ANALISIS DAN DETEKSI *FRAUD* PADA DATA PANGGILAN MENGGUNAKAN ALGORITMA *NAÏVE BAYES* PADA PT XYZ

ANALYSIS AND DETECTION FRAUD ON DATA CALL USING NAÏVE BAYES ALGORITHM IN PT XYZ

Ilna Nuzla Putri[1], Rachmadita Andeswari[2] , Edi Sutoyo[3]
[1,2,3]Program Studi S1 Sistem Informasi,
Fakultas Rekayasa Industri, Universitas Telkom
ilnanuzlaputri@telkomuniversity.ac.id[1], andreswari@telkomuniversity.ac.id[2], edisutoyo@telkomuniversity.ac.id[3]

Abstrak

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 maupun suatu organisasi tertentu yang tujuannya agar mendapatkan layanan tersebut serta menghindari biaya layanan dan pelacakan rekaman tagihan yang dilakukan secara ilegal.

Tujuan penelitian ini yaitu mendeteksi nomor panggilan yang terdeteksi sebagai SIM*Box fraud* yang telah merugikan pihak PT XYZ yang mempunyai tugas dalam menangani masalah *fraud* tersebut. Penelitian ini dilakukan dengan menggunakan *data mining* dan menggunakan algoritma *naïve bayes.*

Data mining merupakan suatu teknik yang memanfaatkan data dalam jumlah besar agar dapat memperoleh informasi berharga yang dapat dimanfaatkan dalam pengambilan keputusan penting. *Naïve Bayes* merupakan algoritma yang dapat digunakan untuk memprediksi nomor telepon yang terdeteksi sebagai SIM*Box fraud* yang bisa dikategorikan sebagai *fraud* dan *Not fraud.* Dengan menggunakan *data mining,* khususnya pada klasifikasi untuk prediksi menggunakan algoritma na*ïve bayes* dapat dilakukan prediksi terhadap nomor panggilan telepon yang terdeteksi sebagai *fraud* dari data panggilan. Hasil pengujian dengan menggunakan algoritma *naïve bayes* menunjukkan nilai akurasi tertinggi adalah 87.2% dengan nilai *macro average precision* adalah 90%, nilai *macro average recall* adalah 86% dan *macro average f1-score* adalah 87%. Sedangkan nilai akurasi yang paling rendah adalah 85.2%. Yang berarti menunjukkan bahwa implementasi algoritma *naïve bayes* merupakan salah satu algoritma terbaik untuk diterapkan dalam memprediksi data panggilan *fraud* pada PT XYZ.

Kata kunci: *Telecom fraud, Data Mining,* Algoritma *Naïve Bayes,* prediksi SIM*Box fraud*

Abstract

Telecommunications fraud is one of the acts or activities that use telecommunications that is carried out illegally and intentionally in various forms of fraud, payment or incorporation by certain people or organizations that require services that are accompanied by service fees and payments made with assistance carried out illegally.

The purpose of this study is to submit a number of calls submitted as SIMBox fraud owned by PT XYZ who have the task in the matter of fraud. This research was conducted using data mining and using the naïve Bayes algorithm.

Data mining is a technique that utilizes large amounts of data to obtain information that can be used in making important decisions. Naïve Bayes is an algorithm that can be used to predict phone numbers that are predicted to be SIMBox fraud which can be categorized as fraud and not fraud. By using data mining, specifically in the classification for predictions using the naïve Bayes algorithm, predictions can be made on telephone numbers called fraud from call data. The test results using the Naïve Bayes algorithm an average value of 87.2% with an average macro precision value of 90%, the average value of macro recall is 86% and the average macro f1-score is 87%. While the lowest value is 85.2%. What is meant by the application of the naïve Bayes algorithm is one of the best algorithms to be applied in predicting fraudulent call data at PT XYZ.

Keyword: *Telecom fraud, Data Mining, Naïve Bayes Algorithm, SIM Box fraud prediction*

1. Pendahuluan

Pada perkembangan tekonologi yang semakin modern, banyak orang yang ingin melakukan penipuan ataupun kecurangan untuk mendapatkan keuntungan yang merugikan banyak pihak [1]. Menurut *Association of Certified Fraud Examiners* (ACFE, 2020), *"Fraud is the use one's occupation for personal enrichment through the deliberate misuse or application of the employing organization's resources or assets"* yang artinya adalah "suatu tindakan untuk memperkaya diri melalui penyalahgunaan yang dilakukan secara sengata atau penggunaan sumber daya organisasi atau asset-asetnya [2] Secara umum, *fraud* merupakan suatu tindakan penggunaan fasilitas telekomunikasi secara ilegal yang sengaja melakukannya dengan berbagai cara bentuk kecurangan, penipuan ataupun juga penggelapan oleh orang maupun perusahaan tertentu yang tujuannya adalah untuk menghindari biaya layanan atau pelacakan rekaman tagihan yang dilakukan secara ilegal [3]. Saat ini, banyak orang yang memanfaatkan SIM *Box*

agar bisa memanipulasi nomor-nomor yang berasal dari luar negeri yang seharusnya menggunakan *roaming internasional*, tetapi dengan menggunakan SIM *Box* hitungannya menjadi panggilan lokal [4].

Menurut hasil monitoring perangkat Pos dan Informatika dari Dirjen SDPPI (Direktorat Jenderal Sumber Daya dan Perangkat Pos dan Informatika) Kominfo, SIM*Box fraud* 2.26% sebagai alat dan perangkat telekomunikasi *illegal*. Dampak dari SIM*Box fraud* bagi perusahaan adalah kerugian pendapatan (*revenue loss*) dan *security compromise* [5]. Kecurangan atas *fraud* sudah sangat meningkat, dengan meningkatnya kemajuan teknologi. Oleh karena itu, untuk mendeteksi data kecurangan SIM *Box* tersebut dapat menggunakan proses *data mining* [7]. *Data mining* merupakan pengumpulan data, pemakaian data historis untuk menentukan keteraturan, pola atau hubungan dalam data berukuran besar [8].

Untuk melakukan deteksi SIM*Box fraud* dapat dilakukan dengan berbagai algoritma, tetapi algoritma yang digunakan pada penelitian ini ini adalah algoritma *Naïve Bayes. Naïve Bayes* adalah model prediksi yang digunakan untuk menghasilkan model klasifikasi. Hasil penelitian ini adalah untuk mengukur performa dari algoritma *Naïve Bayes* dalam mendeteksi nomor *fraud* pada data panggilan.

2. Dasar teori

2.1 *Telecom Fraud*

Telecom Fraud adalah pencurian layanan telekomunikasi (telepon, ponsel, komputer) atau penggunaan layanan telekomunikasi untuk melakukan bentuk penipuan lainnya. Para korban termasuk konsumen, bisnis, dan penyedia layanan komunikasi [28].

2.2 SIM *Box Fraud*

SIM *Box* adalah sebuah perangkat yang mampu memanipulasi nomor ponsel dari pengguna ke penerima. Penipuan SIM *Box* hanya mengalihkan panggilan yang dikenakan biaya pengakhiran dan juga penipu SIM *Box* dapat menghentikan jaringan panggilan di jaringan perusahaan [5].

2.3 *Call Detail Record* (CDR)

Call Detail Record (CDR) merupakan data yang mencatat transaksi yang dilakukan oleh pengguna jasa telekomunikasi. CDR mengidentifikasi panggilan yang dilakukan atau diterima, dengan menyediakan beberapa data yaitu tanggal panggilan, waktu Panggilan dimulai dan berakhir, durasi panggilan, nomor panggilan masuk dan menerima panggilan serta biaya tol yang ditambahkan melalui jaringan atau biaya untuk layanan operator [22].

2.4 *Big Data*

Big data merupakan istilah yang diberikan pada kumpulan data yang berukuran sangat besar dan kompleks, sehingga tidak memungkinkan untuk diproses menggunakan perangkat pengelola *database* konvensional maupun aplikasi pemroses data lainnya [23].

2.5 *Pentaho Data Integration* (PDI)

Pentaho adalah kumpulan aplikasi *Business Intelligence* yang bersifat *free open source software* (FOSS) dan berjalan di atas platform Java. *Pentaho Data Integration* (PDI) atau *Kettle* merupakan *software Open Source* dari Pentaho yang dapat digunakan untuk mengintegrasikan data [19].

2.6 *Python*

Python merupakan bahasa pemrograman yang bersifat interpretatif. Dibandingkan dengan bahasa pemrograman lainnya, Python dianggap sebagai bahasa pemrograman yang menjanjikan peluang dimasa depan setelah Java [24]. Bahasa Pemrograman Python bersifat *open source.* Bahasa pemrograman ini dioptimalisasikan untuk *software quality, developer productivity, program portability,* dan *component integration* [25].

2.7 Data Mining

Data mining adalah suatu proses penerapan metode untuk menemukan nilai tambah dan informasi dari kumpulan data dengan jumlah yang besar [10].

Klasifikasi adalah jenis data yang membantu dalam memprediksi label kelas sampel sampai harus diklasifikasikan. Dalam klasifikasi terdiri dari data *training* dan data *testing.* Data *training* digunakan untuk membentuk sebuah model *classifier.* Sedangkan data *testing* digunakan untuk mengukur sejauh mana *classifier* berhasil melakukan klasifikasi dengan benar [1].

2.8 Algoritma *Naïve Bayes*

Naïve Bayes Classifier (NBC) merupakan sebuah metoda klasifikasi yang berakar pada teorema *Bayes*. Metode pengklasifikasian dengan menggunakan metode probabilitas dan statistik yang dikemukakan oleh ilmuan Inggris Thomas *Bayes*, yaitu memprediksi peluang yang ada di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai Teorema Bayes. Ciri yang signifikan dari *Naïve Bayes* adalah asumsi yang sangat kuat (naif) akan independensi dari masing-masing kejadian [11]. Persamaan dari teorema bayes adalah:

$$P\ (H|X) = \frac{P(X|H)P(H)}{\mathrm{P(X)}} \quad (1)$$

Keterangan:

X	:	Data dengan parameter yang belum diketahui.
H	:	Hipotesis keputusan
P(H|X)	:	Probabilitas keputusan *H* berdasarkan kondisi parameter X
P(H)	:	Probabilitas jumlah keputusan *H*
P(X|H)	:	Probabilitas parameter *X* berdasarkan kondisi keputusan H
P(X)	:	Probabilitas Parameter (*X*)

Keuntungan dari penggunaan metode ini hanya membutuhkan jumlah data pelatihan (*training data*) yg kecil untuk menentukan estimasi parameter yang diperlukan dalam proses pengklasifikasian [30]. Kelebihan dari penerapan algoritma pengklasifikasi *Naïve Bayes* adalah dapat mengurangi data noise pada dataset yang berukuran besar [29]:

Alur kerja algoritma *naïve bayes* adalah sebagai berikut [13]:

1. Baca data *training*
2. Menghitung jumlah probabilitas, apabila kondisi data adalah numerik, maka:
   a. Mencari nilai *mean* dan *standar deviasi* pada masing-masing parameter yang merupakan data numerik.
   b. Mencari nilai probabilitas yaitu dengan cara menghitung jumlah data yang sesuai dari kategori yang sama kemudian dibagi dengan jumlah data pada kategori tersebut.
3. Setelah itu akan mendapatkan nilai dalam table mean, standar deviasi dan probabilitas.

Terdapat 3 macam *Naïve Bayes*, diantaranya adalah [13]:

1. *Bernoulli Naïve Bayes*
   Bernoulli Naïve Bayes merupakan algoritma yang digunakan untuk deteksi data berupa *binary* (0 dan 1).
2. *Gaussian Naïve Bayes*
   Gaussian Naïve Bayes adalah algoritma yang digunakan untuk mendeteksi data berupa diskrit dengan menggunakan distribusi normal.
3. *Multinomial Naïve Bayes*
   Multinomial Naïve Bayes merupakan sebuah dokumen vektor bilangan bulat yang unsur-unsurnya menunjukkan *term frequency* yang sesuai dalam dokumen.

Proses klasifikasi algoritma didasarkan pada empat komponen, yaitu sebagai berikut [27]:

a. Kelas, yaitu *variable* dependen yang mempresentasikan label yang terdapat pada objek.
b. *Predictor*, yaitu *variable* yang direpresentasikan oleh karakteristik (parameter) data.
c. *Training dataset*, merupakan satu set data yang berisi nilai dari kedua komponen diatas yang digunakan untuk menentukan kelas yang cocok berdasarkan predictor.
d. *Testing dataset*, merupakan data baru yang akan diklasifikasikan oleh model yang telah dibuat dan akurasi klasifikasi dievaluasi.

2.9 *Confusion Matrix*

Confusion Matrix merupakan suatu metode yang digunakan untuk mengukur kinerja pada suatu metode klasifikasi. Hasil prediksi akan dibandingkan dengan kelas asli. *Confusion Matrix* mengevaluasi kinerja model klasifikasi berdasarkan pada kemampuan akurasi prediktif suatu model. *Confusion Matrix* juga memberikan keputusan yang diperoleh dalam *training* dan *testing* dan memberikan penilaian dalam *performance* klasifikasi berdasarkan objek dengan benar atau salah [14]. *Confusion Matrix* dapat dilihat pada Tabel 1.

Tabel 1 Confusion Matrix

Classification		***Predicted Class***	
		Positive	**Negative**
Actual Class	**Positive**	A (True Positive = TP)	B (False Positive = FP)
	Negative	C (False Negative) = FN)	D (True Negative= TN)

Keterangan:

TP = Prediksi positif yang positif
FN = Prediksi positif yang negatif
FP = Prediksi negatif yang positif
TN = Prediksi negatif yang negatif

Dari *confusion Matrix* pada Tabel I dapat dilakukan perhitungan lebih lanjut untuk mendapatkan persamaan sebagai berikut:

1. *Precision*
   Precision adalah tingkat ketepatan hasil dari klasifikasi dan jumlah keseluruhan pengenalan yang dilakukan sistem. Perhitungan *precision* dapat dirumuskan sebagai berikut:

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

   Dimana, TP merupakan *True Positive*, FN merupakan *False Negative.*
2. *Recall*
   Recall dinyatakan dalam jumlah data yang benar diklasifikasi dalam sebuah kelas dibagi dengan jumlah total dalam kelas tersebut. Perhitungan *recall* dapat dirumuskan sebagai berikut:

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

   Dimana, TP merupakan *True Positive*, FP merupakan *False Positive.*
3. F1-*Score*
   F1 *score* digunakan untuk mengevaluasi rata-rata *precision* dan *recall* hasil klasifikasi. Perhitungan F1 *Score* dapat dirumuskan sebagai berikut:

$$\text{F1 } score = 2 * \frac{(Recall * Precision)}{(Recall + Precision)} \quad (4)$$

   Dimana, Perhitungannya adalah 2 dikali dengan *precision* dikali dengan hasil *recall* dan dibagi dengan jumlah *precision* ditambah *recall.*
4. *Accuracy*
   Accuracy adalah jumlah data yang diklasifikasikan benar dibagi jumlah keseluruhan data. Perhitungan *accuracy* dapat dirumuskan sebagai berikut:

$$Accuracy = \frac{TP + TN\ (All\ True)}{TP + TN + FP + FN} \quad (5)$$

Umumnya, model yang dibangun dapat memprediksi dengan benar pada semua data yang di latihnya, akan tetapi ketika model berhadapan dengan data uji, barulah kinerja model dari sebuah klasifikasi ditentukan [20].

3. Pembahasan

3.1 Data Cleansing

Data yang diperoleh masih mengandung beberapa data yang tidak diperlukan sehingga, diperlukan *cleansing data* untuk menghilangkan data yang tidak diperlukan [18]. Pada penelitian ini memerlukan beberapa parameter saja yang digunakan untuk mendeteksi SIM *Box fraud.* Sehingga parameter yang tidak digunakan harus dihilangkan.

3.2 Labelling Data

Labelling Data bertujuan untuk membagi data ke dalam dua kelas, yaitu *fraud* dan *Not fraud.* Dataset yang digunakan pada penelitian ini, merupakan data panggilan pada bulan Juni, Juli dan Agustus tahun 2017 sebanyak 6.985.327 data yang belum memiliki label.

Labelling data pada data panggilan bulan Juni, Juli dan Agustus tahun 2017 dilakukan dengan menggunakan beberapa parameter yang sangat mempengaruhi dalam melakukan pendeteksian *SIM Box fraud*. *Labelling data* dilakukan berdasarkan 3 parameter, yaitu:

1. Nomor panggilan yang menerima panggilan telepon (*B_Number*)
   Berdasarkan parameter *B_Number*, apabila nomor yang melakukan panggilan telepon (*A_Number*) melakukan panggilan ke nomor (*B_Number*) dengan beberapa nomor yang sama, maka nomor tersebut merupakan nomor yang terdeteksi sebagai *Not fraud*. Beda hal nya dengan nomor yang merupakan *fraud*, artinya *A_Number* melakukan panggilan ke nomor *B_Number* dengan nomor yang berbeda-beda.
2. Lama panggilan yang dilakukan (*Duration*)
   Berdasarkan parameter *Duration*, rata-rata durasi yang dihasilkan pada data panggilan selama bulan Juni, Juli dan Agustus tahun 2017 adalah sebesar 14109.496 *desisecond*. Apabila durasi panggilan yang dilakukan diatas rata-rata, maka nomor telepon tersebut dideteksi sebagai SIM *Box fraud*. Sedangkan apabila durasi panggilan totalnya dibawah rata-rata, maka nomor telepon tersebut

dideteksi sebagai *Not fraud*. Total rata-rata durasi panggilan yang dihasilkan setiap bulan nya akan selalu berbeda-beda sesuai dengan lama panggilan yang dilakukan.

3. Waktu panggilan yang dilakukan (*Calling_Time*).
   Berdasarkan parameter *Calling_Time*, apabila nomor telepon melakukan panggilan selama 5 hari berturut-turut maka nomor tersebut dideteksi sebagai SIM *Box fraud*. Beda hal nya dengan nomor yang *Not fraud*, apabila nomor telepon tersebut tidak melakukan panggilan kurang dari 5 hari berturut-turut maka nomor tersebut dideteksi sebagai *Not fraud*

Berikut merupakan contoh hasil *labelling data* yang telah dilakukan berdasarkan 3 parameter yang mempengaruhi *fraud* dan *Not fraud* yang dapat dilihat pada Tabel IV.8.

Tabel 2 Hasil Labelling Data

No	Nomor Telepon	Durasi	Kesamaan B_Number	Total_ Calling _Time	Tanggal Calling_ Time	Jumlah Urutan Calling_ Time	Label
1	318538809	106758	4	0	[ ]	[ ]	*Not Fraud*
2	318962338	311028	2	0	[ ]	[ ]	*Not Fraud*
3	361702106	3670	0	0	[ ]	[ ]	*Not Fraud*
4	315680083	6853	1	0	[ ]	[ ]	*Not Fraud*
5	361811344	1327551	3	4	[28/06/2017, 19/07/2017, 31/07/2017, 10/08/2017]	[5, 9, 5, 5]	*Fraud*
6	341410375	157101	4	0	[ ]	[ ]	*Not Fraud*
7	342801708	2068	0	0	[ ]	[ ]	*Not Fraud*
8	317508300	19226	0	0	[ ]	[ ]	*Fraud*
9	341320905	47868	0	0	[ ]	[ ]	*Fraud*
10	3707508210	17783	0	0	[ ]	[ ]	*Fraud*
11	318538851	587792	17	1	[24/07/2017]	[5]	*Fraud*
12	3159173459	8173503	0	1	[22/07/2017]	[13]	*Fraud*
13	3158208800	8433878	173	5	[04/06/2017, 01/07/2017, 10/07/2017, 23/07/2017, 07/08/2017]	[21, 7, 11, 13, 9]	*Fraud*
14	341494739	323518	5	0	[ ]	[ ]	*Not Fraud*
15	315047662	24416	1	0	[ ]	[ ]	*Not Fraud*

3.3 Klasifikasi

Klasifikasi dapat membangun suatu model yang mampu mengklasifikasikan suatu objek berdasarkan parameter-parameternya. Dalam klasifikasi terdiri dari *data training* dan *data testing. Data training* digunakan untuk membentuk sebuah model *classifier*. Sedangkan data *testing* digunakan untuk mengukur sejauh mana *classifier* berhasil melakukan klasifikasi dengan benar

3.4 Hasil Prediksi Data Panggilan

Hasil prediksi data panggilan bulan Juni, Juli dan Agustus tahun 2017 menggunakan algoritma *Bernoulli Naïve bayes* dapat ditunjukkan pada Tabel 3.

Tabel 3 Hasil Prediksi Data Panggilan

No	Nomor Telepon	Durasi	Kesamaan B_Number	Total_ Calling_ Time	True Label	Predicted Label
1	318[redacted]809	106758	4	0	*Not Fraud*	*Not Fraud*
2	318[redacted]338	311028	2	0	*Not Fraud*	*Not Fraud*
3	361[redacted]106	3670	0	0	*Not Fraud*	*Fraud*
4	315[redacted]083	6853	1	0	*Not Fraud*	*Not Fraud*
5	361[redacted]344	1327551	3	4	*Fraud*	*Fraud*
6	341[redacted]375	157101	4	0	*Not Fraud*	*Not Fraud*

No	Nomor Telepon	Durasi	Kesamaan B_Number	Total_ Calling_ Time	True Label	Predicted Label
7	342 708	2068	0	0	*Not Fraud*	*Fraud*
8	317 300	19226	0	0	*Fraud*	*Fraud*
9	341 905	47868	0	0	*Fraud*	*Fraud*
10	370 3210	17783	0	0	*Fraud*	*Fraud*
11	318 351	587792	17	1	*Fraud*	*Fraud*
12	315 3459	8173503	0	1	*Fraud*	*Fraud*
13	315 3800	8433878	173	5	*Fraud*	*Fraud*
14	341 739	323518	5	0	*Not Fraud*	*Not Fraud*
15	315 362	24416	1	0	*Not Fraud*	*Not Fraud*

3.5 Skenario Pengujian

Pada skenario pengujian ini, dilakukan percobaan data *training* yaitu data *fraud* dan data bukan *fraud* menggunakan algoritma *Naïve Bayes.* Skenario pengujian dengan membandingkan performansi akurasi *Naïve Bayes*, dengan proporsi sebagai berikut:

a. Skenario 1
   Pada skenario ini dilakukan penelitian dengan menggunakan data panggilan bulan Juni, Juli dan Agustus tahun 2017 yang dibagi menjadi data *testing* dan data *training* dengan proporsi 90% data *training* dengan jumlah datanya sebanyak 4500 data dan 10% data *testing* dengan jumlah datanya adalah 500 data.
b. Skenario 2
   Pada skenario ini dilakukan penelitian dengan menggunakan data panggilan bulan Juni, Juli dan Agustus tahun 2017 yang dibagi menjadi data *testing* dan data *training* dengan proporsi 80% data *training* dengan jumlah datanya sebanyak 4000 data dan 20% data *testing* dengan jumlah datanya adalah 1000 data.
c. Skenario 3
   Pada skenario ini dilakukan penelitian dengan menggunakan data panggilan bulan Juni, Juli dan Agustus tahun 2017 yang dibagi menjadi data *testing* dan data *training* dengan proporsi 75% data *training* dengan jumlah datanya sebanyak 3750 data dan 25% data *testing* dengan jumlah datanya adalah 1250 data.

4. Evaluasi

Setelah dilakukan tahapan implementasi, tahapan selanjutnya adalah melakukan pengujian pada performansi sistem yang telah dibuat. Performansi sistem dengan mengukur nilai akurasi dari setiap skenario pengujian yang telah dilakukan.

4.1 Skenario Pengujian

Berdasarkan pada strategi pengujian diatas, jumlah perbandingan data *training* dan data *testing* akan dibagi menjadi beberapa bagian, dimulai dari rasio 90:10, 90% merupakan data *training* dan 10% merupakan data *testing*, rasio 80:20 dan rasio 75:25. Setelah didapatkan hasil akurasi dari perbandingan tiap bagian data, kemudian didapatkan titip optimal dari algoritma *Naïve Bayes* yaitu dengan menggunakan rasio 90:10. Pengukuran akurasi tersebut menggunakan *confusion matrix.*

4.2 Dataset

Data yang akan digunakan pada penelitian ini adalah data panggilan bulan Juni, Juli dan Agustus tahun 2017 sebanyak 6.985.327 data. Pada penelitian ini akan menggunakan data sample sebanyak 2500 data berlabel *fraud* dan 2500 data *Not fraud.* Dengan total 5000 data, data akan dibagi dengan beberapa rasio. Tabel 4 menunjukkan jumlah pembagian data *training* dan data *testing* masing-masing rasio.

Tabel 4 Pembagian Data Training dan Data Testing

Skenario	Rasio (%)	*Data Training*	*Data Testing*	Total Data
1	90:10	4500	500	5000
2	80:20	4000	1000	
3	75:25	3750	1250	

5. Hasil Pengujian

Skenario pengujian, dilakukan dengan membandingkan performansi akurasi algoritma *naïve bayes*, dengan proporsi sebagai berikut:

5.1 Hasil Skenario 1

Hasil pengujian dengan rasio 90:10 dengan menggunakan data panggilan yaitu untuk mencari nilai performansi dari parameter *accuracy*, *recall* dan *precision* dari *confusion matrix* yang didapatkan. Hasil analisis menggunakan *Confusion Matrix* pada skenario 1 ditunjukkan pada Tabel 5.

Tabel 5 Confusion Matrix Skenario 1

Predicted / True	***Fraud***	***Not Fraud***	**JUMLAH**
Fraud	**266**	0	266
Not Fraud	**64**	**170**	234
JUMLAH	**330**	**170**	**500**

Berdasarkan Tabel 5 *Confusion Matrix* Skenario 1, nomor panggilan *Not fraud* memiliki nilai tertinggi pada nilai *True Positive* (TP), dan nilai terendah pada nilai *False Negative* (FN). Nilai pada *True Positive* (TP) yang tinggi menunjukkan banyaknya kelas yang bernilai "*Not fraud*" dan diklasifikasikan dengan tepat sebagai "*Not fraud*". Sedangkan *False Negative* (FN) yang rendah menunjukkan sistem masih memiliki kesalahan maupun *error* dalam mengklasifikasikan data, dimana kelas, "*Fraud*" diklasifikasikan sebagai "*Not fraud"* oleh *classifier.*

Setelah Tabel *Confusion Matrix* terbentuk, maka dari tabel tersebut dapat dilakukan perhitungan *precision, recall, f1-score* dan *accuracy* berdasarkan nilai *True Positive* (TP), *False Positive* (FP), *True Negative* (TN), dan *False Negative* (FN). Nilai pada *precision, recall, f1-score* dan *accuracy* akan dihitung menggunakan beberapa persamaan. Pada Tabel 6 merupakan tampilan hasil *classification* pada Skenario 1 dalam memprediksi nomor data panggilan *fraud* dan *Not fraud.*

Tabel 6 Hasil Classification Skenario 1

Kategori Prediksi	***Precision***	***Recall***	***f1-score***	**Support**	**Akurasi**
Not fraud	81%	100%	89%	266	87.2%
F*raud*	100%	73%	84%	234	
Macro Average	**90%**	**86%**	**87%**	**500**	

Berdasarkan pada Tabel 6, Nilai akurasi akan menggambarkan tingkat klasifikasi yang tepat terhadap suatu dataset, yaitu nilai dengan kelas *"Fraud"* yang diklasifikasikan sebagai *"Fraud"* dan kelas "*Not fraud*" diklasifikasikan sebagai "*Not fraud*". Apabila mengacu pada probabilitas maka nilai akurasi yang dihasilkan menunjukkan tingkat probabilitas klasifikasi yang tepat. Nilai akurasi yang dihasilkan pada Skenario 1 adalah 87.2%.

Nilai *precision* adalah nilai yang menunjukkan rasio data yang dilabeli sebagai *"Not fraud"* memang bernilai sebagai "*Not fraud*". *Macro average precision* merupakan hasil rata-rata perhitungan *precision fraud* dan *Not fraud* yang hasilnya adalah sebesar 90%. Hal tersebut disebabkan karena nilai *False Positive* (FP) lebih rendah dibandingkan dengan nilai *True Positive* (TP) sehingga menghasilkan nilai *precision* yang tinggi.

Nilai *recall* adalah nilai yang menunjukkan rasio data yang diklasifikasikan secara relevan. *Macro average recall* merupakan hasil rata-rata perhitungan *recall fraud* dan Bukan *fraud* yaitu sebesar 86%. Nilai *recall* mendapatkan hasil yang tinggi karena memiliki *False Negative* (FN) yang paling kecil. Nilai *f1-score* digunakan untuk mengevaluasi hasil rata-rata *precision* dan *recall* hasil klasifikasi.

5.2 Analisis Hasil Pengujian

Berdasarkan skenario pengujian, data yang telah dibagi menjadi 2 bagian yaitu data *training* dan data *testing*, rasio dengan 90:10 yang merupakan 90% data *training* dan 10% pada skenario 3 menghasilkan akurasi yang paling tinggi, yaitu 89.0 % dibandingkan dengan skenario 2 dan skenario 3 dengan penjabaran hasil akurasi dapat dilihat pada Tabel 12 dibawah ini.

Tabel 7 Hasil Pengujian

Skenario	Rasio (%)	*Data Training*	*Data Testing*	*Macro Average Precision*	*Macro Average Recall*	*Macro Average Precision*	Hasil Akurasi
1	90:10	4500	500	90%	86%	87%	87.2%
2	80:20	4000	1000	89%	87%	86%	86.7%
3	75:25	3750	1250	88%	85%	85%	85.2%

Pada Tabel 12 diatas menunjukkan bahwa semakin berkurangnya data *training*, maka tingkat akurasi akan cenderung semakin meningkat. Hasil tertinggi diperoleh ketika pada skenario 1 menggunakan rasio 90% merupakan data *training* dan 10% merupakan data *testing* dengan *macro average precision* bernilai 90%, *macro average recall* bernilai 86% dan *macro average f1-score* bernilai 87%.

Dari hasil pengujian yang telah dilakukan didapat bahwa metode *Bernoulli Naïve Bayes* menghasilkan nilai yang baik pada pengujian skenario 1 dengan nilai akurasi yang dihasilkan adalah 87.2%. Sedangkan untuk pengujian skenario 2 dan skenario 3 menghasilkan akurasi yang tidak begitu jauh nilai akurasinya dengan skenario 1 dengan nilai akurasi yang dihasilkan adalah 86.7% dan 85.2%. Seluruh skenario dengan menggunakan metode *Bernoulli Naïve bayes* menghasilkan nilai akurasi yang baik yaitu diatas 80%.

Berdasarkan hasil dari tiga skenario yang di uji, menunjukkan bahwa *Bernoulli Naïve bayes* merupakan salah satu metode terbaik untuk memprediksi data panggilan *fraud* dan *Not fraud.* Metode dipilih berdasarkan akurasi terbesar dari hasil ketiga skenario tersebut. Tingkat akurasi diukur dengan cara membagi jumlah prediksi benar dengan jumlah data yang diprediksi kemudian akan dikalikan dengan 100%. Dari hasil pengukuran akurasi, algoritma *Bernoulli Naïve bayes Classifier* mampu melakukan prediksi data panggilan *fraud* pada PT PYZ dengan baik.

6. Kesimpulan

Berdasarkan penelitian yang telah dilakukan, dapat disimpulkan bahwa:

1. Pada penelitian ini melakukan analisis dan deteksi pada data panggilan SIM*Box fraud.* Data yang digunakan pada penelitian ini adalah data panggilan bulan Juni, Juli dan Agustus tahun 2017. Dengan pelabelan dengan membandingkan 3 parameter yang terkait dalam mendeteksi SIM*Box fraud* yaitu *B_Number, Calling_Time* dan *Duration.*
2. Kemudian melakukan tahapan pemisahan data *training* dan data *testing.* Data *training* digunakan untuk membentuk sebuah model *classifier* dan data *testing* digunakan untuk mengukur sejauh mana *classifier* berhasil melakukan klasifikasi dengan benar. Pada pemisahan data *training* dan data *testing* akan melakukan klasifikasi menggunakan algoritma *Bernoulli Naïve Bayes*.
3. Pengujian yang telah dilakukan menggunakan metode *Bernoulli Naïve Bayes,* untuk proses klasifikasi didapatkan nilai *accuracy, recall, precision* dan *f1-score* dari evaluasi dengan *confusion matrix*, yaitu 89.0% untuk nilai *accuracy* skenario 1, 86.7% nilai *accuracy* skenario 2 sedangkan 85.2% nilai *accuracy* untuk skenario 3. Sedangkan untuk *macro average precision* terbesar untuk skenario 1, yaitu dengan nilai 90%, *macro average recall* bernilai 86% dan *macro average f1-score* bernilai 87%. Hasil pengujian yang telah dilakukan menghasilkan nilai yang baik pada pengujian skenario 1 dengan nilai akurasi adalah 87.2 %.
4. Berdasarkan hasil dari tiga skenario yang di uji, menunjukkan bahwa *Bernoulli Naïve bayes* merupakan salah satu metode terbaik untuk memprediksi data panggilan SIM*Box fraud.* Metode dipilih berdasarkan akurasi terbesar dari hasil ketiga skenario tersebut. Dari hasil pengukuran akurasi, algoritma *Bernoulli Naïve bayes Classifier* mampu melakukan prediksi data panggilan *fraud* pada PT PYZ dengan baik.

DAFTAR PUSTAKA

[1] L. P. Utomo, "Kecurangan Dalam Laporan Keuangan “Menguji Teori Fraud Triangle”," *Jurnal Akuntansi dan Pajak*, 2018.

[2] ACFE, "What Is Fraud?," *ACFE (Assosiate of Certified Fraud Examiners)*, 2020.

[3] N. Sulisrudatin, "Analisa Kasus CyberCrime Bidang Perbankan Berupa Modus Pencurian Data Kartu Kredit," *Jurnal Ilmiah Hukum Dirgantara*, 2018.

[4] Diskominfo, 2017. [Online]. Available: https://www.postel.go.id/downloads/59/20180716114822-Lakip-SDPPI-2017-Final.pdf.

[5] D. A. Susilo, "Deteksi Kecurangan Pada Jaringan Telekomunikasi Menggunakan Metode Data Mining," *Jurnal Tesis IPB*, 2006.

[6] D. K. A. Puri, "Strategi Pengembangan Unit Anti Fraud PT Bank BPD DIY dalam Meminimalkan Fraud,"

Skripsi Fakultas Ekonomi, 2018.

[7] C. M. D. Rosario Taek, "Fraud Detection pada Transaksi Perbankan Menggunakan Algoritma C4.5," *Skripsi Program Studi Teknik Informatika*, 2019.

[8] F. Nurchalifatun, "Penerapan Metode Asosiasi Data Mining menggunakan Algoritma Apriori untuk Mengetahui Kombinasi Antar Itemset pada Pondok Kopi," *UDiNus Repository*, 2015.

[9] R. McLeod, Jr and G. P. Schell, Sistem Informasi Manajemen, 2007.

[10] H. Widayu, S. D. Nasution, N. Silalahi and Mesran, "Data Mining untuk Memprediksi Jenis Transaksi Nasabah pada Kopersi Simpan Pinjam dengan Algoritma C4.5," *Jurnal Media Informatika Budidarma*, 2017.

[11] M. H. Widianto, "Algoritma Naive Bayes," *Binus University*, 2019.

[12] P. S. A. F. U. M. Chung, Studi Kasus Sistem Informasi Manajemen: Volume 1, 2018.

[13] M. Rifa'i, "Peramalan Cuaca di Kabupaten Bandung Menggunakan Algoritma Naive Bayes," *Open Library Telkom University*, 2018.

[14] N. Amalia, "Penerapan Teknik Data Mining Untuk Klasifikasi Ketepatan Waktu Lulus Mahasiswa Teknik Informatika Universitas Telkom Menggunakan Algoritma Naive Bayes Classifier," *Open Library Telkom University*, 2015.

[15] M. R. AlBougha, "Comparing Data Mining Classification Algortihms in Detection of Simbox Fraud," *Journal Department of Information Systems*, 2016.

[16] A. D. Kusumadety, "Analisis Boosting pada Decision Tree dengan Studi Kasus Klasifikasi Daerah Pelanggan Telekomunikasi Berdasar data Calling Detail Record (CDR) Boosting Analysis in Decision Tree with Case Study Classification of Telecommunication Customer Area Based onCall," *Open Library Telkom University*, 2008.

[17] B. Maryanto, "Big Data dan Pemanfaatannya dalam Berbagai Sektor," *Media Informatika Vol.16 No.2*, 2017.

[18] B. A. Setyawan and D. Pratidana, "Penerepan Teknik Web Scraping dalam Aplikasi Komparasi Harga Komponen dan Perakitan Komputer berbasis Web," *Universitas Multimedia Nusantara*, 2015.

[19] Mardiana, M. A. Muhammad , Y. Mulyani and D. Despa, "Sikronisasi dan Integrasi Database Heterogen Sistem Perpustakaan Unila," *Jurnal Perpustakaan dan Informasi Ilmiah*, 2017.

[20] M. Suhartanto, "Pembuatan Website Sekolah Menengah Pertama Negeri 3 Delanggung," *Jurnal Speed (Sentra Penelitian Engineering dan Edukasi)* , 2012.

[21] I. Warman and R. Ramdaniansyah, "Analisis Perbandingan Kinerja Query Database Management System (DBMS) Antara MySQL 5.7.26 dan MariaDB 10.1," *Jurnal Teknik Dan Informatika*, 2018.

[22] J. Enterprise , Trik Cepat Menguasai Pemrograman Python, Jakarta: PT Elex Media Komputindo, 2016.

[23] F. A. Pahar, W. Suadi and B. J. Santoso, "Implementasi Aplikasi Anti-Virus Berbasis Python-Fuse dan Clamav Pada Sistem Operasi Unix," *Jurnal Teknik Informatika*.

[24] M. Ridwan, H. Suyono and M. Sarosa, "Penerapan Data Mining Untuk Evaluasi Kinerja Akademik Mahasiswa menggunakan Algoritma Naive Bayes Classifier," *Jurnal EECCIS Vol.7*, 2013.

[25] A. Khoirunnisa, "Analisis dan Implementasi Perbandingan Algoritma C4.5 dengan Naive Bayes untuk Prediksi Penawaran Produk," *Open Library Telkom University*, 2016.

[26] T. Nexus, "Detecting and preventing telecom fraud," *Paper Trans Nexus*, 2020.

[27] A. Saleh, "Penerapan Data Mining dengan Metode Klasifikasi Naive Bayes untuk Memprediksi Kelulusan Mahasiswa Dalam Mengikuti English Proficiency Test," *Jurnal Teknik Informatika Universitas Potensi Utama*, 2015.

[28] E. Manalu, F. A. Sianturi and R. M. Manalu, "Penerapan Algoritma Naive Bayes untuk Memprediksi Jumlah Produksi Barang Berdasarkan Dara Persediaan dan Jumlah Pemesanan Pada CV. Papadan Mama Pastries," *Jurnal Mantik Penusa*, 2017.

[29] S. R. Afif, P. Sukarno and M. A. Nugroho , "Analisis Perbandingan Algoritma Naive Bayes dan Decision Tree untuk Deteksi Seriangan Denial of Service (DoS) pada Arsitektur Software Defined Network (SDN)," *e-Proceeding of Engineering*, 2018.