PENGELOMPOKAN MAHASISWA BERBASIS CATEGORICAL VARIABLES MENGGUNAKAN METODE K-MODES CLUSTERING

Heribertus Yulianton[1], Felix Andreas Sutanto[2], Sri Mulyani[3]
[1,2]Program Studi Teknik Informatika,[3]Program Studi Manajemen Informatika
Fakultas Teknologi Informasi, Universitas Stikubank
[1]heri@edu.unisbank.ac.id, [2]felix@edu.unisbank.ac.id, [3]srimulyani@edu.unisbank.ac.id

ABSTRAK

Universitas Stikubank setiap tahunnya menerima mahasiswa dari berbagai wilayah dan berbagai latar belakang yang berbeda. Karena berasal dari lingkungan yang berbeda, profile mahasiswa adalah hal menarik yang perlu dianalisa.

Penelitian ini bertujuan untuk melakukan pengelompokan terhadap data mahasiswa Universitas Stikubank berbasis atribut (Categorical Variables) dengan menggunakan teknik K-Modes Clustering. Metode K-Modes mempunyai beberapa kelebihan, yaitu bisa diterapkan untuk data kategorikal, dalam menghasilkan klaster prosesnya lebih rinci, waktu komputasi pembentukan klaster lebih singkat, dan unggul dalam klasterisasi pada data berdimensi banyak.

Diharapkan dari penelitian ini dapat diketahui kelompok-kelompok mahasiswa berdasarkan kategori-kategori atau atribut-atribut yang mereka miliki. Hasil pengelompokan dapat digunakan untuk keperluan pemetaan mahasiswa, promosi dimasa mendatang dan hal-hal lain yang berguna untuk manajemen universitas.

Kata Kunci: *clustering, k-modes, categorical variables*

1. PENDAHULUAN

Mahasiswa adalah salah satu aset yang dimiliki oleh setiap perguruan tinggi termasuk Universitas Stikubank. Karena berasal dari lingkungan yang berbeda, profile mahasiswa adalah hal menarik yang perlu dianalisa.

Penelitian ini bertujuan untuk melakukan pengelompokan terhadap data mahasiswa Universitas Stikubank berbasis atribut (Categorical Variables) dengan menggunakan teknik K-Modes Clustering. Metode K-Modes mempunyai beberapa kelebihan, yaitu bisa diterapkan untuk data kategorikal, dalam menghasilkan klaster prosesnya lebih rinci, waktu komputasi pembentukan klaster lebih singkat, dan unggul dalam klasterisasi pada data berdimensi banyak. Penelitian di bidang klasterisasi data kategorikal sudah mulai berkembang, walaupun perkembanganya masih jauh lebih sedikit dibanding klasterisasi pada tipe data numerik. Data kategorikal secara alami tidak bisa diperlakukan sebagai data numerik karena ada beberapa operasi dalam data numerik yang tidak bisa di lakukan dalam data kategorikal seperti mean dan median. Sebagai contoh atribut data kategorikal adalah atribut berdomain jenis kelamin (pria, wanita), domain agama (Islam, Kristen, Katolik, Hindu dan sebagainya), dan domain etnis (mongoloid, kaukasoid, negroid).

Diharapkan dari penelitian ini dapat diketahui kelompok-kelompok mahasiswa berdasarkan kategori-kategori atau atribut-atribut yang mereka miliki. Hasil pengelompokan dapat digunakan untuk keperluan pemetaan mahasiswa, promosi dimasa mendatang dan hal-hal lain yang berguna untuk manajemen universitas.

2. TINJAUAN PUSTAKA

Penelitian di bidang klasterisasi data kategorikal sudah mulai berkembang, walaupun perkembanganya masih jauh lebih sedikit dibanding klasterisasi pada tipe data numerik. Data kategorikal secara alami tidak bisa diperlakukan sebagai data numerik karena ada beberapa operasi dalam data numerik yang tidak bisa di lakukan dalam data kategorikal seperti mean dan median.

KModes merupakan pengembangan dari metode KMeans agar dapat di gunakan untuk klasterisasi data kategorikal. K-Modes menggunakan sebuah ukuran jarak (dissimilarity) berupa kecocokan suatu nilai atribut tiap dimensi terhadap titik pusat klaster, menggantikan mean dengan modus, dan menggunakan metode berbasis frekuensi untuk memutakhirkan modus dalam proses meminimalkan jarak (dissimilarity) dari seluruh data ke pusat klaster masing-masing. Karena KModes yang dikembangkan Huang merupakan pengembangan K-Means, maka K-Modes mempunyai karakteristik dan kelemahan yang sama dengan K-Means. Kelemahan tersebut adalah keakuratan hasil klaster sangat tergantung dari penentuan titik awal pusat klaster sehingga sensitif terhadap penentuan titik awal.

Tanti dan Deden [1] mengimplementasikan Algoritma K-modes untuk Penentuan Prioritas Rehabilitasi Daerah Aliran Sungai Berdasarkan Parameter Lahan Kritis. Data yang diolah hanya empat parameter sesuai dengan Permenhut Nomor P.32/MenhutII/2009. Data penelitian ini bertipe kategorik sesuai dengan data yang dibutuhkan dalam metode yang diusulkan yaitu skor penutupan lahan, skor lereng, skor erosi, skor produktivitas dan skor manajemen. Pada studi kasus penentuan DAS prioritas di Kabupaten Wonogiri pada kawasan hutan lindung yang didasarkan atas empat parameter dengan menerapkan algoritma k-modes memberikan hasil pengelompokan yang baik.

Fatma dan Irwan [2] melakukan clustering untuk mengetahui jenis masakan daerah yang populer pada website resep online cookpad.com.Metode clustering yang dipilih adalah k-modes karena cocok digunakan pada data kategorikal. Berdasar metode Elbow, jumlah cluster yang ideal adalah k=4 dan k=8. Jumlah cluster k=4 menghasilkan kelompok yang lebih umum, sedangkan k=8 menghasilkan kelompok yang lebih spesifik.

Ahmad dkk [3] menerapkan k-modes dengan validasi Davies Bouldin Index dalam menentukan karakteristik kanal Youtube di indonesia menurut Socialblade. Jumlah klaster terbaik dapat diperiksa menggunakan Davies-Bouldin Index (DBI).Data pada penelitian ini menggunakan data primer berupa kuesioner tentang pertimbangan pemilihan menari sebagai kegiatan ekstrakurikuler. Penelitian ini menggunakan 5 variabel, yaitu pilihan kegiatan menari, ajakan menari, banyaknya ikut lomba, rutinitas latihan tiap 1 minggu dan lama latihan.Berdasarkan hasil perhitungan dan pembahasan dapat disimpulkan bahwa cluster paling optimal dengan menggunakan K-Modes 2 cluster. Nilai validitas Davies-Bouldin Index (DBI) yang dihasilkan sebesar 0,52.

3. METODE PENELITIAN

Pada penelitian ini, data yang diambil adalah data mahasiswa yang masuk pada tahun 2018 sampai 2019. Data diambil dari sistem informasi Smart Campus Universitas Stikubank.Ada banyak atribut pada data mahasiswa, namun untuk keperluan proses hanya digunkan atribut Prodi, Kelamin, Darah, Kota, PendidikanAyah, PekerjaanAyah, PendidikanIbu, PekerjaanIbu, dan JurusanSekolah saja. Selain pemilihan atribut, data mahasiswa yang akan diproses hanya data mahasiswa pada jenjang D3 dan S1 saja, data mahasiswa S2 tidak diikut sertakan dalam penelitian ini. Hal ini karena tujuan akhir dari penelitian ini untuk kebutuhan promosi jenjang D3 dan S1. Setelah data mahasiswa S2 dibersihkan, terdapat 2011 data mahasiswa D3 dan S1. Selain itu, untuk proses data mining dilakukan pembersihan pada data yang atributnya kosong atau tidak terisi. Data tersebut tidak diikutkan dalam proses pengelompokan. Akhirnya didapatkan 1684 record mahasiswa.

Pada penelitian ini, peneliti memanfaatkan Colaboratory milik Google yang dapat diakses melalui https://colab.research.google.com. Library Python yang digunakan dalam penelitian ini adalah Numpy, Pandas, Mathplotlib dan Seaborn. Numpy dan Pandas digunakan untuk membantu proses pengolahan data dengan K-Modes, sedangkan Mathplotlib dan Seaborn digunakan untuk memvisualisasikan data.Untuk pengolahan data mahasiswa dengan K-Modes, digunakan bantuan library Scikit-learn. Scikit-learn adalah library untuk machine learning dan pemodelan statistik termasuk classification, regression, clustering and dimensionality reduction.

Proses klastering dengan K-Modes diawali dengan mempersiapkan data agar mudah digunakan dalam pemrosesan. Library Scikit-learn dapat digunakan untuk mengubah secara otomatis nilai data-data yang bersifat kategorikal menjadi suatu nilai numerik. Sebelas atribut mahasiswa akan diubah nilainya menjadi kode yang lebih praktis untuk pemrograman. Hasil konversi nilai-nilai atribut dapat dilihat pada gambar 1.

	ProdiID	Kelamin	DARAH	Kota	PendidikanAyah	PekerjaanAyah	PendidikanIbu	PekerjaanIbu	JurusanSekolah	TahunLulus	PAGI
0	10	1	3	146	6	0	6	0	29	16	1
1	12	1	3	146	6	0	6	1	143	16	0
2	8	1	3	146	0	0	6	1	2	17	1
3	9	1	3	146	7	0	6	1	86	18	1
4	12	0	0	83	8	1	5	1	196	11	0
5	0	1	0	41	5	1	5	1	60	14	0
6	11	0	0	116	5	1	5	1	209	16	1
7	10	1	0	166	5	1	5	1	60	17	1
8	9	1	0	41	5	1	5	1	60	17	1
9	8	1	0	25	7	1	5	1	60	17	1

Gambar 1. Konversi Nilai

Penentuan titik awal dalam algoritma K-Modes dapat diambil secara random, tetapi hal tersebut dapat menyebabkan terjadinya iterasi yang tidak dapat diprediksi jumlah dan akurasinya. Pada penelitian ini, akan menggunakan metode Cao.Pusat klaster akan memberikan dampak secara langsung terhadap informasi klaster di akhir iterasi [4]. Pada penelitiannya, Cao juga membahas tentang penggunaan algoritma MaxMin dengan mengembangkan penggunaan jumlah rata-rata frekuensi untuk mengganti penggunaan inisialisasi secara random pada algoritma MaxMin yang bermanfaat dalam menghasilkan pusat klaster yang akan digunakan untuk mengganti pusat klaster pada algoritma k-modes.

Pada penelitian ini untuk menentukan jumlah klaster yang ideal dilakukan dengan membandingkan cost setiap jumlah klaster. Untuk perbandingannya digunakan sampel k mulai 1 sampai 5. Dari jumlah 1 klaster hingga 5 klaster akan dicari yang paling efektif. Cost dihitung seperti pada gambar 2, dan didapatkan hasil sebanyak 2 klaster.

```
Init: initializing centroids
Init: initializing clusters
Starting iterations...
Run 1, iteration: 1/100, moves: 0, cost: 9508.0
Init: initializing centroids
Init: initializing clusters
Starting iterations...
Run 1, iteration: 1/100, moves: 66, cost: 8883.0
Run 1, iteration: 2/100, moves: 51, cost: 8883.0
Init: initializing centroids
Init: initializing clusters
Starting iterations...
Run 1, iteration: 1/100, moves: 52, cost: 8683.0
Init: initializing centroids
Init: initializing clusters
Starting iterations...
Run 1, iteration: 1/100, moves: 148, cost: 8356.0
Run 1, iteration: 2/100, moves: 30, cost: 8356.0
```

Gambar 2. Perhitungan Cost

4. HASIL DAN PEMBAHASAN

4.1. Proses K-Modes

Setelah dilakukan pemilihan metode inisialisasi dan jumlah klaster yang tepat, dilakukan prediksi pengelompokan mahasiswa dengan dua klaster. Hasil prediksi kemudian digabungkan dengan data awal mahasiswa agar mendapatkan informasi yang jelas. Prosesnya dapat dilihat pada gambar 3.Hasil prediksi dari klaster mahasiswa dapat dilihat pada gambar 4. Data mahasiswa telah ditambahkan sebuah field yaitu cluster_predicted, yang artinya setiap record dengan kemiripian data tertentu telah dikelompokkan pada klaster yang sama juga.Klaster pertama memiliki 1367 anggota yang tersebar dari awal hingga akhir record. Sedangkan klaster kedua hanya memiliki 317 anggota yang tersebar diantara record ke 4 hingga 1679.

```
[15] #choosing K=2
     km_cao = KModes(n_clusters=2, init = "Cao", n_init = 1, verbose=1)
     fitClusters_cao = km_cao.fit_predict(mhsw)

     Init: initializing centroids
     Init: initializing clusters
     Starting iterations...
     Run 1, iteration: 1/100, moves: 66, cost: 8883.0
     Run 1, iteration: 2/100, moves: 51, cost: 8883.0

[16] fitClusters_cao

     array([0, 0, 0, ..., 0, 0, 0], dtype=uint16)

[17] mhsw = mhsw_cust_copy.reset_index()

[18] clustersDf = pd.DataFrame(fitClusters_cao)
     clustersDf.columns = ['cluster_predicted']
     combinedDf = pd.concat([mhsw, clustersDf], axis = 1).reset_index()
     combinedDf = combinedDf.drop(['index', 'level_0'], axis = 1)
```

Gambar 3. Proses K-Modes

```
[29] combinedDf.head(10)
```

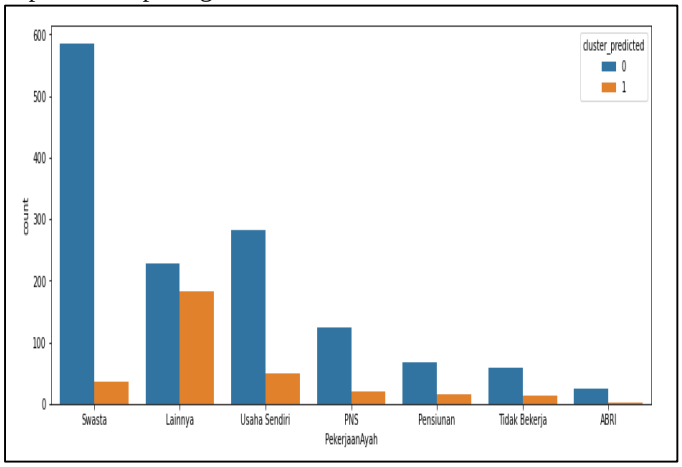
	ProdiID	Kelamin	DARAH	Kota	PendidikanAyah	PekerjaanAyah	PendidikanIbu	PekerjaanIbu	JurusanSekolah	TahunLulus	PAGI	cluster_predicted
0	55	W	O	SEMARANG	Tamat SMA	ABRI	Tamat SMA	ABRI	Animasi	2018	Y	0
1	83	W	O	SEMARANG	Tamat SMA	ABRI	Tamat SMA	Lainnya	TEKNIK ELEKTRONIKA	2018	N	0
2	51	W	O	SEMARANG	Diploma	ABRI	Tamat SMA	Lainnya	ADMINISTRASI PERKANTORAN	2019	Y	0
3	52	W	O	SEMARANG	Tamat SMP	ABRI	Tamat SMA	Lainnya	MULTIMEDIA	2020	Y	0
4	83	P	A	KENDAL	Tidak Tamat SD	Lainnya	Tamat SD	Lainnya	Teknik Furniture	2013	N	1
5	11	W	A	GROBOGAN	Tamat SD	Lainnya	Tamat SD	Lainnya	IPA	2016	N	1
6	62	P	A	PAGARALAM	Tamat SD	Lainnya	Tamat SD	Lainnya	USAHA PERJALANAN WISATA	2018	Y	1
7	55	W	A	TEGAL	Tamat SD	Lainnya	Tamat SD	Lainnya	IPA	2019	Y	1
8	52	W	A	GROBOGAN	Tamat SD	Lainnya	Tamat SD	Lainnya	IPA	2019	Y	1
9	51	W	A	Bumiayu	Tamat SMP	Lainnya	Tamat SD	Lainnya	IPA	2019	Y	1

Gambar 4. Hasil Prediksi Klaster

4.2. Visualisasi Hasil

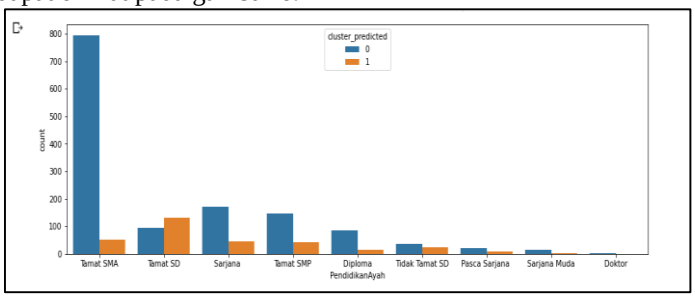
Tujuan dari klasterisasi mahasiswa pada penelitian ini adalah untuk mendapatkan gambaran profile mahasiswa dilihat dari aspek program studi, jenis kelamin, golongan darah, asal kota, pendidikan ayah, pekerjaan ayah, pendidikan ibu, pekerjaan ibu, jurusan sekolah, tahun lulus dan kelompok.

Berdasarkan pekerjaan ayah, dapat dilihat bahwa hasil tertinggi untuk pekerjaan ayah adalah pegawai swasta. Visualisasi klaster dapat dilihat pada gambar 5.



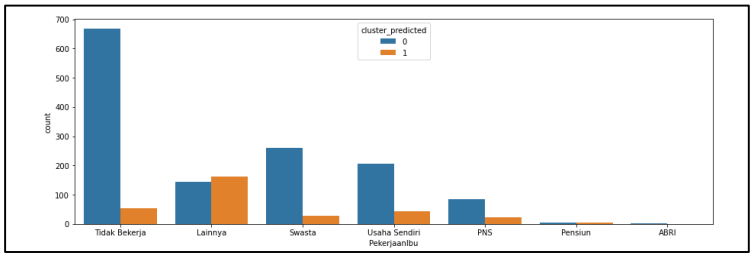
Gambar 5. Klaster Berdasarkan Pekerjaan Ayah

Berdasarkan pendidikan ayah, dapat dilihat bahwa hasil tertinggi untuk pendidikan ayah adalah tamatan SMA. Visualisasi klaster dapat dilihat pada gambar 6.



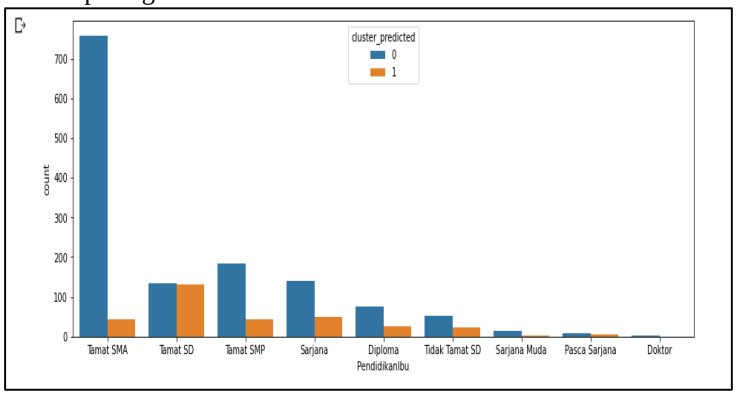
Gambar 6. Klaster Berdasarkan Pendidikan Ayah

Berdasarkan pekerjaan ibu, dapat dilihat bahwa hasil tertinggi untuk pekerjaan ibu adalah tidak bekerja. Visualisasi klaster dapat dilihat pada gambar 7.



Gambar 7. Klaster Berdasarkan Pekerjaan Ibu

Berdasarkan pendidikan ibu, dapat dilihat bahwa hasil tertinggi untuk pendidikan ibu adalah tamatan SMA. Visualisasi klaster dapat dilihat pada gambar 8.



Gambar 7. Klaster Berdasarkan PendidikanIbu

5. KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian penulis menyimpulkan sebagai berikut:

1. Metode K-Modes dapat digunakan dengan mudah dengan library yang ada pada bahasa Python.
2. Sebarang mahasiswa Universitas Stikubank terbagi menjadi 2 klaster berdasarkan program studi, pekerjaan dan pendidikan orang tua. Dari hasil penelitian, dominasi klaster 1 cukup signifikan. Dapat disimpulkan bahwa sebagian besar gambaran profile mahasiswa adalah klaster 1.
3. Hasil pengolahan K-Modes menunjukkan bahwa klaster pertama menunjukkan jumlah terbesar dari keseluruhan data. Pada klaster pertama terdapat aspek-aspek yang cukup significant yaitu, pekerjaan ayah didominasi oleh pegawai swasta, ibu sebagian besar tidak bekerja, dan pendidikan ayah dan ibu terbanyak adalah tamatan SMA..

Berdasarkan paparan dan hasil penelitian yang telah dijelaskan maka diperlukan beberapa masukan yang dapat menjadi bahan untuk meningkatkan dan/atau menyempurnakan penelitian yang sudah dilakukan, antara lainpenentuan titik awal centroid dapat dikembangkan dengan algoritma yang lain.

DAFTAR PUSTAKA

[1] Tanti Yulianita, Deden Istiawan. Implementasi Algoritma K-modes untuk Penentuan Prioritas Rehabilitasi Daerah Aliran Sungai Berdasarkan Parameter Lahan Kritis. The 6th University Research Colloquium 2017 Universitas Muhammadiyah Magelang.

[2] Fatma Indriani, Irwan Budiman. 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 (JTIIK) Vol. 4, No. 4, Desember 2017.

[3] Ahmad Badruttamam, Sudarno, Di Asih I Maruddani. Penerapan Analisis Klaster K-Modes Dengan Validasi Davies Bouldin Index Dalam Menentukan Karakteristik Kanal Youtube Di Indonesia (Studi Kasus: 250 Kanal YouTube Indonesia Teratas Menurut Socialblade). JURNAL GAUSSIAN, Volume 9, Nomor 3, Tahun 2020.

[4] Cao, Fuyuan., Liang, Jiye & Bai, Liang. 2009. A New Initialization Method for Categorical data Clustering. An International Journal : Expert System with Application. Elsevier. Vol. 36, DOI:10.1016/j.eswa.2009.01.060.