

Klasifikasi Tingkat Keparahan Korban Kecelakaan Lalu Lintas Di Kota Samarinda Menggunakan Algoritma *K-Nearest Neighbor* dan *Naive Bayes*

Classification of The Severity Of Traffic Accident Victims In The City of Samarinda Uses The K-Nearest Neighbor and Naive Bayes Algorithms

Nabila Abda Salsabila^{1, a)}, Rito Goejantoro¹, dan Syaripuddin²

¹Laboratorium Statistika Komputasi, FMIPA Universitas Mulawarman

²Laboratorium Matematika Komputasi, FMIPA Universitas Mulawarman

^{a)}Corresponding author: abdasalsabilanabila@gmail.com

ABSTRACT

Classification is the process of evaluating data objects to be included in a particular class from a number of available classes. The K-Nearest Neighbor algorithm is one of the algorithms used to classify an object against a new object based on its K nearest neighbors. Naive Bayes is a classification of data using probability based on the Bayes theorem with strong independence assumptions. This study aims to compare the accuracy of the classification results on traffic accident victim data in Samarinda City using the K-Nearest Neighbor algorithm and the Naive Bayes algorithm. The data used is data on the severity of traffic accident victims in Samarinda City from 2020 to 2021 with death and non-death classes and uses 6 independent variables, namely age, gender, victim's role, victim's vehicle, road status, and condition weather. The measurement of accuracy in classifying the K-Nearest Neighbor algorithm and the Naive Bayes algorithm uses a classification performance matrix. Based on the results of the study, the accuracy of the classification results of the K-Nearest Neighbor algorithm was obtained at 75.86%, while the Naive Bayes algorithm obtained an accuracy rate of 79.31%. From the results of this analysis, it can be concluded that the Naive Bayes algorithm works better than the K-Nearest Neighbor algorithm in classifying the severity of traffic accident victims in Samarinda City.

Keywords : *classification, k-nearest neighbor, naive Bayes, the severity of traffic accident victims*

1. Pendahuluan

Data dengan volume yang sangat besar hanya akan menjadi sampah di memori jika tidak diubah menjadi sebuah informasi. Teknik untuk mengubah sejumlah besar data menjadi informasi atau pengetahuan dikenal sebagai *data mining* (Larose & Larose, 2014). *Data mining* dibagi menjadi beberapa kelompok berdasarkan fungsi yang dapat dilakukan yaitu deskripsi, estimasi, prediksi, klustering, asosiasi, dan klasifikasi (Larose & Larose, 2014).

Salah satu teknik analisis yang menarik untuk dibahas adalah klasifikasi. Klasifikasi merupakan proses mengevaluasi objek data untuk dimasukkan ke dalam kelas tertentu dari sejumlah kelas yang tersedia. Dalam klasifikasi terdapat proses pelatihan yang harus dilakukan yaitu membangun model untuk mengklasifikasikan data, sehingga diketahui kelas mana yang dimiliki objek data akan dimasukkan sesuai dengan model yang tersimpan di memori. Proses pelatihan untuk membangun model tersebut diperlukan suatu algoritma. Terdapat berbagai algoritma pelatihan yang sudah dikembangkan, seperti *K-Nearest Neighbor*, *Artificial Neural Network* (ANN), *Naive Bayes*, *Support Vector Machine* (SVM), dan sebagainya (Prasetyo, 2012).

Algoritma *K-Nearest Neighbor* adalah salah satu algoritma paling populer serta memiliki keuntungan yaitu dapat diproses dengan mudah dan sederhana (Harrington, 2012). Algoritma *K-Nearest Neighbor* merupakan algoritma yang melakukan pengklasifikasian dengan cara menghitung kedekatan jarak suatu data dengan data lain (Prasetyo, 2012). Sedangkan algoritma *Naive Bayes* merupakan suatu pengklasifikasian data menggunakan probabilitas berdasarkan teorema Bayes yang ditemukan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi peluang di masa mendatang berdasarkan pengalaman di masa sebelumnya. Teorema tersebut dikombinasikan dengan *naive* di mana kondisi antar variabel diasumsikan saling bebas (Nofriansyah dkk, 2016). Keunggulan dari algoritma *K-Nearest Neighbor* dan *Naive Bayes* dalam melakukan klasifikasi terkenal memiliki tingkat keakuratan yang baik (Novalia dkk, 2020).

Beberapa penelitian terdahulu mengenai algoritma *K-Nearest Neighbor* dan *Naive Bayes* dilakukan oleh Wisdayani dkk (2019) tentang penerapan algoritma *K-Nearest Neighbor* dalam klasifikasi tingkat keparahan korban kecelakaan lalu lintas di Kabupaten Jawa Tengah diperoleh nilai akurasi sebesar 88,82%. Syarli dan Muin (2016) telah melakukan penelitian tentang metode *Naive Bayes* untuk prediksi kelulusan (studi kasus: data mahasiswa baru perguruan tinggi), hasil dari penelitian ini mendapatkan persentase keakuratan yang

tinggi yaitu sebesar 94%. Selanjutnya, Novalia dkk (2020) telah melakukan penelitian tentang perbandingan metode klasifikasi *Naive Bayes* dan *K-Nearest Neighbor* (studi kasus: status kerja penduduk di Kabupaten Kutai Kartanegara Tahun 2018) didapatkan hasil akurasi yang tinggi dengan menggunakan metode klasifikasi *Naive Bayes* dan *K-Nearest Neighbor* masing-masing 90,08% dan 94,66%.

Algoritma klasifikasi *K-Nearest Neighbor* dan *Naive Bayes* dapat diaplikasikan dalam berbagai bidang, salah satunya adalah di bidang lalu lintas. Indonesia merupakan salah satu negara dengan tingkat persentase kecelakaan lalu lintas yang cukup tinggi. Data Kepolisian RI yang dikutip oleh Subandi dkk (2021) menyebutkan bahwa sepanjang tahun 2009 terjadi sedikitnya 57.726 kasus kecelakaan lalu lintas di jalan raya. Kecelakaan lalu lintas merupakan kejadian di jalan yang tidak disengaja melibatkan kendaraan dengan atau tanpa pengguna jalan lain sehingga mengakibatkan kerugian bagi korban kecelakaan lalu lintas.

Kota Samarinda merupakan ibukota Provinsi Kalimantan Timur, memiliki tingkat kecelakaan lalu lintas yang cukup tinggi. Badan Pusat Statistik Provinsi Kalimantan Timur mencatat bahwa selama tahun 2020, jumlah kecelakaan di Kota Samarinda menempati angka kecelakaan tertinggi dibandingkan dengan kabupaten atau kota lainnya di Provinsi Kalimantan Timur. Jumlah kecelakaan lalu lintas di Kota Samarinda pada tahun 2019 adalah sebanyak 109 kasus yang menyebabkan jatuhnya korban sebanyak 164 orang. Dari jumlah korban tersebut di antaranya yang meninggal dunia sebanyak 47 orang, luka berat sebanyak 49 orang dan luka ringan sebanyak 68 orang. Sedangkan pada tahun 2020 jumlah kecelakaan lalu lintas sebanyak 96 kasus menyebabkan jatuhnya korban sebanyak 147 orang.

Pengklasifikasian tingkat keparahan korban kecelakaan lalu lintas di Kota Samarinda dirasa perlu untuk dilakukan, sehingga dengan mengetahui hasil klasifikasi tingkat keparahan korban kecelakaan lalu lintas diharapkan dapat membantu dalam mengolah penyajian informasi dan menjadi acuan bagi berbagai pihak yang berkepentingan dalam membuat kebijakan terkait kecelakaan lalu lintas, serta dapat meminimalisasi jumlah kecelakaan dan dapat memberikan penanganan alternatif pada korban kecelakaan lalu lintas, khususnya di Kota Samarinda. Berdasarkan latar belakang tersebut, maka penulis tertarik mencoba mengaplikasikan algoritma *K-Nearest Neighbor* dan *Naive Bayes* serta membandingkan keoptimalan kedua algoritma tersebut dalam mengklasifikasikan tingkat keparahan korban kecelakaan lalu lintas di Kota Samarinda.

2. Tinjauan Pustaka

2.1 Data mining

Data mining adalah serangkaian proses untuk menggali informasi yang selama ini tidak diketahui secara manual dari suatu basis data. Informasi diperoleh dengan cara mengekstraksi dan mengenali pola penting atau menarik dari suatu data yang terdapat pada basis data (Vulandari, 2017).

Menurut Larose dan Larose (2014), *data mining* dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan yaitu:

- a. Deskripsi
Deskripsi dapat memberikan kemungkinan penjelasan untuk menggambarkan suatu pola atau kecenderungan yang terdapat dalam data.
- b. Estimasi
Langkah dalam estimasi hampir mirip dengan klasifikasi, kecuali variabel target yang digunakan lebih ke arah numerik dibandingkan kategori.
- c. Prediksi
Prediksi hampir sama dengan estimasi, nilai dari hasil estimasi ada di masa sekarang sedangkan hasil dari prediksi menunjukkan bahwa sesuatu mungkin terjadi di masa yang mendatang.
- d. Klasterisasi
Klasterisasi dilakukan dengan membagi keseluruhan data pengamatan menjadi kelompok-kelompok yang memiliki kemiripan.
- e. Asosiasi
Asosiasi dalam *data mining* bertujuan untuk mengidentifikasi kategori dalam variabel yang sering muncul bersamaan.
- f. Klasifikasi
Dalam melakukan klasifikasi peneliti menggunakan variabel target kelas. Beberapa teknik yang digunakan dalam klasifikasi dan estimasi dapat juga digunakan (dalam keadaan yang tepat) untuk melakukan prediksi.

2.2 Probabilitas Bersyarat

Probabilitas terjadinya kejadian A dengan syarat bahwa B adalah terjadinya atau akan terjadi disebut probabilitas bersyarat atau $P(A|B)$. Probabilitas bersyarat terjadinya B dengan syarat A telah terjadi, dapat dirumuskan sebagai berikut (Novilla dkk, 2019):

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \tag{1}$$

dengan:

$P(B|A)$ = probabilitas bersyarat B terjadi dengan syarat A telah terjadi

$P(A \cap B)$ = probabilitas awal B dengan petunjuk A terjadi secara simultan

$P(A)$ = probabilitas awal A terjadi tanpa memandang kelas apapun.

2.3 Klasifikasi

Klasifikasi merupakan suatu pekerjaan menilai objek data untuk memasukkannya ke dalam kelas tertentu dari sejumlah kelas yang tersedia. Dalam klasifikasi terdapat dua proses utama yang dilakukan, yaitu pembangunan model sebagai prototipe untuk disimpan sebagai memori dan penggunaan model tersebut untuk melakukan pengenalan atau klasifikasi atau prediksi pada suatu objek data lain agar diketahui di kelas mana objek data tersebut dalam model yang sudah disimpannya (Prasetyo, 2012).

2.4 Algoritma K-Nearest Neighbor

Algoritma *K-Nearest Neighbor* merupakan salah satu metode dalam *data mining* yang paling sering digunakan, metode ini sangat sederhana dan mudah untuk diimplementasikan. Pengklasifikasian dilakukan berdasarkan kemiripan suatu data dengan data yang lain. Klasifikasi dilakukan terhadap objek berdasarkan data pembelajaran yang jaraknya dekat dengan objek tersebut.

Pendefinisian jarak antara objek dengan tetangganya dapat dihitung dengan menggunakan jarak Euclid, seperti ditunjukkan persamaan berikut:

$$d(x_{ik}, y_{jk}) = \sqrt{\sum_{k=1}^r (x_{ik} - y_{jk})^2} \tag{2}$$

dengan $d(x_{ik}, y_{jk})$ adalah jarak Euclid data *training* ke- i dengan data *testing* ke- j , x_{ik} adalah nilai variabel bebas ke- k dari data *training* ke- i , $i = 1, 2, \dots, p$, y_{jk} adalah nilai variabel bebas ke- k dari data *testing* ke- j , $j = 1, 2, \dots, q$, dan r adalah banyaknya variabel bebas.

Semakin dekat jarak kedua data maka semakin besar kemiripannya dan sebaliknya semakin jauh jarak kedua data tersebut maka semakin kecil kemiripannya. Dalam hal ini jumlah tetangga terdekat dinyatakan dengan K yang dilibatkan dalam penentuan prediksi label kelas pada data *testing*. Dilakukan *voting* kelas berdasarkan K tetangga terdekat yang telah terpilih. Kelas yang memiliki suara tetangga terbanyak yang akan diberikan label kelas hasil prediksi pada data *testing* tersebut (Prasetyo, 2014).

2.5 Algoritma Naive Bayes

Algoritma *Naive Bayes* merupakan salah satu teknik yang sering digunakan dalam *data mining*. Klasifikasi data menggunakan probabilitas berdasarkan teorema Bayes yang ditemukan oleh ilmuwan Inggris Thomas Bayes, yaitu dapat memprediksi peluang di masa mendatang berdasarkan kejadian sebelumnya. Teorema tersebut dikombinasikan dengan asumsi independensi yang sangat tinggi (*naive*) di mana kondisi antar variabel diasumsikan saling bebas. Klasifikasi *Naive Bayes* diasumsikan bahwa ada tidaknya ciri tertentu dari sebuah kelas tidak terdapat hubungan dengan ciri dari kelas lainnya (Nofriansyah dkk, 2016).

Menurut Prasetyo (2012), secara umum persamaan dari teorema Bayes adalah sebagai berikut.

$$P(C|F) = \frac{P(C)P(F|C)}{P(F)} \tag{3}$$

dengan:

$P(C|F)$: probabilitas terjadinya C dengan syarat F telah terjadi

$P(C)$: probabilitas terjadinya C

$P(F|C)$: probabilitas terjadinya F dengan syarat C telah terjadi

$P(F)$: probabilitas terjadinya F

Penjelasan teorema *Naive Bayes*, perlu diketahui bahwa proses klasifikasi memerlukan sejumlah petunjuk (variabel) untuk menentukan kelas yang sesuai bagi objek yang dianalisis tersebut. Oleh karena itu, teorema Bayes di atas disesuaikan sebagai berikut:

$$P(C|F_1, F_2, \dots, F_n) = \frac{P(C)P(F_1, F_2, \dots, F_n|C)}{P(F_1, F_2, \dots, F_n)} \tag{4}$$

dengan variabel C mempresentasikan kelas, sementara variabel F_1, F_2, \dots, F_n mempresentasikan sejumlah variabel yang dibutuhkan untuk melakukan pengklasifikasian. Rumus tersebut dapat menjelaskan bahwa peluang masuknya objek dengan variabel tertentu dalam kelas C (*posterior*) adalah peluang kemunculan kelas C sebelum masuknya objek (*prior*), dikali dengan peluang munculnya variabel-variabel objek yang

masuk kelas C (*likelihood*), dibagi dengan peluang kemunculan variabel-variabel pada objek secara umum (*evidence*). Oleh karena itu, rumus di atas dapat juga ditulis secara sederhana sebagai berikut:

$$Posterior = \frac{prior \times likelihood}{evidence} \quad (5)$$

nilai *posterior* tersebut nantinya akan dibandingkan dengan nilai *posterior* dari kelas lainnya untuk menentukan kelas suatu objek yang akan diklasifikasikan dengan memilih kelas yang memiliki *posterior* terbesar. Nilai *evidence* selalu tetap dan merupakan pembagi pada setiap kelasnya sehingga dalam perhitungan *posterior* cukup dengan mengalikan nilai *prior* dengan *likelihood*. Rumus Bayes tersebut dapat dijabarkan lebih lanjut dengan menjabarkan $P(C | F_1, F_2, \dots, F_n)$ menggunakan aturan perkalian sebagai berikut:

$$\begin{aligned} P(C | F_1, F_2, \dots, F_n) &= P(C)P(F_1, F_2, \dots, F_n | C) \\ &= P(C)P(F_1 | C)P(F_2, \dots, F_n | C, F_1) \\ &= P(C)P(F_1 | C)P(F_2 | C, F_1)P(F_3 | C, F_1, F_2) \\ &\quad \dots P(F_n | C, F_1, F_2, F_3, \dots, F_{n-1}) \end{aligned} \quad (6)$$

Terlihat bahwa hasil penjabaran tersebut menyebabkan faktor-faktor syarat yang mempengaruhi nilai probabilitas semakin banyak dan semakin kompleks, yang hampir mustahil untuk dianalisis satu persatu. Akibatnya, perhitungan tersebut sulit untuk dilakukan. Asumsi *naive* digunakan bahwa masing-masing variabel (F_1, F_2, \dots, F_n) saling bebas (*independent*) satu sama lain. Dengan asumsi tersebut maka berlaku suatu kesamaan sebagai berikut:

$$P(F_i | F_j) = \frac{P(F_i \cap F_j)}{P(F_j)} = \frac{P(F_i)P(F_j)}{P(F_j)} = P(F_i), \text{ untuk } i \neq j \quad (7)$$

sehingga diperoleh,

$$P(F_i | C, F_j) = P(F_i | C) \quad (8)$$

Persamaan (8) dapat diambil kesimpulan bahwa asumsi independensi *naive* tersebut menjadikan syarat peluang menjadi lebih sederhana, sehingga perhitungan menjadi mungkin untuk dapat dilakukan. Selanjutnya, $P(C | F_1, F_2, \dots, F_n)$ dapat dijabarkan secara sederhana menjadi,

$$\begin{aligned} P(C | F_1, F_2, \dots, F_n) &= P(C)P(F_1 | C)P(F_2 | C)P(F_3 | C) \dots P(F_n | C) \\ &= P(C) \prod_{i=1}^n P(F_i | C) \end{aligned} \quad (9)$$

persamaan (9) merupakan model dari teorema Bayes yang selanjutnya akan digunakan dalam proses klasifikasi. Menurut Saleh (2015), algoritma *Naive Bayes* mempunyai alur sebagai berikut:

- a. Membaca data *training*
- b. Menghitung nilai probabilitas setiap variabel pada setiap kelasnya
- c. Menentukan probabilitas akhir untuk data yang ingin diuji, yaitu sebagai berikut:
 - i. Menghitung probabilitas akhir untuk setiap kelas. Persamaan yang digunakan sebagai berikut:

$$\prod_{i=1}^n P(F_i | C) \quad (10)$$

- ii. Nilai dari persamaan (10) dimasukkan untuk mendapatkan probabilitas akhir. Persamaan yang digunakan berdasarkan Persamaan (9).

2.6 Data Training dan Data Testing

Menurut Prasetyo (2014), pengklasifikasian membagi data menjadi data *training* dan data *testing*. Data atau vektor yang telah diketahui sebelumnya label kelas dan digunakan untuk membangun model *classifier* disebut dengan data *training*. Sedangkan data *testing* merupakan data atau vektor yang belum diketahui (dianggap belum diketahui) label kelasnya menggunakan model *classifier* yang sudah dibangun.

Jumlah dari data *training* dan data *testing* dapat dihitung menggunakan Persamaan (11) dan Persamaan (12).

$$\text{Jumlah data } training = \text{proporsi data } training \times n \quad (11)$$

$$\text{Jumlah data } testing = n - \text{jumlah data } training \quad (12)$$

dengan, n adalah jumlah seluruh data.

2.7 Pengukuran Tingkat Akurasi

Menurut Prasetyo (2012), sebuah sistem yang melakukan klasifikasi diharapkan dapat melakukan klasifikasi semua *set* data dengan benar. Akan tetapi, kinerja suatu sistem tidak bisa bekerja 100% benar. Oleh karena itu, sebuah sistem klasifikasi kinerjanya juga harus diukur. Umumnya, cara mengukur kinerja klasifikasi menggunakan matriks konfusi (*confusion matrix*).

Matriks konfusi merupakan tabel hasil kerja klasifikasi. Tabel 1 merupakan contoh matriks konfusi yang melakukan klasifikasi dengan masalah biner (dua kelas), hanya terdapat dua kelas yaitu 0 dan 1. Setiap sel f_{ij} dalam matriks menyatakan jumlah data dari kelas i yang hasil prediksinya masuk ke kelas j . Misalnya, sel f_{11} adalah jumlah data dalam kelas 1 yang secara benar dipetakan masuk ke kelas 1, dan f_{10} adalah data dalam kelas 1 yang dipetakan secara salah ke kelas 0.

Tabel 1. Matriks Konfusi untuk Klasifikasi Dua Kelas

		Kelas hasil prediksi (j)	
		Kelas = 1	Kelas = 0
Kelas asli (i)	Kelas = 1	f_{11}	f_{10}
	Kelas = 0	f_{01}	f_{00}

Berdasarkan isi matriks konfusi di atas, dapat diketahui bahwa jumlah data dari masing-masing kelas yang diprediksi secara benar yaitu $(f_{11} + f_{00})$, dan data yang diklasifikasikan secara salah yaitu $(f_{10} + f_{01})$. Kuantitas matriks konfusi dapat diringkas menjadi dua nilai, yaitu akurasi dan laju *error*. Dengan mengetahui jumlah data yang diklasifikasikan secara benar maka dapat diketahui tingkat akurasi dari hasil prediksi yang dilakukan, dan mengetahui jumlah data yang diklasifikasikan secara salah maka dapat diketahui laju *error* dari hasil prediksi yang dilakukan. Dua kuantitas ini digunakan sebagai matriks kinerja klasifikasi.

Perhitungan nilai akurasi dapat digunakan rumus sebagai berikut (Prasetyo, 2012).

$$\begin{aligned}
 \text{Akurasi} &= \frac{\text{Jumlah data yang diprediksi secara benar}}{\text{Jumlah prediksi yang dilakukan}} \times 100\% \\
 &= \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}} \times 100\%
 \end{aligned}
 \tag{13}$$

Semua algoritma klasifikasi berusaha untuk membentuk model yang mempunyai akurasi yang tinggi. Model yang dibangun umumnya dapat memprediksikan dengan benar pada semua data yang menjadi data *training*-nya, tetapi ketika model berhadapan dengan data *testing* barulah kinerja model dari sebuah algoritma klasifikasi ditentukan (Prasetyo, 2014).

2.8 Kecelakaan Lalu Lintas

Menurut Peraturan Pemerintah (PP) Nomor 43 Tahun 1993 pada pasal 93 tentang lalu lintas dan angkutan jalan, kecelakaan lalu lintas adalah suatu peristiwa di jalan yang tidak diduga dan tidak sengaja melibatkan kendaraan dengan atau pengguna jalan lain yang mengakibatkan korban manusia ataupun kerugian harta benda. Kecelakaan lalu lintas disebabkan oleh banyak faktor, terdapat empat faktor utama penyebab terjadinya kecelakaan antara lain faktor manusia, faktor kendaraan, faktor jalan, dan faktor lingkungan (Fitriah dkk, 2012).

3. Hasil Penelitian dan Pembahasan

3.1 Data Penelitian

Data yang digunakan dalam penelitian ini merupakan data korban kecelakaan lalu lintas yang tercatat di Kepolisian Resor Kota (POLRESTA) Samarinda pada tahun 2020 sampai tahun 2021. Teknik analisis dalam mengklasifikasikan data menggunakan algoritma *K-Nearest Neighbor* dan *Naive Bayes*. Adapun variabel terikat adalah tingkat keparahan korban kecelakaan lalu lintas dan faktor penyebab terjadinya kecelakaan antara lain usia, jenis kelamin, peran korban, kendaraan korban, status jalan, dan kondisi cuaca.

3.2 Algoritma *K-Nearest Neighbor*

Pada proses klasifikasi algoritma *K-Nearest Neighbor* terdapat beberapa alur yang harus dilakukan yaitu menentukan nilai parameter K , menghitung jarak Euclid antara data *training* dan data *testing*, kemudian mengurutkan jarak dari hasil perhitungan. Kelas yang memiliki anggota terbanyak pada data *testing* merupakan hasil klasifikasi akhir yang digunakan untuk menentukan termasuk dalam kelas tidak meninggal dunia ataukah kelas meninggal dunia berdasarkan hasil prediksi klasifikasi. Data yang digunakan dalam perhitungan klasifikasi adalah menggunakan data *training* 90% sebanyak 262 data, sedangkan data *testing* 10% sebanyak 29 data, serta jumlah tetangga terdekat yang digunakan yaitu 1, 3, 5, 7, dan 9.

Pengukuran tingkat akurasi algoritma *K-Nearest Neighbor* dilakukan dengan menghitung jumlah data yang diklasifikasikan dengan benar, klasifikasi yang dihasilkan dapat dilihat pada Tabel 2.

Tabel 2. Hasil Klasifikasi Algoritma *K-Nearest Neighbor*

Klasifikasi Awal	Prediksi Klasifikasi 1-NN	
	MD	TMD
MD	8	5*
TMD	3*	13

Berdasarkan Tabel 2 pada algoritma *K-Nearest Neighbor* dapat diketahui bahwa tingkat keparahan korban kecelakaan lalu lintas Kota Samarinda menggunakan 1-NN diperoleh hasil yaitu dari 13 korban yang mengalami tingkat keparahan korban kecelakaan dengan kelas meninggal dunia, terdapat 8 korban tepat diklasifikasikan mengalami tingkat keparahan korban kecelakaan dengan kelas meninggal dunia dan sisanya 5 korban tidak tepat diklasifikasikan. Sedangkan dari 16 korban dengan kelas tidak meninggal dunia, terdapat 13 korban tepat diklasifikasikan mengalami tingkat keparahan korban kecelakaan dengan kelas tidak meninggal dunia dan 3 korban lainnya tidak tepat diklasifikasikan, sehingga diperoleh tingkat akurasi menggunakan algoritma *K-Nearest Neighbor* adalah sebesar 72,41%. Dengan cara yang sama, nilai akurasi dari hasil prediksi klasifikasi untuk setiap *K* yang lain yaitu 3, 5, 7, dan 9 juga dihitung.

Nilai akurasi yang dihasilkan untuk setiap nilai *K* pada algoritma *K-Nearest Neighbor* dapat dilihat pada Tabel 3.

Tabel 3. Perbandingan Nilai Akurasi Setiap Nilai *K*

K-NN	Akurasi
1-NN	72,41%
3-NN	75,86%
5-NN	75,86%
7-NN	72,41%
9-NN	72,41%

Berdasarkan Tabel 3 dapat dilihat nilai akurasi hasil prediksi kelas pada data *testing* menunjukkan nilai akurasi yang lebih tinggi berada pada 3-NN dan 5-NN, sehingga dapat disimpulkan bahwa klasifikasi tingkat keparahan korban kecelakaan lalu lintas dengan menggunakan algoritma *K-Nearest Neighbor* akan menghasilkan akurasi yang optimal apabila menggunakan *K* = 3 dan *K* = 5. Oleh karena itu, nilai akurasi 3-NN atau 5-NN akan dibandingkan dengan nilai akurasi yang dihasilkan oleh pengklasifikasian menggunakan algoritma *Naive Bayes*.

3.3 Algoritma *Naive Bayes*

Dalam proses klasifikasi algoritma *Naive Bayes* terdapat tiga alur yaitu membaca data *training*, menghitung nilai probabilitas setiap variabel pada setiap kelasnya dan menentukan probabilitas akhir. Hasil perhitungan probabilitas akhir digunakan untuk menentukan data termasuk dalam kelas meninggal dunia atau tidak meninggal dunia. Data yang digunakan pada proses perhitungan klasifikasi adalah data *training* dengan proporsi data yaitu untuk data *training* masing-masing 90% sebanyak 262 sampel, dan data *testing* sebanyak 29 sampel. Hasil pengklasifikasian yang diperoleh menggunakan algoritma *Naive Bayes* dapat dilihat pada Tabel 4.

Tabel 4. Hasil Klasifikasi Algoritma *Naive Bayes*

Klasifikasi Awal	Prediksi Klasifikasi	
	MD	TMD
MD	8	5*
TMD	1*	15

Berdasarkan Tabel 4, pada algoritma *Naive Bayes* dapat diketahui bahwa tingkat keparahan korban kecelakaan lalu lintas Kota Samarinda diperoleh hasil yaitu dari 13 korban yang mengalami tingkat keparahan korban kecelakaan dengan kelas meninggal dunia, terdapat 8 korban tepat diklasifikasikan mengalami tingkat keparahan korban kecelakaan dengan kelas meninggal dunia dan sisanya 5 korban tidak tepat diklasifikasikan. Sedangkan dari 16 korban dengan kelas tidak meninggal dunia, terdapat 15 korban tepat diklasifikasikan mengalami tingkat keparahan korban kecelakaan dengan kelas tidak meninggal dunia dan 1 korban tidak tepat diklasifikasikan, sehingga diperoleh tingkat akurasi menggunakan algoritma *Naive Bayes* adalah sebesar 79,31%.

3.4 Perbandingan Hasil Klasifikasi

Setelah diperoleh hasil pengukuran tingkat akurasi pada algoritma *K-Nearest Neighbor* dan algoritma *Naive Bayes*, kemudian nilai akurasi tersebut dibandingkan dan diambil nilai terbesar untuk

dijadikan sebagai algoritma yang lebih baik dalam melakukan pengklasifikasian data korban kecelakaan lalu lintas Kota Samarinda pada tahun 2020 sampai tahun 2021. Perbandingan nilai akurasi kedua algoritma dapat dilihat pada Tabel 5.

Tabel 5. Perbandingan Tingkat Akurasi

Algoritma	Akurasi
<i>K-Nearest Neighbor</i>	75,86%
<i>Naive Bayes</i>	79,31%

Berdasarkan Tabel 5 dapat diketahui bahwa pada algoritma *K-Nearest Neighbor* menunjukkan ketepatan pengklasifikasian dalam memprediksi klasifikasi sebesar 75,86%. Sedangkan, pada algoritma *Naive Bayes* menunjukkan ketepatan klasifikasi dalam memprediksi klasifikasi sebesar 79,31%. Hal ini menunjukkan bahwa algoritma *Naive Bayes* bekerja lebih baik dibandingkan algoritma *K-Nearest Neighbor* dalam mengklasifikasikan tingkat keparahan korban kecelakaan Kota Samarinda dilihat dari nilai akurasi yang lebih tinggi.

4. Kesimpulan

Pengklasifikasian tingkat keparahan korban kecelakaan lalu lintas Kota Samarinda pada tahun 2020 sampai tahun 2021 menggunakan algoritma *K-Nearest Neighbor* dengan menggunakan K optimal 3-NN dan 5-NN adalah dari 13 korban mengalami tingkat keparahan korban kecelakaan yang meninggal dunia, terdapat 7 korban tepat diklasifikasikan mengalami tingkat keparahan korban kecelakaan dengan status meninggal dunia dan sisanya 6 korban tidak tepat diklasifikasikan, sedangkan dari 16 korban yang tidak meninggal dunia, terdapat 15 korban tepat diklasifikasikan mengalami tingkat keparahan korban kecelakaan yang tidak meninggal dunia dan 1 korban tidak tepat diklasifikasikan.

Pengklasifikasian tingkat keparahan korban kecelakaan lalu lintas Kota Samarinda pada tahun 2020 sampai dengan tahun 2021 menggunakan algoritma *Naive Bayes* memberikan hasil klasifikasi yaitu dari 13 korban yang mengalami tingkat keparahan korban kecelakaan yang meninggal dunia, terdapat 8 korban tepat diklasifikasikan mengalami tingkat keparahan korban kecelakaan yang meninggal dunia dan sisanya 5 korban tidak tepat diklasifikasikan. Sedangkan dari 16 korban dengan kelas tidak meninggal dunia, terdapat 15 korban tepat diklasifikasikan mengalami tingkat keparahan korban kecelakaan yang tidak meninggal dunia dan 1 korban tidak tepat diklasifikasikan.

Hasil pengukuran tingkat akurasi klasifikasi tingkat keparahan korban kecelakaan lalu lintas Kota Samarinda pada tahun 2020 sampai tahun 2021 menggunakan algoritma *K-Nearest Neighbor* dengan data *training* 90% memperoleh hasil ketepatan pengklasifikasian dalam memprediksi klasifikasi sebesar 75,86%. Sedangkan, pada algoritma *Naive Bayes* memperoleh hasil ketepatan klasifikasi dalam memprediksi klasifikasi sebesar 79,31%. Hal ini menunjukkan bahwa algoritma *Naive Bayes* bekerja lebih baik dibandingkan algoritma *K-Nearest Neighbor* dalam mengklasifikasikan tingkat keparahan korban kecelakaan lalu lintas Kota Samarinda pada tahun 2020 sampai tahun 2021 dilihat dari nilai akurasi yang lebih tinggi.

Referensi

- BPS Provinsi Kalimantan Timur. (2021). *Jumlah Korban Kecelakaan Lalu Lintas, 2018-2020*. Samarinda: BPS Provinsi Kalimantan Timur.
- Fitriah, W. W., Mashuri, M., & Irhamah. (2012). Faktor-faktor yang mempengaruhi keparahan korban kecelakaan lalu lintas di Kota Surabaya dengan pendekatan Bagging regresi logistik ordinal. *Jurnal Sains dan Seni ITS*, 253-258.
- Harrington, P. (2012). *Machine Learning in Action*. Shelter Island: Manning Publications.Co.
- Larose, D. T., & Larose, C. D. (2014). *Discovering Knowledge in Data: An Introduction to Data Mining*. New Jersey: John Wiley & Sons, Inc.
- Nofriansyah, D., Erwansyah, K., & Ramadhan, M. (2016). Penerapan Data Mining dengan Algoritma Naive Bayes Classifier untuk Mengetahui Minat beli Pelanggan terhadap Kartu Internet XL (Studi Kasus di CV. Sumber Utama Telekomunikasi). *Jurnal Ilmiah Saindikom*, 81-92.
- Novalia, V., Goejantoro, R., & Sifriyani. (2020). Perbandingan Metode Klasifikasi Naive Bayes dan K-Nearest Neighbor (Studi Kasus: Status Kerja Penduduk di Kabupaten Kutai Kartanegara Tahun 2018). *Jurnal Eksponensial*, 159-165.
- Novilla, D. A., Goejantoro, R., & Amijaya, F. D. (2019). Klasifikasi Data Nasabah Asuransi Dengan Menggunakan Metode Naive Bayes (Studi kasus: PT. Prudential Life Jalan Mt. Haryono Samarinda). *Jurnal Eksponensial*, 95-101.
- Pemerintah Republik Indonesia. (1993). *Peraturan Pemerintah Nomor 43 Tahun 1993 Pasal 93 tentang Lalu Lintas dan Angkutan Jalan*. Diambil kembali dari Pemerintah Republik Indonesia.

- Prasetyo, E. (2012). *Data Mining : Konsep dan Aplikasi Menggunakan MATLAB*. Yogyakarta: CV. Andi Offset.
- Prasetyo, E. (2014). *Data Mining: Mengolah Data Menjadi Informasi Menggunakan Matlab*. Yogyakarta: CV. Andi Offset.
- Saleh, A. (2015). Implementasi Metode Klasifikasi Naive Bayes Dalam Memprediksi Besarnya Penggunaan Listrik Rumah Tangga. *Citec Journal*, 207-217.
- Subandi, A., Purnomo, T. W., & Aulia, S. M. (2021). Efektifitas Latihan Basic Life Support (BLS) pada Masyarakat Awam (Forum Anak Provinsi Jambi) untuk Penanganan Korban Kecelakaan Lalulintas di Kota Jambi. *JUPIIS*, 108-117.
- Syarli, & Muin, A. A. (2016). Metode Naive Bayes untuk Prediksi Kelulusan (Studi Kasus: Data Mahasiswa Baru Perguruan Tinggi). *Jurnal Ilmiah Ilmu Komputer*, 22-26.
- Vulandari, R. T. (2017). *Data Mining: Teori dan Aplikasi Rapidminer*. Yogyakarta: Gava Media.
- Wisdayani, D. S., Nur, I. M., & Wasono, R. (2019). Penerapan Algoritma K-Nearest Neighbor dalam Klasifikasi Tingkat Keparahan Korban Kecelakaan Lalu Lintas di Kabupaten Jawa Tengah. *Prosiding Unimus*, 373-380.