ISSN Cetak : 2302-6227
Jurnal Insand Comtech, Vol. 4, No. 2, Oktober 2019 ISSN Online : 2580-488X

SEGMENTASI MAHASISWA DENGAN '*UNSUPERVISED*' ALGORITMA GUNA MEMBANGUN STRATEGI MARKETING PENERIMAAN MAHASISWA

Arif Abriyanto[1)], Natalia Damastuti[2)]
[1,2] Fakultas Ilmu Komputer Universits Narotama
Email : [1)] arifabriyanto12@gmail.com, [2)] natalia.damastuti@narotama.ac.id

ABSTRAK

Perkembangan teknologi informasi yang demikian cepat akan menghasilkan suatu data yang besar dan heterogen yang dapat dimanfaatkan dalam berbagai bidang tidak terkecuali dalam bidang pendidikan. Proses pendaftaran peserta didik baru atau mahasiswa baru akan menghasilkan suatu data mahasiswa mulai dari profil mahasiswa sampai dengan kegiatan proses belajar. pengolahan data yang benar akan dapat membantu mendapatkan suatu informasi yang akurat. Dengan memanfaatkan suatu algoritma pembelajaran mesin dalam melakukan segmentasi data mahasiswa akan dihasilkan informasi terkait prediksi penerimaan mahasiswa baru. K-Means Clustering dilakukan untuk mengelompokkan data mahasiswa berdasarkan tiga atribut yaitu wilayah asal pendaftar, program studi dan umur mahasiswa. Hasil dari pengolahan data klaster mahasiswa yang terbentuk adalah tiga cluster, dengan cluster pertama 1112 mahasiswa, cluster kedua 825 mahasiswa dan cluster ketiga sejumlah 744 mahasiswa. Berdasarkan klasterisasi yang dihasilkan diharapkan mampu memberikan rekomendasi kepada kegiatan marketing didalam menjaring calon mahasiswa baru.

Kata Kunci : K-Means Clustering, Segmentasi, Klaster, Mahasiswa

ABSTRACT

The development of information technology that runs so fast will produce a large and heterogeneous data which can be utilized in various fields including the field of education. Process of registering new college students will produce a student data ranging from student profiles to the learning process activities. Correct data processing will be able to help obtaining accurate information. By utilizing a machine learning algorithm in segmenting student data information will be generated information related to the prediction of new student admissions. K-Means Clustering is conducted to group student data based on three attributes, which are the area of origin of the applicant, the study program and the age of the student. The results of data processing formed by the student cluster are three clusters, with the first cluster of 1112 students, the second cluster of 825 students and the third cluster of 744 students. Based on the result of clustering is expected to be able to provide recommendations to marketing activities in recruiting prospective new students.

Keywords : Clustering K-Means, Segmentation, Cluster, Student

ISSN Cetak : 2302-6227
ISSN Online : 2580-488X

PENDAHULUAN

Semakin bertambahnya tahun teknologi informasi akan lebih muda untuk berkembang dengan cepat dan hampir mencangkup disegala bidang kehidupan. Kemajuan ini dapat menghasilkan tersedianya data yang sangat besar dan banyak mulai dari bidang industri, ekonomi, dan pendidikan serta berbagai bidang kehidupan lainnya. Penerapan teknologi informasi dalam dunia pendidikan juga dapat menghasilkan banyak data yang berlimpah dari siswa yang mengikut prosesi pendidikan dan kegiatan belajar mengajar. Pada institusi pendidikan perguruan tinggi, sejumlah data dapat diperoleh berdasarkan data historis, sehingga data akan bertambah secara terus menerus misalnya data pendaftar mahasiswa baru.

Jumlah perguruan tinggi swasta di Indonesia pada tahun 2016 adalah sebanyak 3.940 (Dikti, 2016). Sebagai perguruan tinggi yang melakukan pengelolaan dana sendiri maka setiap perguruan tinggi harus aktif mempromosikan keunggulan setiap perguruan tinggi masing-masing sehingga mendapatkan mahasiswa baru. semakin banyak penerimaan mahasiswa baru pada satu perguruan tinggi maka semakin besar juga pendapatan financial perguruan tinggi tersebut sehingga kegiatan belajar mengajar dapat dilakukan secara optimal sehingga strategi dalam marketing perlu dilakukan. Salah satu strategi yang dapat dilakukan adalah dengan mengetahui potensi terbesar berdasarkan wilayah pendaftar pada suatu universitas tertentu.

Berdasarkan latar belakang tersebut maka diperlukannya suatu klasterisasi berdasarkan jenis data yang dimiliki. Pengelompokkan dilakukan dengan menggunakan metode pembelajaran mesin dengan '*unsupervised*' algoritma. Salah satu metode ini adalah menggunakan K-Means. Tujuan dari penelitian ini adalah untuk mengetahui segmentasi dari mahasiswa berdasarkan kriteria umur, program studi dan wilayah tempat tinggal mahasiswa.

Klasterisasi

Klastering sangat penting dalam suatu aplikasi data mining, sebagai contoh untuk menggali data ilmu pengetahuan, mengakses informasi dan text mining. Klasterisasi juga diterapkan dalam pencarian informasi di internet. Mesin web pencari akan mengelompokkan ratusan data yang cocok dengan kata kunci yang dimasukkan (Roni, 2015a),

Klastering merupakan bagian terpenting dari analisa data dimana dataset akan dipartisi berdasarkan kriteria-kriteria tertentu yang sejenis dalam dataset tersebut (Unnati R. Raval, Chaita Jani, 2015). Salah satu metode clustering yang sangat populer dan banyak dipelajari untuk meminimalkan kesalahan clustering untuk titik ruang Euclidean disebut K-Means clustering (Beta Estri Adiana, 2018a)

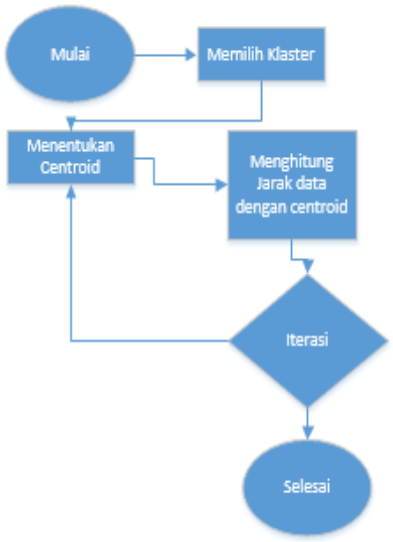
K-Means Algoritma

K-Means merupakan algoritma yang paling banyak digunakan dalam berbagai aplikasi kecil mulai dari tingkat sederhana sampai dengan tingkat menengah karena kemudahan implementasinya. Ide dasar algoritma ini adalah meminimalkan *Sum of Square Error* (SSE) antara obyek data dengan sejumlah k centroid. Algoritma ini bekerja dengan 4 langkah , yaitu himpunan data yang akan dikelompokkan dipilih sejumlah k obyek secara acak sebagai centroid awal, setiap obyek yang bukan centroid akan dimasukkan kedalam klaster terdekat berdasarkan ukuran jarak tertentu, setiap centroid diperbaiki berdasarkan rata-rata dari obyek yang ada dalam setiap klasternya dan dilakukan berulang-ulang sampai dengan semua centroid stabil atau konvergen. (Dr. Suyanto, 2017) (Youguo Li, 2012a). Secara garis besar alur dari algoritma ini dapat digambarkan seperti gambar 1.

Keluaran dari K-Means bergantung pada centroid awal yang ditentukan secara acak. Oleh karena itu K-Means harus dilakukan berulang dengan centroid awal yang berbeda untuk menghasilkan centroid

ISSN Cetak : 2302-6227
ISSN Online : 2580-488X

akhir yang bernilai bagus (Preeti Panwar, 2016a).



Gambar 1. Alur K-Means

METODOLOGI PENELITIAN

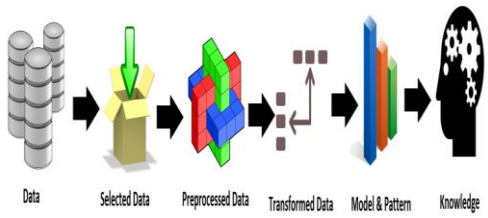
Pengumpulan Data

Pengumpulan data berupa suatu peranan tentang sifat, keadaan, kegiatan tertentu dan sejenisnya. Pengumpulan data dilakukan untuk mendapatkan suatu informasi yang dibutuhkan untuk mencari tujuan penelitian. Data penelitian didapatkan dari Bagian Adminitrsi dan Registrasi Universitas Narotama. Berdasarkan dari pengamatan pada tahun 2015 sampai 2018 data pendaftaran mahasiswa baru di Universitas Narotama mencapai 2681 mahasiswa dari semua prodi S1. Data yang berisi nama mahasiswa, nim, alamat rumah, dan alamat sekolah kemudian akan melalui pengolahan data hingga menghasilkan dataset yang siap diklasifikasi.

Adapun jenis data yang kami gunakan dalam penelitian ini adalah Data Primer yaitu data yang diperoleh langsung dari pihak kepala BAR. Dalam hal ini diperoleh langsung dari kepala BAR. Dalam hal ini diperoleh data pendaftaran mahasiswa baru dari tahun 2015 sampai 2018 dan menganalisa jumlah pendaftar.

Prapemrosesan

Pada tahap ini melakukan pengubahan data dikarenakan data tersebut masih belum sempurna seperti data tidak lengkap dan tidak konsisten. Tujuan utama dalam tahap prapemrosesan data adalah melakukan pembersihan, penambahan, pengurangan dan penyusunan data menjadi terstruktur sesuai kebutuhan pada proses *mining*.



Gambar 2. Tahapan Penambangan Data

- *Selected Data:* untuk meningkatkan akurasi dan kualitas hasil data mining, data diambil dari pendaftaran mahasiswa baru antara tahun 2015-2018 sebanyak 2681
- *Preprocessed Data:* pada tahap ini melakukan pembersihan data karena data yang diperoleh dari perusahaan umumnya masih kotor seperti data yang kurang lengkap dan atribut yang tidak relevan, data yang masih kotor menyebabkan hasil yang kurang akurat dan terjadinya eror saat pemrosesan data sehingga pada proses ini perlu dilakukan pembersihan data. Pada proses ini peneliti menghapus data yang tidak lengkap yaitu produk yang tidak bersinambung dikarenakan produk tersebut sudah tidak di produksi dan varibel serta kolom yang tidak memiliki nilai.
- *Transformation:* tahap ini melakukan perubahan data, data yang didapatkan dari perusahaan masih menggunakan format yang tidak sesuai, karena pada saat pemrosesan data hanya bisa menerima input data kategorikal dan tipe data numerik sehingga data yang

ISSN Cetak : 2302-6227
ISSN Online : 2580-488X

digunakan harus berbentuk format excel dengan ekstensi CSV.

- *Model & Patterns:* tahap ini merupakan proses pembentukan pola yaitu mengolah data menjadi kelompok dengan menggunakan metode klasterisasi. Proses ini bertujuan untuk menentukan kategori produk yang diminati dengan menggunakan algoritma *K-Means*(Yang and Wang, 2017).

Tahapan metode *K-Means* sebagai berikut dalam bentuk *pseudocode*.

Input: D (Data), K (jumlah klaster)

1) Tentukan jumlah kelompok (K)
2) Inisialisasi nilai pusat secara acak

Proccess:

Hitung setiap titik data dalam D dengan nilai pusat terdekat. Data yang telah dihitung akan menjadi kelompok K.

3) Hitung ulang posisi nilai pusat.

Repeat:

Ulangi langkah *process* sampai tidak ada lagi perubahan keanggotaan titik data.

Output:

Data yang sudah dikelompokkan berdasarkan perhitungan jarak minimum. Untuk menghitung setiap objek data dengan nilai pusat menggunakan rumus Euclidean Distance sebagai berikut:

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

Dimana :

d = jarak antara x dan y
x = nilai pusat
y = data pada atribut
j = setiap data
n = jumlah data
xj = nilai pusat ke j
yj = objek data ke j

- *Knowledge:* Pada tahap ini mengetahui dari hasil penggalian data. Hasil tersebut digunakan sebagai sumber pengetahuan untuk digunakan pada tujuan penelitian ini dan mempresentasikan pengetahuan dalam bentuk yang mudah dipahami.

Data yang berjenis nominal seperti kota asal dan program studi harus dilakukan proses inisialisasi data terlebih dahulu ke dalam bentuk angka atau numerikal. Untuk melakukan inisialisasi kota asal dapat dilakukan dengan:

a. Pada kota asal mahasiswa terlebih dahulu dilakukan pembagian wilayah-wilayah menjadi beberapa bagian wilayah.

Tabel 1. Contoh Inisialisasi wilayah

Kota Asal	Frekuensi	Inisiasi
Surabaya	1852	1
Sidoarjo	223	2
Gresik	73	3
Ponorogo	15	5
Trenggalek	8	6
Tulungagung	2	7
blitar	11	8
kediri	34	9
malang	11	10
Lumajang	3	11
probolinggo	2	16
pasuruan	10	17
luar jawa	63	34
luar negri	121	35

Tabel 1 merupakan inisiasi berdasarkan atribut wilayah. Terdapat 35 wilayah baik di wilayah Jawa Timur ataupun di luar wilayah tersebut. Sebanyak 35 kode wilayah yang dihasilkan dari proses inisiasi. Proses selanjutnya adalah inisiasi dari atribut program studi seperti yang ditunjukkan dalam tabel 2.

Tabel 2. Hasil inisiasi berdasarkan atribut program studi

Progam Studi	Frekuensi	Inisiasi

ISSN Cetak : 2302-6227
ISSN Online : 2580-488X

Akuntasi	335	1
Manajemen	614	2
Hukum	385	3
Teknik Sipil	563	4
Sistem Komputer	158	5
Sistem Informasi	327	6
Teknik Informatika	300	7

Dan Yang Terakhir adalah inisiasi umur dan diubah kedalam bentuk angka atau numerikal.

Tabel 3. Inisiasi Atribut Umur Mahasiswa

Umur	Frekuensi	Inisial
1998-0000 (21-00)	670	1
1994-1997 (22-25)	1267	2
0000-1993 (00-26)	744	3

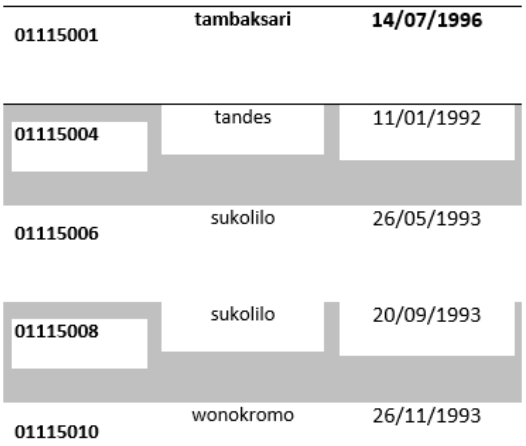
HASIL DAN PEMBAHASAN

Eksperimen dilakukan dengan menggunakan perangkat lunak *WEKA* 3.8, program tersebut digunakan untuk membantu penelitian dalam proses mengelompokkan data ke dalam klaster dengan menggunakan *tool cluster* dan salah satu metode algoritma klasterisasi K-Means. Pengujian menggunakan data tabel dengan mengubah data awal menjadi penyusunan sebagai berikut:

1) Atribut yang digunakan pada penelitian ini berjumlah 3 atribut yaitu Prodi,wilayah,umur
2) Penelitian ini menggunakan data dengan jumlah 2681.

Tahap awal yang dilakukan adalah melakukan seleksi data. Data yang diperoleh dari Biro Admisi dan Registrasi merupakan data mentah yang memiliki multi atribut, sehingga diperlukan seleksi data berdasarkan tujuan yang ingi dicapai. Atribut yang diseleksi hanya 3 atribut yaitu Nomor induk mahasiswa yang menyatakan program studi, alamat dan tanggal lahir mahasiswa. Hasil seleksi data dapat disampaikan pada Gambar 3.

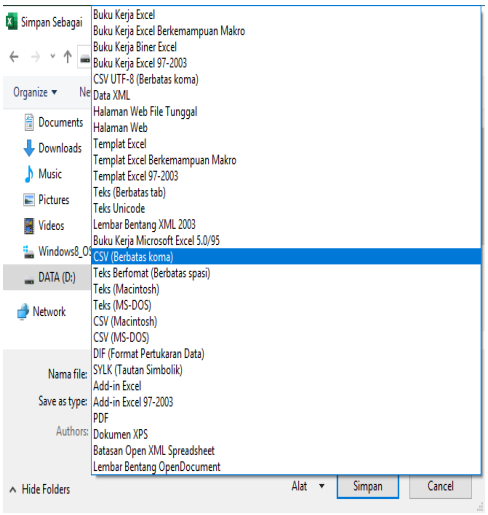
Tahap selanjutnya mengubah delimeter yang dipakai untuk membatasi atau memisahkan data dikarenakan perangkat lunak yang digunakan pada penelitian ini yaitu weka harus menggunakan delimeter koma sedangkan perangkat lunak yang di gunakan sebagai pembaca file dan visualisasi data menggunakan titik koma.



Gambar 3. Hasil Seleksi Data

Sebelum mengubah delimeter file data file harus disimpan dalam bentuk CSV dikarenakan ekstensi file tersebut umumnya digunakan perangkat lunak pengolahan data yang tidak mengandung rumus maupun format lainya. Selanjutnya mengubah delimeter data, perangkat lunak yang digunakan untuk mengubah delimeter data yaitu *notepad* ++ dikarenakan notepad ++ mudah digunakan untuk *mereplace* data, berikut proses perubahan dan hasil delimeter.

ISSN Cetak : 2302-6227
ISSN Online : 2580-488X



Gambar 4. Proses Delimeter

Hasil dari proses delimeter ini dapat disampaikan pada gambar berikut

```
jurusan;wilayah;umur
1;1;2
1;1;3
1;1;3
1;1;3
1;1;3
1;1;3
1;1;3
1;1;3
1;1;3
1;1;3
1;3;3
1;1;3
1;1;3
1;1;3
1;1;3
1;1;3
1;1;3
1;1;3
1;2;3
1;1;3
1;1;3
1;1;3
1;1;3
1;1;3
```

Gambar 5..Hasil Dilimiter

Gambar 5 dihasilkan setelah mengubah delimeter awal yang memiliki format *ekstensi* file data yang digunakan dari CSV ke ARFF dikarenakan perangkat lunak WEKA hanya mengenali atau menggunakan *ekstensi* tersebut. Untuk mengubah *ekstensi* data tersebut menggunakan perangkat lunak WEKA sendiri dengan memilih menu *tools* kemudian pilih ARFFviewer untuk membaca file data ke dalam perangkat lunak WEKA dengan memilih file data yang akan diubah kemudian menyimpan data dengan memilih ekstensi ARFF

Prosesing Data

Proses pengelompokkan data di awali dengan melakukan inisialisasi nilai tengah suatu objek data (centroid), nilai *centroid* digunakan untuk perhitungan jarak objek data ke dalam klaster, menentukan *centroid* di awali dengan menetapkan *centroid* awal setelah itu *centroid* awal digunakan untuk menentukan *centroid* akhir, untuk menemukan *centroid* awal metode K-Means secara acak yang di dapat dari nilai objek data. Hasil penentuan *centroid* awal yang akan digunakan dapat dilihat pada tabel berikut.

Tabel 4. Centroid Awal

Initial starting points (random):			
Cluster 0	3	2	2
Cluster 1	5	1	2
Cluster 2	3	`1	3

Nilai pada objek data yang akan digunakan sebagai *centorid* awal kemudian nilai tersebut digunakan untuk melakukan perhitungan jarak terdekat (iterasi), perhitungan iterasi menggunakan rumus *euclidean distance* dengan tujuan menemukan nilai *centroid* akhir yang akan digunakan sebagai nilai pusat klaster. *Centroid* akhir yang di dapatkan dari hasil perhitungan iterasi dapat dilihat sebagai berikut.

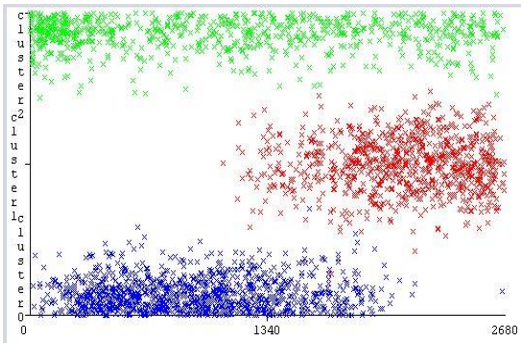
Tabel 5. Centroid Akhir

Attribute	0	1	2
Jurusan	2.5414	5.6206	3.1613
Wilayah	5.6763	5.2695	6.2445
Umur	1.8471	1.3939	3

ISSN Cetak : 2302-6227
ISSN Online : 2580-488X

Proses ini menempatkan objek data ke dalam beberapa klaster. Data akan di masukkan ke dalam klaster dengan melakukan iterasi ulang berdasarkan jarak minimum dari nilai *centroid* dengan nilai setiap objek data dan ditetapkan pusat *centroid* terdapat pada cluster 2.

Hasil dari klasterisasi yang telah dilakukan dapat disampaikan secara visual sebagaimana tampak pada gambar 6.



Gambar 6. Visualisasi Hasil

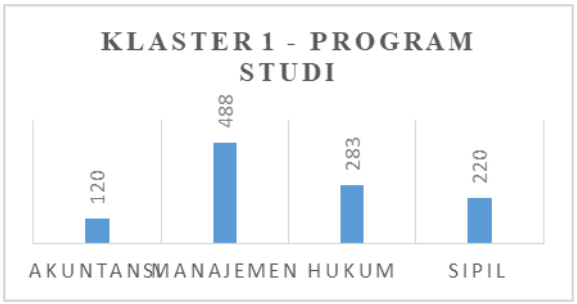
Pada gambar diatas dapat dilihat hasil penyebaran data ke dalam *cluste*r sesuai dengan anggota yang telah dihitung berdasarkan jarak terdekat dari data ke nilai pusat atau centroid. Pada gambar di atas menunjukkan bahwa *cluster* 0 berwarna biru berada di posisi bawah merupakan kumpulan objek data yang memiliki nilai persebaran yang terbatas atau rendah. selanjutnya *cluster* 1 berwarna merah berada di posisi tengah merupakan kumpulan objek yang memiliki nilai sedang karna persebaran hampir merata, sedangkan *cluster* 2 berwarna hijau posisi atas merupakan kumpulan objek data yang besar dan merata.

Hasil pengelompokan kedekatan atribut dengan jarak kedekatan titik pusat dengan data mahasiswa.

Tabel 6. Hasil Klaster

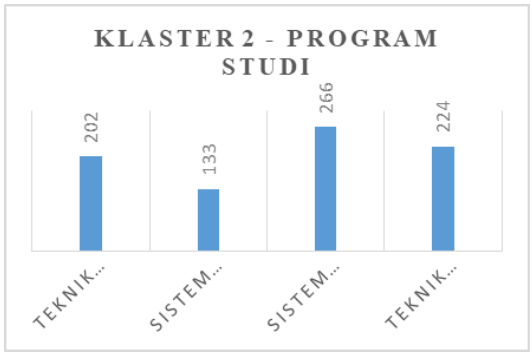
Klaster	Jumlah	Persen
0	1112	(41%)
1	825	(31%)
2	744	(28%)

Pada tabel diatas dapat disimpulkan hasil ke tiga klaster secara terperinci seperti gambar dibawah.

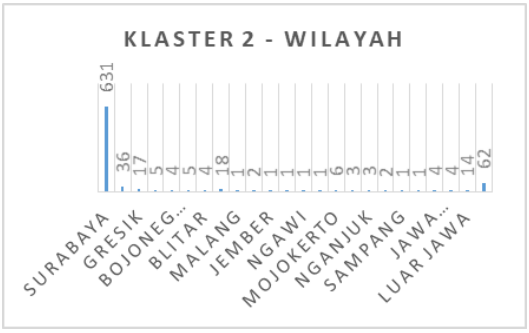


Gambar 7. Grafik Hasil Klaster 1

Pada cluster 1 ini dihasilkan 1112 (42%) mahasiswa dengan pendaftar terbanyak di kota Surabaya degan 733 mahasiswa dan yang paling sedikit di beberapa daerah yaitu Jember, Banyuwangi, Situbondo dan progam studi yang paling banyak diminati ialah Akuntansi, Manajemen, Hukum, dan Teknik Sipil dengan rata rata umur 23-25 tahun.

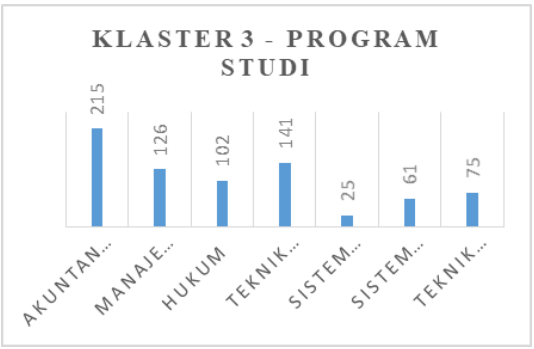


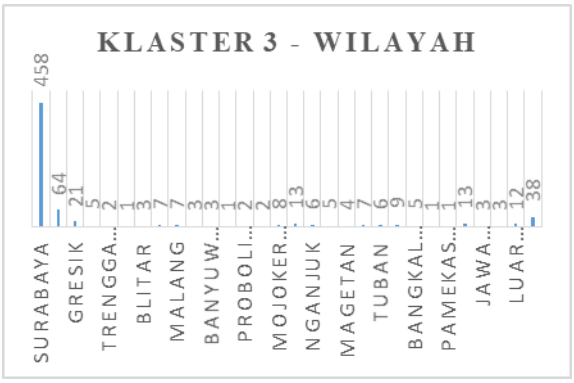
ISSN Cetak : 2302-6227
ISSN Online : 2580-488X



Gambar 8. Grafik Hasil Klaster 2

Pada kluster 2 dihasilkan 825 (31%) seperti di gambar 8 mahasiswa dengan pendaftar terbanyak di kota Surabaya dengan jumlah pendaftar terbanyak di kota surabaya dengan jumlah 631 dan yang paling sedikit di beberapa daerah antara lain Malang, Jember, dan Pasuruan dengan progam studi yang paling diminati adalah Teknik Sipil, Sistem Komputer, Sistem Informasi, dan Teknik Informatika dengan rata – rata umur pendaftar 19-25 tahun





Gambar 9. Grafik Hasil Klaster 3

Pada cluster 3 dihasilkan 744 mahasiswa (28%) dan ditetapkan juga sebagai pusat *centroid* dengan jumlah pendaftar terbanyak masih di kota Surabaya dengan 458 pendaftar dan yang paling sedikit pendaftar terdapat di daerah yang hampir sama antara lain Tulungagung, Situbondo, Sampan, Pamekasan.

Dari data hasil clustering yang telah dilakukan di atas, maka dapat ditentukan beberapa cara untuk tim marketing.

a. Promosi dengan mengirim tim marketing ke daerah daerah yang sesuai dengan program studi yang paling banya6tk diminati.

b. Pada persebaran wilayah yang paling sedikit peminat tim marketing bisa langsung datang ke sekolah sekolah yang ada pada daerah sepi peminat sehingga dapat meningkatkan penerimaan mahasiswa baru di Universitas Narotama.

KESIMPULAN

Setelah dilakukan pengelompokan data mahasiswa melalui persebaran wilayah berdasarkan potensi Progam Studi dan wilayah menggunakan K-Means clustering terbentuk tiga cluster yaitu, cluster satu dengan jumlah 1112 mahasiswa dengan rata-rata umur 23-25 tahun, cluster dua dengan jumlah 825 mahasiswa dengan rata-rata umur 19-25 tahun, dan cluster tiga dengan jumlah 744 mahasiswa dengan rata-rata umur 25 tahun keatas.

Strategi untuk meningkatkan pendaftaran bagi calon mahasiswa baru yang tepat sasaran untuk setiap wilayah berdasarkan cluster yang terbentuk adalah dengan mengirim tim marketing yang sesuai dengan program studi yang paling banyak diminati dan melakukan promosi berdasarkan potensi wilayah yang paling banyak peminat mahasiswa yang telah mendaftar dengan melakukan penyelarasan menggunakan promotion mix dan dengan melihat rata-rata pendaftar pada setiap klaster.

ISSN Cetak : 2302-6227
ISSN Online : 2580-488X

DAFTAR PUSTAKA

Beta Estri Adiana. (2018a). Analisis Segmentasi Pelanggan Menggunakan Kombinasi RFM Model dan Teknik Clustering. *JUTEI*, 23-32.

Dikti, D. J. (2016). Jakarta.

Dr. Suyanto, S. M. (2017). *Data Mining untuk Klasifikasi dan Klasterisasi Data.* Bandung: Informatika.

Irwansyah. (2016). Implementasi Data Mining untuk Menentukan Persediaan Stok Burger Menggunakan Metode K-Means Clustering. *RUBSI*.

Preeti Panwar. (2016a). Image Segmentation using K-means clustering and Thresholding. *International Research Journal of Engineering and Technology* .

Roni, A. (2015a). Penerapan Metode K-Means Untuk Clustering Mahasiswa Berdasarkan Nilai Akademik dengan Weka Interface. *https://www.researchgate.net/publication/329831347*.

Unnati R. Raval, Chaita Jani. (2015). Implementing and Improvisation of K-Means Clustering. *International Journal of Computer Science and Mobile Computing*.

Youguo Li. (2012a). A Clustering Method Based on K-Means Algorithm. *International Conference on Solid State Devices and Materials Science.* China: Elsevier.