ISSN: 2339-2541
JURNAL GAUSSIAN, Volume 10, Nomor 3, Tahun 2021, Halaman 359 - 366
Online di: https://ejournal3.undip.ac.id/index.php/gaussian/

PENERAPAN *k-MODES CLUSTERING* DENGAN VALIDASI DUNN *INDEX* PADA PENGELOMPOKAN KARAKTERISTIK CALON TKI MENGGUNAKAN R-GUI

Hanik Malikhatin[1*], Agus Rusgiyono[2], Di Asih I Maruddani[3]
[1,2,3] Departemen Statistika, Fakultas Sains dan Matematika, Universitas Diponegoro
***hanikmalikhatinn@gmail.com**

ABSTRACT

Prospective TKI workers who apply for passports at the Immigration Office Class I Non TPI Pati have countries destinations and choose different PPTKIS agencies. Therefore, the grouping of characteristics prospective TKI needed so that can be used as a reference for the government in an effort to improve the protection of TKI in destination countries and carry out stricter supervision of PPTKIS who manage TKI. The purpose of this research is to classify the characteristics of prospective TKI workers with the optimal number of clusters. The method used is k-Modes Clustering with values of k = 2, 3, 4, and 5. This method can agglomerate categorical data. The optimal number of clusters can be determined using the Dunn Index. For grouping data easily, then compiled a Graphical User Interface (GUI) based application with RStudio. Based on the analysis, the optimal number of clusters is two clusters with a Dunn Index value of 0,4. Cluster 1 consists of mostly male TKI workers (51,04%), aged ≥ 20 years old (91,93%), with the destination Malaysia country (47%), and choosing PPTKIS Surya Jaya Utama Abadi (37,51%), while cluster 2, mostly of male TKI workers (94,10%), aged ≥ 20 years old (82,31%), with the destination Korea Selatan country (77,95%), and choosing PPTKIS BNP2TKI (99,78%).

Keywords: prospective TKI workers, cluster, k-Modes Clustering, categorical data, Dunn Index, GUI

1. PENDAHULUAN

Calon TKI yang membuat paspor di Kantor Imigrasi Kelas I Non TPI Pati memiliki tujuan negara serta memilih lembaga PPTKIS yang berbeda-beda. Oleh karena itu, diperlukan pengelompokan karakteristik calon TKI agar dapat dijadikan referensi bagi pemerintah dalam upaya meningkatkan perlindungan TKI di negara tujuan serta melakukan pengawasan yang lebih ketat terhadap PPTKIS yang mengurus TKI. Analisis *cluster* menjadi pilihan peneliti untuk mengetahui bagaimana pengelompokan Calon TKI berdasarkan karakteristiknya.

Pada umumnya analisis *cluster* yang sering digunakan adalah analisis *cluster k-Means*. Pada penerapan metode *k-Means Clustering* ini, data yang digunakan adalah data numerik yang berbentuk angka (Ediyanto *et al.*, 2013). Sementara data calon TKI yang membuat paspor di Kantor Imigrasi Kelas I Non TPI Pati tergolong ke dalam data berskala kategorik sehingga diperlukan metode analisis *cluster* yang dapat mengolah data berskala kategorik. Salah satunya yaitu metode *k-Modes Clustering*.

k-Modes Clustering merupakan suatu metode *clustering* yang dikembangkan dari metode *k-Means Clustering*, sehingga metode *k-Modes Clustering* bersifat efisien seperti metode *k-Means Clustering* namun digunakan pada data yang bersifat kategorik (Indriani dan Budiman, 2017). Dalam penentuan jumlah *cluster* yang optimal menggunakan metode validasi Dunn *Index*. Metode validasi Dunn *Index* ini memberikan skor terbaik untuk algoritma *clustering* yang menghasilkan *cluster* dengan kemiripan tinggi dalam suatu *cluster* namun kemiripan yang rendah antar *cluster-cluster* (Santoso *et al.*, 2020). Dalam mempermudah pengelompokkan data, maka disusun aplikasi berbasis *Graphical User Interface* (GUI) dengan RStudio.

2. TINJAUAN PUSTAKA

2.1. *Clustering*

Analisis *cluster* merupakan suatu analisis statistik yang bertujuan memisahkan objek ke dalam beberapa kelompok yang mempunyai sifat berbeda di antara kelompok satu dengan kelompok yang lain. Tujuan utama dari analisis *cluster* adalah mengelompokkan objek-objek berdasarkan kesamaan karakteristik di antara objek-objek tersebut. Objek bisa berupa produk (barang dan jasa), benda (tumbuhan atau lainnya), serta orang (responden, konsumen atau yang lain). Objek tersebut akan diklasifikasikan ke dalam satu atau lebih *cluster* (kelompok) sehingga objek-objek yang berada dalam satu *cluster* akan mempunyai kemiripan satu dengan yang lain (Heriyati dan Kurniatun, 2020).

Clustering data dapat menggunakan metode *non-hierarchical cluster* dan *hierarchical cluster*. Pada *non-hierarchical cluster*, peneliti menentukan jumlah *cluster* yang diinginkan terlebih dahulu. Setelah itu proses *clustering* dilakukan untuk mengetahui karakteristik (kemiripan) yang ada pada masing-masing *cluster*. Sedangkan pada metode *hierarchical cluster*, dilakukan *clustering* terhadap dua atau lebih objek yang memiliki karakteristik yang paling dekat secara terus-menerus hingga terbentuk hirarki dari yang paling mirip sampai yang paling tidak mirip (Hidayat dan Istiadah, 2011).

2.2. *k-Means Clustering*

k-Means adalah metode *clustering* berbasis jarak yang membagi data ke dalam sejumlah *cluster* dan algoritma ini hanya bekerja pada data numerik. Algoritma *k-Means* sangat terkenal karena kemudahan dan kemampuannya untuk *clustering* data yang besar dan data *outlier* dengan sangat cepat (Wahyudi *et al.*, 2020).

2.3. Variabel Kategorik

Variabel kategorik merupakan salah satu variabel yang nilainya dari sekumpulan kategori yang sering diberi label angka. Variabel kategorik mempunyai dua tipe skala pengukuran yaitu skala ordinal dan skala nominal (Nugraha, 2014).

2.4. *k-Modes Clustering*

k-Modes Clustering merupakan pendekatan nonparametrik untuk mendapatkan *cluster* dari data kategorik (Huang dan Ng, 2003). *k-Modes Clustering* pertama kali diperkenalkan oleh Huang (1998) sebagai suatu metode *clustering* yang dikembangkan dari metode *k-Means Clustering*, sehingga metode *k-Modes Clustering* bersifat efisien seperti metode *k-Means Clustering* namun digunakan pada data yang bersifat kategorik. Modifikasi yang dilakukan terhadap metode *k-Means Clustering* yaitu:

1. Jarak antara dua titik data X dan Y adalah jumlah fitur pada X dan Y yang nilainya berbeda (*simple dissimilarity measure*), secara formal dirumuskan seperti berikut ini:
   $$d_1(X,Y) = \sum_{j=1}^{m} \delta(x_j, y_j) \quad (1)$$
   dengan
   $$\delta(x_j, y_j) = \begin{cases} 0 & (x_j = y_j) \\ 1 & (x_j \neq y_j) \end{cases} \quad (2)$$
   x_j dan y_j adalah nilai fitur ke-j dari data X dan Y, serta m adalah jumlah fitur.
2. Mengubah rata-rata (*means*) menjadi modus (*modes*).
3. Menggunakan frekuensi untuk mencari modus dalam proses pembentukan *centroid*.

Langkah-langkah *k-Modes Clustering* berdasarkan (Huang, 2008) adalah sebagai berikut:

1. Menentukan k *cluster* dari data dan melakukan inisialisasi *cluster* secara acak.
2. Mengalokasikan objek data pada *cluster* terdekat berdasarkan *simple dissimilarity measure. Update* tiap modus *cluster* setelah tiap alokasi.
3. Setelah semua objek data dialokasikan ke suatu *cluster*, memeriksa kembali nilai *dissimilarity* tiap objek terhadap modus. Jika suatu objek data memiliki modus terdekat berada pada *cluster* lain, maka memindahkan objek ke *cluster* yang sesuai dan *update* modus kedua *cluster*.
4. Mengulangi langkah 3 sampai tidak ada objek data yang berubah *cluster* (Indriani dan Budiman, 2017).

2.5. Dunn *Index*

Dunn *Index* merupakan salah satu metode validasi *cluster*. Metode ini memberikan skor terbaik untuk algoritma *clustering* yang menghasilkan *cluster* dengan kemiripan tinggi dalam suatu *cluster* namun kemiripan yang rendah antar *cluster-cluster*. Dunn *Index* bertujuan untuk mengidentifikasi *cluster* yang terpisah dengan baik. Metode ini menghitung rasio jarak antar *cluster* minimal dengan jarak intra-*cluster* maksimal. Dengan demikian semakin tinggi nilai Dunn *Index*, maka semakin optimal jumlah *cluster* yang dihasilkan. Menurut Ansari (2011), Dunn *Index* dirumuskan sebagai berikut:

$$D = \left\{ \frac{min_{\ 1\le i\le k}\ _{i+1\le j\le q}\left(d(C_i,C_j)\right)}{max_{1\le l\le q}d(C_l)} \right\} \qquad (3)$$

dengan

$d(C_i, C_j)$: ukuran kedekatan antara *cluster* i dan *cluster* j.

$d(C_l)$: ukuran kedekatan antar anggota dalam *cluster* l (Pratiwi *et al.*, 2019).

2.6. RStudio dan *Graphical User Interface* (GUI)

RStudio merupakan *tool* pemrograman atau *integrated development environment* (IDE) bahasa R yang memiliki antarmuka lebih baik daripada RGui (Faisal dan Nugrahadi, 2019).

RStudio dapat digunakan untuk pengolahan metode *k-Modes Clustering* dan Dunn *Index*. Pada pengolahan *k-Modes Clustering*, diperlukan sintaks RStudio sebagai berikut:

```
kmodes(data, modes, iter.max = 10, weighted = FALSE, fast = TRUE)
```

(https://www.rdocumentation.org).

Sedangkan pada pengolahan Dunn *Index*, diperlukan sintaks RStudio sebagai berikut:

```
dunn(distance = NULL, clusters, Data = NULL, method = ”euclidean”)
```

(https://www.rdocumentation.org).

Graphical User Interface (GUI) adalah tampilan grafis yang memudahkan *user* atau pengguna berinteraksi dengan perintah teks (Chapman, 2001 dalam Kurniastuti dan Andini, 2018). *Packages* utama yang dibutuhkan dalam pembuatan GUI salah satunya adalah *shiny* yang terdiri dari 3 komponen yaitu *User Interface* (*ui*), *Server*, dan *ShinyApp* (https://medium.com).

3. METODE PENELITIAN

3.1. Sumber Data, Variabel Penelitian, dan Alat Analisis

Data yang digunakan dalam penelitian tugas akhir ini adalah data sekunder yang diperoleh dari hasil rekapan Bidang Informasi dan Komunikasi Keimigrasian Kantor Imigrasi Kelas I Non TPI Pati. Data tersebut merupakan data pemohon paspor calon TKI pada tahun 2019 di daerah kerja utama yang meliputi Kabupaten Pati, Kabupaten Rembang, Kabupaten Blora, dan Kabupaten Jepara.

Variabel yang digunakan disajikan dalam tabel 1 sebagai berikut:

Tabel 1. Variabel Penelitian

Variabel	Nama Variabel
X_1	Jenis Kelamin
X_2	Umur
X_3	Negara yang Dituju Calon TKI
X_4	PPTKIS (Pelaksana Penempatan Tenaga Kerja Indonesia Swasta) yang Mengurus Calon TKI

Data penelitian diolah menggunakan *software* Microsoft Excel 2010 dan RStudio 1.0.153.

3.2. Metode Analisis

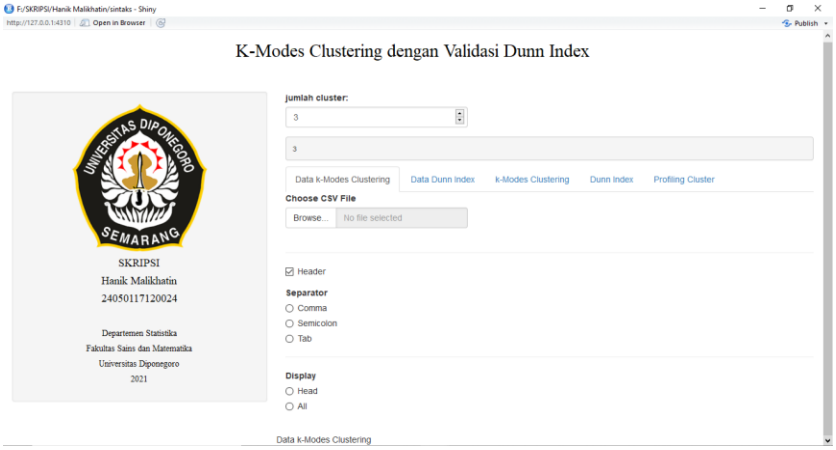
Langkah-langkah analisis data yang akan dilakukan adalah sebagai berikut:

1. Melakukan *pre-processing* dan pelabelan data.
2. Melakukan transformasi data ke dalam bentuk data biner.
3. Menyusun *User Interface* dan *Server* pada GUI.
4. *Running* Program GUI.
5. Memasukkan data.
6. Menentukan nilai *k* (jumlah *cluster*) yaitu *k* = 2, 3, 4, dan 5.
7. Melakukan *k-Modes Clustering*.
8. Melakukan validasi Dunn *Index*l.
9. Memilih *k* optimal.
10. Melakukan *profiling* hasil *k-Modes Clustering* dengan *k* optimal.

4. HASIL DAN PEMBAHASAN

4.1. Pembuatan *Graphical User Interface* (GUI)

Program GUI yang dibangun dalam penelitian ini terdiri dari beberapa bagian yaitu: halaman judul, input data *k-Modes Clustering*, input data Dunn *Index*, hasil *k-Modes Clustering*, hasil Dunn *Index*, dan hasil *Profiling Cluster*. *Packages* utama yang dibutuhkan dalam pembuatan GUI pada penelitian ini adalah *shiny* yang terdiri dari 3 komponen yaitu *User Interface* (*ui*), *Server*, dan *ShinyApp*. Setelah kedua fungsi pada paket *shiny* yaitu *ui* dan *server* disatukan dengan perintah *ShinyApp*(*ui*,*server*), maka dapat dijalankan melalui jendela RStudio dengan mengklik tombol *Run App* dan menghasilkan tampilan GUI pada Gambar 1.



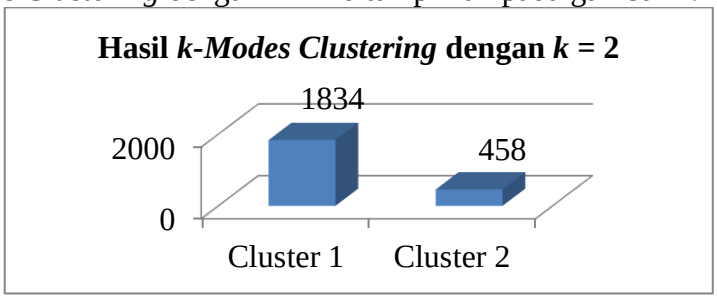
Gambar 1. Tampilan Awal GUI

Pengelompokan data menggunakan *k-Modes Clustering* dengan validasi Dunn *Index* dapat dijalankan dengan memilih data melalui mengklik tombol *Browse* pada bagian Data *k-Modes Clustering* dan Data Dunn *Index* serta memilih jumlah *cluster* sesuai keinginan.

4.2. Analisis *k-Modes Clustering*

Pada proses *clustering* menggunakan metode *k-Modes Clustering* dilakukan pada berbagai nilai *k* (*k* = 2, 3, 4, dan 5).

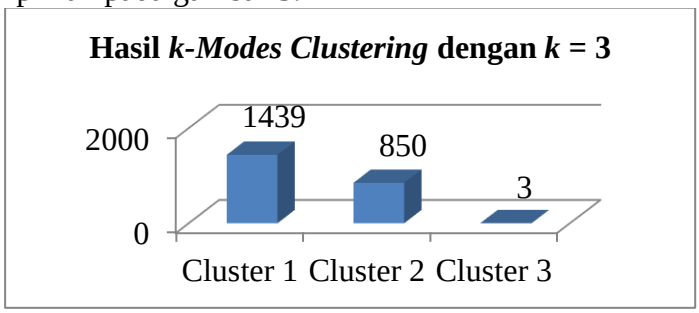
Hasil *k-Modes Clustering* dengan *k* = 2 ditampilkan pada gambar 2.



Gambar 2. Histogram Frekuensi Hasil *k-Modes Clustering* dengan *k* = 2

Pada gambar 2 terlihat bahwa anggota *cluster* 1 sebanyak 1.834 orang dan anggota *cluster* 2 sebanyak 458 orang.

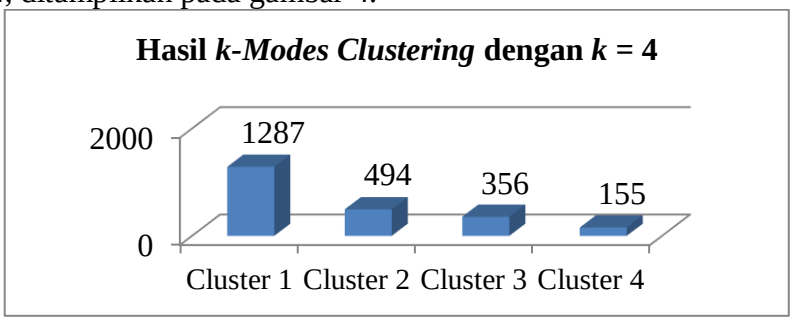
Pada *k* = 3, ditampilkan pada gambar 3.



Gambar 3. Histogram Frekuensi Hasil *k-Modes Clustering* dengan *k* = 3

Pada gambar 3 terlihat bahwa anggota *cluster* 1 sebanyak 1.439 orang, anggota *cluster* 2 sebanyak 850 orang, dan anggota *cluster* 3 sebanyak 3 orang.

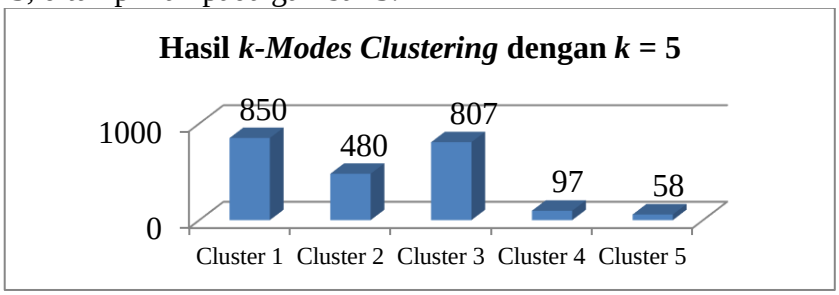
Pada k = 4, ditampilkan pada gambar 4.



Gambar 4. Histogram Frekuensi Hasil *k-Modes Clustering* dengan k = 4

Pada gambar 4 terlihat bahwa anggota *cluster* 1 sebanyak 1.287 orang, anggota *cluster* 2 sebanyak 494 orang, anggota *cluster* 3 sebanyak 356 orang, dan anggota *cluster* 4 sebanyak 155 orang.

Pada k = 5, ditampilkan pada gambar 5.



Gambar 5. Histogram Frekuensi Hasil *k-Modes Clustering* dengan k = 5

Pada gambar 5 terlihat bahwa anggota *cluster* 1 sebanyak 850 orang, anggota *cluster* 2 sebanyak 480 orang, anggota *cluster* 3 sebanyak 807 orang, anggota *cluster* 4 sebanyak 97 orang, dan anggota *cluster* 5 sebanyak 58 orang.

4.3. Dunn *Index*

Berdasarkan hasil olahan RStudio, diperoleh hasil pada tabel 2.

Tabel 2. Tabel Nilai Dunn *Index*

k	Nilai Dunn *Index*
2	0,4
3	0
4	0
5	0

Berdasarkan tabel 6 dapat disimpulkan bahwa *cluster* dengan k = 2 lebih optimal dibandingkan *cluster* dengan k = 3, 4, dan 5. Hal ini dikarenakan nilai Dunn *Index* pada *cluster* dengan k = 2 paling besar dibandingkan yang lainnya.

4.4. *Profiling Cluster* dengan *k* optimal (*k* = 2)

Hasil pengolahan metode *k-Modes Clustering* dengan k = 2 menggunakan RStudio diperoleh jumlah *cluster* 1 sebanyak 1.834 orang calon TKI dan *cluster* 2 sebanyak 458 orang calon TKI.

Berdasarkan hasil olahan RStudio pada *profiling cluster* 1, contoh datanya ditampilkan pada tabel 3.

Tabel 3. Tabel Anggota *Cluster* 1 dengan k = 2

No	Nama	X1	X2	X3	X4
1	Sutiyah	2	2	13	9
2	Febri Puguh Trisnani	2	2	13	9

3	Zubaidah	2	2	13	33
4	Tri Enggar Budiarti	2	2	13	33
5	Ahmad Jauhari	1	2	13	43
6	Lestari	2	2	13	28
7	Ruhyani	2	2	13	28
8	Nur Akimil Laela	2	2	5	28
9	Asmini	2	2	5	28
10	Lutfi Fitriani	2	2	13	28
.	.	.	.	.	.
.	.	.	.	.	.
1830	Dicky Prasetyo Aji	1	2	13	43
1831	Dandang Handoko	1	2	13	43
1832	Surini	2	2	12	7
1833	Suharti Bt Sukarjo Karyono	2	2	12	9
1834	Leginah	2	2	12	7

Kemudian mencari frekuensi masing-masing variabel pada *profiling cluster* 1. Setelah dicari frekuensinya, dapat disimpulkan bahwa *cluster* 1 sebagian besar terdiri dari calon TKI yang berjenis kelamin laki-laki, berumur ≥ 20 tahun, dengan negara yang dituju Negara Malaysia dan memilih PPTKIS Surya Jaya Utama Abadi.

Sedangkan pada *profiling cluster* 2, contoh datanya ditampilkan pada tabel 4.

Tabel 4. Tabel Anggota *Cluster* 2 dengan $k = 2$

No	Nama	X1	X2	X3	X4
1	Haryanto	1	2	7	8
2	Suparti	2	2	7	8
3	Ahmad Rokhimul Ibad	1	1	6	8
4	Ipung Supriyanto	1	2	6	8
5	Mohamad Abdul Rohman	1	2	6	8
6	Okik Setiyawan	1	2	6	8
7	Pariyono	1	2	13	8
8	Mohammad Junaedi	1	2	6	8
9	Prehatin	1	1	6	8
10	Dewi Purwati	2	2	6	8
.	.	.	.	.	.
.	.	.	.	.	.
454	Yeni Husada	1	2	7	8
455	Bambang Sugiyanto	1	1	7	8
456	Muhammad Ridwan	1	2	7	8
457	Adi Kuncoro	1	2	7	8
458	Ahmad Rusdiono	1	2	7	8

Kemudian mencari frekuensi masing-masing variabel pada *profiling cluster* 2. Setelah dicari frekuensinya, dapat disimpulkan bahwa *cluster* 2 sebagian besar terdiri dari calon TKI yang berjenis kelamin laki-laki, berumur ≥ 20 tahun, dengan negara yang dituju Negara Korea Selatan dan memilih PPTKIS BNP2TKI.

5. KESIMPULAN

Setelah dilakukan analisis *cluster* menggunakan metode *k-Modes Clustering* dan Dunn *Index* pada 2.292 pemohon paspor calon TKI pada tahun 2019 di daerah kerja utama, dapat disimpulkan bahwa *Graphycal User Interface* (GUI) dapat mempermudah pengguna untuk melakukan komputasi *k-Modes Clustering* dengan validasi Dunn *Index* menggunakan jumlah *cluster* sesuai keingingan. Hasil dari *k-Modes Clustering* dengan Validasi Dunn *Index* terbentuk dua *cluster* dengan banyaknya *cluster* 1 sebanyak 1.834 orang dan *cluster* 2 sebanyak 458 orang. Jumlah *cluster* terbaik yang dihasilkan menggunakan metode *k-Modes Clustering* adalah dua *cluster* ($k = 2$) dengan memiliki nilai Dunn *Index* paling

besar yaitu 0,4. Pada *cluster* 1 paling banyak terdiri dari calon TKI yang berjenis kelamin laki-laki (51,04%), berumur ≥ 20 tahun (91,93%), dengan negara yang dituju Malaysia (47%), dan memilih PPTKIS Surya Jaya Utama Abadi (37,51%), sedangkan *cluster* 2 paling banyak terdiri dari calon TKI yang berjenis kelamin laki-laki (94,10%), berumur ≥ 20 tahun (82,31%), dengan negara yang dituju Korea Selatan (77,95%), dan memilih PPTKIS BNP2TKI (99,78%).

DAFTAR PUSTAKA

DataCamp. -. *kmodes: K-Modes Clustering*. Tersedia: https://www.rdocumentation.org/packages/klaR/versions/0.6-15/topics/kmodes (diakses pada tanggal 6 April 2021).

DataCamp. -. dunn: Dunn Index. Tersedia: https://www.rdocumentation.org/packages/clValid/versions/0.6-9/topics/dunn (diakses pada tanggal 7 April 2021).

Ediyanto, Mara, M. N. dan Satyahadewi, N. 2013. Pengklasifikasian Karakteristik Dengan Metode K-Means Cluster Analysis. *Buletin Ilmiah Mat. Stat. dan Terapannya (Bimaster)* Vol. 02, No. 2, Hal: 133–136.

Faisal, M. R. dan Nugrahadi D. T. 2019. *Belajar Data Science Klasifikasi dengan Bahasa Pemrograman R*. Kalimantan Selatan: Scripta Cendekia.

Hakim, R. B. F. 2019. *Bermain dengan R Shiny*. Tersedia: https://medium.com/@986110101/bermain-dengan-r-shiny-b3430fc7ae5f (diakses pada tanggal 19 Juli 2021).

Heriyati, P. dan Kurniatun, T. C. 2020. *Analisa Triple Helix pada Industri Fashion di Jakarta*. Pasuruan: Qiara Media.

Hidayat, T. dan Istiadah, N. 2011. *Panduan Lengkap Menguasai SPSS 19 untuk Mengolah Data Statistik Penelitian*. Jakarta Selatan: MediaKita.

Huang, Z. dan Ng, M. K. 2003. A Note on K-modes Clustering. *Journal of Classification* Vol. 20, No. 2, Hal: 257–261.

Indriani, F. dan Budiman, I. 2017. K-Modes Clustering untuk Mengetahui Jenis Masakan Daerah yang Populer pada Website Resep Online (Studi Kasus: Masakan Banjar di cookpad.com. *Jurnal Teknologi Informasi dan Ilmu Komputer* Vol. 4, No. 4, Hal: 290–296.

Kurniastuti dan Andini. 2018. Perancangan Program Penentuan Histogram Citra dengan Graphical User Interface (GUI). *Applied Technology and Computing Science Journal* Vol. 1, No. 1, Hal:-.

Nugraha, J. 2014. *Pengantar Analisis Data Kategorik: Metode dan Aplikasi Menggunakan Program R*. Yogyakarta: Deepublish.

Pratiwi, S. I., Widiharih, T. dan Hakim, A. R. 2019. Analisis Klaster Metode *Ward* dan *Average Linkage* dengan Validasi *Dunn Index* dan Koefisien Korelasi *Cophenetic* (Studi Kasus: Kecelakaan Lalu Lintas Berdasarkan Jenis Kendaraan Tiap Kabupaten/Kota di Jawa Tengah Tahun 2018). *Jurnal Gaussian* Vol. 8, No. 4, Hal: 486-495.

Santoso, B., Azis, A. I. S. dan Zohrahayaty. 2020. *Machine Learning & Reasoning Fuzzy Logic Algoritma, Manual, Matlab, & Rapid Miner*. Yogyakarta: Deepublish.

Wahyudi, M., Masitha, M., Saragih, R. dan Solikhun, S. 2020. *Data Mining: Penerapan Algoritma K-Means Clustering dan K-Medoids Clustering*. Medan: Yayasan Kita Menulis.