PENERAPAN DATA MINING MENGGUNAKAN POHON KEPUTUSAN DENGAN ALGORITMA C4.5 DALAM MENENTUKAN KECELAKAAN PENERBANGAN

Andreas Chandra

Teknik Informatika STMIK AMIKOM Yogyakarta
Jl Ring road Utara, Condongcatur, Sleman, Yogyakarta 55281
*Email : andreaschaandra@*yahoo.com

Abstrak

Penerbangan adalah hal yang sering dilakukan pada jaman yang sangat modern ini. Mobilitas masyarakat menuntut transportasi yang lebih cepat, namun setiap transportasi memiliki resiko masing-masing, karena semakin tingginya frekuensi penerbangan maka akan terdapat pula resiko kecelakaan pada transportasi pesawat. Paper ini menjelaskan tentang pohon keputusan dengan algoritma C4.5.

Pohon keputusan merupakan metode klasifikasi yang sangat kuat dan populer, metode pohon keputusan mengubah fakta yang sangat besar menjadi pohon keputusan yang merepresentasikan aturan. Pohon keputusan juga beguna untuk eksplorasi data, menemukan hubungan tersembunyi antara sejumlah calon variabel input dengan sebuat variabel target.

Pengujian yang dilakukan mengalami penyerdehanaan variabel dari dataset, variable yang digunakan hanya flighttype, investigation type, aircraft demange dan number of engines. Dan mengambil flight type sebagai hasil akhir.

Kata kunci: *C4.5, klasifikasi, penerbangan, pohon keputusan.*

1. Pendahuluan

Data mining dapat dilihat sebagai hasil dari evolusi alami dari teknologi informasi. Perkembangan basis data dan industri manajemen data memiliki beberapa fungsi penting diantaranya adalah *data collection and database creation*, *data management* serta *advanced data analysis*. Dalam *data mining*, perlu menemukan pengetahuan dalam bentuk pola yang nantinya akan diekstrak menjadi informasi yang akan bermanfaat untuk selnajutnya dilakukan interpretasi terhadap data tersebut.

Data mining dapat dilihat sebagai hasil dari evolusi alami dari teknologi informasi. Perkembangan basis data dan industri manajemen data memiliki beberapa fungsi penting diantaranya adalah *data collection and database creation*, *data management* serta *advanced data analysis*. Dalam *data mining*, perlu menemukan pengetahuan dalam bentuk pola yang nantinya akan diekstrak menjadi informasi yang akan bermanfaat untuk selnajutnya dilakukan interpretasi terhadap data tersebut.

Data mining dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan, yaitu deskripsi, estimasi, prediksi, klasifikasi, pengelompokan, dan asosiasi. Pohon keputusan termasuk kedalam klasifikasi. Klasifikasi merupakan suatu pekerjaan menilai objek data untuk memasukkannya kedalam kelas tertentu dari sejumlah kelas yang tersedia. Dalam klasifikasi ada dua pekerjaan uama yang dilakukan, yaitu (1) pembangunan model sebagai prototype untuk disimpan sebagai memori dan (2) penggunaan model tersebut untuk melakukan pengenalan/klasifikasi/prediksi pada suatu objek data lain agar diketahui di kelas mana objek data tersebut dalam model yang sudah disimpannya. [4] Teknik utama dalamdata mining yaitu klasifikasi dan prediksi, pengelompokan, *outlier detection* aturan asosiasi, *sequence analysis*, time series analisis dan *text mining*.[6]

Data Mining sering juga disebut *knowledge discovery in database* (KDD), adalah kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar. Keluaran dari *data mining* ini bisa dipakai untuk memperbaiki pengambilan keputusan dimasa depan. Sehinggan istilah *pattern recognition* sekarang jarang digunakan karena ia termasuk bagian dari *data mining*.[2]

Proses *knowledge discovery in database* (KDD) secara garis besar dapat dijelaskan sebagai berikut: (1) *Data Selection*, Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai. Data hasil seleksi yang akan digunakan untuk proses *data mining*. Disimpan dalam suatu berkas, terpisah dari basis data operasional. (2) *Pre-processing / Cleaning*, sebelum proses *data mining* dapat dilaksanakan, perlu dilakukan proses *cleaning* pada data yang menjadi fokus KDD. Proses *cleaning* mencakup antara lain membuat duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data, seperti kesalahan cetak (tipografi).

(3) *Transformation*, *Coding* adalah proses transformasi pada data yang telah dipilih, sehingga data tersebut

sesuai untuk proses *data mining*. Proses *coding* dalam KDD merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicar dalam basis data. (4) *Data Mining*, *data mining* adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode atau algoritma dalam *data mining* sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan. (5) *Interpretation / Evaluation*, pola informasi yang dihasilkan dari proses data mining perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagian dari poses KDD yang disebut *interpretation*. Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya.[3]

Teknik klasfikasi merupakan suatu pendekatan sistematis untuk membangun model klasifikasi dari suatu himpunan atau masukan. Tiap teknik menggunakan suatu algoritma pembelajaran untuk mendapatkan suatu model yang paling memenuhi hubungan antara himpunan atribut dan label klas dalam data masukan. Tujuan dari algoritma pembelajaran adalah utuk membangun model yang secara umum berkemampuan baik, yaitu model yang dapat memprediksi label kelas dari record yang tidak diketahui kelas sebelumna dengan akurat. [1] Sebuah pohon keputusan digunakan untuk mengklasifikasikan data dengan melewati rekor data ke simpul daun dari pohon keputusan menggunakan nilai dari variabel atribut dan menempatkan nilai target dari simpul daun untuk data pencatatan. [5]

2. Pembahasan

Pohon keputusan merupakan representasi sederhana dari teknik klasifikasi untuk sejumlah kelas berhingga, dimana simpul internal maupun simpul akar ditandai dengan nama atribut, rusuk-rusuknya diberi lebel nilai atribut yang mungkin dan simpul daun ditandai dengan kelas yang berbeda.[1]

Decision tree sesuai digunakan untuk kasus-kasus dimana outputnya bernilai diskrit. Walaupun banyak variasi model decision tree dengan tingkat kemampuan dan syarat yang berbeda, pada umumnya beberapa ciri kasusu berikut cocok untuk diterapkan *decision tree*[2] :

1. *Data/example* dinyatakan dengan pasangan atribut dan nilainya.
2. *Label/output* data biasanya bernilai diskri
3. *Data* mempunyai *missing value*.

Pohon keputusan juga berguna unuk mengeksplorasi data, menemukan hubungan tersembunyi antara sejumlah calon variabel input dengan sebuah variabel target. Pohon keputusan juga dapat diunakan untk mengestimasi nilai dasar variabel *continue* meskipun ada beberapa teknik yang lebih sesuai untuk kasus ini[3].

Secara umum algoritma C4.5 untuk membangun pohon keputusan adalah sebagai berikut.

a. Pilih atrbu sebagai akar.
b. Buat cabang untuk tiap-tiap nilai.
c. Bagi kasus dalam cabang.
d. Ulangi proses untuk setiap cabang sampai semua kasus pada cangan memiliki kelas yang sama.

Dataset penerbangan memiliki 41 variabel.

```
"cartodb_id"
"the_geom"
"amateurbuilt"
"field_1"
"flighttype"
"incidentlocation"
"weathercondition"
"schedule"
"reportstatus"
"registrationnumber"
"purposeofflight"
"model"
"make"
"location"
"investigationtype"
"injuryseverity"
"fardescription"
"eventid"
"enginetype"
"country"
"broadphaseofflight"
"airportname"
"airportcode"
"aircraftdamage"
"aircraftcategory"
"aircarrier"
"accidentnumber"
"severity_injured"
"severity_fatal"
"pctdeath"
"pctinjured"
"totalperson"
"totaluninjured"
"totalseriousinjuries"
"totalminorinjuries"
"totalfatalinjuries"
"numberofengines"
"longitude"
"latitude"
"publicationdate"
"eventdate"
```

Gambar 1.*Variabel dalam dataset*

Namun dataset penerbangan yang diuji dalam algoritma C4.5 pohon keputusan didapat dari ramdayz.carto.com. https://ramdayz.carto.com/tables/flights/public hanya 4

variabel yang digunakan. Data yang digunkan sebanyak 13777, dengan sampel di tabel 1.

Tabel 1.*Data Penerbangan*

No	Flight type	Investigation type	Aircraft damage	Num of Eng
1	Non-commercial	Accident	Substantial	1
2	Non-commercial	Accident	Substantial	1
3	Non-commercial	Accident	Substantial	1
4	Non-commercial	Accident	Substantial	1
5	Non-commercial	Accident	Substantial	1
6	Non-commercial	Accident	Substantial	1
7	Non-commercial	Accident	Substantial	1
8	Non-commercial	Accident	Substantial	1
9	Non-commercial	Accident	Substantial	1
10	Non-commercial	Accident	Substantial	1

Kesimpulan dataset terhadap variabel flighttype, investigationtype, aircraftdemage, numberofengines dapat dilihat sebagai berikut

```
Flighttype

Commercial     : 1023
Non-commercial:12754

Investigationtype

Accident:13469
Incident:  308

Aircraftdamage

Destroyed  : 1825
Minor      :  400
Substantial:11552

Numberofengines

Min.   : 1.000
1st Qu.: 1.000
Median : 1.000
Mean   : 1.132
3rd Qu.: 1.000
Max.  :24.000
```

Gambar 2 *Summary Data Penerbangan*

Setelah dilakukan proses data menggunakan algoritma C4.5 dengan fungsi J48 pada *package* RWeka.

```
=== Summary ===

Correctly Classified Instances
12872               93.4311 %

Incorrectly Classified Instances
905                 6.5689 %

Kappa statistic
0.2626

Mean absolute error
0.1088

Root mean squared error
0.2332

Relative absolute error
79.0865 %

Root relative squared error
88.9476 %

Total Number of Instances
13777


=== Confusion Matrix ===

     a     b   <-- classified as
   179   844 |     a = Commercial
    61 12693 |     b = Non-commercial
```

Gambar 3.*Summary Klasifikasi Algoritma C4.5*

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

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

Nilai gain dan entropy terhadap tipe penerbangan

Node			Jml Kasus
1	TOTAL		13777
	Investigation type		
		Accident	13469
		Incident	308
	Aircraft demage		
		Destroyed	1825
		Minor	400
		Substansial	11552
	number of engines		
		One	12075
		Two	1647
		Three	25
		Four	28
		Eightteen	1
		twentyfour	1

Gambar 4 *Nilai Gain Node 1*

Node			commercial	non-commercial
1	TOTAL		1023	12745
	Investigation type			
		Accident	874	12595
		Incident	149	159
	Aircraft demage			
		Destroyed	146	1679
		Minor	173	227
		Substansial	704	10848
	number of engines			
		One	419	11656
		Two	562	1085
		Three	25	0
		Four	17	11
		Eightteen	0	1
		twentyfour	0	1

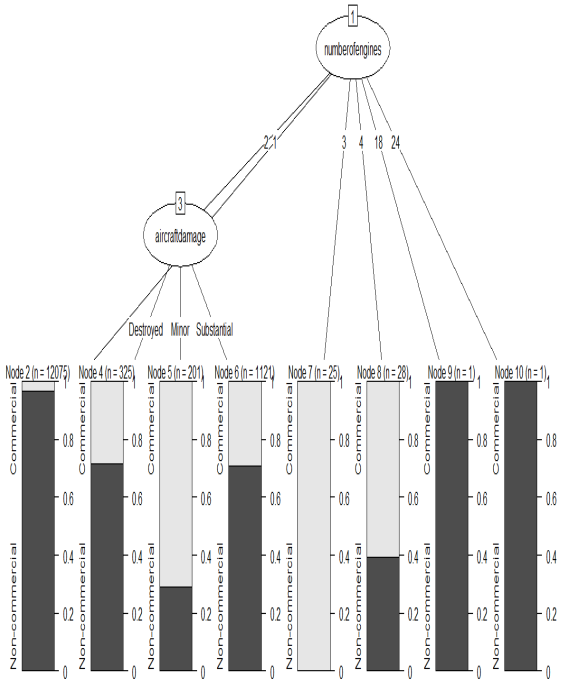
Gambar 5. *Jumlah variabel lain terhadap flight type*

Setalah mengkalkulasi jumlah jumlah terhadap flight type, kemudian menghitung nilai entropy dan gain. Nilai gain yang paling tinggi akan menjadi akar dalam pohon keputusan.

Node			Entropy	Gain
1	TOTAL		0.382472	
	Investigation type			0.02132347
		Accident	0.3465571	
		Incident	0.9992395	
	Aircraft demage			0.02285781
		Destroyed	0.4021792	
		Minor	0.9868132	
		Substansial	0.3311723	
	number of engines			0.07923136
		One	0.2174393	
		Two	0.9259875	
		Three	0	
		Four	0.9666186	
		Eightteen	0	
		twentyfour	0	

Gambar 6. *Nilai Entropy dan Gain*

Dari gambar diatas terlihat bahwa number of engines memiliki nilai gain yang paling tinggi. Oleh karena itu variable number of engines menjadi akar dalam pohon keputusan. Perhitungan diatas merupakan perhitungan pada node pertama. Setelah node berakhir dihasilkan pohon keputusan menyeluruh.
Pohon keputusan yang dibuat oleh algoritma C4.5 adalah sebagai berikut:



Gambar 7. *Pohon Keputusan*

3. Kesimpulan

Hasil yang di dapat dari algoritma C4.5 menggunakan dataset penerbangan terdapat 8 kolom dan ditiap kolom terdapat pilihan *Commercial* dan *Non-Commercial* dan

terdapat 4 level di pohon keputusan, penelitian ini diharapkan dapat menggambarkan pemilihan keputusan berdasarkan dataset yang digunakan sehingga mengetahui mana saja penerbanga yang mengalami kecelakaan.

Daftar Pustaka

[1] F. A. Hermawati, "Data Mining", ANDI, 2013
[2] B. Sentosa, "Data Mining Teknik Pemnafaatan Data untuk Keperluan Bisnis", Graha Ilmu, 2007.
[3] Kusrini, E. T. Luthfi, "Algoritma Data Mining", Penerbit ANDI, R. Frinkel, R. Taylor, R. Bolles, R. Paul, "An overview of AL, programming system for automation," *in Proc. Fourth Int. Join Conf Artif.Intel*., pp. 758-765, Sept. 3-7, 2006.
[4] A.S.R. Ansori, M. Hariadi, W. Endah,"Pemodelan Retakan Tiga Dimensi Akibat Ledakan Untuk Serious Games", *in Proc. Semnasteknomedia 2013*, pp.13-1, Januari 13,2013.

Biodata Penulis

Andreas Chandra, saat ini sedang menempuh pendidikan sarjana 1 program studi teknik informatika di STMIK Amikom Yogyakarta.

ISSN : 2302-3805