Clustering Mahasiswa Untuk Evaluasi Kinerja Perguruan Tinggi Menggunakan Algoritma *K-Modes* dan *K-Prototypes*

1st Ahmad Habib Hasan Zein
Dept. mathematics
Gadjah Mada University
Indonesia, Yogyakarta
ahmad.habib.hasan@mail.ugm.ac.id

2nd Farla Pricilla Fatima
Dept. mathematics
Gadjah Mada University
Indonesia, Yogyakarta
farlapricilla@mail.ugm.ac.id

3rd Muhammad Zaki Nurkholis
Dept. mathematics
Gadjah Mada University
Indonesia, Yogyakarta
zaki.n@mail.ugm.ac.id

4th Petra Abdi Paskalisa
Dept. mathematics
Gadjah Mada University
Indonesia, Yogyakarta
petra.abdi.paskalisa@mail.ugm.ac.id

***Abstract*—Di era globalisasi ini, sumber daya manusia menjadi komponen penting guna mendukung persaingan global. Perguruan tinggi sebagai salah satu lembaga pendidikan wajib turut serta dalam upaya peningkatan kualitas sumber daya manusia (SDM) melalui pelaksanaan pendidikan bermutu. Evaluasi kinerja merupakan salah satu bentuk upaya perguruan tinggi dalam meningkatkan kualitas pendidikannya. Dalam menjalankan evaluasi kinerja tersebut, terdapat beberapa aspek yang mempengaruhi, salah satunya adalah mahasiswa. Karakteristik kemampuan dan prestasi yang dimiliki mahasiswa menjadi tolak ukur kualitas suatu perguruan tinggi. Pada penelitian ini, dilakukan analisis *cluster* menggunakan algoritma *K-Modes* dan *K-Prototypes* untuk mengelompokan mahasiswa menjadi tiga kategori, yaitu unggul, normal, dan kurang unggul. Diperoleh *cost* dengan menggunakan algoritma *K-Modes* adalah sebesar 35616.0 dan *K-Prototypes* sebesar 21799.8. dari ketiga cluster yang terbentuk hasil yang diperoleh adalah perguruan tinggi tersebut memiliki frekuensi anggota *cluster* yang relatif seimbang untuk ketiga kategori tersebut. Hal itu menunjukkan bahwa perguruan tinggi tersebut sudah cukup memenuhi pendidikan bermutu. Peningkatan mutu perguruan tinggi tersebut masih harus dilakukan dengan melukan evaluasi dan perbaikan dari segi kegiatan kemahasiswaan.**

Keywords: clustering, mahasiswa, perguruan tinggi, k-modes, k-prototypes

I. PENDAHULUAN

Sumber daya manusia (SDM) merupakan komponen atau individu produktif yang berfungsi menggerakan suatu organisasi. Sumber daya manusia menjadi hal yang krusial di era globalisasi ini, dimana untuk dapat beradaptasi dan berkembang di era globalisasi ini, diperlukan upaya peningkatan sumber daya manusia. Salah satu indikator utama kualitas sumber daya manusia adalah pendidikan. Pendidikan diyakini dapat menjadi katalisator yang memiliki peran penting dalam peningkatan kualitas sumber daya manusia. Strategi untuk meningkatkan kualitas sumber daya manusia dapat dilakukan dengan meningkatkan mutu pendidikan. Makna dari kualitas mutu pendidikan tersebut mengacu pada proses dan hasil pendidikan. Pendidikan diharapkan mampu menghasilkan lulusan yang kompeten dan memiliki kemampuan yang relevan dengan tuntutan dunia kerja.

Pendidikan tinggi merupakan jenjang pendidikan formal setelah pendidikan menengah. Berbeda dengan pendidikan menengah, pendidikan tinggi tidak hanya menyelenggarakan kegiatan pembelajaran, tetapi juga penelitian dan pengabdian kepada masyarakat. Pendidikan tinggi diharapkan mampu menghasilkan lulusan yang kompeten. Perguruan tinggi sebagai suatu lembaga pendidikan tinggi diharuskan untuk meningkatkan kualitas yang dapat diwujudkan melalui pelaksanaan pendidikan bermutu. Salah satu upaya bagi perguruan tinggi untuk meningkatkan mutu pendidikan adalah melalui evaluasi kinerja. Evaluasi kinerja ini dapat dilakukan dengan mempertimbangkan beberapa aspek, salah satunya adalah mahasiswa. Peningkatan kemampuan dan prestasi mahasiswa menjadi salah satu tolak ukur kualitas suatu perguruan tinggi.

Berdasarkan uraian di atas, pada penelitian ini akan dilakukan *clustering* dari beberapa indikator yang tersedia untuk mengelompokkan mahasiswa ke dalam kelompok unggul, normal, dan kurang unggul. Indikator yang dipertimbangkan adalah status kerja mahasiswa, biaya kuliah, unit kegiatan mahasiswa (UKM), organisasi kampus, lama kuliah, dan fakultas.

Clustering adalah suatu proses untuk mengelompokkan data ke dalam beberapa *cluster* atau kelompok sehingga data dalam *cluster* memiliki tingkat kemiripan yang maksimum dan data antar *cluster* memiliki kemiripan yang minimum. Algoritma *clustering* yang digunakan adalah *K-Prototype* dan *K-Modes* karena data yang digunakan bersifat numerik dan kategorik. Hasil dari *clustering* dengan kedua algoritma tersebut akan digunakan sebagai bahan evaluasi kinerja suatu perguruan tinggi.

II. PEMBAHASAN

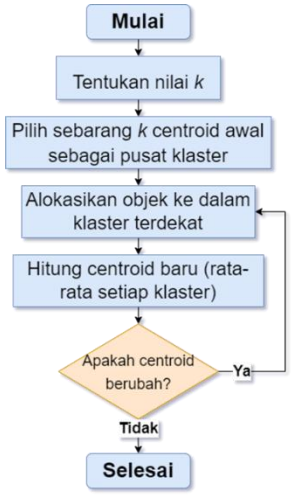
Landasan Teori

Clustering merupakan salah satu teknik data *mining* dengan tujuan memisahkan atau memecah data ke dalam beberapa kelompok berdasarkan karakteristik tertentu. *Clustering* membagi beberapa data tersebut berdasarkan tingkat kemiripan, yaitu data dalam satu *cluster* akan memiliki karakteristik yang sama atau kemiripan yang tinggi, sedangkan dengan *cluster* lain akan memiliki karakteristik yang berbeda atau kemiripan yang rendah.

Tujuan dari analisis *cluster* adalah meminimalkan variasi data dalam suatu *cluster* serta memaksimumkan variasi data antar *cluster*-nya. Dalam data *mining clustering* bukan termasuk dalam analisis klasifikasi, karena data yang terkandung dalam analisis ini tidak mempunyai label yang merupakan variabel respon. Analisis *cluster* banyak digunakan untuk mencari pola tersembunyi dalam data, kemudian hasilnya dapat dimanfaatkan sebagai informasi baru dalam analisis yang lebih lanjut.

Salah satu metode analisis *clustering* yang umum dikenal dan digunakan untuk mengelompokkan data yang heterogen adalah Algoritma *K-Means*. Algoritma *K-Means* adalah metode *clustering* non-hirarki yang mengelompokkan data ke dalam beberapa jumlah *cluster*.

Algoritma *K-Means clustering* menurut Lloyd (1982) adalah sebagai berikut:



Gambar 1. Diagram Alur Algoritma *K-Means*

1. Menentukan jumlah *k cluster* yang akan dibentuk.
2. Pilih sebanyak *k centroid* secara acak. Pemilihan *centroid* pada awal inisiasi dapat dilakukan dengan berbagai cara. Salah satunya adalah dengan memilih secara acak *k* observasi dari data dan menjadikannya sebagai *centroid* awal.
3. Kelompokkan data ke dalam *k* klaster dimana masing-masing klaster berisi observasi yang paling dekat dengan *centroid* klaster tersebut (atau dengan kata lain, alokasikan observasi ke dalam centroid terdekat). Perhitungan kedekatan antara setiap observasi dengan *centroid* klaster dapat dilakukan dengan menghitung jarak *Euclidean* antara keduanya. Terdapat metode perhitungan jarak lainnya yang dapat digunakan seperti jarak Manhattan, namun hal ini secara tidak langsung tidak akan selalu meminimumkan variansi dalam kelompok.
4. Hitung nilai *centroid* yang baru, yakni dengan mencari rata-rata (*mean*) dari data dalam setiap klaster. Perhitungan rata-rata dilakukan seperti pada umumnya.
5. Lakukan perulangan langkah (2) dan (3) hingga didapatkan hasil yang konvergen. Konvergensi dapat ditandai dengan tidak berubahnya nilai *centroid*, tidak berubahnya data di setiap klaster, maupun dengan ukuran lainnya seperti inertia. Algoritma *K-Means* klasik akan berhenti ketika diperoleh bahwa data dalam setiap *centroid* tidak berubah dalam dua iterasi berturut-turut. Perlu diperhatikan bahwa algoritma ini tidak menjamin terjadinya konvergensi. Salah satu cara mengatasinya adalah dengan membatasi banyaknya perulangan langkah (2) dan (3), atau dengan memberikan batas atas iterasi.

Algoritma *K-Means Clustering* menggunakan jarak *Euclidean* sebagai ukuran kedekatan antara objek dengan klasternya. Madhulatha (2012) menjelaskan jarak euclidean dapat dihitung dengan mencari kuadrat jarak dari masing-masing variabel, menjumlahkannya, dan mengambil akar kuadrat dari jumlahannya. Secara matematis dinyatakan sebagai berikut.

$$d = \sqrt{\sum_{j=1}^{n} (x_j - y_j)^2}$$

Dimana x $(x_1, x_2, \dots)$ dan y $(y_1, y_2, \dots)$ adalah dua objek yang diukur jaraknya.

Metode *K-Means* sendiri memiliki beberapa kelemahan, karena pada perhitungannya memanfaatkan rata-rata sebagai perhitungan *centroid* baru secara tidak langsung algoritma *K-Means* tidak cukup baik digunakan terhadap data pencilan. Lalu metode *K-means* sendiri hanya efektif digunakan untuk data yang bertipe numerik.

Untuk mengatasi permasalahan data bertipe non-numerik, Huang (1997) mengusulkan metode *clustering* yang dinamakan dengan *K-Modes* untuk data bertipe kategorik dan *K-Prototype* untuk data bertipe campuran numerik dan kategorik.

Algoritma *K-Modes* pertama kali dikenalkan oleh Zhexue Huang sebagai pengembangan lebih luas dari algoritma *K-Means*. *K-Modes* ditujukan untuk kebutuhan *clustering data sets* dengan data bertipe kategorik, dimana perhitungan jarak dalam algoritma *K-Means* (ataupun algoritma berbasis jarak lainnya) tidak tepat untuk digunakan karena dalam meminimalkan fungsi *cost*, nilai rata-rata *cluster* harus diubah secara iteratif. Proses transformasi nilai kategorik ke dalam numerik tidak selalu memberikan hasil yang informatif mengingat bahwa nilai kategorik belum tentu dapat diurutkan.

Huang juga memperkenalkan algoritma yang serupa bernama *K-Prototypes*. Tidak seperti *K-Modes*, algoritma *K-Prototypes* menggabungkan proses *K-Means* dan *K-Modes* untuk *clustering* data *sets* dengan tipe numerik dan kategorik. Keduanya memanfaatkan *dissimilarity measure* (ukuran ketakmiripan) berupa kuadrat jarak *Euclidean* dua objek untuk nilai numerik dan banyaknya kategori yang tak sama diantara dua objek untuk nilai kategorik. Pemilihan *centroid* setiap *cluster* ditetapkan melalui nilai modus (frekuensi tertinggi) dengan tujuan yang sama seperti *K-Means*, yakni meminimumkan fungsi *cost*.

Misalkan s^r adalah ukuran ketakmiripan variabel numerik dan s^c variabel kategorik. Pada *K-Prototypes*, nilai ukuran *dissimilarity* variabel numerik dan kategorik adalah sebuah kombinasi linear dengan bobot sebesar γ, dinyatakan sebagai $s^r + \gamma s^c$. Pemberian bobot dilakukan untuk menyeimbangkan bagian numerik dan kategorik agar model tidak berbias kepada salah satu dari keduanya.

Beberapa pendekatan lain telah dilakukan, seperti mengubah setiap atribut kategorik ke dalam bentuk biner (0 dan 1) dan memperlakukannya layaknya variabel numerik dalam *K-Means* (Ralambondrainy, 1995). Perlakuan ini diharapkan dapat memberikan *cluster* yang lebih baik tanpa meninggalkan efisiensi dari algoritma *K-Means*. Akan tetapi, metode ini sulit diterapkan pada *datasets* dengan ukuran besar dimana banyaknya variabel biner yang dibentuk adalah sebanyak kategori yang ada. Selain itu, semua nilai *centroid* yang berada diantara 0 dan 1 tidak bermakna karena hanya nilai 0 dan 1 saja yang didefinisikan. Di sisi lain, pengembangan terhadap *K-Modes* dan *K-Prototypes* telah banyak dilakukan, sebagai contoh Pham D.T. (2011) menciptakan algoritma bernama RANKPR (*Random Search with K-Prototypes Algorithm*) yang mengimplementasikan *random search* untuk mencari solusi ekstrema yang lebih baik secara terus menerus.

Algoritma *K-Modes* dan *K-Prototypes* memanfaatkan *dissimilarity measure* (ukuran ketakmiripan) sederhana yang didefinisikan sebagai berikut.

Misal X, Y dua objek dengan *m* atribut kategorik. Ukuran ketakmiripan antara X dan Y didefinisikan sebagai banyaknya ketidakcocokan dari atribut kategorik dari kedua objek. Semakin kecil banyaknya ketidakcocokan ini, maka semakin mirip pula kedua objek tersebut. Secara matematis,

$$d_1(X,Y) = \sum_{j=1}^{m} \delta(x_j\ ,\ y_j)$$

dimana

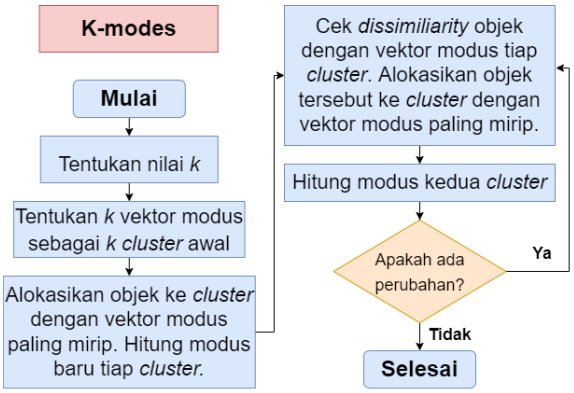
$$\delta(x_j, y_j) = \begin{cases} 0 & (x_j = y_j) \\ 1 & (x_j \neq y_j) \end{cases}$$

Selanjutnya, *cost function* yang digunakan apabila memerhatikan ukuran ketakmiripan diatas (Huang 1998) adalah

$$P(W,\ Q)\ =\ \sum_{l=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{m} w_{i,l}\ \delta(x_{i,j},\ q_{l,j})$$

dimana $w_{i,l} \in W$ adalah matriks partisi $n \times k$, $Q = \{Q_1, Q_2\ ,\ \dots\ , Q_k\}$ adalah himpunan objek dalam domain yang sama. Tujuan dari *K-Modes* adalah untuk meminimalkan fungsi *cost* tersebut sebaik mungkin.

Dalam meminimalkan fungsi *cost*, dilakukan perubahan terhadap algoritma k-means, seperti penggunaan ukuran ketakmiripan sederhana dan penggunaan modus data daripada rata-rata untuk perhitungan dan pemilihan *centroid.* Huang mendefinisikan algoritma dengan tahapan-tahapan berikut untuk meningkatkan efisiensi komputasi :



Gambar 2. Diagram Alir Algoritma *K-Modes*

1. Menentukan *k* vektor modus awal, masing-masing untuk sebuah *cluster*

$$c_1^{(t)},\ c_2^{(t)},\ \dots,\ c_k^{(t)}$$

2. Kelompokkan setiap objek (berupa vektor data) dari X ke dalam *cluster* dengan vektor modus yang paling mirip sehingga didapatkan partisi berbentuk

$$X = X_1^{(t)} \cup X_2^{(t)} \cup \dots \cup X_k^{(t)}$$

3. Hitung nilai modus yang baru pada setiap *cluster* setelah dilakukan pengelompokan objek

$$c_1^{(t+1)},\ c_2^{(t+1)},\ \dots,\ c_k^{(t+1)}$$

4. Setelah seluruh objek dikelompokkan ke dalam satu *cluster*, lakukan pengecekan *dissimilarity* objek dengan vektor modus saat ini. Apabila ditemukan objek yang lebih mirip dengan modus *cluster* lain, pindahkan objek tersebut ke *cluster* tersebut dan hitung ulang modus kedua *cluster*.

$$X = X_1^{(t+1)} \cup X_2^{(t+1)} \cup \dots \cup X_k^{(t+1)}$$

5. Kembali ke langkah ke-4 hingga tidak ada perubahan

$$X_i^{(t+1)} = X_i^{(t)},\ i\ = 1..k$$

Sama halnya dengan *K-Means*, algoritma *K-Modes* tidak menjamin diperolehnya solusi optimal global, sering kali hanya diperoleh solusi optimal lokal. Pemilihan nilai modus awal dan urutan objek dalam data sets akan berpengaruh terhadap hasil akhir *clustering*. Terdapat beberapa variasi dari

algoritma ini, seperti yang dijelaskan Funderlic, R.E. et al. (2004), yaitu kriteria untuk pemberhentian iterasi dan *tie-breaking* untuk frekuensi modus yang sama. Variasi-variasi dari *K-Modes* ini ditujukan untuk meningkatkan efisiensi iterasi ataupun untuk mempercepat diperolehnya konvergensi.

Pada metode *K-Prototype*, perbedaan mendasar terdapat pada perhitungan *dissimilarity*, yang didefinisikan sebagai kombinasi linear antara komponen numerik (jarak *Euclidean*) dan kategorik *(mismatched*).

$$d_2(X, Y) = \sum_{j=1}^{p} (x_j - y_j)^2 + \gamma \sum_{j=p+1}^{m} \delta(x_j, y_j)$$

Nilai γ digunakan sebagai bobot untuk menghindari bias terhadap salah satu komponen. Selanjutnya, *cost function* yang digunakan dalam *K-Prototypes* adalah sebagai berikut.

$$P(W, Q) = \sum_{l=1}^{k} \left(\sum_{i=1}^{n} w_{i,l} \sum_{j=1}^{p} (x_{i,j} - q_{l,j})^2 + \gamma \sum_{i=1}^{n} w_{i,l} \sum_{j=p+1}^{m} \delta(x_{i,j}, q_{l,j}) \right)$$

Misal

$$P_l^r = \sum_{i=1}^{n} w_{i,l} \sum_{j=1}^{p} (x_{(i,j)} - q_{(l,j)})^2$$

$$P_l^c = \gamma \sum_{i=1}^{n} w_{i,l} \sum_{j=p+1}^{m} \delta(x_{(i,j)} - q_{(l,j)})$$

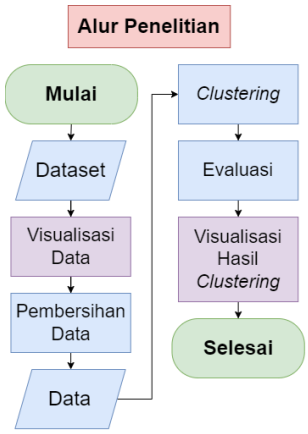
Cost function di atas selanjutnya dapat ditulis dalam bentuk

$$P(W, Q) = \sum_{l=1}^{k} (P_l^r + P_l^c)$$

Perhatikan P_l^r dan P_l^c tidak bernilai negatif, maka meminimumkan *cost function* ekivalen dengan meminimumkan P_l^c dan P_l^r untuk $1 \leq l \leq k$. Proses iterasi dalam algoritma *K-Prototypes* sama seperti *K-Modes* untuk mencari solusi optimal sebaik mungkin. Oleh karenanya, *K-Prototypes* juga memiliki kelemahan *K-Modes*, seperti tidak dijaminnya mencapai konvergensi global.

Metode Penelitian

Penelitian ini akan dibangun secara sistematik agar dapat digunakan sebagai pedoman dengan tujuan mendapatkan hasil yang diinginkan dan tidak menyimpang dari yang telah ditetapkan. Adapun diagram alur proses dalam melakukan penelitian ini adalah sebagai berikut :



Gambar 3. Diagram Alur Penelitian

A. Dataset

Pada penelitian ini digunakan *dataset* mahasiswa angkatan 2007-2009 di suatu perguruan tinggi yang sudah dinyatakan lulus. Dataset tersebut terdiri dari 11499 observasi mahasiswa dan 11 variabel dengan 9 variabel kategorik dan 2 variabel numerik. Variabel kategorik pada *dataset* tersebut, yaitu *Nama*, *Gender*, *Tinggal_Dengan*, *Status_Kerja*, *Biaya*, *Alamat*, *UKM*, *Organisasi_Kampus*, dan *Fakultas*. Sementara itu variabel numerik pada *dataset* tersebut adalah *Tgl_Daftar_Kuliah* dan *Lama_Kuliah*. Berikut merupakan rincian dari variabel-variabel berikut :

Tabel 1. Deskripsi Variabel

Nama Kolom	Deskripsi
Nama	Nama Siswa/Mahasiswa
Gender	Jenis Kelamin
Tinggal_Dengan	Tempat Siswa/Mahasiswa tinggal
Status_Kerja	Apakah siswa/mahasiswa bekerja atau tidak
Biaya	Biaya kuliah
Tgl_Daftar_Kuliah	Tahun daftar kuliah
Alamat	Alamat siswa/mahasiswa
UKM	Jenis UKM yang diikuti ketika kuliah
Organisasi_Kampus	Apakah mahasiswa mengikuti organisasi kampus
Lama_Kuliah	Lama mahasiswa menyelesaikan kuliah (dalam tahun)
Fakultas	Fakultas asal mahasiswa

B. Visualisasi

Visualisasi data merupakan proses penyajian data ke dalam bentuk grafik yang berisi informasi agar dapat lebih mudah dimengerti. Proses ini memungkinkan peneliti mendapatkan pengetahuan lebih banyak mengenai data yang digunakan. Pada penelitian ini dilakukan visualisasi untuk melihat bentuk dan pola dari

data sampel yang digunakan oleh peneliti, serta menelaah metode analisis yang dapat digunakan.

C. Pra-pemrosesan Data

Pra-pemrosesan data merupakan proses yang dilakukan sebelum pemrosesan data. Proses ini dilakukan untuk meningkatkan kualitas perhitungan dalam melakukan *clustering*. Tahapan-tahapan dalam pra-pemrorsesan data adalah sebagai berikut :

1. Seleksi Atribut atau Variabel

   Penelitian ini menggunakan 6 atribut, di antaranya Status_Kerja, Biaya, UKM, Organisasi_Kampus, Lama_Kuliah, dan Fakultas. Atribut-atribut yang digunakan tersebut disesuaikan dengan tujuan penelitian ini yaitu mengelompokan mahasiswa menjadi tiga kategori yaitu unggul, normal, dan kurang unggul. Atribut Tgl_Daftar_Kuliah tidak digunakan karena bersifat *redundant* dengan Lama_Kuliah. Atribut Nama tidak digunakan karena memiliki nilai berbeda untuk setiap observasi sehingga tidak mengandung informasi yang berharga dalam analisis ini. Atribut lainnya yaitu Gender, Tinggal_Dengan, dan Alamat tidak digunakan karena diasumsikan tidak mempengaruhi kualitas mahasiswa.
2. *Data Cleaning*

   Dataset yang digunakan oleh peneliti memuat beberapa data hilang. Untuk itu peneliti melakukan imputasi pada data yang digunakan yaitu dengan menggunakan strategi *constant* (merubah nilai hilang menjadi kategori baru). Data yang hilang tersebut diasumsikan oleh peneliti merupakan kategori lain yang tidak terdapat pada proses pengisian kuisioner atau survei, dan bukan merupakan akibat dari kesalahan teknis.
3. Transformasi Data

   Dari *dataset* yang digunakan, sebagian besar variabel berbentuk kategorik. Pada tahapan ini akan dilakukan perubahan data kategorik yang awalnya tidak bisa diolah secara matematis menjadi data yang bisa diolah. Proses transformasi ini dilakukan tanpa merubah tipe data.

D. *Clustering* Algoritma *K-Modes* dan *K-Prototypes*

Proses *clustering* dilakukan dengan menggunakan pustaka sumber terbuka (*library open source*) yang tersedia dalam bahasa pemrograman Python, dimana pustaka yang digunakan (kmodes) menggunakan algoritma yang sejalan dengan Huang, dengan beberapa modifikasi untuk meningkatkan efisiensi setiap iterasi. Implementasi algoritma tersebut diterapkan ke dalam 11499 observasi data sampel. Dalam penelitian ini, diasumsikan nilai *k* (banyaknya *cluster*) sebesar 3.

E. Evaluasi

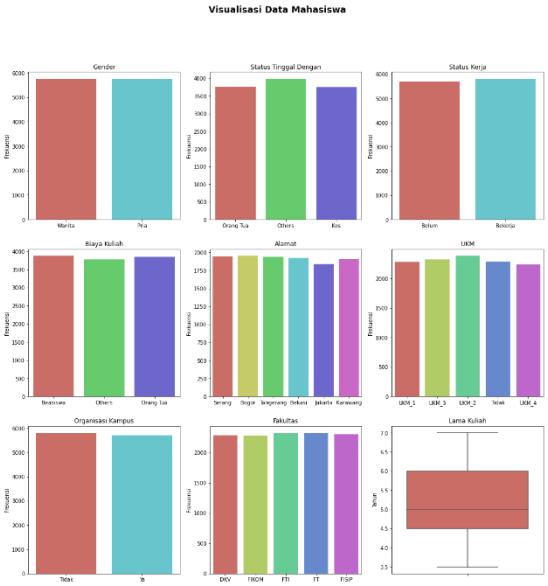
Proses selanjutnya setelah hasil *cluster* didapatkan, kedua algoritma dibandingkan secara objektif, kemudian memilih salah satu yang paling optimal dengan menggunakan nilai *cost* terendah.

F. Visualisasi Hasil *Cluster*

Pada penelitian ini akan dilakukan proses visualisasi hasil *cluster*. Proses ini memudahkan peneliti untuk melihat persebaran kategori pada masing-masing cluster yang diperoleh.

Hasil dan Pembahasan

Visualisasi



Gambar 4. Visualisasi Dataset Mahasiswa Angkatan 2007 -2009

Persebaran kategori pada dataset mahasiswa angkatan 2007-2009 di suatu perguruan tinggi yang sudah dinyatakan lulus dapat dilihat pada Gambar 3. Dari 11499 mahasiswa ditemukan sebanyak 5750 adalah wanita dan 5749 adalah pria. Pada kategori-kategori lain seperti *Status_Tinggal_Dengan*, *Status_Kerja*, *Biaya_Kuliah*, *Alamat*, *UKM*, *Organisasi_Kampus*, dan *Fakultas* juga memiliki persebaran yang relatif seimbang. Rata-rata lama kuliah mahasiswa adalah 5.2 tahun, dengan nilai terlama 7 tahun dan tercepat 3.5 tahun.

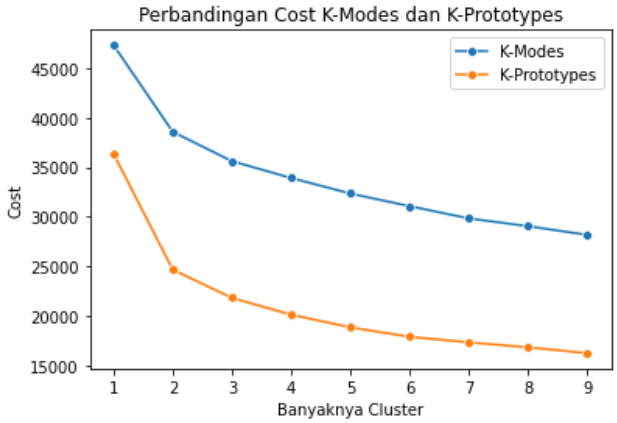
Persebaran data yang relatif uniform di atas menunjukkan bahwa informasi dari data akan sulit diperoleh hanya melalui inspeksi visual saja. Pola tersembunyi dalam data diharapkan dapat ditemukan melalui analisis *cluster*, sehingga diperoleh hasil akhir berupa *cluster* yang mampu memberikan informasi berharga untuk menjelaskan karakteristik mahasiswa yang terlibat.

Clustering Algoritma *K-Modes* dan *K-Prototypes*

Tabel 2. Nilai *Cost K-Modes* dan *K-Means*

Cluster	**Cost (*K-Modes*)**	**Cost (*K-Prototype*)**
1	47334.0	36360.9
2	38582.0	24647.3
3	35616.0	21799.8
4	33920.0	20106.6

5	32347.0	18821.9
6	31073.0	17873.0
7	29833.0	17314.5
8	29052.0	16810.5
9	28169.0	16224.8



Gambar 5. Perbandingan Cost *K-Modes* dan *K-Prototypes*

Tabel 3. *Centroid Cluster K-Modes* dan *K-Prototypes*

Metode	*Cluster*	Status Kerja	Biaya	UKM	Organisasi Kampus	Lama Kuliah	Fakultas
K-Modes	First	Belum	Others	UKM_1	Ya	5.5	DKV
	Second	Belum	Beasiswa	UKM_4	Ya	4.5	FIKOM
	Third	Bekerja	Orang Tua	Tidak	Tidak	4.5	FT

Metode	*Cluster*	Status Kerja	Biaya	UKM	Organisasi Kampus	Lama Kuliah	Fakultas
K-Prototypes	First	Belum	Orang Tua	UKM_2	Ya	5.2	FTI
	Second	Bekerja	Beasiswa	UKM_4	Tidak	4.0	FISIP
	Third	Bekerja	Orang Tua	Tidak	Tidak	6.5	FIKOM

Semakin banyak jumlah *cluster* umumnya akan menurunkan nilai *cost*-nya. Hal ini dikarenakan 'jarak' (atau dalam kasus ini adalah *similarity*) antara objek dengan centroid terdekat semakin kecil seiring bertambahnya jumlah *cluster*. Penambahan jumlah *cluster* tidak selalu baik karena hasil *cluster* tidak representatif terhadap kondisi data yang ada.

Nilai *cost* pada model *K-Modes* konsisten lebih tinggi dibandingkan *K-Prototype* untuk nilai *k* yang sama pada rentang 1 hingga 9. Hal ini boleh jadi dikarenakan dalam algoritma *K-Modes*, seluruh variabel diubah menjadi kategorik karena perhitungannya murni menggunakan modus data kategorik, dimana telah diketahui bahwa variabel lama kuliah tidaklah kategorik. Selain itu, adanya pembobotan pada fungsi *cost* algoritma *K-Prototype* berpengaruh positif pada hasil *clustering* ini, mengingat bahwa kedua metode memiliki tahapan yang sama.

Nilai *cost* mulai melandai setelah $k = 2$. Pemilihan nilai *k* berdasarkan intuisi metode *elbow* adalah sedemikian hingga penambahan nilai *k* tidak menurunkan *cost* secara signifikan (atau sedemikian hingga terjadi *diminishing return* untuk *k* yang lebih besar). Dalam kasus ini, nilai $k = 2$ dan $k = 3$ tepat untuk dipilih karena nilai *cost* untuk *k* yang lebih besar tidak menguntungkan. Hasil ini cocok dengan asumsi awal penelitian yang mengharapkan banyaknya *cluster* yaitu $k = 3$.

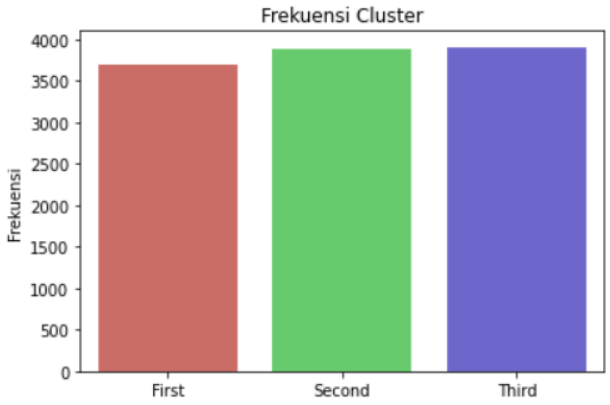
K-Prototype mampu memberikan *cluster* yang, secara objektif, lebih baik daripada *K-Modes*. Namun perlu diperhatikan bahwa algoritma *K-Prototype* jauh lebih lambat dibandingkan *K-Modes*. Seiring bertambahnya ukuran data, maka efisiensi komputasi ini perlu diperhatikan. Boleh jadi dalam kasus lainnya dengan skala yang lebih besar, efisiensi komputasi pada *K-Modes* lebih menguntungkan daripada nilai *cost* yang lebih rendah pada *K-Prototype*.

Perbedaan karakteristik centroid antara *K-Modes* dan *K-Prototype* menunjukkan bahwa kedua metode memiliki proses yang konvergen ke nilai yang berbeda. Sebagai contoh, terlihat bahwa *K-Modes* tidak mampu memisahkan *cluster* berdasarkan variabel Lama Kuliah. Hal ini bisa jadi disebabkan oleh algoritma *K-Modes* yang memperlakukan Lama Kuliah sebagai variabel kategorik dimana nilai modus tidak cocok untuk digunakan. Perbedaan konvergensi ini dapat menjelaskan mengapa *cost* pada *K-Modes* konsisten lebih tinggi dan dalam kasus ini, *K-Prototype* memberi hasil yang lebih baik.

Hasil analisis *cluster* ini hanya berdasar pada dua algoritma yang relatif sederhana dan efisien, yaitu *K-Modes* dan *K-Prototypes*. Diperlukan adanya analisis lebih lanjut dengan menggunakan algoritma-algoritma lainnya sebagai pembanding untuk *K-Prototype* dan *K-Modes*.

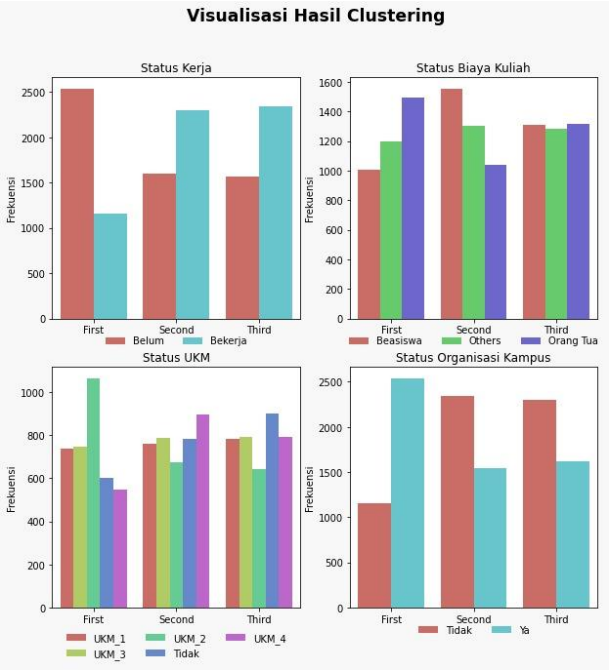
Pembahasan selanjutnya akan menggunakan hasil dari *K-Prototype* yang lebih baik daripada *K-Modes*.

Visualisasi Hasil *Cluster*



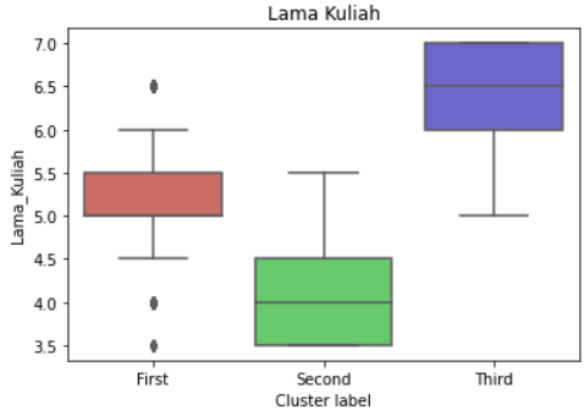
Gambar 6. Persebaran Data Tiap Cluster

Setelah dilakukan *clustering* menggunakan algoritma *K-Prototype* dengan pemilihan $k = 3$, diperoleh frekuensi masing-masing *cluster* yang relatif sama. Cluster pertama berisikan 3697 mahasiswa, cluster kedua berisi 3893 mahasiswa, dan cluster ketiga berisi 3909 mahasiswa. Hasil ini bisa jadi merupakan pertanda bahwa *clustering* tidak berhasil memisahkan data dengan baik. Untuk itu dilakukan inspeksi visual terhadap beberapa variabel yang menarik dan ditunjukkan pada bagian selanjutnya.



Gambar 7. Grafik Variabel Masing-Masing *Cluster*

Selain itu, pada ketiga *cluster* yang terbentuk dapat dilihat masing-masing persebaran kategorinya. *Cluster* First didominasi oleh mahasiswa yang belum bekerja, biaya kuliah yang ditanggung oleh orang tua, lebih memilih mengikuti UKM 2, dan mengikuti organisasi di kampus. *Cluster* Second didominasi oleh mahasiswa yang sudah bekerja, mendapatkan beasiswa untuk berkuliah, lebih memilih mengikuti UKM 4, dan tidak mengikuti organisasi kampus. *Cluster* Third didominasi oleh mahasiswa yang sudah bekerja, lebih memilih untuk tidak mengikuti UKM, tidak mengikuti organisasi kampus, serta relatif seimbang antara biaya kuliah yang berasal dari beasiswa dan orang tua.



Gambar 8. Boxplot Lama Kuliah Tiap *Cluster*

Pada atribut lama kuliah mahasiswa, boxplot *cluster* Third berada di atas dua *cluster* lainnya, menunjukkan bahwa nilai median, kuartil 1, dan kuartil 2 lebih tinggi daripada *cluster* lainnya. *Cluster* Third memiliki rata-rata lama kuliah yang lebih lama dibanding *cluster* lainnya yaitu 6.5 tahun. Selanjutnya disusul oleh *cluster* First dengan rata-rata 5.2 tahun dan *cluster* Second dengan rata-rata 4 tahun.

III. KESIMPULAN

Berdasarkan hasil analisis dan pembahasan, diperoleh bahwa algoritma *K-Prototypes* lebih baik digunakan pada penelitian ini dibanding *K-Modes* karena nilai *cost* yang konsisten lebih rendah untuk setiap nilai $1 \leq k \leq 9$. Selain itu diperoleh juga kesimpulan sebagai berikut:

1. Dari hasil *clustering* diperoleh tiga *cluster* mahasiswa dengan karakter *cluster* First adalah kelompok mahasiswa dengan mayoritas belum bekerja, mayoritas menggunakan biaya orang tua untuk biaya kuliah, mayoritas mengikuti UKM dengan jenis UKM yaitu UKM_2, mayoritas mengikuti organisasi kampus, serta memiliki rata-rata lama kuliah 5.2 tahun. Karakter *cluster Second* adalah mahasiswa dengan mayoritas sudah bekerja, mayoritas mendapat beasiswa untuk biaya kuliah, mayoritas mengikuti UKM dengan jenis UKM yaitu UKM_4, mayoritas tidak mengikuti organisasi kampus, serta memiliki rata-rata lama kuliah 4 tahun. Karakter *cluster* Third adalah mahasiswa dengan mayoritas sudah bekerja, relatif seimbang antara biaya kuliah dari orang tua, beasiswa, dan sumber lainnya, mayoritas tidak mengikuti UKM, mayoritas tidak mengikuti organisasi kampus, serta memiliki rata-rata lama kuliah 6.5 tahun.
2. Berdasarkan ketiga *cluster* mahasiswa yang terbentuk, *cluster* First merepresentasikan mahasiswa normal, *cluster* Second merepresentasikan mahasiswa unggul, *cluster* Third merepresentasikan mahasiswa kurang unggul. Hal itu didasari dengan penjelasan pada poin sebelumnya.
3. Pada perguruan tinggi tersebut, dari 11499 observasi mahasiswa, terdapat 3893 mahasiswa yang dikatakan unggul, 3697 mahasiswa normal, dan 3909 mahasiswa kurang unggul. Hal itu menunjukkan tidak ada perbedaan signifikan antara ketiga kategori mahasiswa.

Dari hasil penelitian yang telah dilakukan dapat dikatakan bahwa perguruan tinggi tersebut masih cukup baik. Namun, perguruan tinggi tersebut disarankan untuk tetap melakukan peningkatan mutu pendidikan dengan melakukan evaluasi dan perbaikan dari segi kegiatan kemahasiswaan secara berkala. Selain itu, peneliti mengharap adanya penelitian yang lebih lanjut menggunakan algoritma lain atau melakukan pengembangan terhadap metode yang dilakukan dalam penelitian ini untuk digunakan sebagai pembanding performa *clustering* data mahasiswa. Hal ini dikarenakan penelitian ini hanya terbatas pada penggunaan algoritma K-Modes dan K-Prototypes yang relatif sederhana dalam menentukan *cluster* mahasiswa.

REFERENCES

[1] D.T. Pham dan Maria Mar Suarez-Alvarez (2011), "Random search with k-prototypes algorithm for clustering mixed datasets," Proceedings of The Royal Society A Mathematical Physical and Engineering Sciences 467(2132):2387-2403.

[2] Manuel Fritz, Michael Behringer, dan Holger Schwarz (2020), "LOG-Means: efficiently estimating the number of clusters in large datasets," Proc. VLDB Endow. 13, 12, 2118–2131.

[3] R.E. Funderlic, et al. (2004), "Convergence and other aspects of the *k-modes* algorithm for clustering categorical data."

[4] T. Soni Madhulatha (2012), "An overview on clustering methods," IOSR Journal of Engineering, Apr. 2012, Vol. 2(4) pp: 719-725.

[5] Zhexue Huang (1997), Clustering large data sets with mixed numeric and categorical values.

[6] Zhexue Huang (1998), "Extensions to the k-means algorithm for clustering large data sets with categorical values," Data Mining and Knowledge Discovery 2, 283–304.