



METODE K-MEANS UNTUK PENGELOMPOKAN PROVINSI DI INDONESIA BERDASARKAN VARIABEL YANG BERPENGARUH PADA PERUBAHAN INDEKS PEMBANGUNAN MANUSIA TAHUN 2020

Oleh

Wara Pramesti

Program Studi Statistika UNIPA Surabaya

Email: warapra@unipasby.ac.id

Abstrak

Indeks Pembangunan Manusia (IPM) merupakan indikator komposit untuk mengukur capaian pembangunan kualitas hidup manusia. Indeks Pembangunan Manusia untuk menekankan pentingnya manusia beserta sumber daya yang dimilikinya dalam pembangunan. Indeks ini terbentuk dari rata-rata ukur capaian tiga dimensi utama pembangunan manusia, yaitu umur panjang dan hidup sehat, pengetahuan, dan standar hidup layak. Penelitian ini membahas tentang pengelompokan provinsi-provinsi di Indonesia berdasarkan variabel-variabel yang berpengaruh terhadap perubahan indeks pembangunan manusia dengan menggunakan metode *K-Means Cluster*. Metode regresi *Stepwise* digunakan untuk menentukan variabel-variabel yang signifikan berpengaruh terhadap perubahan Indeks Pembangunan Manusia di Indonesia. *Cluster* ditentukan sebanyak 2, *Cluster-1* terdiri dari 25 provinsi, *cluster-2* sembilan provinsi. Presentase SMA Menurut Provinsi dan Status Sekolah pada kedua cluster sangat berbeda.

Kata Kunci: *Stepwise K-Means & Indeks Pembangunan Manusia*

PENDAHULUAN

Indeks Pembangunan Manusia (IPM) merupakan salah satu ukuran kualitas yang dapat digunakan untuk mengetahui sejauh mana kualitas pembangunan manusia yang telah berhasil dicapai. IPM terbentuk dari rata-rata ukur capaian tiga dimensi utama pembangunan manusia, yaitu umur panjang dan hidup sehat, pengetahuan, dan standar hidup layak. Dimensi umur Panjang dan hidup sehat diwakili oleh Persentase Perempuan Pernah Kawin Berusia 15-49 Tahun Yang Proses Kelahiran Terakhirnya Ditolong Oleh Tenaga Kesehatan Terlatih, Persentase Rumah Tangga menurut Provinsi dan Memiliki Akses terhadap Sanitasi Layak, Persentase Rumah Tangga menurut Provinsi Tipe Daerah dan Sumber Air Minum Layak, dan Persentase Rumah Tangga menurut Provinsi, Tipe Daerah dan Sumber Air Minum Layak. Dimensi pengetahuan diwakili oleh Angka Partisipasi Kasar (APK) Perguruan Tinggi (PT) Menurut Provinsi, Presentase SD Menurut Provinsi dan Status Sekolah,

Presentase SMA Menurut Provinsi dan Status Sekolah, dan Presentase SMP Menurut Provinsi dan Status Sekolah. Dimensi standar hidup layak diwakili oleh Persentase Tenaga Kerja Formal Menurut Provinsi, dan Presentase Angkatan Kerja.

Sejak tahun 2010 IPM Indonesia terus tumbuh secara konsisten. Pada tahun 2020 pertumbuhan IPM di tingkat nasional daerah menghadapi tantangan dengan tumbuh melambat akibat pandemi COVID-19 yang melanda Indonesia dan sebagian besar negara di dunia. Perlambatan pertumbuhan IPM umumnya disebabkan oleh melambatnya pertumbuhan umur harapan hidup dan pendidikan, serta menurunnya pengeluaran riil per kapita sebagai akibat dari kontraksi pertumbuhan ekonomi [2]. Melihat perlambatan pertumbuhan IPM yang melanda Indonesia maka perlu untuk dilakukan pengelompokan provinsi-provinsi berdasarkan variabel-variabel yang berpengaruh signifikan terhadap perubahan IPM Indonesia.



K-Means merupakan salah satu metode pengelompokan yang sederhana. *K-Means* ini membagi data dalam sejumlah *cluster* berdasarkan keserupaan (*similarity*) yang terdapat pada kumpulan data. Penelitian terdahulu [9] berjudul Analisis *Cluster* Dengan Menggunakan Metode *K-Means* Untuk Pengelompokan Kabupaten/Kota Di Provinsi Maluku Berdasarkan Indikator Indeks Pembangunan Manusia Tahun 2014, menyimpulkan terbentuknya 3 *cluster* yang ditentukan. Analisis *K-Means Cluster* Indeks Pembangunan Manusia Di Indonesia, menyimpulkan bahwa Hasil analisis pengelompokan menggunakan metode *K-Means Cluster* berdasarkan indikator penyusun Indeks Pembangunan Manusia tahun 2017 diperoleh 2 kelompok (*cluster*) yang terbentuk. Karakteristik untuk *cluster* 1 yaitu berisi provinsi dengan umur harapan hidup (UHH) yang relatif rendah, harapan lama sekolah (HLS) yang relatif rendah, rata-rata lama sekolah (RLS) yang relatif rendah, dan pengeluaran disesuaikan yang rendah juga, sedangkan pada *cluster* 2 berisi provinsi dengan umur harapan hidup (UHH) relatif tinggi, harapan lama sekolah (HLS) yang relatif tinggi, rata-rata lama sekolah (RLS) yang relatif tinggi, dan pengeluaran disesuaikan yang relatif tinggi juga [8].

Pengelompokan provinsi-provinsi di Indonesia dalam penelitian ini, didasarkan pada variabel-variabel yang berpengaruh terhadap perubahan IPM di Indonesia, bukan berdasarkan faktor pembentuk IPM, sehingga diharapkan dapat memacu laju pertumbuhan IPM di Indonesia.

LANDASAN TEORI

Analisis Regresi

Analisis regresi digunakan untuk mengetahui pola hubungan secara matematis antara satu variabel terikat dengan satu atau lebih variabel bebas [3].

Persamaan garis regresi dapat ditulis dalam bentuk matriks sebagai berikut:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1k} \\ 1 & X_{21} & X_{22} & \dots & X_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

dan disederhanakan menjadi:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (1)$$

\mathbf{Y} menyatakan vektor variabel terikat berukuran $n \times 1$, \mathbf{X} menyatakan matriks variabel bebas berukuran $n \times (k + 1)$, $\boldsymbol{\beta}$ menyatakan vektor parameter model berukuran $(k + 1) \times 1$ dan $\boldsymbol{\varepsilon}$ menyatakan vektor error berukuran $n \times 1$. Estimasi parameter $\boldsymbol{\beta}$ dengan meminimalkan RSS ($\boldsymbol{\beta}$) = 0 kemudian diturunkan terhadap $\boldsymbol{\beta}$, hasilnya sebagai berikut:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \quad (2)$$

Pengujian Parameter Model Regresi Klasik

Pengujian parameter ini untuk mengetahui apakah parameter regresi sama dengan nol secara statistik atau tidak, baik secara serentak maupun secara parsial.

Langkah pengujian secara serentak

$$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$$

$$H_1: \text{minimal ada satu } \beta_k \neq 0,$$

Statistik uji yang digunakan :

$$F_{\text{hit}} = \frac{SS_{\text{reg}}/k}{\text{RSS}/(n-k-1)} = \frac{(\hat{y}_i - \bar{y})^2/k}{(y_i - \hat{y})^2/(n-k-1)} \sim F_{\alpha, (k, n-k-1)} \quad (3)$$

H_0 ditolak apabila $F_{\text{rasio}} > F_{\alpha; k; (n-k-1)}$ atau $p\text{-value} < \alpha$ dengan k adalah banyaknya parameter, dan dapat diartikan minimal ada satu



β_k yang tidak sama dengan nol atau jika p -value $< \alpha$.

Pengujian secara parsial

Langkah pengujian secara parsial sebagai berikut:

$$H_0: \beta_k = 0$$

$$H_1: \beta_k \neq 0,$$

Statistik uji yang digunakan:

$$t_{\text{hit}} = \frac{\hat{\beta}_k}{\text{se}(\hat{\beta}_k)} \sim t_{\frac{\alpha}{2}, n-k-1}$$

(4)

H_0 ditolak apabila $|t_{\text{hit}}| > t_{\left(\frac{\alpha}{2}, n-k-1\right)}$, atau p -value $< \alpha$

Error dalam analisis regresi diasumsikan identik, independen dan berdistribusi normal. Multikolinieritas adalah terjadinya korelasi antara variabel bebas dalam analisis regresi. Multikolinieritas tidak boleh terjadi dalam analisis regresi, tetapi multikolinieritas akan selalu ada dalam setiap model regresi dengan dua variabel bebas atau lebih.

Multikolinieritas yang memerlukan perhatian, adalah terjadinya korelasi antar variabel bebas yang sangat tinggi, karena dampak yang ditimbulkan adalah tidak dapat dihitungnya nilai koefisien regresi dan varians koefisien regresi akan semakin membesar, sehingga menyebabkan selang kepercayaan menjadi lebar dan berakibat penaksiran parameter regresi menjadi tidak akurat [7]. Indikasi terjadinya kasus multikolinieritas antara lain, apabila ditemukan nilai koefisien determinasi yang tinggi, hasil pengujian secara serentak signifikan, tetapi banyak parameter regresi yang tidak signifikan dan nilai Varians Inflasi Faktor lebih besar dari 10 ($VIF > 10$). VIF dapat dicari dari persamaan berikut:

$$VIF = \frac{1}{(1 - R_i^2)} \quad (5)$$

R_i^2 : koefisien determinasi variabel bebas ke i terhadap variabel-variabel bebas lainnya dalam regresi.

Perbaikan pelanggaran asumsi multikolinieritas antara lain dapat

menggunakan metode pemilihan regresi terbaik, yaitu metode Stepwise.

Metode Regresi Stepwise

Metode regresi (*stepwise regression*) merupakan prosedur statistik untuk menentukan variabel mana yang akan dimasukkan ke dalam persamaan regresi. Prosedur stepwise pada dasarnya merupakan kombinasi dari prosedur eliminasi *backward* dan prosedur eliminasi *forward*.

Regresi *stepwise* sering digunakan pada regresi linier berganda dengan banyak variabel bebas. digunakan juga untuk mencari persamaan regresi terbaik dan untuk mengatasi multikolinieritas.

Algoritma Regresi *Stepwise* [1]:

1. Menghitung korelasi semua variabel bebas terhadap variabel terikat. Variabel bebas dengan nilai korelasi tertinggi dimasukkan dalam model regresi (syarat uji F menunjukkan bahwa variabel bebas yang dimasukkan model berpengaruh signifikan). Jika tidak signifikan, maka berhenti dan mengambil persamaan sebagai yang terbaik.
2. Menghitung korelasi parsial semua variabel bebas tanpa menyertakan variabel bebas yang telah masuk model. Variabel bebas dengan korelasi parsial tertinggi dimasukkan ke dalam persamaan.
3. Setelah pengujian parameter, perhatikan peningkatan R^2 dan nilai-F parsial atau nilai-t parsial untuk kedua variabel yang ada di dalam persamaan. Nilai-F parsial terendah kemudian dibandingkan dengan nilai F tabel, dan variabel bebas tersebut dipertahankan apabila nilai-F parsial terendah lebih besar dari F tabel atau nilai-t parsial lebih besar dari nilai-t tabel, dan dikeluarkan dari persamaan apabila hasil pengujian tidak signifikan
4. Kembali ke langkah b



5. Jika tidak ada variabel yang dapat dikeluarkan atau dimasukkan, maka proses akan terhenti.

Pengujian asumsi residual identik dilakukan dengan uji *Glejser*, residual independen dengan uji *Durbin Watson* [5] dan normalitas residual menggunakan *Anderson Darling*

Metode K-Means

K-Means merupakan salah satu metode data *clustering* nonhierarki yang berusaha mempartisi data yang ada dalam bentuk satu atau lebih *cluster*. Data dengan karakteristik sama dikelompokkan kedalam satu *cluster* yang sama. Asumsi semua pengukuran bersifat numerik

Ukuran Kemiripan dan Ketakmiripan

Syarat jarak yang digunakan untuk mengukur ketakmiripan antar 2 objek a dan b, dinotasikan dengan $d(a,b)$, :

$d(a, b) \geq 0$, $d(a, a) = 0$, $d(a, b) = d(b, a)$, $d(a, b)$ akan meningkat seiring semakin tidak mirip kedua objek a dan b. $d(a,c) \leq d(a,b) + d(b,c)$

Algoritmanya K-Means sebagai berikut :

1. Tentukan besarnya k, yaitu banyaknya kelompok yang akan dibuat, dan tentukan juga *centroid* di tiap kelompok.
2. Hitung jarak setiap objek dengan setiap *centroid*.
3. Hitung kembali rataaan (*centroid*) untuk kelompok yang baru terbentuk.
4. Ulangi langkah 2 sampai tidak ada lagi pemindahan objek antar kelompok.

Untuk menentukan jarak terdekat setiap data terhadap masing-masing *centroid* digunakan rumus *Euclidean Distance* (jarak Euclid), berikut :

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

(6)

$d(x, y)$ = jarak *Euclid*

x_i = Objek 1

y_i = Objek 2

x_i, y_i dua data yang dihitung jaraknya dan p merupakan dimensi data yang digunakan. Titik pusat cluster dapat dicari dengan persamaan berikut :

$$C_{m(q)} = \frac{1}{n_m} \sum_{i=1}^{n_m} x_{i(q)} \tag{7}$$

$C_{m(q)}$ pusat kelompok ke m dari variable ke p, $m = 1, 2, \dots, k$, k adalah jumlah cluster, x_i nilai pengamatan obyek ke-i variable ke-q, $q = 1, 2, \dots, p$

METODE PENELITIAN

Data dalam penelitian ini merupakan data sekunder yang diperoleh dari Badan Pusat Statistik Indonesia tahun 2021.

Variabel Penelitian :

Variabel terikat dalam penelitian ini yaitu Indeks Pembangunan Manusia menurut provinsi di Indonesia tahun 2020 dan variable bebas yang diduga berpengaruh terhadap Indeks Pembangunan Manusia menurut provinsi di Indonesia sebagai berikut :

Tabel 1. Variabel dan Definisi Operasional Variabel

Nomor	Variabel	Definisi Operasional Variabel
1	x_1	Persentase Tenaga Kerja Formal Menurut Provinsi (Persen), 2020
2	x_2	Presentase Ang Kerja 2020
3	x_3	Proporsi Lapangan Kerja Informal Sektor Non-Pertanian Menurut Provinsi, 2020
4	x_4	Persentase Perempuan Pernah Kawin Berusia 15-49 Tahun Yang Proses Kelahiran Terakhirnya Ditolong Oleh Tenaga Kesehatan Terlatih Menurut Provinsi, 2020



5	x_5	Persentase Rumah Tangga menurut Provinsi dan Memiliki Akses terhadap Sanitasi Layak (Persen), 2020
6	x_6	Persentase Rumah Tangga menurut Provinsi, Tipe Daerah dan Sumber Air Minum Layak (Persen), 2020
7	x_7	Angka Partisipasi Kasar (APK) Perguruan Tinggi (PT) Menurut Provinsi, 2020
8	x_8	Persentase SD Menurut Provinsi dan Status Sekolah 2020
9	x_9	Persentase SMA Menurut Provinsi dan Status Sekolah 2020
10	x_{10}	Persentase SMP Menurut Provinsi dan Status Sekolah 2020

Tahap pertama penelitian diawali dengan menentukan persamaan garis regresi taksiran, kemudian pengujian parameter regresi secara serentak dan parsial, pemilihan persamaan terbaik dengan metode Stepwise, dilanjutkan pengujian asumsi residual. Langkah selanjutnya melakukan pengelompokan provinsi berdasarkan variable-variabel yang berpengaruh terhadap Indeks Pembangunan Manusia

HASIL DAN PEMBAHASAN

Analisis regresi digunakan untuk mengetahui hubungan secara matematis antara satu variable terikat dengan satu atau lebih variable bebas. Parameter regresi diestimasi dengan metode *Ordinary Least Square* (OLS), menghasilkan persamaan regresi taksiran sebagai berikut :

$$\hat{y} = 35,456 + 0,304x_1 + 0,040x_2 + 0,145x_3 + 0,143x_4 + 0,016x_5 - 0,052x_6 + 0,163x_7 + 0,014x_8 - 0,058x_9 +$$

Selanjutnya dilakukan pengujian secara serentak dan parsial. Hasil pengujian secara serentak memperoleh nilai $F_{rasio} = 19.542$ dan nilai signifikansinya 0,00, artinya paling tidak

terdapat satu parameter yang mempunyai nilai tidak nol. Untuk mengetahui, parameter mana yang tidak sama dengan nol, dilakukan pengujian secara individu.

Tabel 2. Hasil Estimasi Parameter Regresi

Model	Koefisien	Statistik Uji (t)	Sig	VIF
Konstanta	35.456	3.137	.005	
x_1	.304	3.508	.002	9.967
x_2	.040	.344	.734	1.709
x_3	.145	1.395	.176	5.165
x_4	.143	2.112	.046	3.659
x_5	.016	.318	.753	3.781
x_6	-.052	-1.216	.236	2.372
x_7	.163	5.199	.000	1.654
x_8	.014	.446	.660	2.001
x_9	-.058	-.915	.369	16.205
x_{10}	.009	.135	.894	17.490

Hasil pengujian individu dapat dilihat pada Tabel 2. Parameter yang signifikan hanya β_1 , β_4 , dan β_7 . Kolom nilai VIF menunjukkan bahwa terdapat 2 variabel yang memiliki nilai $VIF > 10$, dan 1 variabel memiliki nilai $VIF = 9,967$ mendekati nilai 10. Hal ini mengindikasikan terjadinya kasus multikolinieritas.

Untuk mengatasi hal tersebut, maka dilakukan pemilihan model terbaik dengan menggunakan metode Regesi *Stepwise*. Hasil analisis dapat dilihat pada Tabel 2. Model regresi yang dipilih, adalah model regresi taksiran yang mengandung variabel x_1, x_7, x_4, x_9 . Pemilihan variabel bebas tersebut berdasarkan koefisien determinasi (R^2) yang paling tinggi dan memiliki simpangan (S) yang paling kecil, yaitu 2,110.

Tabel 3. Hasil Analisis dengan Stepwise

Variabel	F	Sig.	S	R^2
x_5	48,092	0,00	6,273	0,600
$x_5 x_1$	36,405	0,00	4,840	0,701
$x_5 x_1 x_7$	36,374	0,00	3,611	0,784
$x_5 x_1 x_7 x_4$	40,558	0,00	2,627	0,848



x_1 x_7 x_4	54,889	0,00	2,581	0,846
x_1 x_7 x_4	52,265	0,00	2,110	0,878
x_9				

Uji multikolinieritas dilakukan terhadap keempat variable bebas tersebut, dan hasilnya tertera dalam Tabel 4. Nilai VIF yang diperoleh semua lebih kecil dari 10. Hal ini menunjukkan bahwa tidak terjadi multikolinieritas, sehingga dapat ditentukan persamaan garis regresi terbaik taksiran.

Tabel 4. Nilai VIF

Model	VIF
x_1	1.308
x_7	1.011
x_4	1.246
x_9	1.118

Persamaan garis regresi taksiran terbaik yang diperoleh:

$$\hat{y} = 40,96 + 0,196x_1 + 0,151x_7 + 0,213x_4 + 0,044x_9$$

dengan $R^2 = 0,878$

Pengujian secara serentak dilakukan untuk mengetahui apakah semua parameter regresi sama dengan nol atau tidak. Hasil pengujian secara serentak dapat dilihat pada Tabel :

Tabel 5. Hasil Uji Serentak

Model	Jumlah Kuadrat	db	Rata-rata Kuadrat	F	Sig
Regresi	441,211	4	110,303	52,265	0,00
Residual	61,204	29	2,110		
Total	502,414	33			

Uji serentak menghasilkan nilai F yang tinggi, yaitu 52,262 dan nilai signifikansi 0,00, berarti paling tidak ada satu nilai parameter yang tidak sama dengan nol. Untuk mengetahui parameter yang tidak bernilai nol, dilakukan pengujian secara individu atau uji parsial, dan hasilnya sebagai berikut :

Tabel 6. Hasil Uji Parsial

Model	Koefisien	Stat Uji (t)	Sig
Konstanta	40.956	11.426	.000
x_1	.196	6.536	.000
x_7	.151	6.427	.000
x_4	.213	5.645	.000
x_9	-.044	-2.773	.010

Hasil pengujian secara parsial pada Tabel 6, menunjukkan bahwa semua parameter tidak bernilai nol secara statistik, karena masing-masing mempunyai nilai signifikansi kurang dari α yang bernilai 0,05. Jadi dapat dikatakan bahwa persentase tenaga kerja formal menurut provinsi (x_1), angka partisipasi kasar perguruan tinggi (x_7), persentase perempuan pernah kawin berusia 15-49 tahun yang proses kelahiran terakhirnya ditolong oleh tenaga kesehatan terlatih (x_4), dan presentase SMA menurut provinsi dan status sekolah (x_9) berpengaruh secara parsial terhadap Indeks Pembangunan Manusia.

Pengujian asumsi residual identik dilakukan dengan uji *Glesjer*, yaitu dengan meregresikan residual absolut dengan semua variable bebas, dan apabila uji parsial menyatakan bahwa semua parameter tidak signifikan, maka asumsi residual identik dipenuhi. Tabel 6 menunjukkan hasil pengujian residual identik, dan nilai signifikansi yang diperoleh semua lebih besar dari α yang bernilai 0,05. Jadi dapat dikatakan bahwa asumsi residual identik dipenuhi.

Tabel 7. Hasil Uji Residual Identik

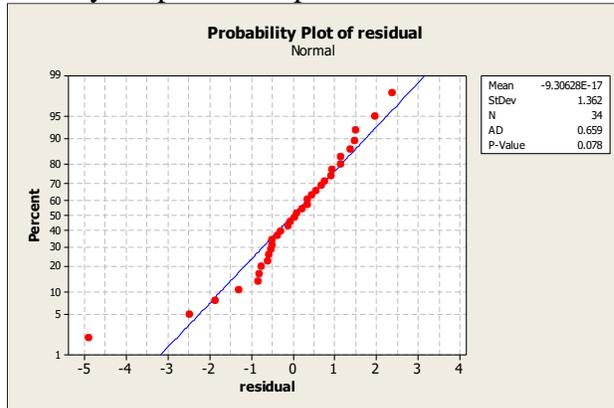
Model	Stat Uji (t)	Sig
Konstanta	1,045	0,304
x_1	1,452	0,157
x_7	0,093	0,927
x_4	-1,308	0,201
x_9	0,595	0,556

Statistik uji residual independen sebesar 2,072, lebih kecil dari $4 - dL$ ($2,072 < 2,79$),



maka dapat dikatakan bahwa asumsi residual independen dipenuhi.

Anderson Darling digunakan untuk pengujian asumsi residual berdistribusi normal. Hasilnya dapat dilihat pada Grafik 1



Grafik1. Grafik Hasil Uji Normalitas

Hasil uji normalitas residual dengan menggunakan *Anderson Darling*, diperoleh nilai $p\text{-value} = 0,078$, lebih besar dari $\alpha 5\%$, maka dapat disimpulkan bahwa residual berdistribusi normal.

Koefisien determinasi $R^2 = 0,878$, artinya perubahan indeks pembangunan manusia dapat dijelaskan oleh persentase tenaga kerja formal menurut provinsi (x_1), angka partisipasi kasar perguruan tinggi (x_7), persentase perempuan pernah kawin berusia 15-49 tahun yang proses kelahirannya ditolong oleh tenaga kesehatan terlatih (x_4), dan presentase SMA menurut provinsi dan status sekolah (x_9) sebesar 87,8% dan sisanya dijelaskan oleh variable lain yang tidak masuk dalam model.

Hasil pengujian secara serentak dan parsial, juga pengujian residual semua sudah memenuhi syarat, maka model regresi yang diperoleh sudah baik

Pengelompokkan variabel yang berpengaruh terhadap Indeks Pembangunan Manusia

Analisis *Cluster K-means* digunakan untuk pengelompokan provinsi di Indonesia berdasarkan hasil analisis regresi Stepwise. Analisis dilakukan dengan menggunakan metode *cluster K-means* yang ditentukan

menjadi 2 kelompok dari 34 provinsi di Indonesia. Hasil pengelompokan ke tiga puluh empat provinsi sebagai berikut :

Cluster 1 :

Cluster 1 terdiri dari 25 provinsi, yaitu : Aceh, Sumatera Barat, Riau, Jambi, Sumatera Selatan, Bengkulu, Lampung, Kep. Bangka Belitung, Nusa Tenggara Barat, Nusa Tenggara Timur, Kalimantan Barat, Kalimantan tengah, Kalimantan Timur, Kalimantan Utara, Sulawesi Utara, Sulawesi Tengah, Sulawesi Selatan, Sulawesi Tenggara, Gorontalo, Sulawesi Barat, Maluku, Maluku Utara, Papua Barat, dan Papua.

Cluster 2 :

Cluster 2 terdiri dari 9 provinsi, yaitu : Sumatera Utara, Riau, DKI Jakarta, Jawa Barat, Jawa Tengah, DI Yogyakarta, Jawa Timur, Banten dan Bali.

Dari tabel *output Final Cluster Centers*, dapat dijelaskan sebagai berikut :

Tabel 8. Final Cluster Centers

Variabel	Cluster	
	1	2
X_1	36.17	46.83
X_7	33.62	35.59
X_4	92.08	97.50
X_9	38.22	66.94

Cluster 1 :

Karakteristik provinsi- provinsi yang memiliki Persentase Tenaga Kerja Formal, Angka Partisipasi Kasar Perguruan Tinggi Menurut Provinsi, Persentase Perempuan Pernah Kawin Berusia 15-49 Tahun Yang Proses Kelahirannya Ditolong Oleh Tenaga Kesehatan Terlatih dan Presentase SMA Menurut Provinsi dan Status Sekolah, yang lebih dari rata-rata populasi provinsi yang diteliti, hal ini dapat dilihat dari nilai positif yang terdapat pada Tabel 8 dalam keseluruhan variabel

Cluster 2 :

Karakteristik provinsi- provinsi yang memiliki Persentase Tenaga Kerja Formal, Angka Partisipasi Kasar Perguruan Tinggi Menurut Provinsi, Persentase Perempuan Pernah Kawin Berusia 15-49 Tahun Yang



Proses Kelahiran Terakhirnya Ditolong Oleh Tenaga Kesehatan Terlatih dan Presentase SMA Menurut Provinsi dan Status Sekolah, berada di atas rata-rata populasi provinsi yang diteliti.

Cluster-2 merupakan pengelompokan dari provinsi-provinsi yang memiliki rata-rata presentase lebih tinggi daripada *cluster-1*, maka *cluster 1* merupakan kelompok yang memiliki rata-rata presentase lebih rendah dan *cluster-2* merupakan kelompok dengan rata-rata presentase lebih tinggi.

Setelah terbentuk 3 *cluster*, langkah selanjutnya melihat apakah variabel-variabel yang telah terbentuk dalam *cluster* mempunyai perbedaan pada setiap *cluster* dengan menggunakan Analisis Varians berikut :

Tabel 9. ANAVA

Variabel	Cluster		Error		F	Sig.
	Rata-rata Kuadrat	db	Rata-rata Kuadrat	db		
X1	752.753	1	71.922	32	10.466	.003
X7	25.494	1	120.718	32	.211	.049
X4	194.428	1	51.516	32	3.774	.041
X9	5458.456	1	122.324	32	44.623	.000

Berdasarkan Tabel 9, semakin besar F_{rasio} suatu variabel dengan nilai signifikansinya 0,000, maka semakin besar pula perbedaan variabel tersebut pada kedua *cluster*. Misal F_{rasio} terbesar, yaitu 44,623 pada Presentase SMA Menurut Provinsi dan Status Sekolah dengan nilai signifikansinya 0,000 adalah berarti (nyata). Artinya variabel nilai Presentase SMA Menurut Provinsi dan Status Sekolah sangat membedakan karakteristik kedua *cluster* tersebut, atau dapat dikatakan Presentase SMA Menurut Provinsi dan Status Sekolah pada kedua *cluster* sangat berbeda, demikian pula dengan Persentase Tenaga Kerja Formal Menurut Provinsi, sedangkan Persentase Perempuan Pernah Kawin Berusia 15-49 Tahun Yang Proses Kelahiran Terakhirnya Ditolong Oleh Tenaga Kesehatan Terlatih dan Presentase SMA Menurut Provinsi dan Status Sekolah berbeda pada kedua *cluster*.

PENUTUP

Kesimpulan

1. Model terbaik yang terbentuk dengan menggunakan metode *Stepwise* sebagai berikut:

$$\hat{y} = 40,96 + 0,196x_1 + 0,151x_7 + 0,213x_4 - 0,044x_9$$

$R^2 = 0,878$, perubahan IPM pada provinsi-provinsi di Indonesia dapat dijelaskan oleh Persentase Tenaga Kerja Formal Menurut Provinsi, Angka Partisipasi Kasar Perguruan Tinggi Menurut Provinsi, Persentase Perempuan Pernah Kawin Berusia 15-49 Tahun Yang Proses Kelahiran Terakhirnya Ditolong Oleh Tenaga Kesehatan Terlatih dan Presentase SMA Menurut Provinsi dan Status Sekolah sebesar 87,8% dan sisanya dapat dijelaskan oleh faktor lain yang tidak masuk ke dalam model.

2. Hasil pengelompokan ke tiga puluh empat provinsi dengan menggunakan metode *K-Mean* berdasarkan variabel-variabel yang berpengaruh signifikan terhadap IPM terdiri dari 2 *cluster*, dan *cluster-1* beranggotakan 25 provinsi dan *cluster-2* sembilan provinsi.

DAFTAR PUSTAKA

- [1] Achmad E dan Rahma F. 2020. *Analisis Regresi*. Teori dan Aplikasi dengan R. Malang:UB Press.
- [2] Badan Pusat Statistik Indonesia. 2020. *Indeks Pembangunan Manusia Tahun 2020*. Jakarta
- [3] Draper, N.R., and Smith, H., 1992. *Analisis Regresi Terapan edisi kedua* (Terjemahan oleh Bambang Sumantri). Jakarta : Gramedia.
- [4] Du.W.dkk., 2016. *A New Projection-based K-Means Initialization Algorithm*. Proceedings of 2016 IEE Chinese Guidance, Navigation and Control Conference. 978-1-4673-8318-9/16/S31.00©2016 IEEE.



-
- [5] Gujarati, Damodar N. 2015. *Dasar-Dasar Ekonometrika. Buku II. Edisi Kelima*. Jakarta Selatan: Salemba Empat.
- [6] Hu.J.dkk., 2017. *Improved K-Means Algorithm Based on Hybrid Fruit Fly Optimization and Differential Evolution*. 978-1-5090-2508-4/17/\$31.00©2017 IEEE.
- [7] Joko A.N., Pray P.H.N, (2016). *Setetes Ilmu Regresi Linier*. Malang: Media Nusa Creative
- [8] Muchlas Santoso (2019). *Analisis K-Means Cluster Indeks Pembangunan Manusia Di Indonesia*. Skripsi. Jurusan Matematika Fakultas Matematika Dan Ilmu Pengetahuan Alam Universitas Negeri Semarang
- [9] M. W. Talakua¹, Z. A. Leleury², A. W. Talluta. 2017. *Analisis Cluster Dengan Menggunakan Metode K-Means Untuk Pengelompokan Kabupaten/Kota Di Provinsi Maluku Berdasarkan Indikator Indeks Pembangunan Manusia Tahun 2014*. Barekeng Jurnal Ilmu Matematika dan Terapan. Desember 2017 . Volume 11 Nomor 2 .Hal. 119 – 128
- [10] Sembiring, R.K. 1995, *Analisis Regresi*, Bandung: ITB



HALAMAN INI SENGAJA DIKOSONGKAN