ISSN : 2654-766X

Penerapan Algoritma *K-Nearest Neighbor* dalam Klasifikasi Tingkat Keparahan Korban Kecelakaan Lalu Lintas di Kabupaten Jawa Tengah

Application of K-Nearest Neihbor Algorithm in the Classification of Severity of Traffic Victims in Pati, Central Java

Dwi Selvy Wisdayani [1], Indah Manfaati Nur [2], Rochdi Wasono[3]
Universitas Muhammadiyah Semarang, Semarang
dwiselvy6288@gmail.com[1] indahmnur@unimus.ac.id[2] rochdi@unimus.ac.id[3]

Riwayat Artikel: Dikirim; Diterima; Diterbitkan

Abstrak

Kecelakaan lalu lintas merupakan masalah yang membutuhkan penanganan serius karena besarnya kerugian yang mengakibatkan korban manusia dan kerugian harta benda. Klasifikasi dapat diselesaikan dengan menggunakan teknik data mining. Untuk mengklasifikasikan tingkat keparahan kecelakaan lalu lintas, peneliti menerapkan algoritma *K-Nearest Neighbor*. *K-Nearest Neighbor* dipilih karena metode tersebut tangguh terhadap data noise dan *K-Nearest Neighbor* memiliki tingkat akurasi yang tinggi. Hasil klasifikasi dalam penelitian ini untuk mengetahui kinerja algoritma dalam memprediksi berdasarkan nilai akurasi, recall, error, precision dan f-measure. Hasil dari penelitian ini diperoleh K-Nearest Neighbor memiliki nilai akurasi sebesar 88.82 %, nilai recall sebesar 60.43 %, nilai error sebesar 11.18 %, nilai precision sebesar 64.37 % dan nilai f-measure sebesar 62.33 %. Sehingga agoritma *K-Nearest Neighbor* baik digunakan dalam klasifikasi tingkat keparahan kecelakaan lalu lintas di Pati Jawa Tengah.

Kata kunci: kecelakaan lalu lintas, klasifikasi, *K-Nearest Neighbor*

Abstract

Traffic accidents are a problem that requires serious treatment because of the large number of losses resulting in human casualties and property losses. The classification can be solved using data mining techniques. To classify the severity of a traffic accident, the researcher applies the K-Nearest Neighbor algorithm. K-Nearest Neighbor was chosen because the method is robust against data noise and K-Nearest Neighbor has a high degree of accuracy. The results of the classification in this study to determine the performance of the algorithm in predicting based on the value of accuracy, recall, error, precision and f-measure. The results of this study obtained K-Nearest Neighbor has an accuracy value of 88.82%, a recall value of 60.43%, an error value of 11.18%, a precision value of 64.37% and an f-measure value of 62.33%. So that the K-Nearest Neighbor agorithm is well used in the classification of the severity of traffic accidents in Pati, Central Java.

***Keywords**: traffic accidents, classification, K-Nearest Neighbor*

PENDAHULUAN

Kecelakaan lalu lintas menurut UU RI No. 22 Tahun 2009 adalah suatu peristiwa di jalan yang tidak diduga dan tidak disengaja melibatkan kendaraan dengan atau tanpa pengguna jalan lain yang mengakibatkan korban manusia dan kerugian harta benda (Wordpress, 2016). Provinsi Jawa Tengah merupakan salah satu provinsi yang memiliki jumlah kepadatan penduduk tertinggi di Indonesia, berdasarkan hasil Data Sensus Jumlah Penduduk Tahun 2010-2020 tersebut didapatkan jumlah penduduk sebesar 34.490.835 jiwa yang tersebar di 35 Kabupaten/Kota dengan di dominasi penduduk terbanyak di Kabupaten Brebes, Kabupaten Cilacap, dan Kota Semarang (Badan Pusat Statistik, 2017).

Kepolisian Republik Indonesia memiliki data-data kecelakaan lalu lintas hasil dari pencatatan setiap peristiwa kecelakaan yang terjadi. Data-data tersebut perlu dikelola dalam

ISSN : 2654-766X

suatu basis data untuk memudahkan proses penggalian informasi-informasi yang ada di dalamnya (Dewi, 2018). Berdasarkan data *Traffic Accidents, Victims and Loss in Region of Police of* Jawa Tengah, Kabupaten Pati memiliki tingkat kecelakaan lalu lintas yang tinggi, maka diperlukan sebuah penelitian tentang pola tingkat keparahan korban kecelakaan lalu lintas.

Data mining yang membuat data-data kecelakaan menjadi sumber untuk suatu model yang bisa digunakan untuk memprediksi suatu kejadian. Berdasarkan kebutuhan akan pencarian informasi tentang kecelakaan yang melibatkan beberapa kriteria yang tidak bisa ditentukan sebelumnya, maka penggunakan metode data mining merupakan solusi yang layak untuk diajukan (Yunanto, Hariadi, & Purnomo, 2012). Klasifikasi merupakan salah satu teknik dalam data mining. Klasifikasi adalah pemprosesan untuk menemukan sebuah model atau fungsi yang menjelaskan dan mencirikan konsep atau kelas data, untuk kepentingan tertentu. Penelitian – penelitian terdahulu mengenai *K-Nearest Neigbor* Pengaruh Nilai K pada Metode K-Nearest Neighbor (KNN) terhadap Tingkat Akurasi Identifikasi Kerusakan Jalan (A.Anggraini, 2018)

METODE

1. Statistika Deskriptif

Statistika deskriptif adalah metode-metode yang berkaitan dengan pengumpulan dan penyajian suatu gugus data sehingga memberikan informasi yang berguna. Dengan statistika deskriptif, kumpulan data yang diperoleh akan tersaji dengan rapi serta dapat memberikan informasi yang diperoleh dari analisis deskriptif antara lain ukuran pemusatan data, ukuran penyebaran data, serta kecenderungan suatu gugus data (Walpole, 1993).

2. Data Mining

Data mining merupakan suatu proses untuk menemukan informasi dari jumlah data yang besar (Zaki & Meira, 2014). Menurut *Gertner Group* Data Mining adalah suatu proses menemukan hubungan yang berarti, pola, dan kecenderungan dengan memeriksa dalam sekumpulan besar data yang tersimpan dalam penyimpanan dengan menggunakan teknik pengenalan pola seperti teknik statistik dan matematika" (Larose, 2006).

3. Klasifikasi

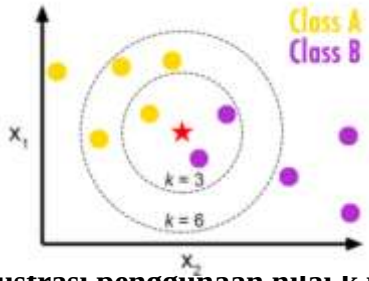
Klasifikasi adalah salah satu bagian dari data mining yang dapat digunakan untuk menggambarkan dan membedakan kelas data. (Farid et all, 2014). Klasifikasi merupakan salah satu teknik dalam data mining. Klasifikasi adalah pemprosesan untuk menemukan sebuah model atau fungsi yang menjelaskan dan mencirikan konsep atau kelas data, untuk kepentingan tertentu. Klasifikasi adalah tugas dasar dari analisis data yang berfungsi memberikan label kelas untuk kasus yang dijelaskan oleh satu set atribut.

4. Algoritma K-Nearest Neighbor

K-Nearest Neighbor merupakan salah satu algoritma pembelajara mesin sederhana. Hal ini hanya didasarkan pda gagasan bahwa suatu objek yang 'dekat' satu sama lain juga akan memiliki karakteristik yang mirip. Ini berarti jika kita mengetahui ciri-ciri dari dalah satu objek, maka kita juga dapat memprediksi objek lain berdasarkan tetangga terdekatnya. K-NN adalah improvisasi lanjutan dari teknik klasifikasi *Nearest Neighbor*. Hal ini didasarkan pada gagasan bahwa setiap contoh baru dapat diklasifikasikan oleh suara mayoritas dari k tetangga, di mana k adalah bilangan bulat positif, dan biasanya dengan jumlah kecil (Khamis et all, 2014). dibagi menjadi dua fase, yaitu pembelajaran (*training*) dan klasifikasi. Pada fase pembelajaran, algoritma ini hanya melakukan penyimpanan vektor-vektor fitur dan klasifikasi dari data pembelajaran. Pada fase klasifikasi, fitur-fitur yang sama dihitung untuk data yang akan diuji coba (yang klasifikasinya tidak diketahui). Jarak dari vektor yang baru ini terhadap seluruh vektor data pembelajaran dihitung, dan sejumlah nilai k

ISSN : 2654-766X

buah neighbor yang paling dekat diambil. Sebuah titik akan diprediksi jenisnya berdasarkan pada klasifikasi terbanyak dari neighbor di sekitarnya (Dewilde, 2012).



Gambar 1. Ilustrasi penggunaan nilai k pada metode KNN

Nilai k yang terbaik untuk KNN tergantung pada data. Secara umum, nilai k yang tinggi akan mengurangi efek noise pada klasifikasi, tetapi membuat batasan antara setiap klasifikasi menjadi lebih kabur. Nilai k yang bagus dapat dipilih optimasi parameter, misalnya dengan menggunakan cross-validation. Kasus khusus di mana klasifikasi diprediksikan berdasarkan data pembelajaran yang paling dekat (dengan kata lain, k = 1) disebut algoritma nearest neighbor.

Algoritma *K-Nearest Neighbor* adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. Untuk mendefinisikan jarak antara dua titik pada data *training* (x) dan titik pada data *testing* (y) maka digunakan rumus *Euclidean*, seperti yang ditunjukkan pada persamaan (2.6)

$$D(x,y) = \sqrt{\sum_{k}^{n} (x_k - y_k)^2}$$

Dengan D adalah jarak antara titik pada data *training* x dan titik data *testing* yang akan diklasifikasi, dimana x = x1, x2,…,xi dan y = y1, y2,…yi dan I merepresentasikan nilai atribut serta n merupakan dimensi atribut (M.J. Islam dkk , 2011).

5. Confusion Matrix

Confusion matrix adalah suatu metode yang biasanya digunakan untuk melakukan perhitungan akurasi pada konsep data mining. *Confusion matrix* digambarkan dengan tabel yang menyatakan jumlah data uji yang benar diklasifikasikan dan jumlah data uji yang salah diklasifikasikan.

Tabel 1 *Confusion Matrix*

		Clasified as	
Correct Clasification		+	-
	+	*True Positive*	*False Positif*
	-	*False Negative*	*True Negative*

True positive adalah jumlah *record* positif yang berhasil diklasifikasikan sebagai positif, sedangkan *false positive* merupakan *record* positif yang salah diklasifikasikan menjadi negatif. Sedangkan *false negative* merupakan *record* negatif yang salah diklasifikasikan sebagai *record* positif, dan untuk *true negative* adalah *record* negatif yang

ISSN : 2654-766X

berhasil diklasifikasikan sebagai *record* negatif. Metode pengujian *confusion matrix* dapat menghasilkan perhitungan dengan 5 output, diantaranya yaitu

$$\text{Akurasi} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}} \text{ x } 100\ \%$$

$$\text{Presisi} = \frac{\text{TP}}{\text{TP} + \text{FP}} \text{x } 100\ \%$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \text{x } 100\ \%$$

$$\text{F-Measure} = 2 \text{ x} \frac{\text{precission X recall}}{\text{precission} + \text{recall}}$$

$$Error = \frac{FN + FP}{TP + TN + FP + FN} \ x\ 100\ \%$$

6. Ukuran Performansi Klasifikasi

Ukuran performasi termasuk ke dalam tahapan evaluasi. Terdapat beberapa ukuran performansi untuk teknik klasifikasi yaitu recall, precision, F-Measure dan accuracy. Semakin tinggi tingkat *akurasi, precision, recall dan f-measure* maka algoritma yang dihasilkan dengan metode tersebut semakin baik dalam melakukan klasifikasi. Berdasarkan data yang didapat akan dihitung akurasi, precision,recall dan f-measure (Witten & Frank, 2006).

7. Cross-Validation

Cross-Validation merupakan metode statistik validasi silang yang dilakukan dengan melakukan evaluasi serta perbandingan. Metode ini dilakukan dengan cara membagi data menjadi dua segmen. Segmen pertama untuk melatih model yaitu data training sedangkan segmen kedua untuk memvalidasi model yaitu data uji atau data testing. *Cross Validation* yang populer kerjanya, dataset dibagi menjadi sejumlah K-buah partisi maka dilakukan sebanyak K-kali eksperimen. Pada masing-masing eksperimen, digunakan data partisi ke-K sebagai data testing dan partisi yang lain sebagai data training.

8. Kurva ROC/Curva ROC

Kurva ROC adalah salah satu teknik yang dapat memvisualisasikan, mengorganisasi dan memilih classifier berdasarkan performanya ((Vuk & Curk, 2006). *Recevier Operating Characteristic* (ROC) merupakan hasil dari pengukuran klasifikasi dalam bentuk 2 dimensi,dimana garis horizontal menggambarkan nilai false positif dan garis vertikal mengambarkan nilai true positive ((Vercellis, 2006). Kurva ROC dibagi dalam dua dimensi, dimana tingkat TP di plot pada sumbu Y dan tingkat FP di plot pada sumbu X. Tetapi untuk merepresentasikan grafis yang menentukan klasifikasi mana yang lebih baik, digunakan metode yang menghitung luas daerah dibawah kurva ROC yang disebut AUC (A*rea Under the ROC Curve*) yang diartikan sebagai probabilitas.

AUC mengukur kinerja diskriminatif dengan memperkirakan probabilitas output dari sampel yang dipilih secara acak dari populasi positif atau negatif, semakin besar AUC, semakin kuat klasifikasi yang digunakan. Karena AUC adalah bagian dari daerah unit persegi, nilainya akan selalu antara 0,0 dan 1,0.

Pada penelitian ini, tabel kontingensi yang digunakan untuk menganalisis ROC yaitu tabel *Confusion Matrix* dua kelas. AUC sering digunakan untuk mengukur kualitas classifier ROC dilihat berdasarkan akurasi dengan rentang yang diperlihatkan pada Tabel 2 (Gorunescu, 2011).

Tabel 2. Nilai Kualitas Classifier

Rentang Akurasi	Klasifikasi

ISSN : 2654-766X

0.90 – 1.00	*Excellent*
0.80 – 0.90	*Good*
0.70 – 0.80	*Fair*
0.60 – 0.70	*Poor*
0.50 – 0.60	*Failure*

METODE PENELITIAN

1. Sumber Data

Data yang digunakan dalam penelitian ini adalah data sekunder yang diperolehdari kantor SATLANTAS jumlah kecelakaan lalu lintas yang terjadi di Kabupaten Pati pada thun 2017.

2. Variabel Penelitian

Variabel yang digunakan dalam penelitian ini ada dua variabel yaitu variabel respon/label dan variabel predictor/atribut. Variabel respon dalam penelitian ini adalah tingkat keparahan korban, sedangkan variabel prediktor dalam penelitian ini adalah tanggal kejadian, jenis kejadian, jenis pekerjaan, jenis kecelakaan, jenis kelamin, usia, jenis kendaraan, pendidikan, peran korban, faktor korban, faktor manusia, alat keselamatan.

3. Analisis Data

Langkah-langkah metode K-Nearest Neighbor adalah sebagai berikut :

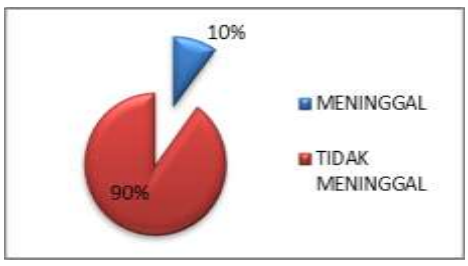
1. Menyiapkan data set
2. Menentukan parameter K (jumlah tetangga paling dekat)
   a. Pada penelitian ini nilai K telah ditentukan yaitu 3
3. Menghitung kuadrat jarak *euclid* (*euclidean distance*) masing-masing objek terhadap data sampel yang diberikan :

$$D(x,y) = \sqrt{\sum_{k}^{n}(x_k - y_k)^2}$$

4. Mengurutkan objek-objek tersebut ke dalam kelompok yang mempunyai jarak euclidean terkecil
5. Mengumpulkan kategori Y (klasifikasi *nearest neighbor*)
6. Dengan menggunakan kategori mayoritas,maka dapat hasil klasifikasi

HASIL DAN PEMBAHASAN

1. Analisis Deskriptif



Gambar 2. Persentase Tingkat Keparahan Korban Kecelakaan Lalu lintas

Langkah-langkah algoritma K-NN yaitu sebagai berikut :

1. Menentukan parameter K (jumlah tetangga paling dekat)

Pada penelitian ini niai k adalah 3. Dipilihnya nilai k berdasarkan ketentuan apabila klasifikasi genap maka ambil nilai k ganjil dan niai k >1.

2. Menghitung kuadrat jarak *euclid (euclidean distance*) masing-masing objek terhadap data sampel yang diberikan X1 sampai X11 berturut-turut : 1,0,2,3,4,0,1,0,2,1,3

Dari hasi perhitungan data testing dikurangi data training maka di dapatkan hasil jarak euclidean sebagai berikut :

Tabel 3 Jarak *Euclidean*

Data		Data		Data	
1	5.13	**15**	5.13	**29**	5.13
2	6.01	**16**	6.01	**30**	6.01
3	5.95	**17**	5.95	**31**	5.95
4	5.80	**18**	5.80	**32**	5.80
5	5.80	**19**	5.80	**33**	5.80
6	6.10	**20**	6.10	**34**	6.10
7	5.87	**21**	5.87	**35**	5.87
8	5.53	**22**	5.53	**36**	5.53
9	6.15	**23**	6.15	**37**	6.15
10	6.01	**24**	6.01	**38**	6.01
11	5.96	**25**	5.96	**.**	.
12	6.12	**26**	6.12	**.**	.
13	6.10	**27**	6.10	**.**	**.**
14	6.12	**28**	6.12	**241**	6.12

Tabel 4*Confusion Matrix* KNN

	kenyataan meninggal	kenyataan tidak meninggal	Class precision
prediksi meninggal	6	9	40.00 %
prediksi tidak meninggal	18	208	92.04 %
Class recall	25.00 %	98.62 %	

Berdasarkan tabel 4.7 Perhitungan akurasi data training 241 data, 6 fiklasifikasi prediksi meninggal dan ternyata meninggal. Sebanyak 18 diprediksi tidak meninggal tetapi ternyata meninggal dan sebanyak 208 diprediksi sesuai tidak meninggal. Dari TP< TN< FP< FN diatas juga dilakukan perhitungan *confussion matrix*, diperoleh hasil akurasi model *K-Nearest Neighbor* sebesar 88.82 %, nilai *recall* sebesar 60.43 %, nilai *precision* sebesar 64.40 %, nilai *error* sebesar 11.18 % dan *f-measure* 62.33 %.

ISSN : 2654-766X

KESIMPULAN

1. Tingkat keparahan kecelakaan lalu lintas di Kabupaten Pati Jawa Tengah memiliki arakteristik secara umum yaitu komposisi korban kecelakaan terluka didominasi oleh laki-laki kategori remaja dan dewasa baik SLTA hingga Wiraswasta sebagai pengendara. Kecelakaan terjadi pada kisaran (06.00-12.00 WIB) di waktu kejadian harian, terbilang ketika mengendarai SPM sering tidak menggunakan alat keselamatan dan lengah mengakibatkan kecelakaan depan-samping.
2. Diketahui bahwa kinerja sistem berdasarkan data sampel yang digunakan menghasilkan data prediksi yang besar dibanding dengan yang tidak sesuai menghasilkan akurasi 88.82 %, precision 64.40 %, nilai error 11.18 %, recall 60.43 %dan f-measure 62.33 % sehingga dapat dikatakan bahwa *K-Nearest Neighbor* baik dalam mengkalsifikasikan tingkat keparahan kecelakaan lalu lintas.

DAFTAR PUSTAKA

A.Anggraini, I. A. (2018). Pengaruh Nilai K pada Metode K-Nearest Neighbor (KNN) terhadap Tingkat Akurasi Identifikasi Kerusakan Jalan. *Rekayasa Sipil, Vol.7 No.2.*, 63-69.

Badan Pusat Statistik. (2017). *Data Sensus Jumlah Penduduk Pertengahan Tahun Hasil Proyeksi Penduduk.*

Dewilde, B. (2012). Classification of Hand-written Digits.

Farid et all. (2014). Hybrid decision tree and naive bayes classifiers for multi-class classification tasks. *ELSEVIER*, 1937-1946.

Gorunescu, F. (2011). *Data Mining : Concepts, models and techniques.* Verlag Berlin Heidelberg: Springer.

Khamis et all, H. S. (2014). Application of K-Nearest Neighbor Classification in Medical Data Mining. *JICT*, 121-128.

Larose, D. (2006). Data Mining Methods and Models. *Spring*, Vol 131.

Vercellis, C. (2006). *Business Intelegence: Data mining and Optimization for Decision Making.* Milano, Italy: Wiley.

Vuk, M., & Curk, T. (2006). ROC Curve, Lift Chart and Calibration Plot. *Metodoloski zvezki, Vil.3 No.1*, 89-108.

ISSN : 2654-766X

Witten, I., & Frank, E. (2006). *Review of " Data Mining: Practical Machine Learning Tools and Techniques" by Witten and Frank.* Francisco Azuaje: BioMedCentral.

Wordpress. (2016). *Pengertian kecelakaan lalu lintas.* Jateng: https://lakarestadps.wordpress.com.

Yunanto, W., Hariadi, M., & Purnomo, H. M. (2012). Pemetaan Kecelakaan Lalu Lintas Berbasis Klasifikasi Naive Bayes dengan Parameter Infrastruktur Jalan. *Seminar on Intelligent Technology and its Applications (SITIA)*, 13.

Zaki, M. J., & Meira, W. (2014). *Data Mining and Analysis : Fundamental Concepts and ALgorithms.* New York: British Library.