Available online at https://ejournal.almaata.ac.id/index.php/IJUBI
Indonesian Journal of Business Intelligence
Volume 3 | Issue 1 | June (2020)
IJUBI
Indonesian Journal
---- of ----
Business Intelligence

KOMPARASI METODE NAÏVE BAYES DAN C4.5 DALAM KLASIFIKASI LOYALITAS PELANGGAN TERHADAP LAYANAN PERUSAHAAN

Musthofa Galih Pradana[1], Pujo Hari Saputro[2]
[1 2]Program Studi Informatika, Fakultas Komputer, Universitas Alma Ata Yogyakarta
[1]mgalihpradana@almaata.ac.id, [2]pujo@almaata.ac.id
[12]Jl. Brawijaya 99, Yogyakarta

Keywords:
Data Mining, Classification, Naïve Bayes, C.45.

Abstract

The existence of customers for the course of a business is very important. Customers have a tendency to continue to subscribe to the company or to stop subscribing. One technique that can be used to identify trends in customer loyalty is data classification. Based on customer-owned data the company can do data processing or data mining by classifying loyal and non-loyal customers. There are many methods that can be applied for data classification, including the Naïve Bayes algorithm and C4.5. Both of these methods produce different accuracy when used for the data classification process. Two scenarios are used in the process of testing both algorithms, scenarios for dividing data in testing and training data and testing scenarios using cross validation. The results of these two scenarios show that the C4.5 method is superior to the Naïve Bayes method with scenario 1 accuracy of 78.6086% and scenario 2 accuracy of 78.61%.

Kata Kunci:
Data Mining, Klasifikasi, Naïve Bayes, C.45.

Abstrak

Keberadaan pelanggan bagi jalannya sebuah usaha sangat penting. Pelanggan memiliki kecenderungan yakni untuk tetap lanjut berlangganan dengan perusahaan atau sebaliknya berhenti berlangganan. Salah satu teknik yang dapat digunakan untuk mengidentifikasi kecenderungan loyalitas pelanggan adalah dengan klasifikasi data. Berdasarkan data pelanggan yang dimiliki perusahaan dapat dilakukan pengolahan data atau data mining dengan mengkelompokan pelanggan yang loyal dan yang tidak loyal. Ada banyak metode yang dapat diterapkan untuk klasifikasi data, diantaranya adalah algortima Naïve Bayes dan C4.5. Kedua metode ini menghasilkan akurasi yang berbeda ketika digunakan untuk proses klasifikasi data. Digunakan 2 skenario dalam proses pengujian kedua algoritma, skenario membagi data dalam data testing dan training serta skenario pengujian menggunakan cross validation. Hasil kedua skenario ini menunjukan bahwa metode C4.5 lebih unggul dibandingkan dengan metode Naïve Bayes dengan akurasi skenario 1 sebesar 78,6086 % dan skenario 2 akurasi sebesar 78,61%.

Pendahuluan

Mengindentifikasi pelanggan yang setia merupakan salah satu tantangan bagi sebuah perusahaan. Keberadaan pelanggan bagi sebuah perusahaan sangat penting dan vital. Pelanggan adalah kunci keberhasilan dari sebuah usaha yang dijalankan perusahaan. Tingkat kesuksesan dan keberhasilan perusahaan diukur dari tingkat profit yang didapatkan. Profit yang tinggi akan sangat dipengaruhi oleh jumlah pelanggan aktif yang melakukan transaksi dengan perusahaan. Contoh nyata dari kondisi ini adalah data perusahaan musik Spotify yang mengalami kenaikan jumlah pelanggan (pelanggan yang loyal) sebanyak 9 juta pelanggan ada kuartal IV tahun 2018, hal ini secara langsung juga ikut mendongkrak pendapatan perusahaan sebesar 11% dari kuartal sebelumnya. Berdasarkan fakta tersebut maka proses identifikasi loyalitas pelanggan menjadi penting. Cara yang dapat digunakan untuk mengindentifikasi pelanggan yang setia adalah dengan mengklasifikasi data.

Tujuan dari klasifikasi data digunakan untuk mengelompokan jumlah pelanggan yang loyal

dan yang tidak loyal. Menurut Emmett C. Murphy dan Mark A. Murphy untuk mendapatkan pelanggan baru akan jauh lebih sulit dibandingkan dengan mempertahankan pelanggan lama, serta biaya yang dikeluarkan perusahaan lima kali lipat lebih banyak dibandingkan dengan memuaskan dan mempertahankan pelanggan lama. Data hasil klasifikasi dapat dijadikan acuan bagi perusahaan untuk menentukan langkah selanjutnya yang akan ditempuh. Salah satu contoh yang dapat dilakukan perusahaan adalah dengan melakukan kegiatan promosi dengan lebih tepat dengan identifikasi lebih dini mengenai loyalitas pelanggan.

Pada penelitian ini dataset yang digunakan adalah dataset pelanggan sebuah perusahaan yang bergerak di bidang telekomunikasi. Berdasarkan dataset yang ada, data akan doiolah dengan teknik klasifikasi, atau dalam kata lain data akan diklasifikasikan menggunakan algoritma C4.5 dan menggunakan algoritma Naïve Bayes. Kedua teknik klasifikasi ini akan dibandingkan satu sama lain dan dicari yang memiliki akurasi paling tinggi dalam mengklasifikasi data dengan menggunakan 2 skenario pengujian.

Skenario pengujian yang pertama menggunakan teknik pembagian data training dan data testing dari dataset yang sudah ada untuk mencari akurasi paling tinggi. Ada 4 percobaan pada pengujian pertama. Pengujian kedua dilakukan dengan menggunakan teknik cross validation untuk mencari nilai akurasi. Jumlah nilai K yang digunakan sebanyak 5 nilai. Kedua skenario pengujian tersebut akan dibandingkan atau dikomparasi untuk dicari mana algoritma yang memiliki akurasi yang lebih baik dari keduanya.

Landasan Teori

1. Tinjauan Pustaka

Penelitian tentang penggunaan algoritma C4.5 dan Naïve Bayes pernah dilakukan oleh Yogiek Indra Kurniawan yang di publikasi di Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) dengan kesimpulan yang diambil semakin banyakdata training yang digunakan, maka nilai precision, recall dan accuracy akan semakin meningkat, selain itu untuk menentukan algoritma terbaik yang akan dipakai di sebuah kasus, harus melihat kriteria, variable maupun jumlah data di kasus tersebut. [1]

Selanjutnya penelitian tentang algoritma C4.5 dilakukan oleh Erlan Darmawan yang dimuat dalam JOIN (Jurnal Online Informatika), pada penelitian penulis melakukan development sistem dimana digunakan untuk menentukan kandidat calon siswa di sekolah menengah atas dengan menerapkan model pengembangan waterfall. [2]

Penelitian yang ditulis oleh Mochamad Idris dalam ICENIS 2019, E3S Web of Conferences adalah penelitian yang terbit pada tahun 2019 menuliskan bahwa algoritma C4.5 menghasilkan akurasi sebesar 90% dalam melakukan klasifikasi dan melakukan kombinasi menggunakan teknik forward chaining. [3]

Penelitian terkait pernah dilakukan oleh Andri Wijaya dan Abba Suganda Girsang dengan judul The Use Of Data Mining For Prediction Of Customer Loyalty, dipublikasikan dalam CommIT (Communication & Information Technology), May 31. Penelitian ini mengambil kesimpulan bahwa Algoritma Naïve Bayes menghasilkan akurasi sebesar 76 %. [4]

Penelitian rujukan berikutnya adalah A Survey on Naive Bayes Based Prediction of Heart Disease Using Risk Factors dengan penulis Sohana Saiyed, Nikita Bhatt and Amit P. Ganatra di International Journal of Innovative and Emerging Research in Engineering. Penelitian ini berkesimpulan bahwa Algoritma Naïve bayes mendapatkan nilai akurasi 86.38%. [5]

Penelitian ini dapat dikembangkan juga dalam bentuk sistem pendukung keputusan. Berdasarkan data dari olah data mining yang dapat di buat sebuah decision support system. [6]

2. Dasar Teori

a. Data Mining

Data mining merupakan proses menggali nilai dari sebuah data yang sudah ada secara otomatis.[7] Ada banyak teknik yang dapat diterapkan dalam proses data mining.

b. Algoritma C4.5

Algoritma C4.5 merupakan algoritma yang digunakan dalam membentuk decision tree. Algoritma C4.5 adalah salah satu algoritma dalam induksi decision tree yaitu ID3. Prosedur algoritma ID3, input berupa sampel training, label training dan atribut. Algoritma C4.5 ini merupakan pengembangan dari ID3. [8]

c. Algoritma Naïve Bayes

Naïve Bayes merupakan sebuah model klasifikasi statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu

kelas. Naïve Bayes didasarkan pada teorema bayes yang memiliki kemampuan klasifikasi serupa dengan decision tree dan neural network [5]. Adapun rumus kedekatan pada Naïve Bayes sebagai berikut [9] :

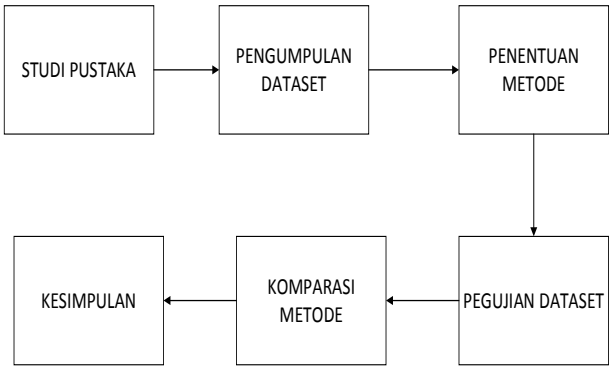
$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (1)$$

Keterangan :

X : Data dengan class yang belum diketahui
H : Hipotesis data X merupakan suatu class spesifik
P(H|X) : Probabilitas hipotesis H berdasar kondisi X (posteriori probability)
P(H) : Probabilitas hipotesis H (prior probability)
P(X) : Probabilitas dari X

Metode

Metode penelitian yang diterapkan adalah metode kuantitatif dengan mengolah dataset pelanggan perusahaan yang ada untuk menghasilkan sebuah informasi baru yang dapat bermanfaat. Adapun alur penelitian ini ditunjukan pada Gambar 1.



Gambar 1. Alur Penelitian

Alur dimulai dari studi pustaka mencari penelitian yang relevan kemudian dilanjutkan dengan mencari dataset yang cocok untuk penelitian. Berikutnya dataset yang telah ditemukan akan diolah menggunakan metode yang ditentukan untuk mencari akurasi dari metode yang diterapkan, dalam penelitian ini digunakan atau dibandingkan akurasi dari kedua metode dalam mengklasifikasi data. Hasil akurasi setelah dibandingkan akan menghasilkan mana metode atau algoritma yang memiliki keunggulan atau memiliki tingkat akurasi yang lebih baik satu dengan yang lainnnya. Baru dapat ditarik kesimpulan berdasarkan eksperimen atau percobaan dan pengujian yang telah dilakukan.

Hasil dan Pembahasan

1. Pengolahan Dataset

Dataset yang diolah pada penelitian ini adalah sejumlah 7043 data dengan rincian data yang ditunjukan pada Table 1.

Table 1. Dataset

No	Nama
1.	Jenis Kelamin
2.	Warga Lokal
3.	Rekan
4.	Tanggungan
5.	Masa Jabatan
6.	Layanan Telepon
7.	Multi Layanan
8.	Layanan Internet
9.	Keamanan Online
10.	Cadangan Online
11.	Perlindungan Perangkat
12.	Teknologi
13.	Layanan Streaming TV
14.	Layanan Streaming Film
15.	Kontrak
16.	Tagihan
17.	Metode Pembayaran
18.	Biaya Bulanan
19.	Total Biaya
20.	Loyalitas

Masing-masing data tersebut merupakan dasar informasi yang digunakan dalam penelitian ini, data tersebut merupakan acuan dalam menentukan loyalitas pelanggan. Dataset akan diolah menggunakan metode/algoritma Naïve Bayes dan C4.5 untuk menghasilkan informasi berupa akurasi metode dalam mengklasifikasikan data.

Pengolahan data dilakukan dengan membagi data menjadi data testing dan data training. Adapun pembagian data training dan data testing dilakukan dengan detail yang ditunjukan pada Table 2.

Table 2. Pembagian Data Training dan Testing

Pengujian	Naïve Bayes		C4.5	
	Training	Testing	Training	Testing
1	4226	2.817	4226	2.817
2	3521	3.522	3521	3.522
3	2817	4.226	2817	4.226
4	2113	4.930	2113	4.930

Dari pembagian data training dan data testing tersebut akan menghasilkan akurasi kebenaran dalam melakukan klasifikasi data

2. Metode Naïve Bayes

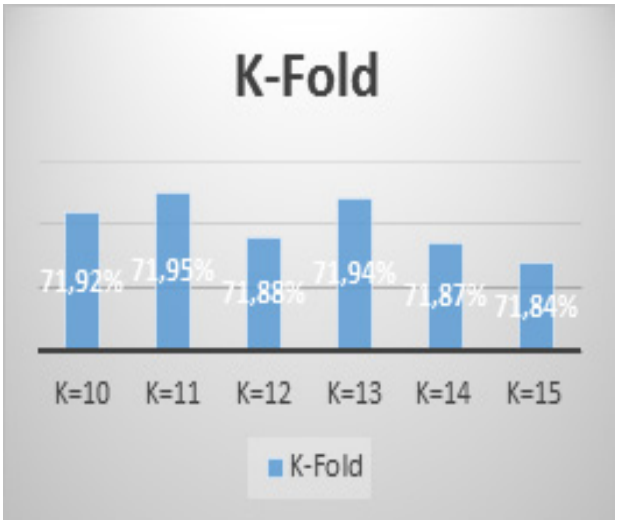
Pengujian metode Naïve Bayes dalam pembagian data training dan data testing sesuai yang telah dilakukan pada proses sebelumnya menghasilkan nilai Precision, Recall, dan F Measure yang ditunjukan pada Table 3.

Table 3. Pengujian Data Testing dan Training

Pengujian	Precision	Recall	F-Measure	Accuracy
1	0.800	0.719	0.735	71.8883
2	0.796	0.719	0.735	71.9398
3	0.790	0.709	0.726	70.891
4	0.793	0.709	0.727	70.8945

Pada pengujian yang telah dilakukan menggunakan metode Naïve Bayes maka prosentase akurasi paling tinggi ditunjukan pada pengujian 2 dengan nilai akurasi sebesar 71,9398 dengan jumlah data training sebesar 35.21 dan data testing sebesar 3.522.

Percobaan selanjutnya adalah dengan menguji algoritma menggunakan teknik cross validation. Pengujian cross validation adalah proses membagi data secara acak ke dalam beberapa bagian. Pengujian dilakukan sebanyak 5 kali. Akurasi pengujian ditunjukan pada Gambar 2.



Gambar 2. Pengujian Cross Validation

Pada hasil pengujian algoritma Naïve Bayes didapatkan hasil bahwa nilai K tertinggi didapatkan pada K=13 dengan akurasi sebesar 71,94%. Nilai pada K=13 dijabarkan lebih lanjut pada Table 4.

Table 4. Detail Akurasi K=13

Precision	Recall	F-Measure
0.797	0.719	0.735

3. Metode C4.5

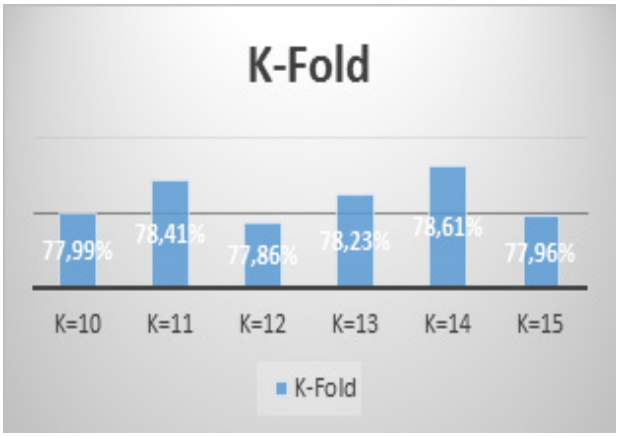
Pengujian metode C4.5 dalam pembagian data training dan data testing sesuai yang telah dilakukan pada proses sebelumnya menghasilkan nilai Precision, Recall, dan F Measure yang ditunjukan pada Table 5.

Table 5. Pengujian Data Testing dan Training

Pengujian	Precision	Recall	F-Measure	Accuracy
1	0.777	0.784	0.780	78.4193
2	0.769	0.778	0.773	77.762
3	0.766	0.772	0.769	77.2453
4	0.780	0.786	0.782	78.6086

Pada pengujian yang telah dilakukan menggunakan metode C4.5 maka prosentase akurasi paling tinggi ditunjukan pada pengujian 4 dengan nilai akurasi sebesar 78.6086 dengan jumlah data training sebesar 2113 dan data testing sebesar 4.930.

Pengujian yang kedua menggunakan teknik cross validation. Percobaan yang dilakukan menggunakan nilai yang sama dengan Algoritma Naïve Bayes, yaitu K=10 sampai dengan K=15. Akurasi yang dihasilkan algoritma C4.5 ditunjukan pada Gambar 3.



Gambar 3. Pengujian Cross Validation C4.5

Pada hasil pengujian algoritma C4.5 didapatkan hasil bahwa nilai K tertinggi didapatkan pada K=14 dengan akurasi sebesar 78,61%. Nilai pada K=14 dijabarkan lebih lanjut pada Table 6.

Table 6. Detail Akurasi K=14

Precision	Recall	F-Measure
0.775	0.786	0.778

Komparasi Metode

Berdasarkan percobaan yang telah dilakukan, didapatkan masing-masing hasil dari setiap

metode/algoritma. Adapun rangkuman hasil dari setiap percobaan ditunjukan pada Table 7 .

Table 7. Rangkuman Hasil Percobaan

Jenis Pengujian	Naïve Bayes	C4.5
Data Training & Data Testing	Nilai terbaik sebesar 71,9398 pada skenario pengujian ke-2	Nilai terbaik sebesar 78,6086 pada skenario pengujian ke-4
Cross Validation	Nilai terbaik pada K=13 dengan akurasi sebesar 71,94%.	Nilai terbaik pada K=14 dengan akurasi sebesar 78,61%.

Kesimpulan dan Saran

Kesimpulan yang dapat ditarik dari penelitian ini sebagai berikut :

a. Akurasi tertinggi pengujian klasifikasi dengan pembagian data training dan data testing adalah algoritma C4.5 dengan akurasi sebesar 78,6086.
b. Akurasi tertinggi pengujian klasifikasi dengan Cross Validation adalah algoritma C4.5 dengan akurasi sebesar 78,61%.
c. Algoritma C4.5 memiliki akurasi yang lebih tinggi berdasarkan 2 skenario pengujian yang telah dilakukan.
d. Nilai akurasi dari pengujian skenario pertama dan kedua masing-masing metode memiliki gap yang tidak terlalu jauh.

Adapun saran dari penelitian ini adalah sebagai berikut :

a. Dapat dilakukan skenario pengujian tambahan agar lebih meningkatkan kembali tingkat keabsahan pengujian.
b. Dapat dilakukan dengan menggunakan dataset yang lain untuk mencari hasil yang lebih baik dengan algoritma yang sama.
c. Dapat dilakukan pembuatan prototipe dalam bentuk software untuk pengembangan penelitian selanjutnya.

Referensi

[1]. Kurniawan, Y.I., 2018. Perbandingan Algoritma Naive Bayes dan C. 45 Dalam Klasifikasi Data Mining. Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK), 5(4), pp.455-464.

[2]. Darmawan, E., 2018. C4. 5 Algorithm Application for Prediction of Self Candidate New Students in Higher Education. Jurnal Online Informatika, 3(1), pp.22-28.

[3]. Idris, M. and Suseno, J.E., 2019. Implementation of C4. 5 Algorithm and Forward Chaining Method for Higher Education Performance Analysis. In E3S Web of Conferences (Vol. 125, p. 21002). EDP Sciences.

[4]. Wijaya, A. and Girsang, A.S., 2015. Use of Data Mining for Prediction of Customer Loyalty. CommIT (Communication and Information Technology) Journal, 10(1), pp.41-47.

[5]. Saiyed, S., Bhatt, N. and Ganatra, A.P., 2016. A Survey on Naive Bayes Based Prediction of Heart Disease Using Risk Factors. International Journal of Innovative and Emerging Research in Engineering, 3(2), pp.111-115.

[6]. E. T. L. Musthofa Galih Pradana, Kusrini, "PERBANDINGAN METODE WEIGHTED PRODUCT DAN SIMPLE ADDITIVE WEIGHTING DALAM SELEKSI PENGURUS FORUM ASISTEN (STUDI KASUS : UNIVERSITAS AMIKOM YOGYAKARTA)," vol. 4, no. 2, 2019.

[7]. Vivek Kale. Enterprise Performance Intelligence and Decision Patterns. CRC Press. 2018

[8]. Saxena, K. and Sharma, R., 2015, May. Efficient heart disease prediction system using decision tree. In International Conference on Computing, Communication & Automation (pp. 72-77). IEEE.

[9]. Budi Santosa. 2007. Data Mining Teknik Pemanfaatan Data Untuk Keperluan Bisnis. Graha Ilmu.