ANALISIS DAN DETEKSI FRAUD PADA DATA PANGGILAN MENGGUNAKAN ALGORITMA DECISION TREE PADA PT XYZ

ANALYSIS AND DETECTION OF FRAUD ON CALL DATA USING DECISION TREE ALGORITHM IN PT XYZ

Kurnia Dwi Wiranti Rahayu[1], Rachmadita Andreswari[2], Edi Sutoyo[3]
[1,2,3]Program Studi S1 Sistem Informasi,
Fakultas Rekayasa Industri, Universitas Telkom
kurniadwi@student.telkomuniversity.ac.id[1], andreswari@telkomuniversity.ac.id[2],
edisutoyo@telkomuniversity.ac.id[3]

Abstrak

Di era saat ini, perkembangan teknologi yang begitu cepat membuat banyaknya teknologi-teknologi canggih yang bermunculan. Salah satunya yaitu munculnya kecanggihan teknologi di bidang telekomunikasi. Telecom fraud merupakan suatu tindakan atau aktivitas penggunaan fasilitas telekomunikasi yang dilakukan secara ilegal dan disengaja dalam berbagai bentuk kecurangan, penipuan, atau pun penggelapan oleh orang atau organisasi tertentu yang tujuannya adalah untuk mendapatkan layanan tersebut dan menghindari biaya layanan atau pelacakan rekaman tagihan yang dilakukan secara ilegal. Salah satu jenis dari *telecom fraud* adalah SIM Box fraud. Untuk mengetahui adanya SIM Box Fraud, dapat dilakukan dengan cara melakukan analisis terhadap data nomor telepon milik salah satu provider. Metode yang digunakan adalah data mining dengan menggunakan algoritma Decision Tree – CART. Tahapan dalam melakukan deteksi dan prediksi dalam penelitian ini adalah preprocessing data, labelling, klasifikasi, dan evaluasi. Model yang dihasilkan diuji dan dievaluasi dengan melihat nilai akurasi, presisi, recall, dan f1-measure. Pada penelitian ini didapatkan akurasi tertinggi algoritma Decision Tree – CART. sebesar 95,6%.

Kata kunci: Nomor telepon, SIM Box Fraud, Decision Tree, Data Mining

Abstract

In the current era, the rapid development of technology makes emerging technologies emerge. One of them is technological sophistication in the field of telecommunications. Telecommunications fraud is one of the acts or activities that use telecommunications carried out illegally and intentionally in various forms of fraud, debate, or embezzlement by certain people or organizations intended to obtain these services and redeem payment services or payments made illegally. One type of telecommunications fraud is a SIM Box fraud. To find out about the SIM Box Fraud, it can be done by analyzing the telephone number data of one of the providers. The method used is data mining using the Decision Tree - CART algorithm. The stages in making detection and prediction in this study are data preprocessing, labeling, classification, and evaluation. The resulting model is tested and evaluated by looking at the value of accuracy, precision, recall, and f1-measure. In this study, the highest accuracy experiment Decision Tree - CART was obtained. by 95.6%.

Keyword: Customer Loyaty, Phone Number, SIM Box, Fraud, Decision Tree, Data Mining

1. Pendahuluan

Di era saat ini, perkembangan teknologi yang begitu cepat membuat banyaknya teknologi-teknologi canggih yang bermunculan. Salah satunya yaitu munculnya kecanggihan teknologi di bidang telekomunikasi. Telekomunikasi adalah teknik pengiriman atau penyampaian informasi dari satu tempat ke tempat lain. Pesatnya perkembangan telekomunikasi di Indonesia mengakibatkan hilangnya batas-batas jarak dan mereduksi perbedaan antara masyarakat di daerah perkotaan dengan perdesaan. Perbedaan waktu, lokasi, serta heterogenitas karakteristik penduduk tidak menjadi hambatan dalam kecepatan informasi saat ini.

Menurut *Cambridge Advanced Learner's Dictionary*, *fraud* adalah sebuah penipuan atau kecurangan yang disengaja yang dimaksudkan untuk mendapatkan keuntungan sementara penipuan dalam komunikasi dapat didefinisikan sebagai pencurian layanan dan penyalahgunaan suara serta jaringan data penyedia telekomunikasi [1]. Karena perkembangan teknologi modern yang telah meningkat baru-baru ini, kegiatan penipuan seperti kartu kredit *e-commerce* [2], [3], pencucian uang, penipuan perbankan online [4], intrusi komputer dan penipuan telekomunikasi juga telah meningkat secara dramatis yang mengakibatkan hilangnya uang dalam jumlah yang besar di seluruh dunia [5], [6]. *SIM Box Fraud* memiliki dampak yang berbeda pada operator telekomunikasi. Beberapa dampak

utamanya antara lain kerugian pendapatan karena terminasi panggilan, tidak dapat diaksesnya layanan dan *missing call backs*, dan kehilangan gambar karena kualitas layanan yang buruk [7].

Banyak cara atau metode yang dapat dilakukan untuk mendeteksi/mengklasifikasikan nomor telepon yang melakukan kecurangan, salah satunya dengan melakukan *data mining*. Metode *data mining* ini memanfaatkan data yang tersimpan pada sistem dari salah satu penyedia jasa dan jaringan telekomunikasi di Indonesia. Data yang dikelola oleh penyedia jasa dan jaringan telekomunikasi tersebut bisa diolah lalu divisualisasikan menggunakan metode *data mining*. *Data mining* adalah suatu proses yang menggunakan teknik matematika, statistika, kecerdasan buatan, dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai data yang besar [8].

Setelah mendapatkan data panggilan dari salah satu penyedia jasa layanan telekomunikasi, maka dilakukan data cleansing dan dilanjutkan dengan melakukan deteksi dan prediksi untuk mengetahui nomor telepon yang terindikasi melakukan kecurangan dengan menggunakan algoritma *Decision Tree*. Selain itu juga bisa mengetahui tingkat akurasi dari algoritma *Decision Tree* dalam melakukan klasifikasi dalam jumlah data yang besar.

2. Dasar teori dan metodologi

2.1 *SIM Box Fraud*

SIM Box adalah penyaluran *traffic incoming* international yang tidak melalui jalur normal interkoneksi internasional. Namun melalui jalur lain yang pada umumnya menggunakan teknologi *VoIP (Voice Over Internet Protocol)*. *SIM Box* merupakan sebuah perangkat yang mampu memanipulasi nomor ponsel dari pengguna ke penerima. Panggilan dari luar negeri bisa dimanipulasi seolah-olah menjadi panggilan dalam negeri.

Gambar 1 Cara Kerja SIM Box [9]

Cara kerja dari *SIM Box* yaitu seperti pada gambar II.1 di atas. Lokasi A merupakan lokasi yang berada di luar negeri, sedangkan lokasi B merupakan lokasi yang berada di Indonesia. *SIM Box* dikendalikan dengan menggunakan GUI. Sebelumnya, trafik dari luar negeri tersebut masuk melalui jaringan internet. Setelah itu dibutuhkan server *(SIM Box)* yang berguna untuk menyalurkan trafik ke seluruh perangkat yang ada. Sehingga jika tujuannya ke operator A, maka *server* akan menyalurkan ke *SIM Box* yang sudah ditentukan. Nomor panggilan yang akan muncul pada telepon penerima nantinya berupa nomor lokal, bukan menggunakan nomor luar negeri.

2.2 *Data Mining*

Data mining adalah suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan di dalam database. *Data mining* adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan dari berbagai basis data yang besar [8].

Menurut Gartner Group, *data mining* adalah suatu proses menemukan hubungan yang berarti, pola, dan kecenderungan dengan memeriksa dalam sekumpulan data yang besar yang tersimpan dalam penyimpanan dengan menggunakan teknik pengenalan pola seperti teknik statistik dan matematika [10].

2.3 Algoritma *Decision Tree*

Decision Tree merupakan suatu metode klasifikasi dan prediksi yang berbentuk seperti struktur pohon yang mana pada setiap *internal node* merupakan pengujian terhadap suatu atribut, setiap cabang menyatakan hasil dari pengujian tersebut dan *leaf node* menyatakan label pada setiap kelasnya. *Node* teratas dari *Decision Tree* ini disebut dengan *root* [11].

Dalam *Decision Tree*, terdapat beberapa metode, antara lain ID3, C4.5, dan CART [10]. Dan juga terdapat tigas jenis ukuran pemilihan atribut yang digunakan, yaitu *gain ratio, information ratio,* dan *gini* index. *Gain ratio* digunakan dalam metode C4.5, *information gain* digunakan untuk metode ID3, sedangkan *gini* index digunakan dalam metode CART. Selama penelitian ini, metode yang akan digunakan adalah CART (*Classification and Regression Trees*) yang menggunakan *Gini Index* sebagai metrik. *Gini index* salah satu ukuran pemilihan atribut. *Gini index* digunakan untuk menentukan

kemurnian dari suatu kelas tertentu yang terbelah disepanjang atribut tertentu. *Gini index* didefinisikan sebagai berikut:

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

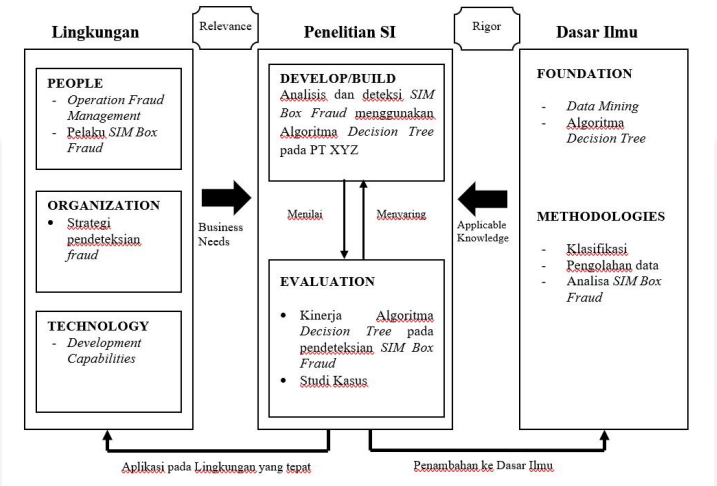
Dimana
Gini = *impurity* dari suatu partisi
C = banyaknya indeks
P_i = probabilitas suatu objek yang diklasifikasikan ke kelas tertentu.

CART mempunyai beberapa kelebihan dibandingkan dengan metode klasifikasi lainnya, antara lain hasil dari klasifikasi menggunakan metode ini lebih mudah untuk diinterpretasikan, lebih akurat, dan lebih cepat dalam perhitungannya. Selain itu, CART juga bisa diimplementasikan pada data-data dalam jumlah besar, variable yang sangat banyak dan dengan skala variable campuran melalui prosedur pemilihan biner [12].

3. Metodologi Penelitian

3.1 Model Konseptual

Model Konseptual ini merupakan model konseptual terhadap penelitian yang dilakukan, mencakup lingkungan, dasar ilmu, dan penelitian. Model konseptual pada penelitian ini dapat dilihat pada gambar dibawah ini:

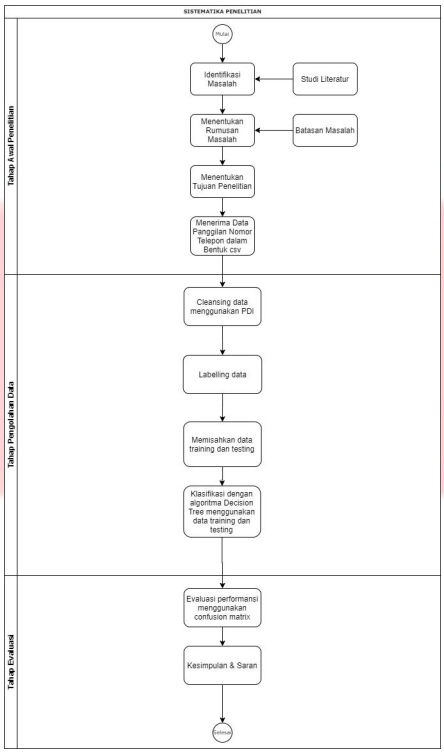


Gambar 2 **Model Konseptual**

Dalam penelitian ini melibatkan mencakup kasus pada sebuah data yang tersimpan pada unit *revenue assurance* PT XYZ. Pada unit tersebut memiliki histori data panggilan pada bulan Juni hingga Agustus tahun 2017, penelitian ini memanfaatkan metode *data mining* dengan menerapkan algoritma *Decision Tree* yang memanfaatkan *call data record* atau catatan data panggilan yang terekam pada sistem untuk diolah sehingga menghasilkan analisa dan pendeteksian nomor-nomor yang terindikasi sebagai *SIM Box Fraud*.

3.2 Sistematika Penelitian

Model Sistematika ini menunjukkan tahapan dalam proses penelitian. Sistematika penelitian dibagi menjadi tiga tahapan yaitu tahapan identifikasi awal penelitian, tahapan pengolahan data, dan tahapan evaluasi.



Gambar 3 Sistematika Penelitian

1. Tahap Awal Penelitian
   Pada tahap awal penelitian ini, hal pertama yang dilakukan yaitu menentukan latar belakang penelitian ini. Kemudian melakukan perumusan masalah berdasarkan latar belakang tersebut. Lalu menentukan tujuan dan manfaat dari penelitian ini serta menentukan batasan masalah agar pembahasan tidak keluar dari topik penelitian dan terlalu luas untuk dibahas sehingga akan memberikan hasil sesuai dengan yang diharapkan yaitu berupa deteksi cepat nomor-nomor yang terindikasi *SIM Box fraud* dengan menggunakan metode *data mining*.
2. Tahap Pengolahan Data
   Pada tahap metode ini telah didapatkan data panggilan (c*all detail record*) selama tiga bulan dalam bentuk *file* .csv dimana data tersebut digunakan dalam *data mining*. Selanjutnya dilakukan proses pembersihan data menggunakan *Pentaho Data Integration* untuk membuang atau menghapus data yang tidak dibutuhkan dalam penelitian ini. Setelah data sudah bersih, maka akan terotomatis masuk ke dalam *database* MySQL lalu di *export* dengan format CSV. Kemudian data di *import* ke dalam *Jupyter* dan dilakukan pelabelan data yaitu *fraud* atau *not fraud* berdasarkan parameter yang sudah ditentukan dari proses pembersihan data sebelumnya. Label data dikatakan *fraud* apabila A_NUMBER menelepon B_NUMBER yang berbeda-beda pada setiap melakukan panggilan. Selanjutnya akan dikatakan fraud juga apabila A_NUMBER melakukan aktivitas panggilan kepada B_NUMBER minimal 5 hari berturut-turut dan juga durasi diatas rata-rata yaitu 14,109 *ds*. Selanjutnya dilakukan klasifikasi dengan menggunakan algoritma *Decision Tree*.
3. Tahap Evaluasi
4. Pada tahap terakhir ini, data panggilan yang sudah diklasifikasi dan prediksi pada tahap sebelumnya akan dievaluasi menggunakan *confusion matrix* dan juga ditarik kesimpulan untuk menentukan nomor telepon yang benar-benar melakukan kecurangan *SIM Box*. Tujuannya agar dapat memberikan informasi yang berguna bagi perusahaan agar dapat mengurangi dampak dari *SIM Box fraud.*

4. Pembahasan

4.1 Dataset dan Labelling

Data yang digunakan pada penelitian ini adalah data nomor telepon yang melakukan panggilan pada salah satu penyedia jasa telekomunikasi di Indonesia yang berbentuk file dengan format csv. Jumlah data mentah yang digunakan sebanyak 6.985.327 yang merupakan total dari data selama bulan Juni hingga Agustus 2017. Rasio perbandingan yang antara data *training* dan *testing* pada penelitian ini adalah 90:10, 80:20, dan 75:25. Maksud dari salah satu rasio tersebut adalah 90% data akan dijadikan data *training* dan 10% data akan dijadikan data *testing*. Pada penelitian ini juga hanya menggunakan 5000 data yang bersifat *random* dari total data yang sudah dilakukan pelabelan *fraud* dan *not fraud*.

Tabel 1 Perbandingan Data training dan Data testing

Rasio (%)	Data *training*	Data *Testing*	Total Data
90:10	4500	500	5000
80:20	4000	1000	
75:25	3750	1250	

Pada penelitian ini, labelling dilakukan secara otomatis berdasarkan parameter yang sudah dipilih sebelumnya dengan menggunakan *tools* Jupyter Notebook dan bahasa pemrograman *Python*. Pelabelan dibagi menjadi 2 yaitu *fraud* dan *not fraud*. Berikut adalah contoh hasil pelabelan.

Nomor	Durasi	KESAMAAN_B_NUMBER	TOTAL_CALLING_TIME	Label
[redacted]	586288	16	1	Fraud
[redacted]	873469	17	1	Fraud
[redacted]	17232	0	0	Fraud
[redacted]	38084	0	0	Fraud
[redacted]	50141	0	0	Fraud
[redacted]	591811	3	1	Fraud
[redacted]	88423	2	0	Not Fraud
[redacted]	5911	0	0	Not Fraud
[redacted]	114592	2	0	Not Fraud
[redacted]	12712	0	0	Not Fraud
[redacted]	148950	2	0	Not Fraud

Gambar 4 Hasil labelling

4.2 Hasil Kategori Klasifikasi

Data yang sudah dilakukan labelling selanjutnya diklasifikasikan menggunakan algoritma *Decision Tree* pada *tools* Jupyter Notebook. Selanjutnya didapatkan hasil dari masing-masing rasio pada tabel berikut ini:

Tabel 2 Hasil Klasifikasi

Skenario	Ratio (%)	Data Training	Data Testing	Macro Average Precision	Macro Average Recall	Macro Average f1-Score	Accuracy
1	90:10	4500	500	85%	82%	83%	83%
2	**80:20**	**4000**	**1000**	**96%**	**96%**	**96%**	**95.6%**
3	75:25	3750	1250	88%	85%	85%	85.2%

Dari hasil data pada table diatas, data yang diklasifikasikan dengan rasio 80:20 menghasilkan akurasi algoritma *Decision Tree* tertinggi yaitu sebesar 95.6% dengan menggunakan *tools* Jupyter Notebook.

4.3 Evaluasi Performansi

Pengukuran evaluasi performansi diantaranya presisi, *recall*, dan *f1-measure*. Presisi merupakan suatu metode pengujian dengan melakukan perbandingan antara jumlah informasi yang relevan dengan jumlah seluruh informasi yang relevan maupun tidak. *Recall* merupakan suatu metode pengujian yang membandingkan antara jumlah informasi relevan dengan jumlah seluruh informasi relevan yang tersedia. *F1-Measure* merupakan suatu metode pengujian yang mengkombinasi antara rata-rata dari *precision* dan *recall* yang berbanding lurus dengan nilai keduanya.

Tabel 3 Hasil Evaluasi Performansi

No	Kategori Prediksi	Presisi	Recall	F1-Score
1	**Fraud**	94%	98%	96%
2	**Bukan Fraud**	98%	93%	95%
Macro Average		96%	96%	96%

5. Kesimpulan

Berdasarkan hasil penelitian yang dilakukan, kesimpulan yang didapatkan antara lain:

1. Dari penelitian yang telah dilakukan dapat disimpulkan bahwa tidak semua parameter dari *Call Data Record* bisa dijadikan sebagai parameter untuk mendeteksi dan memprediksi nomor telepon yang terindikasi sebagai *SIM Box Fraud*.
2. Dari pengujian yang dilakukan pada data panggilan bulan Juni hingga Agustus 2017 dengan menggunakan algoritma *Decision Tree* – CART, nilai akurasi tertinggi yaitu 95,6% dengan rasio 80:20, presisi 96%, *recall* 96%, *f1-measure* 96%. Hal ini bisa dikatakan bahwa algoritma *Decision Tree* – CART mampu melakukan prediksi yang baik pada data panggilan dengan jumlah data yang cukup besar.

DAFTAR PUSTAKA

[1] C. S. Hilas dan J. N. Sahalos, "User profiling for fraud detection in telecommunication networks," *5th Int. Conf. Technol. Autom.*, 2005.

[2] V. Ganji dan S. Mannem, "Credit card fraud detection using anti-k nearest neighbor algorithm," *Int. J. Comput. Sci. Eng.*, 2012.

[3] S. Jha, M. Guillen, dan J. Christopher Westland, "Employing transaction aggregation strategy to detect credit card fraud," *Expert Syst. Appl.*, 2012.

[4] W. Wei, J. Li, L. Cao, Y. Ou, dan J. Chen, "Effective detection of sophisticated online banking fraud on extremely imbalanced data," *World Wide Web*, 2013.

[5] R. J. Bolton dan D. J. Hand, "Statistical fraud detection: A review," *Statistical Science*. 2002.

[6] Y. Sahin, S. Bulkan, dan E. Duman, "A cost-sensitive decision tree approach for fraud detection," *Expert Syst. Appl.*, 2013.

[7] F. Mola, "Analysis and Detection Mechanisms of SIM Box Fraud in The Case of Ethio Telecom Thesis," 2017.

[8] E. Turban, J. Aronson, dan T. Llang, *Decision Support Systems and Intelligent Systems*. 2003.

[9] V. Airn, "Analysis and Detection of SIM box," *Int. J. Adv. Res. Ideas Innov. Technol.*, 2018.

[10] D. T. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining*. 2005.

[11] X. Wu *et al.*, "Top 10 algorithms in data mining," *Knowl. Inf. Syst.*, 2008.

[12] R. J. Lewis, D. Ph, dan W. C. Street, "An Introduction to Classification and Regression Tree (CART) Analysis," *2000 Annu. Meet. Soc. Acad. Emerg. Med.*, 2000.

.