PENERAPAN NAÏVE BAYES UNTUK PREDIKSI CUSTOMER CHURN (STUDI KASUS: PT HUTCHISON 3 INDONESIA)

Rifky Alfarez[1], Rianto[2], Vega Purwayoga[3]
Program Studi Informatika, Fakultas Teknik, Universitas Siliwangi
Jalan Siliwangi No 24, Kelurahan Kahuripan, Kecamatan Tawang, Kota Tasikmalaya
207006052@student.unsil.ac.id[1], rianto@unsil.ac.id[2], vega.purwayoga@unsil.ac.id[3]

Abstrak

Penelitian ini memiliki maksud untuk mengembangkan model prediksi dengan memanfaatkan metode Naive Bayes. Pemodelan algoritma Naive Bayes dilakukan dengan menerapkan penggunaan Jupyter Notebook dan mengacu pada dataset yang terdiri dari atribut durasi langganan, frekuensi transaksi, tingkat kepuasan, dan status churn. Pada tahap awal, dilakukan eksplorasi data untuk memahami distribusi dan karakteristik atribut. Kemudian, dilakukan pengolahan data dengan menghapus kolom yang tidak relevan, memisahkan dataset menjadi data pelatihan dan data pengujian. Berikutnya, dilakukan pemodelan Naive Bayes dengan menghitung probabilitas kemunculan setiap nilai atribut untuk kelas churn true dan false. Probabilitas ini digunakan dalam perhitungan prediksi kelas churn berdasarkan atribut yang diberikan. Setelah model Naive Bayes terbentuk, dilakukan evaluasi performa model menggunakan metrik evaluasi. Evaluasi dilakukan dengan membandingkan kelas prediksi dengan kelas aktual pada data pengujian. Evaluasi menunjukkan model Naive Bayes menghasilkan akurasi sebesar 91,3%. Presisi, recall, dan F1-score sebesar 95% menunjukkan kemampuan model dalam mengklasifikasikan data churn dengan tingkat keakuratan yang tinggi.

Kata Kunci : Customer churn, Naïve bayes, Prediksi.

Abstract

The objective of this study is to create a predictive model by employing the Naive Bayes method. The Naive Bayes algorithm is implemented using Jupyter Notebook and applied to a dataset containing attributes such as subscription duration, transaction frequency, satisfaction level, and churn status. Initially, data exploration is conducted to gain insights into the attribute distribution and characteristics. Subsequently, data processing entails eliminating irrelevant columns and dividing the dataset into separate training and testing sets. The Naive Bayes modeling is then conducted by calculating the probabilities of attribute values for both true and false churn classes. These probabilities are used to predict the churn class based on the given attributes. Once the Naive Bayes model is established, its performance is assessed using evaluation metrics. The evaluation involves comparing the predicted class with the actual class in the testing set. The evaluation results demonstrate an accuracy of 91.3% for the Naive Bayes model. The precision, recall, and F1-score, which are all at 95%, indicate the model's high accuracy in classifying churn data.

***Keyword** : Customer churn, Naïve bayes, Predict.*

PENDAHULUAN

Customer churn adalah sebuah istilah yang digunakan untuk menggambarkan keadaan di mana pelanggan berhenti menggunakan produk atau layanan suatu perusahaan (Senthilnayaki et al., 2021). Faktor yang menyebabkan hal ini terjadi seperti harga yang lebih mahal dari pesaing, kurang adanya promo ataupun penawaran, dan kualitas layanan yang kurang memuaskan. Customer churn bisa menyebabkan dampak negatif pada keuangan dan reputasi perusahaan, sehingga perusahaan perlu mengidentifikasi dan mencegah churn pelanggan (Ahn et al., 2006).

Naive Bayes adalah metode klasifikasi yang memanfaatkan teori probabilitas untuk memprediksi suatu kejadian. Dalam konteks prediksi customer churn, metode ini akan menghitung peluang terjadinya churn atau tidak churn berdasarkan informasi yang ada pada dataset, seperti durasi penggunaan layanan, jumlah panggilan atau SMS, jumlah tagihan, dan lain sebagainya (Huang et al., 2012).

Telah dilakukan studi literatur mengenai masalah customer churn yang dihadapi oleh perusahaan telekomunikasi. Pemahaman yang lebih menjadi tujuan dari studi literatur, serta untuk mengetahui juga faktor-faktor apa saja yang memengaruhi permasalahan tersebut. Selain itu, studi literatur juga

bertujuan untuk melihat batasan-batasan yang ada pada penelitian sebelumnya yang dapat menjadi peluang penelitian baru.

Pada penelitian yang dilakukan Yulianto dan Firmansyah (2021) dilakukan prediksi tingkat churn pelanggan dalam bisnis ritel. Diambil data dan dilakukan proses klasifikasi terlebih dahulu lalu data pelatihan dan data pengujian dihitung oleh metode naïve bayes dan menghasilkan akurasi sebesar 80%. Tingkat akurasi tersebut menjadi batasan dalam penelitian yang dilakukan Yulianto dan Firmansyah (2021) yang mana tingkat akurasi masih bisa ditingkatkan dengan melakukan peningkatan data training.

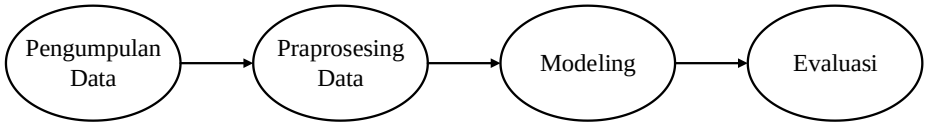
Oleh karena itu, penelitian ini difokuskan dalam peningkatan akurasi menggunakan metode naïve bayes dengan cara meningkatkan jumlah data traning. Dengan adanya peningkatan akurasi prediksi, diharapkan perusahaan dapat mengidentifikasi faktor-faktor apa yang berkontribusi terhadap terjadinya customer churn, mengambil keputusan yang lebih baik dalam mengelola hubungan dengan pelanggan, dan mengurangi dampak negatif dari churn pelanggan pada bisnis perusahaan.

PENELITIAN RELEVAN

Penelitian yang dilakukan Yulianto dan Firmansyah (2021) dilakukan prediksi tingkat churn pelanggan dalam bisnis ritel. Didapatkan akurasi sebesar 80%, yang mana hasil tersebut masih bisa ditingkatkan dengan menambah data training. Selain itu penelitian yang dilakukan Novendri (2021) mengembangkan model prediksi dengan akurasi yang mampu mencapai 83,02% untuk melakukan prediksi customer churn. Penelitian sebelumnya telah menjadi referensi utama dalam penelitian ini, di mana fokus ditujukan untuk akan difokuskan dalam meningkatkan tingkat akurasi.

METODE PENELITIAN

Pada penelitian ini, terdapat empat tahapan yang dilakukan, yaitu pengumpulan data, pemrosesan data, pemodelan, dan evaluasi. Gambar 1 menunjukkan setiap tahapan yang ada.



Gambar 1. Alur penelitian

Pengumpulan Data

Data dikumpulkan melalui proses penyebaran kuesioner yang dilaksanakan secara online pada sebuah grup facebook.

```
(1000, 5)
```

	age	durasi_langganan(hari)	frekuensi_transaksi	tingkat_kepuasan	churn
0	45	740	7	10	False
1	54	514	3	10	False
2	81	820	5	10	True
3	85	420	8	10	False
4	45	483	10	2	False
5	48	467	10	5	True
6	19	537	1	7	False
7	58	143	10	9	False
8	25	257	3	4	False
9	56	785	9	6	True

Gambar 2. Jumlah baris dan kolom

Tabel 1. Keterangan setiap atribut

No	Atribut	Keterangan
1	age	Umur pengguna
2	Durasi_langganan(hari)	Durasi lamanya pengguna menggunakan layanan dalam satuan hari
3	Frekuensi_transaksi	Seberapa sering pengguna melakukan transaksi dalam rentan 1-10
4	Tingkat_kepuasan	Tingkat kepuasan pengguna dalam rentang 1-10
5	churn	Meninggalkan layanan atau tidak (true/false)

Pra-prosesing Data

Praprosesing data ini bertujuan untuk mempersiapkan data agar dapat diolah oleh algoritma Naive Bayes dengan lebih baik. Tahapan-tahapan praprosesing data yang akan dilakukan yaitu cleaning data dan klasifikasi (Chandrasekar & Qian, 2016).

	durasi_langganan(hari)	frekuensi_transaksi	tingkat_kepuasan	churn
0	740	7	10	False
1	514	3	10	False
2	820	5	10	True
3	420	8	10	False
4	483	10	2	False
5	467	10	5	True
6	537	1	7	False
7	143	10	9	False
8	257	3	4	False
9	785	9	6	True

Gambar 3. *Cleaning data*

Pada (Gambar 3) dilakukan cleaning data yaitu untuk menghilangkan kolom yang tidak relevan, pada proses ini dilakukan penghapusan pada kolom age.

	durasi_langganan(hari)	frekuensi_transaksi	tingkat_kepuasan	churn
0	sedang	sedang	tinggi	False
1	sedang	rendah	tinggi	False
2	panjang	sedang	tinggi	True
3	pendek	tinggi	tinggi	False
4	pendek	tinggi	rendah	False
5	pendek	tinggi	sedang	True
6	sedang	rendah	sedang	False
7	pendek	tinggi	tinggi	False
8	pendek	rendah	sedang	False
9	sedang	tinggi	sedang	True

Gambar 4. Klasifikasi data

Pada (Gambar 4) dilakukan proses klasifikasi data yaitu melakukan pengkategorian pada data yang awalnya berbentuk numerik diubah menjadi bentuk kategori.

Pemodelan

Pemodelan algoritma naïve bayes dilakukan menggunakan jupyter notebook. Teorema Bayes merupakan dasar yang digunakan dalam pemrograman untuk mengimplementasikan algoritma Naive Bayes (Mustafa et al., 2018). Rumus yang terkandung dalam teorema Bayes adalah:

$P(A|B) = (P(B|A) * P(A))/P(B)$

Rumus ini menggambarkan peluang kejadian A terjadi ketika B telah terjadi. Peluang tersebut dihitung berdasarkan peluang B terjadi jika A telah terjadi (P(B|A)), peluang A terjadi (P(A)), dan peluang B terjadi (P(B)). Dalam konteks Naive Bayes, rumus ini digunakan untuk menghitung probabilitas kelas atau label tertentu (A) berdasarkan fitur atau atribut yang diamati (B). Dengan menggunakan rumus ini, kita dapat mengestimasi peluang suatu kejadian atau klasifikasi berdasarkan informasi yang ada. Rumus Bayes merupakan landasan matematis yang mendasari algoritma Naive Bayes dan digunakan untuk membuat prediksi atau klasifikasi berdasarkan data yang diamati (Tang et al., 2016). Penerapannya dalam jupyter notebook adalah sebagai berikut:

```
# Menentukan probabilitas
P_B_given_A = 0.9  # Probabilitas kejadian B jika kejadian A terjadi
P_A = 0.3  # Probabilitas kejadian A
P_B = 0.6  # Probabilitas kejadian B

# Menghitung probabilitas kejadian A jika kejadian B terjadi
P_A_given_B = (P_B_given_A * P_A) / P_B

# Menampilkan hasil
print("P(A|B) =", P_A_given_B)
```

✓ 0.0s

```
P(A|B) = 0.45000000000000007
```

Gambar 5. Contoh penerapan naïve bayes pada jupyter notebook

Evaluasi

Untuk mengevaluasi hasil dari algoritma Naive Bayes, terdapat beberapa metrik yang umum digunakan:

Akurasi (Accuracy): Mengukur persentase kebenaran prediksi model terhadap data yang diamati. Perbandingan antara jumlah prediksi yang benar dengan jumlah total prediksi dinyatakan sebagai akurasi.

Presisi (Precision): Mengevaluasi tingkat kebenaran dari prediksi positif. Perhitungan presisi melibatkan pembagian antara jumlah prediksi positif yang benar dengan jumlah total prediksi positif.

Recall (Recall): Mengukur sejauh mana model dapat menemukan atau mengenali kasus positif. Recall dinyatakan sebagai perbandingan antara jumlah prediksi positif yang benar dengan jumlah total kasus positif yang sebenarnya.

F1-Score: Menggabungkan presisi dan recall menjadi satu skor yang seimbang. F1-Score dinyatakan sebagai rata-rata harmonik antara presisi dan recall (Muzakir & Wulandari, 2016)..

Pada penelitian ini, proses evaluasi akan dilakukan menggunakan confusion matrix.

HASIL DAN PEMBAHASAN

Data training digunakan untuk melatih model agar menghasilkan model prediksi untuk *customer churn*. Setelah itu, sebagai implementasi dari algoritma, data pengujian dimasukkan pada model tersebut untuk melakukan prediksi churn pelanggan. Diketahui contoh kasus jika durasi langganan = pendek, frekuensi transaksi = sedang, dan tingkat kepuasan = tinggi. Maka dilakukan beberapa tahapan sebagai berikut.

 e-ISSN : 2715-8756

Probabilitas Kemunculan Setiap Atribut

churn	False	True
durasi_langganan(hari)		
panjang	134	108
pendek	217	226
sedang	150	165

```
Sum of True: 499
Sum of False: 501
```

Gambar 6. Durasi langganan

Probabilitas kemunculan setiap nilai untuk atribut durasi langganan:
P(durasi langganan=pendek | churn=true) = 217/499
P(durasi langganan=pendek| churn=false) = 226/501

churn	False	True
frekuensi_transaksi		
rendah	161	159
sedang	192	214
tinggi	148	126

```
Sum of True: 499
Sum of False: 501
```

Gambar 7. Frekuensi transaksi

Probabilitas kemunculan setiap nilai untuk atribut frekuensi transaksi:
P(frekuensi transaksi=sedang | churn=true) = 214/499
P(frekuensi transaksi=sedang| churn=false) = 192/501

churn	False	True
tingkat_kepuasan		
rendah	153	160
sedang	186	196
tinggi	162	143

```
Sum of True: 499
Sum of False: 501
```

Gambar 8. Tingkat kepuasan

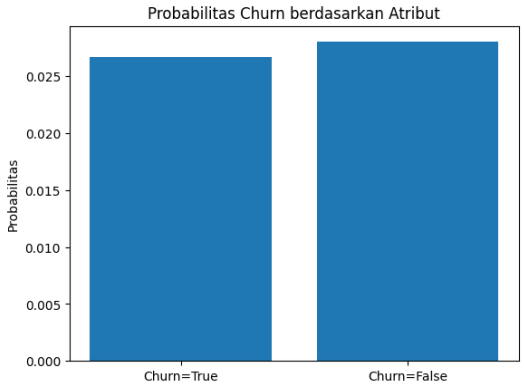
Probabilitas kemunculan setiap nilai untuk atribut tingkat kepuasan:
P(tingkat kepuasan=tinggi| churn=true) = 143/499
P(tingkat kepuasan=tinggi| churn=false) = 162/501

Jumlahkan semua variabel true dan false:
True = P(durasi langganan=pendek | churn=true) * P(frekuensi transaksi=sedang | churn=true) * P(tingkat kepuasan=tinggi | churn=true) * P(churn=true)
True = (217/499) * (214/499) * (143/499) * (499/1000) = 0.026669105746563265

False = P(durasi langganan=pendek | churn=false) * P(frekuensi transaksi=sedang | churn=false) * P(tingkat kepuasan=tinggi | churn=false) * P(churn=false)
False = (226/501) * (192/501) * (162/501) * (501/1000) = 0.028005880454659543

```
Probabilitas Churn=True: 0.026669105746563265
Probabilitas Churn=False: 0.028005880454659543
```

Gambar 9. Perhitungan menggunakan jupyter notebook



Gambar 10. Visualisasi hasil perhitungan

Nilai probabilitas tertinggi terletak pada variabel false yang berarti pelanggan tersebut tidak akan melakukan churn.

Evaluasi

Akurasi data diukur menggunakan confusion matrix sebagai berikut.

Tabel 2. Confusion matrix

n = 115	**Aktual : Positif(1)**	**Aktual : Negatif (0)**
Prediksi : Positif(1)	TP = 95	FP = 5
Prediksi : Negatif (0)	FN = 5	TN = 10
	100	15

Akurasi merupakan metrik yang menunjukkan seberapa tepat data diklasifikasikan oleh model.
Rumus akurasi = (TP+TN) / (TP+FP+FN+TN)
= (95+10)/ (115)
= 0.913
= 0.913*100 % = 91,3 %

Presisi mencerminkan sejauh mana hasil prediksi dari model sesuai dengan data yang diminta atau diharapkan.
Rumus presisi = (TP) / (TP + FP)
= 95 / (95 + 5)
= 0.95
= 0.95 * 100% = 95%

Recall mencerminkan seberapa berhasil model dalam mendeteksi atau menemukan kembali informasi yang relevan.
Rumus recall = TP / (TP + FN)
= 95 /(95+5)
= 0.95
= 0.95 * 100% = 95%

F1-Score mengukur perbandingan rata-rata yang tereksplisit antara presisi dan recall. Akurasi digunakan sebagai referensi performa algoritma saat jumlah data False Negatif dan False Positif relatif seimbang (simetris). Namun, jika jumlahnya tidak seimbang, F1-Score lebih disarankan sebagai ukuran referensi yang lebih tepat (Xu et al., 2020).

F-1 Score = (2 * Recall * Precision) / (Recall + Precision)
= (2* 0.95 * 0.95) / (0.95 + 0.95)
= 1.805 / 1,9
= 0.95 *100%
= 95%

SIMPULAN

Dari pemodelan naïve bayes ini menghasilkan akurasi sebesar 91,3% menunjukkan tingkat kesesuaian antara kelas prediksi dengan kelas aktual. Hal ini mengindikasikan performa yang baik dalam memprediksi kelas data. Presisi sebesar 95% menggambarkan kemampuan model untuk memberikan prediksi yang benar terhadap kelas positif. Recall sebesar 95% menggambarkan kemampuan model dalam mengidentifikasi keseluruhan jumlah data yang sesuai dengan kelas positif. F1-Score sebesar 95% merupakan nilai gabungan antara presisi dan recall, yang menunjukkan keseimbangan antara kedua metrik tersebut. Berdasarkan hasil evaluasi tersebut, dapat disimpulkan bahwa model klasifikasi memiliki performa yang baik dalam memprediksi kelas data dengan tingkat akurasi yang tinggi. Tingginya presisi, recall, dan F1-Score juga menunjukkan kemampuan model dalam mengklasifikasikan data dengan baik tanpa mengorbankan kualitas prediksi. Namun, tingkat akurasi masih dapat ditingkatkan lagi dengan menambah data training ataupun menggunakan algoritma lain.

DAFTAR PUSTAKA

Ahn, J., Han, S., & Lee, Y. (2006). Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry. *Telecommunications Policy*, *30*(10–11), 552–568. https://doi.org/10.1016/j.telpol.2006.09.006

Chandrasekar, P., & Qian, K. (2016). *The Impact of Data Preprocessing on the Performance of a Naive Bayes Classifier*. https://doi.org/10.1109/compsac.2016.205

Huang, B., Kechadi, T., & Buckley, B. S. (2012). Customer churn prediction in telecommunications. *Expert Systems With Applications*, *39*(1), 1414–1425. https://doi.org/10.1016/j.eswa.2011.08.024

Mustafa, M. S., Ramadhan, M., & Thenata, A. P. (2018). Implementasi Data Mining untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier. *Citec (Creative Information Technology) Journal*. https://doi.org/10.24076/citec.2017v4i2.106

Muzakir, A., & Wulandari, R. (2016). Model Data Mining sebagai Prediksi Penyakit Hipertensi Kehamilan dengan Teknik Decision Tree. *Scientific Journal of Informatics*. https://doi.org/10.15294/sji.v3i1.4610

Novendri, R. (2021). IMPLEMENTASI DATA MINING UNTUK MEMPREDIKSI CUSTOMER CHURN MENGGUNAKAN ALGORITMA NAIVE BAYES. *eProceedings of Engineering*, *8*(2). https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/14678

Senthilnayaki, B., Swetha, M., & Nivedha, D. (2021). CUSTOMER CHURN PREDICTION. *International Advanced Research Journal in Science, Engineering and Technology*, *8*(6), 527–531. https://doi.org/10.17148/iarjset.2021.8692

Tang, B., Kay, S., & He, H. (2016). Toward Optimal Feature Selection in Naive Bayes for Text Categorization. *IEEE Transactions on Knowledge and Data Engineering*, *28*(9), 2508–2521. https://doi.org/10.1109/tkde.2016.2563436

Xu, J., Zhang, Y., & Miao, D. (2020). Three-way confusion matrix for classification: A measure driven view. *Information Sciences*, *507*, 772–794. https://doi.org/10.1016/j.ins.2019.06.064

Yulianto, A., & Firmansyah. (2021). Prediksi Customer Churn Pada Bisnis Retail Menggunakan Algoritma Naïve Bayes. *Remik: Riset Dan E-jurnal Manajemen Informatika Komputer*, *6*(1), 41–47. https://doi.org/10.33395/remik.v6i1.11196