PENERBITAN ARTIKEL ILMIAH MAHASISWA

Universitas Muhammadiyah Ponorogo

ANALISA KECELAKAAN LALU LINTAS MENGGUNAKAN METODE ALGORITMA C4.5 DAN *NAÏVE BAYES*

(STUDI KASUS DI KABUPATEN PONOROGO)

Tien Rubiyanti, Ida Widaningrum, Andy Triyanto

Jurusan Teknik Informatika, Fakultas Teknik, Universitas Muhammadiyah Ponorogo

Jalan Budi Utomo 10 Ponorogo

Email : tienrubiyanti38@gmail.com

Abstrak

Penelitian ini menerapkan metode data mining yang kemudian dilakukan komparasi dari dua metode yang berbeda yaitu Algoritma C4.5 dan Naïve Bayes yang dikaji untuk memperoleh nilai akurasi yang terbaik berdasarkan data kecelakaan lalu lintas yang ada di Kota Ponorogo untuk mengetahui penyebab kecelakaan lalu lintas dengan kategori faktor pengemudi, faktor jalan, faktor cuaca dan faktor kendaraan. Kedua algoritma tersebut dibantu dengan perangkat lunak Weka yang berbasis Open Source (GPL) dan berengine Java. Maka hasil pengujian kedua algoritma tersebut diuji menggunakan Confusion Matrix dan Kurva ROC (Receiver Operating Characteristic), menunjukkan algoritma C4.5 nilai Accuracy sebesar 88.2609% dan nilai AUC (Area Under Curve) sebesar 0.996 untuk algoritma Naïve Bayes nilai accuracy sebesar 86.5217% dan nilai AUC (Area Under Curve) sebesar 0.9496. Sehingga algoritma C4.5 merupakan metode yang terbaik sebab nilai Accuracy dan AUC yang tertinggi.

Kata kunci : *Kecelakaan Lalu Lintas, Weka, Algoritma C4.5, Naïve Bayes*

1. PENDAHULUAN

Kabupaten Ponorogo merupakan daerah yang ada di Provinsi Jawa Timur 200 Km sebelah barat daya dari provinsi, dan sekitar 800 Km sebelah timur dari ibu kota Negara Indonesia. Kabupaten Ponorogo berada pada 111°7’ hingga 111° 52’ Bujur Timur dan 7° 49’ hingga 8° 20’ Lintang Selatan. Data jumlah penduduk Kabupaten Ponorogo yang dihasilkan dari proyeksi BPS yaitu sebesar 865.809 jiwa pada tahun 2014 [1].

Indonesia merupakan salah satu Negara dengan tingkat presentase kecelakaan yang cukup tinggi. Menurut Dinas Perhubungan, kecelakaan lalu lintas menjadi penyebab kematian nomor tiga di Indonesia setelah serangan jantung dan stroke.

Menurut data dari pihak Satlantas Polres Ponorogo, pada pertengahan Desember tahun 2016, tercatat 334 kasus kecelakaan lalu lintas dan korban yang meninggal mencapai 120 korban jiwa. Angka tersebut meningkat dari tahun 2015 lalu, korban jiwa 108, dan

mengalami kenaikan 12 orang. Sementara untuk jumlah kasus di tahun 2016 mengalami penurunan dari tahun sebelumnya, yakni pada 2015 yang lalu ada 530 kasus. Dari data tersebut maka diperlukan adanya upaya untuk mengurangi jumlah kecelakaan. Sebagai langkah awal maka diperlukan untuk mengolah data tersebut, sehingga variabel awal dari pemicu terjadinya kecelakaan di Kabupaten Ponorogo dapat diketahui.

Data mining bertujuan untuk menemukan pola dan aturan dalam basis data yang berukuran besar sehingga dapat dipakai dalam mengambil suatu keputusan. Salah satu teknik yang sering digunakan dalam data mining adalah Klasifikasi, sebab dari teknik ini tujuannya adalah mengelompokkan suatu objek ke dalam kelas tertentu berdasarkan pola kelas tertentu. Metode Klasifikasi yang paling sering dipakai antara lain Decision tree, Rule Based, dan Naïve Bayessian karena mudah untuk diinterpretasi dalam kehidupan nyata.

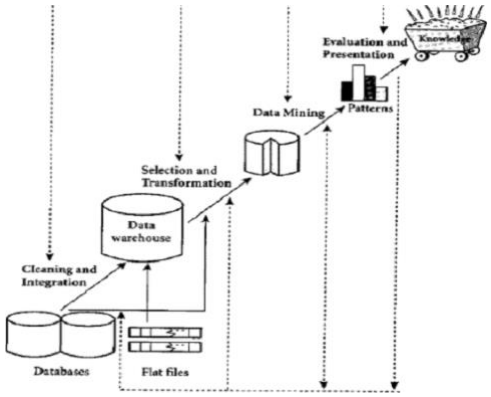
Melihat permasalahan yang ada maka peneliti menggunakan algoritma C4.5 dan *naïve bayes* sebagai pengambilan keputusan untuk mengetahui penyebab kecelakaan lalu lintas di Ponorogo, sehingga dapat mengurangi pemicu terjadinya kecelakaan lalu lintas.

Adapun tujuan dari penelitian ini adalah mengetahui hasil pengujian dari kedua algoritma yang paling akurat dalam mengolah informasi data kecelakaan lalu lintas, pada dasarnya untuk mengurangi pemicu terjadinya kecelakaan lalu lintas di Ponorogo, sehingga dapat memberikan penanganan alternatif pada kasus kecelakaan lalu lintas. Dan penelitian ini menggunakan perangkat lunak *Weka* salah satu aplikasi data mining berbasis *open source* (GPL) dan berengine Java.

2. LANDASAN TEORI

a. Data Mining

Menambang data atau *data mining* merupakan suatu metode yang dipakai untuk mengekstraksi suatu informasi tersembunyi yang bersifat prediktif pada suatu basis data, hal tersebut diakui oleh perusahaan – perusahaan sebagai teknologi yang sangat potensial [8]. *Knowledge Discovery in Database* sebutan lain dari *data mining* yang merupakan pencarian pola, keteraturan, atau hubungan dalam data berukuran besar dengan cara pengumpulan pemakaian data historis [7].



Gambar 2.1 Tahap Data Mining

Dalam KDD terdapat beberapa proses yang meliputi [4] antara lain :

1) *Data Cleaning* (Pembersihan Data)
2) *Data Integration* (Intregasi Data)
3) *Data Selection* (Seleksi Data)
4) *Data Transformation* (Transformasi Data)
5) *Data Mining*
6) *Pattern Evaluation* (Evaluasi Pola)
7) *Knowledge Presentation* (Presentasi Pengetahuan)

b. Kecelakaan

Menunurt Undang – undang No. 4 tahun 1992 Kecelakaan lalu lintas adalah suatu peristiwa atau kejadian yang tidak disengaja yang melibatkan kendaraan dengan atau tanpa pemakai jalan lainnya, dan mengakibatkan korban jiwa yang mengalami luka ringan, luka berat dan meninggal dunia juga kerugian material [4].

c. Algoritma C4.5

Algoritma C4.5 merupakan suatu metode yang dipakai untuk mengklasifikasi pada teknik data mining. Klasifikasi ialah suatu teknik untuk mengetahui sekumpulan fungsi atau aturan yang dipakai untuk menentukan suatu kelas data yang satu dengan lainnya berfungsi untuk menyatakan sebuah objek.

Metode ini sangat terkenal sebab dapat mengklasifikasi serta memberikan hubungan antar variabel. Algoritma C4.5 ini digunakan untuk membuat sebuah *decision tree* (pohon keputusan).

Tahap – tahap untuk membentuk pohon keputusan sebagai berikut :

1) Mempersiapkan *data training* yang diperoleh dari data histori lalu dikelompokkan sesuai dengan kelasnya.
2) Tentukan akar pohonnya, dengan cara menghitung nilai *gain* dari atribut tersebut dan nilai *gain* yang tertinggi dijadikan sebagai akar pertama. Sebelum menghitung nilai *gain* maka terlebih dahulu menhitung nilai *entropy*, berikut rumus mencari nilai *entropy* [6] :

$$\text{Entropy (S)} \sum_{i=1}^{n} - pi * log2\ pi \qquad (2.3)$$

Keterangan:

S : Himpunan kasus

n : Jumlah partisi S

p_i : Proporsi S_i terhadap S

3) Selanjutnya menghitung nilai *gain* dengan menerpakan rumus :

$$\text{Gain(S. A)} = \text{Entropy(S)} - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * \text{Entropy } (S_i)$$

Keterangan:

S : Himpunan kasus

A : Parameter (atribut)

n : Jumlah partisi atribut A

$|S_i|$: Jumlah kasus pada partisi ke i

|S| : Jumlah kasus dalam S

4) Kemudian ulangi tahap 2 sampai semua recordnya terpatisi.
5) Adapun pada proses dari partisi akan berhenti saat:
   a) Jika semua record pada simpul N menperoleh kelas yang sama
   b) Tidak terdapat atribut atau variabel dalam *record* yang dipartisi lagi
   c) Dan tidak terdapat *record* pada cabang yang kosong

d. *Naïve Bayes*

Metode *naïve bayes* merupakan salah satu teknik data mining yang sangat populer dalam mengklasifikasi sebuah data yang berukuran besar yang kemudian dapat dipakai untuk memprediksi probabilitas keanggotaan suatu *class*.

Teorema *bayes* dapat memprediksi peluang dimasa depan sesuai dengan kejadian sebelumnya. Teorema bayes, dimana *X* dijabarkan oleh sekumpulan *n* atribut dari beberapa hipotesis, sehingga data *X* termasuk sebuah *class C* [3]. Dengan teorema *bayes*, sebagai berikut :

$$P(H|X) = \frac{\text{P(X|H).P(H)}}{\text{P(X)}} \qquad (2.1)$$

Keterangan :

X : Data class yang belum diketahui

H : Hipotesis data

P(H|X) : Probabilitas hipotesis H berdasar kondisi X (posteriori probabilitas)

P(H) : Probabilitas hipotesis H (prior probabilitas)

P(X|H) : Probabilitas X berdasarkan pada hipotesis H (2.2)

P(X) : Probabilitas X

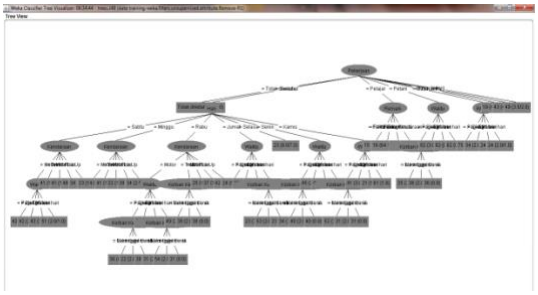
3. PEMBAHASAN

Dalam penelitian ini sample yang digunakan untuk mengolah sejumlah 230 data

kecelakaan dan 116 data kecelakaan digunakan untuk menguji hasil data training.

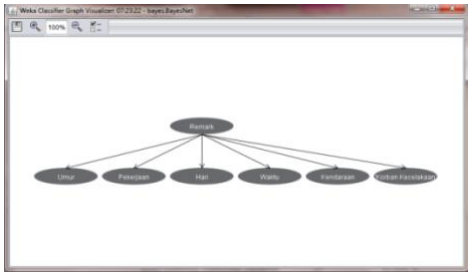
a. Algoritma C4.5

Dari hasil perhitungan *entropy* dan *gain*, bahwa atribut waktu mempunyai nilai *gain* tertinggi.



Gambar 3.1 Pohon Keputusan Hasil dengan tools Weka

b. Naïve Bayes



Gambar 3.2 Remark Penyebab Kecelakaan *Naïve Bayes* dengan Weka

Hasil dari Weka remark penyebab kecelakaan lalu lintas terdiri dari beberapa variabel antara lain umur, pekerjaan, hari, waktu, kendaraan dan korban kecelakaan. Dan remark itu merupakan faktor kecelakaan yang meliputi faktor jalan, faktor pengemudi, faktor cuaca dan faktor kendaraan.

c. Pengujian Algoritma

Dalam pengujian algoritma tersebut yang akan diuji adalah tingkat nilai akurasi dari kedua algoritma tersebut dengan cara mengolah dari data uji. Dan hasil nilai akurasi algoritma C4.5 88.2609% dan *naïve bayes* 86.5217%.

1) Confusion Matrix

Tabel 1 *Confusion Matrix* untuk Algoritma C4.5

```
=== Confusion Matrix ===

   a   b   c   d   <-- classified as
   0  22   0   0 |   a = Faktor Jalan
   0 203   0   0 |   b = Faktor Pengemudi
   0   3   0   0 |   c = Faktor Cuaca
   0   2   0   0 |   d = Faktor Kendaraan
```

Table diatas adalah hasil dari *confusion matrix* untuk metode algoritma C4.5 menggunakan data *training* berjumlah 230 data.

Tabel 2 *Confusion Matrix* untuk *Naïve Bayes*

```
=== Confusion Matrix ===

   a   b   c   d   <-- classified as
   1  21   0   0 |   a = Faktor Jalan
   5 198   0   0 |   b = Faktor Pengemudi
   0   3   0   0 |   c = Faktor Cuaca
   0   2   0   0 |   d = Faktor Kendaraan
```

Table diatas adalah hasil dari *confusion matrix* untuk metode naïve bayes dengan menggunakan data *training* sejumlah 230 data.

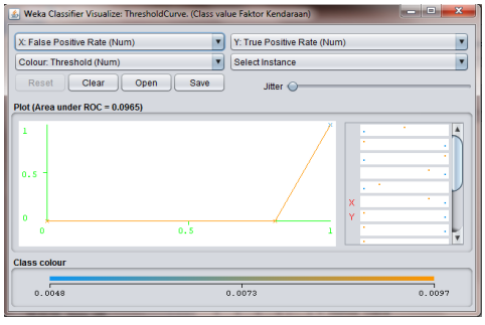
Maka langkah selanjutnya mencari nilai *accuracy*, *precision* dan *recall* dari kedua algoritma tersebut pada tabel 3

Tabel 3 Perbandingan Nilai *Accuracy*, *Precision*, dan *Recall*

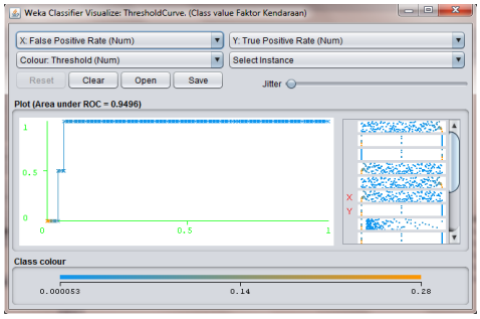
	C4.5	*Naïve Bayes*
Accuracy	88.2609 %	86.5217 %
Precision	0.779 %	0.796 %
Recall	0.883 %	0.865 %

2) Kurva ROC

Berikut ini adalah perbandingan dari kedua algoritma dengan melihat kurva ROC pada gambar dibawah ini :



Gambar 3.3 Kurva ROC algoritma C4.5



Gambar 3.4 Kurva ROC Metode Naïve *Bayes*

Dari gambar 3.3 dan gambar 3.4 merupakan kurva ROC penjabaran dari *Confusion Matrix* pada tabel 1. *False positive* adalah garis horizontal dan *true positive* adalah garis vertikal.

Hasil perhitungan dari nilai AUC pada kedua algoritma tersebut :

Tabel 4 Perbandingan Nilai AUC

	C4.5	*Naïve Bayes*
AUC	0.9965	0.9496

d. Analisa Hasil Perbandingan

Sehingga dapat disimpulkan dari tabel 3 bahwa hasil perbandingan kedua metode tersebut dan algoritma C4.5 yang memiliki nilai paling tinggi dibandingkan dengan *naïve bayes*.

Tabel 5 Perbandingan Nilai *Accuracy* dan AUC

	C4.5	*Naïve Bayes*
Accuracy	88.2609 %	86.5217 %
AUC	0.9965	0.9496

Untuk klasifikasi data mining, maka ada beberapa kelompok untuk menentukan nilai AUC [2] antara lain :

1) 0.90 – 1.00 klasifikasi sangat baik
2) 0.80 – 0.90 klasifikasi baik
3) 0.70 – 0.80 klasifikasi cukup
4) 0.60 – 0.70 klasifikasi buruk
5) 0.50 – 0.60 klasifikasi salah

Sehingga dapat disimpulkan dari tabel 5 bahwa algoritma C4.5 dan naïve bayes masuk kedalam klasifikasi sangat baik sebab nilai AUC-nya antara 0.90 – 1.00.

e. Perapan Algoritma Terpilih

Algoritma C4.5 adalah algoritma yang terpilih sehingga apabila terdapat data baru maka tinggal menginputkan.

Tabel 6 Data Baru untuk *Testing*

Umur	Pekerjaan	Hari	Waktu	Kendaraan	Korban Ke	Remark
Tidak dike	Tidak dike	Sabtu	Pagi hari	Motor	Luka ringa	Faktor Jalan
23	Swasta	Minggu	Siang hari	Motor	Luka ringa	Faktor Pengemudi
38	Swasta	Rabu	Pagi hari	Motor	Luka ringa	Faktor Pengemudi
Tidak dike	Tidak dike	Minggu	Pagi hari	Motor	Luka ringa	Faktor Pengemudi
24	Swasta	Jumat	Pagi hari	Motor	Luka ringa	Faktor Jalan
22	Swasta	Selasa	Siang hari	Motor	Luka ringa	Faktor Jalan
40	Swasta	Minggu	Dini hari	Motor	Meningga	Faktor Pengemudi
14	Pelajar	Minggu	Pagi hari	Motor	Meningga	Faktor Pengemudi
27	Swasta	Selasa	Siang hari	Motor	Luka ringa	Faktor Pengemudi
17	Pelajar	Selasa	Malam ha	Motor	Meningga	Faktor Cuaca
15	Pelajar	Sabtu	Malam ha	Motor	Luka ringa	Faktor Jalan
16	Pelajar	Selasa	Siang hari	Motor	Luka ringa	Faktor Pengemudi
12	Pelajar	Minggu	Dini hari	Motor	Luka ringa	Faktor Pengemudi
50	Swasta	Senin	Pagi hari	Motor	Meningga	Faktor Pengemudi
82	Petani	Jumat	Siang hari	Motor	Meningga	Faktor Jalan
34	Polri	Sabtu	Siang hari	Truk	Luka ringa	Faktor Pengemudi
19	Mahasisw	Selasa	Siang hari	Motor	Meningga	Faktor Pengemudi
53	Swasta	Minggu	Pagi hari	Motor	Luka ringa	Faktor Pengemudi
31	Swasta	Jumat	Pagi hari	Motor	Meningga	Faktor Pengemudi
20	Swasta	Senin	Malam ha	Motor	Luka ringa	Faktor Cuaca
19	Mahasisw	Rabu	Malam ha	Motor	Luka ringa	Faktor Jalan
Tidak dike	Tidak dike	Jumat	Dini hari	Truk	Luka ringa	Faktor Jalan
Tidak dike	Tidak dike	Senin	Pagi hari	Motor	Luka ringa	Faktor Pengemudi
15	Pelajar	Senin	Siang hari	Motor	Luka ringa	Faktor Pengemudi
17	Pelajar	Jumat	Pagi hari	Motor	Luka ringa	Faktor Pengemudi

Tingkat akurasi dari data baru tersebut sebesar 79.3103% dan *confusion matrix* sebagai berikut :

```
=== Confusion Matrix ===

  a  b  c  d   <-- classified as
  0 19  0  0 |  a = Faktor Jalan
  0 92  0  0 |  b = Faktor Pengemudi
  0  3  0  0 |  c = Faktor Cuaca
  0  2  0  0 |  d = Faktor Kendaraan
```

4. KESIMPULAN DAN SARAN

a. Kesimpulan

Dalam penelitian ini membahas tentang menganalisa kecelakaan lalu lintas dengan menggunakan algoritma C4.5 dan *Naïve Bayes* untuk mengetahui penyebab kecelakaan lalu lintas. Faktor yang menyebabkan kecelakaan yang ada di Ponorogo paling banyak disebabkan oleh faktor pengemudi dengan presentase 89%. Dan kebanyakan kecelakaan lalu lintas terjadi pada pagi hari saat hari kerja, jenis kendaraan yang mengalami kecelakaan tertinggi adalah sepeda motor dan korban kecelakaan dengan luka ringan merupakan jumlah yang sangat banyak.

Dari kedua algoritma tersebut pengujiannya menerapkan model CRISP-DM dan melakukan perbandingan untuk menentukan diantara kedua model tersebut mana yang mempunyai tingkat akurasi tertinggi kemudian diterapkan pada data kecelakaan untuk mengetahui faktor penyebab kecelakaan lalu lintas. Dalam mengukur kinerja dari kedua metode tersebut maka menggunakan *Confusion Matrix* dan Kurva ROC, algoritma C4.5 yang memiliki nilai *accuracy* dan ROC yang tinggi dibandingkan *Naïve Bayes*.

Sehingga dapat disimpulkan metode algoritma C4.5 merupakan metode yang terbaik dalam mengklasifikasikan data. Maka algoritma C4.5 dapat memecahkan masalah untuk permasalahan dalam mengetahui faktor penyebab kecelakaan lalu lintas di Kabupaten Ponorogo.

b. Saran

Adapun saran dari hasil penelitian ini sebagai berikut :

1) Penelitian selanjutnya disarankan untuk menerapkan metode selain C4.5 dan *Naïve Bayes*.

2) Penelitian selanjutnya mengumpulkan lebih lengkap data kecelakaan untuk memetakan kelompok kecelakaan lalu lintas di Ponorogo.

3) Diharapkan untuk pihak kepolisian di Kabupaten Ponorogo agar lebih memberikan pengawasan lalu lintas pada jalan sepi karena cenderung terjadinya kecelakaan lalu lintas.

4) Diharapkan untuk pihak kepolisian di Kabupaten Ponorogo agar memberikan sosialisasi disekolah – sekolah mengenai keselamatan dalam lalu lintas.

5) Diharapkan untuk pemerintah agar memperbaiki dan memfasilitasi jalan seperti lampu penerangan, rambu – rambu lalu lintas dan jalan yang rusak dan berlubang karena itu merupakan salah satu penyebab terjadinya kecelakaan.

5. DAFTAR PUSTAKA

[1] Badan Pusat Statistik (2015), Ponorogo Dalam Angka 2015. Hlm 3 dan 43 (Katalog BPS : 1102001.3502)

[2] Gorunescu, Florin, Data Mining (2011): Concepts, Models, and Techniques, Verlag Berlin Heidelberg : Springer.

[3] Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques Third Edition.* Waltham: Morgan Kaufmann.

[4] Harahap, G. (1995), *Masalah Lalu lintas dan Pengembangan Jalan (DPU)*. Bandung.

[5] Kabir, M. F., Rahman, C. M., Hossain, A., & Dahal, K. (2011). Enhanced Classification Accuracy on Naive Bayes Data Mining Models.

[6] Kusrini, & Emha Taufiq Luthfi. (2009). *Algoritma Data Mining.* Yogyakarta: Andi.

[7] Santoso, Budi (2007). Data mining Teknik Pemanfaatan Data untuk Keperluan Bisnis, Yogyakarta : Graha Ilmu

[8] Sulistiana Feri, Juju Dominikus (2010), Data Mining Meramalkan Bisnis Perusahaan, Jakarta