

Komparasi Algoritme C4.5 Dan *Naïve Bayes* Dalam Klasifikasi Produk *Zam-Zam Time* Berdasarkan Tingkat Kepuasan Pelanggan

Dwi Puspa Martiyaningsih^{1*}, Rima Dias Ramadhani², dan Atika Ratna Dewi³

Program Studi Teknik Informatika, Institut Teknologi Telkom Purwokerto, Purwokerto, Indonesia

*e-mail *Corresponding Author*: 19102214@ittelkom-pwt.ac.id

Abstract

The grouping of Zam-Zam Time products based on the level of customer satisfaction is carried out using the C4.5 and Naïve Bayes classification algorithms. Algorithm classification for Zam-Zam Time products is carried out to find out which products are classified as Best Selling or Less Selling. The purpose of this study is to measure and analyze the best algorithm for handling data on the level of customer satisfaction. Zam-Zam Time is classified as a best seller or not a best seller. The method used in this study was data preprocessing by distributing questionnaires and labeling taken from private or primary data from Zam-Zam Time itself as well as the results of a questionnaire of 400 customer respondents, then a classification analysis process was carried out. The results of the performance of the C4.5 Algorithm in the classification of Zam-Zam Time products are classified as Best Selling or Less Selling, namely with a Training data accuracy value of 98%, computation time of 0.003989458084106445 seconds, Testing data accuracy value of 96%, commutation time of 0.001993417739868164 seconds, with the 8th max_depth, and while Naïve Bayes Data Accuracy Value Training 90% computing time 0.0049860477447509766 seconds, Data Testing 85%, computing time 0.0019948482513427734 seconds.

Keywords: Classification; Customer Satisfaction; C4.5 Algorithm; Naïve Bayes; Zam-Zam Time

Abstrak

Pengelompokan produk Zam-Zam Time berdasarkan tingkat kepuasan pelanggan dilakukan menggunakan klasifikasi algoritme C4.5 dan *Naïve Bayes*. Klasifikasi algoritme pada produk *Zam-Zam Time* dilakukan untuk mengetahui produk tergolong *laris* atau *kurang laris*. Tujuan dari penelitian ini mengukur dan analisis algoritme terbaik dalam menangani data tingkat kepuasan pelanggan *Zam-Zam Time* tergolong *laris* atau *kurang laris*. Metode yang digunakan dalam penelitian ini adalah dilakukan *preprocessing* data dengan penyebaran kuesioner dan pelabelan yang diambil dari data privat atau primer dari *Zam-Zam Time* itu sendiri serta hasil kuesioner sebanyak 400 responden pelanggan, kemudian dilakukan proses analisis klasifikasi. Hasil kinerja Algoritme C4.5 dalam klasifikasi produk *Zam-Zam Time* tergolong *Laris* atau *Kurang Laris* yaitu dengan Nilai Akurasi data Training 98%, waktu komputasi 0.003989458084106445 detik, nilai akurasi data Testing 96%, waktu komutasi 0.001993417739868164 detik, dengan *max_depth* ke-8, sedangkan *Naïve Bayes* Nilai Akurasi data *Training* 90% waktu komputasi 0.0049860477447509766 detik, data *Testing* 85%, waktu komputasi 0.0019948482513427734 detik.

Kata Kunci: Kepuasan Pelanggan; Algoritme C4.5; Naïve Bayes; Klasifikasi; Zam-Zam Time

1. Pendahuluan

Zam-Zam Time merupakan Izin Usaha Mikro Kecil (IUMK) yang bergerak dibidang bisnis minuman. *Zam-Zam Time* saat ini memiliki banyak *outlet* yang tersebar di Purwokerto dan sekitarnya. Banyaknya *outlet* menandakan semakin banyak pelanggan yang membeli produk *Zam-Zam Time* maka pelayanan dan produk yang dijual harus memiliki kualitas baik sehingga dapat membantu tercapainya bisnis yang diinginkan. Tingkat kepuasan pelanggan bisa dilihat dari cara pelayanan dan penilaian pada produk itu sendiri dengan produk sesuai pada kebutuhan dan ekspetasi pelanggan. *Zam-Zam Time* memiliki jumlah produk sebanyak 8

varian rasa. Belum adanya penilaian pelayanan dan produk pada Zam–Zam Time membuat Zam–Zam Time kesulitan untuk mengevaluasi barista dan apakah produk Zam-Zam Time termasuk dalam golongan Laris atau Kurang Laris. Kepuasan pelanggan terhadap Zam-Zam Time merupakan salah satu faktor ketika Zam-Zam Time akan melakukan peningkatan mutu pelayanan dan menentukan bahwa produk Zam-Zam Time tergolong laris atau kurang laris. Berdasarkan masalah ini dibutuhkan sebuah teknik klasifikasi tingkat kepuasan pelanggan terhadap produk Zam -Zam Time. Produk Zam-Zam Time tergolong laris atau kurang laris.

Di era digitalisasi sekarang ini pengolahan data akan lebih mudah jika dilakukan dalam bidang data mining yang merupakan sebuah teknik menemukan korelasi yang bermakna dengan jumlah data besar yang disimpan pada repositori menggunakan teknologi statistika dan matematika. Naïve Bayes [1][2] yaitu sebuah metode klasifikasi statistika yang terstruktur digunakan dalam memprediksi probabilitas suatu kelas berdasarkan distribusi data dalam dokumen klasifikasi statistik. Perhitungan *Teorema Bayes* yang sederhana sehingga dapat mengurangi kerumitan perhitungan menjadi perkalian sederhana dari probabilitas [3] metode yang simpel dan waktu komputasi yang singkat [4]. C4.5 [5][6] merupakan suatu algoritme yang dimanfaatkan untuk membangun sebuah pohon keputusan serta digunakan dalam pembelajaran mesin dan penambangan data. C4.5 memiliki keunggulan tahan terhadap data *noise*, dapat menangani variabel diskrit, kontinu [7] serta hilang dan bisa memangkas cabang pohon keputusan yang lain, cabang yang memvisualisasikan hasil dari atribut yang *Testing* dalam sebuah kelas sehingga pohon keputusan dengan mudah diubah menjadi aturan klasifikasi dan dipetakan nilai atribut ke kelas baru, pohon keputusan yang mudah dimengerti, dibangun, dan tidak membutuhkan sampel data yang banyak seperti algoritme klasifikasi lainnya.

Berdasarkan kelebihan Algoritme C4.5 dan *Naïve Bayes* maka pada penelitian ini diperlukan sebuah komparasi dari kedua algoritme tersebut untuk menemukan algoritme baik yang terlihat pada nilai akurasi yang merupakan hasil dan waktu komputasi yang diperlukan masing–masing algoritme dalam proses pengklasifikasian tingkat kepuasan pelanggan terhadap produk Zam-Zam Time. Adapun tujuan dari penelitian ini yaitu untuk mengukur dan menganalisis algoritme terbaik dengan melihat hasil nilai akurasi dan waktu komputasi antara Naïve Bayes dan C4.5 dalam klasifikasi produk Zam-Zam Time berdasarkan tingkat kepuasan pelanggan. Hasil penelitian ini dapat digunakan sebagai bahan evaluasi Zam-Zam Time sehingga dapat meningkatkan hasil omset.

2. Tinjauan Pustaka

Penelitian sebelumnya oleh Nawangsih [8] tentang Menentukan Klasifikasi Produk Terlaris Pada Penjualan Pulsa menggunakan Algoritme *Naïve Bayes*. Permintaan pulsa yang meningkat membuat RA *Cell* sulit mengetahui stok produk yang tinggi pembelainya. Jumlah stok pulsa yang tersedia memiliki pengaruh yang tinggi terhadap peningkatan omset dengan tingkat kepuasan pelanggan. Sehingga dihasilkan bahwa penjualan laris dan tidak laris pada produk pulsa konter RA *cell* yaitu pulsa Telkomsel menggunakan Algoritme *Naïve Bayes* mendapatkan hasil akurasi sebesar 97,50%, nilai *precision* 100 % dan nilai *recall* 93,48%.

Penelitian kedua dari Nurhidayati [9] dan Kawan-kawan menyatakan bahwa Algoritma C4.5 berhasil diterapkan untuk menentukan produk terlaris faktor pestisida yang diminati pelanggan. Kepuasan pelanggan pada persediaan pestisida yang dicari menyebabkan dapat meningkatnya jumlah omset yang diperoleh. Metode Klasifikasi dengan Algoritme *Decision Tree* C4,5 menerapkan 10 *cross validation* dan mendapatkan hasil Akurasi tertinggi yaitu k=8 dengan nilai 97,43%.

Penelitian ketiga dari Lalo [10] dan kawan-kawan menyatakan Algoritme C4.5 berhasil diterapkan untuk klasifikasi produk pada penjualan barang pada swalayan Dutalia. Atribut yang dilihat adalah jenis, harga, jumlah jual, waktu jual, momen jual, dan status penjualan. Berdasarkan pengujian yang dilakukan pada data training dan testing akurasi yang dihasilkan C4.5 mencapai 100%.

Penelitian keempat dari Abdullah [11] dan kawan-kawan pada implementasi klasifikasi produk terlaris menggunakan Algoritma *Naïve Bayes* pada toko Prapti. Penelitian ini menggunakan data penjualan produk yang mana dihasilkan 4 atribut yang dapat mempengaruhi klasifikasi yaitu triwulan 1, 2, 3 dan 4. Hasil klasifikasi pada penelitian ini yaitu *accuracy* 83,3%, *precision* 84,2% dan *recall* 88,9%.

Pada penelitian kami, Klasifikasi dilakukan menggunakan 2 algoritme sehingga dilakukan komparasi dan mendapatkan algoritme dengan kinerja terbaik untuk menunjang proses evaluasi Zam-Zam Time. Berdasarkan penelitian sebelumnya mengenai Algoritme C4.5 dan *Naïve Bayes* merupakan algoritme yang memiliki hasil akurasi yang baik. Algoritme C4.5 memiliki keunggulan dalam proses pembentukan pohon keputusan dan *Naïve Bayes* memiliki keunggulan proses perhitungan yang sederhana dan bisa menangani pada data kecil. Kebaruan dari penelitian ini dengan penelitian sebelumnya adalah dataset yang bersifat privat atau primer pada Zam-Zam Time yang sudah tervalidasi kebenarannya. Penambahan kategori pengukuran faktor kepuasan pelanggan pada produk Zam-Zam Time. Adanya waktu komputasi menjadi bagian pengukuran pada penelitian ini, pada Algoritme yang digunakan dalam proses klasifikasi. Proses analisis klasifikasi kinerja Algoritme dengan adanya waktu komputasi sehingga membuat proses validasi kinerja semakin akurat.

3. Metodologi

3.1 Subjek dan Objek Penelitian

Subjek penelitian ini yaitu Algoritme C4.5 dan *Naïve Bayes* yang mana akan diterapkan pada analisis algoritme mana yang terbaik dalam mengklasifikasikan produk Zam –Zam time tergolong Laris dan Kurang Laris berdasarkan tingkat kepuasan pelanggan. Objek penelitian ini yaitu hasil kuesioner tingkat kepuasan pelanggan pada produk Zam-Zam Time. Dataset kuesioner Zam-Zam Time menggunakan Label kelas status yaitu Laris dan Kurang Laris.

3.2 Teknik Pengambilan Data

Teknik pengumpulan dataset langsung dengan dilakukan penyebaran kuesioner tingkat kepuasan pelanggan pada produk Zam-Zam Time. Kuesioner dilakukan menggunakan media *Google Form* dengan teknik sampling yaitu peneliti menentukan karakteristik sampelnya yang disesuaikan dengan kebutuhan penelitian [12] . Populasi penjualan Zam-Zam Time rata–rata selama 6 bulan (Juli – Desember 2021) sebanyak 7166 pcs. Perhitungan menggunakan rumus *Slovin* dengan nilai E = 5% maka jika diterapkan pada kondisi lapangan sebagai berikut:

$$n = \frac{N}{1+N(E)^2} = \frac{7166}{1+7166(0,05)^2} = \frac{7166}{18,9} = 379,15 \quad (1)$$

dihasilkan ukuran sampel yang dibutuhkan minimal 379 orang untuk dapat mewakili populasi penjualan. Tabel 1 merupakan deskripsi dataset yang digunakan.

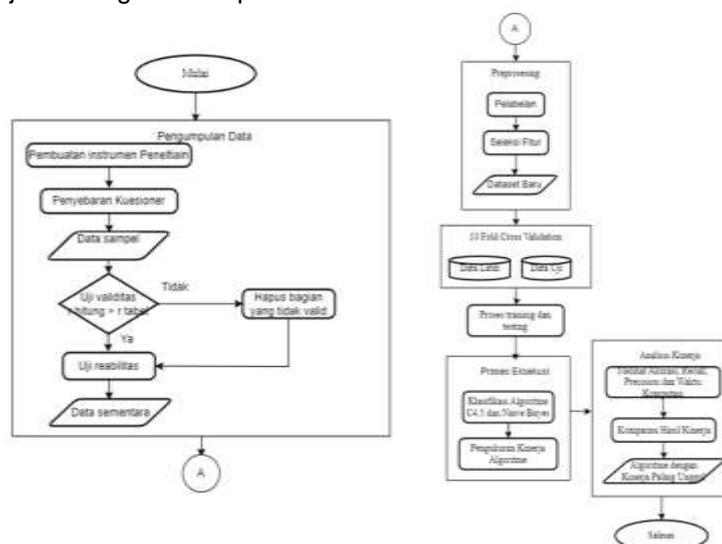
Tabel 1 Deskripsi Dataset

Atribut	Inisialisasi	Kategori
Bagaimana Pendapat anda mengenai beragamnya Varian rasa dari produk Zam-Zam time yang ditawarkan dari pihak Zam-Zam time?	Q1	
Bagaimana Pendapat anda mengenai kesesuaian kemasan, menarik dan rapi yang digunakan pihak Zam-Zam Time terhadap produk Zam-Zam Time?	Q2	Penilaian terhadap Produk
Bagaimana pendapat anda mengenai kualitas rasa produk Zam-Zam Time ketika proses pemesanan dan konsumsi memiliki selang waktu yang cukup lama?	Q3	
Bagaimana pendapat anda mengenai harga yang dipatok Zam-Zam Time dengan rasa produk Zam -Zam Time yang diberikan?	Q4	
Bagaimana Pendapat anda mengenai kebersihan dari Outlet ketika anda melakukan transaksi pembelian produk Zam-Zam Time?	Q5	
Bagaimana pendapat anda mengenai ketanggapan barista dalam melayani anda ketika melakukan pembelian produk Zam-Zam Time?	Q6	

Atribut	Inisialisasi	Kategori
Bagaimana pendapat anda mengenai kecepatan pelayanan barista dalam melayani anda saat membeli produk Zam-Zam Time?	Q7	
Bagaimana pendapat anda mengenai promosi - promosi yang diberikan dari pihak Zam-Zam time selama ini saat anda membeli produk Zam-Zam Time?	Q8	Penilaian terhadap Barista
Bagaimana perasaan anda ketika akan merekomendasikan produk Zam-Zam time kepada orang lain?	Q9	

3.3 Alur Penelitian

Alur penelitian ini setelah menentukan topik, masalah, tujuan, dan manfaatnya. Dilakukan pengambilan dataset yang dibagi untuk data *Training* dan data *Testing* menerapkan *10-Fold Cross Validation* dengan Algoritme *Naïve Bayes* dan *C4.5*. Proses akhir dengan melakukan analisis kinerja nilai akurasi dan waktu komputasi algoritme yang digunakan. Gambar 1 menunjukkan diagram alur penelitian.



Gambar 1. Diagram Alur Penelitian

1) Dataset

Hasil kuesioner yang dibagikan kepada pelanggan atau masyarakat yang telah melakukan pembelian produk Zam-Zam Time di berbagai outlet mendapat data kotor sebanyak 400 responden dengan 9 atribut terhadap penilaian produk Zam-Zam Time. Oleh sebab itu, data yang dihasilkan dengan penyebaran kuesioner yang dilakukan oleh peneliti sudah memenuhi syarat sehingga dataset dapat dilanjutkan pada tahap pengolahan data.

2) Preprocessing Data

Preprocessing pada penelitian ini dilakukan dengan menginisialisasikan dari pertanyaan sebenarnya pada kuesioner menjadi kode inisialisasi seperti yang terlihat pada Tabel 1. Inisialisasi yang digunakan oleh peneliti yang kemudian akan dilakukan *pelabelan* dilanjutkan dengan seleksi fitur pada atribut yang digunakan. Penambahan kelas dengan nama Status yang mana pada penelitian ini berupa pembagian klasifikasi Laris dan Kurang Laris terhadap produk Zam-Zam Time. Pada kelas Status peneliti melakukan diskusi dan proses validasi kepada pakar yaitu pihak Zam-Zam Time yang mana ketika penilaian produk dikatakan Laris memiliki penilaian produk dan pelayanan dengan minimal pelanggan merasa puas yaitu ≥ 27 ketika melakukan pembelian produk. Sebaliknya jika pelanggan merasa kurang puas yaitu < 27 pada produk dan pelayanan maka akan menunjukkan klasifikasi produk Kurang Laris.

Dilakukanya seleksi fitur atribut menggunakan *Correlation Attribute Evaluation* pada dataset yang digunakan dengan melihat nilai korelasi pada masing-masing atribut. Didapatkan hasil data bersih sebanak 400 kemudian dilakukan proses pembagian data menggunakan

fungsi *train_test_split* sehingga akan didapatkan nilai akurasi unggul yang akan diterapkan pada kedua Algoritme yaitu C4.5 dan Naïve Bayes dengan jumlah data *Training* sebanyak 85% dan data *Testing* 15 %.

3) Algoritme C4.5

Algoritme C4.5 banyak penerapan dalam membuat pohon keputusan memiliki tujuan untuk menambah nilai akurasi dari prediksi yang ada dengan hasil yang mudah dimengerti [13]. Algoritme C4.5 memiliki beberapa cabang dan masing–masing mewakili atribut hingga semua terpenuhi dan berakhir [14]. Tahap dalam membuat pohon keputusan pada Algoritme C4.5 terlihat pada Gambar 2 atau seperti dibawah ini yaitu:

1. Mempersiapkan data yang dibagi 2 yang diterapkan untuk data *Training* dan *Testing*.
2. Tentukan akar dari setiap pohon, menghitung nilai *Information Gain* tertinggi dari masing–masing atribut yang digunakan untuk ambil nilai *index entropy* terendah, seperti yang terlihat pada persamaan (2).

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

Ket:

S: kumpulan kasus

pi: Atribut

3. Hitung nilai *Information Gain* dengan rumus, seperti yang terlihat persamaan (3):

$$Information\ Gain(S,A) = Entropy(S) - \sum_{i=1}^N \frac{|s_i|}{|s|} * Entropy(s_i) \quad (3)$$

Ket:

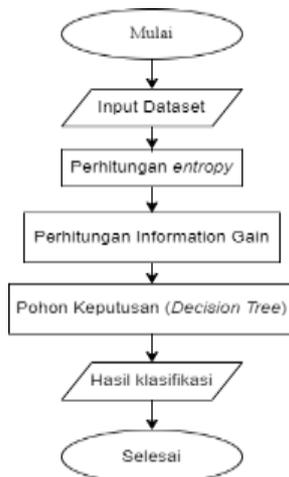
S: Himp kasus

A: Atribut

N: Jmlh partisi atribut A

|si|: Jmlh kasus partisi ke-i

|s|: jmlh kasus pada S



Gambar 2 Diagram Alur Algoritme C4.5

4) Algoritme Naïve Bayes

Naïve Bayes mengacu pada asumsi yang dibuat oleh metode, bahwa pengaruh nilai satu atribut pada probabilitas klasifikasi tertentu tidak bergantung pada nilai atribut lainnya. Kemampuan dan probabilitas bersyarat dalam satu rumus, yang dapat kita gunakan menghitung probabilitas dari setiap kemungkinan klasifikasi secara bergantian, kemudian memilih klasifikasi dengan nilai terbesar [15]. *Naïve bayes* menerapkan cab matematika dan dipahami sebagai teori probabilitas dalam mencari peluang dari banyaknya kemungkinan yang ada, dan tetap memperhatikan frekuensi pada klasifikasi data *Traning* [16].

Tipe *Naïve Bayes Bernoulli* memiliki pokok hasil ya/tidak, variabel digunakan tipe boolean. Multinomial *Naive Bayes* Sebagian besar diterapkan pada klasifikasi dokumen, menerapkan tipe data variabel kata. Gaussian *Naive Bayes* dapat menangani data diskrit atau kontinu dengan data akan diasumsikan Sebagai sampel distribusi gaussian. Oleh karena itu, pada penelitian ini menggunakan Gaussian *Naive Bayes* yang disesuaikan pada tipe data yang

digunakan pada penelitian ini [17]. Konsep Naïve bayes menerapkan teorema kuno yang ditemukan *Thomas Bayes* pada abad ke 18 yang dinyatakan seperti yang terlihat pada persamaan (4).

$$P(X|H) = \frac{P(H|X)P(X)}{P(H)} ; P(H) \neq 0 \quad (4)$$

Ket:

H = Data kelas belum diketahui

X = Hipotesis dari B yang merupakan suatu class spesifik

P(A|H) = Prob H maka A

P(B|H) = Prob H maka B

P(B) = Probs B

P(A)= Prob A

Proses hitung P (B_i) dimana probabilitas *prior* pada masing–masing sub kelas B menghasilkan seperti yang terlihat pada persamaan (5).

$$P(B_i) = \frac{S_i}{s} \quad (5)$$

Keterangan:

S_i: Total data *Training* dari kategori

s: Total data

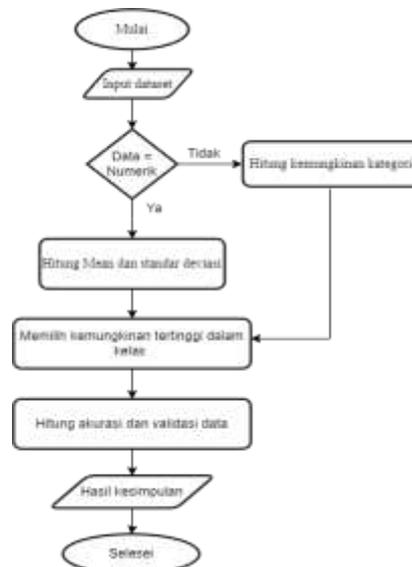
Distribusi *Gaussian* digunakan untuk menunjukkan kemungkinan bersyarat pada atribut numerik dari kelas P(A_i|B) dengan ciri dua parameter yaitu mean μ dan varian σ^2 . Dengan kemungkinan bersyarat kelas B_j untuk atribut A_j seperti pada persamaan (6):

$$P(A_i = a_i | B = b_j) = \frac{1}{\sigma_{ij} \sqrt{2\pi}} \exp \left(-\frac{(a_i - \mu_{ij})^2}{2\sigma_{ij}^2} \right) \quad (6)$$

Ket:

σ = Standar Deviasi

μ = mean



Gambar 3 Diagram Alur Algoritme Naive Bayes

5) 10-Fold Cross Validation

K – Fold Cross Validation didefinisikan teknik dalam *Training Testing* yang digunakan objek kecil berfungsi menilai kinerja algoritme [18]. Jika dataset terdiri N objek dibagi bagian K sama besar, K terdiri dari nilai 1 – 10 jika N tidak habis dibagi K maka bagian akhir memiliki nilai lebih sedikit daripada K lainnya. K run dilakukan masing–masing bagian secara bergantian digunakan sebagai data *Training* dan K lainnya digunakan sebagai data uji. Kemudian jumlah objeknya diklasifikasikan dengan benar yang akan dibagi dengan jumlah N yang memberikan tingkat seluruh akurasi P, atau persamaan kesalahan standar [15]. Nilai *Cross Validation*

diterapkan 10-fold, dimana *10-fold validation* akan melakukan perulangan. Gambar 4 adalah *10-Fold Cross Validation*.

Validation	Dataset									
1	1	2	3	4	5	6	7	8	9	10
2	1	2	3	4	5	6	7	8	9	10
3	1	2	3	4	5	6	7	8	9	10
4	1	2	3	4	5	6	7	8	9	10
5	1	2	3	4	5	6	7	8	9	10
6	1	2	3	4	5	6	7	8	9	10
7	1	2	3	4	5	6	7	8	9	10
8	1	2	3	4	5	6	7	8	9	10
9	1	2	3	4	5	6	7	8	9	10
10	1	2	3	4	5	6	7	8	9	10

Training Testing

Gambar 4 10-Fold Cross Validation

6) Confusion Matrik

Confusion matrix diterapkan model klasifikasi yang memiliki nilai terbaik melihat dari perhitungan akurasi, *precision* dan *recall* [19]. *Confusion matrix* yaitu instrumen tabel yang digunakan menganalisis seberapa besar ketepatan atau kinerja suatu model algoritme dapat digunakan untuk memahami *tuple* data yang berbeda dalam sebuah klasifikasi.

	Positif	Negatif	
Positif	TP	FN	TP + FN
Negatif	FP	TN	FP + TN
	TP + FP	FN + TN	

Gambar 5 Model Confusion Matrix

Nilai yang dikeluarkan dengan *Confusion Matrix* [20] [21] adalah

1. Akurasi merupakan sebuah nilai yang melambangkan suatu ketepatan, ketelitian dan keakuratan pada suatu hasil proses klasifikasi secara benar menggunakan algoritme, terlihat pada persamaan (7).

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN} \tag{7}$$

2. *Precision* merupakan tingkat ketepatan hasil sebuah model algoritme yang akan dibandingkan nilai Laris / Kurang Laris dengan total data label Laris / Kurang Laris, seperti yang terlihat pada persamaan (8).

$$\text{Precision} = \frac{TP}{TP+FP} \tag{8}$$

3. *Recall* merupakan parameter kelengkapan suatu model algoritme yang membandingkan total data yang benar – benar bernilai Laris / Kurang Laris (9).

$$\text{Recall} = \frac{TP}{TP+FN} \tag{9}$$

4. Hasil dan Pembahasan

Pada penelitian ini dilakukan klasifikasi menerapkan model Algoritme yaitu C4.5 dan *Naïve Bayes*. Pembagian data yang digunakan untuk proses *Training* sebanyak 10 % - 90 % dan *Testing* sebanyak 10% - 90%. Tabel 2 merupakan proses pembagian data.

Tabel 2 Dataset Bersih

Percobaan	Proses	Data Training %	Data Testing %	Average Cross Validation Score		Akurasi	
						C4.5	Naïve Bayes
1	Training	90	10	0.9319	0.9043	100	86
	Testing					95	85
2	Training	85	15	0.9345	0.9068	98	90
	Testing					95	85
3	Training	75	25	0.9184	0.9026	93	89

Percobaan	Proses	Data Training %	Data Testing %	Average Validation Score	Cross Score	Akurasi	
						C4.5	Naïve Bayes
4	Testing	70	30	0.9184	0.9176	96	87
	Training					96	89
5	Testing	65	35	0.9295	0.9100	95	91
	Training					93	88
6	Testing	55	45	0.9333	0.8763	94	90
	Training					92	89
7	Testing	45	55	0.9500	0.9375	92	90
	Training					94	81
8	Testing	35	65	0.8600	0.9400	93	89
	Training					89	87
9	Testing	30	70	0.8833	0.8917	90	89
	Training					83	88
10	Testing	25	75	0.9500	0.8333	89	88
	Training					82	74
						85	75

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	STATUS
0	3	3	3	4	3	3	3	3	3	1
1	3	3	3	2	3	3	3	3	3	0
2	3	3	3	3	3	3	3	3	3	1
3	3	3	3	3	3	3	3	3	2	0
4	3	4	4	3	4	4	3	4	3	1
...
395	3	3	3	3	3	3	3	3	3	1
396	4	3	3	3	3	3	3	3	3	1
397	3	3	3	3	3	2	2	3	3	0
398	3	3	3	3	3	3	3	3	4	1
399	3	3	3	3	3	3	3	3	4	1

400 rows x 10 columns

Gambar 6 Dataset Bersih

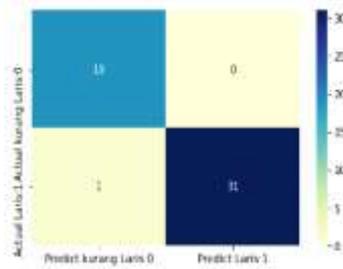
Gambar 6 merupakan dataset bersih. Jumlah dataset yang digunakan setelah melalui pembagian data pada penelitian ini yaitu sebanyak 85% maka 340 digunakan pada set data *Training*. Data tersebut digunakan pada pembuatan dan pengujian model Algoritme C4.5 dan Naïve Bayes. 15% data lainnya yang diterapkan pada set data *testing* terdapat data sebanyak 60 digunakan pada set data *Testing*.

Pada pemodelan C4.5 yang digunakan pada penelitian ini memiliki waktu eksekusi dalam performasi sebanyak 0.012967824935913086 detik. Diterapkan *max_depth* yang bertujuan untuk mengontrol model pada proses pembelajaran sesuai dengan kedalaman pohon. Hasil pengujian C4.5 menggunakan *Max_depth* =8 sesuai pada Tabel 3.

Tabel 3 Max_Depth

Max_depth	Akurasi Train%	Akurasi Test%
1	80	75
2	82	76
3	94	83
4	88	90
5	94	93
6	92	95
7	94	96
8	98	96
9	90	95
10	96	96

Klasifikasi C4.5 dilakukan penerapan *10-Fold Cross Validation*. Dilakukannya proses validasi dengan *10-Fold Cross Validation*, hasil rata-rata dari penerapan *10-Fold Cross Validation* adalah 0.9345. Diterapkan *Confusion Matrix* pada penelitian ini untuk mengetahui persebaran data yang dibentuk dari hasil pemodelan. Didapatnya *Confusion Matrix* yang terbentuk dapat dilihat pada Gambar 7.



Gambar 7 Confusin Matrix Training C4.5

Hasil perhitungan sesuai dengan hasil akurasi *Training* pada *python* terlihat pada Gambar 8.

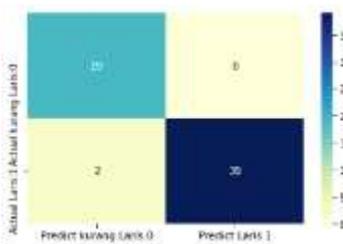
Tingkat Akurasi Algoritma C4.5

	precision	recall	f1-score	support
Kurang Laris	0.95	1.00	0.97	19
Laris	1.00	0.97	0.98	32
accuracy			0.98	51
macro avg	0.97	0.98	0.98	51
weighted avg	0.98	0.98	0.98	51

Tingkat Akurasi : 98 persen

Gambar 8 Akurasi Training C4.5

Berdasarkan pada penerapan model C4.5 didapatkan *Confusion Matrix* seperti yang terlihat pada Gambar 9.



Gambar 9 Confusin Matrix Testing C4.5

Hasil dari penerapan model dapat terlihat pada Gambar 10, dimana pada Gambar 10 memiliki informasi yang memuat waktu komputasi, nilai akurasi, recall, dan precision.

Waktu Komputasi 0.001993417739868164

Tingkat Akurasi Algoritma C4.5

	precision	recall	f1-score	support
Kurang Laris	0.90	1.00	0.95	19
Laris	1.00	0.95	0.97	41
accuracy			0.97	60
macro avg	0.95	0.98	0.96	60
weighted avg	0.97	0.97	0.97	60

Tingkat Akurasi : 96 persen

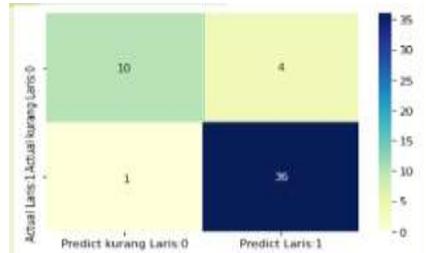
Gambar 10 Hasil Akurasi *Testing* Model C4.5

Hasil Prediksi Status Produk Zam-Zam Time pada data teratas dapat terlihat pada Gambar 11.

	Actual	Predict
252	0	0
220	0	1
62	1	1
303	0	0
209	1	1

Gambar 11 Hasil Prediksi teratas C4.5

Pemodelan Naïve Bayes yang digunakan pada penelitian ini memiliki waktu komputasi sebanyak 0.003991127014160156 detik. Sama halnya dengan C4.5 pada model *Naïve Bayes* juga diterapkan pengujian *10-Fold Cross Validation*. Hasil dari penerapan *10-Fold Cross Validation* yaitu rata-rata 0.9031. Didapatnya *Confusion Matrix* yang terbentuk dapat dilihat pada Gambar 12.



Gambar 12 *Confusion Matrix Naïve Bayes Training*

Berdasar pada perhitungan akurasi, *precision*, dan *recall* selaras dengan menggunakan python seperti yang terlihat Gambar 13.

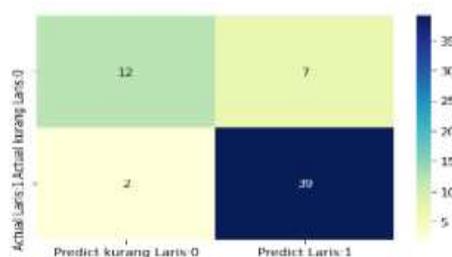
```

Tingkat Akurasi Algoritma NAIVE BAYES
precision recall f1-score support
Kurang Laris 0.91 0.71 0.80 14
Laris 0.98 0.97 0.94 37
accuracy 0.90 51
macro avg 0.90 0.84 0.87 51
weighted avg 0.90 0.90 0.90 51

Tingkat Akurasi : 90 persen
    
```

Gambar 13 Hasil Akurasi *Training Naïve Bayes*

Berdasarkan pada penerapan model *Naïve Bayes* didapatkan *Confusion Matrix* seperti yang terlihat pada Gambar 14.



Gambar 14 *Confusion Matrix Naïve Bayes Testing*

Hasil dari model *Naïve Bayes* dalam pada penelitian terlihat pada Gambar 9, yang mana pada Gambar 9 memuat informasi waktu komputasi, akurasi, *recall*, dan *precision*.

```

Waktu Komputasi 0.0030298233032226562
Tingkat Akurasi Algoritma NAIVE BAYES
precision recall f1-score support
Kurang Laris 0.86 0.63 0.73 19
Laris 0.85 0.95 0.90 41
accuracy 0.85 60
macro avg 0.85 0.79 0.81 60
weighted avg 0.85 0.85 0.84 60

Tingkat Akurasi : 85 persen
    
```

Gambar 15 Hasil akurasi *Testing Naïve Bayes*

Hasil Prediksi Status Produk Zam-Zam Time pada data teratas dapat terlihat pada Gambar 16.

	Actual	Predict
73	1	0
111	1	1
190	1	1
276	1	1
165	0	1

Gambar 16 Hasil Prediksi Teratas Naive Bayes

Hasil komparasi proses *Training* dataset pada model dapat dilihat pada Tabel 4.

Tabel 4 Hasil Komparasi Data *Training*

Parameter	C4.5	Naïve Bayes
Akurasi	98%	90%
Waktu Komputasi	0.003989458084106445	0.0049860477447509766
Recall Laris	97%	97%
Recall Kurang Laris	100%	71%
Precision Laris	100%	90%
Precision Kurang Laris	95%	91%

Dilakukan pengujian *Testing* pada model klasifikasi dan terbukti tetap lebih baik Algoritme C4.5, hal ini selaras dengan penelitian [22] [23] banyaknya data *Training* pada proses pembelajaran membuat model akan semakin pintar sehingga pada pengujian dihasilkan nilai akurasi yang unggul. Hasil komparasi kinerja Algoritme klasifikasi C4.5 dengan *Naïve Bayes* dilihat pada Tabel 5.

Tabel 5. Hasil Komparasi Data *Testing*

Parameter	C4.5	Naïve Bayes
Akurasi	96%	85%
Waktu Komputasi	0.001993417739868164	0.0019948482513427734
Recall Laris	95%	95%
Recall Kurang Laris	100%	63%
Precision Laris	100%	85%
Precision Kurang Laris	90%	86%

Tabel 5 menunjukkan bahwa C4.5 mendapatkan nilai akurasi unggul. Hal tersebut dipengaruhi oleh pembagaiman data yang tersebar dengan optimal, jumlah dataset yang digunakan, serta hasil perhitungan dan penerapan yang tepat pada proses pembentukan model yang digunakan [24]. Algoritme C4.5 pada proses perhitungan dengan mencari *entropy* pada masing-masing atribut sehingga didapat *Information Gain* tertinggi, baik untuk menangani data statistik selaras dengan penelitian [25], informasi di dalam pohon yang menampilkan faktor-faktor yang terbentuk selaras dengan penelitian [26], mengatasi data diskrit atau kontinu dengan baik, kecepatan pada proses klasifikasi memiliki sifat prediktif sejalan dengan penelitian [27]. Penelitian ini sejalan dengan penelitian-penelitian [28][29][30] yang menyatakan bahwa Algoritme C4.5 memiliki nilai akurasi yang tinggi. Faktor yang mempengaruhi pengambilan keputusan yaitu Q5 (kebersihan), Q9 (rekomendasi pelanggan), Q2 (kesesuaian kemasan) sejajar dengan Q4 (harga). Keterhubungan nilai kepuasan pelanggan terhadap Laris atau Kurang Larisnya produk Zam-Zam Time dapat terlihat ketika pelanggan merasakan dan memiliki pengalaman yang puas saat membeli produk sehingga pelanggan akan melakukan Repeat Order produk Zam-Zam Time.

5. Simpulan

Simpulan dari penelitian ini didapatkan hasil kinerja dan validasi dari Algoritme C4.5 unggul dari *Naïve Bayes* yaitu Nilai Akurasi C4.5 pada data Training sebanyak 98%, waktu komputasi 0.00398945 detik, data Testing sebanyak 96%, waktu komputasi 0.00199341 detik pada kedalaman pohon ke-8, tergolong pada *excellent classification* dan hasil kinerja Algoritma *Naïve Bayes* Nilai Akurasi pada Training sebanyak 90%, waktu komputasi 0.00498604 detik, data Testing sebanyak 85%, waktu komputasi 0.0019948482 tergolong *Good Classification* dalam mengklasifikasikan faktor Laris dan Kurang Larisnya produk Zam-Zam Time.

Daftar Referensi

- [1] T. Imandasari, E. Irawan, A. Perdana Windarto, A. Wanto, And S. A. Tunas Bangsa Pematangsiantar Jln Jendral Sudirman Blok No, "Prosiding Seminar Nasional Riset Information Science (Senaris) Algoritme Naive Bayes Dalam Klasifikasi Lokasi Pembangunan Sumber Air," No. September, Pp. 750–761, 2019.
- [2] E. Indrayuni, "Klasifikasi Text Mining Review Produk Kosmetik Untuk Teks Bahasa Indonesia Menggunakan Algoritme Naive Bayes," *J. Khatulistiwa Inform.*, Vol. 7, No. 1, Pp. 29–36, 2019, Doi: 10.31294/Jki.V7i1.1.
- [3] T. S. Jaya, "Klasifikasi Nanas Layak Jual Dengan Metode Naive Bayes Classifier Dan K-Nearest Neighbor," Vol. 13, No. 1, Pp. 2221–2229, 2021.
- [4] E. Setiani And W. Ce, "Text Classification Services Using Naive Bayes For Bahasa Indonesia," In *Proceedings Of 2018 International Conference On Information Management And Technology, Icimtech 2018*, 2018, Pp. 361–366. Doi: 10.1109/Icimtech.2018.8528258.
- [5] V. Anestiviya, A. Ferico, And O. Pasaribu, "Analisis Pola Menggunakan Metode C4.5 Untuk Peminatan Jurusan Siswa Berdasarkan Kurikulum (Studi Kasus : Sman 1 Natar)," *J. Teknol. Dan Sist. Inf.*, Vol. 2, No. 1, Pp. 80–85, 2021, [Online]. Available: <Http://Jim.Teknokrat.Ac.Id/Index.Php/Jtsi>
- [6] P. B. N. Setio, D. R. S. Saputro, And Bowo Winarno, "Klasifikasi Dengan Pohon Keputusan Berbasis Algoritme C4.5," *Prism. Pros. Semin. Nas. Mat.*, Vol. 3, Pp. 64–71, 2020.
- [7] E. Purwaningsih, "Seleksi Mobil Berdasarkan Fitur Dengan Komparasi Metode Klasifikasi Neural Network, Support Vector Machine, Dan Algoritme C4.5," *J. Pilar Nusa Mandiri*, Vol. Xii, No. 2, Pp. 153–160, 2016, [Online]. Available: <Https://Ejournal.Nusamandiri.Ac.Id/Index.Php/Pilar/Article/View/269>
- [8] I. Nawangsih, & A. Setyaningsih, "Penerapan Algoritma Naive Bayes Untuk Menentukan Klasifikasi Produk Terlaris Pada Penjualan Pulsa". *Jurnal SIGMA*, vol. 10, no. 4, pp. 9-15, 2019.
- [9] A. Nurhidayati, "Klasifikasi Penjualan Obat Pertanian Laris Dan Kurang Laris Pada UD Cahaya Tani Menggunakan Metode Decission Tree". *Jurnal Informatika dan Teknologi*, vol. 2, no. 2, pp. 101-108, 2019.
- [10] A.K. Lalo, P. Batarius, & Y.C.H. Siki, "Implementasi Algoritma C4. 5 Untuk Klasifikasi Penjualan Barang di Swalayan Dutalia". *Jurnal Teknik Informatika UNIKA Santo Thomas*, vol. 6, no. 1, pp. 1-12, 2021.
- [11] R.W. Abdullah, D. Hartanti, H. Permatasari, A.W. Septyanto, & Y.A. Bagaskara, "Penerapan Data Mining untuk Memprediksi Jumlah Produk Terlaris Menggunakan Algoritma Naive Bayes Studi Kasus (Toko Prapti)". *Jurnal Informatika Global*, vol. 13, no. 1, pp. 34-41, 2022
- [12] A. Hanafiah, Sutedja, & I. Ahmaddien, *Pengantar Statistika*. Bandung: Widina Bhakti Persada Bandung, 2020.
- [13] R. H. Alawiah, Saifullah, And I. S. Damanik, "Analisis Kepuasan Konsumen Terhadap Pelayanan Bengkel Menggunakan Metode Algoritme C4.5," *J. Penerapan Sist. Inf. (Komputer Manajemen)*, Vol. 2, No. 1, Pp. 31–38, 2021, Doi: 10.29406/Jpr.V4i1.789.
- [14] D. R. Adhy, "Rancang Bangun Sistem Prediksi Varietas Padi Yang Cocok Dengan Lahan Menggunakan Metode Data Mining Algoritme C4. 5," *Jurnal Ilmiah Sains, Teknologi Dan Rekayasa*. 2021. [Online]. Available: <Http://Saintesa.Sttybsi.Ac.Id/Index.Php/Saintesa/Article/View/3>
- [15] D. J. Hand, *Principles Of Data Mining*, Vol. 30, No. 7. 2007. Doi: 10.2165/00002018-200730070-00010.
- [16] Fatmawati, "Perbandingan Algoritme Klasifikasi Data Mining Model C4.5 Dan Naive Bayes Untuk Prediksi Penyakit Diabetes," *J. Techno Nusa Mandiri*, Vol. Xiii, No. 1, P. 50, 2016.
- [17] Budi Raharjo, *Pembelajaran Mesin (Machine Learning)*. Semarang: Yayasan Prima Agus Teknik, 2016. [Online]. Available: <Https://Www.Codepolitan.Com/Mengenai-Teknologi-Machine-Learning-Pembelajaran-Mesin>
- [18] F. T. Admojo And Ahsanawati, "Klasifikasi Aroma Alkohol Menggunakan Metode Knn," *Indones. J. Data Sci.*, Vol. 1, No. 2, Pp. 34–38, 2020, Doi: 10.33096/ljodas.V1i2.12.
- [19] T. A. M. Putri, U. Enri, And B. N. Sari, "Ijcit (Indonesian Journal On Computer And Information Technology) Analisis Algoritme Naive Bayes Classifier Untuk Klasifikasi Tweet Pelecehan Seksual Dengan #Metoo," *Ijcit (Indonesian J. Comput. Inf. Technol.*, Vol. 5, No.

- 2, Pp. 126–135, 2020, [Online]. Available: <https://Ejournal.Bsi.Ac.Id/Ejurnal/Index.Php/ljcit/Article/View/8636>
- [20] A. Rosadi And D. Gustiana, “Analisis Sentimen Berdasarkan Opini Pengguna Pada Media Twitter Terhadap Bpjs Menggunakan Metode Lexicon Based Dan Naïve Bayes Classifier,” *J. Ilm. Komputasi*, Vol. 20, No. 1, Pp. 39–52, 2021, Doi: 10.32409/Jikstik.20.1.401.
- [21] D. Normawati And S. A. Prayogi, “Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter,” *J. Sains Komput. Inform.*, Vol. 5, No. 2, Pp. 697–711, 2021.
- [22] W. Widystuti And J. B. B. Darmawan, “Pengaruh Jumlah Data Set Terhadap Akurasi Pengenalan Dalam Deep Convolutional Network,” *Konf. Nas. Sist. Inf.*, Pp. 8–9, 2018.
- [23] I. Budiman, Muliadi, And R. Ramadina, “Penerapan Fungsi Data Mining Klasifikasi Untuk Prediksi Masa Studi Mahasiswa Tepat Waktu Pada Sistem Informasi Akademik Perguruan Tinggi,” *J. Jupiter*, Vol. 7, No. 1, Pp. 39–50, 2015.
- [24] Ainurrohmah, “Akurasi Algoritme Klasifikasi Pada Software Rapidminer Dan Weka,” *Prisma*, Vol. 4, Pp. 493–499, 2021, [Online]. Available: <https://Journal.Unnes.Ac.Id/Sju/Index.Php/Prisma/>
- [25] N. S. B. Kusrorong, D. R. Sina, And N. D. Rumlaklak, “Kajian Machine Learning Dengan Komparasi Klasifikasi Prediksi Dataset Tenaga Kerja Non-Aktif,” *J-Icon*, Vol. 7, No. 1, Pp. 37–49, 2019.
- [26] G. Lukhayu Pritalia, “Penerapan Algoritme C4.5 Untuk Penentuan Ketersediaan Barang E-Commerce,” *Indones. J. Inf. Syst.*, Vol. 1, No. 1, Pp. 47–56, 2018, Doi: 10.24002/ljis.V1i1.1727.
- [27] S. Ucha Putri, E. Irawan, F. Rizky, S. Tunas Bangsa, P. A. -Indonesia Jln Sudirman Blok No, And S. Utara, “Implementasi Data Mining Untuk Prediksi Penyakit Diabetes Dengan Algoritme C4.5,” *Januari*, Vol. 2, No. 1, Pp. 39–46, 2021.
- [28] F. M. Hana, “Klasifikasi Penderita Penyakit Diabetes Menggunakan Algoritme Decision Tree C4.5,” *J. Sist. Komput. Dan Kecerdasan Buatan*, Vol. 4, No. 1, pp. 32-39, 2020.
- [29] N. Nurhidayati And A. Alimuddin, “Klasifikasi Penjualan Obat Pertanian Laris Dan Kurang Laris Pada Ud Cahaya Tani Menggunakan Metode Decission Tree,” *Infotek J. Inform. Dan Teknol.*, Vol. 2, No. 2, Pp. 101–108, 2019, Doi: 10.29408/Jit.V2i2.1489.
- [30] M. Muhamad, A. P. Windarto, And S. Suhada, “Penerapan Algoritme C4.5 Pada Klasifikasi Potensi Siswa Drop Out,” *Komik (Konferensi Nas. Teknol. Inf. Dan Komputer)*, Vol. 3, No. 1, Pp. 1–8, 2019, Doi: 10.30865/Komik.V3i1.1688.