JUSTIN
Jurnal Sistem dan Teknologi Informasi
p-ISSN : 2460-3562 / e-ISSN : 2620-8989

DOI: 10.26418/justin.v8i3.40982
Vol. 08, No. 3, Juli 2020

Integrasi Metode *Decision Tree* dan SMOTE untuk Klasifikasi Data Kecelakaan Lalu Lintas

Afrilio Franseda[a1] Wawan Kurniawan [a2] Sita Anggraeni [a3] Windu Gata [a4]

[a]*Program Studi Informatika STMIK Nusa Mandiri*
Jl. Damai No. 8 Warung Jati Barat (Margasatwa) Jakarta Selatan 12540

[1]14002355@nusamandiri.ac.id
[2]14002315@nusamandiri.ac.id
[3]sita.sia@nusamandiri.ac.id
[4]windu@nusamandiri.ac.id

Abstrak

Kecelakaan lalu lintas merupakan suatu peristiwa yang tidak dapat diprediksi dengan pasti dan dapat mengakibatkan korban jiwa, korban luka ringan, korban luka berat atau kerugian materil seperti benda berharga. Permasalahan ini terjadi di seluruh dunia, tidak terkecuali Australia Selatan yang merupakan salah satu wilayah di Australia. Tercatat bahwa wilayah tersebut memiliki total kecelakaan yang memakan korban 4.953 pada tahun 2018. Oleh karena itu, dibutuhkan analisis untuk mengantisipasi kecelakaan agar tidak terulang kembali kejadian dengan faktor yang sama. Salah satu solusi untuk permasalahan ini yaitu diperlukan metode klasifikasi untuk mengelompokkan faktor-faktor yang mempengaruhi kecelakaan lalu lintas. Metode klasifikasi yang digunakan sebagai pengolah data adalah metode *Decision Tree*. Metode pada permasalahan ketidakseimbangan kelas menggunakan metode *Synthetic Minority Over-sampling Technique* (SMOTE). Untuk proses dalam meningkatkan evaluasi pada penelitian ini menggunakan proses *Knowledge Discovery in Database* (KDD). Pengujian dilakukan dengan tiga desain model yaitu *Split Validation Decision Tree* dan SMOTE diperoleh akurasi 69.23%. Pengujian menggunakan *Cross Validation Decision Tree* dan SMOTE diperoleh akurasi 63.56%. Pengujian menggunakan *Decision Tree* dan SMOTE *Split Data* diperoleh akurasi 71.12% dengan perbandingan 1:9. Sehingga, setelah ketiga desain model tersebut dibandingkan, maka *Decision Tree* dan SMOTE *Split Data* mendapatkan akurasi yang paling baik. Selain itu diperoleh pula presisi 89.71% (3:7) dan *area under curve* (AUC) sebesar 0.773 (1:9). Penelitian ini masuk kedalam kategori *fair classification* (cukup).

Kata kunci: *Data Mining*, Klasifikasi, *Decision Tree*, SMOTE, Kecelakaan Lalu Lintas

Integration of Decision Tree and SMOTE Methods for Classification of Traffic Accidents Data

Abstract

Traffic accidents are events that cannot be predicted with certainty and can result in casualties, minor injuries, serious injuries, or material losses such as valuable objects. This problem occurs throughout the world, including South Australia which is one of the regions in Australia. It is recorded that the area had a total of 4,953 casualties in 2018. Therefore the analysis is needed to anticipate the accident so that it does not happen again with the same factors. One solution to this problem is the classification method needed to classify the factors that affect traffic accidents. The classification method used for data processing is the Decision Tree method. The method for class imbalance problems uses the method of Synthetic Minority Over-sampling Technique (SMOTE). For the process of increasing evaluation in this study using the Knowledge Discovery in Database (KDD) process. The test was carried out with three model designs namely Split Validation Decision Tree and SMOTE model design obtained an accuracy of 69.23%. Testing use Cross Validation Decision Tree and SMOTE obtained an accuracy of 63.56%. Testing using the Decision Tree and SMOTE Split Data obtained an accuracy of 71.12% with a ratio of 1:9. So, after the three design models are compared, the split Decision Tree and SMOTE Split Data get the best accuracy. Also, a precision of 89.71% (3:7) and area under the curve (AUC) were obtained of 0.773 (1:9). This research belongs to the fair classification category.

Keywords: Data Mining, Classification, Decision Tree, SMOTE, Traffic Accident

I. PENDAHULUAN

Kecelakaan lalu lintas merupakan permasalahan yang dihadapi oleh setiap negara di seluruh dunia, tidak

terkecuali Australia. Khusus di Australia Selatan, menurut data yang dihimpun oleh Pemerintah Australia Selatan yaitu *Department of Planning, Transport and Infrastructure* pada tahun 2018 tercatat total kecelakaan 13.599 dengan rincian 75 kecelakaan fatal, 485 kecelakaan berat, 4.393 kecelakaan ringan dan 8.646 kerusakan meteril atau benda berharga. Ini mengakibatkan 80 korban meninggal, 576 luka berat dan 5.468 luka ringan. Berdasarkan data tersebut diperlukan analisis mengenai model yang menghasilkan akurasi terbaik berdasarkan faktor-faktor yang berkontribusi dalam terjadinya kecelakaan lalu lintas [1]. Kecelakaan lalu lintas merupakan suatu kejadian dimana sebuah kendaraan bermotor atau lebih bertabrakan dengan benda lain dan menyebabkan kerusakan. Kadang kecelakaan ini dapat mengkibatkan luka-luka atau kematian manusia atau binatang. Kecelakaan lalu lintas merupakan kejadian yang sulit untuk diprediksi walau telah diantisipasi, sehingga kita tidak tahu kapan dan dimana akan terjadinya [2].

Menurut Organisasi Kesehatan Dunia (WHO), faktor penyebab tertinggi terjadinya kecelakaan lalu lintas adalah *human error* atau kesalahan manusia itu sendiri. Pendekatan yang dilakukan melalui sistem keamanan yang dirancang untuk keselamatan diperjalanan yang bertujuan untuk memastikan keselamatan pengguna jalan, seperti batas kecepatan minimal dan maksimal kendaraan bermotor, namun *human error* tetep menjadi permasalahan yang harus dicari solusinya. Sistem ini akan bekerja apabila manusia yang terlibat patuh terhadap aturan yang telah dibuat. Penyebab kedua adalah ngebut atau pengemudi yang berkendara melebihi batas maksimal. Peningkatan kecepatan rata-rata secara langsung terkait baik dengan kemungkinan kecelakaan terjadi maupun beratnya konsekuensi dari kecelakaan itu. Penyebab selanjutnya infrastruktur jalan dimana desain jalan dapat memiliki dampak yang besar pada keselamatannya. Idealnya, jalan harus dirancang dengan dipertimbangkan berdasarkan keselamatan semua pengguna jalan. Ini berarti memastikan bahwa ada fasilitas yang memadai untuk pejalan kaki, pengendara sepeda, dan pengendara sepeda motor. Langkah-langkah seperti jalan setapak, jalur bersepeda, titik persimpangan yang aman, dan langkah-langkah penenangan lalu lintas lainnya bisa menjadi penting untuk mengurangi risiko cedera di antara pengguna jalan [3].

Untuk dapat mengklasifikasikan data yang diperoleh berdasarkan penyebab kecelakaan yang terjadi di Australia Selatan, diperlukan model atau metode pengolah data menggunakan *data mining* yang memiliki jumlah yang sangat besar dan untuk meminimalisir terulang kembali kecelakaan dari penyebab yang sama. Metode yang digunakan pada penelitian ini adalah metode klasifikasi *Decision Tree*.

Data mining adalah proses analisa data untuk mengekstraksi, mengidentifikasi pola, serta menemukan korelasi suatu informasi dengan cara menambang *repository* data yang begitu besar menggunakan teknik, metode atau algoritma tertentu dengan tujuan untuk menggali informasi penting. Pada *Knowledge Discovery in Database* (KDD), *data mining* merupakan salah satu bagian yang paling penting yang bertugas untuk mengekstrak pola atau model dari data dengan menggunakan suatu algoritma yang spesifik. Terdapat empat kelompok yang termasuk kedalam *data mining* diantaranya yaitu model prediksi atau klasifikasi, analisis kelompok atau *clustering*, analisis asosiasi dan deteksi anomali [4],[5].

Klasifikasi adalah salah satu model dalam *data mining*. Model klasifikasi merupakan teknik memprediksi data, membuat prediksi nilai dari suatu data yang hasilnya telah ditemukan berasal dari data yang berbeda. Tujuan dari model ini yaitu memprediksi nilai dari suatu variabel yang tidak diketahui dari variabel lain yang telah diberikan. Klasifikasi sering disebut *supervised learning* karena kelas yang telah ditentukan sebelum memproses data [6],[7].

Decision Tree atau yang disebut dengan pohon keputusan merupakan model dari klasifikasi. Bentuknya yang seperti struktur pohon merepresentasikan atribut setiap data yang diproses. Istilah *Decision Tree* yaitu mendeskripsikan tiap – tiap kelas untuk menemukan pola atau fungsi dengan tujuan untuk melakukan klasifikasi atau prediksi data yang belum mempunyai kelas. Metode ini sangat popular karena dapat dipakai pada banyak bidang. Beberapa algoritma yang sering yang menerapkan *Decision Tree* yaitu C4.5, ID3, dan *Random Forest* [8],[9].

Synthetic Minority Over-sampling Technique (SMOTE) merupakan pendekatan untuk menyeimbangkan data sampel pada kelas yang memiliki ketidakseimbangan berlebihan (mayoritas) dengan fokus terhadap kelas minoritas, dengan tujuan meningkatkan kinerja dari metode klasifikasi. Pada SMOTE kemungkinan terjadi *overfitting* yaitu data pada kelas minoritas yang terduplikasi [10], [11].

Pada penelitian sebelumnya terkait *data mining* tentang kecelakaan lalu lintas terdapat beberapa metode dan pembahasan yaitu klasifikasi atau prediksi, klusterisasi dan asosiasi. Melalui metode *K-Means Clustering* untuk mengidentifikasi lokasi kecelakaan lalu lintas dengan frekuensi yang tinggi dan faktor yang mempengaruhi kecelakaan tersebut [12], menggunakan metode *Hierarchical Clustering* untuk memprediksi tingkat keparahan kecelakaan lalu lintas [13], penelitian ini mengunakan 3 metode yaitu *Hierarchical Clustering*, *Random Forest* dan *Regression Tree* (CART) untuk mengidentifikasi dan menganalisa *cluster* yang berbahaya, serta mengidentifikasi *variable* yang memiliki dampak besar terkait kecelakaan [14]. Pada metode asosiasi dengan menggunakan algoritma *Fp-Growth*, tidak hanya terbatas pada faktor seperti waktu kejadian pada kecelakaan, jenis kecelakaan dan jalan, namun variabel seperti usia, pekerjaan dan jenis kelamin ikut membentuk pola terjadinya kecelakaan [15].

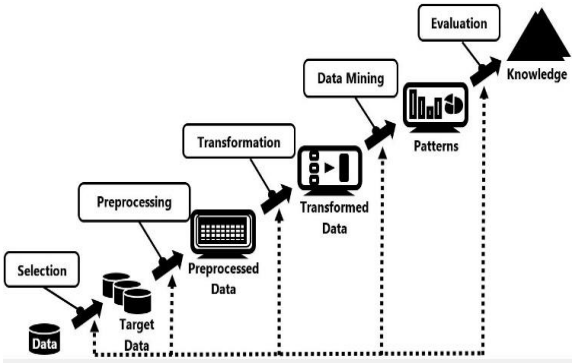
Pada penelitian dengan menggunakan *Classification*, dilakukan penelitian terkait faktor yang berkontribusi dalam kecelakaan parah dan dengan menggunakan beberapa model prediksi. Data yang digunakan yaitu catatan keselamatan yang disediakan oleh pemerintah

Inggris (STATS19). Penelitian ini memiliki tujuan untuk memprediksi atau memperkirakan keparahan suatu kecelakaan dengan menggunakan tiga algoritma klasifikasi yaitu *Bayesian Network, Decision Tree* dan *Artificial Neural Network* (ANN), melakukan eksplorasi faktor-faktor mendasar yang dapat mempengaruhi kecelakaan dan melakukan komparasi terkait model yang telah di prediksi [16]**.** Penelitian lain menerapkan metode klasifikasi bernama *Bootstrap Aggregating* (*Bagging*) pada kasus kecelakaan lalu lintas yang berlokasi di Surabaya dan menggunakan data pada tahun 2011. *Bootstrap Aggregating* merupakan metode untuk meningkatkan akurasi dari klasifikasi atau prediksi. Dalam hal ini metode analisis yang digunakan adalah *Ordinal Logistic Regression*. Terdapat tiga kategori yang digunakan dalam penelitian ini, yaitu luka ringan, luka berat dan meninggal dunia. Dengan demikian, tercipta suatu hubungan anatara faktor penyebab kecelakaan dengan keparahan korban [17]. Berbeda dengan *bagging*, namun masih menggunakan *Ordinal Logistic Regression*, penelitian ini menerapkan *Fuzzy K-Nearest Neighbor in Every Class* (FK-NNC) sebagai *classifier*. Cara kerja metode ini yaitu mencari tetangga terdekat setiap kelas dari data yang telah diolah. Data dalam hal ini yaitu data kecelakaan lalu lintas yang terjadi di Kota Semarang [18]. Dalam penelitian lainnya, klasifikasi yang digunakan adalah metode *boosting* yaitu metode yang melakukan kombonasi-kombinasi pada beberapa *classifier* atau dengan nama lain *ensamble* yang diterapkan di Kota Palu. Dengan menggunakan delapan buah variabel untuk memprediksi korban kecelakaan lalu lintas yang salah satunya merupakan variabel respon dan sisanya sebagai *predictor*. Dengan tujuan memberikan informasi kepada Kepolisian Kota Palu berdasarkan data kecelakaan lalu lintas pada tahun 2018 [19].

Oleh karena itu berdasarkan dari uraian diatas, maka dilakukan penelitian dengan menggunakan salah satu metode yang ada pada *data mining*. Tujuan dari penelitian ini untuk mengetahui desain model pada *Decision Tree* terintegrasi dengan metode SMOTE yang memiliki akurasi paling baik berdasarkan faktor-faktor yang mempengaruhi terjadinya kecelakaan lalu lintas di Australia Selatan, dengan melakukan pengujian akurasi menggunakan tiga desain model yaitu *Split Validation Decision Tree* dan SMOTE, *Cross Validation Decision Tree* dan SMOTE dan *Decision Tree* dan SMOTE *Split Data*. Hasil dari penelitian ini diharapkan bermanfaat untuk kepolisian atau instansi terkait dengan keselamatan pengguna jalan, transportasi dan lain – lain.

II. METODOLOGI

Metode penelitian memiliki tahap-tahap yang akan digunakan serta perancangan dalam melakukan implementasi metode *Decision Tree* untuk klasifikasi data kecelakaan lalu lintas di Australia Selatan. Tahapan metode penelitian ditunjukkan pada Gambar 1.



Gambar 1. Metode proses *Knowledge Discovery in Database*

Dengan menggunakan proses *Knowledge Discovery in Database* (KDD) terdapat tahap-tahap yang akan dilalui oleh data kecelakaan lalu lintas yaitu *selection, preprocessing, transformation, data mining* dan *evaluation*. Proses KDD merupakan *interative process*, dimana langkah evaluasi dapat ditingkatkan, disempurnakan dan mendapat hasil yang berbeda dan lebih tepat.

A. *Selection*

Tahap ini merupakan tahap yang pertama dalam penelitian, operasi yang dilakukan yaitu mengekstrak *dataset* yang didapatkan atau dikumpulkan berdasarkan hubungan yang lebih luas (*universal*) [20]. Dalam penelitian ini *dataset* yang akan dipilih adalah data sekunder yang telah disediakan oleh situs penyedia *repository*. Berikut langkah – langkah seleksi data kecelakaan [21]:

- Kunjungi *website* pemerintah Australia Selatan.
- Pilih menu *dataset*.
- Pilih menu *road crash data*.
- Unduh *dataset road crash*.
- Ubah format data *comma separated values* (CSV) ke dalam bentuk *excel*.

Selanjutnya, *dataset* akan diproyeksikan, diseleksi atau digabungkan apabila lebih dari satu *dataset*.

B. *Preprocessing*

Pada tahap *preprocessing*, operasi yang dilakukan adalah mendeteksi dan melakukan koreksi data-data yang salah, menghapus data yang berlebihan, serta menambahakan atau menggantikan data yang hilang. Sehingga ketika data telah dibersihkan, maka data tersebut akan memberikan peluang lebih besar untuk keberhasilan dalam data mining [20].

C. *Transformation*

Pada tahap transformasi, *dataset* yang sebelumnya telah melalui tahap seleksi dan *preprocessing* dapat dipakai dalam pengembangan dalam *data mining*, karena generasi data menjadi lebih baik. Langkah yang termasuk dalam transformasi yaitu transformasi atribut, pengurangan dimensi, gabungan data (*aggregation*) dan lain- lain [20]. Tahap ini menjadi sangat penting untuk keberlangsungan proses *Knowledge Discovery in Database* (KDD).

D. Data Mining

Pada *data mining*, tahap ini merupakan inti dari proses KDD. Operasi dari *data mining* akan megnekstraksi *dataset* kecelakaan lalu lintas [20]. Pilihan algoritma yang digunakan dalam penelitian ini adalah menggunakan algoritma *Decision Tree*, ekstraksi data akan ditampilkan dalam bentuk akurasi. Orientasi dari operasi ini adalah pemasangan model karena algoritma *Decision Tree* merupakan bagian dari *classification*. *Decision Tree* merupakan salah satu model klasifikasi yang dapat menggunakan variabel data *nominal* dan *numerical*. Di dalam *data mining* juga akan dilakukan klasifikasi data kecelakaan lalu lintas berdasarkan faktor yang mempengaruhinya.

Ada berbagai macam algoritma pada metode atau teknik *Decision Tree* yaitu algoritma ID3, algoritma *random forest*, algoritma CHAID, algoritma C4.5 dan lainnya. Khusus untuk algoritma C4.5 adalah salah satu algoritma yang sering dipakai karena merupakan pengembangan dari algoritma pendahulunya yaitu algoritma ID3. Berikut langkah-langkah algoritma C4.5 dalam membangun *Decision Tree* adalah sebagai berikut [22]:

- Pilih atribut sebagai akar.
- Buat cabang untuk tiap-tiap nilai.
- Bagi kasus dalam cabang.
- Ulangi proses untuk setip proses cabang sampai semua kasus pada cabang memiliki kelas yang sama.

Perbedaan antara algoritma C4.5 yang merupakan perbaikan dari algoritma ID3 terletak pada *pruning* yaitu pemangkasan *outlier* atau *noise* data untuk kebutuhan akurasi pada klasifikasi dan prediksi. Algoritma C4.5 menghitung *gain ratio* untuk masing-masing atribut, dan atribut yang memiliki nilai yang tertinggi akan dipilih sebagai simpul. Penggunaan gain ratio ini memperbaiki kelemahan dari ID3 yang menggunakan *information gain* [22].

$$Gain(S,A) = Entropy(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy(Si) \quad (1)$$

Keterangan:
S = Himpunan kasus
A = Atribut
n = Jumlah partisi atribut A
|Si| = Jumlah kasus pada partisi ke-i
|S| = Jumlah kasus dalam S

Sementara itu, perhitungan nilai entropi dapat dilihat pada persamaan 2 berikut ini:

$$Entropy(S) = -\sum_{i=1}^{n} pi * log_2\ pi \quad (2)$$

Keterangan:
S = Himpunan kasus
A = Atribut
n = Jumlah partisi S
pi = Proporsi dari Si terhadap S

E. Evaluation

Pada tahap terakhir yaitu evaluasi, dilakukan pengujian akurasi dan visualisasi dari data yang telah diekstrak dan diolah oleh algoritma *Decision Tree* dan SMOTE. Maka, diperlukan *tools* yang dapat menerapkan tahap ini. Pada penelitian ini, *tools* yang digunakan adalah *Rapidminer Studio* v.9.6. *Rapidminer* tidak hanya melakukan visualisasi data, namun juga menguji akurasi, presisi, serta klasifikasi benar dan klasifikasi yang salah.

Adapun akurasi merupakan rasio prediksi benar pada keseluruhan data.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (3)$$

Presisi adalah rasio prediksi benar positif (TP) yang dibandingkan dengan keseluruhan hasil yang diprediksi positif

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

III. HASIL DAN PEMBAHASAN

A. Selection

Data yang diperoleh untuk tujuan suatu penelitian dikelompokkan menjadi dua jenis sesuai dengan sumber data yaitu data primer dan data sekunder. Data yang diperoleh dari penelitian ini adalah data sekunder laporan tahun 2018 kecelakaan lalu lintas *The Department of Planning, Transport and Infrastructure* (DPTI) Australia Selatan. Jumlah data yang dihimpun yaitu 13.599 data [23]. Dengan menggunakan model klasifikasi *Decision Tree* dan algoritma SMOTE sebagai penyeimbang data dari kelas minoritas, diharapkan dapat membantu dan menjadi solusi untuk meminimalisir terjadinya kecelakaan di Australia Selatan.

B. Preprocessing

Sumber dari data yang telah diseleksi maka dilakukan *preprocessing*, pada penelitian ini maka kolom yang tidak dipakai untuk pengolahan data harus dihapus atau dibuang, karena pada beberapa kolom tersebut juga terdapat *missing value.*

TABEL I
ATRIBUT KECELAKAAN LALU LINTAS

Atribut	Variabel	Kategori
Area Speed	*Numerical*	-
Road Surface	*Nominal*	*Sealed Unsealed*
Moisture Condition	*Nominal*	*Dry Wet*
Weather Condition	*Nominal*	*Raining Not Raining*
DayNight	*Nominal*	*Daylight Night*
Entity Code	*Nominal*	*Driver Rider Passenger Pedestrian Animal*
Severity	*Nominal*	*Property Damage Only*

		Serious Injury *Minor Injury* *Fatal*
Traffic Control	*Nominal*	*No Control* *Traffic Signals* *Roundabout* *Give Way Sign* *Stop Sign*

Pada tabel I. terdapat 8 atribut yaitu *area speed, road surface, moisture condition, weather condition, daynight, entity code, severity* dan *traffic control*. *Severity* digunakan sebagai *class label* karena termasuk kedalam *polynominal* yaitu keterangan dari atribut lebih dari dua. Dalam *dataset* terdiri data dari satu atau lebih variabel. Pada penelitian ini variabel yang digunakan ada dua yaitu *numerical* dan *nominal*. Dari tujuh atribut pendukung tersebut, satu-satunya variabel data yang *numerical* adalah *area speed*. *Area speed* merupakan batas kecepatan pada waktu dan lokasi terjadinya kecelakaan. Berdasarkan *dataset*, tercatat bahwa kecepatan yang terjadi yaitu antara 5 sampai dengan 110 km/jam. Atribut kedua yaitu *road surface* yakni variabel data nominal yang merupakan jenis permukaan jalan saat terjadi kecelakaan. Atribut ketiga *moisture condition* yaitu kelembapan permukaan jalan. Dimana permukaan jalan kering atau basah yng bisa mebuat jalanan menjadi licin. Atribut keempat *weather condition* yaitu kondisi cuaca ketika terjadi kecelakaan. Cuaca hujan dapat mengakibatkan pengelihatan dari pengemudi tidak setajam ketika cuaca cerah. Atribut kelima *daynight* yaitu kondisi pencahayaan saat terjadi kecelakaan. Biasanya kecelakaan terjadi pada malam hari, karena pencahayaan yang minim. Atribut keenam *entity code* yaitu entitas yang terlibat dalam kecelakaan. Keterlibatan seperti pengemudi, penumpang, pejalan kaki, hewan yang melintasi jalan atau entitas lainnya yang dapat mengganggu pengemudi. Atribut yang terakhir *traffic control* yaitu kontrol lalu lintas saat terjadi kecelakaan. Lampu merah, kuning, hijau atau tanda kereta api akan lewat dapat menjadi faktor kecelakaan.

C. Transformation

Data kecelakaan lalu lintas telah yang melalui proses data *cleansing* dan *data integration* yang tergabung kedalam *preprocessing*, tahap selanjutnya yaitu *transformation.*

TABEL II

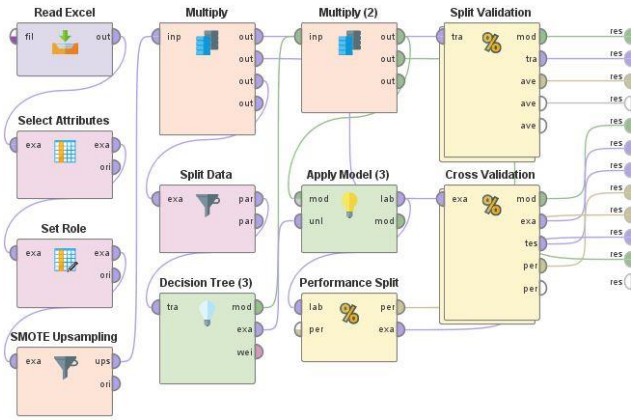
TRANSFORMASI ATRIBUT KELAS

Data Sebelum Transformasi				**Data Setelah Transformasi**	
Severity				***Crashes***	
Property Damage only	*Minor Injury*	*Serious Injury*	*Fatal*	*Property Damage Only*	*Casua-lties*

Pada tabel II. atribut *severity* (keparahan) kecelakaan lalu lintas yang dijadikan kelas dalam klasifikasi dan merupakan informasi awal data dengan memiliki kategori *property damage only* (kerusakan barang berharga), *minor injury* (korban luka ringan), *serious injury* (korban luka berat) dan *fatal* (kematian), diubah menjadi atribut *crashes* yang memiliki kategori *property damage only* (kerusakan barang berharga) dan *casualties* (korban). Transformasi ini bertujuan untuk melihat AUC pada model *Decision Tree*. AUC digunakan untuk menunjukkan akurasi dan membandingkan hasil klasifikasi secara visual.

D. Data Mining

Sebelum mengetahui hasil dan visualisasi dari penelitian berdasarkan faktor penyebab kecelakaan, langkah awal yang perlu dilakukan yaitu merancang model, agar bisa menampilkan akurasi, presisi dan juga hasil kemungkinan dalam bentuk grafik yang diolah oleh *tools rapidminer*. Berikut desain model yang ditampilkan pada gambar 2.



Gambar 2. Model proses pengujian

Pada gambar 2. *dataset* yang telah diperoleh diubah kedalam format *excel* dan membentuk sebuah tabel yang menampilkan kolom-kolom yang berisi atribut. Kemudian dilakukan pemilihan atribut berdasarkan faktor kecelakaan yang menyebabkan korban atau kerusakan barang berharga. Lalu pilih atribut *crashes* sebagai label. Selanjutnya, melakukan peningkatan kinerja pada model *Decision Tree* dengan menggunakan SMOTE *Upsampling*. Terakhir, penentuan akurasi dibagi menjadi 3 bagian, yaitu menggunakan *Split Validation Decision Tree* dan SMOTE, *Cross Validation Decision Tree* dan SMOTE dan *Decision Tree* dan SMOTE *Split Data* dan membandingkan desain model tersebut dengan mencari akurasi yang paling baik.

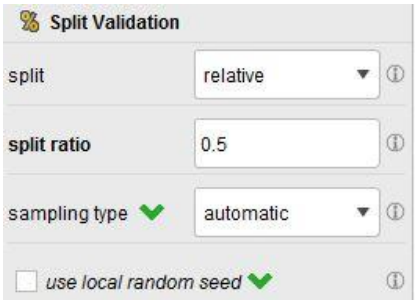
E. Evaluation

Penelitian ini bertujuan untuk mengetahui pengujian model desain yang terbaik dari integrasi metode klasifikasi *Decision Tree* dan teknik penyeimbang data yaitu SMOTE berdasarkan faktor yang mempengaruhi kecelakaan lalu lintas, sehingga akurasi menjadi meningkat. *Output* dari evaluasi ini berupa akurasi perhitungan data dan visualisasi data.

1) Split Validation Decision Tree dan SMOTE

Pengujian pertama dilakukan pada desain model *Split Validation Decision Tree* dan SMOTE dengan melakukan

validasi data dengan membagi data tesebut mejadi dua bagian, data bagian pertama bernama *data training* dan bagian kedua yaitu *data testing*. Perubahan dilakukan pada *split ratio*.



Gambar 3. Pembagian rasio split validation

Split ratio ini mejadi parameter yang spesifik pada *data training*. Perubahan rasio atau perbandingana dilakukan dalam rentang 0.1 (1:9) sampai dengan 0.9 (9:1).

TABEL III

PERBANDINGAN HASIL PENGUJIAN

Rasio	Akurasi
0.1 (1:9)	67.51%
0.2 (2:8)	68.88%
0.3 (3:7)	68.81%
0.4 (4:6)	68.99%
0.5 (5:5)	68.05%
0.6 (6:4)	69.23%
0.7 (7:3)	69.05%
0.8 (8:2)	68.26%
0.9 (9:1)	67.71%

Dari tabel III. diperoleh hasil pengujian data training dengan akurasi paling baik yaitu rasio 0.6 (6:4) yaitu 69.23%. Untuk melihat *confusion matrix* menggunakan *Split Validation Decision Tree* dan SMOTE ditampilkan pada tabel IV.

TABEL IV

CONFUSION MATRIX

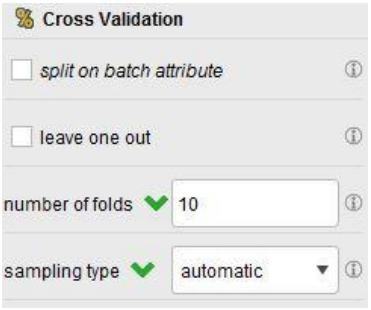
True:	***Property Damage Only***	***Casualties***
Property Damage Only:	3042	1735
Casualties:	357	1664

Confusion matrix digunakan untuk evaluasi klasifikasi model *Decision Tree* berdasarkan akurasi prediktif. Pada tabel didapatkan hasil sebagai berikut:

- Jumlah data positif dan terdeteksi benar atau *True Positive* (TP) yaitu 3042.
- Jumlah data negatif dan terdeteksi salah atau *False Positive* (FP) yaitu 1735.
- Jumlah data positif namun terdeteksi salah atau *False Negative* (FN) yaitu 357.
- Jumlah data negatif namun terdeteksi benar atau *True Negative* (TN) yaitu 1664.

2) Cross Validation Decision Tree dan SMOTE

Pengujian kedua dengan menggunakan desain model *Cross Validation Decision Tree* dan SMOTE dilakukan untuk menguji kinerja model klasifikasi yaitu *Decision Tree*. *Decision Tree* dilatih oleh dua buah subproses yaitu subproses *training* dan subproses *testing*.



Gambar 4. Jumah *fold cross validation*

Pada gambar 4. Angka 10 *fold* pada *cross validation* dipilih karena angka tersebut merupakan rekomendasi yang terbaik untuk model. Cara kerja nya yaitu dibagi menjadi 10 *fold* berarti terdapat 10 *subset data* untuk melakukan evaluasi pada *Decision Tree.*

TABEL V

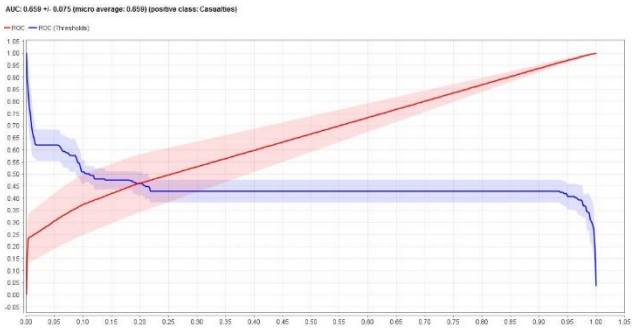
CONFUSION MATRIX

True:	***Property Damage Only***	***Casualties***
Property Damage Only:	7781	5476
Casualties:	717	3022

Berdasarkan tabel V. diperoleh *confusion matrix* yang terdapat pada model *Cross Validation Decision Tree* dan SMOTE. Pada tabel didapatkan hasil sebagai berikut:

- Jumlah data positif dan terdeteksi benar atau *True Positive* (TP) yaitu 7781.
- Jumlah data negatif dan terdeteksi salah atau *False*
- *Positive* (FP) yaitu 5476.
- Jumlah data positif namun terdeteksi salah atau *False Negative* (FN) yaitu 717.
- Jumlah data negatif namun terdeteksi benar atau *True Negative* (TN) yaitu 3022.

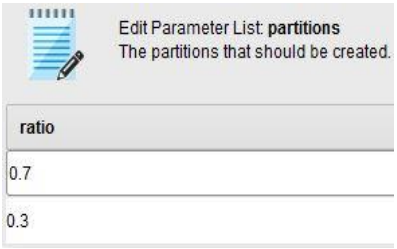
Perhitungan akurasi yang dihasilkan dari pemodelan *Cross Validation Decision Tree* dan SMOTE adalah 63.56%. Presisi yang yang dihasilkan yaitu 82.08%. Sementara itu untuk kurva ROC digunakan untuk menerangkan ketepatan keseluruhan dari data yang diuji, sehingga nilai AUC (*Area Under Curve*) sebesar 0.659. Pengujian pada desain model ini menunjukkan AUC yang tidak bisa mencapai 70% atau lebih untuk bisa masuk kedalam kategori *fair classification.*



Gambar 5. AUC *cross validation*

3) Decision Tree dan SMOTE Split Data

Pada pengujian dengan menggunakan model *Decision Tree* dan SMOTE *Split Data* dilakukan *split* yaitu membagi data menjadi 2 *subset* atau partisi.



Gambar 6. *Split data*

Split data merupakan parameter yang penting dalam model klasifikasi. Parameter ini menentukan jumlah dari partisi data dan rasio setiap partisi dengan pemilihan secara acak dan otomatis oleh *rapidminer*. Pada gambar rasio yang diuji antara 0.1 (1:9) hingga 0.9 (9:1).

TABEL VI

PERBANDINGAN HASIL PENGUJIAN

Rasio	Akurasi *Decision Tree*	Akurasi *Decision Tree* + SMOTE	Presisi *Decision Tree* + SMOTE	AUC *Decision Tree* + SMOTE
0.1 (1:9)	65.97%	71.12%	83.06%	0.773
0.2 (2:8)	65.45%	70.47%	87.58%	0.763
0.3 (3:7)	64.91%	70.44%	89.71%	0.757
0.4 (4:6)	64.69%	70.17%	88.36%	0.754
0.5 (5:5)	65.06%	70.43%	88.99%	0.755
0.6 (6:4)	64.55%	70.48%	89.40%	0.756
0.7 (7:3)	64.53%	70.53%	88.52%	0.759
0.8 (8:2)	64.68%	70.21%	84.09%	0.755
0.9 (9:1)	64.72%	70.05%	84.06%	0.751

Berdasarkan tabel VI. Pengujian yang dilakukan berdasarkan jumlah data pada tiap-tiap rasio. Sebagai perbandingan, dilakukan pengujian *Decision Tree* dengan *Decision Tree* + SMOTE. Hasil terbaik dari pengujian *Decision Tree* hanya mencapai 65.97%. Namun, dengan mengintegrasikan dengan SMOTE, akurasi meningkat lebih dari 5%. Maka akurasi yang terbaik yaitu 71.12% pada rasio 0.1 (1:9) yang memiliki presisi 83.06% dan AUC 0.773, serta jumlah *example data* yang digunakan yaitu 1.700 dari total 16.996 data yang telah dilakukan perhitungan menggunakan model SMOTE *Upsampling*.

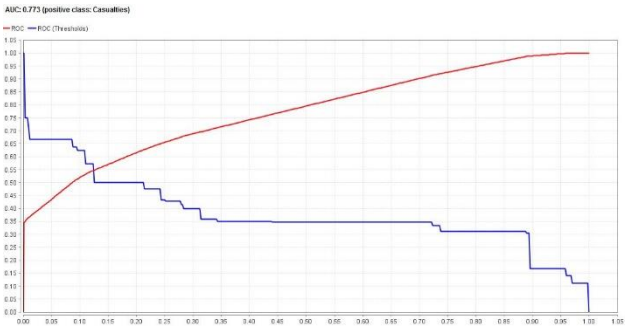
TABEL VII

CONFUSION MATRIX

True:	***Property Damage Only***	***Casualties***
Property Damage Only:	758	399
Casualties:	92	451

Confusion matrix pada perhitungan *Decision* Tree dan SMOTE *Split Data* diperoleh hasil sebagai berikut:

- Jumlah data positif dan terdeteksi benar atau *True Positive* (TP) yaitu 758.
- Jumlah data negatif dan terdeteksi salah atau *False Positive* (FP) yaitu 399.
- Jumlah data positif namun terdeteksi salah atau *False Negative* (FN) yaitu 92.
- Jumlah data negatif namun terdeteksi benar atau *True Negative* (TN) yaitu 451.



Gambar 7. AUC *split data*

Pada kurva ROC yang ada pada model *Decision Tree* dan SMOTE *Split Data*, nilai AUC yang diperoleh yaitu 0.773. Berdasarkan penilaian AUC, grafik tersebut termasuk kedalam kriteria *fair classification* (0.7 – 0.8) atau cukup. Berarti model *Decision Tree* layak digunakan pada *dataset* kecelakaan lalu lintas.

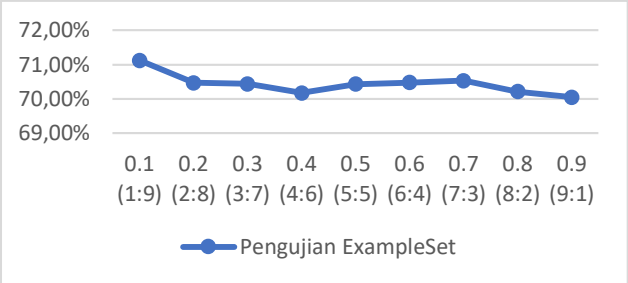
Berdasarkan hasil dari ketiga pengujian desain model, berikut perbandingan akurasi terbaik dari tiap-tiap model:

TABEL VIII

PERBANDINGAN DESAIN MODEL

No.	Model Pengujian	Akurasi
1	*Split Validation Decision Tree* dan SMOTE	69.23%
2	*Cross Validation Decision Tree* dan SMOTE	63.56%
3	*Decision Tree* dan SMOTE *Split Data*	71.12%

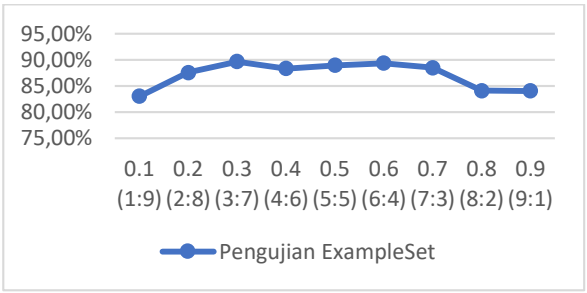
Perbandingan dari tiga desain model tersebut menunjukkan bahwa model *Decision Tree* dan SMOTE *Split Data* mendapatkan hasil akurasi yang paling baik. Berikut grafik dari akurasi, presisi dan AUC.



Gambar 8. Grafik pengujian akurasi

Berdasarkan grafik dari hasil pengujian menunjukan bahwa nilai tertinggi yang berjumlah 1.700 data (1:9) memiliki nilai akurasi tertinggi yaitu 71.12%.

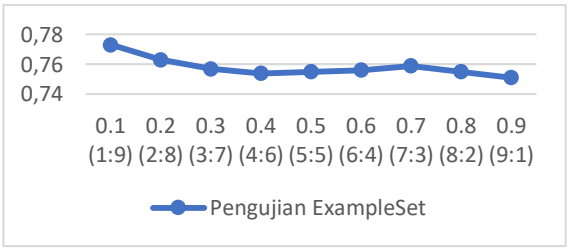
Untuk grafik dari pengujian presisi ditampilkan pada gambar 9.



Gambar 9. Grafik pengujian presisi

Berdasarkan grafik dari hasil pengujian menunjukan bahwa nilai tertinggi yang berjumlah 5.098 data (3:7) memiliki nilai presisi tertinggi yaitu 89.71%.

Grafik dari pengujian AUC ditampilkan pada gambar 10.



Gambar 10. Grafik pengujian AUC

Hasil pengujian menunjukan bahwa *example data* berjumlah 1.700 (1:9) memiliki nilai AUC tertinggi yaitu 0.773.

IV. KESIMPULAN

Dari hasil penelitian dapat disimpulkan bahwa, pengujian yang dilakukan dengan melakukan perbandingan desain model yaitu *Split Validation Decision Tree* dan SMOTE dengan akurasi 69.23%, *Cross Validation Decision Tree* dan SMOTE dengan akurasi 63.56%, menunjukkan dua pengujian tersebut masuk ke dalam kriteria *poor classification* atau buruk dan tidak layak untuk dijadikan model klasifikasi *dataset* pada penelitian ini. Sedangkan pengujian menggunakan desain model *Decision Tree dan* SMOTE *Split Data* dengan akurasi 71.12% (1:9), presisi 89.71 dan AUC 0.773, ini masuk ke dalam kriteria *fair classification* atau cukup. Kriteria ini merupakan yang terendah untuk bisa dikatakan layak pada penelitian klasifikasi dalam *data mining*.

Berdasarkan hasil penelitian, saran untuk pengembangan dari penelitian ini adalah dapat menggunakan metode klasifikasi *data mining* yang lain seperti metode klasifikasi *ensemble* yaitu menggabungkan beberapa metode sebagai solusi klasifikasi untuk mendapatkan hasil terbaik. Penelitian ini juga dapat dikembangkan dengan menambahkan kategori terkait kecelakaan lalu lintas.

DAFTAR PUSTAKA

[1] the G. of S. A. The Department of Planning, Transport and Infrastructure, "Road Crashes in South Australia Statistical Summary of Road Crashes & Casualties in 2018," 2018. [Online]. Available: https://www.dpti.sa.gov.au/towardszerotogether/road_crash_facts/sa_crashes#reports. [Accessed: 20-Apr-2020].

[2] A. D. Saputra, "Studi Tingkat Kecelakaan Lalu Lintas Jalan di Indonesia Berdasarkan Data KNKT (Komite Nasional Keselamatan Transportasi) dari Tahun 2007-2016," *War. Penelit. Perhub.*, 2018.

[3] WHO, "Road traffic injuries," 2020. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries. [Accessed: 20-May-2020].

[4] Y. Mardi, "Data Mining : Klasifikasi Menggunakan Algoritma C4.5," *J. Edik Inform.*, 2017.

[5] K. Fatmawati and A. P. Windarto, "Data Mining: Penerapan Rapidminer Dengan K-Means Cluster Pada Daerah Terjangkit Demam Berdarah Dengue (DBD) Berdasarkan Provinsi," *Comput. Eng. Sci. Syst. J.*, 2018.

[6] R. L. Hasanah, M. Hasan, W. E. Pangesti, F. F. Wati, and W. Gata, "Klasifikasi Penerima Dana Bantuan Desa Menggunakan Metode KNN (K-Nearest Neighbor)," *J. Techno Nusa Mandiri*, 2019.

[7] A. B. E. D. Ahmed and I. S. Elaraby, "Data Mining: A prediction for performance improvement using classification," *World J. Comput. Appl. Technol.*, 2014.

[8] F. Dwi Meliani Achmad, Budanis, Slamat, "Klasifikasi Data Karyawan Untuk Menentukan Jadwal Kerja Menggunakan Metode Decision Tree," *J. IPTEK*, 2012.

[9] A. Shiddiq, R. K. Niswatin, and I. N. Farida, "Ahmad Shiddiq Analisa Kepuasan Konsumen Menggunakan Klasifikasi Decision Tree Di Restoran Dapur Solo (Cabang Kediri)," *Gener. J.*, 2018.

[10] A. N. Kasanah, M. Muladi, and U. Pujianto, "Penerapan Teknik SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, 2019.

[11] A. Bisri and R. Rachmatika, "Integrasi Gradient Boosted Trees dengan SMOTE dan Bagging untuk Deteksi Kelulusan Mahasiswa," *J. Nas. Tek. Elektro dan Teknol. Inf.*, 2019.

[12] S. Kumar and D. Toshniwal, "A data mining approach to characterize road accident locations," *J. Mod. Transp.*, 2016.

[13] M. Taamneh, S. Taamneh, and S. Alkheder, "Clustering-based classification of road traffic accidents using hierarchical clustering and artificial neural networks," *Int. J. Inj. Contr. Saf. Promot.*, 2017.

[14] M. A. Raihan, M. Hossain, and T. Hasan, "Data mining in road crash analysis: the context of developing countries," *Int. J. Inj. Contr. Saf. Promot.*, 2018.

[15] R. Fitria, W. Nengsih, and D. H. Qudsi, "Implementasi Algoritma FP-Growth Dalam Penentuan Pola Hubungan Kecelakaan Lalu Lintas," *J. Sist. Inf.*, 2017.

[16] Y. Castro and Y. J. Kim, "Data mining on road safety: Factor assessment on vehicle accidents using classification models," *Int. J. Crashworthiness*, 2016.

[17] W. W. Fitriah and M. Mashuri, "Faktor-Faktor yang Mempengaruhi Keparahan Korban Kecelakaan Lalu Lintas di Kota Surabaya dengan Pendekatan Bagging Regresi Logistik Ordinal," *J. Sains dan Seni ITS*, 2012.

[18] C. Silvia, Y. Wilandari, and A. Hoyyi, "Ketepatan Klasifikasi Tingkat Keparahan Korban Kecelakaan Lalu Lintas Menggunakan Metode Reresi Logistik Ordinal dan Fuzzy K-Nearest Neighbor In Every Class," *None*, 2015.

[19] L. Susiana, I. T. Utami, and J. Junaidi, "Penerapan Metode Boosting Pada Cart Untuk Mengklasifikasikan Korban Kecelakaan Lalu Lintas Di Kota Palu," *Nat. Sci. J. Sci. Technol.*, 2019.

[20] P. Ristoski and H. Paulheim, "Semantic Web in data mining and knowledge discovery: A comprehensive survey," *Journal of Web Semantics*. 2016.

[21] F. Syukmana *et al.*, "Predicting Relegation Clubs in Italian Serie A with Method based C4.5 Decision Tree Algorithm," in *Journal of Physics: Conference Series*, 2020.

[22] A. Muzakir and R. A. Wulandari, "Model Data Mining sebagai Prediksi Penyakit Hipertensi Kehamilan dengan Teknik Decision Tree," *Sci. J. Informatics*, 2016.

[23] the G. of S. A. The Department of Planning, Transport and Infrastructure, "2018_DATA_SA_Crash," 2018. [Online]. Available: https://data.sa.gov.au/data/dataset/road-crash-data/resource/45ceb7e8-59bd-4492-b107-8379752ea597. [Accessed: 11-Apr-2020].