Analisis *Churn* Pelanggan Produk *Fashion Campus* Menggunakan Metode RFM *Analysis* dan Algoritma Naïve Bayes (Studi Kasus Yayasan Bakti Achmad Zaky)

Fauzan Kamil[1], Ratna Salkiawati[1,*], Allan D. Alexander[1,]

[1,3] *Informatika, Ilmu Komputer, Universitas Bhayangkara Jakarta Raya*
Jl. Raya Perjuangan, Margamulya, Bekasi, Indonesia

[1]fauzan.kamil19@mhs.ubharajaya.ac.id, [1*]ratna_tind@dsn.ubharajaya.ac.id, [1]allan@ubharajaya.ac.id

Intisari— **Pelanggan yang loyal memiliki dampak positif bagi perusahaan, baik melalui pembelian berulang maupun rekomendasi produk kepada orang lain. Namun, dalam dunia bisnis, terdapat berbagai faktor yang dapat mempengaruhi keputusan pelanggan untuk beralih ke pesaing, seperti harga, kualitas produk, dan pelayanan. Oleh karena itu, penting bagi perusahaan untuk memahami perilaku pelanggan dan mengambil keputusan strategis guna mempertahankan dan meningkatkan loyalitas mereka. Dalam penelitian ini, metode RFM (*Recency, Frequency, Monetary*) dan Naïve Bayes digunakan untuk menganalisis perilaku pelanggan dan memprediksi churn (berhenti berlangganan) pelanggan. Pendekatan CRISP-DM digunakan dalam langkah-langkah penyelesaiannya, mencakup tahapan pemahaman bisnis, pemahaman data, persiapan data, pemodelan, evaluasi, dan implementasi. Melalui analisis RFM, berhasil diidentifikasi 5 segmen pelanggan yang berbeda, yaitu *at-risk customers*, *best customers*, *lost customers*, *loyal customers*, dan *promising customers*. Setiap segmen memiliki karakteristik dan kecenderungan perilaku yang berbeda, memberikan wawasan berharga bagi perusahaan dalam memahami kebutuhan dan preferensi pelanggan. Hasil evaluasi Naïve Bayes menunjukkan bahwa model yang disimpan dalam format pickle memiliki performa yang setara dengan model yang telah diuji sebelumnya. Tingkat akurasi, *recall*, dan *f1-score* model tersebut sekitar 0.81 atau 81%, menunjukkan tingkat keakuratan yang baik dalam memprediksi churn pelanggan. Penggunaan model pickle memberikan keuntungan bagi perusahaan dalam hal efisiensi waktu dan biaya.**

Kata kunci— Pelanggan, Churn, RFM Analysis, Naïve Bayes, CRISP-DM

Abstract— ***Loyal customers have a positive impact on the company, either through repeated purchases or product recommendations to others. However, in the business world, there are various factors that can influence customers decisions to switch to competitors, such as price, product quality, and service. Therefore, it is important for companies to understand customer behavior and make strategic decisions in order to maintain and increase their loyalty. In this study, RFM (recency, frequency, and monetary) and Naïve Bayes methods were used to analyze customer behavior and predict customer churn. The CRISP-DM approach is used in the steps of its completion, covering the stages of business understanding, understanding data, data preparation, modeling, evaluation, and implementation. Through RFM analysis, five different customer segments were identified: at-risk customers, best customers, lost customers, loyal customers, and promising customers. Each segment has different characteristics and behavioral tendencies, providing valuable insights for companies to understand customer needs and preferences. The results of Naïve Bayes evaluation showed that models stored in pickle format had equivalent performance to models previously tested. The accuracy, recall, and f1-score of the models are about 0.81, or 81%, indicating a good level of accuracy in predicting customer churn. The use of the pickle model gives the company an advantage in terms of time and cost efficiency.***

Keywords— Customer, Churn, RFM Analysis, Naïve Bayes, CRISP-DM

I. PENDAHULUAN

Pelanggan adalah aset berharga bagi sebuah perusahaan, dan menjaga kepuasan dan kepercayaan mereka merupakan hal yang penting agar tetap loyal terhadap produk atau layanan yang ditawarkan [1]. Namun, dalam dunia bisnis, terdapat berbagai faktor yang dapat mempengaruhi keputusan pelanggan untuk beralih ke pesaing, seperti harga, kualitas produk, dan pelayanan. Oleh karena itu, penting bagi perusahaan untuk memahami perilaku pelanggan dan mengambil tindakan strategis guna mempertahankan dan meningkatkan loyalitas mereka. Salah satu cara yang dilakukan oleh perusahaan untuk mempertahankan pelanggan adalah dengan melakukan analisis churn. Churn pelanggan adalah fenomena ketika pelanggan perusahaan atau customer tidak lagi melakukan pembelian atau berinteraksi dengan perusahaan. Jika churn pelanggan tinggi, maka menunjukkan banyak pelanggan yang tidak tertarik untuk membeli barang atau jasa dari perusahaan tersebut [2].

Yayasan Bakti Achmad Zaky memiliki fokus pengembangan kewirausahaan dan mengoperasikan Fashion Campus, sebuah platform e-commerce yang menyediakan berbagai produk fashion untuk pemuda-pemudi Indonesia. Fashion Campus telah berhasil mempertahankan pelanggan

dengan kerja sama brand lokal yang diminati. Namun, pada tahun 2022, Fashion Campus mengalami penurunan jumlah pelanggan yang berhenti menggunakan layanan mereka, yang disebut dengan churn. Penurunan ini dipengaruhi oleh faktor-faktor seperti harga, kualitas produk, pelayanan, serta kehadiran kompetitor dengan variasi produk yang lebih menarik.

Untuk mempertahankan pelanggan dan mengatasi tantangan churn, Fashion Campus perlu melakukan pengolahan data yang tepat terkait dengan perilaku pelanggan dan mengidentifikasi faktor-faktor yang berpengaruh pada churn. Selain itu, dibutuhkan model prediksi menggunakan metode RFM dan algoritma klasifikasi Naïve Bayes untuk memperkirakan kemungkinan churn pelanggan berdasarkan data historis. Pendekatan ini akan membantu Fashion Campus mengambil tindakan strategis dalam meningkatkan loyalitas pelanggan.

II. BACKGROUND/LATAR BELAKANG

RFM *(Recency, Frequency, Monetary*) Analysis adalah sebuah teknik analisis data yang digunakan untuk mengelompokkan pelanggan berdasarkan tiga faktor utama, yaitu *Recency*, *Frequency*, dan *Monetary*. RFM *Analysis* bertujuan untuk memahami dan mengelompokkan pelanggan berdasarkan seberapa sering mereka membeli produk atau jasa, seberapa besar jumlah uang yang mereka habis kan, dan seberapa baru transaksi terakhir mereka dilakukan [3]. Terdapat beberapa kelas pelanggan berdasarkan nilai RFM. Kelas-kelas tersebut mencakup:

a. *Best Customers*: Pelanggan dengan skor RFM tertinggi yang sering melakukan pembelian dan menghasilkan pendapatan yang tinggi bagi perusahaan.
b. *Loyal Customers*: Pelanggan dengan skor RFM tinggi yang sering melakukan pembelian dan loyal terhadap produk atau layanan perusahaan.
c. *Promising Customers*: Pelanggan dengan skor RFM yang menunjukkan potensi untuk menjadi pelanggan loyal di masa depan.
d. *At-Risk Customers*: Pelanggan dengan skor RFM yang menunjukkan adanya risiko pelanggan tidak setia di masa depan.
e. *Lost Customers*: Pelanggan yang telah lama tidak ber transaksi dengan perusahaan dan kemungkinan besar tidak akan ber transaksi lagi [5].

Naïve Bayes merupakan salah satu metode machine learning yang umum digunakan dalam klasifikasi data. Metode ini didasarkan pada perhitungan probabilitas dari suatu data yang akan diklasifikasikan. Konsep dasar yang digunakan oleh Naïve Bayes adalah Teorema Bayes, yang menyatakan bahwa probabilitas suatu kejadian terjadi pada waktu yang akan datang, bergantung pada probabilitas kejadian tersebut terjadi di masa lalu dan kondisi saat ini. Naïve Bayes bekerja dengan menghitung probabilitas untuk setiap kemungkinan kelas, berdasarkan nilai-nilai fitur atau atribut dari data yang diberikan. Kemudian, kelas dengan probabilitas tertinggi akan dipilih sebagai hasil klasifikasi. Metode ini sering digunakan dalam pengenalan pola, klasifikasi dokumen, spam filtering, dan sebagainya [4].

Churn, atau sering disebut sebagai tingkat pergantian pelanggan, merupakan indikator penting dalam manajemen bisnis modern. Churn merujuk pada proses ketika pelanggan atau konsumen berhenti menggunakan produk atau layanan yang ditawarkan oleh suatu perusahaan. Kondisi ini dapat berdampak signifikan pada kesehatan bisnis, karena mengakibatkan penurunan pendapatan, keuntungan, dan stabilitas.

Sejumlah faktor telah diidentifikasi sebagai penggerak churn dalam berbagai industri. Pertama, faktor kualitas layanan berperan kunci dalam mempengaruhi keputusan pelanggan untuk tetap berlangganan atau beralih ke pesaing. Studi oleh Keaveney (1995) menekankan pentingnya kepuasan pelanggan dalam meminimalkan churn [6].

Selain itu, faktor harga juga memiliki peran yang signifikan. Penelitian oleh Reinartz dan Kumar (2002) menunjukkan bahwa peningkatan harga dapat mengakibatkan peningkatan churn, sementara strategi penetapan harga yang bijak dapat membantu mempertahankan pelanggan [7].

Selain faktor-faktor tersebut, faktor psikologis seperti loyalitas merek dan preferensi pelanggan juga berpengaruh besar. Melalui penelitian Chaudhuri dan Holbrook (2001), ditemukan bahwa pelanggan yang memiliki loyalitas merek yang kuat lebih cenderung tetap berlangganan [8] .

Analisis churn dapat dilakukan dengan berbagai metode. Metode yang umum digunakan termasuk regresi logistik, analisis survival, dan algoritma pembelajaran mesin seperti Naïve Bayes dan Random Forest. Penelitian oleh Verbeke et al. (2012) memberikan contoh aplikasi metode analisis churn menggunakan algoritma pembelajaran mesin dalam industri telekomunikasi [9] .

III. METODOLOGI PENELITIAN

A. Tahapan Penelitian

Tahapan metodologi penelitian akan diuraikan secara umum sebagai berikut:

1. Identifikasi Masalah: Melakukan pengumpulan data, peneliti berhasil mengidentifikasi masalah yang ada di Fashion Campus.
2. Rumusan Masalah: Menentukan identifikasi masalah akan membantu mendapatkan informasi yang diperlukan dalam perumusan masalah.
3. Metode Pengumpulan Data: Tahapan yang digunakan dalam mengumpulkan data atau informasi yang diperlukan dalam penelitian, yaitu.
    a. Metode Observasi: Mengamati dan mencari informasi pada dataset yang telah disediakan oleh Yayasan Bakti Achmad Zaky.
    b. Studi Pustaka: Pengumpulan data dengan cara mengumpulkan informasi dari berbagai dokumen yang terkait dengan topik penelitian seperti jurnal, buku, dan sebagainya.
4. Metode Pendekatan: Penelitian ini dilakukan dengan menggunakan metode pendekatan Cross-Industry Standard Process for Data Mining (CRISP-DM)

5. Penerapan: Tahap ini melibatkan proses penelitian dan pelaksanaan dari tahap-tahap persiapan yang telah dilakukan sebelumnya, seperti yang telah dijelaskan sebelumnya, yaitu sebagai berikut:
   a. *Business Understanding*
   b. Data *Understanding*
   c. Data *Preparation*
   d. *Modeling*
   e. *Evaluation*
   f. *Deployment*

B. Metode Analisis

Dalam penelitian ini, digunakan dua metode analisis, yaitu RFM Analysis dan Naïve Bayes, untuk menganalisis perilaku pelanggan dan memprediksi churn. RFM Analysis digunakan untuk segmentasi pelanggan berdasarkan tiga faktor, yaitu recency (waktu terakhir pelanggan melakukan transaksi), frequency (frekuensi transaksi), dan monetary (jumlah uang yang dibelanjakan). Sementara itu, Naïve Bayes digunakan untuk memprediksi pelanggan yang kemungkinan akan mengalami churn dalam periode 6 bulan, berdasarkan selisih antara tanggal terakhir dalam dataset dan tanggal pembelian terakhir.

Pemilihan periode waktu 6 bulan dalam analisis churn didasarkan pada kebiasaan industri yang umum, karena periode tersebut dapat memberikan gambaran yang cukup baik tentang perilaku pelanggan dalam jangka waktu yang lebih panjang [10]. Dalam waktu 6 bulan, pelanggan yang tidak melakukan transaksi kemungkinan besar sudah kehilangan minat terhadap produk atau layanan perusahaan. Oleh karena itu, periode waktu tersebut dapat membantu dalam memprediksi apakah pelanggan akan mengalami churn atau tidak.

Dalam pembangunan model prediksi menggunakan algoritma Naïve Bayes, RFM Analysis digunakan sebagai salah satu fitur. Penggunaan RFM Analysis sebagai fitur memungkinkan variabel prediktor dalam model Naïve Bayes untuk memprediksi apakah pelanggan akan mengalami churn atau tidak.

IV. Hasil dan Pembahasan

Berikut adalah pengolahan data menggunakan bahasa pemrograman Python yang mengikuti alur pada CRISP-DM.

A. Business Understanding

Fashion Campus sedang mengembangkan bisnis penjualan pakaian bekas yang masih layak pakai. Meskipun jumlah pengguna semakin meningkat, sebagian besar di antaranya tidak organic, yang diperoleh hanya untuk memenuhi target jumlah pengguna. Sebagai hasilnya, banyak pengguna tidak kembali ke platform untuk melakukan transaksi atau pembelian. Oleh karena itu, perlu dicari cara untuk meningkatkan kualitas pengguna dan mendorong pertumbuhan organik bisnis Fashion Campus.

B. Data Understanding

Dalam tahap ini akan dilakukan pemahaman mengenai karakteristik data, termasuk jenis datanya, kualitasnya, ukuran, dan distribusi nilainya. Berikut merupakan data understanding dari masing-masing dataset.

```
df_customer = pd.read_csv('E:\Skripsi\Fauzan Kamil\data\customer.csv')
df_product = pd.read_csv('E:\Skripsi\Fauzan Kamil\data\product.csv', on_bad_lines='skip')
df_trans = pd.read_csv('E:\Skripsi\Fauzan Kamil\data\\transactions.csv')
```

Gambar 1. Data Understanding

Gambar di atas menjelaskan membaca file dengan library pandas, pada file “customer.csv” dibaca dan disimpan ke dalam data frame df_customer. Kemudian, file “product.csv” dibaca dan disimpan ke dalam data frame df_product. Pada pembacaan file “product.csv”, parameter “on_bad_lines” diatur sebagai “skip”, yang artinya jika terdapat baris yang tidak valid, maka baris tersebut akan dilanjutkan. Terakhir, file “transactions.csv” dibaca dan disimpan ke dalam data frame df_trans.

1. Data Customer: Kumpulan informasi tentang pelanggan Fashion Campus yang telah mendaftar, dengan jumlah baris sebanyak 100.000 dan terdiri dari 15 kolom, berikut merupakan dataset customer.
2. Data Produk: Kumpulan informasi tentang produk yang ditawarkan oleh Fashion Campus, dengan jumlah baris sebanyak 44.424 dan terdiri dari 10 kolom, berikut merupakan dataset produk.
3. Data Transaksi: Kumpulan informasi tentang transaksi yang berhasil dan transaksi yang gagal pada Fashion Campus, dengan jumlah baris sebanyak 856.584 dan terdiri dari 14 kolom, berikut merupakan dataset transaksi

Kolom yang memiliki nilai kosong atau missing value serta kolom yang tipe datanya tidak sesuai akan diatasi pada tahap data preparation.

C. Data Preparation

Data Preparation melibatkan serangkaian langkah atau tahapan yang dilakukan oleh seorang peneliti untuk mengubah data mentah menjadi data yang berkualitas. Tahapan-tahapan yang dilakukan dalam proses ini meliputi beberapa langkah sebagai berikut;

a. Membuat RFM

RFM adalah sebuah metode analisis yang digunakan untuk mengategorikan dan mengelompokkan pelanggan berdasarkan tiga faktor utama: Recency, Frequency, dan Monetary. Konsep RFM digunakan untuk membantu bisnis dalam segmentasi pelanggan, penentuan strategi pemasaran, dan pengambilan keputusan bisnis lainnya.

1. Recency

	customer_id	recent_days
0	3	97
1	7	98
2	8	224
3	9	212
4	11	293

Gambar 2. Recency

Penghitungan *recency*, yaitu menghitung jumlah hari terakhir kali pelanggan melakukan pembelian dari

hari terakhir transaksi di dataset. Pertama-tama, code mengelompokkan dataset berdasarkan customer_id dan mengambil tanggal transaksi terbaru untuk setiap customer. Kemudian, code menghitung selisih antara tanggal transaksi terbaru dan tanggal transaksi terakhir untuk setiap customer, dan mengonversinya menjadi jumlah hari.

2. Frequency

customer_id	frequency
3	52
7	1
8	7
9	6
11	1

Gambar 3. Frequency

DataFrame “frequency” berisi informasi tentang berapa kali setiap pelanggan melakukan transaksi pada periode tertentu. Dengan demikian, dapat dilihat frekuensi atau jumlah transaksi yang dilakukan oleh setiap pelanggan berdasarkan “customer_id”. Data tersebut menunjukkan contoh dari DataFrame “frequency” yang terdiri dari kolom “customer_id” dan “frequency”. Misalnya, pelanggan dengan “customer_id” 3 memiliki frekuensi transaksi sebanyak 52 dan pelanggan dengan “customer_id” 7 hanya memiliki satu transaksi.

3. Monetary

customer_id	monetary
3	23122776
7	966126
8	3898561
9	2638665
11	197533

Gambar 4. Monetary

Dilakukan agregasi dengan menjumlahkan total_amount (jumlah uang yang dibayarkan oleh pelanggan) pada setiap customer_id menggunakan fungsi sum(). Hasil akhirnya ditampung dalam dataframe baru dengan nama monetary. Data tersebut menunjukkan contoh dari DataFrame monetary yang terdiri dari kolom “customer_id” dan “monetary”. Misalnya, pelanggan dengan “customer_id” 3 memiliki total_amount sebesar 23.122.776.

4. RFM Score

R_rank_norm	F_rank_norm	M_rank_norm	RFM_Score
95.07	92.24	92.24	4.63
95.02	11.15	11.15	1.19
54.08	55.59	55.59	2.77
57.64	52.11	52.11	2.65
40.53	11.15	11.15	0.78

Gambar 5. Normalisasi RFM dan RFM Score

Dilakukan peringkat atau ranking terhadap pelanggan berdasarkan nilai recency, frequency, dan monetary. Peringkat ini dinormalisasi menjadi skala 0-100 untuk menghindari bias antara variabel. Selanjutnya, dihitung RFM Score dengan bobot yang telah ditentukan untuk masing-masing variabel. RFM Score adalah indikator komposit yang mencerminkan nilai keseluruhan dari recency, frequency, dan monetary.

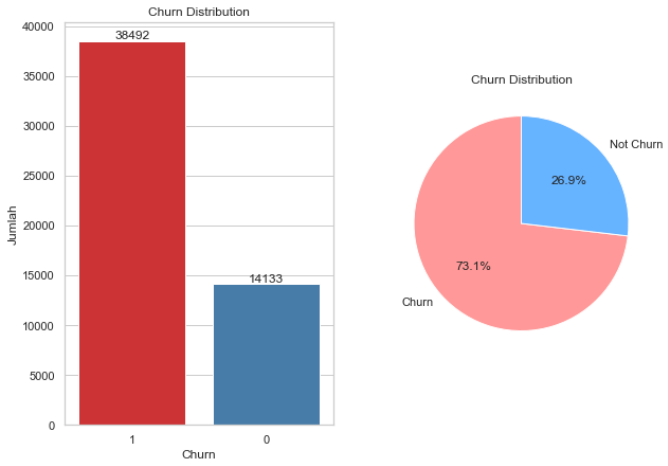
5. Segmentasi Customer

RFM_Score	Customer_segment
4.63	Best Customers
1.19	Lost Customers
2.77	At-Risk Customers
2.65	At-Risk Customers
0.78	Lost Customers

Gambar 6. Segmentasi Customer

Segmen pelanggan ditentukan berdasarkan rentang nilai RFM Score yang telah ditentukan, seperti “Best Customers”, “Loyal Customers”, “Promising Customers”, “At-Risk Customers”, dan “Lost Customers”.

b. Status Churn Customer

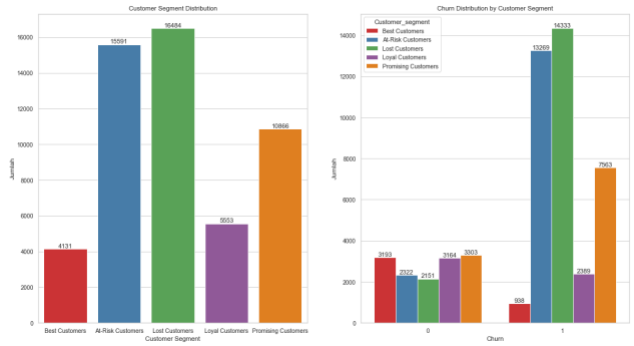


Gambar 7. Status Customer Churn

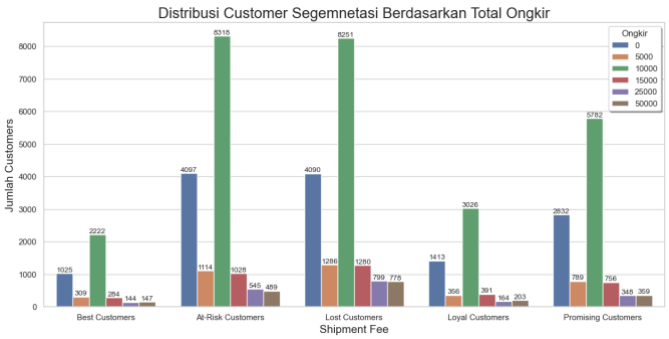
Dari total 52.625 pelanggan yang aktif melakukan transaksi, terdapat dua kategori utama, yaitu Churn dan Not Churn. Dalam kategori Churn, terdapat sebanyak 73,1% atau 38.495 pelanggan. Hal ini mengindikasikan bahwa sebagian besar pelanggan Fashion Campus mengalami Churn, yaitu ketika mereka tidak melakukan transaksi dalam periode waktu 6 bulan terakhir. Sedangkan dalam kategori Not Churn, terdapat sebanyak 26,9% atau 14.133 pelanggan yang tetap aktif bertransaksi.

c. Exploratory Data Analysis (EDA)

Hasil visualisasi menunjukkan lima segmen pelanggan menarik pada Fashion Campus: “At-Risk Customers” (2.322 pelanggan yang tidak mengalami Churn dan 13.269 pelanggan yang mengalami Churn), “Best Customers” (3.193 pelanggan tetap aktif dan 938 pelanggan yang mengalami Churn), “Lost Customers” (2.151 pelanggan yang tidak mengalami Churn dan 14.333 pelanggan yang mengalami Churn), “Loyal Customers” (3.164 pelanggan tetap aktif dan 2.389 pelanggan yang mengalami Churn), dan “Promising Customers” (3.303 pelanggan yang tidak mengalami Churn dan 7.563 pelanggan yang mengalami Churn). Visualisasi tersebut memberikan informasi tentang perilaku dan kategori pelanggan dalam hal Churn, dengan segmen “At-Risk Customers” dan “Lost Customers” menunjukkan risiko tinggi terhadap Churn, sementara segmen “Best Customers” dan “Loyal Customers” memiliki pelanggan yang tetap aktif dan setia. Segmen “Promising Customers” menunjukkan potensi pertumbuhan yang baik, tetapi membutuhkan perhatian untuk meminimalkan Churn dan mempertahankan pertumbuhan positif.



Gambar 8. Jumlah Segmentasi Customer



Gambar 9. Customer Segemnetasi Berdasarkan Total Ongkir

Berdasarkan grafik di atas, dapat dilihat bahwa shipment fee yang paling banyak digunakan oleh setiap segmentasi pelanggan adalah dengan harga 10.000, kecuali pada segmentasi Best Customers dan Loyal Customers, dimana harga yang paling banyak digunakan adalah 0. Hal ini menunjukkan bahwa setiap segmentasi pelanggan memiliki preferensi harga shipment fee yang berbeda-beda. Selain itu, dapat dilihat bahwa jumlah pengguna shipment fee dengan harga yang lebih tinggi seperti 25.000 dan 50.000 relatif sedikit dibandingkan dengan harga yang lebih rendah.

d. Feature Selection

	feature	Churn
0	Churn	1.000000
1	recent_days	0.767668
2	freq_diff	0.513053
3	mean_pembelian	0.313238
4	selisih_tanggal	0.044225
5	tenure	0.044225
6	promo_amount	0.016281
7	hari	0.015547
8	home_location_lat	0.012874
9	year	0.003867
10	home_location_long	-0.002080
11	customer_id	-0.003274
12	shipment_location_long	-0.025907
13	shipment_location_lat	-0.034450
14	shipment_fee	-0.039077
15	id	-0.039317
16	price	-0.046691
17	total_amount	-0.067130
18	qty	-0.119525
19	total_promo	-0.257022
20	order_count	-0.322379
21	F_rank_norm	-0.322816
22	M_rank_norm	-0.322816
23	frequency	-0.322816
24	type_buy_tahun	-0.327205
25	purchase_year	-0.327205
26	ongkir	-0.348671
27	Customer_segment	-0.386327
28	monetary	-0.395831
29	RFM_Score	-0.403212
30	monthly_spend	-0.567645
31	R_rank_norm	-0.767668

Gambar 10. Hasil Korelasi

Berdasarkan hasil korelasi di bawah, terdapat beberapa fitur yang dapat digunakan sebagai prediktor, yaitu recent_days, freq_diff, frequency, monetary, purchase_year, ongkir, Customer_segment, RFM_Score, monthly_spend, dan mean_pembelian. Hal ini

dikarenakan fitur-fitur tersebut memiliki korelasi yang tinggi dengan variabel Churn, dengan nilai korelasi yang dimulai dari 0.32. Oleh karena itu, fitur-fitur ini sangat penting untuk dipertimbangkan dalam membangun model untuk memprediksi Churn Selection.

e. Imbalance

Imbalance adalah kondisi di mana jumlah data pada setiap kelas atau label tidak seimbang atau tidak proporsional [11]. Hal ini dapat menyebabkan masalah dalam proses training model karena model cenderung lebih condong untuk memprediksi label dengan jumlah data yang lebih banyak. Atribut Churn terlihat tidak seimbang atau imbalance, karena jumlah data dengan Churn = 1 (38492) jauh lebih banyak daripada jumlah data dengan Not Churn = 0 (14133). Hal ini dapat menyebabkan masalah ketika melakukan pemodelan atau analisis data, karena model atau algoritma yang digunakan cenderung cenderung memiliki bias terhadap kelas mayoritas (dalam hal ini, Churn=1).

```
1 X, y = SMOTE().fit_resample(X, y)
2 print(sorted(Counter(y).items()))
```

✓ 0.5s

```
[(0, 38492), (1, 38492)]
```

Gambar 11. *Handling Imbalance* Data

SMOTE (Synthetic Minority Over-sampling Technique) digunakan untuk menangani masalah ketidakseimbangan kelas (imbalanced class) pada data, di mana kelas minoritas memiliki jumlah yang jauh lebih sedikit daripada kelas mayoritas. Pada code tersebut, variabel X dan y digunakan untuk menyimpan feature dan target pada dataset yang ingin diaplikasikan SMOTE. Kemudian, SMOTE().fit_resample(X, y) digunakan untuk melakukan proses oversampling pada dataset, di mana kelas minoritas akan disintesis kembali untuk membuat sampel baru hingga jumlah sampel dari kedua kelas menjadi seimbang.

D. Modeling

a. Splitting data

Splitting data adalah proses untuk membagi dataset menjadi 3 bagian yaitu train set, validation set, dan test set [12]. Data training memiliki 49269 data, sedangkan data test memiliki 15397 data dan data validation memiliki 12318 data. Selain itu, data predictor atau variabel independen (X) memiliki 10 fitur atau kolom, sementara data target atau variabel dependen (y) hanya memiliki satu kolom. Pembagian data menjadi tiga subset tersebut dilakukan untuk memastikan bahwa model yang dibuat tidak hanya mampu memprediksi dengan baik pada data training, tetapi juga pada data yang belum pernah dilihat sebelumnya, yaitu data test dan validation.

b. Model Naïve Bayes

Dalam tahap ini, algoritma yang digunakan adalah Naïve Bayes. Model Naïve Bayes akan mempelajari kemungkinan masing-masing fitur pada data dan kemudian melakukan prediksi terhadap hasil kelas berdasarkan fitur-fitur tersebut.

```
1 model_nb = GaussianNB()
2 model_nb.fit(X_train, y_train)
```

✓ 0.4s

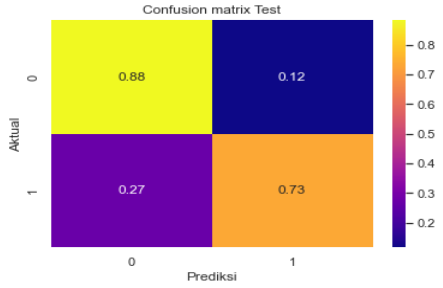
```
GaussianNB()
```

Gambar 12. Training Model Naïve Bayes

Membuat model Naive Bayes dan melakukan training model menggunakan data training yang telah diproses sebelumnya. GaussianNB() merupakan fungsi yang digunakan untuk membuat model Naive Bayes dengan Gaussian Distribution, yaitu distribusi probabilitas dimana data berada di sekitar nilai rata-rata (mean).

E. Evaluation

Evaluation atau evaluasi adalah proses untuk mengukur sejauh mana model yang telah dibuat dapat bekerja dengan baik dalam memprediksi atau mengklasifikasikan data yang belum pernah dilihat sebelumnya [13]. Evaluasi dapat dilakukan dengan berbagai metode, seperti confusion matrix, accuracy score, precision, recall, dan F1 score.



Gambar 13. Confusion Matrix

	precision	recall	f1-score	support
0	0.77	0.88	0.82	7692
1	0.86	0.73	0.79	7705
accuracy			0.81	15397
macro avg	0.82	0.81	0.81	15397
weighted avg	0.82	0.81	0.81	15397

Gambar 14. Classification Report

Pada hasil classification report di atas, terdapat dua kelas yang dievaluasi yaitu kelas 0 dan kelas 1. Precision untuk kelas 0 adalah 0.77, menunjukkan bahwa 77% prediksi positif untuk kelas 0 adalah benar. Recall untuk kelas 0 adalah 0.88, menandakan bahwa model dapat mengidentifikasi 88% dari semua nilai positif untuk kelas 0. F1-score untuk kelas 0 adalah 0.82, yang merupakan rata-rata harmonik antara precision dan recall. Jumlah sampel pada kelas 0 adalah 7692, sedangkan pada kelas 1 adalah 7705. Accuracy pada data test adalah 0.81, menunjukkan bahwa model dapat memprediksi dengan tepat 81% dari semua data.

F. Deployment

Tahap deployment model melibatkan pengujian dan validasi kembali model yang telah dilatih dan diuji untuk memastikan performanya tetap konsisten dan dapat diandalkan di lingkungan produksi.

```
# Prediksi
pred = model_pickle.predict(X_test)
pred[:5]
```

✓ 0.1s

```
array([0, 1, 1, 1, 0], dtype=int64)
```

Gambar 15. Prediksi model pickle

	precision	recall	f1-score	support
0	0.77	0.89	0.83	7692
1	0.87	0.73	0.80	7705
accuracy			0.81	15397
macro avg	0.82	0.81	0.81	15397
weighted avg	0.82	0.81	0.81	15397

Gambar 16. Hasil classification report dari model pickle

Classification report di atas memberikan informasi tentang performa model pickle dalam melakukan klasifikasi pada dua kelas yang ada, yaitu kelas 0 dan kelas 1. Dalam model pickle, model memiliki precision sebesar 0.77 untuk kelas 0 dan 0.87 untuk kelas 1, serta recall sebesar 0.89 untuk kelas 0 dan 0.73 untuk kelas 1. F1-score untuk kelas 0 adalah 0.83 dan untuk kelas 1 adalah 0.80. Akurasi model mencapai 0.81, yang merupakan rata-rata dari keakuratan prediksi pada kedua kelas. Dalam keseluruhan, model memiliki performa yang baik dalam mengklasifikasikan data dengan rata-rata f1-score sebesar 0.81

V. KESIMPULAN

Berdasarkan penelitian yang dilakukan, ditemukan beberapa temuan penting terkait dengan churn. Pertama, terdapat 5 segmen pelanggan yang dapat dikelompokkan berdasarkan recency, frequency, dan monetary. Terdapat segmen "At-Risk Customers" (15.591 pelanggan), "Best Customers" (4.131 pelanggan), "Lost Customers" (16.484 pelanggan), "Loyal Customers" (5.553 pelanggan), dan "Promising Customers" (10.866 pelanggan). Kedua, faktor-faktor seperti recent_days, freq_diff, frequency, monetary, purchase_year, ongkir, Customer_segment, RFM_Score, monthly_spend, dan mean_pembelian memiliki pengaruh terhadap churn pelanggan. Faktor-faktor ini dapat memengaruhi keputusan pelanggan untuk churn. Terakhir, model pickle dapat digunakan secara efektif untuk memprediksi churn pelanggan di masa depan dengan tingkat akurasi, recall, dan f1-score sekitar 0.81 atau 81%. Penggunaan model pickle membantu menghemat waktu dan biaya dalam pelatihan ulang model.

REFERENSI

[1] G. P. Hafidz and R. U. Muslimah, "Pengaruh Kualitas Layanan, Citra Merek, Kepercayaan Pelanggan Dan Kepuasan Pelanggan Terhadap Loyalitas Pelanggan Produk Herbalife," vol. 1, 2023.

[2] L. I. Muhammad, S. Agus, and G. Windu, "Prediksi Tingkat Pelanggan Churnpada Perusahaan Telekomunikasi Dengan Algoritma Adaboost".

[3] Firmansyah and Y. Agus, "Prediksi Customer Churn Pada Bisnis Retail Menggunakan Algoritma Naïve Bayes," Riset dan E-Jurnal Manajemen Informatika Komputer, vol. 6, no. 1, 2021, doi: 10.33395/remik.v4i1.11196.

[4] P. Singal and D. R. S. Chhillar, "International Journal of Computer Science and Mobile Computing A Review on GPS and its Applications in Computer Science," 2014. [Online]. Available: www.ijcsmc.com

[5] K. Tsiptsis Antonios Chorianopoulos WILEY, "Data Mining Techniques in CRM: Inside Customer Segmentation," 2018.

[6] S. M. Keaveney, "Customer Switching Behavior in Service Industries: An Exploratory Study," J Mark, vol. 59, no. 2, pp. 71–82, Apr. 1995, doi: 10.1177/002224299505900206.

[7] W. Reinartz and V. Kumar, "The Mismanagement of Customer Loyalty," Harv Bus Rev, vol. 80, pp. 86–94, 125, Aug. 2002.

[8] A. Chaudhuri and M. Holbrook, "The Chain of Effects From Brand Trust and Brand Affect to Brand Performance: The Role of Brand Loyalty," J Mark, vol. 65, pp. 81–93, Apr. 2001, doi: 10.1509/jmkg.65.2.81.18255.

[9] W. Verbeke, K. Dejaeger, D. Martens, J. Hur, and B. Baesens, "New insights into churn prediction in the telecommunication sector: A profit driven data mining approach," Eur J Oper Res, vol. 218, no. 1, pp. 211–229, Apr. 2012, doi: 10.1016/j.ejor.2011.09.031.

[10] N. Lu, H. Lin, J. Lu, and G. Zhang, "A customer churn prediction model in telecom industry using boosting," IEEE Trans Industr Inform, vol. 10, no. 2, pp. 1659–1665, 2014, doi: 10.1109/TII.2012.2224355.

[11] F. D. Astuti and F. N. Lenti, "Implementasi SMOTE Untuk Mengatasi Imbalance Class Pada Klasifikasi Car Evolution Menggunakan K-NN," Jurnal JUPITER, vol. 13, no. 1, pp. 89–98, 2021.

[12] J. Wira and G. Putra, "Pengenalan Konsep Pembelajaran Mesin dan Deep Learning Edisi 1.4 (17 Agustus 2020)," 2020.

[13] M. A. Wiratama and W. M. Pradnya, "Optimasi Algoritma Data Mining Menggunakan Backward Elimination untuk Klasifikasi Penyakit Diabetes," Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI), vol. 11, no. 1, p. 1, Apr. 2022, doi: 10.23887/janapati.v11i1.45282.