# Penerapan Metode Principle Component Analysis (PCA) untuk Clustering Data Kunjungan Wisatawan Mancanegara ke Indonesia

**Elly Muningsih[1], Noor Hasan[2], Gunawan Budi Sulistyo[3]**
Universitas Bina Sarana Informatika
elly.emh@bsi.ac.id[1], noor.nhs@bsi.ac.id[2], gunawan.gnw@bsi.ac.id[3]

**Abstrak** - Sektor pariwisata menjadi salah satu penyumbang devisa negara yang paling besar. Kunjungan wisatawan mancanegara ke Indonesia mencapai 16,1 juta sepanjang tahun 2019. Karenanya kunjungan wisatawan mancanegara menjadi suatu hal yang sangat penting. Pada penelitian ini akan dilakukan *clustering* atau pengelompokan data kunjungan wisatawan asing menjadi 5 kelompok untuk kategori negara dengan kunjungan sangat tinggi, tinggi, cukup tinggi, rendah dan rendah sekali. Pengolahan data yang dilakukan menggunakan metode *clustering K-Means* dan metode reduksi dimensi *Principle Component Analysis (PCA)*. Dari pengolahan data yang dilakukan didapatkan hasil *modelling K-Means* yang digabung dengan metode PCA menghasilkan nilai evaluasi *Index Davies Bouldin (DBI)* yang lebih kecil atau lebih baik yaitu sebesar 0,310 dibanding *modelling K-Means* saja yang mendapatkan hasil nilai DBI sebesar 0,382. Tools yang digunakan dalam pengolahan data adalah *RapidMiner*. Hasil *clustering* diharapkan bisa menjadi referensi pihak terkait untuk memaksimalkan promosi wisata ke luar negeri.
Kata Kunci : clustering, K-Means, Principle Component Analysis, Index Davies Bouldin

***Abstract*** *- The tourism sector is one of the country's biggest foreign exchange earners. Foreign tourist visits to Indonesia reached 16.1 million during 2019. Therefore foreign tourist visits become a very important thing. In this study clustering will be carried out or grouping data on foreign tourist visits into 5 groups for the category of countries with very high, high, high enough, low and very low visits. Data processing was performed using the K-Means clustering method and the Principle Component Analysis (PCA) dimension reduction method. From the data processing, K-Means modeling results combined with the PCA method resulted in a smaller or better Davies Bouldin Index (DBI) evaluation value of 0.310 compared to K-Means modeling alone which obtained a DBI value of 0.382. The tools used in data processing are RapidMiner. The results of clustering are expected to be a reference for related parties to maximize the promotion of overseas tourism.*
*Keywords : clustering, K-Means, Principle Component Analysis, Davies Bouldin Index*

## I. Pendahuluan

Sektor pariwisata di Indonesia sudah dirintis menjadi salah satu sektor strategis melalui kampanye Visit Indonesia Year 1991 yang kemudian ditetapkan sebagai sektor prioritas pembangunan dalam Nawa Cita 2014-2019. Hal ini tertuang dalam Renstra Kementerian Pariwisata 2015 - 2019 (Kementerian Pariwisata, 2015). Kontribusi nyata sektor pariwisata dalam perekonomian merupakan peran sektor pariwisata sebagai sektor strategis prioritas pembangunan (Hapsari & Nuryakin, 2019a). Pariwisata Indonesia memiliki potensi yang sangat beragam di seluruh wilayah Indonesia. Potensi pariwisata yang ada telah dikemas sedemikian rupa menjadi produk pariwisata mulai dari wisata alam (wisata bahari, ekowisata, wisata petualangan), wisata budaya (wisata warisan budaya dan sejarah, wisata belanja dan kuliner, wisata kota dan desa) serta wisata buatan manusia (wisata MICE, olahraga, dan objek wisata terintegrasi). Beragamnya produk pariwisata Indonesia ini menjadi daya tarik bagi wisatawan, baik wisatawan mancanegara maupun wisatawan domestik utuk berwisata sesuai minat yang digemarinya (Hapsari & Nuryakin, 2019b).

Kaitannya dengan wisatawan mancanegara atau biasa disebut Wisman, menurut Badan Pusat Statistik (BPS, 2019) dijelaskan bahwa Wisman adalah setiap pengunjung yang mengunjungi suatu negara di luar tempat tinggalnya, didorong oleh satu atau beberapa keperluan tanpa bermaksud memperoleh penghasilan di tempat yang dikunjungi dan lamanya kunjungan tersebut tidak lebih dari 12 (dua belas) bulan, yang mencakup dua kategori yaitu: Wisatawan (tinggal paling sedikit 24 jam, akan tetapi tidak lebih dari 12 bulan di tempat yang dikunjungi, dengan antara lain berlibur/rekreasi, olahraga, bisnis, menghadiri pertemuan, studi, dan kunjungan dengan alasan kesehatan), dan Pelancong / *Excursionist* (tinggal kurang dari 24 jam di tempat yang dikunjungi, termasuk setiap pengunjung yang tiba di suatu negara dengan kapal atau kereta api, dimana mereka tidak menginap di akomodasi yang tersedia di negara tersebut).

Kunjungan wisatawan mancanegara ke Indonesia bersumber dari pencatatan yang dilakukan oleh imigrasi pada pintu-pintu masuk laut (Pelabuhan), udara (Bandara), dan darat (Pos lintas batas - PLB), serta data yang

berbasis *Mobile Positioning Data* (MPD) (Pariwisata, Industri, & Regulasi, 2019).

Menurut data Kementerian Pariwisata dan Ekonomi Kreatif, data kunjungan Wisman ke Indonesia melalui seluruh pintu masuk bulan April 2020 berjumlah 160.042 kunjungan atau mengalami penurunan sebesar -87,44% dibandingkan bulan April 2019 yang berjumlah 1.274.231 kunjungan. Hal ini bisa dipahami karena adanya Pandemi Covid 19 yang melanda hampir di seluruh dunia termasuk Indonesia dan berdampak besar pada kunjungan Wisman.

Sudah lama sektor pariwisata khususnya kunjungan Wisman menjadi salah satu penopang perekonomian di Indonesia (Surtiningsih, Furqon, & Adinugroho, 2018). Karena itu kunjungan Wisman menjadi prioritas untuk diperbaiki sistem maupun layanannya agar tujuan utama sektor pariwisata bisa menjadi "core economy" atau penyumbang terbesar devisa negara bisa terwujud. Karena hal itu salah satu cara yang bisa dilakukan adalah dengan melakukan pengelompokan atau clustering data kunjungan wisatawan asing ke Indonesia agar diketahui negara mana saja yang kunjungan wisman nya tinggi dan rendah sehingga bisa menjadi rujukan untuk melakukan promosi wisata yang lebih intens dan kontinyu setelah pandemi Covid berakhir.

Penelitian ini akan mengolah data kunjungan Wisman tahun 2017-2020 yang diambil dari data BPS menjadi beberapa cluster dengan tools RapidMiner. Pada penelitian yang dilakukan (Linda Maulida-referensi), peneliti melakukan analisis penerapan datamining dalam mengelompokkan jumlah kunjungan wisatawan asing ke Prov. DKI Jakarta menggunakan K-Means. Sumber data penelitian berasal dari BPS Prov. DKI Jakarta. Data penelitian yang digunakan adalah jumlah pengunjung wisatawan tahun 2007-2013 sesuai dengan BPS Prov. DKI Jakarta. Data dikelompokkan menjadi 3 cluster yaitu C1= jumlah kunjungan wisatawan tinggi, C2=jumlah kunjungan wisatawan sedang dan C3= jumlah kunjungan wisatawan rendah.

Teknik data mining yang digunakan pada penelitian ini adalah metode clustering K-Means dengan penerapan algoritma reduksi dimensi Principle Component Analysis (PCA). Metode K-Means merupakan salah satu metode terbaik dan paling popular dalam algoritma *clustering* dimana K-Means mencari partisi yang optimal dari data dengan meminimalkan kriteria jumlah kesalahan kuadrat dengan prosedur iterasi yang optimal (Muningsih & Kiswati, 2015).

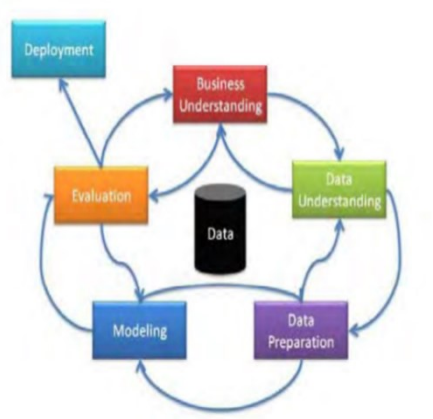
Algoritma dasar *clustering* data menggunakan metode *K-Means* dapat dilakukan dengan cara yaitu (Muningsih & Kiswati, 2018):

1. Menentukan jumlah *cluster* atau banyak kelompok
2. Inisialisai *k* sebagai pusat *cluster* (beri nilai random)
3. Alokasikan setiap data atau obyek ke cluster terdekat. Kedekatan dua obyek ditentukan berdasarkan jarak antar kedua obyek tersebut.
4. Hitung kembali pusat *cluster* dengan anggota *cluster* yang baru. Pusat *cluster* adalah rata-rata semua data atau obyek dalam *cluster.*
5. Tugaskan lagi setiap obyek memakai pusat *cluster* yang baru. Jika pusat *cluster* sudah tidak berubah lagi, maka proses peng*cluster*-an sudah selesai.
6. Kembali ke langkah 3 sampai pusat *cluster* tidak berubah lagi

Sedangkan Principal Component Analysis (PCA) merupakan salah satu bentuk teknik untuk mengambil data dimensi tinggi kemudian menggunakan dependensi antara variabel untuk mewakilinya secara lebih sistematis, membentuk dimensi rendah tanpa kehilangan banyak informasi yang ada dalam dataset (Dash, Nayak, & Prasad Das, 2014). PCA adalah metode reduksi dimensi yang umum dan banyak digunakan (Luo, Chen, & Jian, 2018). Penelitian ini akan melakukan perbandingan atau komparasi antara penerapan metode K-Means tanpa dan dengan teknik reduksi dimensi PCA untuk menemukan nilai Davies-Bouldin Index (DBI) yang terkecil sebagai nilai yang terbaik. Hasil dari pengelompokan data kunjungan wisatawan asing adalah jumlah dan negara yang termasuk dalam 5 kategori yaitu cluster negara dengan kunjungan wisman sangat tinggi, tinggi, cukup tinggi, rendah dan rendah sekali. Hasil penelitian diharapkan bisa menjadi rujukan pihak terkait dalam melakukan promosi keluar negeri.

## II. Metode Penelitian

Untuk membangun model pada penelitian ini digunakan metode CRISP-DM (Cross-Industry Standard Process for Data Mining). Metode ini memiliki enam fase atau tahapan seperti yang ditampilkan pada gambar 1 :




Sumber : (Sastry & Babu, 2013)
Gambar 1. Tahapan pada CRISP-DM

Pada penelitian ini, tahapan yang dilakukan adalah mengumpulkan dataset, memilih atribut yang relevan, membangun model clustering tanpa dan dengan algoritma reduksi dimensi PCA, menggunakan model clustering untuk pengelompokan data dan evaluasi model.

### 2.1. Dataset

Dataset yang digunakan adalah data kunjungan wisatawan mancanegara (wisman) dari tahun 2017 s/d 2020 yang diambil dari website BPS (www.bps.go.id) dengan jumlah data 241 record (negara) dan atribut berjumlah 40. Record berisi informasi kunjungan wisman dari 241 negara di dunia dan atribut berisi jumlah kunjungan wisman tiap bulan dmulai bulan Januari 2017 sampai bulan April 2020. Informasi lengkap data kunjungan Wisman yang digunakan adalah :

1. Kebangsaan : Brunei Darussalam, Malaysia, Philiphines, Singapore dan seterusnya..
2. Data jumlah kunjungan : Januari, Februari, Maret, April, Mei dan seterusnya.

### 2.2. Preprocessing Data

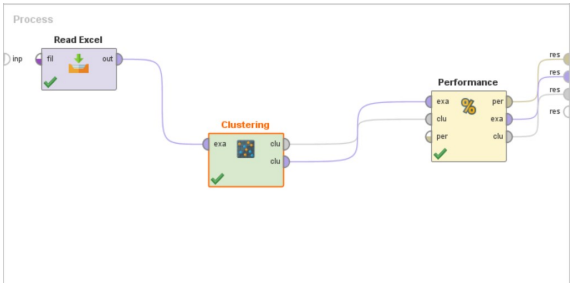
Dari dataset yang ada dilakukan reprocessing data yaitu pemilihan record dan antribut yang akan digunakan. Record yang digunakan adalah data kunjungan Wisman yang minimal ada 100 jumlah kunjungan pada salah satu atribut atau bulan yang digunakan. Selain itu dilakukan juga pengubahan kode untuk record dan atribut yang digunakan. Hasil dari reprocessing data menghasilkan record (negara) sebanyak 133 dan atribut berjumlah 39. Tipe data untuk atribut yang digunakan adalah integer. Data lengkap yang diolah ditampilkan pada gambar 2 berikut ini :

| Row No. | id_negara | a1 | a2 | a3 | a4 | a5 | a6 | a7 |
|---|---|---|---|---|---|---|---|---|
| 1 | N1 | 1713 | 1440 | 2591 | 1993 | 1897 | 1365 | 1518 |
| 2 | N2 | 149488 | 142045 | 172097 | 171536 | 176147 | 176276 | 174367 |
| 3 | N3 | 24536 | 23163 | 24054 | 27855 | 27567 | 21973 | 28649 |
| 4 | N4 | 118433 | 91255 | 135441 | 135619 | 117526 | 134699 | 114722 |
| 5 | N5 | 8677 | 8785 | 9673 | 14420 | 13147 | 8432 | 14652 |
| 6 | N6 | 5062 | 5217 | 5423 | 6323 | 6507 | 6763 | 9722 |
| 7 | N7 | 255 | 229 | 237 | 334 | 698 | 283 | 626 |
| 8 | N8 | 466 | 452 | 417 | 692 | 623 | 311 | 701 |
| 9 | N9 | 3817 | 3244 | 3546 | 5762 | 4202 | 3945 | 3988 |
| 10 | N10 | 22339 | 16982 | 20553 | 19733 | 21448 | 26880 | 19266 |

Sumber : Penulis (2020)
Gambar 2. Preprocessing Dataset

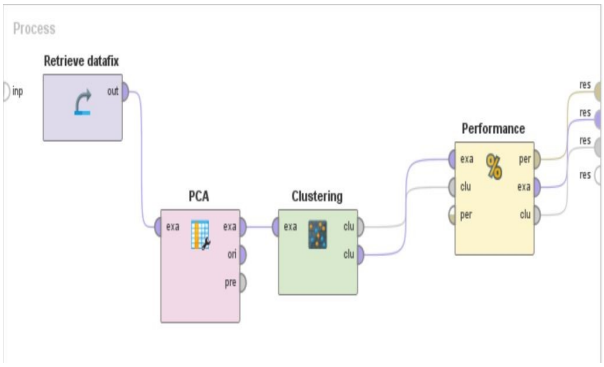
### 2.3. Modelling dan Evaluasi

Pada tahapan ini, dilakukan beberapa langkah untuk modelling menggunakan tools RapidMiner. Ada 2 model yang dibuat dengan tools RapidMiner, yaitu yang pertama adalah Model K-Means dengan jumlah cluster 5 ditampilkan pada gambar 3 seperti dibawah ini :




Sumber : Penulis (2020)
Gambar 3. Proses model clustering K-Means

Model yang kedua adalah model clustering menggunakan metode reduksi dimensi Principle Component Analysis (PCA) yang ditampilan pada gambar 4 seperti berikut ini :




Sumber : Penulis (2020)
Gambar 4. Proses model clustering K-Means + PCA

Langkah dan tahapan dalam proses modelling dan evaluasi adalah :

a. Membuat model clustering dengan metode K-Means dimana jumlah cluster adalah 5 dan dicari nilai DBI-nya

b. Dilanjutkan dengan membuat model clustering metode K-Means + PCA dimana jumlah cluster adalah 5 dan setting pada PCA adalah dimensionality keep variance dengan varian thershold adalah 0,95 yang merupakan nilai standar pada tools RapidMiner dimana semua komponen dengan varians kumulatif yang lebih besar dari ambang yang diberikan dihapus dari ExampleSet dan ambang ditentukan oleh parameter ambang varians. Dari proses yang dilakukan kemudian diambil nilai DBI-nya.
c. Dari 2 model yang ada kemudian dilakukan perbandingan atau komparasi hasil nilai DBI antara model K-Means dan K-Means+PCA. Nilai DBI yang terkecil menunjukkan hasil yang paling baik.

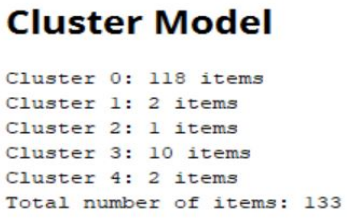
## III. Hasil dan Pembahasan

Proses pengolahan data yang dilakukan dua kali menggunakan model yang berbeda. Yang dilakukan pertama adalah pengolahan data dengan modelling K-Means dan diketahui nilai DBI-nya. Kemudian dilakukan pengolahan data yang kedua dengan modelling K-Means+PCA. Dari pengolahan data yang dilakukan, dihasilkan informasi seperti pada tabel 1 sebagai berikut :

Tabel 1. Nilai DBI

| No | Modelling | Nilai DBI |
|---|---|---|
| 1 | K-Means | 0,382 |
| 2 | K-Means + PCA | 0,310 |

Dari hasil pengolahan diketahui nilai DBI modelling K-Means+PCA adalah lebih kecil dibanding nilai DBI modelling K-Means. Hal ini membuktikan bahwasanya metode reduksi dimensi PCA mampu mendapatkan hasil yang lebih baik. Sehingga model clustering yang digunakan untuk pengelompokan data kunjungan wisman ke Indonesia menggunakan model K-Means+PCA. Adapun hasil clustering ditampilkan pada gambar 4 berikut ini :




Sumber : Penulis (2020)
Gambar 4 : Hasil Clustering

Dari gambar diatas dapat dijelaskan bahwa untuk cluster 0 memiliki anggota 118 negara, cluster 1 memiliki anggota 2 negara, cluster 2 memiliki anggota 1 negara, cluster 3 memiliki 10 anggota negara dan cluster 4 memiliki anggota 2 negara.

Untuk nilai centroid masing-masing cluster ditampilkan pada tabel 5 seperti dibawah ini :

| Attribute | cluster_0 | cluster_1 | cluster_2 | cluster_3 | cluster_4 |
|---|---|---|---|---|---|
| pc_1 | -47309.297 | 617089.832 | 1251730.688 | 125110.038 | 922743.152 |

Sumber : Penulis (2020)
Gambar 5 : Nilai Centroid Tiap Cluster

Kategori kunjungan wisatawan mancanegara (wisman) dibedakan berdasarkan kelompok atau clusternya dapat dilihat sebagai berikut :
1. Cluster 0 : kunjungan wisman sangat rendah
2. Cluster 1 : kunjungan wisman cukup tinggi
3. Cluster 2 : kunjungan wisman sangat tinggi
4. Cluster 3 : kunjungan wisman rendah
5. Cluster 4 : kunjungan wisman tinggi

Dan anggota negara dari masing-masing cluster dapat dilihat pada tabel 2 berikut ini :

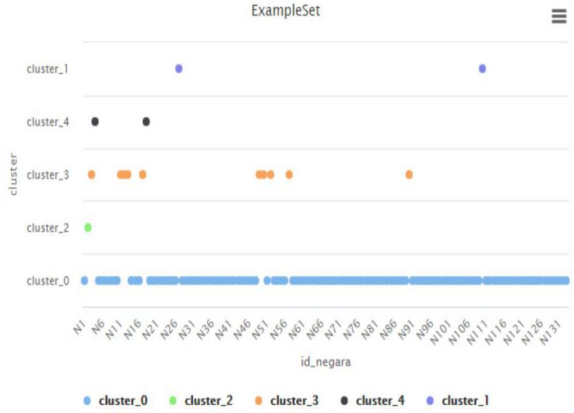
Tabel 2 : Data Negara per cluster

| Cluster | Anggota | Negara |
|---|---|---|
| 0 | N1, N5, N6, N7, N8, N9, N10, N14, N15, N16, N19, N20, N21, N22….. | Brunei Darussalam, Thailand, Vietnam, Laos, Kamboja, Myanmar, Hongkong, Pakistan, Bangladesh, Srilanka, Afganistan, Bhutan,Kazakstan, Kirgistan…… |
| 1 | N27, N110 | Timor Leste, Australia |
| 2 | N2 | Malaysia |
| 3 | N3, N11, N12, N13, N17, N49, N50, N52, N57, N90. | Philippines, India, Japan, South Korea, Taiwan, Prancis, Jerman, Belanda, Inggris dan Amerika. |
| 4 | N4, N18 | Singapore, China |

Dari data diatas diketahui bahwa, Malaysia menjadi negara dengan kunjungan wisman tertinggi (sangat tinggi). Kemudian ada negara Singapore dan China menjadi negara kunjungan wisman tertinggi kedua (tinggi). Disusul ada Timor Leste dan Australia menjadi negara dengan kunjungan wisman tertinggi ketiga (]cukup tinggi). Kemudian ada 10 negara dengan kategori kunjungan wisman yang rendah yaitu Filipina, India, Jepang, Korea Selatan, Taiwan, Prancis, Jerman, Belanda, Inggris dan Amerika. Dan ada 118 negara

lainnya menjadi kategori kunjungan wisatawan mancanegara rendah sekali.

Dari hasil clustering kemudian ditampilkan grafik penyebaran anggota cluster pada gambar 6 dibawah ini :




Sumber : Penulis (2020)

Gambar 6. Penyebaran anggota cluster

## IV. Kesimpulan

Dari pengolahan data yang sudah dilakukan, didapatkan hasil dan kesimpulan bahwa modelling K-Means+PCA menghasilkan nilai DBI yang lebih baik dibanding modelling K-Means. Hasil clustering menghasilkan pengelompokan negara-negara dengan kunjungan wisman sangat tinggi, tinggi, cukup tinggi, rendah dan sangat rendah. Karena keterbatasan waktu dan tenaga, Peneliti menyadari bahwa hasil dari penelitian ini masih jauh dari kata sempurna karena hal tersebut maka untuk penelitian berikutnya bisa dilakukan komparasi dengan algoritma reduksi dimensi yang lain:

## V. REFERENSI

Dash, P., Nayak, M., & Prasad Das, G. (2014). Principal Component Analysis using Singular Value Decomposition for Image Compression. *International Journal of Computer Applications*, *93*(9), 21–27. https://doi.org/10.5120/16243-5795

Hapsari, V. J., & Nuryakin, C. (2019a). ANALISIS PROFIL WISATAWAN MANCANEGARA YANG KELUAR MELALUI PINTU SOEKARNO HATTA DAN NGURAH RAI. *Jurnal Kepariwisataan Indonesia*, *12*(September), 17–30.

Hapsari, V. J., & Nuryakin, C. (2019b). MELALUI PINTU SOEKARNO HATTA DAN NGURAH RAI, *12*(September), 17–30.

Luo, S., Chen, T., & Jian, L. (2018). Using principal component analysis and least squares support vector machine to predict the silicon content in blast furnace system. *International Journal of Online Engineering*, *14*(4), 149–162. https://doi.org/10.3991/ijoe.v14i04.8397

Muningsih, E., & Kiswati, S. (2015). Penerapan Metode K-Means untuk Clustering Produk Online Shop dalam Penentuan Stok Barang. *Jurnal Bianglala Informatika*, *3*(1), 10–17.

Muningsih, E., & Kiswati, S. (2018). Sistem Aplikasi Berbasis Optimasi Metode Elbow Untuk Penentuan Clustering Pelanggan. *Joutica*, *3*(1), 117. https://doi.org/10.30736/jti.v3i1.196

Pariwisata, K., Industri, A., & Regulasi, D. (2019). IDENTIFIKASI POTENSI KUNJUNGAN WISATAWAN MANCANEGARA PADA POS LINTAS BATAS BUILALO DI PROVINSI NUSA TENGGARA TIMUR Identifying Potential Tourist Arrivals From Cross-Border Builalo Post's In East Nusa Tenggara Province Addin Maulana, *13*(2), 67–78.

Sastry, S. H., & Babu, P. M. S. P. (2013). Implementation of CRISP Methodology for ERP Systems, *2*(05), 203–217.

Surtiningsih, L., Furqon, M. T., & Adinugroho, S. (2018). Prediksi Jumlah Kunjungan Wisatawan Mancanegara Ke Bali Menggunakan Support Vector Regression dengan Algoritma Genetika, *2*(8), 2578–2586.