IMPLEMETASI K-MEANS CLUSTERING PADA RAPIDMINER UNTUK ANALISIS DAERAH RAWAN KECELAKAAN

Brilian Rahmat C.T.I.[*1], Agum Agidtama Gafar[2], Nurul Fajriani[3], Umar Ramdani[4], Fitria Rihin Uyun[5], Yuwanda Purnamasari P.[6], Natalis Ransi[7]
[*1,2,3,4,5,6,7]Jurusan Teknik Informatika, Universitas Halu Oleo – Kendari – Sulawesi Tenggara
email : [*1]**it.brilian@gmail.com**, [2]superagum@gmail.com, [3]nfajriani96@gmail.com, [4]umarramdhani24@gmail.com, [5]fitriauyuun@gmail.com, [6]yuwandapurnamasari@gmail.com, [7]natalis.ransi@uho.ac.id,

Abstrak

Kecelakaan lalu lintas kerap menjadi masalah utama dalam pemerintahan dan sosial karena dapat menyebabkan kerugian dari segi biaya dan keselamatan manusia.DataMining telah terbukti sebagai teknik yang dapat dipercaya untuk menganalisa data kecelakaan lalu lintas dan memberikan hasil yang produktif. Kebanyakan analisis data kecelakaan lalu lintas, hanya terfokus mengidentifikasi faktor-faktor yang mempengaruhi seberapa parah kecelakaan tersebut. Terkadang, kecelakaan terjadi lebih sering pada suatu lokasi tertentu. Analisis pada lokasi tersebut dapat membantu mengidentifikasi penyebab terjadinya kecelakaan yang membuat kecelakaan lalu lintas lebih sering terjadi di lokasi tersebut. Dari 2620 data kecelakaan yang tercatat di dalam basis data Resor Kendari, data tersebut diseleksi menjadi 500 data. Data tersebut kemudian dianalisis menggunakan algoritma K-Means Clustering dengan bantuan aplikasi RapidMiner Studio. Hasil analisis menunujukan frekuensi tingkat kecelakaan di tiap lokasi beserta waktu-waktu rawan yang berpotensi terjadi kasus kecelakaan.

Kata kunci— K-Means Clustering, Kecelakaan Lalu Lintas, RapidMiner.

1. PENDAHULUAN

Kecelakaan lalu lintas kerap menjadi masalah utama dalam pemerintahan dan sosial karena dapat menyebabkan kerugian dari segi biaya dan keselamatan manusia. Oleh [1] menyatakan bahwa setiap tahun di seluruh dunia, ada lebih dari 1,2 juta orang meninggal dan 50 orang terluka akibat kecelakaan.Sebuah studi yang dilakukan oleh [2] menyatakan bahwa penyebab utama kematian setelah kardiovascular adalah kecelakaan lalu lintas jalan.

Kecelakaan lalu lintas dipengaruhi beberapa faktor yang diakibatkan karena kondisi pengendara, karakteristik jalan, lingkungan dan cuaca. Sampai saat ini, daerah rawan kecelakaan makin meningkat yang mengakibatkan banyak korban. Untuk dapat mengelompokkan data dan memberikan *list* (daftar) daerah rawankecelakaan yang dapat dijadikan informasi bagi pengendara. Metode pengelompokkan *K-Means* digunakan untuk mengelompokkandata-data yang memiliki ciri yang sama dan mengelompokkannya kedalam sebuah klaster.

Oleh karena itu,dengan menggunakan salah satu metode data mining, kami dapat mengelompokkan daerah rawan kecelakaan berdasarkan metode K-Means Clustering.

2. METODE PENELITIAN

2.1. Dataset

Dataset yang digunakan bersumber dari data kecelakaan yang tercatat pada basis data Resor Kendari. Sebanyak 2620 kasus kecelakaan tercatat di dalam basis data tersebut. Data tersebut kemudian diseleksi menjadi 500 kasus kecelakaaan pada rentang tahun 2010-2011 sebagai sampel dari data. Dari 500 data tersebut, diperoleh171 lokasi kecelakaan yang berbeda.

Untuk atribut yang digunakan pada data ini dapat dilihat pada Tabel 1. Atribut-atribut tersebut dipilih berdasarkaan faktor-faktor

yang dianggap dapat mempengaruhi kecelakaan secara signifikan dan juga mencocokan dengan atribut-atribut yang digunakan pada data panggilan darurat ambulan tentang kecelakaan yang juga pernah digunakan untuk klasifikasi daerah rawan kecelakaan [3].

Tabel 1. Daftar Atribut yang Digunakan

No.	Nama Attribut	Nilai
1	Hari	1,....,7
2	Bulan	1,....,12
3	Jam	0,....,23
4	Lokasi (no.lokasi)	1,....,171
5	Karakteristik Jalan	lurus, tikungan, pertigaan, perempatan, tanjakan
6	Cuaca	cerah, mendung, hujan
7	Keadaan Jalan	beraspal, tidak beraspal, rusak
8	Keadaan Lalu Lintas	ramai, agak ramai, sedang, agak sepi, sepi
9	Kelurahan (no. kelurahan)	1,....,71
10	Kecamatan (no. kecamatan)	1,....,16
11	Lingkungan Sekitar	kawasan pemukiman, kawasan pertokoan (mall), pusat perbelanjaan (pasar), tempat hiburan, kawasan wisata, lain-lain
12	Daerah	kab/kota, propinsi, nasional,desa
13	Jenis Kecelakaan	depan-samping, depan-depan, depan-belakang, tabrak manusia, tungal, tabrak lari, samping-samping, beruntun, lain-lain
14	Kondisi Pengendara	batas kecepatan, tidak tertib, lengah, pengaruh alkohol, lelah, mengantuk, sakit

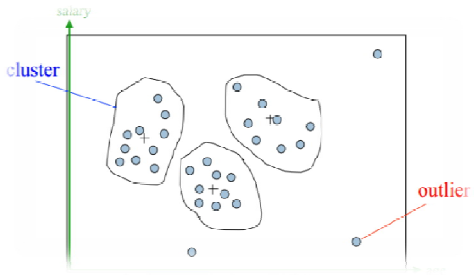
Tahapan dalam melakukan data mining salah satunya adalah preprosesing data yaitu data perlu di bersihkan sebelum diproses, hal ini terjadi karena biasanya data yang akan digunakan belum baik.

Teknik atau metode yang digunakan dalam data preprocessing, diantaranya:

1. Data cleaning

Data cleaning adalah menghilangkan nilai-nilai data yang salah, memperbaiki kekacauan data dan memeriksa data yang tidak konsisten. Adapun beberapa teknik membersihkan data, yakni mengisi missing value dan mengidentifikasi atau membuang outlier.

a. **Missing value** adalah informasi yang tidak tersedia untuk sebuah objek (kasus). Missing value terjadi karena informasi untuk sesuatu tentang objek tidak diberikan, sulit dicari, atau memang informasi tersebut tidak ada. Missing value pada dasarnya tidak bermasalah bagi keseluruhan data, apalagi jika jumlahnya hanya sedikit, misal hanya 1 % dari seluruh data. Namun jika persentase data yang hilang tersebut cukup besar, maka perlu dilakukan pengujian apakah data yang mengandung banyak missing tersebut masih layak diproses lebih lanjut ataukah tidak.
b. **Data outlier** (pencilan) adalah data yang secara nyata berbeda dengan data-data yang lain.



Gambar 1 Missing Value dan Data Outlier (*Sumber : (missing value & outlier) file.upi.edu/Direktori/FPIPS/.../Modul_Analisis_Missing_Value_%26_Outlier.pdf*)

2. Data integrasi

Menggabungkan data dari beberapa sumber (database, data cube, atau file) ke dalam penyimpanan data yang sesuai.

3. Data transformasi

Normalisasi dan pengumpulan data sehingga menjadi sama.

4. Data reduksi

Menguraikan data ke dalam bentuk yang lebih kecil ukurannya tetapi tetap menghasilkan hasil analitis yang sama.

5. Data diskretisasi

Bagian dari data reduksi tetapi memiliki arti penting tersendiri, terutama untuk data numerik.

2.2. K-Means Clustering

Metode *K-Means* adalah salah satu metode dalam fungsi c*lustering* atau pengelompokan. Menurut (Larose, 2005) c*lustering* mengacu pada pengelompokkan data, observasi atau kasus berdasar kemiripan objek yang diteliti. Sebuah *cluster* adalah suatu kumpulan data yang mirip dengan lainnya atau ketidakmiripan data pada kelompok lain. Sedangkan (Xu & Wunsch II, 2009) menjelaskan bahwa *clustering* adalah membagi objek data (bentuk, entitas, contoh, ketaatan, unit) ke dalam beberapa jumlah kelompok (grup, bagian atau kategori).

Du (2010) menjelaskan bahwa klasterisasi adalah proses membagi data yang tidak berlabel menjadi kelompok - kelompok data yang memiliki kemiripan. Misalkan K adalah jumlah klaster, C merupakan label klaster, dan P merupakan dataset. Klasterisasi harus memenuhi kriteria berdasarkan Persamaan (1), (2) dan (3).

$$C_i \neq \Phi, \forall i \in \{1,2,..., K\} \quad (1)$$

$$C_i \cap C_j = \Phi, \forall i \neq jandi, j \in \{1,2,..., K\} \quad (2)$$

$$\bigcup_{i=1}^{K} Ci = P \quad (3)$$

Algoritma K-Means merupakan algoritma klasterisasi yang mengelompokkan data berdasarkan titik pusat klaster (centroid) terdekat dengan data. Tujuan dari K-Means adalah pengelompokkan data dengan memaksimalkan kemiripan data dalam satu klaster dan meminimalkan kemiripan data antar klaster. Ukuran kemiripan yang digunakan dalam klaster adalah fungsi jarak. Sehingga pemaksimalan kemiripan data didapatkan berdasarkan jarak terpendek antara data terhadap titik centroid.

Tahapan awal yang dilakukan pada proses klasterisasi data dengan menggunakan algoritma K-Means adalah pembentukan titik awal centroid cj Pada umumnya pembentukan titik awal centroid dibangkitkan secara acak. Jumlah centroid cj yang dibangkitkan sesuai dengan jumlah klaster yang ditentukan di awal. Setelah k centroid terbentuk kemudian dihitung jarak tiap data xi dengan centroid ke-j sampai k dinotasikan dengan d(xi,cj). Terdapat beberapa ukuran jarak yang digunakan sebagai ukuran kemiripan suatu instance data, salah satunya adalah jarak Euclid. Perhitungan jarak Euclidean seperti pada Persamaan 4.

$$d(Xi, Cj) = \sqrt{\sum_{i=1}^{N} (Xi - Cj)^2} \quad (4)$$

Duran dan Odell (1974) menyatakan jika d(Xi,Cj) semakin kecil, kesamaan antara dua unit pengamatan semakin dekat. Syarat menggunakan jarak Euclid adalah jika semua fitur dalam dataset tidak saling berkorelasi. Jika terdapat fitur yang berkorelasi maka menggunakan konsep jarak Mahalanobis.

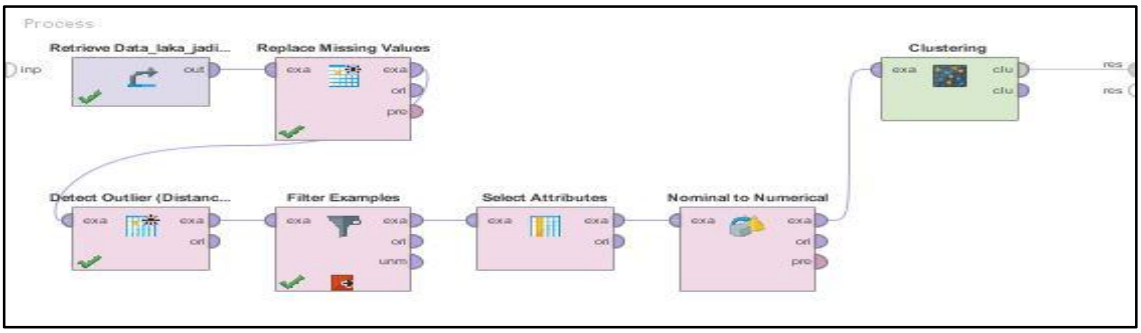
Agusta (2007) menyatakan kelanjutan dari jarak tersebut dicari yang terdekat sehingga data akanmengelompok berdasarkan centroid yang paling dekat. Tahap berikutnya adalah update titik centroid dengan menghitung rata-rata jarakseluruh data terhadap centroid. Selanjutnya akan kembali lagi ke proses awal. Iterasi iniakan diulangi terus sampai didapatkan centroidyang konstan artinya titik centroid sudah tidakberubah lagi. Atau iterasi dihentikan berdasarkan jumlah iterasi maksimal yang ditentukan.

3. HASIL DAN PEMBAHASAN

RapidMiner merupakan software/perangkat lunak untuk pengolahan data. Dengan menggunakan prinsip dan algoritma data mining, RapidMiner mengekstrak pola-pola dari data set yang besar dengan mengkombinasikan metode statistika, kecerdasan buatan dan database.

RapidMiner memudahkan penggunanya dalam melakukan perhitungan data yang sangat banyak dengan menggunakan operator-operator. Operator ini berfungsi untuk memodifikasi data. Data dihubungkan dengan node-node pada operator kemudian kita hanya tinggal menghubungkannya ke node hasil untuk melihat hasilnya. Hasil yang diperlihatkan RapidMiner pun dapat ditampilkan secara visual dengan grafik. Menjadikan RapidMiner adalah salah satu software pilihan untuk melakukan ekstraksi data dengan metode-metode data mining.

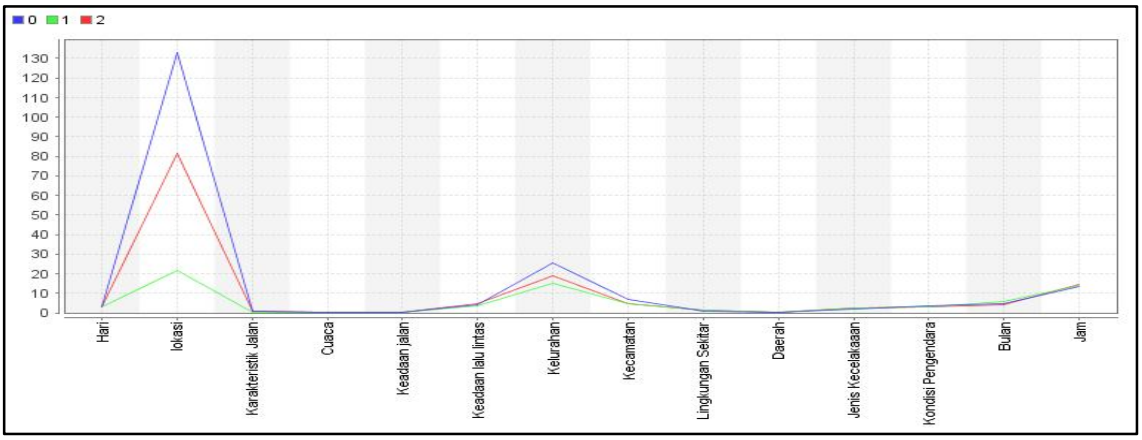
Pada contoh kasus analisis data kecelakaan, kami menggunakan konfigurasi data dan operator seperti yang dapat terlihat pada Gambar 2.



Gambar 2. Konfigurasi RapidMiner

Data awalnya dilakukan preproses terlebih dahulu yaitu dengan mengganti atribut-atribut yang memiliki nilai kosong dengan mengambil rata-rata nilai dari atribut tersebut. Setelah itu data dilakukan proses penghilangan outlier. Ditentukan 10 outlier yang dieksklusikan dari cluster/kelompok data yang lain. Kemudian data diseleksi dengan mengambil data yang bukan outlier yang berpotensi masuk ke dalam cluster tertentu. Sebelum data benar-benar di-cluster, terlebih dahulu, dilakukan proses diskritisasi data dengan mengubah data nominal menjadi numerik dikarenakan dalam proses clustering, data yang diproses haruslah data numerik. Setelah itu data kemudian dapat benar-benar di-cluster. Pada proses pengelompokkan itu sendiri, kami menentukan 3 titik centroid secara random/acak.

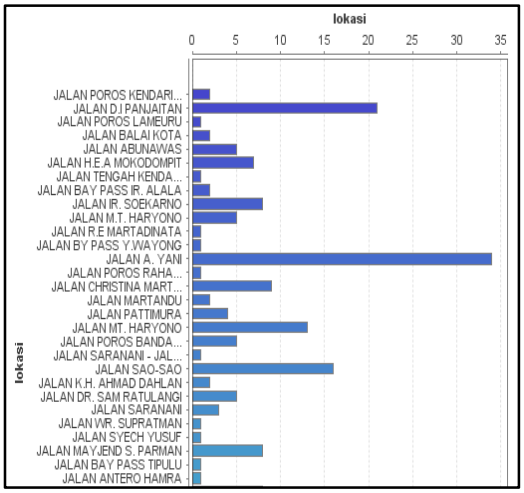
Dengan konfigurasi operator demikian, diketahui bahwa hasil klasterisasi data menunjukkan 490 data yang dikelompokkan ke 3 cluster. Cluster pertama berjumlah 82 kasus, cluster kedua berjumlah 295 kasus dan cluster ketiga berjumlah 113 kasus. Hubungan tiap kasus dengan titik centroid dari tiap cluster dapat dilihat pada Gambar 3. Gambar 4 berisi grafik yang menunjukan hubungan centroid dari tiap cluster.



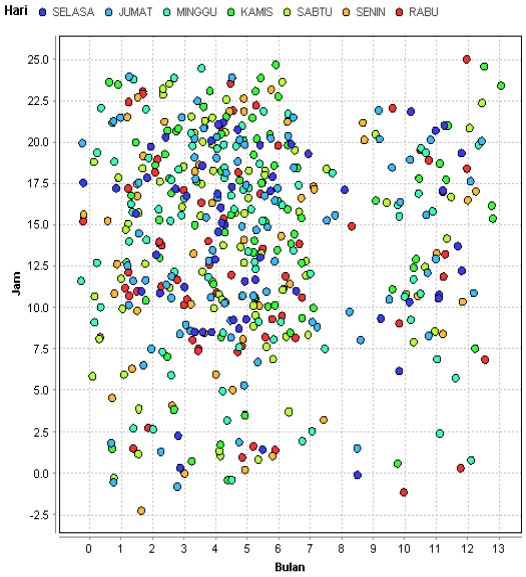
Gambar 1. Hubungan Titik Centroid dari Tiap Cluster

Analisis daerah rawan kecelakaan menggunakan RapidMiner dapat dilakukan dengan mudah. Kita juga dapat mengetahui jumlah kasus kecelakaan yang terjadi di tiap lokasi yang berbeda. Untuk itu, data tersebut dapat dilihat pada gambar 3 yang menunjukan grafik frekuensi kecelakaan tiap lokasi. Dari total 171 lokasi yang ada, diketahui bahwa Jl. A.Yani memiliki tingkat frekuensi kecelakaan mencapai 34 kasus.

Tidak hanya itu, kami juga menganalisis pengaruh dari atribut-atribut temporal (dalam hal ini bulan, hari, dan jam). Hubungan atribut tersebut juga dapat dilihat pada Gambar 5 dimana gambar tersebut berisi grafik hubungan antara bulan, hari dan jam terjadinya kecelakaaan.



Gambar 4. Frekuensi Lokasi Kejadian



Gambar 5. Hubungan antara Bulan, Jam, dan Hari Kecelakaan

Dari grafik tersebut, dapat diketahui bahwa kecelakan di kota Kendari sering terjadi pada bulan Januari hingga bulan Juli pada hari-hari kerja seperti hari Senin-Jum'at dari jam 10 pagi hingga jam 10 malam.

4. KESIMPULAN

Dari hasil penelitian ini, dapat disimpulkan bahwa analisis data kecelakaan menggunakan aplikasi RapidMiner dapat mengekstraksi beberapa informasi yang dibutuhkan untuk mengelompokkan data kecelakaan menjadi 3 buah kelompok/cluster dari 500 contoh data kecelakaan. Hasil ekstraksi data juga menunjukan tingkat frekuensi kecelakaan pada tiap lokasi kejadian. Serta mengekstraksi hubungan antara bulan, hari, dan waktu terjadinya kecelakaan lalu lintas di kota Kendari.

5. SARAN

Penelitian ini masih memiliki banyak kekurangan. Diharapkan kepada para peneliti yang lain untuk dapat menggunakan penelitian ini sebagai bahan ilmiah untuk melanjutkan analisis lokasi rawan kecelakaan

DAFTAR PUSTAKA

[1] Beshah, T. dan S. Hill. *Mining road traffic accident data to improve safety: Role of road-related factors on accident severity in Ethiopia*. 2010.

[2] Bener and D. Crundall, "Road traffic accidents in the United Arab Emirates compared to Western countries," *Advances in Transportation Studies in an international Journal Section A 6*, 2005.

[3] K. Sachin dan Durga Toshniwal. *A data mining approach to characterize road accident locations*. 2016.