Journal of Emerging Information System and Business Intelligence 6(1) p.49-62 (2025)

Journal of Emerging Information System and Business Intelligence
ISSN: 2774-3993
Journal homepage: https://ejournal.unesa.ac.id/index.php/JEISBI/

Implementation of the Support Vector Machine (SVM) Algorithm in Predicting Transaction Cancellations at Shopee E-commerce

Ridhotul Maulidia[1], Wiyli Yustanti[2]
[1,2]*Universitas Negeri Surabaya, Surabaya, Indonesia*
ridhotul.20090@mhs.unesa.ac.id, *wiyliyustanti@unesa.ac.id*

ABSTRACT

In today's digital era, online shopping through e-commerce platforms such as Shopee is increasingly in demand. However, the problem of transaction cancellation is still a big challenge for sellers as it can lead to financial losses and operational instability. This research uses the Support Vector Machine (SVM) algorithm to predict transaction cancellations to improve efficiency and reduce the risk of loss for sellers, with data taken from the transaction history of Toko Nafystore.id on Shopee. The CRISP-DM method is applied in the data analysis process, including data cleaning, encoding, normalization, as well as the application of Principal Component Analysis (PCA) for dimension reduction. The SMOTE oversampling technique was used to handle data imbalance, while model evaluation was performed using K-Fold Cross-Validation with various SVM kernels, such as linear, polynomial, RBF, and sigmoid. The results showed that the linear kernel provided the best performance with 95.57% accuracy. The main factors affecting transaction cancellation include total payment, estimated shipping discount, and receipt number. This study makes an important contribution in understanding the factors of transaction cancellation in Shopee and can be applied to other e-commerce platforms to improve the reliability of online transactions.

Keyword: transaction cancellation, Support Vector Machine (SVM), CRISP-DM, data preparation, PCA, K-Fold Cross-Validation.

Article Info:	**Corresponding Author**
Article history:	Ridhotul Maulidia
Received November 12, 2024	Universitas Negeri Surabaya, Surabaya, Indonesia
Revised February 5, 2025	ridhotul.20090@mhs.unesa.ac.id
Accepted March 8, 2025	

1. INTRODUCTION

The development of information technology has encouraged the growth of e-commerce such as Shopee, which changes people's shopping patterns and provides convenience, especially for housewives who can shop or do business without having to leave the house. Research by [1] shows that the adoption of digital technology by households significantly increases the frequency of online transactions, especially on e-commerce platforms that provide flexible payment systems such as COD (Cash on Delivery).

However, sellers often face transaction cancellation constraints, especially in COD payment systems, which can lead to financial losses and operational disruptions. According to [2], the transaction cancellation rate on the COD system reaches more than 30% in some online stores, which is a serious challenge in maintaining smooth operations and cash flow.

This study uses the Support Vector Machine (SVM) method to predict transaction cancellations on Shopee, with a dataset of customer transaction history at Nafystore.id stores that sell various hampers products. SVM was chosen because of its excellence in classifying non-linear data to analyze transaction cancellation patterns in more depth.

The results of this study are expected to help sellers in managing cancellation risk, become a reference for other e-commerce to improve transaction efficiency, and encourage technological

ISSN: 2774-3993
innovation in transaction risk management to increase customer satisfaction and digital economy growth.

2. METHODS

This research uses the CRISP-DM (Cross-Industry Standard Process for Data Mining) method which is conceptually good.[7] This method consists of six stages, namely: Business understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment."[8] This cycle of stages is shown in Figure 1.

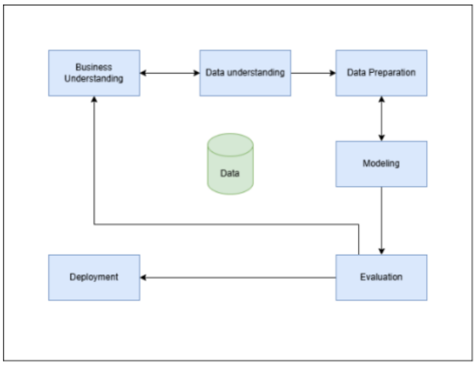

Figure 1. CRISP-DM (Cross Industry Standard Process for Data Mining

A. Business Understanding

The initial stage in the CRISP-DM method is business understanding, which aims to reduce the transaction cancellation rate in Shopee, improve customer satisfaction, as well as identify the main factors causing cancellations. This analysis of cancellation patterns helps sellers in taking effective measures, such as improving delivery reliability and clarifying product descriptions. In addition to cancellation prediction, this research also explores hidden patterns in consumer behavior, including the influence of payment methods. Collaboration with sellers is necessary so that the prediction model developed fits operational needs and provides added value to the business.

B. Data Understanding

In the data understanding stage, an evaluation of the dataset covering the transaction history of the Nafystore.id store on Shopee during the period January-December 2023, with 49 attributes and 6,236 rows of data, was conducted. This process includes identifying issues such as missing values, duplication, or outliers, as well as exploring the data to understand distributions and patterns that may affect modeling results. This stage ensures that the data is ready to be used for further analysis. The following is a description of each column of the dataset in Table 1.

Table 1. DATASET OF EACH ATTRIBUTE

No	Atribute Name	Description
1	Source.Name	Source the order came from
2	Order No. Identification	number of each order
3	Order Status Current	Current status of the order
4	Cancellation/Return Status	Status of order cancellation made by the buyer
5	Receipt No. Tracking	Receipt of delivery status
6	Shipping Option	Shipping method selected by the buyer
7	Delivery to counter/pick-up	Indication that the buyer picked up the order at a specific counter or location
8	Order Must Be Delivered Before (Avoid delays)	Delivery deadline to avoid delays

9	Delivery Time	Set Delivery schedule set
10	Order Time	Time the order was first created by the buyer
11	Time Payment Made Time	Time made that the buyer makes payment
12	Parent SKU	Product unique code
13	Product Name	A product name that the buyer has ordered
14	SKU Reference	Number Reference number related to the SKU product
15	Variation Name	Name of the product variation
16	Initial Price	Initial price of the product before discount or rebate
17	Price After Discount	Price of the product after the discount is applied
18	Quantity	The number of products purchased
19	Total Product Price	Total price of all products purchased
20	Total Discount	Total discount value applied to the order
21	Discount From Seller	Amount of discount given by the seller
22	Discount From Shopee	Total discount which Shopee has given
23	Product Weight	The individual weight of each product.
24	Number of Products Ordered	The total quantity of products that have been ordered.
25	Total Weight	The combined weight of all products in an order.
26	Voucher Covered by Seller	The total discount amount offered by the seller in the form of a voucher.
27	Coin Cashback	The total cashback amount granted to the buyer in the form of coins.
28	Voucher Covered	The portion of the voucher amount subsidized by Shopee.
29	Discount Package Additional	Extra discounts applied to specific product bundles.
30	Discount Package (provided by Shopee)	Extra discounts offered by Shopee for selected product bundles.
31	Discount Package (Provided by Seller)	Additional discount from the seller for product bundles.
32	Shopee Coin Discount	Price reduction in the form of Shopee coins.
33	Credit Card Discount	Special discount applied when using a credit card.
34	Shipping Fee Paid by Buyer	Shipping cost paid by the buyer.
35	Estimated Shipping Fee Discount	Estimated shipping cost reduction.
36	Return Shipping Fee	Shipping cost that must be paid for product returns.
37	Total Payment	Total amount the buyer needs to pay.
38	Estimated Shipping Fee	Estimated shipping cost to be paid.
39	Buyer's Note	Special notes provided by the buyer.
40	Order Notes	Additional notes related to the order.
41	Username (Buyer)	Buyer's username.
42	Recipient’s Name	Recipient's name.
43	Phone Number	Recipient’s phone number.
44	Shipping Address	Complete address where the order will be delivered.
45	City/Regency	City or regency of the recipient.
46	Province	Province of the recipient.

47	Order Completion Time	Time when the order was completed.
48	Payment Method	Payment method chosen by the buyer.
49	Returned Quantity	Quantity of items returned to the seller

C. Data Preparation

The data preparation process consists of eight stages, to ensure that the data to be processed is clean and ready for processing algorithms, in this study using the Support Vectord Machine (SVM) algorithm. Data preparation steps in this research include:

1. Label Binary

   In this study, the test data is labeled 1 for canceled orders and 0 for non-canceled orders. This process helps the algorithm distinguish between the two results. The implementation was done using the NumPy library, with the `np.where()` function to convert labels to binary format based on certain conditions. This approach simplifies data processing and analysis in Machine Learning modeling.

2. Handling Missing Values and Inconsistent

   To address missing values, the fillna() method from the pandas library was used to replace empty values with statistical measures such as the mean, median, mode, or other predefined values [9]. Alternatively, rows containing missing values can be removed using the dropna() method to reduce noise in the dataset. For inconsistent values, the replace() method was used to correct or standardize erroneous entries, such as converting 'NA' to 'NaN' or ensuring a uniform date format.

3. Exploratory Data Analysis

   This stage involves exploring the dataset to understand its attributes and identify patterns, trends, and anomalies that may influence the analysis results. To visualize the correlation between variables within the DataFrame, the matplotlib and seaborn libraries were used to create a heatmap. This heatmap helps in identifying relationships between variables and selecting relevant features for further modeling.

4. Encoding Categorical Variables

   The first step in encoding categorical variables is to import the LabelEncoder module from Scikit-learn, which converts categorical data into numerical format. The fit_transform() method from LabelEncoder was used to transform each categorical variable into a numerical representation. The categorical variables used in this study are presented in Table 2.

Table 2. VARIABLES OF CATEGORIES

No.	Categorical Column Name
1	Order Status
2	Shipping Option
3	Order Creation Time
4	Payment Time
5	Payment Method
6	Product Name
7	Variation Name
8	SKU Reference Number
9	Product Weight
10	Total Weight
11	Bundle Discount
12	City/Regency
13	Province

ISSN: 2774-3993

5. Normalizing numerical variables

   To standardize numerical data, a natural logarithm (ln) transformation is used. This technique helps to reduce the impact of outliers, equalize the scale of variables, and improve the accuracy of the model by minimizing bias. Before application, the dataset was checked for zero or negative values, as the logarithmic transformation is undefined for such values.

6. Resampling

   The dataset exhibited class imbalance, as shown in Figure 2, which could negatively impact model accuracy by biasing predictions toward the majority class. To mitigate this issue, the Synthetic Minority Oversampling Technique (SMOTE) was applied. This method generates synthetic examples for the minority class, reducing the risk of overfitting and improving classification performance [10].

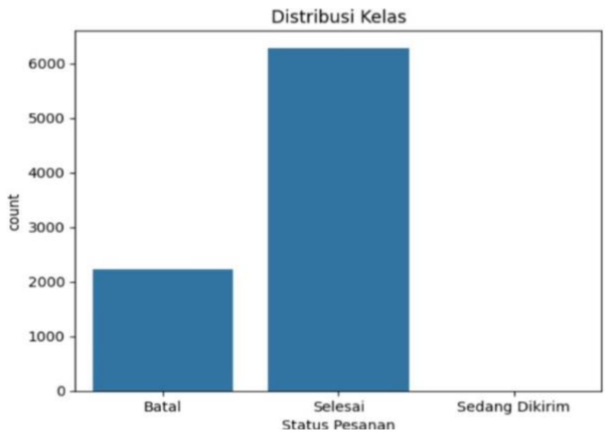

Figure 2. Distribution Class Before Resampling

7. Dimension Reduction

   At this stage, Principal Component Analysis (PCA) is used to improve classification accuracy by reducing the dimensionality of the data while retaining important information. PCA aims to simplify datasets that have many related variables, while maintaining maximum variance. This method helps improve computational efficiency and reduce redundancy in data representation.

8. Train-Test Split

   Prior to modeling, variables were identified with Cancellation as the dependent variable (Y), which is the focus of prediction. Meanwhile, DataFrame X contains all independent variables, except Cancellation, that play a role in influencing the order cancellation decision. This step is important to ensure a clear separation between the predicted variable and the factors that influence it. After defining these variables, the dataset was divided into two parts:

   - 80% Training Data – Used for model training.
   - 20% Testing Data – Used for model evaluation.

D. Modeling

The modeling process uses the Support Vector Machine (SVM) algorithm with K-Fold Cross-Validation to improve model performance. The dataset is divided into several subsets, where each subset alternates as test data, so the model is tested with various combinations of data. The average performance of this process provides a more stable model evaluation. SVM constructs a hyperplane to optimally separate the two classes, ensuring good generalization and reducing the risk of overfitting. Figure 3 illustrates the workflow from data input to model evaluation:

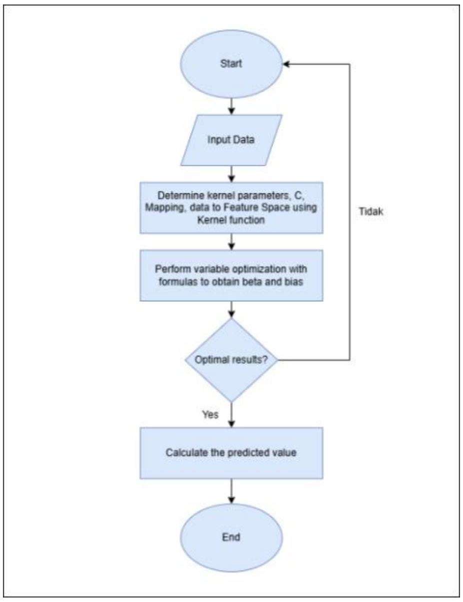

Figure 3. Flowchart Support Vector Machine (SVM)

The modeling process with Support Vector Machine (SVM) and K-Fold Cross-Validation starts by dividing the test data into several subsets, where each subset is alternately used as test and training data. Averaging the results of this process improves the stability of the model performance. The SVM then forms a hyperplane to separate the two classes, ensuring the model can generalize well and reducing the risk of overfitting. Figure 3 shows the flow from the data input stage to model evaluation.

E. Evaluation

In the evaluation stage, the performance of the Shopee transaction cancellation prediction model is measured using the main evaluation metrics, namely accuracy, precision, recall, and F1-Score calculated based on the Confusion Matrix. Accuracy shows the percentage of correct predictions against all data, precision measures the accuracy of predictions of cancellations that actually occur, recall measures how well the model detects all cases of cancellation, and F1-Score provides a balance between precision and recall. This evaluation is to see the effectiveness of the model and consider adjusting the model or hyperparameters with business objectives. The following are the formulas for accuracy, precision, recall, and F1-Score:

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

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

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

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

Figure 4. Confussion Matrix Formula

F. Deployment

At the deployment stage, to build a Graphical User Interface (GUI) transaction cancellation system on the Shopee e-commerce website based on this research will use one of the SDLC (Systems Development Life Cycle) methods, namely waterfall. Fig. 4 below is the system flow of the GUI that will be built:

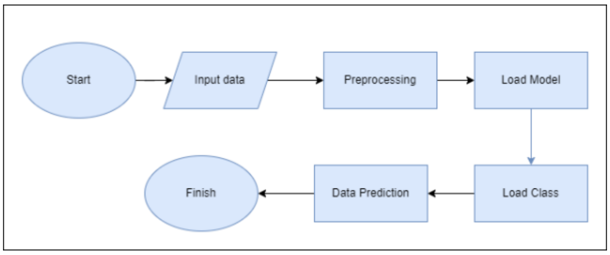

Figure 5. System Flow

3. RESULTS AND DISCUSSION

This chapter will present the results of research using the CRISP-DM method. This chapter will describe the results of the Data preparation, Modeling, Evaluation, and Deployment stages thoroughly. Meanwhile, the Business understanding and Data Understanding stages have been discussed in detail in the previous chapter.

A. Data Preparation

In this data cleaning stage, we used the 'fillna()', 'dopna()', and replace methods. The volume of data did not change at 8,501 rows after the cleaning process. This shows that the dataset has only replaced inconsistent values with alternative values using the 'replace' method and filled in missing values using the 'fillna()' method. The following is a dataset that shows the relationship of the label 'status_order' as depicted in Fig. 6.

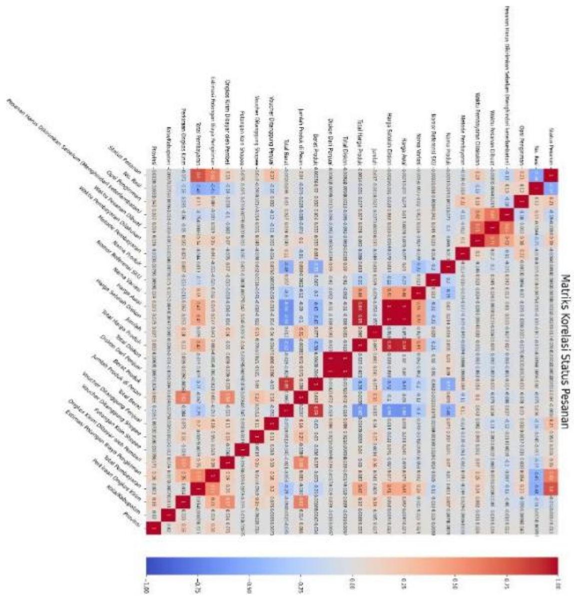

Figure 6. Correlation of Label Order Status

In Fig. 6, there are three attributes that are highly correlated with the 'Order Status' variable: Receipt No., Total Payment, and Estimated Fee Deduction. In addition to these three attributes, other attributes also have a correlation with the 'Order Status' variable, although they have a weaker correlation than the three previous attributes. But even though it has a weak correlation, it can also have an influence on predictions, where interactions between variables can be considered. After the encoding and normalization process, the PCA technique is applied to reduce the dimensionality of the data which can improve the efficiency of the model.[12] To handle data imbalance, an oversampling technique is applied so that the model can more reliably predict changes in order status.

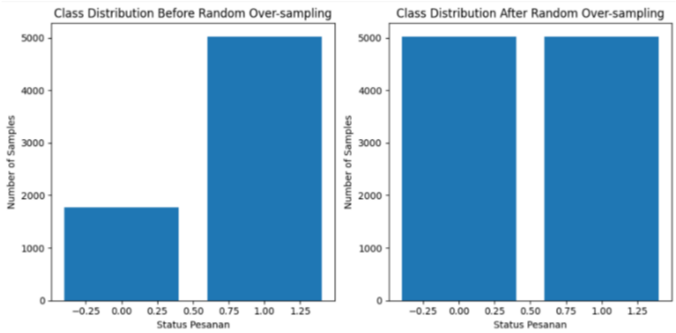

Figure 7. Data Distribution After Oversampling

After resampling using the oversampling method as illustrated in Fig.6 above, the testing data has data imbalance with 5,026 samples and category 0 (success) and 1,774 samples in category 1 (cancel). After the oversampling process, the number of samples of both categories became balanced, i.e. 5,026. This result is to ensure that the model will not be biased towards either class and is ready to proceed to the modeling stage. In addition, using the PCA technique to reduce the dimensions of the dataset to generate eigenvalues [13], which are presented in table 3 and Fig. 8.

Table 3. LIFT OF EIGENVALUES FOR EACH COMPONENT

Component	*Eigenvalue*
1	4.7421
2	3.3179
3	2.5184
4	2.1152
5	1.8470
6	1.7583
7	1.5219
8	1.2879
9	0.9980
10	0.9738
11	0.8858
12	0.7804
13	0.6814

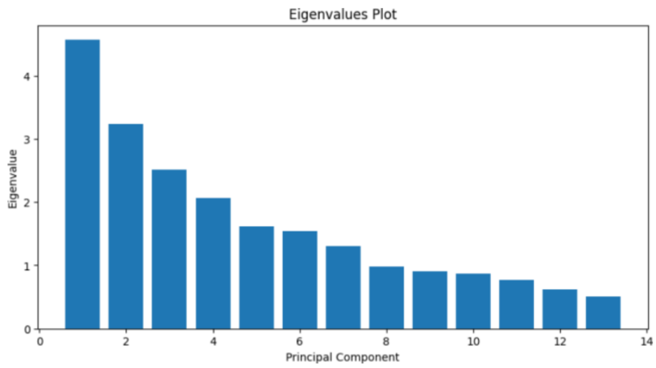

Figure 8. Eigenvalues Graphic

The eigenvalue shows the contribution of each component in explaining the variation of the data. Based on Table III, the first (4.5670) and second (3.2304) components have the largest role in explaining the variation of the dataset, as shown in the increasing cumulative variance ratio in Figure 8. Although other components still contribute, their role is decreasing, so only the first few components are considered sufficient for analysis, as shown in Figures 9 and 10.

ISSN: 2774-3993

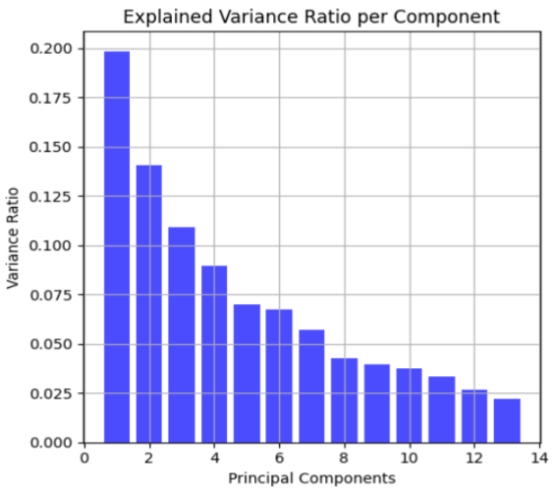

Figure 9. Explained Varience Ratio

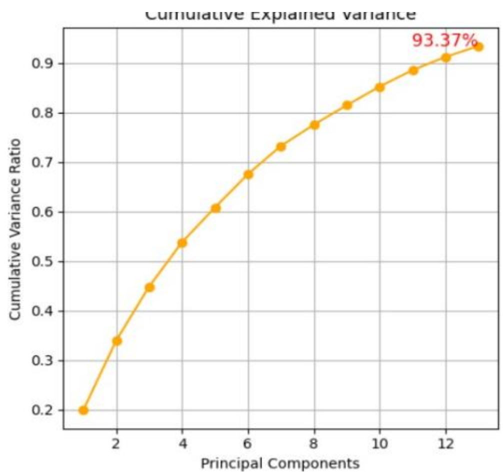

Figure 10. Cumulative Explained Variance

Fig. 9 and Fig. 10 show that the first component, with an explained variance ratio of 0.200, explains 20% of the variation in the data, making it the most significant component in this analysis. The second component adds up to 34% of the total variation, indicating that the second component also plays an important role. The third component, with a variance ratio of 0.110, adds up to 46% in total, although its contribution is smaller than the two principal components. The fourth to sixth components add a smaller proportion with values of 0.090, 0.075, and 0.065 respectively. Overall, the first 13 components explain 93.37% of the variation in the data, providing an overview of the patterns in the dataset. From this stage, the dimensions were obtained as presented in Table 4.

TABLE 4 VARIABLES CORRELATION WITH PRINCIPAL COMPONENT

No.	Variabel	Komponen	Nilai
1	Order Time Made	3	0.5252
2	Payment Method	7	0.5593
3	Shopee Coin Discount	8	0.9543
4	SKU Reference Number	9, 10	0.6931
5	Shopee Covered Voucher	11	0.5635
6	Seller Covered Voucher	12	0.6765
7	Shipping Cost Paid by Buyer	13	0.6378

From table 4, it can be seen that there are seven variables that have a value above 0.5, namely Order Time Created, Payment Method, Shopee Coin Discount, SKU

Reference Number, Shopee Covered Voucher, Seller Covered Voucher, and Shipping Cost Paid by Buyer. These variables are distributed across several principal components, making significant contributions to explaining the variations in the data. With these values, these variables are able to explain the overall data and provide more accurate prediction results for canceled transactions on the Shopee platform.

B. Modeling

Next, to obtain a good model, the best kernel was selected based on accuracy, precision, and recall metrics. Then the results are obtained as presented in Fig. 11

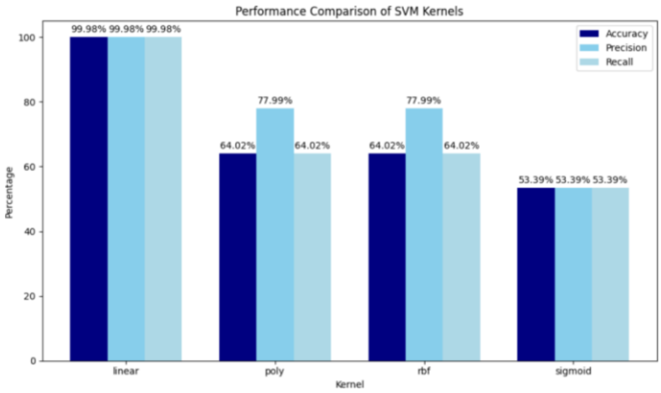

Figure 11. Graphs of Kernel Test Results

Fig. 11 illustrates the test results for the four kernels used: linear, polynomial (poly), radial basis function (RBF), and sigmoid kernels. The linear kernel shows the best performance with an average accuracy, precision, and recall of 99.98%. the polynomial (poly) and Radial Basis Function (RBF) kernels show lower performance with an average of about 64.02%. the sigmoid kernel has the lowest accuracy with an average of 53.39%. overall, the linear kernel is proven to be the most effective for predicting canceled transactions on the dataset. Then next, in the application of the Support Vector Machine (SVM) model, testing was carried out by applying K-fold Cross Validation using 3, 5, and 10 folds to reduce the risk of underfitting and overfitting. [14] Then from the trial, the following results were obtained:

Table 5 AVERAGE K-FOLD CROSS VALIDATION TEST

Fold	**Mean Train**	**Mean Test**
3	0.9532	0.9552
5	0.9532	0.9557
10	0.9532	0.9557

From table 5, the SVM model using 3, 5, and 10 folds as test material produces stable and consistent performance in each iteration. Hence, 10 fold cross validation is used in the final evaluation because this method provides more accurate and stable results. Then next, the SVM model evaluation technique is carried out using the accuracy, precision, recall, and F-1 Score metrics. Which produces the following test results:

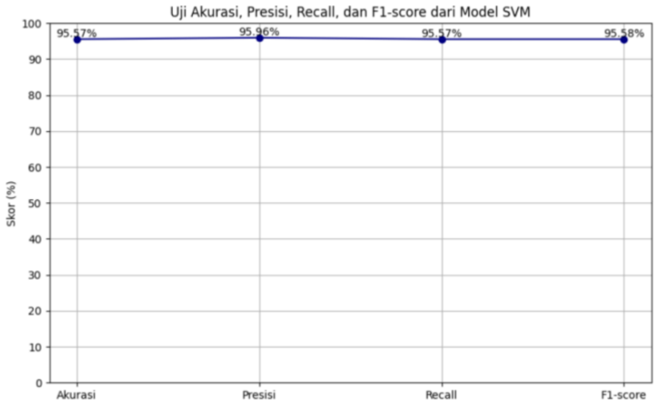

Figure 12. Result of VSM graph

From the SVM model performance test results illustrated in Fig. 12, obtained accuracy of 95.57%, precision of 95.96%, recall of 95.57%, and F1-Score of 95.58% from these results it can be concluded that there is no indication of a problem, because the model shows stable and accurate performance on both training data and test data.

C. Evaluation

Furthermore, model evaluation is carried out using the Confusion Matrix technique with the following results:

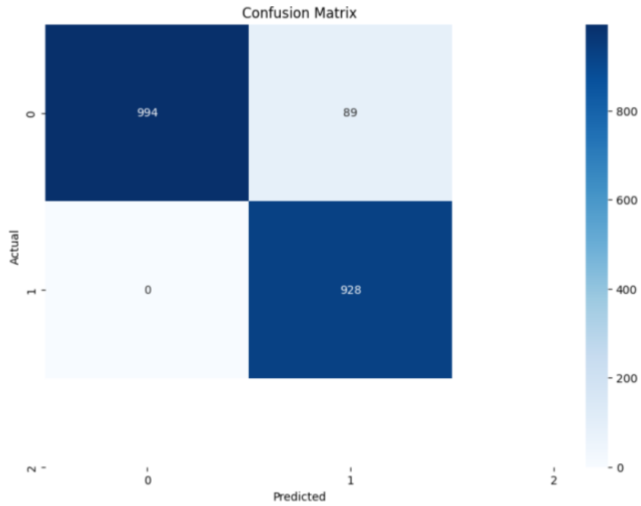

Figure 13. Confusion Results Matrix

Based on the evaluation, Figure 13 shows that the model successfully predicted 994 samples correctly as the first class (True Positive/TP). However, there were 89 samples from the second class that were misclassified as the first class (False Positive/FP), and 0 samples from the second class that were not identified as the second class (False Negative/FN). In addition, the model also successfully predicted 928 samples correctly as the third class (True Negative/TN). Although there were some False Positives, the model still performed well in data classification. This stage was conducted to provide a detailed overview of the model's performance in accurately classifying the data using Confusion Matrix.

D. Deployment

In the deployment section, the transaction cancellation prediction system is built using streamlit. [The model is stored using the 'joblib' library in pickle file format. In its development, this system uses the waterfall method. [Each stage is completed thoroughly before proceeding to the next stage. With the waterfall method, each stage is developed systematically and coherently, ensuring good integration between the Machine Learning model and the web interface. Fig. 14 below is the flow of the waterfall method used in the deployment stage of this research:

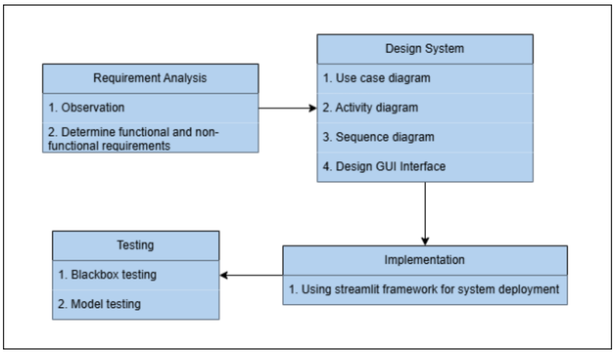

Figure 14. Waterfall Flow

After going through the stages of deployment, Fig. 15 to Fig. 17 shows the GUI view of the system built:

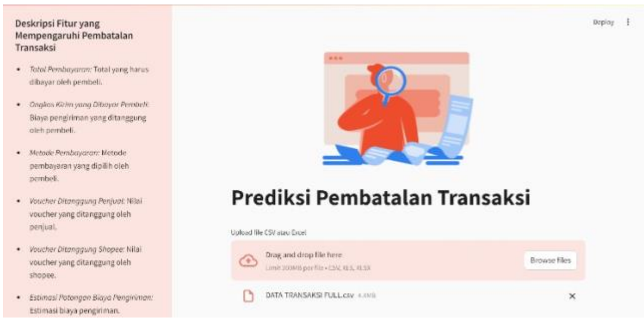

Figure 15. System Output 1

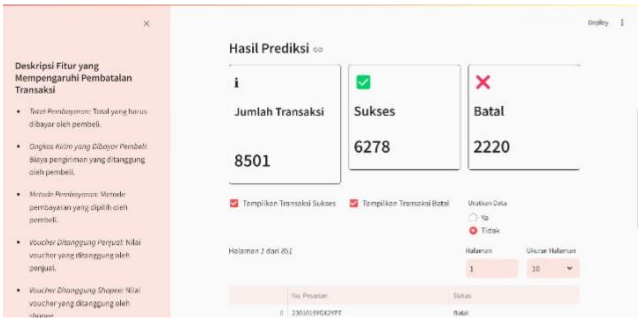

Figure 16. System Output 2

Figure 17. System Output 3

CONCLUSION

Based on the research conducted, several conclusions were obtained. The CRISP-DM method proved to be quite effective in analyzing and modeling e-commerce data through the stages of Data Preparation, Modeling, Evaluation, and Deployment. In the Data Preparation stage, Data Preprocessing is carried out such as data cleaning, encoding, normalization, and handling class imbalance using SMOTE. In addition, PCA is applied to reduce the dimensionality of the data. In the Modeling stage, Support Vector Machine (SVM) algorithm is used with evaluation through K-Fold Cross-Validation (3, 5, and 10 folds). The evaluation results show that SVM with linear kernel has a good performance with an average accuracy of 95.57%, as well as precision 95.96% and recall 95.58% on training and test data. The developed model has passed the Deployment stage using Streamlit, with integration of preprocessing and models stored using Joblib. This research provides significant benefits for sellers on the Shopee e-commerce platform, as it helps them analyze the causes of transaction cancellation. With the developed prediction system, sellers can identify patterns and factors that influence cancellations, thereby improving their service quality and business strategy.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to Dr. Wiyli Yustanti, S.Si., M.Kom. for her valuable guidance and support during this research. I also appreciate the lecturers and staff of the Department of Information Systems, Surabaya State University, who have created an inspiring academic environment. Finally, I would like to thank my family and friends for their continuous support and motivation throughout this journey.

REFERENCES

[1] M. Rafki Nazar, A. Timotius Oloando, M. A. Putri, C. Berri, dan M. Tazkia, “Pengaruh Perkembangan Teknologi terhadap E-Commerce.”

[2] R. S. Priyambada, N. A. Ramadani, dan A. Risko Faristiana, “Perubahan Budaya Ibu Rumah Tangga Pasca Maraknya E-Commerce Di Indonesia,” vol. 1, no. 3, hlm. 72–85, 2023, doi: 10.59024/simpati.v1i3.222.

[3] M. Wibisana, J. Neltje, dan D. Fitriana, “Perlindungan Hukum Bagi Pelaku Usaha Terhadap Tindakan Pembatalan Pembayaran Oleh Konsumen E- Commerce Dalam Sistem Cash on Delivery (COD) Menurut Undang- Undang Nomor 8 Tahun 1999 Tentang Perlindungan Konsumen,” vol. 17, no. 2, hlm. 437– 464, 2023, doi: 10.31599/krtha.v17i2.2330.

[4] J. Homepage *dkk.*, “MALCOM: Indonesian Journal of Machine Learning dan Computer Science Prediction System for Determine The Number of Drug Orders using Linear Regression,” vol. 2, hlm. 62–70, 2022.

[5] A. Sanjurjo-de-No, A. M. Pérez-Zuriaga, dan A. García, “Analysis and prediction of injury severity in single micromobility crashes with Random Forest,” *Heliyon*, vol. 9, no. 12, Des 2023, doi: 10.1016/j.heliyon.2023.e23062.

[6] X. Xiahou dan Y. Harada, “B2C E-Commerce Customer Churn Prediction Based on K-Means and SVM,” *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 17, no. 2, 2022, doi: 10.3390/jtaer17020024.

[7] A. Massahiro, S. Instituto, D. P. Tecnológicas, D. S. Paulo, dan F. Univer-, “The

Evolution of CRISP-DM for Data Science : Methods , Processes and The evolution of CRISP-DM for Data Science : Methods ," no. October, 2024, doi: 10.13140/RG.2.2.22493.42721.

[8] S. N. Kane, A. Mishra, dan A. K. Dutta, "Preface: International Conference on Recent Trends in Physics (ICRTP 2016)," *J Phys Conf Ser*, vol. 755, no. 1, 2016, doi: 10.1088/1742-6596/755/1/011001.

[9] J. N. Myhre, M. Tejedor, I. K. Launonen, A. El Fathi, dan F. Godtliebsen, "In- silico evaluation of glucose regulation using policy gradient reinforcement learning for patients with type 1 diabetes mellitus," *Applied Sciences (Switzerland)*, vol. 10, no. 18, hlm. 1–21, 2020, doi: 10.3390/APP10186350.

[10] S. I. Figure, "● --), - •••), - □ -)," no. 1.

[11] I. Markoulidakis, G Kopsiaftis, I. Rallis, dan I. Georgoulas, "Multi-class confusion matrix reduction method and its application on net promoter score classification problem," dalam *The 14th pervasive technologies related to assistive environments conference*, 2021, hlm. 412–419.

[12] V. A. Madjid, F. A. Mahdi, C. A. Lukito, D. P. A. Nofri, dan D. S. Prasvita, "Pengaruh Principal Component Analysis terhadap Akurasi Model Machine Learning dengan Algoritma Artificial Neural Network untuk Prediksi Kebangkrutan Perusahaan,"
Senamika, vol. 2, no. 2, hlm. 110–119, 2021, [Daring]. Tersedia pada: https://conference.upnvj.ac.id/index.php/senamika/article/view/1804

[13] S. Dewi dan M. A. I. Pakereng, "Implementasi Principal Component Analysis Pada K-Means Untuk Klasterisasi Tingkat Pendidikan Penduduk Kabupaten Semarang," *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 8, no. 4, hlm. 1186–1195, 2023, doi: 10.29100/jipi.v8i4.4101.

[14] F. Tempola, M. Muhammad, dan A. Khairan, "Perbandingan Klasifikasi Antara KNN dan Naive Bayes pada Penentuan Status Gunung Berapi dengan K-Fold Cross Validation," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 5, no. 5, hlm. 577– 584, 2018.

[15] B. Gavrilović dan J. Mitrović, "Comparative analysis of the traffic accidents in the territory of the city of Užice for 2021 and 2022 using open data and the Streamlit application," *Vojnotehnicki glasnik*, vol. 71, no. 3, hlm. 616–633, 2023, doi: 10.5937/vojtehg71-44007.

[16] S. Supiyandi, M. Zen, C. Rizal, dan M. Eka, "Perancangan Sistem Informasi Desa Tomuan Holbung Menggunakan Metode Waterfall," *JURIKOM (Jurnal Riset Komputer)*, vol. 9, no. 2, hlm. 274, 2022, doi: 10.30865/jurikom.v9i2.3986.