promosi dan menciptakan prioritas untuk 10 universitas yang menjadi target. Diharapkan promosi dapat efisien dan efektif. Promosi mampu menghasilkan calon peserta yang berkualitas tanpa mengurangi jumlah keikutsertaan mahasiswa. Prioritas diharapkan akan memisahkan mahasiswa yang dinilai sesuai dengan standar perusahaan atau tidak. Prioritas akan dilakukan dengan cara mengelompokkan data peserta dan melihat persentase dari masing–masing universitas.

## Metode Penelitian

Pada bagian ini akan dibahas metode-metode yang akan digunakan untuk menyelesaikan permasalahan dalam penelitian ini yaitu analisis deskriptif,

---

<sup>1</sup>Fakultas Teknologi Industri, Jurusan Teknik Industri, Universitas Kristen Petra. Jl. Siwalankerto 121-131, Surabaya 60236. Email: kevincg160198@gmail.com.

*k-means algorithm, elbow method, silhouette method, crosstabs and chi-square.*

**Analisis Deskriptif**

Analisis deskriptif adalah metode analisis dengan mengelompokkan dan menampilkan data secara visual sesuai dengan tujuan dan kegunaan diagram atau grafik yang dimiliki. Alat bantu yang dapat digunakan pada analisis deskriptif antara lain adalah tabel dan diagram. Macam-macam diagram dan data yang diproyeksikan adalah seperti pada Tabel 1.

**Tabel 1.** Jenis diagram (Hariyanti [1])

Data	Jenis Diagram
Data Nominal	Diagram batang
Data Kontinyu	Diagram garis Diagram batang Diagram titik dan tren
Data Ordinal	Diagram batang Diagram titik
Standar Deviasi	Diagram garis Diagram titik dan garis
Distribusi	Diagram histogram
Frekuensi	Diagram garis
Korelasi	Diagram <i>scatter-plot</i>

**K-Means Algorithm**

*K-means* merupakan algoritma *clustering* atau analisis *cluster* yang dimulai dengan pemilihan secara acak nilai *k*. Nilai *k* merupakan jumlah *cluster* yang dibentuk. Penetapan nilai-nilai *k* dapat dilakukan secara *random*. Penentuan nilai *k* yang akan menentukan juga jumlah pusat dari *cluster* atau disebut *centroid* (Witten *et al.* [2]).

Dasar algoritma *k-means* adalah sebagai berikut (Santosa [3]):

1. Tentukan nilai *k* sebagai jumlah *cluster* yang ingin dibentuk.
2. Bangkitkan titik pusat *cluster* awal secara *random*.
3. Hitung jarak setiap data ke masing-masing titik pusat *cluster* menggunakan rumus antar dua objek atau *euclidean distance*. Perhitungan *euclidean distance* dapat dilihat pada Persamaan 1:

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{1}$$

Keterangan:

- $d_{ij}$  : Jarak antara titik *i* dengan titik *j*
- $x_i$  : Letak data *i* di sumbu *x*
- $x_j$  : Letak data *j* di sumbu *x*
- $y_i$  : Letak data *i* di sumbu *y*
- $y_j$  : Letak data *j* di sumbu *y*

4. Mengelompokkan data berdasarkan jarak terdekat antara data dengan titik pusat *cluster*.

5. Tentukan posisi titik pusat baru dengan cara menghitung nilai pusat dari data-data yang ada pada titik pusat yang sama.
6. Kembali ke langkah 3 jika posisi titik pusat baru dengan titik pusat lama tidak sama.

**Elbow Method**

*Elbow method* adalah sebuah metode yang digunakan untuk menentukan jumlah *cluster* yang optimal dari *dataset*. Metode *elbow* mengambil ide dasar dari *k-means* untuk meminimalkan total *intra-cluster variation* atau *Within Cluster Sum of Square* (WCSS) melalui pemilihan nilai *k* yang *random*. Metode ini dilihat secara visual dengan melihat perubahan drastis yang berbanding terbalik dengan nilai *k* sebelumnya, maka nilai sebelum terjadinya perubahan tersebut dianggap sebagai jumlah *cluster* yang paling tepat. Penambahan *cluster* baru tidak terlalu berpengaruh dalam membedakan data yang ada (Kordinariya *et al.* [4]). Perhitungan WCSS dapat dilihat pada Persamaan 2:

$$WCSS = \sum_{k=1}^K \sum_{x_i \in S_k} \|x_i - c_k\|_2^2 \tag{2}$$

Keterangan:

- K* : Jumlah *cluster* yang terbentuk
- $x_i$  : Data *i* pada kelompok *cluster* *S<sub>k</sub>*
- $c_k$  : Nilai *centroid cluster* ke *k*

**Silhouette Method**

*Silhouette method* merupakan metode untuk melihat kualitas dari *cluster*. Metode *silhouette* merupakan gabungan dari dua metode. *Cohesion* adalah mengukur relasi antara objek-objek yang ada dalam sebuah *cluster*. *Separation* adalah mengukur seberapa jauh sebuah *cluster* terpisah dengan *cluster* lain. Tahapan dalam perhitungan metode *silhouette coefficient* (Ahmed *et al.* [5]). Perhitungan *silhouette coefficient* dapat dilihat pada Persamaan 3:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{3}$$

Keterangan:

- $s(i)$  : *Silhouette coefficient*
- $a(i)$  : Rata-rata data di *cluster* A
- $b(i)$  : Jarak minimum dari kohesi data
- i* : Jumlah *cluster* yang terbentuk

**Crosstabs and Chi-Square**

Tujuan dari analisis adalah melihat hubungan variabel-variabel yang diuji. Jika terdapat dua variabel yang diuji dapat digunakan tabel kontingensi. Tabel kontingensi akan menjelaskan secara deskriptif apakah terdapat perbedaan yang besar antara data observasi dengan jumlah data yang diekspektasikan. Pengujian statistik yang digunakan untuk melihat hubungan antar variabel adalah pengujian dengan metode *chi-square* (Sarwono [6]).

Nilai *chi-square* yang melebihi nilai dari *chi-square* pada tabel, mengidentifikasi penolakan terhadap hipotesis nol. Penolakan ini mengartikan bahwa ada interaksi atau saling mempengaruhi antar variabel (Montgomery *et al.* [7]).

## Hasil dan Pembahasan

### Analisis Deskriptif

Data yang digunakan dalam membentuk prioritas terhadap universitas yang telah mengikuti dan menjadi target promosi dari program magang adalah data dari tahun 2017-2019. Terdapat perbedaan sistem dalam perusahaan yang menyebabkan kesulitan dalam menggunakan data di tahun-tahun sebelumnya. Peserta program INTERN pada tahun 2017-2019 merupakan mahasiswa dari 10 universitas yang ditargetkan dan mahasiswa dari universitas lain di luar target. Universitas lain yang di luar target dikelompokkan sesuai daerah asal universitas tersebut. Mahasiswa yang mendaftar dan sesuai syarat akan diberikan kesempatan mengikuti tahap tes *online*. Jumlah peserta yang mendaftar beserta jumlah peserta yang lulus tahap tes *online* dapat dilihat pada Tabel 2.

**Tabel 2.** Jumlah peserta dan persentase lulus

Universitas	Total Peserta	Persentase Lulus Tes <i>Online</i>
A	249	30%
B	406	31%
I	530	60%
<i>Others in East and Central Java</i>	211	42%
<i>Others in Outside Java</i>	34	62%
<i>Others in West Java</i>	419	47%
<i>Overseas</i>	34	79%
PE	316	45%
PR	260	59%
U	465	46%
UA	298	52%
UE	381	30%
UND	182	45%
UNR	430	38%

Universitas-universitas yang terdapat pada Tabel 2 merupakan 10 universitas yang menjadi promosi dan juga 4 kelompok daerah dari universitas-universitas di luar target promosi. Terdapat universitas target yang memiliki persentase lulus tertinggi sebesar 60% yaitu Universitas I, sedangkan juga terdapat

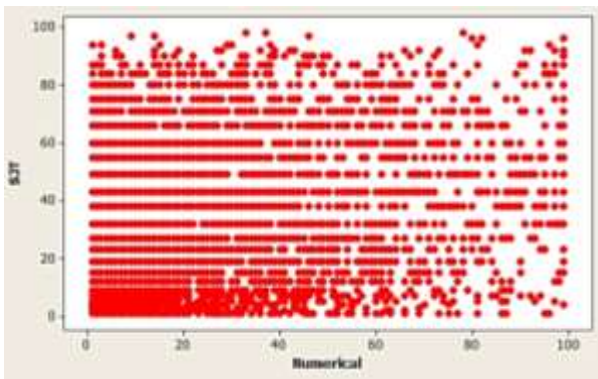
universitas target dengan persentase terendah sebesar 30% yaitu Universitas A dan UE. Universitas target yang memiliki peserta terbesar adalah Universitas I dengan 530 mahasiswa. Jumlah terkecil adalah Universitas UND dengan 182 mahasiswa. Melalui persentase yang ada, terlihat bahwa ada universitas-universitas yang memiliki persentase lulus tahap tes *online* dengan setengah atau lebih dari mahasiswa universitas tersebut yang mendaftar lulus tahap tes *online*. Terdapat juga keadaan yang sebaliknya. Keadaan pada jumlah mahasiswa yang mengikuti program juga sama, terdapat universitas yang mampu mengirimkan bahkan hingga dua kali lipat dibanding universitas yang lain. Selain melihat dari tahap tes *online*, bila dilihat sampai tahap akhir hasilnya juga sama. Persentase universitas dalam lulus tahap akhir dapat dilihat pada Tabel 3.

**Tabel 3.** Persentase lulus tahap rekrutmen

Universitas	Persentase Lulus Tahap Rekrutmen
A	2%
B	3%
I	32%
<i>Others in East and Central Java</i>	0%
<i>Others in Outside Java</i>	0%
<i>Others in West Java</i>	10%
<i>Overseas</i>	5%
PE	10%
PR	11%
U	15%
UA	5%
UE	5%
UND	2%
UNR	2%

Tabel 3 menunjukkan terdapat kategori universitas yang belum pernah lulus hingga ke tahap rekrutmen. *Others in East and Central Java*, *Others in Outside Java* adalah pengelompokkan universitas di luar target berdasarkan daerah yang belum pernah ada kandidatnya yang lulus hingga tahap rekrutmen. Universitas yang telah menjadi target selama program ini berjalan seperti Universitas A juga hanya memiliki 1 mahasiswa yang pernah lulus hingga tahap rekrutmen dalam tahun 2017-2019. Hal ini dapat menjadi pemikiran awal untuk mengelompokkan data pendaftar yang ada, khususnya asal universitas mereka untuk memilah manakah universitas yang perlu menjadi fokus promosi ataupun dilakukan evaluasi terhadap bentuk promosinya. Data juga menunjukkan terdapat 4 universitas target yang memiliki persentase lulus tahap rekrutmen tertinggi yaitu I, PE, PR, dan U. Universitas yang lain memiliki

persentase di bawah 10%. Persebaran data lulus ini menunjukkan adanya kecenderungan beberapa universitas memiliki persentase lebih besar dibandingkan yang lain. Fokus pada hasil mahasiswa yang lulus hingga tahap tes *online*. Mahasiswa yang mengikuti program INTERN baru dapat dinyatakan lulus dari tahap *online* tes dan masuk kepada tahap selanjutnya ketika nilai di kedua tes yang ada minimal adalah 30 di kedua nilainya. Data dari kedua nilai untuk seluruh peserta dapat dilihat pada Gambar 1.



**Gambar 1.** *Scatter plot* nilai SJT dan *Numerical*

Gambar 1 menunjukkan posisi dari setiap data dengan menggunakan dua nilai tes dalam bentuk *scatter plot*. Nilai pada sumbu x merupakan *Numerical* dan sumbu y adalah SJT. Peserta yang lulus merupakan mereka yang memiliki nilai di sumbu x dan y lebih atau sama dengan 30. Terlihat persebaran data lebih condong ke arah kedua nilai rendah daripada ke arah ketika sumbu x dan sumbu y memiliki nilai yang lebih tinggi. Melihat dari Gambar 1 tidak terlihat kecenderungan dari universitas yang ada terhadap dua nilai tes yang dimiliki.

Keperluan mengetahui informasi tersebut dalam analisis deskriptif maka dilakukan pengecekan terhadap rata-rata dari nilai yang dimiliki seluruh mahasiswa dari masing-masing universitas. Rata-rata nilai akan digunakan untuk melihat gambaran besar dari data mahasiswa yang ada. Nilai rata-rata akan mewakili rata-rata hasil dari mahasiswa pada universitas tersebut terhadap hasil yang akan didapatkan. Rata-rata akan dinilai baik jika di kedua nilai memiliki rata-rata di atas standar perusahaan yaitu 30. Jika di bawah 30 artinya rata-rata mahasiswa dari universitas tersebut gagal dalam tes *online*. Hal itulah yang ingin dihindari oleh perusahaan dan lebih berfokus pada universitas dengan nilai rata-rata di atas standar perusahaan. Data rata-rata untuk setiap universitas terhadap nilai SJT dapat dilihat pada Tabel 4.

**Tabel 4.** Rata-rata nilai SJT dan *Numerical*

Universitas	SJT	Numerical
A	27,16	20,29
B	34,48	16,75
I	38,20	47,52
<i>Others in East and Central Java</i>	32,85	16,31
<i>Others in Outside Java</i>	32,00	18,47
<i>Others in West Java</i>	32,06	19,71
<i>Overseas</i>	55,00	51,53
PE	38,41	31,66
PR	41,91	39,38
U	33,42	27,19
UA	30,68	19,86
UE	35,94	26,44
UND	31,63	19,54
UNR	30,20	21,36

Tabel 4 menunjukkan rata-rata nilai SJT yang tertinggi terdapat pada mahasiswa *Overseas* dengan nilai sebesar 55,00. Universitas dengan nilai rata-rata terendah adalah mahasiswa yang berasal dari Universitas A dengan nilai 27,16. Perbedaan yang cukup jauh dalam analisis deskriptif menunjukkan adanya indikasi perbedaan kualitas. Universitas A yang memiliki rata-rata nilai di bawah standar perusahaan memiliki arti bahwa rata-rata dari mahasiswa universitas A yang mengikuti tes *online* memiliki nilai di bawah standar. Nilai *Numerical* pada Tabel 4 menunjukkan universitas yang memiliki nilai tertinggi adalah kategori *Overseas* dengan nilai 51,53. Rata-rata nilai terendah pada *Numerical* adalah *Others in East and Central Java* dengan nilai 16,31. Pada kasus nilai tes *Numerical* terdapat kategori universitas yang lebih banyak memiliki nilai rata-rata di bawah standar yang diinginkan perusahaan. Terdapat 10 universitas dari 14 yang ada, memiliki nilai rata-rata di bawah standar perusahaan. Secara sekilas hal ini menunjukkan adanya kecenderungan gagal dari mahasiswa terletak pada tes *Numerical*.

**Analisis K-Means Cluster**

Analisis *cluster* berfungsi untuk melakukan pengelompokan data berdasarkan kesamaan atau sifat yang diinginkan. Analisis yang dipakai adalah algoritma *k-means*, pembentukan kelompok pada algoritma ini dilakukan dengan menentukan titik tengah (*centroid*) dan akan mengelompokkan data berdasarkan jarak terkecil dari setiap data ke *centroid* yang ada. Algoritma *k-means* akan membangkitkan nilai *centroid* secara *random*. *Centroid* akan digunakan untuk iterasi yang berfungsi untuk

membuktikan kestabilan model dari data yang digunakan dengan jumlah *cluster* atau *k* yang diminta. Iterasi dilakukan hingga tidak terdapat perubahan nilai *centroid*. Pengerjaan *clustering* data ini dilakukan dengan menggunakan program SPSS dan juga Minitab. Nilai *k* yang dipilih pada algoritma *k-means* adalah 8, dan akan dibentuk grafik dari WCSS dengan pengelompokan 1 sampai dengan 8. Setelah itu dilihat iterasi apakah model ini stabil atau tidak.

**Metode Elbow**

Nilai WCSS didapatkan melalui pengolahan algoritma *k-means* menggunakan program Minitab dengan melakukan operasi pada setiap nilai *k* dan melakukan pencatatan nilai WCSS. Pencatatan pada nilai WCSS dilakukan hingga nilai *k*=8. Pencatatan akan dijadikan sumber membuat grafik, yang akan membantu dalam melihat di titik mana atau nilai *k* berapa yang berkurang penurunan pada WCSS. Hasil grafik WCSS dapat dilihat pada Gambar 2.



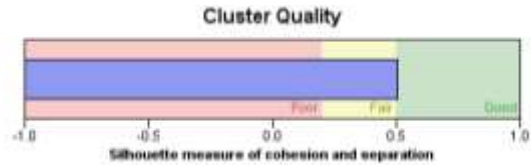
Gambar 2. Grafik line WCSS

Pada Gambar 2 terlihat grafik menunjukkan adanya pengurangan angka penurunan secara visual dimulai dari nilai *k*=3 hingga *k*=4. Pada kedua titik inilah yang secara visual atau metode *elbow* diperkirakan sebagai jumlah kelompok terbaik untuk sebaran data SJT dan *Numerical* yang dimiliki. Kelemahan metode *elbow* adalah penilaian dilakukan secara visual, sehingga sulit untuk menentukan ketika penurunan yang terjadi tidak menunjukkan perubahan hanya pada satu titik. Metode *silhouette* akan digunakan untuk membantu melihat kualitas dari model *cluster* jika nilai *k* adalah 3 dan 4.

**Metode Silhouette**

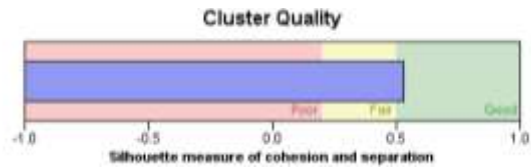
Metode *silhouette* digunakan untuk membantu melihat kualitas model *cluster* dengan melakukan perhitungan secara kohesi dan separasi pada pengelompokan data. Dilakukan dua perhitungan dengan pembentukan 3 *cluster*

dan 4 *cluster*. Koefisien yang dihasilkan dari nilai *k*=3 dapat dilihat pada Gambar 3.



Gambar 3. Nilai koefisien *silhouette* 3 cluster

Kualitas model dengan nilai *k*=3 memiliki nilai dekat dengan batas sebuah model dikatakan *fair* atau lebih dari 0,5 seperti pada Gambar 5. Nilai ini yang digunakan sebagai pembandingan kualitas terhadap model yang memiliki nilai *k*=4. Koefisien yang dihasilkan dari nilai *k*=4 dapat dilihat pada Gambar 4.



Gambar 4. Nilai koefisien *silhouette* 4 cluster

Gambar 4 menunjukkan model dengan nilai *k*=4 memiliki kualitas yang lebih baik dibandingkan dengan model dengan nilai *k*=3. Kualitas yang lebih baik tersebut menjadi alasan untuk memilih melakukan analisis *k-means cluster* dengan 4 *cluster*. Pada analisis *cluster* dengan pembentukan 4 kelompok maka akan didapati iterasi seperti pada Tabel 5.

Tabel 5. Iterasi *k-means* 4 cluster

Iterasi	Perubahan <i>Centroid Cluster</i>			
	1	2	3	4
1	23,69	36,94	34,86	37,09
2	2,62	3,27	3,13	4,96
3	1,45	1,70	1,70	3,16
4	0,97	1,24	0,91	1,95
5	0,44	0,56	0,54	1,21
6	0,32	0,78	0,37	1,08
7	0,40	0,54	0,37	0,86
8	0,23	0,53	0,31	0,72
9	0,26	0,59	0,28	0,80
10	0,22	0,52	0,26	0,50
11	0,05	0,19	0,00	0,22
12	0,04	0,05	0,00	0,12
13	0,05	0,00	0,00	0,13
14	0,00	0,00	0,00	0,00

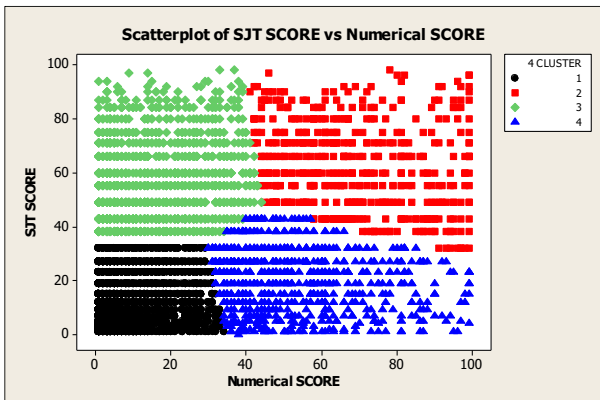
Tabel 5 menunjukkan iterasi yang terjadi pada 4 kelompok yang terbentuk. Iterasi ini berakhir pada iterasi ke 14 dengan semua kelompok telah mencapai nilai 0. Kelompok yang terbentuk telah membagi dan memiliki kelompok data yang stabil. Iterasi juga menunjukkan bahwa *centroid* dari 4 kelompok yang

terbentuk mengalami perpindahan sebesar angka yang ditunjukkan pada Tabel 5 secara *euclidean*. Selanjutnya fokus dalam pembagian kelompok akan dilakukan dengan membagi data dalam 4 kelompok. Hasil analisis *cluster* menghasilkan 4 kelompok dengan kesamaan yang berbeda. Kelompok yang terbentuk memiliki titik *centroid* yang dapat dilihat pada Tabel 6.

Tabel 6. Titik *centroid* 4 *cluster*

Score	Cluster			
	1	2	3	4
SJT	14	63	57	21
Numerical	11	69	17	53

Tabel 6 menunjukkan pengelompokkan dengan  $k=4$  memiliki nilai *centroid* tertinggi pada *cluster* 2. Nilai ini adalah satu-satunya nilai *centroid* yang memenuhi syarat untuk lanjut ke fase berikutnya, yaitu dengan batas 30 pada kedua nilai tes. *Cluster* yang lain juga memiliki sifat lain yang merupakan kombinasi dari keadaan pada nilai SJT dan *Numerical*. Kelompok yang terbentuk akan mengelompokkan keseluruhan data dalam 4 kelompok sesuai dengan kedekatan setiap data dengan masing-masing *centroid* dari *cluster* yang terbentuk. Pengelompokkan data terhadap *cluster* yang telah terbentuk dapat dilihat pada Gambar 5.

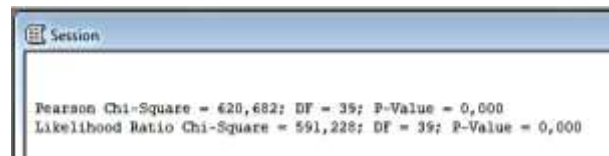


Gambar 5. Scatter plot data dengan *cluster*

Gambar 5 menunjukkan *scatter plot* data dengan nilai sumbu x merupakan nilai *Numerical* dan nilai sumbu y merupakan nilai SJT data. *Cluster* 1 dengan warna hitam mengelompokkan data yang berada di ujung kiri bawah artinya kelompok 1 merupakan kumpulan dari mahasiswa yang memiliki nilai rendah di kedua tes. Sedangkan untuk kelompok 2 dengan warna merah merupakan kumpulan mahasiswa dengan nilai di kedua tes memiliki nilai yang tinggi. Kedua *cluster* yang lain dengan warna hijau untuk *cluster* 3 dan biru untuk *cluster* 4 memiliki data dengan kombinasi nilai tinggi atau rendah pada salah satu dari dua jenis tes. Kelompok yang terbentuk memiliki 4 sifat-sifat yang berbeda.

### Analisis Chi-Square

*Cluster* yang terbentuk akan diuji menggunakan analisis *chi-square* dengan kategori mahasiswa yaitu asal universitas. Universitas memiliki 14 kategori untuk semua data. Analisis ini dilakukan untuk melihat hubungan antara asal universitas mahasiswa yang menjadi peserta dengan kategori kelompok yang terbentuk. Jika terdapat hubungan, maka data universitas setiap peserta akan mampu dikelompokkan dari hasil *cluster* yang dilakukan. Hasil analisis *cluster* dapat dilihat pada Gambar 6.



Gambar 6. Output *chi-square* Minitab

Gambar 6 menunjukkan *pearson chi-square* adalah sebesar 620,682 dengan *p-value* kurang dari 0,05. Nilai dari *pearson chi-square* memiliki nilai yang lebih besar dibandingkan nilai dari tabel *chi-square*. Hasil menunjukkan bahwa hipotesis awal yang mengatakan bahwa tidak terdapat hubungan pada variabel yang diuji dapat ditolak, sehingga menyatakan terdapat hubungan pada dua variabel yang diuji. Hubungan dua variabel yang diuji menyebabkan data mahasiswa berdasarkan asal universitasnya dapat dibagi ke dalam 4 kelompok yang terbentuk. Pengelompokkan akan melihat persentase data mahasiswa berdasarkan asal universitasnya pada masing-masing *cluster* terhadap total data mahasiswa yang mendaftarkan pada universitas tersebut. Persentase universitas terhadap *cluster* dapat dilihat pada Tabel 7.

Tabel 7. Persentase universitas terhadap *cluster*

Universitas	Cluster			
	1	2	3	4
A	53%	4%	28%	15%
B	48%	3%	40%	9%
I	18%	27%	20%	34%
Others in East and Central Java	50%	5%	39%	6%
Others in Outside Java	47%	6%	38%	9%
Others in West Java	51%	8%	29%	12%
Overseas	9%	50%	29%	12%
PE	32%	16%	34%	18%
PR	22%	24%	30%	23%
U	40%	11%	32%	17%
UA	53%	6%	32%	9%
UE	37%	9%	34%	20%
UND	50%	5%	32%	13%
UNR	50%	6%	30%	14%

Tabel 7 menunjukkan persentase data mahasiswa dengan asal universitas pada setiap kelompok yang terbentuk. Terdapat beberapa universitas yang memiliki persebaran data dengan kecenderungan pada *cluster* tertentu. Kecenderungan dinilai dengan melihat persebaran data terbanyak universitas tersebut pada 4 kelompok yang terbentuk. Persentase dari data universitas pada setiap kelompok, akan membantu dalam menentukan mahasiswa dari universitas tersebut berada pada *cluster* berapa. Seperti yang terlihat pada Tabel 7 terdapat universitas A yang memiliki persentase terbesar pada *cluster* 1. Universitas B dengan dua persentase terbesarnya terletak pada *cluster* 1 dan 3. Indikasi ini menunjukkan adanya pengelompokan baik ke dalam satu atau banyak kelompok untuk suatu universitas. Pengelompokan universitas terhadap *cluster* dapat dilihat pada Tabel 8.

**Tabel 8.** Pengelompokan universitas terhadap *cluster*

Universitas	Cluster
A	1
B	1 dan 3
I	Tersebar rata
<i>Others in East and Central Java</i>	1
<i>Others in Outside Java</i>	1 dan 3
<i>Others in West Java</i>	1
<i>Overseas</i>	2
PE	1 dan 3
PR	Tersebar rata
U	1 dan 3
UA	1
UE	1 dan 3
UND	1
UNR	1

Pengelompokan pada Tabel 8 menunjukkan terdapat beberapa universitas yang masuk ke lebih dari 1 *cluster* sebagai pusat dari persebaran data. Terdapat juga data universitas yang tersebar rata dalam 4 *cluster* artinya tidak ada kecenderungan data ke dalam salah satu kelompok *cluster* yang dimiliki. Pembagian universitas menurut *cluster* inilah yang akan dijadikan sebagai pertimbangan untuk meletakkan universitas tersebut pada prioritas tertentu dalam promosi program INTERN. Prioritas akan ditentukan berdasarkan *cluster* yang paling sesuai dengan keinginan perusahaan.

Hasil dari analisis *k-means* dengan metode *silhouette* menunjukkan bahwa 4 kelompok yang terbentuk mempunyai sifat yang berbeda 1 dengan yang lain. Kelompok 1 adalah kumpulan mahasiswa dengan nilai rata-rata di bawah standar yang diinginkan perusahaan yaitu 30 pada kedua jenis tes. Kelompok 2 adalah mereka yang memiliki nilai di kedua tes di

atas standar perusahaan. Kelompok 3 adalah kumpulan mahasiswa dengan rata-rata nilai SJT berada di atas standar perusahaan dan nilai *Numerical* di bawah standar perusahaan. Terakhir kelompok 4 adalah kumpulan mahasiswa dengan rata-rata nilai SJT berada di bawah standar perusahaan dan nilai *Numerical* di atas standar perusahaan. Perusahaan dalam melakukan promosi akan berfokus pada universitas yang memiliki tingkat keberhasilan yang tinggi, atau pada hal ini adalah mereka yang memiliki persentase tinggi pada *cluster* 2. Empat universitas yang memiliki persentase tertinggi pada *cluster* 2 adalah I, U, PE, dan PR. Setelah itu promosi akan difokuskan pada *cluster* 3 dan 4 yang memiliki salah satu nilai di atas standar. Terakhir baru dilakukan pada *cluster* 1. Prioritas promosi dapat dilihat pada Tabel 9.

**Tabel 9.** Target promosi berdasarkan analisis *cluster*

Prioritas	Universitas
1 ( <i>cluster</i> 2)	I, PE, PR, U
2 ( <i>cluster</i> 3 dan 4)	B, UE
3 ( <i>cluster</i> 1)	A, UA, UND, UNR

Tabel 9 menunjukkan pembagian prioritas menjadi 3. Prioritas akan berguna untuk fokus perusahaan dalam mengalokasikan dana, waktu, dan tenaga selama proses promosi berlangsung. Promosi akan difokuskan pada universitas yang mampu memberikan calon peserta dengan kualitas yang di atas standar perusahaan untuk menjawab tujuan mencari sumber daya manusia yang berkualitas. Tanpa mengurangi kepentingan tujuan pertama, prioritas juga akan tetap ditujukan terhadap universitas yang telah menjadi universitas target selama ini dengan melakukan evaluasi kedepannya dan juga untuk menjawab tujuan *employer branding*. Penyusunan prioritas ini juga mengesampingkan universitas lain di luar yang menjadi target promosi. Prioritas pertama adalah 4 universitas yang memiliki nilai persentase tertinggi di *cluster* 2. Selanjutnya adalah universitas-universitas lainnya yang berada pada *cluster* 1 dan 3 yang memiliki nilai pada salah satu tes di atas standar perusahaan. Terakhir adalah prioritas untuk universitas yang berada pada *cluster* 1 atau kumpulan mahasiswa dengan kedua nilai di bawah standar perusahaan.

## Simpulan

Program INTERN merupakan program yang memiliki dua tujuan yaitu sebagai wadah untuk mencari sumber daya manusia yang berkualitas dan juga untuk *employer branding*. Manajemen perusahaan dalam memenuhi tujuan melakukan target promosi ke 10 universitas yang tersebar di pulau Jawa. Tim manajemen selain memiliki target promosi, juga membuka pendaftaran untuk seluruh

mahasiswa yang tersebar di Indonesia atau luar negeri tetapi yang merupakan warga negara Indonesia. Selama berjalan program INTERN didapati perbedaan hasil dari mahasiswa masing-masing universitas target. Pada 2017-2019 terdapat 10 universitas yang memiliki rata-rata nilai di bawah standar yang diinginkan perusahaan pada nilai tes *Numerical*. Terdapat juga perbedaan persentase keberhasilan masing-masing universitas. Terdapat 3 universitas yang memiliki persentase lulus lebih dari 50% dan yang lainnya di bawah 50%.

Solusi yang dimiliki oleh PT XYZ adalah dengan melakukan pengelompokan dan prioritas promosi terhadap universitas-universitas yang ada. Terbentuk 4 kelompok dengan kelompok 1 berisi kumpulan mahasiswa yang memiliki rata-rata nilai di bawah standar perusahaan untuk kedua tes. Kelompok 2 memiliki rata-rata nilai di atas standar perusahaan untuk kedua tes. Kelompok 3 memiliki rata-rata nilai di atas standar perusahaan untuk nilai SJT dan di bawah standar perusahaan untuk *Numerical*. Kelompok 4 kebalikan dari kelompok 3. Prioritas promosi yang dilakukan adalah dengan mengutamakan promosi terhadap kelompok 2. Selanjutnya prioritas ditunjukkan untuk kelompok 3 dan 4. Terakhir adalah kelompok 1 dengan tujuan untuk melihat perkembangan program dan akan dilakukan evaluasi untuk tahun-tahun berikutnya.

## Daftar Pustaka

1. Hariyanti, E., *Metodologi Pembangunan Dashboard sebagai Alat Monitoring Kinerja Organisasi Studi Kasus Institut Teknologi Bandung*, Program Pascasarjana, Institut Teknologi Bandung, Bandung, Indonesia, 2008.
2. Witten, I., and Frank, E., *Data Mining Practical Machine Learning Tools and Techniques*, 2<sup>nd</sup> ed., Morgan Kaufmann, San Fransisco, 2005.
3. Santosa, B., *Data Mining: Teknik Pemanfaatan Data untuk Keperluan Bisnis*, Graha Ilmu, Yogyakarta, 2007.
4. Kordinariya, T., and Makwana, D., Review on Determining Number of Cluster in K-Means Clustering, *International Journal of Advance Research in Computer Science and Management Studies*, 1(6), 2013, pp. 90-95.
5. Ahmed, A., and Ashour, W., An Initialization Method for the K-means Algorithm using RNN and Coupling Degree, *International Journal of Computer Applications*, 25, 2011, pp. 1-6.
6. Sarwono, J., *Statistik Itu Mudah: Panduan Lengkap untuk Belajar Komputasi Statistik Menggunakan SPSS 16*, Andi Offset, Yogyakarta, 2009.
7. Montgomery, D. C., and Runger, G. C., *Applied Statistics and Probability for Engineers*, 7<sup>th</sup> ed., Wiley, New Jersey, 2018.