PENERAPAN ALGORITMA K-MEANS CLUSTERING UNTUK PENGELOMPOKAN UNIVERSITAS TERBAIK DI DUNIA

[1]Faisal Dikarya, [2]Sita Muharni

[1,2]STMIK Dharmawacana Metro

faixal279@gmail.com[1], sitamuharni@dharmawacana.ac.id[2]

ABSTRACT

University is a college that teaches various sciences formed from various faculties that establish academic and vocational education and also provides academic degrees in various fields. Almost all tertiary institutions claim to be the best, with many universities today, many considerations will be taken in choosing a university, because each university has different facilities and education systems. so to find out which is the best, an accurate method is needed such as the K-Means algorithm method to classify the best universities from many universities. This study aims to classify the best universities into 3 clusters, namely high, medium and low clusters using 4 attributes including world rank, institution, country, and scores from the top 2000 university data in the world. The data for this study is quoted from the website kaggle.com/datasets. The method used is the K-Means algorithm method which is processed with the rapidminer application. Based on the test results, it can be concluded that the best in cluster 2 is Harvard University with 667 universities, in cluster 1 the University of Haifa with 667 universities, and in cluster 0 National Chung Cheng University with 666 units.

Keywords— Data Mining, RapidMiner, K-Means

ABSTRAK

Universitas ialah perguruan tinggi yang mengajarkan berbagai ilmu pengetahuan yang terbentuk dari beragam fakultas yang mendirikan pendidikan akademik dan vokasi dan juga memberikan gelar akademis dalam beragam bidang. Hampir semua perguruan tinggi mengklaim sebagai yang terbaik, dengan banyaknya universitas saat ini akan membuat banyak pertimbangan dalam memilih universitas, karena setiap universitas memiliki fasilitas dan sistem pendidikan yang berbeda. maka untuk mengetahui mana yang terbaik diperlukan metode yang akurat seperti metode algoritma K-Means untuk mengelompokan universitas terbaik dari banyaknya universitas. Penelitian ini bertujuan untuk mengelompokankan universitas terbaik menjadi 3 *cluster* yaitu *cluster* tinggi, sedang, dan rendah menggunakan 4 atribut antara lain *world rank*, *institution*, *country*, dan *score* dari 2000 data universitas teratas didunia. Data study ini dikutip dari Website kaggle.com/datasets. Metode yang digunakan ialah metode algoritma K-Means yang diolah dengan aplikasi rapidminer. Bersumber dari hasil uji bisa disimpulkan maka yang terbaik di *cluster* 2 adalah Harvard University dengan 667 Universitas, di *cluster* 1 University of Haifa dengan 667 Universitas, dan di *cluster* 0 National Chung cheng University dengan 666 Unitersitas.

Kata Kunci— Data Mining, RapidMiner, K-Means

I. PENDAHULUAN

Pendidikan ialah kebutuhan primer manusia, dalam bersosial dan mencari uang. Berdasarkan perguruan tinggi status keberhasilan mahasiswa juga pengalaman menjadi tolak ukur kualitas mahasiswa untuk memasuki dunia kerja [1]. Pada saat ini universitas berkembang dengan sangat cepat dan banyak universitas yang sudah berdiri diseluruh dunia termasuk Indonesia. Universitas ialah perguruan tinggi yang terdiri dari beragam fakultas, yang mengajarkan berbagai ilmu pengetahuan dan juga membuat pendidikan akademik dan vokasi dalam memberikan gelar akademis diberbagai bidang. Nama universitas sendiri bersumber pada bahasa Latin *universitas magistrorum et scholarium* yang mempunyai arti komunitas guru dan akademisi, sejak dulu universitas telah berdiri lama di Asia dan Afrika [2]. Fungsinya adalah tempat untuk mengembangkan keilmuan, merubah pola pikir, keterampilan sosial dan karakter. Dengan banyaknya pilihan universitas saat ini akan membuat banyak pertimbangan dalam memilih universitas dalam negeri ataupun luar negeri, karena setiap universitas memiliki fasilitas dan sistem pendidikan yang berbeda [3]. Semakin bagus kualitas pendidikan maka akan semakin meningkatkan kualitas universitasnya. Berdasarkan hal tersebut diperlukan metode yang akurat seperti metode algoritma K-Means untuk mengelompokan universitas terbaik dari banyaknya universitas [4]. Algoritma K-Means ialah salah satu dari banyaknya proses yang dipakai dalam kegiatan cluster, metode inilah yang efektif untuk menghasilkan cluster-cluster dari data kecil maupun besar.

Beberapa penelitian terutama clustering sering digunakan untuk mengelompokan data salah satunya yaitu diteliti oleh Venny Novita Sari, Yupianti, dan Dewi Maharani, "Penerapan Algoritma K-Means Dalam Mengelompokan Kualitas Lulusan Mahasiswa Fakultas Ilmu Computer Universitas Dehasan Bengkulu" menurut Nilai akhir, Metode yang dipakai yaitu K-Means clustering, yaitu diolah menggunakan software rapidminer dan dikelompokan berdasarkan karakteristik yang serupa supaya kelompok tidak menjadi tumpang tindih. hasil study didapati bahwa group mahasiswa yang mendapatkan IPK paling tinggi dari ketiga *claster* yaitu *cluster* 2 pada rata-rata IPK 3.3967 sehingga dapat diketahui bahwa *cluster* 2 yakni lulusan mahasiswa yang mempunyai kualitas

nilai terbaik [5].

II. METODE PENELITIAN

Sumber data riset ini diperoleh dari website kaggle.com/datasets yaitu data 2000 universitas teratas didunia dan Atribut yang digunakan ialah *world rank*, *institution*, *country*, dan *score*. Kaggle adalah situs kumpulan data yang berisikan data-data terbuka secara global, serta digunakan untuk mencari kumpulan data tertentu, sebagian besar dataset dalam format CSV [4].

Sebelum mengaplikasikan teknik cluster, terlebih dahulu ditentukan jumlah clusternya salah satu caranya adalah dengan menggunakan kriteria *static Within Sum of Square* (WSS). WSS ialah salah satu kriteria untuk menghitung keragaman data dari cluster yang terbentuk, semakin kecil anekaragam cluster yang terbentuk menandakan bahwa cluster sudah selesai terbentuk. cluster dibagi berdasarkan tingkatnya yaitu tinggi, sedang, dan rendah. Kita dapat membendingkan jumlah cluster yang terbaik untuk menganalisa data.

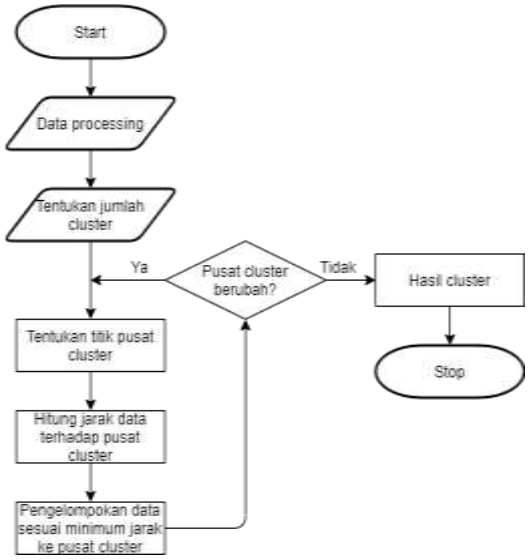
Metode yang digunakan pada penelitian ini ialah metode data mining, data mining merupakan cara pengumpulan dan pengolahan data, data itu sendiri bisa di peroleh dari website Kaggle.com Satudata.com dan lainnya [6]. Untuk diekstrak menjadi informasi penting pada data kecil maupun besar, data mining dibagi menjadi empat, yakni model prediksi, analisis *Cluster*, analisis asosiasi, dan deteksi. Penulis akan menggunakan analisis kelompok(*Cluster*) yang digunakan untuk mengelompokkan universitas.

Selanjutnya menggunakan aplikasi Rapidminer dengan menerapkan Algoritma K-Means, RapidMiner adalah aplikasi atau perangkat lunak yang biasa digunakan sebagai alat dalam ilmu data mining. Rapidminer ialah jawaban bagi user untuk menganalisa data, text, cluster dan prediksi. Kali ini penulis akan menggunakan Aplikasi Rapidminer yang berfokus pada pengelompokkan.

Algoritma K-Means adalah algoritma yang umum dipakai sebagai *clustering* atau pengelompokan data didalam aplikasi Rapidminer. Prinsip Algoritma K-Means ialah menata k *prototype* atau inti massa (*centroid*) dari sekelompok data [7]. Algoritma K-Means ialah algoritma yang memastikan nilai-nilai *cluster* (k) secara acak, selagi nilai tersebut menjadi inti dari *cluster* yang sering disebut *centroid*. Prinsip Kerja dari pengelompokan atau clustering dilakukan secara bertahap. Dan *clustering* satu-satunya pilihan kombinasi suatu objek terhadap objek yang lain [8].

A. *Modeling*

Berikut ini adalah cara perhitungan algoritma K-Means dapat dilihat pada gambar 1 :



Gambar 1. Flowchart alur metode algoritma K-Means

Data 2000 universitas teratas di dunia lalu di proses menggunakan metode algoritma K-Means

Algoritma K-Means sebagai berikut [9]:

a. Tentukan jumlah *cluster* (k), tetapkan pada pusat cluster sembarang.
b. Hitung jarak antara setiap data ke dalam cluster dengan jarak yang paling pendek dengan menggunakan persamaan ukuran jarak *Euclidean distance* dengan persamaan (1) :

$$D_1(X_1, X_1) = \| x_2 - x_1 \| = \sqrt{\sum_{j=1}^{p} \{x_{2j} = x_{1j}\}^2} \quad (1)$$

$D_1(X_1, X_2)$ ialah jarak diantara data ke-i dan data ke-j, X_{2j} adalah koordinat data X_2 pada dimensi j, XX_{1j} adalah kordinat data X_1 pada dimensi j dan P dimensi data

c. Himpunan data ke dalam *cluster* serta jarak yang sangat pendek dengan menggunakan rumus pada persamaan (2):

$$V_{ij} = \frac{\sum_{k=1}^{Ni} x_{kj}}{Ni} \quad (2)$$

V_{ij} adalah data cluster ke – I kolom j, X_{kj} adalah data ke – kkolom ke j, N_i banyaknya anggota cluster ke-i

d. Hitung pada pusat *cluster* yang baru menggunakan persamaan (1) ulangi langkah a) sampai dengan d) hingga tidak terjadi lagi perpindahan data pada *cluster* yang berbeda.

Dari metode ini penulis memilih data dari beberapa jurnal yang di dapatkan di website e-jurnal, yang bersangkutan dengan persoalan yang sedang di bahas dan khususnya dalam penelitian penerapan metode algoritma K-Means sebagai referensi [10].

III. HASIL DAN PEMBAHASAN

Pada proses awal melakukan *modelling*, mencari *modeling*, *read excel* lalu *drag* ke dalam proses [11]. selanjutnya melakukan

import data *excel* 2000 universitas teratas didunia kemudian lakukan *modeling* K-Means, dengan iterasi berlangsung sebanyak 10 kali data dibagi menjadi 3 *cluster* : *cluster* tinggi, sedang dan rendah. Dibaginya 3 cluster lebih cocok digunakan dalam penelitian terkait pengelompokan dari 2000 data universitas dengan metode K-Means *clustering* dibandingkan 4 atau 5 cluster karena dengan 3 cluster lebih mudah dipahami antara universitas yang masuk kategori bagus, sedang, dan rendah. Modeling K-Means dapat dilihat pada gambar 2.

Gambar 2. Proses Modeling K-Means

Melakukan proses visualisasi diagram *step area* dengan memakai *Software* Rapid Miner [12], dapat dilihat pada Gambar 3 :

Gambar 3. *Cluster* Diagram *Step Area*

Gambar 4 memperlihatkan hasil dari *cluster* model terdapat tiga *cluster* yaitu *cluster* 0 : 666 item, *cluster* 1 : 667 item, dan *cluster* 2 : 667 item.

Cluster Model

```
Cluster 0: 666 items
Cluster 1: 667 items
Cluster 2: 667 items
Total number of items: 2000
```

Gambar 4. Hasil Dari *Cluster* Model

Cluster 0 ini terdapat 666 Universitas, dari *cluster* 0 ini yang tertinggi yaitu National Chung cheng Unversity yang berada di Taiwan dan yang terendah adalah Huzhou University yang berada di China. Tabel 1, yang di tampilkan ini hanya berjumlah 10 yang sebenarnya berjumlah 666.

Tabel 1. *Cluster* 0

Institution	country	score
National Chung cheng University	Taiwan	68
National Chengchi University	Taiwan	68
Antonio NariA+o University	Colombia	68
Coventry University	United kingdom	68
UludaAY University	Turkey	68
University of Agriculture Faisalabad	Pakistan	68
University of Ghana	Ghana	68
Okinawa Institute of sience technology	Japan	68
Southem Cross University	Australia	68
……...		
Huzhou University	China	66

Cluster 1 ini terdapat 667 Universitas, dari *cluster* 1 ini yang tertinggi yaitu

University of Haifa yang berada di Israel dan yang terendah adalah University of Centraln Lancashire yang berada di United Kingdom. Tabel 2 yang di tampilkan ini hanya berjumlah 10 yang sebenarnya berjumlah 667.

Tabel 2. *Cluster* 1

Institution	Country	score
University of Haifa	Israel	73
Kanazawa University	Japan	73
University of Wyoming	USA	73
University of Mailaga	Spain	73
Tokyo Medical and Dental University	Japan	73
Istanbul University	Turkey	73
University of Johannesburg	South africa	73
University of Texas at San Antonio	USA	73
University of Loannina	Greece	73
……		
University of Central Lancashire	United kingdom	68

Cluster 2 ini terdapat 667 Universitas, dari *cluster* 2 ini yang tertinggi yaitu Harvard University yang berada di USA, dan yang terendah adalah University of Calabria yang berada di Italy. Tabel 3 yang di tampilkan ini hanya berjumlah 10 yang sebenarnya 667.

Tabel 3. *Cluster* 2

Institution	Country	score
Harvard University	USA	100
Massachusetts Institute of Technology	USA	97
Stanford University	USA	95
University of Cambridge	United kingdom	94
University of Oxford	United kingdom	93
Princeton University	USA	93
Columbia University	USA	92
University of Chicago	USA	92
University of PennsyIvania	USA	91
……		
University of Calabria	Italy	73

IV. SIMPULAN

Kesimpulan dari hasil penelitian pengelompokan universitas terbaik di dunia dari 2000 data menggunakan algoritma k-means *clustering* ialah sebagai berikut:

1. Pengujian dari penelitian ini dilakukan iterasi *clustering* dengan data 2000 universitas teratas di dunia terjadi sebanyak 10 kali terdapat tiga *cluster* yaitu *cluster* rendah, *cluster* sedang, dan *cluster* tinggi.
2. Ketiga cluster tersebut telah dilakukan profiling cluster, cluster tertinggi pada cluster 2 dengan 667 universitas, selanjutnya pada cluster sedang terdapat 667 universitas, dan pada cluster rendah terdapat 666 universitas.
3. Dari 2000 data universitas teratas didunia dapat diketahui yang terbaik adalah Harvard University di *cluster* 2, University of Haifa di *cluster* 1, dan National Chung Cheng University di *cluster* 0.

DAFTAR PUSTAKA

[1] N. Rohmawati, S. Defiyanti, and M. Jajuli, "Implementasi Algoritma K-Means Dalam Pengklasteran Mahasiswa Pelamar Beasiswa," *J. Ilm. Teknol. Infomasi Terap.*, vol. 1, no. 2, 2015.

[2] H. Susanto and S. Sudiyatno, "Data mining untuk memprediksi prestasi siswa berdasarkan sosial ekonomi, motivasi, kedisiplinan dan prestasi masa lalu," *J. Pendidik. Vokasi*, vol. 4, no. 2, pp. 222–231, 2014, doi: 10.21831/jpv.v4i2.2547.

[3] T. H. Sardar and Z. Ansari, "An analysis of MapReduce efficiency in document clustering using parallel K-means algorithm," *Futur. Comput. Informatics J.*, vol. 3, no. 2, pp. 200–209, 2018, doi: 10.1016/j.fcij.2018.03.003.

[4] G. Gustientiedina, M. H. Adiya, and Y. Desnelita, "Penerapan Algoritma K-Means Untuk Clustering Data Obat-Obatan," *J. Nas. Teknol. dan Sist. Inf.*, vol. 5, no. 1, pp. 17–24, 2019, doi: 10.25077/teknosi.v5i1.2019.17-24.

[5] V. N. Sari, Y. Yupianti, and D. Maharani, "Penerapan Metode K-Means Clustering Dalam Menentukan Predikat Kelulusan Mahasiswa Untuk Menganalisa Kualitas Lulusan," *JURTEKSI (Jurnal Teknol. dan Sist. Informasi)*, vol. 4, no. 2, pp. 133–140, 2018.

[6] R. R. Putra and C. Wadisman, "Implementasi Data Mining Pemilihan Pelanggan Potensial Menggunakan Algoritma K Means," *INTECOMS J. Inf. Technol. Comput. Sci.*, vol. 1, no. 1, pp. 72–77, 2018.

[7] E. Irfiani and S. S. Rani, "Algoritma K-Means Clustering untuk Menentukan Nilai Gizi Balita," *JUSTIN (Jurnal Sist. dan Teknol. Informasi)*, vol. 6, no. 4, pp. 165–172, 2018.

[8] C. V. L. N. Abadi and K. L. K. Malang, "BUKU MODUL VISUALISASI DATA MENGGUNAKAN DATA STUDIO," 2022.

[9] M. N. M. Ediyanto and N. Satyahadewi, "Pengklasifikasian Karakteristik Dengan Metode K-Means Cluster Analysis," *Bimaster Bul. Ilm. Mat. Stat. dan Ter.*, vol. 2, no. 02, 2013.

[10] A. Sulistiyawati and E. Supriyanto, "Implementasi Algoritma K-means Clustring dalam Penetuan Siswa Kelas Unggulan," *J. Tekno Kompak*, vol. 15, no. 2, pp. 25–36,

2021.

[11] S. Muharni, S. Andriyanto, and D. Naista, "IMPLEMENTASI DEMPSTER SHAFER UNTUK MENDIAGNOSA GANGGUAN KEHAMILAN PADA IBU," *J. Inform.*, vol. 21, no. 2, pp. 146–160, 2021.

[12] Y. A. Priambodo and S. Y. J. Prasetyo, "Pemetaan Penyebaran Guru di Provinsi Banten dengan Menggunakan Metode Spatial Clustering K-Means (Studi kasus: Wilayah Provinsi Banten)," *Indones. J. Comput. Model.*, vol. 1, no. 1, pp. 18–27, 2018.