

## PENERAPAN ALGORITMA *K-MEANS CLUSTERING* UNTUK PENGELOMPOKAN KECELAKAAN BERKENDARA DI RUAS TOL JAKARTA-CIKAMPEK

Mochamad Rizky Sulaeman<sup>1</sup>, Purwantoro<sup>2</sup>, Tesa Nur Padilah<sup>3</sup>

Universitas Singaperbangsa Karawang<sup>12</sup>

Jl. HS.Ronggo Waluyo, Puseurjaya, Telukjambe Timur, Karawang, Jawa Barat 41361

E-mail : Mochamad.rizky18165@student.unsika.ac.id<sup>1</sup>, Purwantoro.masbro@staff.unsika.ac.id<sup>2</sup>,

Tesa.nurpadilah@staff.unsika.ac.id<sup>3</sup>

### ABSTRAK

Tingginya kecelakaan berkendara di jalan tol merupakan bahan evaluasi. Sehingga menjadi perhatian khusus bagi perusahaan penyedia layanan jalan tol baik milik negara maupun swasta, yang telah membuktikan perhatian tersebut dengan adanya perbaikan, penambahan infrastruktur dan edukasi kepada pengguna jalan guna meminimalisir kecelakaan di jalan tol. Tahapan awal pencegahan kecelakaan berkendara adalah dengan mengetahui faktor-faktor penyebab kecelakaan berkendara yang diperoleh melalui analisa data kecelakaan. Analisa tersebut dapat dilakukan dengan *Data Mining*, yaitu *K-Means Clustering*. *K-Means Clustering* mengelompokkan data menjadi beberapa *cluster* sesuai karakteristik data tersebut. Tahapan klasterisasi yang dilakukan dengan menentukan uji coba banyaknya klaster yaitu dengan menetapkan  $k=3$ ,  $k=5$  dan  $k=7$ , dan *performancenya* menggunakan *Davis Bouldin Index* (DBI). Hasil *cluster* penerapan *K-Means Clustering* dilakukan pengujian untuk menentukan model klaster terbaik yang diuji dan mengacu pada evaluasi *performance* DBI yang mendekati nilai Nol ( nilai Terbaik) secara berurutan sehingga menghasilkan nilai  $k=7$  adalah nilai DBI terbaik yaitu 0.179 sedangkan untu  $k=5$  diperoleh DBI sebesar 0.180 dan  $k=3$  nilai DBI sebesar 0.233.

**Kata kunci :** *K-Means, Clustering, Kecelakaan, Jalan tol*

### ABSTRACTS

*The high number of driving accidents on toll roads is an evaluation material. So that it becomes a special concern for toll road service providers, both state-owned and private, which have proven this concern by improving, adding infrastructure and educating road users to minimize accidents on toll roads. The initial stage of preventing driving accidents is to find out the factors that cause driving accidents obtained through accident data analysis. The analysis can be done with Data Mining, namely K-Means Clustering. K-Means Clustering groups the data into several clusters according to the characteristics of the data. The clustering stage is carried out by determining the number of cluster trials, namely by setting  $k=3$ ,  $k=5$  and  $k=7$ , and the performance is using the Davis Bouldin Index (DBI). The results of the cluster application of K-Means Clustering are tested to determine the best cluster model that tested and refers to the evaluation of DBI performance which approaches the value of Zero (Best value) sequentially so that the value of  $k=7$  is the best DBI value of 0.179 while for  $k=5$  the DBI is 0.180 and  $k=3$  the DBI value is 0.233.*

*Keywords: K-Means, Clustering, Accident, Toll Road*

## 1. PENDAHULUAN

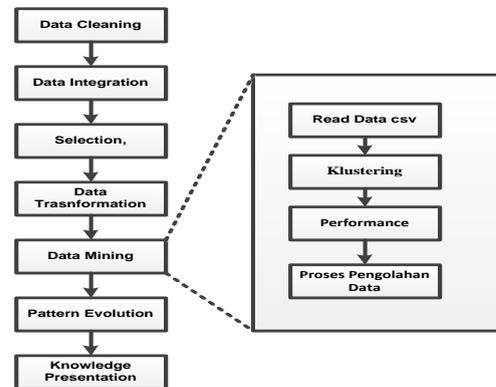
Teknologi hadir sebagai sarana mempermudah manusia dalam menjalankan aktifitasnya sehari-hari. Laju perkembangan teknologi pada semua lini sangat cepat perubahannya. Perkembangan teknologi komputer pada era Revolusi Industri 4.0 sekarang sudah sangat menjadi ketergantungan setiap orang dalam melakukan aktifitas sehari harinya, ditambah dengan kondisi era pandemi Covid 19 menjadikan teknologi komputer sebagai piranti yang seakan wajib dimiliki dan digunakan semua kalangan masyarakat. Perkembangan perangkat keras (*Hardware*) dan perangkat lunak (*Software*) sangat membantu manusia dalam pengelolaan dan manipulasi data pada semua bidang, pengelolaan data dalam proses prediksi, estimasi, klasifikasi, klusterisasi, proses untuk mengetahui hubungan keterkaitan sesuatu dengan sesuatu yang lain (asosiasi) ataupun pengelolaan dan manipulasi data lainnya [1]. Penelitian ini akan mengelompokkan kasus kecelakaan di ruas tol Jakarta-Cikampek berdasarkan hasil riset dari data PT Jasa Marga *Tollroad* Operator Cabang Jakarta-Cikampek. Metode yang digunakan adalah metode *clustering k-means*, Alasan penggunaan algoritma *k-means* karena algoritma ini memiliki ketelitian yang cukup tinggi terhadap ukuran objek, sehingga algoritma ini relatif lebih terukur dan efisien untuk pengolahan objek dalam jumlah besar. Selain itu algoritma *k-means* ini tidak terpengaruh terhadap urutan objek [2]. Jumlah kelompok ditentukan dari evaluasi dilakukan dengan mencari nilai terkecil dari nilai *Davies Bouldin index* sebagai cara untuk optimasi jumlah *cluster* [3].

## 2. METODE PENELITIAN

### 2.1 Knowledge Discovery in Database

KDD adalah salah satu metode yang dapat digunakan dalam melakukan data *mining*, KDD didefinisikan sebagai proses penggunaan metode data *mining* untuk menemukan informasi yang berharga, pola yang ada di dalam data, yang melibatkan algoritma untuk mengidentifikasi pola pada data [4]. Metode Penelitian ini akan menerapkan alur kerja tujuh tahapan terstruktur yang diharapkan mampu

mengarahkan penelitian agar sesuai dengan tujuan. Kerangka kerja akan digunakan adalah *Knowledge Data Discovery* yang tersusun atas *Data Cleaning*, *Data integration*, *Data Selection*, *Data Transformation*, *Data Mining*, *Pattern Evolution*, dan *Knowledge Presentation*.



Gambar 1. Tahapan Knowledge Data Discovery

Tujuh tahapan pada KDD pada gambar 1 yaitu:

1. *Data Cleaning* pada tahap ini adalah tahapan awal dilakukannya pemilihan data yang akan diolah kemudian dipilih data yang dianggap layak pakai.
2. *Data Integration* pada tahap ini proses yang akan dilakukan adalah menyatukan data yang dianggap berulang menjadi satu data.
3. *Data Selection* pada tahap ini proses yang akan dilakukan adalah pemilihan data yang dianggap relevan terhadap analisis.
4. *Data Transformation* pada tahap ini dilakukan proses transformasi data terpilih kedalam bentuk mining procedure.
5. *Data Mining* pada tahap ini proses yang akan dilakukan adalah memproses data dengan berbagai teknik untuk mengekstrak pola-pola potensial menghasilkan data yang bermanfaat, dengan tahapan:
  - a. Read data xls dengan operator membaca file xls yang ditentukan.
  - b. Klustering, proses ini menggunakan algoritma K-means operator untuk pengelompokan data.
  - c. Performance adalah operator yang digunakan untuk evaluasi kinerja metode pengelompokan berbasis

centroid. Operator ini memberikan daftar nilai kriteria berdasarkan centroid klaster.

- d. Proses pengolahan data, proses ini untuk menentukan klaster.
- 6. Pattern Evolution pada tahap ini dilakukan suatu proses dimana mengidentifikasi pola berdasarkan measure yang diberikan, dengan evaluasi kluster menggunakan Davies Bouldin index.
- 7. Knowledge Presentation pada tahap ini dilakukan pemrosesan data yang digambarkan agar lebih mudah dipahami oleh user dan diharapkan bisa diambil tindakan berdasarkan analisis.

**2.2 Rapid Miner**

RapidMiner adalah sebuah perangkat lunak yang bersifat bebas dan bisa digunakan oleh siapa saja (open source), RapidMiner adalah sebuah penyelesaian dalam melakukan penganalisaan terhadap data mining, text mining dan analisis prediksi [5]. RapidMiner adalah sebuah solusi untuk melakukan analisis terhadap data mining, text mining dan analisis prediksi. RapidMiner menggunakan berbagai teknik deskriptif dan prediksi dalam memberikan wawasan kepada pengguna sehingga dapat membuat keputusan yang paling baik. RapidMiner didistribusikan di bawah lisensi dengan nama AGPL (GNU Affero General Public License) tipe 3. RapidMiner tersedia GUI (Graphic User Interface) untuk mendesain pipeline analisis. GUI akan mendapatkan hasil berupa file XML (Extensibel Markup Laguanje) dalam menentukan tahapan analisis yang ingin digunakan dan diterapkan [5].

**2.3 K-Means**

K-Means merupakan suatu metode penganalisaan data atau metode Data mining yang melakukan proses pemodelan tanpa supervise (unsupervised) dan merupakan salah satu metode yang melakukan pengelompokan data dengan sistem partisi [6]. Metode k-means merupakan metode clustering yang paling sederhana dan umum. Dikarenakan k-means mempunyai kemampuan mengelompokkan data dalam jumlah yang cukup besar dengan waktu komputasi yang cepat dan efisien. k-

means merupakan salah satu algoritma klasterisasi dengan metode partisi (partitioning method) yang berbasis titik pusat (centroid) selain algoritma k-medoids yang berbasis objek [6].

**3. HASIL DAN PEMBAHASAN**

Penelitian ini dilakukan untuk klasterisasi kecelakaan di ruas jalan tol sebanyak 384 data menggunakan metodologi KDD dengan beberapa tahapan mulai dari Pengumpulan Data, Data Selection, Data Preprocessing (Cleaning), Data Transformation, Data Mining, dan Evaluation (Interpretation).

Tahapan klasterisasi yang dilakukan dengan menentukan uji coba banyaknya klaster yaitu dengan menetapkan k= 3, k=5 dan k=7, dan performancenya menggunakan Davis Bouldin Index (DBI) yang didalam tools Rapidminer sudah ada parameter yang disediakan untuk DBI tersebut, hal tersebut dilakukan untuk menentukan model klaster terbaik yang diuji dan mengacu pada evaluasi performance DBI yang mendekati nilai Nol ( nilai Terbaik) , sehingga didapat pada nilai k=7 adalah nilai DBI terbaik yaitu 0.179 sedangkan untu k= 5 diperoleh DBI sebesar 0.180 dan k=3 nilai DBI sebesar 0.233.

**Tabel 1. Performance Dengan 3 Klaster**

Avg. within centroid	Davies Bouldin
Avg. within centroid distance: 0.359	Davies Bouldin: 0.233
Avg. within centroid distance_cluster_0: 0.308	
Avg. within centroid distance_cluster_1: 1.190	
Avg. within centroid distance_cluster_2: 0.349	

**Tabel 2. Performance Dengan 5 Klaster**

Avg. within centroid	Davies Bouldin
Avg. within centroid distance: 0.283	Davies Bouldin: 0.180
Avg. within centroid	

distance_cluster_0: 0.287	
Avg. within centroid distance_cluster_1: 0.533	
Avg. within centroid distance_cluster_2: 0.324	
Avg. within centroid distance_cluster_3: 0.395	
Avg. within centroid distance_cluster_4: 0.217	

Tabel 3. Performance Dengan 7 Klaster

Avg. within centroid	Davies Bouldin
Avg. within centroid distance: 0.195	Davies Bouldin: 0.179
Avg. within centroid distance_cluster_0: 0.172	
Avg. within centroid distance_cluster_1: 0.395	
Avg. within centroid distance_cluster_2: 0.217	
Avg. within centroid distance_cluster_3: 0.533	
Avg. within centroid distance_cluster_4: 0.223	
Avg. within centroid distance_cluster_5: 0.342	
Avg. within centroid distance_cluster_6: 0.173	

Dengan demikian klasterisasi dengan nilai k=7 adalah model klasterisasi terbaik yang bisa digunakan dan menghasilkan 42lister: Cluster 0: 111 items, Cluster 1: 9 items, Cluster 2: 70 items, Cluster 3: 3 items, Cluster 4: 16 items, Cluster 5: 10 items, Cluster 6: 165 items dan Total number of items: 384.

Tabel 4. Hasil Klaster

K3	K5	K7
Cluster Model	Cluster Model	Cluster Model
Cluster 0: 273 items	Cluster 0: 255 items	Cluster 0: 111 items
Cluster 1: 18 items	Cluster 1: 3 items	Cluster 1: 9 items
Cluster 2: 93 items	Cluster 2: 47 items	Cluster 2: 70 items
Total number of items: 384	Cluster 3: 9 items	Cluster 3: 3 items
	Cluster 4: 70 items	Cluster 4: 16 items
	Total number of items: 384	Cluster 5: 10 items
		Cluster 6: 165 items
		Total number of items: 384



Gambar 2. Visualisasi Klaster 7 dengan Diagram Scatter Plot

#### 4. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan kesimpulan yang didapatkan dari Klasterisasi tingkat kecelakaan pada ruas jalan tol sebanyak 7 klaster dengan dengan nilai k=7 adalah model klasterisasi terbaik yang bisa digunakan dan menghasilkan klaseter: Cluster 0: 111 items, Cluster 1: 9 items, Cluster 2: 70 items, Cluster 3: 3 items, Cluster 4: 16 items, Cluster 5: 10 items, Cluster 6: 165 items dan Total number of items: 384. Evaluasi yang digunakan menggunakan Davis Boulden index (DBI) dimana nilai yang mendekati Nol adalah

nilai terbaik yang bisa digunakan untuk proses klasterisasi, dalam hal ini adalah pada  $k=7$  dengan nilai DBI sebesar 0,179.

#### **DAFTAR PUSTAKA**

- [1] Rizkinaswara, L. (2020). Revolusi 4.0. *aptika.kominfo.go.id*.
- [2] Simamora B. “Analisis Multivariat Pemasaran”. Jakarta. *PT. Gramedia Pustaka Utama*.(2005)
- [3] Muningsih, E., Maryani, I., Handayani, V. R. “Penerapan Metode K-Means dan Optimisasi Jumlah Cluster Dengan Index Davies Bouldin Untuk Clustering Propinsi Berdasarkan Potensi Desa”. *Jurnal Sains dan Manajemen* 9.1 (2021): 95-100.
- [4] Turban, E., Delen, D., Sharda, R. “Business Intelligence, Analytics, and Data Science: A Managerial Perspective”. Harlow; Munich: Pearson Prentice Hall (2018).
- [5] Luluk Elvitaria, M. H. “Smk Analisis Kesehatan Abdurrab Menggunakan Algoritma”. *Jurnal Teknologi Dan Sistem Informasi Univrab* 2.2 (2017): 220-233.
- [6] Miyamoto, S., Suizu, D. “Fuzzy C-Means Clustering Using Kernel Functions in Support Vector Machines”. *Journal of Advance* (2003).