BAB II

TINJAUAN PUSTAKA

1.1 Statistika Deskriptif

Statistika deskriptif adalah metode-metode yang berkaitan dengan pengumpulan dan penyajian suatu gugus data sehinggamemberikan informasi yang berguna. Statistika deskriptif berhubungan dengan cara menggambarkan, mendeskripsikan atau menyimpulkan data baik secara numerik (misal menghitung rata-rata dan deviasi standar) atau secara grafis (dalam bentuk tabel atau grafik) untuk mendapatkan gambaran sekilas mengenai data tersebut sehingga lebih mudah dibaca dan bermakna.

Statistika deskriptif hanya memberikan informasi mengenai data yang dimiliki dan sama sekali tidak menarik kesimpulan apapun tentang gugus induknya yang lebih besar. Statistika deskriptif biasanya dimunculkan dalam bentuk tabel, diagram,grafik dan dalam bentuk lain seperti di majalah dan koran-koran. Dengan statistika deskriptif, kumpulan data yang diperoleh akan tersaji dengan rapi serta dapat memberikan informasi inti dari kumpulan data yang ada. Informasi yang dapat diperoleh dari analisis deskriptif antara lain ukuran pemusatan data, ukuran penyebaran data, serta kecenderungan suatu gugus data (Walpole, 1993).

Dalam penelitian ini dijelaskan gambaran umum tentang tingkat keparahan korban kecelakaan lalu lintas di Kabupaten Pati Jawa Tengah dengan menggunakan statistik deskriptif berupa diagram. Penggambaran dalam bentuk diagram ini diharapkan mampu mempermudah pembaca dalam memahami dan memperoleh informasi mengenai data yang tersedia.

1.2 Pengertian Kecelakaan

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.

Lalu lintas merupakan pengerakan kendaraan dan orang di ruang lalu lintas jalan, kecelakaan lalu lintas dapat diartikan sebagai suatu peristiwa atau kejadian di jalan raya yang tidak disangka-sangka dan tidak di sengaja yang melibatkan korban manusia atau kerugian harta benda sedangkan korban kecelakaan lalu lintas dapat berupa korban meninggal dunia, luka berat, dan luka ringan.

Kecelakaan lalu lintas merupakan suatu tragedi manusia dimana banyak nyawa manusia di bawah umur 40 tahun yang hilang nyawanya karena terjadi kecelakaan di jalan raya (Zulhendra, 2015). Kecelakaan merupakan tindakan tidak direncanakan dan tidak terkendali, ketika aksi dan reaksi objek, bahan, atau radiasi menyebabkan cedera atau kemungkinan cedera (Heinrich, 1931).

1.3 Faktor-faktor Penyebab Kecelakaan

Kecelakaan lalu lintas dapat terjadi karena beberapa faktor yaitu faktor pengemudi (manusia), lalu lintas, jalan, kendaraan dan lingkungan (Sulistio, 2009).

1.3.1 Tingkat Keparahan Korban

Tingkat keparahan korban berdasarkan korban kecelakaan menitik beratkan pada manusia itu sendiri, kecelakaan ini dapat berupa luka ringan, luka berat maupun meninggal dunia.

1.3.2 Jenis Kelamin

Menurut (Hungu, 2007) jenis kelamin (seks) adalah perbedaan antara perempuan dengan laki-laki secara biologis sejak seseorang lahir.

1.3.3 Usia

Usia merupakan perhitungan waktu yang di mulai dari saat kelahiran seseorang sampai dengan waktu penghitungan usia (Wikipedia, 2018). Sedangkan usia yang dimaksud dalam penelitian ini yaitu usia seseorang saat mengalami kecelakaan lalu lintas, yang dapat di kelompokkan menjadi tiga kategori yaitu kstegori usia anak-anak dan remaja (0-21) tahun, kategori usia dewasa (22-55) tahun dan kategori lanjut Usia ≥ 56 tahun.

1.3.4 Jenis Kecelakaan

Peristiwa terjadinya suatu kecelakaan lalu lintas sangat tergantung dari proses terjadinya mapun faktor penyebabnya. Berdasarkan UU Nomer 14 tahun 1992 tipe atau jenis suatu kecelakaan dapat di bagi menjadi empat kategori peristiwa yaitu tabrakan depan-belakang (Rear End), tabrakan depan-depan (Head On), tabrakan samping-samping, tabrakan depan-samping.

1.3.5 Jenis Kendaraan

Kendaraan adalah suatu alat yang dapat bergerak dijalan yang dibuat oleh manusia dengan tenaga penggerak mesin, serta terdiri dari kendaraan bermotor dan tidak bermotor. Berdasarkan Peraturan Pemerintah Nomer 44 Bab 1 Pasal 1 Tahun 1993 menjelaskan kendaraan bermotor merupakan kendaraan yang digerakkan oleh peralatan teknik yang berada pada kendaraan tersebut, kendaraan bermotor dapat di

bagi menjadi beberapa jenis yaitu sepeda motor, mobil penumpang, mobil bus, mobil barang, kendaraan khusus, kendaraan umum, dan kendaraan tidak bermotor.

1.3.6 Peran Korban

Peran korban merupakan orang yang baik secara individual maupun kolektif telah menderita kerugian termasuk kerugian fisik maupun mental, emosional, dan ekonomi. Pada penelitian ini peran korban diklasifikasikan menjadi pengendara/pengemudi, penumpang kendaraan selain pengendara.dan pengguna jalan non penumpang kendaraan.

1.3.7 Pekerjaan Korban

Pekerjaan korban merupakan suatu hal yang sudah menjadi kegiatan yang dikerjakan oleh seseorang setiap hari guna untuk memenuhi kebutuhannya, contoh pekerjaan korban yaitu pedagang, petani, buruh, karyawan, pelajar/mahasiswa, PNS, POLRI, wiraswasta dll.

1.3.8 Waktu Kejadian

Waktu adalah seluruh rangkaian saat ketika proses, perbuatan, atau keadaan berada atau berlangsung. Dalam hal ini, skala waktu merupakan intervasi antara dua buah keadaan/kejadian.

1.3.9 Tanggal Kejadian

Suatu peristiwa dimana aktivitas atau kejadian perkara sedang berlangsung yang dilakukan secara tiba-tiba maupun tidak di sengaja tanpa melihat situasi dan kondisi lingkungan sekitar.

1.3.10 Alat Keselamatan

Sasaran manajemen kesalamatan dan kesehatan kerja ialah mengurangi dan menghilangkan faktor-faktor yang berperan dalam kejadian kecelakaan dan penyakit akibat kerja ditempat kerja sehingga terwujud duatu tempat kerja yang aman dan sehat yang dapat mendukung proses produksi yang efisien dan produktif.

1.4 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). *Data Mining* memiliki kegunaan untuk menspesifikasikan pola yang harus ditemukan dalam tugas *data mining*. Secara umum tugas *Data Mining* data diklasifikasikan ke dalam dua kategori yaitu deskriptif dan prediktif.

1.5 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. Klasifikasi merupakan teknik penting dalam data

mining, teknik ini dapat memprediksi label kelas. Sehingga tujuan dari klasifikasi adalah kebenaran dalam memprediksi sebuah nilai (Sarangi & Jaglan, 2013).

1.6 Algoritma *Naïve Bayes*

Naïve Bayes dipelajari secara ekstensif sebegai metode dasar proses klasifikasi sejak tahun 1950an (Zheng & Webb, 2000). *Bayes* merupakan teknik prediksi berbasis *probabilistica* sederhana berdasarkan aturan *Bayes* dengan asumsi independensi yang kuat yang artinya antar atribut tidak memiliki hubungan. Ide dasar dari aturan Bayes adalah bahwa hasil dari hipotesis dari suatu peristiwa dapat diperkirakan oleh kejadian lain. Keterkaitan *Naïve Bayes* dengan klasifikasi adalah hipotesis peristiwa tersebut merupakan target atau label dalam klasifikasi sedangkan beberapa kejadian merupakan atribut dalam klasifikasi.

Naïve Bayes menggunakan cabang matematika yang dikenal dengan teori probabilitas untuk mencari peluang terbesar dari kemungkinan klasifikasi, dengan cara melihat frekuensi tiap klasifikasi pada data training (Siradjuddin, 2015). *Naïve Bayes* adalah algoritma yang digunakan dalam statistika untuk menghitung peluang dari suatu hipotesis, *Naïve Bayes* menghitung peluang suatu label berdasarkan pada atribut yang dimiliki dan menentukan label yang memiliki peluang paling tinggi. *Naïve Bayes* mengklasifikasikan label berdasarkan pada probabilitas sederhana dengan mangasumsikan bahwa setiap atribut dalam data tersebut bersifat saling terpisah. *Naïve Bayes* merupakan salah satu metode yang banyak digunakan berdasarkan beberapa sifatnya yang sederhana.

Naïve Bayes merupakan metode klasifikasi yang dapat memprediksi probabilitas sebuah class, sehingga dapat menghasilkan keputusan berdasarkan data

pembelajaran. Dari kelompok pendekatan numeris, *naive bayes* memiliki kelebihan antara lain, sederhana, cepat, dan berakurasi tinggi (Syarifah & Muslim, 2015). *Naïve Bayes* juga mengklasifikasikan data berdasarkan peluang atribut dari setiap label data. Pada model peluang setiap label dan jumlah atribut yang dapat dituliskan seperti persamaan berikut.

$$\boldsymbol{P(C_i|X_1, X_2, \dots X_n)} \quad (2.1)$$

Penghitungan *Naïve Bayes* dapat dijelaskan dengan C_i adalah hipotesis data yang merupakan suatu label $P(C_i|X)$ adalah peluang hipotesis label berdasarkan kategori X *(posteriori probability)*. $P(C_i)$ adalah peluang hipotesis label *(prior probability)*. $P(X|C_i)$ adalah peluang data berdasarkan kategori pada hipotesis label. $P(X)$ adalah jumlah probabilitas data yang nilainya adalah satu. Sehingga didapatkan rumus penghitungan *Naïve Bayes* dituliskan pada persamaan.

$$\boldsymbol{P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}} \quad (2.2)$$

Untuk menentukan peluang dari masing-masing kategori label digunakan persamaan berikut:

$$\boldsymbol{P(C_i|X) = \prod_{k=1}^{n} P\,(X_k|C_i)} \quad (2.3)$$

Untuk menentukan nilai label pada atribut tertentu digunakan persamaan:

$$\boldsymbol{\underset{c \in C}{argmax} = \frac{P(X|C_i)P(C_i)}{P(X)}} \quad (2.4)$$

1.7 Laplace Estimator

Untuk menyiasati supaya hasil probabilitas pada perhitungan *Naive Bayes* tidak bernilai nol dikarenakan tidak adanya data untuk suatu kategori tertentu dalam kelasnya dapat digunakan teknik estimasi yang biasa disebut *Laplace estimato*r atau *Laplacian correction* (Han and kamber, 2006). Dalam teknik ini digunakan

penambahan nilai 1 pada data untuk masing-masing kategori ketika ada kategori yang memiliki nilai 0 sehingga untuk sebanyak k kategori dimana j=1,2,... k dan N= $N = \sum_{j=1}^{n} n_j$ jika masing-masing kategori dalam kelasnya bernilai n_i maka

$$P(X=i)=\frac{n_j+1}{N+banyak\ kategori} \quad (2.5)$$

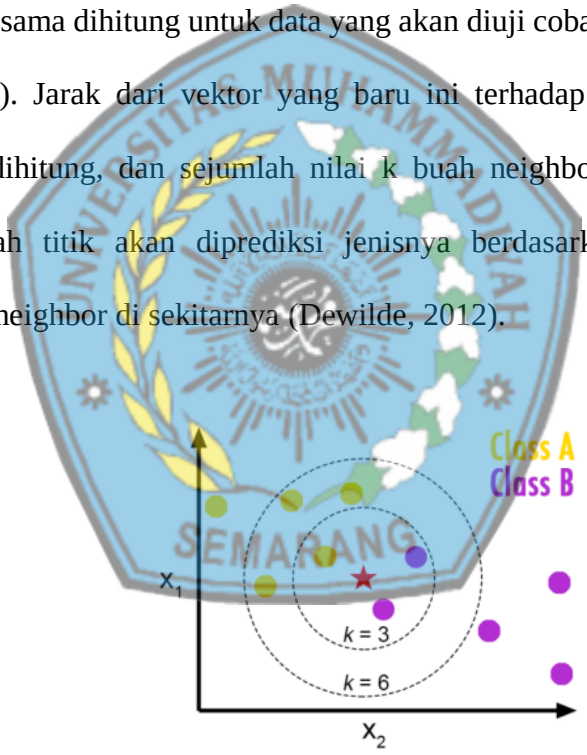
1.8 Algoritma K-Nearest Neighbor (KNN)

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). Algoritma klasifikasi K-NN memprediksi kategori tes sampel sesuai dengan sampel pelatihan k yang merupakan tetangga terdekat dengan sampel uji, dan memasukkan ke dalam kategori yang memiliki kategori probabilitas terbesar (Suguna & Thanushkodi, 2010). Kelebihan K-NN yaitu pelatihan sangat cepat, sederhana dan mudah dipelajari, tahan terhadap data pelatihan yang derau dan efektif jikadata pelatihan besar dan Kelemahan KNN adalah nilai k bias, komputasi kompleks, keterbatasan memori (Mutrofin et all, 2014).

Metode klasifikasi K-NN memiliki beberapa tahap, yang pertama nilai k yang merupaan jumlah k tetangga terdekat yang akan menentukan kueri baru masuk ke kelas mana ditentukan. Tahap kedua, k tetangga terdekat dicari dengan cara

menghitung jarak titik kueri dengan titik training. Tahap ketiga, setelah mengetahui jarak titik training dengan titik kueri, kemudian lihat nilai yang paling kecil. Tahap keempat diambil k nilai terkecil selanjutnya lihat kelasnya. Kelas yang paling banyak merupakan kelas dari kueri baru (Pramesti, 2013).

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



Gambar 2.1 ***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)

$$D(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{\boldsymbol{k}}^{\boldsymbol{n}} (\boldsymbol{x_k} - \boldsymbol{y_k})^2} \qquad (2.6)$$

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).

1.9 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 2.1 **Confusion Matrix**

	Clasified as	
	+	-

Correct	+	*True Positive*	*False Positif*
Clasification	-	*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 berhasil diklasifikasikan sebagai *record* negatif. Metode pengujian *confusion matrix* dapat menghasilkan perhitungan dengan 5 output, diantaranya yaitu

$$Akurasi = \frac{TP+TN}{TP+FN+FP+TN} \; x \; 100 \; \% \quad (2.7)$$

$$Presisi = \frac{\text{TP}}{TP+FP} x \; 100 \; \% \quad (2.8)$$

$$Recall = \frac{\text{TP}}{TP+FN} x \; 100 \; \% \quad (2.9)$$

$$F - Measure = 2 \; x \frac{\text{precission X recall}}{precission+recall} \quad (2.10)$$

$$Error = \frac{FN+FP}{TP+TN+FP+FN} \; x \; 100 \; \% \quad (2.11)$$

1.10 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). Berikut ini adalah penjelasan dari ukuran performani evaluasi (Hossin & Sulaiman, 2015):

a. *Recall :* recall digunakan untuk mengukur fraksi pola positif yang diklasifikasikan dengan benar.
b. *Precision* : presisi digunakan untuk mengukur pola positif yang diprediksi dengan benar dari total pola yang diprediksi di kelas positif.
c. *F-Measure* : suatu ukuran yang menggambarkan rata-rata harmonis antara recall dan nilai presisi.
d. *Acuracy* : suatu ukuran rasio prediksi yang benar terhadap total jumlah sampel dievaluasi.
e. *Error Rate* : kasus yang diidentifikasi salah dengan sejumlah semua kasus.

1.11 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.

1.12 ROC Curve

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. 2 ***Nilai Kualitas*** **Classifier**

Rentang Akurasi	Klasifikasi

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*