

ANALISA SEBARAN MAHASISWA UNIVERSITAS STIKUBANK MENGUNAKAN METODE K-MEANS

Felix Andreas Sutanto ¹, Heribertus Yulianton ², Jeffri Alfa Razaq ³

^{1,2,3}Program Studi Teknik Informatika, Fakultas Teknologi Informasi, Universitas Stikubank
e-mail: ¹felix@unisbank.ac.id, ²heri@unisbank.ac.id, ³mrjf@edu.unisbank.ac.id

Abstrak

Setiap perguruan tinggi selalu berharap mendapatkan mahasiswa baru sebanyak mungkin. Berbagai usaha promosi mungkin dilakukan untuk mendapatkan mahasiswa baru. Ada yang memasang iklan di surat kabar, televisi maupun media baliho, pamflet, brosur dan sebagainya. Namun tidak semua cara bisa dilakukan secara efektif.

Penelitian ini bertujuan untuk melakukan pengelompokan terhadap data mahasiswa Universitas Stikubank dengan memanfaatkan data mining menggunakan teknik Clustering. Diharapkan dari penelitian ini dapat diketahui kelompok-kelompok wilayah promosi yang dapat dijadikan prioritas untuk promosi. Data yang digunakan adalah data mahasiswa Unisbank pada tahun 2018 dan 2019. Metode yang digunakan adalah K-Means dengan algoritma Lloyd. Klaster ditentukan dengan jumlah tertentu. Untuk mengolah data digunakan bahasa R.

Hasil dari penelitian ini adalah kelompok alumni berdasarkan fakultas dan kota asal sekolah. Data tersebut dapat digunakan untuk memetakan wilayah atau kota-kota yang akan dijadikan prioritas untuk promosi.

Kata Kunci: k-means, lloyd, clustering, r

1. PENDAHULUAN

Setiap perguruan tinggi mungkin memiliki pangsa pasar tersendiri, hal ini bisa diteliti dari data mahasiswa atau alumni yang dimiliki. Data tersebut antara lain adalah data pribadi, data rencana studi, dan data hasil studi. Jumlah data yang terakumulasi dari tahun ke tahun perlu dilakukan analisis untuk dapat membuka peluang dihasilkannya informasi yang berguna dalam pembuatan alternatif keputusan bagi manajemen perguruan tinggi.

Penelitian ini bertujuan untuk melakukan pengelompokan terhadap data mahasiswa Universitas Stikubank dengan memanfaatkan data mining menggunakan teknik Clustering. Diharapkan dari penelitian ini dapat diketahui kelompok-kelompok wilayah promosi yang dapat dijadikan prioritas untuk promosi.

Data mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan machine learning untuk mengekstrasi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar. Penelitian tentang BMI menggunakan algoritma K-Means untuk mengelompokkan mahasiswa berdasarkan nilai body mass index (bmi) & ukuran kerangka. Dari hasil penelitian terhadap 20 data sampel diperoleh 3 kelompok mahasiswa berdasarkan nilai BMI dan ukuran rangka, yaitu BMI normal dan kerangka besar, BMI obesitas sedang dan kerangka sedang, BMI obesitas berat dan kerangka kecil [1]. Klasifikasi dengan K-Means juga dilakukan untuk menunjang strategi pemasaran. Penelitian ini menggunakan teknik klasifikasi dengan K-Means untuk mengelompokkan data kategorikal hingga menghasilkan klaster yang lebih stabil [2]. Algoritma K-Means bekerja dengan membagi data ke dalam k buah cluster yang telah ditentukan. Beberapa cara penghitungan jarak yang biasa digunakan yaitu Euclidean distance dan Chebichev Distance [3].

Pada penelitian ini, masalah yang akan dibahas dalam penelitian ini meliputi memilih kelompok data yang akan dijadikan obyek analisa. Dalam hal ini adalah data mahasiswa Universitas Stikubank pada tahun 2018 dan 2019 yang diambil secara acak. Selanjutnya menerapkan algoritma K-Means untuk mendapatkan kelompok wilayah promosi.

2. METODE PENELITIAN

Dalam penelitian ini, metode yang digunakan untuk proses data mining meliputi :

- Eksplorasi : Pemrosesan awal data yang terdiri dari pembersihan data, normalisasi data, transformasi data, penanganan data yang salah, reduksi dimensi, pemilihan subset fitur dan sebagainya.
- Membangun model dan validasi : membangun model dan melakukan validasi berarti melakukan analisis berbagai model dan memilih model dengan prediksi yang terbaik. Dalam hal ini digunakan metode seperti klasifikasi, regresi, analisis cluster, deteksi anomali, analisis asosiasi dan sebagainya.
- Penerapan : penerapan berarti menerapkan model pada data yang baru untuk menghasilkan perkiraan atau prediksi masalah yang diteliti.

3. HASIL DAN PEMBAHASAN

3.1. Persiapan Data

Persiapan awal yang dibutuhkan adalah database mahasiswa. Data diambil melalui Sistem Smart Campus yang dapat diakses di www.unisbank.ac.id/sc. Data diambil melalui query terhadap tabel yang menampung data mahasiswa. Data hasil query disimpan dalam bentuk Excel.

Pembersihan data dilakukan dengan menghapus data yang tidak dibutuhkan dalam penelitian ini, hal ini disebabkan karena tabel mahasiswa terdiri dari informasi dan identitas pribadi secara keseluruhan. Field yang relevan dengan penelitian ini adalah nim, nama, asal sekolah dan kota. Setelah diamati, ternyata banyak data yang tidak lengkap kota asalnya, sehingga peneliti memutuskan agar data tersebut tidak diikutsertakan dalam proses data mining.

Kemudian data mahasiswa yang berasal dari kota yang jauh atau luar pulau juga tidak diikutsertakan, karena tujuan penelitian ini adalah menghasilkan wilayah yang direkomendasikan kepada tim promosi. Dan tim promosi tidak akan menjangkau wilayah luar jawa. Akhirnya didapatkan 372 record alumni dari prodi Sistem Informasi dan Teknik Informatika.

Proses berlanjut dengan mengkonversi kota ke jarak km dan tahun lulus menjadi lama studi. Konversi kota dilakukan dengan bantuan mesin pencari google. Untuk memberikan hasil yang lebih merata, jarak dibuat dalam bentuk skala 1:10. Contoh hasil dari persiapan data seperti pada tabel 1.

Tabel 1. Konversi Data

Nama	Prodi	Skala
Weni Yuliana	15	4,1
Karwito	12	9,5
Diaz Adi Putranto	15	9,5
Hangga Tegar Praditya	15	9,5
Karomatun Nisak	15	9,5
Novia Setyorini	12	9,5
Yunita Fatma Faidha	12	14,8
Rizky Noviananto	15	14,8
Nur`Aini Alfichusna	15	14,8
Fuad Reza Perdana	12	14,8
David Arfianto Prabowo	15	14,8
Nurul Hilal Nasrullah	12	2,5
Shodiq Purna Raharjo	15	8,6
Endar Yugo Utomo	12	8,6
Dewi Jayanti Setiasih	12	8,6
Gerardus Dodi Prasetyo	15	17,3
Ayu Anisa Putri	12	17,3

3.2. Pengolahan Data

Metode K-Means adalah salah satu metode dalam fungsi clustering atau pengelompokan. Clustering mengacu pada pengelompokkan data, observasi atau kasus berdasar kemiripan objek yang diteliti [4]. Sebuah cluster adalah suatu kumpulan data yang mirip dengan lainnya atau ketidakmiripan data pada kelompok lain. Algoritma K-Means dipilih karena mudah untuk diimplementasikan dan dijalankan, relatif cepat dan mudah beradaptasi.

Pengolahan data dilakukan dengan bantuan bahasa R. Dalam bahasa R tersedia beberapa algoritma k-means yang dapat dipilih, namun algoritma yang disebut paling cepat adalah Lloyd, oleh karena itu dalam penelitian ini dicoba untuk menggunakan algoritma tersebut. Karena data yang akan diolah berbentuk excel, maka bahasa R membutuhkan library `xlsx`. Adapun penerapan bahasa R seperti pada gambar 1.

```
library(xlsx)
alumni <- read.xlsx("alumnifti2.xlsx", sheetIndex = 1)
```

Gambar 1. Penerapan Library `xlsx`

Jumlah data yang dibaca oleh R sejumlah 372 data. Dengan perintah summary, dapat diberikan gambaran data awal seperti pada gambar 2.

```

KOTA
SEMARANG :207
KENDAL  : 31
DEMAK   : 23
PURWODADI: 16
PATI    : 15
REMBANG : 13
(Other) : 67

JARAK    SKALA
Min.    : 10.00 Min.    : 1.000
1st Qu.: 10.00 1st Qu.: 1.000
Median  : 10.00 Median  : 1.000
Mean    : 39.24 Mean    : 3.924
3rd Qu.: 65.00 3rd Qu.: 6.500
Max.    :173.00 Max.    :17.300
    
```

Gambar 2. Summary Data Awal

Dari data awal dapat dilihat bahwa mahasiswa yang lulus mulai tahun 2014 sampai 2016 berasal dari kota yang paling banyak adalah Semarang, Kendal, Demak, Purwodadi, Pati, Rembang dan kota-kota lainnya yang jumlahnya tidak lebih dari 13 data untuk masing-masing kota. Sedangkan jarak diukur dari lokasi kampus ke kota. Untuk Semarang diberikan jarak 10 km dan kota dengan jarak terjauh 173 km. Agar nilai data seimbang maka jarak km diubah dalam bentuk skala. Hal ini dimaksudkan agar pada waktu pemrosesan nilainya seimbang dengan nilai prodi.

Proses k-means dimulai dengan pemilihan variabel yang akan digunakan dalam proses. Pada penelitian ini dipilih variabel prodi dan skala jarak kota. Nilai prodi disesuaikan dengan kode yang ditetapkan dalam Smart Campuss Universitas Stikubank, yaitu Sistem Informasi memiliki kode 15 dan Teknik Informatika memiliki kode 12.

Proses k-means menggunakan algoritma Lloyd pada bahasa R menggunakan kode program seperti pada gambar 3.

```

(kmeans.result <- kmeans(mhsw2,4, nstart=1, algorithm = c("Lloyd"), trace=FALSE))
plot(mhsw2[c("LAMA", "SKALA")], col = kmeans.result$cluster)
    
```

Gambar 3. Implementasi K-Means di R

Sebelum proses data mining data yang tidak digunakan dalam proses data mining dibuang dahulu. Dengan perintah head(), contoh data yang siap diproses seperti pada gambar 4.

```

PRODI SKALA
1 15 4.1
2 12 9.5
3 12 9.5
4 12 9.5
5 15 9.5
    
```

Gambar 4. Data Yang Diproses

Kluster ditetapkan sejumlah 4. Pada iterasi pertama, diperlihatkan oleh R centroid diambil dari data ke 24, 28, 53 dan 267. Hal tersebut dapat ditunjukkan pada gambar 5.

```

k-means clustering with 4 clusters
of sizes 53, 28, 267, 24

Cluster means:
      PRODI      LAMA      SKALA
1 13.69811  8.415094  7.054717
2 13.57143  8.642857 11.178571
3 13.06367  8.640449  1.428839
4 13.50000  9.083333 16.300000
    
```

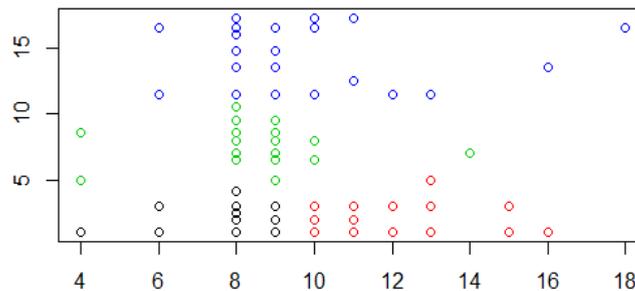
Gambar 5. R Centroid

Proses iterasi diulang hingga tiga kali dan pada perulangan keempat data cenderung tidak berubah. Contoh hasilnya seperti pada gambar 6.

1	2	3	4	
06.01.53.0002	0	0	1	0
06.01.53.0064	1	0	0	0
06.01.53.0167	1	0	0	0
08.01.53.0015	0	0	1	0
08.01.53.0025	0	0	1	0
08.01.53.0027	0	0	1	0
08.01.53.0108	0	0	0	1
08.01.53.0109	0	0	0	1
08.01.53.0119	0	0	1	0
08.01.53.0126	0	0	1	0
08.01.53.0175	0	0	1	0
08.01.53.0192	0	0	1	0
08.01.53.0250	0	0	0	1
08.01.55.0062	0	0	1	0
08.01.55.0143	0	0	1	0

Gambar 6. Hasil K-Means

Adapun plot pengelompokan dapat dilihat pada gambar 7.



Gambar 7. Plot Klustering

Hal yang dapat disimpulkan dari clustering asal mahasiswa dapat dilihat pada tabel 2.

Tabel 2. Hasil Pengelompokan Data

	Kelompok 1	Kelompok 3	Kelompok 2	Kelompok 4
Persentase	58%	16%	14%	12%
Kota	Semarang Kendal Demak	Purwodadi Pati Jepara Batang	Semarang Demak Kendal	Rembang Brebes Tegal Blora
Progdi	TI	TI dan SI	TI	TI dan SI

4. KESIMPULAN

Kesimpulan dari penelitian ini bahwa Kualitas data yang dianalisa sangat mempengaruhi hasil pengelompokan. Jika datanya kurang bervariasi, hasilnya juga kurang bagus. K-Means dapat digunakan dengan mudah untuk mengelompokkan data mahasiswa. Pada penelitian ini terdapat 4 kelompok mahasiswa. Kelompok 1 dan 2 ternyata didominasi oleh program studi TI.

5. SARAN

Saran yang bisa diberikan untuk perkembangan dan penggunaan penelitian ini bahwa Data bisa diperbanyak untuk mendapatkan kemungkinan varian yang lebih banyak. Pengolahan data dapat menggabungkan metode yang lain untuk pengelompokan data.

DAFTAR PUSTAKA

- [1] Tedy Rismawan dan Sri Kusumadewi, *Aplikasi K-Means Untuk Pengelompokan Mahasiswa Berdasarkan Nilai Body Mass Index (BMI) & Ukuran Kerangka* Seminar Nasional Aplikasi Teknologi Informasi, Yogyakarta, 2008, pp. ISSN: 1907-5022
- [2] Totok Suprawoto, *Klasifikasi Data Mahasiswa Menggunakan Metode K-Means Untuk Menunjang Pemilihan Strategi Pemasaran*, Jurnal Informatika dan Komputer (JIKO) – Vol. 1, No. 1, Februari 2016
- [3] Han, J. and Kamber, M. *Data mining: Concepts and Techniques*, Jhon Wiley & Sons, Inc, 2005.
- [4] Larose, Daniel T, *Data Mining Methods and Models*. Hoboken New Jersey: Jhon Wiley & Sons, Inc, 2006.