 ISSN : 2338-2899

Klasifikasi Perilaku Pelanggan Layanan Internet Menggunakan Metode K-Nearest Neighbors

Muhammad Khaibar Akbar[1], Muhammad Nuril Muqit[2], Muhammad Ikram Hidayat[3], dan Zaitun[4]

[1,2,3]Departemen Matematika, Universitas Hasanuddin, [4]Departemen Sains, Institut Teknologi Bacharuddin Jusuf Habibie

[1]akbarmk21h@student.unhas.ac.id, [2]muqitmn21h@student.unhas.ac.id, [3]hidayatmi21h@student.unhas.ac.id, [4]zaitun.zt99@ith.ac.id

***Abstract* — Customer behavior understanding is the key to improving satisfaction, sales, and operational efficiency. To assist internet services in predicting customer behavior, a machine learning approach is used to classify customer behavior. The most relevant variables to our classification results are determined using feature selection. After selecting features as attributes that influence the prediction of customer behavior, they are then classified using the K-Nearest Neighbors (KNN) method. Furthermore, a confusion matrix is used to measure classification performance, with an accuracy level of f1-score being 0.78. It is expected that the results will deepen the understanding of customer behavior and classification based on preferences and purchase history.**

***Keyword* — Internet service, *confusion matrix*, K-Nearest Neighbors (KNN)**

Abstrak — Pemahaman perilaku pelanggan menjadi kunci utama dalam meningkatkan kepuasan, penjualan, dan efisiensi operational. Untuk membantu layanan internet dalam memprediksi perilaku pelanggan, dilakukan dengan menggunakan pendekatanan machine learning dalam mengklasifikasikan perilaku pelanggan. Variabel yang paling relevan dengan hasil klasifikasi kami ditentukan dengan menggunakan feature selection. Setelah memilih fitur-fitur sebagai atribut yang berpengaruh terhadap prediksi perilaku pelanggan, kemudian diklasifikasikan menggunakan metode K – Nearest Neighbors (KNN). Selanjutnya confusion matrix digunakan untuk mengukur performa klasifikasi, tingkat akurasi adalah 0.78. Diharapkan, hasilnya akan mendalamkan pemahaman perilaku pelanggan dan klasifikasi berdasarkan preferensi serta Sejarah pembelian..

Kata kunci — Layanan internet, *confusion matrix*, K - Nearest Neighbors (KNN)

I. PENDAHULUAN

Dalam era digital saat ini, pemahaman perilaku pelanggan menjadi kunci utama bagi penyedia layanan internet untuk meningkatkan kepuasan pelanggan, meningkatkan penjualan, dan meningkatkan efisiensi operasional. Pemahaman mendalam terhadap preferensi dan sejarah pembelian pelanggan memungkinkan penyedia layanan internet untuk menyediakan layanan yang lebih terpersonalisasi dan efektif.

Pentingnya pemahaman perilaku pelanggan telah mendorong penggunaan pendekatan machine learning dalam mengklasifikasikan perilaku pelanggan. Dengan memprediksi perilaku pelanggan, penyedia layanan internet dapat lebih proaktif dalam menyediakan layanan yang sesuai dengan kebutuhan dan preferensi pelanggan.

Sehingga diperlukan suatu riset untuk menentukan klasifikasi perilaku pelanggan yang akan tetap berlangganan ataupun akan berhenti berlangganan berdasarkan preferensi serta sejarah pembelian agar penyedia layanan dapat mengambil langkah yang tepat kedepannya.

Dari uraian pengantar diatas dapat ditemukan titik-titik permasalahan diantaranya, percobaan menggunakan metode *feature selection* untuk menentukan variabel terbaik untuk diuji kemudian menggunakan *machine learning* yang lain yaitu *K - Nearest Neighbors* untuk mengklasifikasian perilaku pelanggan.

Klasifikasi perilaku pelanggan yang dilakukan oleh [1] menggunakan 26 atribut menhasilkan tingkat akurasi 59.02%. Setelah melakukan segmentasi, dengan 11 atribut (16 untuk *K-means*) memiliki hasil yang berbeda yaitu memiliki tingkat akurasi 77.31%

Penelitian yang terkait dengan metode yang menunjukan bahwa tingkat akurasi algoritma dalam pengklasifikasian relative tinggi. Hal ini dibuktikan dengan yang dilakukan oleh Rahmad Robi Waliyansah Perbandingan Akurasi Klasifikasi Citra Kayu Jati Menggunakan Metode Naive Bayes dan k-Nearest Neighbor (k-NN) dengan akurasi 82,7%, sedangkan dalam penelitian ClaudioFresta Suharno Klasifikasi Teks Bahasa Indonesia pada Dokumen Pengaduan Sambat Online menggunakan Metode K-Nearest Neighbors (K-NN) dan ChiSquaredengan akurasi 78%

Permasalahan-permasalah yang telah diteliti oleh peneliti sebelumnya terkait dengan klasifikasi, maka penelitian ini dilakukan untuk mengklasifikasikan dan juga membandingkan akurasi metode. Metode yang digunakan pada penelitian ini adalah *K-Nearest Neighbors* dengan menggunakan *feature selection* untuk memilih variabel yang paling relevan dengan hasil klasifikasi kami.

II. METODOLOGI PENELITIAN

A. Data dan Variabel

Data yang digunakan pada penelitian ini merupakan data sekunder yaitu data dari *kaggle* yang berjudul '*telco customer churn classification*' [2]. Data berupa hasil pengamatan faktor-faktor atau fitur-fitur yang diduga mempengaruhi perilaku pelanggan. Jumlah data yang digunakan sebanyak 7043 data dengan 20 atribut atau variabel.

Variabel dalam penelitian dapat dilihat pada Tabel 1.

Tabel 1. Variabel penelitian

Variabel	Keterangan	Variabel	Keterangan
X_1=gender	Apakah pelanggannya laki-laki atau perempuan	*X11=device protection*	Apakah pelanggan memiliki perlindungan perangkat atau tidak (Ya, Tidak, Tidak ada layanan internet)
X2=senior citizen	Apakah pelanggannya adalah warga lanjut usia atau bukan (1, 0)	*X12=tech support*	Apakah pelanggan memiliki dukungan teknis atau tidak (Ya, Tidak, Tidak ada layanan internet)
X3= partner	Apakah pelanggan mempunyai pasangan atau tidak (Ya, Tidak)	*X13=streaming tv*	Apakah pelanggan memiliki TV streaming atau tidak (Ya, Tidak, Tidak ada layanan internet)
X4=dependents	Apakah pelanggan mempunyai tanggungan atau tidak (Ya, Tidak)	*X14=streaming movies*	Apakah pelanggan memiliki streaming film atau tidak (Ya, Tidak, Tidak ada layanan internet)
X5=tenure	Jumlah bulan pelanggan telah tinggal di perusahaan	*X15= contract*	angka waktu kontrak pelanggan (Bulan ke bulan, Satu tahun, Dua tahun)
X6=phone service	Apakah pelanggan memiliki layanan telepon atau tidak (Ya, Tidak)	*X16=paperless billing*	Apakah pelanggan memiliki tagihan tanpa kertas atau tidak (Ya, Tidak)
X7=multiple lines	Apakah pelanggan memiliki banyak saluran atau tidak (Ya, Tidak, Tidak ada layanan telepon)	*X17=payment method*	Metode pembayaran pelanggan
X8=internet service	Penyedia layanan internet pelanggan (DSL, Fiber optic, No)	*X18=monthly charge*	Jumlah yang dibebankan kepada pelanggan setiap bulan
X9=online security	Apakah pelanggan memiliki keamanan online atau tidak (Ya, Tidak, Tidak ada layanan internet)	*X19=total charge*	Jumlah total yang dibebankan kepada pelanggan
X10=online backup	Apakah pelanggan memiliki cadangan online atau tidak (Ya, Tidak, Tidak ada layanan internet)		
Y = churn (0 = No, 1 = Yes)			

Variabel yang tertera dalam Tabel 1 akan dijadikan atribut dalam penelitian. Sedangkan klasifikasi yang akan digunakan adalah perilaku pelanggang yang melakukan *churn* yang terdiri dari 2 kelas yaitu ya atau tidak.

B. Metode dan Tahapan Analisis

Tahapan analisis yang dilakukan dalam penelitian ini digambarkan pada diagram alur sebagai berikut.

1. Preprocessing

Preprocessing data adalah tahapan dari Data Mining yaitu suatu proses atau tahapan yang dilakukan untuk mengolah data mentah menjadi data yang berkualitas atau inputan yang baik untuk dilanjutkan ke proses selanjutnya.

2. Feature Selection

Feature Selection merupakan tahapan untuk memilih fitur-fitur dalam dataset yang memiliki korelasi terhadap kelasnya. Tujuan utama dari feature selection adalah untuk meningkatkan performa model dengan mengurangi kompleksitas, mengurangi overfitting, dan mempercepat proses pelatihan model. Metode yang dapat digunakan ialah *analisis of variance* (ANOVA) dan *chi square*.

Seleksi fitur ANOVA menggunakan uji-F untuk menentukan apakah variabilitas antara ratarata kelompok fitur lebih besar daripada variabilitas pengamatan di dalam kelompok fitur. UjiF sendiri adalah salah satu uji statistik yang memberikan nilai-f dengan cara menghitung rasio antar varians[3]. *Chi-square* disebut juga dengan Kai Kuadrat. Uji *Chi-square* adalah salah satu jenis uji komparatif non parametris yang dilakukan pada dua variabel, di mana skala data kedua variabel adalah nominal (Sutrisno, 2000). Setelah melakukan *feature selection* fitur yang paling penting yaitu *Tenure, OnlineSecurity, TechSupport, Contract, MonthlyCharges, TotalCharges, Tenure_year.*

3. Membagi data

Data penelitian yang telah melalui *feature selection* kemudian dibagi menjadi 2 yaitu data *training* dan data *testing.* Data training yang telah diambil digunakan untuk membuat model sekaligus uji akurasinya dengan metode K - Nearest Neighbors. Sedangkan data testing digunakan untuk menguji akurasi model yang telah dibuat. Dengan mengambil rasio 80% untuk data *training* dan 20% untuk data *testing*, didapatkan 5634 data sebagai data *training* dan 1409 data *testing.*

4. Klasifikasi

Klasifikasi adalah suatu tugas dalam machine learning yang melibatkan pelatihan model untuk memahami hubungan antara atribut (fitur) dari data latih dan label kelas yang sesuai. Dengan kata lain, tujuannya adalah membangun model yang dapat "mempelajari" pola dari data latih sehingga dapat memberikan prediksi kelas yang akurat untuk data baru yang belum pernah dilihat.

Proses klasifikasi dimulai dengan memberikan model sejumlah besar data latih yang sudah memiliki label kelas yang ditentukan. Model ini kemudian menggunakan data tersebut untuk mengidentifikasi pola atau hubungan antara fitur-fitur yang ada dengan label kelas yang sesuai. Setelah model terlatih, kita dapat menguji keberhasilannya dengan memberikan data baru yang tidak pernah dilihat sebelumnya.

Pada tahap pengujian, model akan melakukan prediksi kelas untuk setiap data baru berdasarkan pola yang telah dipelajarinya dari data latih. Keberhasilan model diukur dengan membandingkan prediksi yang dihasilkan dengan label kelas yang sebenarnya pada data pengujian. Metrik evaluasi, seperti akurasi, presisi, recall, dan F1-score, digunakan untuk mengukur sejauh mana model mampu mengklasifikasikan data dengan benar.

Metode *machine learning* yang digunakan untuk klasifikasi perilaku pelanggan pada penelitian ini adalah *K-Nearest Neighbors.* K-Nearest Neighbor merupakan salah satu metode untuk mengambil keputusan menggunakan pembelajaran terawasi dimana hasil dari data masukan yang baru diklasifikasi berdasarkan terdekat dalam data nilai [4].

Algoritma K-Nearest Neighbor (KNN) adalah sebuah metode untuk melakukan klasifikasi terhadap objek yang berdasarkan dari data pembelajaran yang jaraknya paling dekat dengan objek tersebut. KNN merupakan algoritma supervised learning dimana hasil dari query instance yang baru diklasifikan berdasarkan mayoritas dari kategori pada algoritma KNN. Dimana kelas yang paling banyak muncul yang nantinya akan menjadi kelas hasil dari klasifikasi [5].

Kedekatan didefinisikan dalam jarak metrik, seperti jarak Euclidean. Jarak Euclidean [6] dapat dicari dengan menggunakan persamaan 1 berikut ini:

$$\boldsymbol{Dxy} = \sqrt{\sum_{i=1}^{n}(\boldsymbol{x_i} - \boldsymbol{y_i})^2} \quad (1)$$

Keterangan :
- □ : jarak kedekatan
- □ : data training
- □ : data testing
- □ : jumlah atribut individu antara 1 s.d. □
- □ : fungsi similitary atribut □ antara kasus □ dan kasus □
- □ = Atribut individu antara 1 sampai dengan □

Langkah-langkah untuk menghitung metode KNearest Neighbor [7] antara lain :

1. Menentukan parameter □ (jumlah tetangga paling dekat).
2. Menghitung kuadrat jarak Euclid (queri instance) masing-masing objek terhadap data sampel yang diberikan menggunakan persamaan 1.
3. Kemudian mengurutkan objek-objek tersebut ke dalam kelompok yang mempunyai jarak Euclid terkecil.
4. Mengumpulkan kategori □ (Klasifikasi Nearest Neighbor)
5. Dengan menggunakan kategori Nearest Neighbor yang paling mayoritas maka dapat diprediksi nilai query instance yang telah dihitung.

4. Evaluasi Kinerja

Evaluasi kinerja dilakukan dengan melihat tingkat performa dari pola yang dihasilkan oleh algoritma. *Confusion Matrix* merupakan parameter yang digunakan untuk evaluasi komparasi algoritma tersebut. Jadi, dalam

tahap ini pengujian akan diperoleh nilai akurasi *f1-score*, *recal* dan *precision*.

F1-score adalah metrik gabungan yang mengukur keseimbangan antara *precision* dan *recall* didapatkan dengan persamaan 2. *Recal* adalah proporsi prediksi positif yang benar dari semua data positif yang sebenarnya. *Recall* didapatkan dengan Persamaan 3. *Precision* adalah proporsi prediksi positif yang sebenarnya, di dapatkan dengan Persamaan 4. *Accuracy* adalah proporsi prediksi yang benar (baik positif maupun negatif) dari semua data didapatkan dengan persamaan 5..

$$\square 1 - \square\square\square\square\square = \frac{2 \square \square\square\square\square\square\square\square\square \square \square\square\square\square\square\square}{\square\square\square\square\square\square\square\square\square + \square\square\square\square\square\square} \quad (2)$$

$$\square\square\square\square\square\square = \frac{\square\square}{\square\square + \square\square} \quad (3)$$

$$\square\square\square\square\square\square\square\square\square = \frac{\square\square}{\square\square + \square\square} \quad (4)$$

$$\square\square\square\square\square\square\square\square = \frac{\square\square + \square\square}{\square\square + \square\square + \square\square + \square\square} \quad (5)$$

III. Hasil Dan Pembahasan

A. Analisis Deskriptif

Berdasarkan data yang diperoleh pada penelitian ini, banyaknya data perilaku pelanggan (*churn)* adalah 7043 data. Perilaku pelanggan tersebut terbagi atas 2 yaitu pelanggan meninggalkan layanan (*churn*-nya yes) sebagai kelas 1 dengan jumlah data 1869 dan pelanggan tidak meninggalkan layanan (*churn*-nya no) sebagai kelas 0 dengan jumlah data 5174.

Berdasarkan data untuk masing-masing kelas *churn* data kurang seimbang. Atribut yang paling mempengaruhi rentang harga tersebut berdasarkan *feature selection* ada 7 atribut yaitu *Tenure, OnlineSecurity, TechSupport, Contract, MonthlyCharges, TotalCharges,Tenure_year*. Atribut-atribut tersebut menjadi variabel prediktor dalam penelitian ini dan dapat digunakan untuk menentukan klasifikasi perilaku pelanggan (*churn*).

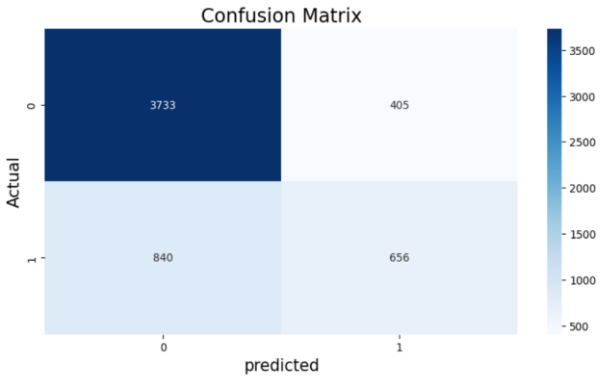
B. Klasifikasi dengan K-Nearest Neighbors

Analisis klasifikasi dilakukan setelah melalui tahap *preprocessing* dan *feature selection*. Analisis ini untuk mengetahui klasifikasi perilaku pelanggan (*churn*) yang dibagi ke dalam dua kategori yaitu ya atau tidak. Tahap pertama penelitian ini yaitu membagi data menjadi dua yaitu data *training* untuk membentuk modelnya dan data *testing* untuk evaluasi kinerjanya dengan persentase 80% dan 20% dari total 7043 data.

Setelah pembagian data, selanjutnya dilakukan klasifikasi perilaku pelanggan (*churn*) menggunakan metode *K-Nearest Neighbors*..

C. Evaluasi Kinerja

Tahap evaluasi kinerja pada penelitian ini menggunakan *confusion matrix*. *Confusion matrix* mengandung informasi yang membandingkan hasil klasifikasi yang dilakukan oleh sistem dengan hasil klasifikasi yang seharusnya. Terdapat 3 istilah dalam evaluasi kinerja menggunakan *confusion matrix* yaitu *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN). *Confusion matrix* dapat dilihat pada Gambar 1.



Gambar 1. *Confusion Matrix*

Dari *confusion matrix* yang telah diperoleh kemudian dicari *f1-score*, *recall* dan *precision* dari hasil klasifikasi menggunakan rumus pada Persamaan 2, Persamaan 3 dan Persamaan 4.

$$\square\square\square\square\square\square\ (\ \square\square\square\square\square\ 0) = \frac{3733}{4138} = 0.90$$

$$\square\square\square\square\square\square\ (\square\square\square\square\square\ 1) = \frac{656}{1496} = 0.44$$

$$\square\square\square\square\square\square\square\square\square\square\ (\square\square\square\square\square\ 0) = \frac{3733}{4573} = 0.80$$

$$\square\square\square\square\square\square\square\square\square\square\ (\square\square\square\square\square\ 1) = \frac{656}{1061} = 0.62$$

$$\square 1 - \square\square\square\square\square\ (\square\square\square\square\square\ 0) = \frac{2 \square 0.90 \square 0.82}{0.90 + 0.82} = 0.86$$

$$\square 1 - \square\square\square\square\square\ (\square\square\square\square\square\ 1) = \frac{2 \square 0.44 \square 0.62}{0.44 + 0.62} = 0.51$$

$$\square\square\square\square\square\square\square\square\square = \frac{4389}{5643} = 0.78$$

Nilai akurasi yang diperoleh adalah 0.78 menunjukkan sebesar 0.78 klasifikasi dapat memprediksi kelas yang benar dari data *testing.* Nilai *f1-score (kelas 0)* adalah 0.86 dan *f1-score (kelas 1)* adalah 0.51 adalah sebesar itu model bekerja untuk kelas tertentu. Nilai *precision (kelas 0)* adalah 0.80 dan *precision (kelas 1)* adalah 0.62, yang artinya sebesar itu model mampu mengidentifikasi positif dari prediksi positif dan seberapa sedikit positif palsu yang dihasilkan. Nilai *recall (kelas 0)* adalah 0.90 dan *recall (kelas 1)* adalah 0.44 yang artinya sebesar itu juga model mampu menangkap semua insiden positif.

IV. KESIMPULAN

Hasil klasifikasi perilaku pelanggan (*churn*) dengan *feature selection* dengan metode *analisis of variance dan chi-square* membuktikan *Tenure, OnlineSecurity, TechSupport,Contract,MonthlyCharges,TotalChargesTenure year* merupakan 7 variabel yang paling relevan untuk model. Kemudian menggunakan metode *machine learning K-Nearest Neighbors* dengan data *training* 80% dan data *testing* 20% diperoleh akurasi sebesar 0.78, *precision* kelas-0 sebesar 0.80, *precision* kelas 1 sebesar 0.62, *recall* kelas 0 sebesar 0.90, *recall* kelas 1 sebesar 0.44, *f1-score* kelas 0 sebesar0.86, *f1-score* kelas 1 sebesar 0.51.. Hasil tersebut dapat dikategorikan sangat baik dan dapat disimpulkan bahwa klasifikasi dengan f*eature selection* menggunakan metode *analisis of variance dan chi-square* dan metode *machine learning K-Nearest Neighbors* dapat meningkatkan akurasi klasifikasi perilaku pelanggan (*churn)*

.

DAFTAR ACUAN

[1] Ela Nur Ela Sari, "Segmentasi Dan Klasifikasi Perilaku Pembayaran Pelanggan Pada Perusahaan Multimedia Dengan Algoritma K-Means Dan C4.5", vol.21, no. 1 Maret 2019.

[2] Telco Customer Churn. Kaggle. Available: https://www.kaggle.com/.

[3] Dimas Ariyoga, "Perbandingan Metode Seleksi Fitur Filter Wrapper dan Embedded pada Klasifikasi Data Nirs Mangga Menggunakan Random Forest dan Support Vekctor Machine", 2022.

[4] Teknomo, K, "What is K-Nearest Neighbor Algoritm ?", 2006.

[5] Avelita, B. , "Klasifikasi_K-Nearest_Neighbor", 2013.

[6] Jiawei Han, M. K.,"Data Mining : Concepts and Techniques. United States of America: Elsevier", 2012.

[7] Ndaumanu, R. I."Analisis Prediksi Tingkat Pengunduran Diri Mahasiswa dengan Metode K-Nearest Neighbor". Jatisi Vol 1, 3, 2014.