Deteksi Anomali Terhadap Pembatalan Transaksi Pada *Platform* Tiktok Shop dengan Algoritma *Categorical Boosting* (Catboost)

Novliyan Dimas Syandika[1], Wiyli Yustanti [2]

[1,2] Sistem Informasi, Teknik Informatika, Universitas Negeri Surabaya
[1]novliyan.19036@mhs.unesa.ac.id
[2]wiyliyustanti@unesa.ac.id

***Abstrak*— Pada era saat ini, jual beli secara *online* semakin populer salah satunya adalah TikTok Shop. Meskipun TikTok Shop menawarkan fitur berbelanja yang menarik bagi konsumen, masalah pembatalan pemesanan tetap menjadi tantangan bagi penjual dalam mengoptimalkan penjualan dan keuntungan. Dalam penelitian ini, digunakan algoritma *Categorical Boosting* (CatBoost) yang telah terbukti efektif dalam memprediksi pembatalan pemesanan. Namun masih terbatas dan belum banyak digunakan dalam *online shop* seperti TikTok Shop. Dalam menangani permasalah imbalaced data digunakan teknik *resampling oversampling* dengan menggunakan metode SMOTE (*Synthetic Minority Oversampling Technique*). Sedangkan untuk mengidentifikasi faktor-faktor yang signifikan dalam pembatalan transaksi digunakan teknik *Principal Component Analysis* (PCA). Data yang digunakan dalam penelitian ini yaitu data riwayat pembelian pelanggan pada *platform* TikTok Shop. Metode yang digunakan pada penelitian ini yaitu CRISP-DM (*Cross Industry Process Model for Data Mining)* yang mencakup *business understanding, data understanding, data preparation, modeling, evaluation,* dan *deployment*. Dalam mengevaluasi model digunakan *Stratified 10-fold Cross-Validation* untuk mengukur kualitas dan efektivitas model prediksi pada variabel target Cancelation dengan hasil didasarkan pada nilai akurasi, recall, presisi, dan F1. Selain itu *Confusion matrix* digunakan juga sebagai alat evaluasi tambahan untuk mengevaluasi performa model yang telah dibentuk. Hasil penelitian menunjukkan bahwa algoritma CatBoost memiliki akurasi yang sangat tinggi (99.7%) dalam mengklasifikasikan transaksi sebagai pembatalan atau bukan pembatalan, dan memiliki presisi, recall, dan f1-score yang sempurna (1.00) untuk kedua kelas. Faktor-faktor yang paling berpengaruh terhadap pembatalan transaksi adalah Payment Method, Regency and City, Order Refund Amount, Variation, Province, Product Category, Shipping Fee After Discount, dan SKU Platform Discount.**

***Kata Kunci*— Catboost, CRISP-DM, Pembatalan, PCA, *Imbalaced*.**

I. PENDAHULUAN

Pada era digital seperti saat ini, bisnis online menjadi pilihan bagi banyak orang untuk membeli dan menjual produk secara *online*. *Online Shop* merupakan tempat untuk memasarkan produk melalui *platform* secara *digital* seperti media sosial ataupun *digital marketplace* dimana pengguna dapat membeli dan juga menjual barang secara *online* [1]. Seperti halnya TikTok, tidak hanya sebatas hiburan video saja, TikTok saat ini sudah memiliki fitur yang bernama TikTok Shop. Tiktok Shop adalah bagian dari sebuah aplikasi yang saat ini menggabungkan marketplace dan media sosial. Di dalamnya, terdapat beberapa faktor yang menjadi pertimbangan konsumen dalam membeli produk melalui TikTok Shop, antara lain kredibilitas, sistem promosi, multifungsi, dan kemudahan. [2]. Menghubungkan pembeli dan penjual di lingkungan yang aman dan nyaman adalah salah satu tantangan terbesar di *marketplace* [3].

Seperti bisnis online pada umumnya, TikTok Shop juga menghadapi masalah pembatalan pemesanan, yang dapat berdampak negatif pada penjual dan pembeli. Oleh karena itu, penting bagi penjual untuk mendeteksi anomali agar dapat mengatasi kemungkinan pembatalan pemesanan secara akurat agar dapat mengambil tindakan yang tepat dan meminimalkan kerugian menggunakan metode *Data Mining*. *Data Mining* merupakan sebuah proses komputasi yang menganalisis kumpulan data, biasanya data yang berukuran besar dengan menggunakan metode statistik secara logis, untuk mendapatkan pengetahuan yang tersembunyi, pola yang masih belum diketahui, dan juga untuk dapat menginformasikan agar dapat dilakukan pengembalian keputusan [4]. *Data Mining* memiliki beberapa peran utama diantaranya adalah Estimasi, *Forecasting*, Klasifikasi, Klastering, dan asosiasi [5]. Dalam melakukan data mining, perlu menggunakan cara atau pemilihan algoritma yang tepat sehingga mendapatkan hasil yang berguna. algoritma *Categorical Boosting* (Catboost) adalah salah satu algoritma *machine learning* yang dapat digunakan untuk mendeteksi anomali terhadap pembatalan transaksi [6]. CatBoost merupakan implementasi *Gradient Boosted Decision Tree* (GBDT) *open source* untuk *Supervised machine learning* yang membawa dua inovasi: *Ordered Target Statistics* dan *Ordered Boosting* [7]. Algoritma ini memanfaatkan teknik boosting untuk meningkatkan akurasi prediksi dengan menggabungkan banyak model yang lemah menjadi satu model yang kuat. Meskipun *Categorical Boosting* (Catboost) terbukti efektif dalam memprediksi pembatalan pemesanan, implementasinya masih terbatas dan belum banyak digunakan dalam *Online Shop platform* seperti TikTok Shop.

Pada penelitian yang akan dilakukan memiliki data yang tidak seimbang pada tiap kelasnya, sehingga menjadi sebuah tantangan terhadap penelitian ini. Data tidak seimbang merupakan keadaan data yang tidak seimbang antar kelas data yang satu dengan kelas data lainnya [8]. Masalah dalam klasifikasi terjadi ketika data tidak seimbang karena

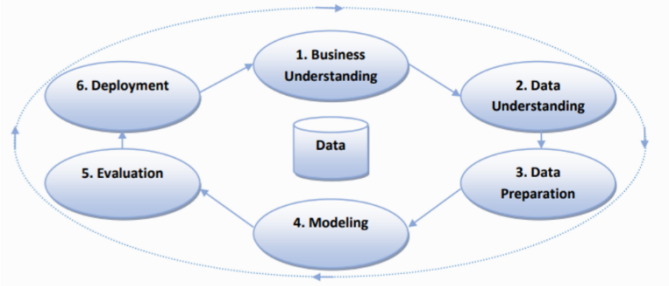
pengklasifikasi lebih cenderung memprediksi kelas data mayoritas yang jumlahnya lebih banyak dibandingkan dengan kelas minoritas yang jumlahnya lebih sedikit. Oleh karena itu, ketidakseimbangan data harus diperhatikan dan diatasi agar pengolahan data dan klasifikasi dapat dilakukan secara tepat. Sehingga diperlukan penelitian untuk mengimplementasikan algoritma *Categorical Boosting* (Catboost) dalam mendeteksi anomali terhadap pembatalan pemesanan pada *platform* TikTok Shop.

Tujuan dari penelitian ini adalah untuk mendeteksi anomali terhadap pembatalan transaksi pada *platform* tiktok dengan algoritma *categorical boosting* (CatBoost). Selain itu, penelitian ini juga bertujuan untuk mengetahui faktor-faktor apa saja yang mempengaruhi pembatalan transaksi pada platform TikTok. Dengan demikian, *Online Shop* dapat mengambil tindakan yang tepat untuk mengurangi tingkat pembatalan pemesanan.

Penelitian ini akan menggunakan *dataset* yang diperoleh dari riwayat pembelian pelanggan pada *platform* TikTok Shop yang diperoleh dari perusahaan PT X. Perusahaan ini merupakan salah satu toko online yang menyediakan berbagai produk *fashion*. *Dataset* ini berisi informasi tentang produk yang dijual, pelanggan yang melakukan pembelian, dan informasi pembatalan pemesanan. Informasi-informasi ini kemudian digunakan untuk melakukan prediksi terhadap kemungkinan pembatalan pemesanan pada toko online tersebut.

II. METODE

Pada penelitian ini menggunakan metode CRISP-DM (Cross-Industry Standard Process for Data Mining) yang bersifat konseptual. Metode ini memiliki enam siklus : Business Understanding, Data Understanding, Data Preparation, Modeling , Evaluation, dan Deployment [4]. Berikut siklus tahapannya pada Gbr. 1:



Gbr. 1 CRISP-DM (Cross Industry Standard Process for Data Mining)

A. Business Understanding

Langkah pertama dalam CRISP-DM adalah *Business Understanding* yang berfokus pada pemahaman tujuan dan kebutuhan sisi bisnis yang kemudian diubah menjadi pengetahuan dalam menentukan definisi masalah utama yang dapat diselesaikan dengan *Data Mining*. Pada tahap ini, dilakukan pemahaman terhadap bisnis yang sedang dijalankan, yaitu *Online Shop* pada *platform* Tiktok Shop, dan permasalahan yang ingin diselesaikan, yaitu deteksi anomali terhadap pembatalan transaksi.

B. Data Understanding

Pada tahap ini dilakukan proses pengenalan terhadap data untuk mengetahui kualitas data, mendapatkan insight awal dari data. Tahap ini merupakan tahap untuk memahami data yang bermasalah untuk menghasilkan model *machine learning* yang lebih baik. Data yang digunakan pada penelitian ini yaitu dataset yang diperoleh dari riwayat pembelian pelanggan pada *platform* TikTok Shop yang diperoleh dari perusahaan PT X. *Dataset* tersebut merupakan data pelanggan pada platfrom TikTok Shop pada tahun 2021,2022, dan 2023 yang terdiri dari 48 atribut dan 8.739 baris . Berikut ini deskripsi setiap kolom dalam dataset pada Tabel I :

TABEL I
DESKRIPSI DATASET SETIAP ATRIBUT

No.	Nama Atribut	Keterangan
1	Order ID	Identitas pada setiap Pemesanan yang dilakukan
2	Order Status	Status proses dari Pemesanan
3	Order Substatus	Substatus dari proses Pemesanan saat ini
4	Cancelation	Penjelasan (Filed to explain) pembatalan pesanan
5	Normal or Pre-order	Produk normal atau pre - order
6	SKU ID	SKU (Stock Keeping Unit) adalah kode unik item barang yang dibuat retailer
7	Seller SKU	SKU penjual yang diinput oleh penjual di sistem produk
8	Product Name	Nama Produk di Platform
9	Variation	Variasi SKUPlatform (Kode unik variasi produk), such as spesific size or color
10	Quantity	Kuantitas SKU terjual dalam pesanan
11	Sku Quantity of return	Kuantitas SKU yg dikembalikan dalam pesanan
12	SKU Unit Original Price	SKU Satuan Harga Asli
13	SKU Subtotal Before Discount	SKU Unit Original Price * Sku Quantity
14	SKU Platform Discount	Total diskon platform dalam ID SKU ini.
15	SKU Seller Discount	Total diskon penjual dalam ID SKU ini.
16	SKU Subtotal After Discount	Subtotal SKU Sebelum Diskon - Diskon Platform SKU - Diskon Penjual SKU.
17	Shipping Fee After Discount	Biaya Pengiriman Asli - Diskon Penjual Biaya Pengiriman - Diskon Platform Biaya Pengiriman.
18	Original Shipping Fee	Biaya asli ongkos kirim
19	Shipping Fee Seller Discount	Diskon biaya pengiriman pesanan dari penjual

No.	Nama Atribut	Keterangan
20	Shipping Fee Platform Discount	Diskon biaya pengiriman pesanan dari platform
21	Taxes	Pajak yang ditanggung pelanggan
22	Order Amount	Total jumlah pesanan yang dibayar pembeli
23	Order Refund Amount	Total jumlah pesanan pengembalian dana dari semua SKU yang dikembalikan.
24	Created Time	Waktu pesanan dibuat
25	Paid Time	Waktu pesanan dibayar
26	RTS Time	Waktu penjual mengklik 'Atur Pengiriman'.
27	Shipped Time	Waktu saat status pesanan berubah menjadi In Transit.
28	Delivered Time	Waktu saat status pesanan berubah menjadi In Terkirim.
29	Cancelled Time	Waktu saat status pesanan berubah menjadi In Dibatalkan.
30	Cancel By	Pesanan dibatalkan oleh siapa
31	Cancel Reason	Alasan pembatalan pesanan
32	Fulfillment Type	Jenis Pemenuhan
33	Warehouse Name	Nama Gudang di PT X.
34	Tracking ID	Nomor pelacakan pesanan
35	Delivery Option	Opsi pengiriman pesanan
36	Shipping Provider Name	Nama shipping provider pesanan
37	Buyer Message	Pesan yang dibuat oleh Pembeli
38	Buyer Username	Username Pembeli
39	Recipient	Penerima barang
40	Country	Negara dari Pembeli
41	Province	Provinsi dari Pembeli
42	Regency and City	Kabupaten dan Kota dari pembeli
43	Payment Method	Metode pembayaran pesanan
44	Weight(kg)	Berat Paket dengan satuan kilogram
45	Product Category	Kategori product SKU
46	Package ID	ID paket pesanan
47	Seller Note	Catatan yang dibuat oleh Pembeli
48	Checked Status	Status yang telah diperiksa

C. Data Preparation

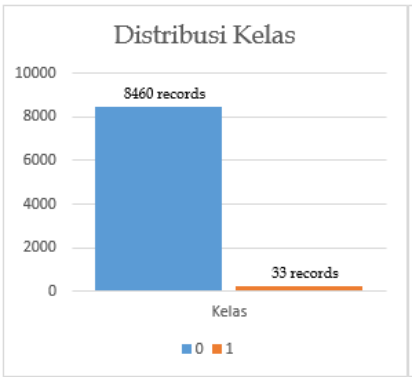
Data Preparation merupakan langkah yang banyak menghabiskan waktu dalam KDD (*Knowledge Discovery in Databases*) [9] . Dimana *Data Mining* merupakan bagian integral dari *Knowledge Discovery in Databases* (KDD). Pada tahap ini, memastikan bahwa data yang akan di olah sudah bersih dan siap untuk diproses algoritma yaitu proses algoritma *Categorical Boosting* (Catboost). Berikut tahapan dari Data Preparation yang dilakukan pada penelitian ini :

1) *Label Binary:* Pada penelitian ini *training dataset* dilabeli dengan metode binary 1 dan 0. Dimana 1 merupakan pelanggan yang melakukan pembatalan pemesanan, sedangkan 0 merupakan pelanggan yang tidak melakukan pembatalan. Untuk melakukan laber *binary classification* digunakan *library* NumPy dengan mengimpor ‘import numpy as np’. Dan untuk mengubah nilai label menjadi 1 dan 0 digunakan fungsi 'np.where()'.
2) *Handling Missing Values and Inconsistent:* Untuk menangani *missing values* atau nilai kosong dapat dilakukan dengan mengganti *missing value* dengan *value* yang sesuai kebutuhan menggunakan metode ‘fillna()’, dan untuk menghapus baris yang mengandung missing value dapat menggunakan metode ‘dropna()’. Sedangkan untuk menangani *inconsistent values* menggunakan metode ‘replace’.
3) *Exploratory Data Analysis:* Tahap ini merupakan proses eksplorasi data yang tujuannya adalah memahami karakteristik data yang akan kita lakukan analisis kedepannya [6] . Selanjutnya dengan mencari korelasi antar variabel dengan menggunakan *library* ‘matplotlib’ dan ‘seaborn’ untuk membuat sebuah *heatmap* dari korelasi antar variabel dalam sebuah *dataframe*.
4) *Encoding Categorical Variables:* Untuk melakukan *encoding categorical* variabel perlu mengimport *library* ‘LabelEncoder’ dari scikit-learn (‘sklearn’) untuk mengkonversi variabel kategorik dalam sebuah dataframe menjadi variabel numerik yang dapat digunakan dalam analisis data. Dalam melakukan *encoding* pada variabel kategorik menggunakan method fit_transform() dari objek LabelEncoder. Berikut variabel kategorikal pada data penelitian pada Tabel II:

TABEL II
KOLOM KATEGORIKAL

No.	Nama Kolom Kategorikal
1.	Product Name
2.	Variation
3.	Created Time
4.	Delivery Option
5.	Province
6.	Regency and City
7.	Payment Method
8.	Product Category

5) *Normalizing Numerical Variabel:* Untuk melakukan normalisasi pada variabel numerik dapat menggunakan log natural (ln). Hal ini dilakukan untuk mengurangi efek dari data yang memiliki range nilai yang besar dan untuk menghindari *skewed* data pada variabel tersebut.
6) *Resampling:* Dalam data ini data mengalami ketidakseimbangan kelas seperti pada Gbr. 2 sehingga dapat mempengaruhi hasil akurasi pada model yang dipilih. Untuk menangani data yang tidak seimbang digunakan teknik *resampling* yaitu *Oversampling* dengan menggunakan metode SMOTE (*Synthetic Minority Oversampling Technique*). Hal ini dapat menghindari overfitting pada kelas mayoritas [10].



Gbr. 2 Kelas yang belum dilakukan oversampling

7) *Dimension Reduction:* Untuk lebih meningkatkan hasil klasifikasi, digunakan metode pengurangan dimensi yaitu menggunakan konsep PCA (*Principal Component Analysis*). PCA (*Principal Component Analysis*) adalah sebuah metode atau teknik yang umum digunakan untuk mengurangi dimensi atau kompleksitas fitur pada data tanpa mengubah esensi data secara signifikan, sehingga data masih dapat mempertahankan informasi yang terkandung di dalamnya [11]. Konsep PCA (*Principal Component Analysis*) didasarkan pada tujuan pengurangan dimensi dari kumpulan data yang terdiri dari beberapa variabel yang berkorelasi satu sama lain sambil mempertahankan variabilitas maksimum dalam kumpulan data [12].
8) *Train-Test Split:* Tahap selanjutnya yaitu menentukan variabel dependen dan variabel independen. Dimana variabel dependen (target) pada penelitian ini yaitu 'Cancelation' diambil dari *data frame* df dan ditampung pada variabel y. Variabel dependen ini merupakan variabel yang akan diprediksi atau dijelaskan oleh variabel-variabel independen dalam *data frame* X yang merupakan dataset tanpa kolom target yaitu ‘'Cancelation''. Setelah menentukan variabel dependen dan independen, kemudian membagi *dataset* menjadi dua bagian: data pelatihan (*train*) dan data pengujian (*test*). Data pelatihan digunakan untuk melatih model, sedangkan data pengujian digunakan untuk menguji seberapa baik model tersebut dapat memprediksi hasil yang benar. Diman *dataset* yang digunakan akan dibagi menjadi dua bagian, yaitu:
    - 80% data akan digunakan sebagai data pelatihan.
    - 20% data akan digunakan sebagai data pengujian.

D. Modeling

Pada proses pemodelan pembelajaran mesin, mengaplikasikan teknik pemodelan dengan menggunakan algoritma *Categorical Boosting* (Catboost). Untuk mendapatkan Hasil model terbaik, selanjutnya dilakukan *fine-tuning hyperparameter* model untuk mendapatkan skor terbaik dengan metode Randomized Search CV. Terdapat komponen model yang disebut dengan *hyperparameter* yang dijadikan sebagai dasar aturan bagaimana mesin akan belajar. Proses menentukan nilai parameter ini disebut dengan optimisasi [6]. Parameter yang digunakan dalam pemodelan menggunakan algoritma *Categorical Boosting* (Catboost) ditunjukan pada Tabel III.

TABEL III
PARAMETER CATEGORICAL BOOSTING (CATBOOST)

Parameter	Penjelasan
iterations	Jumlah pohon yang dapat dibuat
depth	Jumlah kedalaman pohon
Learning_rate	Besaran tingkat pembelajaran
l2_leaf_reg	Koefisien L2 sebagai koefisien regularisasi model

Pada Tabel III, parameter yang akan digunakan tersebut akan dioptimasi secara default dengan melakukan 10 iterasi, dimana berarti akan dibuat 10 model terhadap variasi parameter yang berbeda. Setiap model dilatih dan performa tersebut akan dibandingkan menggunakan metrik yang ingin dioptimalkan, dalam hal ini metrik yang akan digunakan merupakan akurasi. Parameter yang memberikan akurasi terbaik akan dipilih sebagai parameter yang akan digunakan untuk membangun model.

E. Evaluation

Pada tahap evaluasi ini melihat bagaimana kualitas dan efektivitas sebelum model digunakan atau disebarkan. Pada Evaluasi ini menggunakan *Stratified 10-fold Cross-Validation*, teknik ini merupakan salah satu evaluasi model yang banyak digunakan. Selain itu perlu adanya kepastian modifikasi model agar dapat memenuhi tujuan di tahap awal [9] . Penilaian didasarkan pada performa model dalam memprediksi variabel target bernama "Cancelation" yang memiliki nilai biner (0 jika tidak ada pembatalan, dan 1 jika dibatalkan), dengan mempertimbangkan akurasi, recall, presisi, dan skor F1 yang dihasilkan oleh model [13]. Skor akurasi mengukur sebagian kecil dari jumlah prediksi yang benar dari total sampel yang diprediksi [14]. Berikut rumus akurasi, recall, presisi, dan F1 :

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$presisi = \frac{TP}{TP + FP} \quad (2)$$

$$recall = \frac{TP}{TP + FN} \quad (3)$$

$$F_1 = \frac{2 . presisi . recall}{presisi + recall} \quad (4)$$

Selain itu confusion matrix menjadi alat untuk mengevaluasi model yang dibentuk. confusion matrix merupakan dua factorial square matrix, dan seperti yang ditunjukkan pada Tabel IV berikut [14].

TABEL IV
ELEMEN CONFUSION MATRIX

	Actual Positive Class	Actual Negative Class
Predicted Positive Class	Number of True Positive (TP)	Number of True Negative (TN)
Predicted Positive Class	Number of False Positive (FP)	Number of False Negative (FN)

F. *Deployment*

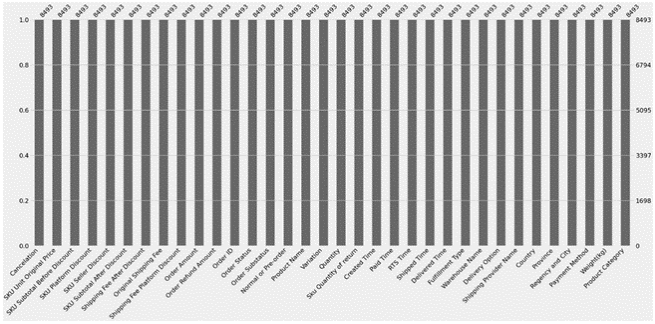
Tahap deployment merupakan tahap implementasi dari model yang telah dihasilkan dengan data baru. Model yang sudah dihasilkan akan disimpan menggunakan metode pickle.dump(). Kemudian, model yang telah tersimpan dapat dimuat kembali menggunakan pickle.load(). Proses pemuatan model ini akan diimplementasikan dalam bentuk aplikasi sederhana. Aplikasi sederhana ini akan dibangun menggunakan Streamlit, sebuah platform yang memungkinkan pembuatan aplikasi web dengan mudah dan cepat.

III. *HASIL DAN PEMBAHASAN*

Pada bab ini merupakan hasil dan pembahasan dari penelitian yang menggunakan metode CRISP-DM. Namun pada tahap *bussiness understanding* dan *data understanding* sudah disampaikan pada bab metodologi, sehingga pada bab ini akan disampaikan hasil dari tahapan *data preparation, modeling, evaluation,* dan *deployment.*

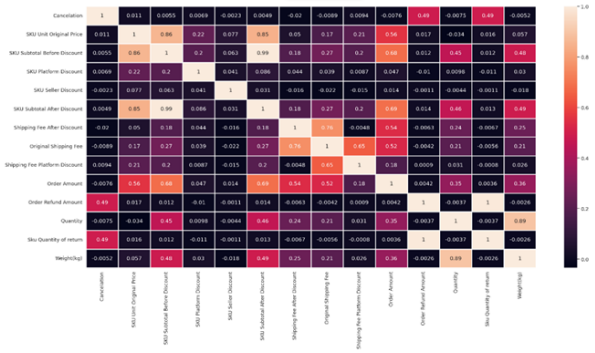
A. *Data Preparation*

Pada tahap ini penanganan terhadap *data cleaning* sudah dilakukan sehingga data sudah bersih dengan menggunakan metode 'fillna()', 'dropna()', dan 'replace', seperti pada Gbr. 3.



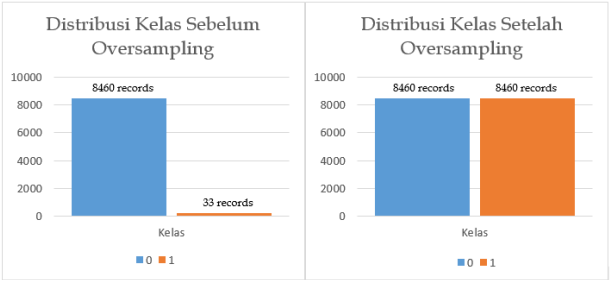
Gbr. 3 Hasil *Data Cleaning*

Pada Gbr. 3 volume data berkurang menjadi 8.493 baris dari 8.739 baris, namun hal tersebut masih dikatakan baik karena data yang berkurang tidak banyak dan merupakan data yang tidak efektif jika digunakan karena dapat mempengaruhi proses analisis. Dari data yang dimiliki terdapat korelasi dengan label pada data ini yaitu 'Cancelation' seperti pada Gbr. 4.



Gbr. 4 Korelasi label 'Cancelation'

Pada Gbr. 4 didapatkan atribut yang berkorelasi dengan label yaitu Sku Quantity of return, Order Refund Amount, Shipping Fee After Discount, SKU Unit Original Price, Shipping Fee Plat,m Discount, Original Shipping Fee, Order Amount, Quantity, SKU Discount, SKU Subtotal sebelum Discount, Weight(kg), SKU Subtotal After Discount, dan SKU Seller Discount. Namun korelasi ini perlu dieksplor lebih agar mendapatkan hasil analisis yang lebih baik. Selanjutnya data yang didapatkan tersebut dilakukan encoding dan normalisasi sebelum nantinya akan di lakukan teknik PCA untuk mengurangi dimensi agar lebih efisien pada hasil analisis. Namun sebelum dilakukannya PCA perlu dilakukan oversampling karena kelas yang dimiliki mengalami imbalance dengan hasil seperti pada Gbr. 5.



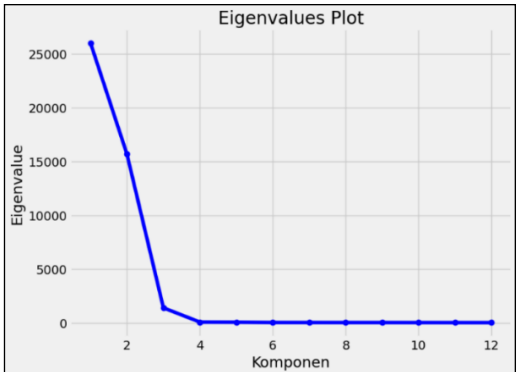
Gbr. 5 Hasil Oversampling Menggunkaan SMOTE

Setelah dilakukan resampling dengan metode *oversampling* didapatkan data yang seimbang seperti pada Gbr. 5. Sehingga proporsi tersebut bisa masuk kedalam kategori normal dan bisa digunakan sebagai data yang siap untuk di model. Selanjutnya penerapan metode PCA untuk mengurangi dimensi dari dataset agar lebih efisien digunakan untuk *modeling* dan didapatkan *eigenvalue* seperti Tabel V dan Gbr. 6.

TABEL V
NILAI EIGENVALUE SETIAP COMPONENT

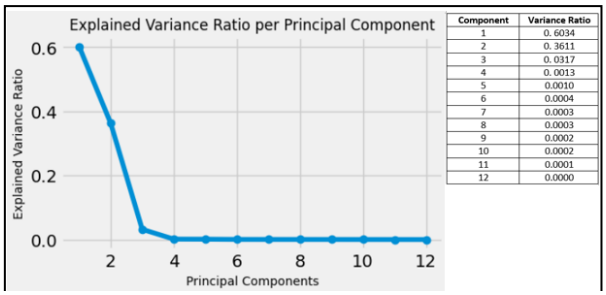
Component	Eigenvalue
1	26033.160237
2	15683.728385
3	1380.386646

4	56.064610
5	45.379722
6	16.670426
7	12.963301
8	10.930204
9	10.141272
10	6.809219
11	3.914702
12	1.649626

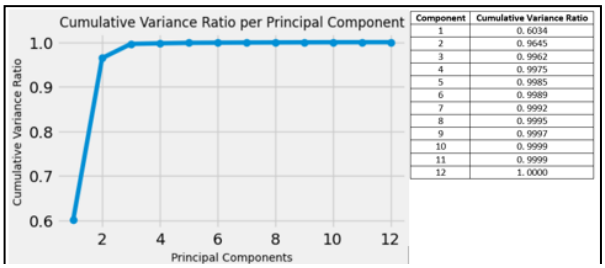


Gbr. 6 Grafik Eigenvalues

Setiap nilai *eigenvalue* menunjukkan seberapa banyak variasi yang dijelaskan oleh komponen tersebut. Semakin besar nilai eigenvalue, semakin besar kontribusi komponen tersebut dalam menjelaskan variasi dalam data. Dari Tabel 6 dapat dilihat bahwa komponen pertama memiliki eigenvalue yang paling besar, diikuti oleh komponen kedua, dan seterusnya. Dan pada Gbr. 6 dapat dilihat bahwa semakin besar nilai *eigenvalue* yang dimiliki faktor tersebut maka semakin besar pengaruhnya, dan sebaliknya jika semakin kecil nilai *eigenvalue* yang dimiliki suatu faktor maka semakin kecil pengaruhnya, dimana pada komponen diatas 12 mendapatkan niai *eigenvalues* dibawah 1. Hal ini menunjukkan bahwa komponen pertama dan kedua menjelaskan sebagian besar variasi dalam data, sedangkan komponen-komponen selanjutnya memiliki *eigenvalue* yang lebih kecil dan menjelaskan variasi yang lebih sedikit. Sedangkan untuk *Variance Ratio* dan *Cumulative Variance Ratio* hasilnya seperti pada Gbr. 7 dan Gbr. 8.



Gbr. 7 Explained Varience Ratio per Principal Component



Gbr. 8 Cumulative Varience Ratio per Principal Component

Pada Gbr. 7 dapat dilihat bahwa dari komponen 1 memiliki nilai yang signifikan dalam menjelaskan variasi dalam data. Namun semakin melandai jika komponen bertambah. Berdasarkan gambar tersebut, *explained variance ratio* untuk komponen 1 (faktor 1) adalah sebesar 0,6034, artinya faktor pertama mampu menjelaskan variansi sebesar 60,34%. Komponen 2 (faktor 2) adalah sebesar 0, 3611, artinya faktor pertama mampu menjelaskan variansi sebesar 36,11%, dan seterusnya hingga komponen 12. Sedangkan pada Gbr. 8 menjelaskan bahwa jumlah kumulatif dari *variance ratio* semakin tinggi jika komponen bertambah. Kumulatif dari variance ratio menunjukan gabungan dari beberapa atau keseluruhan nilai variance ratio dalam menjelaskan variabel-variabel tersebut. Dimana pada komponen 1 dan komponen 2 jika digabungkan akan merepresentasikan 96,45% dari keragaman total. Dari metode ini didapatkan dimensi seperti pada Tabel VI.

TABEL VI
HASIL PENGURANGAN DIMENSI MENGGUNAKAN PCA

	Dimensi
Data Latih	(13536, 9)
Data Uji	(3384, 9)

Setelah menerapkan PCA (*Principal Component Analysis*), dimensi data latih (X_train) berubah menjadi (13536, 9), yang berarti terdapat 13536 sampel (baris) dan 9 atribut (kolom) setelah dilakukan pengurangan dimensi. Sedangkan untuk dimensi data uji (X_test) juga berubah menjadi (3384, 9), yang berarti terdapat 3384 sampel (baris) dan 9 atribut (kolom) pada data uji setelah dilakukan pengurangan dimensi. Namun dari 9 kolom tersebut didapatkan variabel yang menjadi faktor yang berpengaruh terhadap pembatalan transaksi pada *platform* TikTok Shop dengan nilai diatas 0,5 yaitu seperti pada Tabel VII.

TABEL VII
VARIABEL YANG BERPENGARUH TERHADAP PEMBATALAN TRANSAKSI DI TIKTOK SHOP

No.	**Variabel**	**Komponen**	**Nilai**
1.	Payment Method	12	0.9839
2.	Regency and City	2	0.9455
3.	Order Refund Amount	8	0.9421
4.	Variation	1	0.9367
5.	Province	5	0.9208
6.	Product Category	10	0.8655
7.	Shipping Fee After	6	0.8372

	Discount		
8.	SKU Platform Discount	7,9	0.7115
9.	month	9	0.5688

Dari Tabel 8 dapat dilihat terdapat 9 variabel yang memiliki nilai diatas 0,5 yaitu Payment Method, Regency and City, Order Refund Amount, Variation, Province, Product Category, Shipping Fee After Discount, SKU Platform Discount, dan month. 9 variabel tersebut memiliki nilai diatas 0,5 dimana dengan nilai tersebut dapat menjelaskan keseluruhan data dan memberikan hasil prediksi pembatalan transaksi pada Tiktok Shop yang lebih akurat.

B. *Modeling*

Selanjutnya untuk mendapatkan model yang baik digunakan *fine-tuning hyperparameter* dengan 4 parameter, yaitu iterations, depth, learning_rate, dan l2_leaf_reg, dengan nilai parameter seperti pada Tabel VIII.

TABEL VIII
NILAI PARAMETER PADA HYPERPARAMETER TUNING

Parameter	Penjelasan
iterations	500, 1000, 2000
depth	1, 3, 6, 10
Learning_rate	0.012, 0.055, 0.064, 0.1
l2_leaf_reg	1, 3, 5

Secara *default*, Randomized Search CV melakukan 10 iterasi untuk mencari parameter terbaik. Ini berarti 10 model akan dibangun dengan variasi parameter yang berbeda. Setiap model akan dilatih dan hasil kinerjanya akan dibandingkan menggunakan metrik akurasi. Dalam penelitian ini, metrik akurasi digunakan sebagai acuan untuk mengoptimalkan performa model. Parameter yang memberikan akurasi terbaik akan dipilih sebagai parameter utama untuk membangun model. Setelah membangun parameter model secara berulang didapatkan 10 kombinasi *hyperparameter* dengan skor akurasi seperti pada Tabel IX.

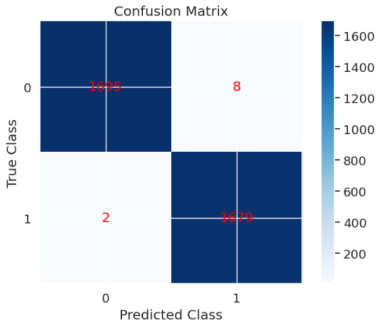
TABEL IX
HASIL 10 KOMBINASI HYPERPARAMETER DENGAN SKOR AKURASI

Rank score	Accuracy	learning_rate	l2_leaf_reg	iterations	depth
1	0.9967	0.064	1	500	10
2	0.9965	0.064	5	1000	6
3	0.9965	0.055	1	1000	10
4	0.9963	0.1	3	500	6
5	0.996	0.012	3	1000	10
6	0.9956	0.064	1	2000	3
7	0.9503	0.1	1	2000	1
8	0.9422	0.064	3	2000	1
9	0.9404	0.055	3	2000	1
10	0.9264	0.064	1	1000	1

Tabel 10 menunjukkan 10 model yang telah dibangun dengan 10 kombinasi *hyperparameter* yang berbeda serta skor akurasi yang didapatkan dari model tersebut. *Hyperparameter* terbaik ditunjukkan pada model dengan Rank Score 1 yaitu dengan iterations=500, learning_rate=0.064, depth=10, l2_leaf_reg= 1. Selanjutnya model yang sudah dioptimasi akan di latih terhadap data latih dan diuji terhadap data uji untuk didapatkan performanya.

C. *Evaluation*

Dalam penelitian ini teknik evaluasi performa menggunakan metode *Stratified 10-fold Cross-Validation*, terdapat 10 iterasi validasi silang, dan untuk setiap iterasi, akurasi model diberikan sebagai skor. Semakin tinggi nilai akurasi maka semakin baik kinerja model. Dengan menggunakan parameter iterations=500, learning_rate=0.064, depth=10, l2_leaf_reg= 1 dimana niali parameter tersebut merupakan parameter terbaik yang didapatkan pada penelitian ini didapatkan hasil seperti yang ditampilkan pada Gbr. 9.



Gbr. 9 Hasil Confusion Matrix

Dalam penemuan nilai *Confusion Matrix* terdapat 1695 prediksi benar untuk kelas 0 (true negative), 8 prediksi salah untuk kelas 0 (false positive), 2 prediksi salah untuk kelas 1 (false negative), dan 1679 prediksi benar untuk kelas 1 (true positive). Dimana hasil performa ditampilkan pada Tabel X berikut :

TABEL X
PERFORMA KLASIFIKASI

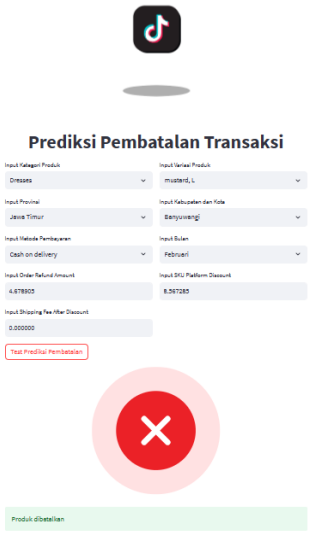
Performa Klasifikasi				
	Precision	Recall	F1-Score	Support
0	1.00	1.00	1.00	1703
1	1.00	1.00	1.00	1681
Akurasi			1.00	3384
Macro avg	1.00	1.00	1.00	3384
Weighted avg	1.00	1.00	1.00	3384

Dalam penelitian ini, akurasi model CatBoost *Classifier* yanh didapatkan adalah sebesar 0.9970 atau 99.7%. Dan untuk Hasil Presisi yang mengukur sejauh mana prediksi positif yang benar, recall mengukur sejauh mana model dapat mengidentifikasi kelas positif dengan benar, dan f1-score dimana rata-rata harmonik dari presisi dan recall menunjukan

bahwa model memiliki presisi, recall, dan f1-score yang sempurna (1.00) untuk kedua kelas.

D. *Deployment*

Tahap *deployment* ini merupakan tahap implementasi dari model yang sudah didapatkan dengan data baru. model yang sudah didapatkan tersebut disimpan dengan menggunakan metode pickle.dump(). Kemudian dari model yang sudah tersimpan tersebut nantinya untuk memuat model digunakan pickle.load(). Dimana pada load model ini di implementasikan dalam bentuk aplikasi sederhana. Aplikasi sederhana yang digunakan menggunakan streamlit seperti pada Gbr. 10.



Gbr. 10 Aplikasi Prediksi Pembatalan Transaksi Tiktok Shop

IV. KESIMPULAN DAN SARAN

Dari hasil penelitian ini dapat disimpulkan bahwa dengan menggunakan metode CRISP-DM dalam penelitian ini yang dimulai dari *business understanding, data understanding, data preparation, modeling, evaluation,* dan *deployment*. Didapatkan hasil akurasi sebesar 0.9970 atau 99.7%. Dan memiliki presisi, recall, dan f1-score yang sempurna (1.00) untuk kedua kelas. Dan dengan menggunakan teknik PCA 12 komponen didapatkan faktor-faktor yang sangat mempengaruhi terjadinya pembatalan atau tidaknya transaksi pada Tiktok Shop adalah Payment Method, Regency and City, Order Refund Amount, Variation, Province, Product Category, Shipping Fee After Discount, SKU Platform Discount, month,dan SKU Platform Discount.

Untuk penelitian berikutnya dapat memperluas sampel data dengan mengumpulkan lebih banyak data dari pengguna TikTok Shop yang melakukan pembatalan transaksi. Selain itu dapat melakukan lebih banyak variasi fold pada proses validasi silang (*cross validation*). Dan juga dapat menggabungkan PCA (*Principal Component Analysis*) dengan teknik seleksi fitur lainnya untuk mengatasi kelemahan yang mungkin dimiliki oleh PCA.

REFERENSI

[1] Pamungkas, Ajar.(2022).Benarkah Online Shop adalah Bisnis Terbaik Masa Kini?.Diakses pada 31 Maret 2023, dari https://majoo.id/solusi/detail/online-shop-adalah

[2] Hasibuan, Z., & Ramadhani, S. (2022). Faktor-Faktor Yang Menjadi Pertimbangan Konsumen Dalam Membeli Produk Pada Fitur Tiktok Shop (Study Pada Pelanggan Tiktok Shop Dikalangan Mahasiswa/I Medan). Syntax Literate: Jurnal Ilmiah Indonesia, 7(12). https://doi.org/10.36418/syntax-literate.v7i12.11193

[3] Dekou, R., Savo, S., Kufeld, S., Francesca, D., & Kawase, R. (n.d.).(2021). Machine Learning Methods for Detecting Fraud in Online Marketplace s. http://https://xuyunzhang.github.io/pstci2021/

[4] North, Matthew A..(2016).Data Mining for the Masses, Second Edition.CreateSpace Independent Publishing Platform

[5] Larose, Daniel T.(2019).Data Mining Methods and Models.New Jersey:John Wiley & Sons, Inc.

[6] Christian, J., Ernawati, I., & Chamidah, N..(2022). Implementasi Penggunaan Algoritma Categorical Boosting (Catboost) Dengan Optimisasi Hiperparameter Dalam Memprediksi Pembatalan Pesanan Kamar Hotel.Jakarta:Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA) https://conference.upnvj.ac.id/index.php/senamika/article/view/2230

[7] Hancock, J. T., & Khoshgoftaar, T. M. (2020). CatBoost for big data: an interdisciplinary review. Journal of Big Data, 7(1). https://doi.org/10.1186/s40537-020-00369-8

[8] Suci, W., & Samsudin, S. (2022). Algoritma K-Nearest Neighbors dan Synthetic Minority

[9] Pradnyana, G.A.,Agustini, Ketut.(2022).Konsep Dasar Data Mining.Tangerang Selatan: Universitas Terbuka. https://pustaka.ut.ac.id/lib/msim4403-data-mining/

[10] Wu, Y., Radewagen, R..(2022).7 Techniques to Handle Imbalanced Data. Diakses pada 04 April 2023, dari https://www.kdnuggets.com/2017/06/7-techniques-handleimbalanced-data.html

[11] Adnyana, I. G. N. D., Arjuna, R. M., Indraini, A. N., & Pasvita, D. S. (2021). Pengaruh Seleksi Fitur Pada Algoritma Machine Learning Untuk Memprediksi Pembatalan Pesanan Hotel.Jakarta:Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA).

[12] Gadekallu, T. R., Khare, N., Bhattacharya, S., Singh, S., Maddikunta, P. K. R., Ra, I. H., & Alazab, M. (2020). Early detection of diabetic retinopathy using pca-firefly based deep learning model. Electronics (Switzerland), 9(2). https://doi.org/10.3390/electronics9020274

[13] Andriawan, Z. A., Purnama, S. R., Darmawan, A. S., Ricko, Wibowo, A., Sugiharto, A., & Wijayanto, F. (2020). Prediction of Hotel Booking Cancellation using CRISP-DM. ICICoS 2020 - Proceeding: 4th International Conference on Informatics and Computational Sciences. https://doi.org/10.1109/ICICoS51170.2020.9299011

[14] Chen, Y., Ding, C., Ye, H., & Zhou, Y. (2022). Comparison and Analysis of Machine Learning Models to Predict Hotel Booking Cancellation. https://doi.org/10.2991/aebmr.k.220307.225