

ALGORITMA *K-MEANS* UNTUK CLUSTERING DALAM PENILAIAN TOKO *E-COMMERCE SHOPEE*

PROPOSAL SKRIPSI

Disusun Oleh:
Felicia Marvela Evanita
NIM: 175150200111082



PROGRAM STUDI TEKNIK INFORMATIKA
JURUSAN TEKNIK INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS BRAWIJAYA
MALANG
2020

DAFTAR ISI

DAFTAR ISI	ii
DAFTAR TABEL.....	iii
DAFTAR GAMBAR.....	iv
BAB 1 PENDAHULUAN.....	1
1.1 Latar Belakang	1
1.2 Identifikasi Masalah.....	2
1.3 Rumusan Masalah	2
1.4 Tujuan	2
1.5 Manfaat	3
1.6 Batasan Masalah	3
1.7 Sistematika Pembahasan.....	3
BAB 2 TINJAUAN PUSTAKA.....	4
2.1 Kajian Pustaka.....	4
2.3 Data Clustering	5
2.4 Algoritma K-Means.....	5
2.5 Silhouette Coefficient	6
BAB 3 METODOLOGI PENELITIAN	8
3.1 Tipe Penelitian	8
3.2 Strategi Penelitian	8
3.3 Teknik Pengumpulan Data.....	8
3.4 Data Penelitian	9
3.5 Implementasi Algoritma	10
3.6 Teknik Pengujian dan Analisis	11

DAFTAR TABEL

Tabel 2.1 Tabel Kajian Pustaka	4
--------------------------------------	---

DAFTAR GAMBAR

Gambar 3.1 Gambar Diagram Teknik Pengumpulan Data	9
Gambar 3.2 Gambar Diagram Implementasi Algoritma	10

BAB 1 PENDAHULUAN

1.1 Latar Belakang

Seiring dengan perkembangan teknologi, akses transaksi jual beli di Indonesia ikut berkembang. Transaksi jual beli yang dulunya hanya dapat dilakukan secara *offline* sekarang dapat dilakukan secara *online*. Transaksi jual beli secara *online* ini memiliki banyak kelebihan, salah satunya transaksi ini dapat dilakukan dari kota yang berbeda. Aplikasi yang melayani jual beli *online* ini biasa disebut *e-commerce*. Menurut riset dari *iprice*, salah satu *e-commerce* dengan pengunjung paling banyak di Indonesia adalah *Shopee* dengan jumlah pengunjung lebih dari 72 juta tiap bulannya pada akhir tahun 2019. (www.iprice.co.id, 2020)

Pada aplikasi *Shopee*, terdapat ribuan toko *online* yang menjual berbagai macam produk, mulai dari makanan, pakaian, hingga kebutuhan-kebutuhan lainnya. Dalam pencarian produk tertentu, pengguna dihadapkan dengan banyaknya pilihan toko yang menjual produk yang sama. Pengguna tentunya akan memilih toko yang menurut mereka mempunyai penilaian yang paling baik. Penilaian dapat dilakukan dengan mempertimbangkan berapa lama toko tersebut sudah berjualan, berapa jumlah pengikut atau *followers* dari toko tersebut, ataupun dari berapa persen kemungkinan pesan dari pengguna akan dibalas. Pengguna harus melakukan penilaian satu persatu kepada tiap toko dan membandingkannya dengan toko yang lain agar mendapatkan toko *online* yang terbaik.

Hingga sekarang belum ada sistem yang dapat melakukan penilaian terhadap toko-toko pada aplikasi *Shopee*. Tetapi terdapat sebuah sistem serupa, yaitu sistem penilaian kinerja dosen di Universitas Dehasen Bengkulu dengan melakukan *clustering* menggunakan algoritma *K-means* (Sartika & Jumadi, 2019). *Clustering* adalah salah satu metode dalam *data mining* dengan tujuan utama mencari pola dari data untuk kemudian mengelompokkan data tersebut berdasarkan pola yang didapatkan, data dengan pola yang sama akan masuk ke *cluster* yang sama (Adolfsson dkk., 2018). Sedangkan algoritma *K-means* adalah algoritma pada *clustering* yang melakukan pengelompokan dengan membandingkan jarak tiap data ke *centroid* tiap *cluster*, dimana nilai *centroid* didapatkan dari rata-rata tiap *cluster* (Chayangkoon & Srivihok, 2016).

Untuk mempermudah pengolahan data-data tersebut dapat dilakukan *clustering* seperti yang dilakukan dalam penilaian kinerja dosen. *Clustering* dapat dilakukan

untuk mengelompokkan toko-toko yang ada di *Shopee* berdasarkan penilaiannya. *Clustering* dapat dilakukan dengan membagi toko-toko tersebut menjadi 4 *cluster*, yaitu; sangat baik, baik, buruk, dan sangat buruk. Algoritma yang dapat digunakan pada penelitian ini adalah algoritma *K-means*. Algoritma ini sangat sering digunakan dalam melakukan *clustering*. Ketelitian dari algoritma *K-means* cukup tinggi dalam pengolahan data yang besar, selain itu urutan dari objek juga tidak mempengaruhi algoritma ini. (Aranda, 2016).

1.2 Identifikasi Masalah

Permasalahan yang ingin diangkat pada penelitian ini adalah permasalahan dalam melakukan penilaian terhadap toko *e-commerce* pada aplikasi *Shopee*. Dimana dalam penelitian ini akan digunakan metode *clustering* dengan algoritma *K-means* untuk menyelesaikan permasalahan tersebut.

1.3 Rumusan Masalah

Berdasarkan permasalahan yang telah dipaparkan, maka rumusan masalah dari penelitian ini adalah:

1. Bagaimana nilai akurasi dari penggunaan algoritma *K-means* dalam sistem penilaian toko yang dibuat?
2. Berapa waktu yang dibutuhkan untuk menjalankan sistem penilaian toko dengan algoritma *K-means* tersebut?
3. Apa kekurangan ataupun kendala dalam menggunakan algoritma *K-means* dalam menyelesaikan permasalahan penilaian toko?

1.4 Tujuan

Tujuan dilakukannya penelitian ini adalah sebagai berikut:

1. Bertujuan untuk mendapatkan nilai akurasi dari penggunaan algoritma *K-means* dari sistem penilaian toko yang dibuat.
2. Bertujuan untuk mengetahui waktu yang diperlukan untuk menjalankan sistem penilaian toko dengan algoritma *K-means* tersebut.
3. Bertujuan untuk mengetahui kekurangan atau kendala dalam menggunakan algoritma *K-means*.

1.5 Manfaat

Manfaat yang dapat diperoleh dari penelitian ini adalah sebagai berikut:

1. Dapat menerapkan pembelajaran yang selama ini didapatkan dari beberapa mata kuliah yang sudah diambil sebelumnya
2. Dapat membantu masyarakat agar lebih akurat dalam memberikan penilaian terhadap toko-toko *online*.
3. Dapat mengembangkan ilmu pengetahuan dengan menerapkan algoritma *K-means* dalam menyelesaikan permasalahan penilaian.

1.6 Batasan Masalah

Batasan masalah dari penelitian ini adalah sebagai berikut:

1. Sistem yang nantinya dibuat hanya dapat memberikan penilaian toko yang ada di *e-commerce Shopee*.
2. *Cluster* yang terbentuk dari metode *clustering* menggunakan algoritma *K-means* adalah 4, yaitu “sangat baik”, “baik”, “buruk”, dan “sangat buruk”.

1.7 Sistematika Pembahasan

Berikut merupakan sistematika penulisan pada penelitian ini:

1. BAB 1 PENDAHULUAN
Bab Pendahuluan berisi uraian tentang permasalahan yang akan diteliti yaitu permasalahan penilaian toko *e-commerce* serta tujuan dan manfaat dilakukannya penelitian tersebut.
2. BAB 2 TINJAUAN PUSTAKA
Bab Tinjauan Pustaka berisi tentang Kajian Pustaka tentang penelitian sebelumnya yang berhubungan dengan penilaian toko *e-commerce* dengan *clustering*. Pada bab ini dicantumkan pula teori-teori yang akan digunakan pada penelitian, yaitu *Data Clustering*, Algoritma *K-Means*, dan Evaluasi dengan *Silhouette Coefficient*.
3. BAB 3 METODOLOGI PENELITIAN
Bab Metodologi Penelitian berisi tentang metode penelitian yang digunakan pada penilaian toko *e-commerce* dengan *clustering*.

BAB 2 TINJAUAN PUSTAKA

2.1 Kajian Pustaka

Penelitian tentang sistem penilaian dengan menggunakan metode *clustering* sudah cukup banyak dilakukan. Salah satu penelitian tentang sistem penilaian adalah penelitian tentang *clustering* penilaian kinerja Dosen menggunakan algoritma *K-means* dengan studi kasus Universitas Dehasen, Bengkulu. Penelitian ini dilakukan oleh Devi Sartika dan Juju Jumadi pada tahun 2019. Dalam penelitiannya, penilaian dosen dilakukan dengan mengelompokkan data dosen ke dalam 4 kelompok penilaian. 4 kelompok penilaian ini adalah kelompok “Dosen Sangat Baik”, “Dosen Baik”, “Dosen Cukup Baik”, dan “Dosen Kurang Baik”. Fitur pada penelitian ini berjumlah 14. Fitur ini didapatkan dari 14 pertanyaan pada kuesioner yang diisi oleh mahasiswa.

Penelitian lain mengenai sistem penilaian dilakukan oleh Fajar Nur Rohmat pada tahun 2018. Penelitian yang ia buat adalah implementasi algoritma *K-means* untuk klasterisasi kinerja akademik mahasiswa. Dalam penelitian ini, penilaian kinerja mahasiswa dilakukan dengan membagi data mahasiswa menjadi 3 kelompok. Hasil implementasi dari penelitian tersebut berupa aplikasi *desktop*. Nilai evaluasi dengan *silhouette coefficient* yang didapatkan dari sistem penilaian kinerja akademik mahasiswa ini adalah 0,1087.

Berikut merupakan tabel dari kajian pustaka:

No	Judul	Objek	Metode
1	<i>Clustering</i> Penilaian Kinerja Dosen Menggunakan Algoritma <i>K-means</i> dengan Studi Kasus Universitas Dehasen, Bengkulu (Sartika & Jumadi, 2019)	Dosen Universitas Dehasen	<i>K-means</i>
2	Implementasi Algoritma <i>K-means</i> untuk Klasterisasi Kinerja Akademik Mahasiswa (Rohmat, 2018)	Mahasiswa tahun ajaran 2012	<i>K-means</i>

Tabel 2.1 Tabel Kajian Pustaka

2.3 Data Clustering

Data Clustering atau pengklasteran data adalah salah satu teknik pada *data mining* yang berguna untuk menganalisis data dan mengelompokkan data atau mempartisi *dataset* ke dalam *subset*. Tujuan dari *data clustering* adalah untuk melakukan pendistribusian objek ke dalam kelompok-kelompok dimana keterhubungan objek pada kelompok yang sama sangat kuat dan dengan keterhubungan objek pada kelompok yang berbeda sangat lemah. (Suyanto, 2017)

2.4 Algoritma K-Means

Algoritma *K-means* adalah salah satu algoritma dalam *data clustering* yang mengelompokkan data menjadi *cluster-cluster*, dimana data-data dengan karakteristik yang sama akan dikelompokkan ke dalam satu *cluster* dan di dalam sebuah *cluster* memiliki tingkat variasi yang kecil. (Suyanto, 2017)

Berikut merupakan langkah-langkah *clustering* dengan menggunakan algoritma *K-means*:

1. Pilih nilai K yang menunjukkan jumlah *cluster* yang ingin dibentuk.
2. Masukkan nilai acak sebanyak K secara acak (*random*) sebagai nilai titik awal pusat dari *cluster*.
3. Alokasikan tiap data ke *cluster* terdekat. Cari nilai jarak dari tiap data ke tiap titik pusat dari *cluster* yang sebelumnya telah ditentukan secara acak. Hal ini dilakukan untuk menentukan kedekatan antara data tersebut dengan tiap *cluster*. Nilai jarak dapat dicari dengan jarak *Euclidean*:

$$D_{ij} = \sqrt{(x_{1i} - x_{1j})^2 + (x_{2i} - x_{2j})^2 + \dots + (x_{ni} - x_{nj})^2}$$

Keterangan:

D_{ij} = jarak data ke i ke pusat cluster j

x_{ni} = data ke i pada atribut data ke n

x_{nj} = data pusat ke j pada atribut ke n

4. Hitung kembali nilai titik pusat *cluster* sesuai dengan keanggotaan yang terbentuk. Titik pusat *cluster* adalah rata-rata semua data yang ada di dalam *cluster*. Nilai rata-rata juga dapat digantikan dengan nilai median.

5. Lakukan langkah ke 3 dan ke 4 sampai nilai titik pusat *cluster* tidak berubah lagi. Perulangan juga dapat dibatasi dengan membatasi jumlah iterasi yang dilakukan.

2.5 Silhouette Coefficient

Metode *silhouette coefficient* bertujuan untuk menguji kualitas dari *cluster* yang dihasilkan dan juga sebagai metode untuk pengesahan suatu *cluster* yang menggabungkan antara metode *cohesion* dan metode *separation*. (Rohmat, 2018) Berikut merupakan tahapan dalam mencari nilai *silhouette coefficient*:

1. Pada tiap objek, hitung nilai rata-rata titik satu dengan semua objek yang berada di dalam *cluster* yang sama. Akan diperoleh nilai rata-rata yang disebut dengan a_i .
2. Pada tiap objek, hitung nilai minimal jarak rata-rata dari titik satu ke titik lain yang berada di *cluster* berbeda. Akan diperoleh nilai rata-rata minimum yang disebut dengan b_i .
3. Setelah semua nilai diketahui, tentukan nilai *silhouette coefficient* dengan rumus berikut.

$$S_i = \frac{b_i - a_i}{\max(a_i - b_i)}$$

Dimana:

S_i = nilai *silhouette coefficient*

a_i = rata-rata jarak titik satu dengan semua data yang berada dalam cluster yang sama

b_i = minimal jarak rata-rata dari titik satu ke titik lain yang berada pada *cluster* yang berbeda

Hasil perhitungan dari nilai *silhouette coefficient* berada pada rentang -1 hingga 1. Hasil *silhouette coefficient* dapat dikatakan baik apabila nilai bernilai positif, yang berarti titik sudah berada pada *cluster* yang benar. Jika hasil *silhouette coefficient* bernilai negatif artinya terjadi *overlapping*, yaitu titik berada di antara dua *cluster*. Berikut merupakan penilaian dari hasil *silhouette coefficient* yang didapatkan:

1. *Strong Structure*
 $0,7 < \text{silhouette coefficient} \leq 1$
2. *Medium Structure*
 $0,5 < \text{silhouette coefficient} \leq 0,7$

3. *Weak Structure*
 $0,25 < \text{silhouette coefficient} \leq 0,5$
4. *No Structure*
 $\text{silhouette coefficient} \leq 0,25$

BAB 3 METODOLOGI PENELITIAN

3.1 Tipe Penelitian

Tipe penelitian ini adalah *non-implementatif*. Penelitian ini hanya bertujuan untuk menganalisis data untuk menyelesaikan permasalahan penilaian toko *e-commerce*. Pendekatan yang digunakan dalam penelitian ini adalah pendekatan analitik. Pendekatan analitik dapat menjelaskan hubungan fitur-fitur pada permasalahan yang sedang diteliti, sehingga hasil yang didapatkan akan dianalisis menjadi informasi yang lebih berguna.

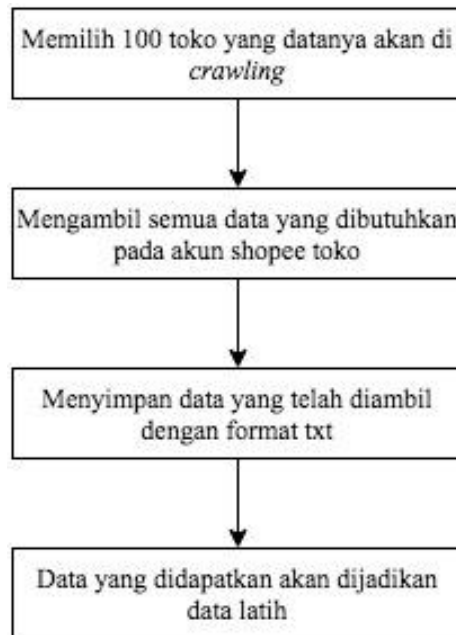
3.2 Strategi Penelitian

Strategi penelitian yang digunakan pada penelitian ini adalah strategi studi kasus. Studi kasus dilakukan untuk mencari penyelesaian dari permasalahan dengan melakukan penilaian pada toko. Penelitian ini difokuskan untuk mencari *cluster* yang tepat untuk tiap data pada data latih.

3.3 Teknik Pengumpulan Data

Pengumpulan data pada penelitian ini menggunakan artikel dari web *e-commerce Shopee*. Artikel ini diambil menggunakan *web crawling*. *Web crawling* ini dilakukan dengan menggunakan salah satu *library* dari python yang dapat mengambil data dari format *file* HTML dan XML, yaitu *BeautifulSoup*. Selain itu, digunakan juga modul *urllib* untuk mengambil data dari *web*. *Web crawling* yang dibentuk hanya dapat digunakan untuk *web* www.shopee.com saja, karena pola HTML dari setiap *web* dapat berbeda-beda. Setelah melakukan *crawling*, data yang diperoleh akan dimasukkan ke dalam *file dokuji.txt* yang nantinya akan digunakan sebagai data latih.

Berikut ini tahap pengumpulan data dengan web crawling.

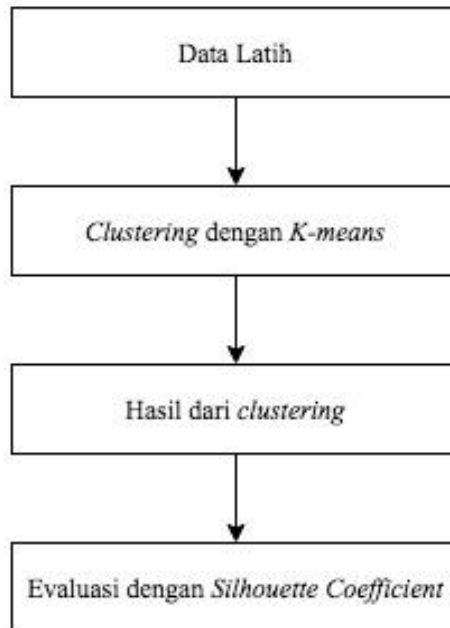


Gambar 3.1 Gambar Diagram Teknik Pengumpulan Data

3.4 Data Penelitian

Pada penelitian ini, data yang digunakan adalah 100 data toko yang ada di *Shopee* dengan fitur jumlah produk, jumlah pengikut, jumlah mengikuti, penilaian, jumlah penilai, performa *chat*, dan waktu bergabung. Pengambilan data 100 toko tersebut dilakukan secara acak.

3.5 Implementasi Algoritma



Gambar 3.2 Gambar Diagram Implementasi Algoritma

Pada implementasi algoritma, dilakukan beberapa proses untuk melakukan penilaian dengan *clustering*. Berikut merupakan proses yang dilakukan:

1. Data Latih

Data latih yang digunakan pada penelitian ini adalah data yang didapatkan dari hasil *web crawling* pada web www.shopee.com. Data yang didapatkan dari *web crawling* tersebut akan disimpan dengan format txt untuk nantinya digunakan pada proses penelitian.

2. *Clustering* dengan *K-means*

Pada tahap ini, akan dilakukan *clustering* dengan algoritma *K-means*. Jumlah *cluster* yang akan digunakan pada penelitian ini adalah 4. *Clustering* dengan metode ini dilakukan dengan mencari jarak tiap data dengan titik pusat awal dengan menggunakan *Euclidean distance*. *Cluster* dari data ditentukan dari jarak terkecil antara data tersebut dengan pusat *cluster*. Pusat *cluster* nantinya akan diperbarui dengan mencari nilai rata-rata dari tiap anggota *cluster* tersebut. Proses ini akan diulang hingga titik pusat dari *cluster* tidak berubah.

3. Hasil dari *clustering*

Setelah menjalankan tahap *clustering* dengan algoritma *K-means* maka akan didapatkan hasil berupa keanggotaan dari tiap *cluster*. Jumlah *cluster* yang terbentuk adalah 4. Dari hasil *clustering* ini akan terlihat karakteristik dari fitur pada tiap *cluster* yang terbentuk.

4. Evaluasi dengan *Silhouette Coefficient*

Untuk memastikan bahwa hasil dari *clustering* sudah cukup baik, akan dilakukan evaluasi dengan menggunakan *Silhouette Coefficient*. Metode ini akan mencari tahu apakah ada data yang *overlapping* atau berada di antara 2 *cluster*. Nilai *silhouette coefficient* yang baik adalah nilai yang mendekati nilai 1.

3.6 Teknik Pengujian dan Analisis

Dalam menjalankan penelitian ini, teknik analisis dan pengujian yang akan dilakukan adalah dengan mencari nilai *silhouette coefficient* sebagai evaluasi dari hasil *clustering* yang terbentuk dengan algoritma *K-means*.

DAFTAR PUSTAKA

- iprice.co.id. 2019. Peta E-Commerce Indonesia. Diambil dari <https://iprice.co.id/insights/mapofecommerce/>. Diakses pada 30 Maret 2020.
- Sarika, Devi & Jumadi, Juju. 2019. Clustering Penilaian Kinerja Dosen Menggunakan Algoritma K-Means (Studi Kasus: Universitas Dehasen Bengkulu). Seminar Nasional Teknologi Komputer & Sains, 703-709. Diambil dari www.prosiding.seminar-id.com/index.php/sainteks/article/view/218.
- Adolfsson, A., Ackerman, M., Brownstein, N.C., 2018. To cluster, or not to cluster: an analysis of clusterability methods. doi: <https://doi.org/10.1016/j.patcog.2018.10.026>.
- Chayangkoon, N. & Srivihok, A., 2016. Two Step Clustering Model for K-Means Algorithm. ICNCC'16, pp. 213-217. Kyoto: ACM.
- Aranda, J., Natasya, WAG. 2016. Penerapan Metode K-Means Cluster Analysis Pada Sistem Pendukung Keputusan Pemilihan Konsentrasi Untuk Mahasiswa International Class STMIK AMIKOM Yogyakarta. Jurnal Karya Ilmiah Teknik Informatika. Volume 4, No 1.
- Suyanto, D. 2017. Data Mining untuk Klasifikasi dan Klasterisasi Data. Bandung: Informatika Bandung.
- Rohmat, F. N. 2018. Implementasi Algoritma K-means untuk Klasterisasi Kinerja Akademik Mahasiswa. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer. Vol. 2, No. 6, hlm. 2243-2251. Malang: Universitas Brawijaya.