# Manuscript

*by* Qr Indonesia

**Submission date:** 30-Oct-2025 12:13PM (UTC+0700)
**Submission ID:** 2562415501
**File name:** Manuscript.docx (3.61M)
**Word count:** 4532
**Character count:** 25937

# Performance Evaluation of Machine Learning Algorithms for Supply Chain Data Classification

*Note: Sub-titles are not captured in Xplore and should not be used

Maniah
*Department of Magister Logistic Management*
*Universitas Logistik dan Bisnis Internasional*
Bandung, Indonesia
maniah@ulbi.ac.id

***Abstract*—Forecasting systems which are data-driven are of great importance in streamlining industrial and business processes during the digital transformation age. Supply chain management (SCM) is among the most significant processes to enhance the efficiency of operations and support strategic decision-making. This study seeks to evaluate the performance of two machine learning based classification algorithms, namely Naive Bayes and the k-Nearest Neighbours (K-NN) algorithm, using data in the supply chain. Some of the most valuable operational attributes, including payment method, customer segment, status of shipment, profit per transaction and customer location, are stored in the database. The data were first cleaned and then transformed into normalized and label-encoded format, after which they were divided into training and testing sets with a ratio of 80:20. The performance of the two algorithms was assessed using accuracy, precision, recall and F1-score. The findings of the research indicate that Naive Bayes is the most promising algorithm, its accuracy and precision are 99.75% and its recall rate is close to 100% in the majority of the classes. These findings show that Naive Bayes is a probabilistic algorithm which better fits the data distribution than a distance-based K-NN algorithm.**

***Keywords—data classification, K-Nearest Neighbours, machine learning, Naïve Bayes, predictive analytics, supply chain.***

## I. INTRODUCTION

During the digital transformation and the growing complexity of the global environment, the level of supply chain management (SCM) has emerged as one of the strategic elements influencing the sustainability of operations of a company [1], [2]. SCM no longer deals mainly with tangible processes like acquisition of raw materials, manufacturing and distribution, but it is much dependent on the capabilities of data processing and analysis to make fast and correct and efficient decisions [3], [4]. A study of 281 actors of MSMEs in Yogyakarta revealed that operational digitalization positively influences sustainable business performance [5]. This validates the relevance of analytical technology and machine learning in aiding data-driven decision-making.

The classification of the data is one of the most important elements that must be taken into consideration when carrying out machine learning in SCM [6]. The classification process allows the companies to group heterogeneous information of a complex nature, including the type of customers, the pattern of demand of the product, potential risks of decline in the delivery time, and even predicting a shift in the market trend [7], [8]. When correctly classified, the strategic decision making will be improved, as well as the operation itself, the number of risks, and the profit per transaction [9]. Technology-based forecasting can serve to minimise forecasting error by as much as 50% [10] and maintenance cost by as much as 30% [11].

A number of global trends in the past five years have added more urgency to the need of having a data-driven approach towards SCM. The COVID-19 crisis revealed how vulnerable the typical global supply chains were because numerous companies were faced with shipping delays, raw material shortages, and demand transformations [12]. More than 90 percent of manufacturing enterprises all over the world faced the disruption of their supply chains during the pandemic [13]. The condition is compelling organisations to establish supply chains that are more resilient, flexible, and sensitive to products of the market. Also, the digitalisation wave is speeding up the use of innovative technologies, including the Internet of Things (IoT), big data, and artificial intelligence (AI) [14]. According to International Data Corporation (IDC), 60 percent of large companies will have AI and IoT integrated into their supply chains by 2026 [15], which will cause a significant growth in the volume of logistics data and will require more innovative ways to process it.

In the industrial sector, rising customer expectations for fast, transparent, and personalised service are further emphasising the importance of data classification [16]. By gaining a better understanding of customer characteristics, companies can tailor distribution strategies while reducing resource waste. However, most previous research still focuses on the technical aspects of classification algorithms [17], [18], [19], [20], [21], with little connection to strategic implications in SCM [22]. Research on the impact of data classification on operational efficiency, supply chain resilience, and digital transformation in the industrial sector remains limited, creating a research gap.

Two popular algorithms in data classification are Naïve Bayes and k-Nearest Neighbors (K-NN). Both are known for their ease of implementation and good performance on various types of data [23]. Naïve Bayes operates on a simple probabilistic principle, assuming feature independence, while K-NN performs classification based on the proximity between data points. These two algorithms have been widely applied in healthcare, text, and pattern recognition, but their application and comparison in the context of SCM are still limited.

Based on the description, this study aims to analyse and compare the performance of the Naïve Bayes and K-NN


ISSN: XXXX-XXXX; DOI:XXXX

algorithms in the context of SCM. The research focuses on measuring the accuracy of SCM data classification using both algorithms, as well as determining the most effective method to support strategic decision-making. Through this study, the research is expected to contribute both academically and practically, enriching the literature on machine learning-based classification and supporting more innovative, more efficient, and adaptive SCM practices in the digital age.

## II. RESEARCH METHOD

### A. Research Design

This research is comparative quantitative in character and is founded on the experiments of data classification. This study analyzes using the Naive Bayes and the k-Nearest Neighbors (K-NN) algorithms in the supply chain management (SCM) data. Public datasets were used to test the model algorithm and the results were evaluated in terms of accuracy, precision, recall and F1 score [9]. Quantitative approach has been selected due to the fact that it gives objective and testable results. In the meantime, the comparative approach will be required to decide which algorithm is good at classifying SCM data. Figure 1 below presents the flow of the research.

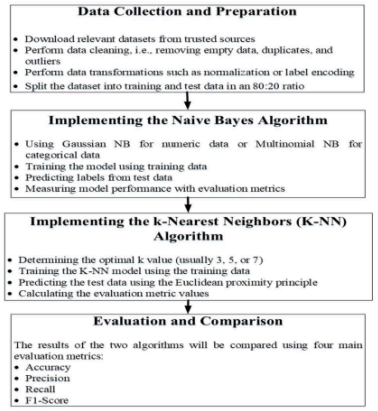



Fig. 1. Research Flow

### B. Data Collection Stage

The data utilized is a secondary data in the form of structured datasets pertaining to supply chain activities. Such variables as the type of product, shipping (on-time and late), customer demand, warehouse placement, supply chain risk (low, medium, high) and delivery time are included in this data. The data has been taken in open repositories, including Kaggle, the UCI Machine Learning Repository, or OpenML, which has been popularly used in earlier studies [10].

The data collection and preparation stage involves the fact of collecting datasets based on the trustworthy sources and data cleaning in order to eliminate all the cases of empty, duplicate, and outlier data [11]. Further data manipulations like normalisation and label encoding were also done to ready the data to be processed by the algorithm. The dataset was then separated into training and testing data in an 80 and 20, respectively, in order to make the evaluation of the models fair [12].

### C. Data Analysis Implementation Phase

The application step of the Naive Bayes Algorithm is done by having a Gaussian Naive Bayes [24], which applies to physical data (numerical) or Multinomial Naive Bayes, which applies to categorical data. The algorithm is trained with training data, then applied to make predictions on the test data. The outcomes of the prediction are subsequently checked against certain metrics to establish the functionality of the model. Meanwhile, the implementation stage of the k-Nearest Neighbors (k-NN) Algorithm starts by identifying which value is the most appropriate as k (usually 3, 5, or 7) [25]. Training data is then used to train k-NN, which then predicts the test data based on Euclidean distance theory. The evaluation metric values are then determined to judge the performance of the k-NN algorithm [26].

### D. Evaluation and Comparison Stage

It is the most important component of the research because it will compare the two algorithms regarding various four evaluation measures. The most relevant metrics that are important to consider include Accuracy, Precision, Recall, and F1-score. This comment provides an impression of the most suitable and optimal algorithm to use in the case of supply chain management data classification [27].

## III. RESULTS AND DISCUSSION

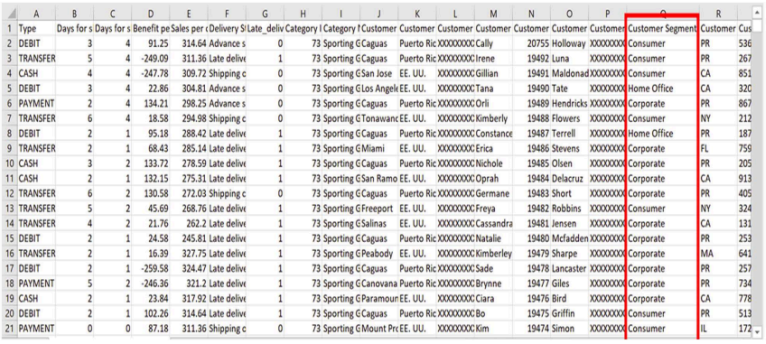
### A. Data Collection

The initial source of data collection in this study was the identification of credible and relevant sources of datasets to the context of supply chain management. The data was retrieved in Kaggle and in UCI machine learning repository and the overall data was about 500 data points including the needed variables like customer segment, payment method, order and delivery date, delivery status, profit per sale and customer location. The data were chosen according to some criteria such as the completeness of features, an acceptable sample size, and the usage license that allows applying it to research. All the details connected with the source, file format, and the date of download are presented clearly in order to ensure authenticity and reproducibility.

The second process is to check and authenticate the data to determine the quality and consistency of data. This is done by getting rid of missing values, duplicate data, and identifying outliers which may influence model accuracy. Also, the target variable of Customer Segment was also checked on class distribution to predict the possible class imbalances. In case there exist important disparities, such sampling techniques as stratified sampling or oversampling can be taken into account. In the case of categorical data, e.g. methods of payment and shipping status, label encoding was done. In the case of numerical data, such as the profit and lead time, normalisation was employed to allow the scale to be consistent.

Once the data had been cleaned other useful features were derived including lead time (the gap between the order date and the delivery date), day of the week, month and even patterns of the seasons that would affect customer behaviour. Location data are also made to be adjusted to regional groupings to simplify the number of categorical variables that may have been too many. This data was then separated into two, in the ratio of 80: 20, i.e., training data to be used in

constructing the model, and test data to be used to quantify the performance of the model on new data. The whole process of data collection and preparation was adequately recorded in terms of versioning datasets as well as the storage of raw files and preprocessed result files in different locations contributing to the possibility of replicating the experiment and making the research findings answerable.



| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Type | Days for s | Days for s | Benefit pe | Sales per ( | Delivery St | Late_deliv | Category I | Category I | Customer | Customer | Customer | Customer | Customer | Customer | Customer | Customer Segment | Customer Cus |
| 2 | DEBIT | 3 | 4 | 91.25 | 314.64 | Advance s | 0 | 73 | Sporting G | Caguas | Puerto Ric | XXXXXXXX | Cally | 20755 | Holloway | XXXXXXXX | Consumer | PR 536 |
| 3 | TRANSFER | 5 | 4 | -249.09 | 311.36 | Late deliv | 1 | 73 | Sporting G | Caguas | Puerto Ric | XXXXXXXX | Irene | 19492 | Luna | XXXXXXXX | Consumer | PR 267 |
| 4 | CASH | 4 | 4 | -247.78 | 309.72 | Shipping c | 0 | 73 | Sporting G | San Jose | EE. UU. | XXXXXXXX | Gillian | 19491 | Maldonad | XXXXXXXX | Consumer | CA 851 |
| 5 | DEBIT | 3 | 4 | 22.86 | 304.81 | Advance s | 0 | 73 | Sporting G | Los Angele | EE. UU. | XXXXXXXX | Tana | 19490 | Tate | XXXXXXXX | Home Office | CA 320 |
| 6 | PAYMENT | 2 | 4 | 134.21 | 298.25 | Advance s | 0 | 73 | Sporting G | Caguas | Puerto Ric | XXXXXXXX | Orli | 19489 | Hendricks | XXXXXXXX | Corporate | PR 867 |
| 7 | TRANSFER | 6 | 4 | 18.58 | 294.98 | Shipping c | 0 | 73 | Sporting G | Tonawanc | EE. UU. | XXXXXXXX | Kimberly | 19488 | Flowers | XXXXXXXX | Consumer | NY 212 |
| 8 | DEBIT | 2 | 1 | 95.18 | 288.42 | Late deliv | 1 | 73 | Sporting G | Caguas | Puerto Ric | XXXXXXXX | Constance | 19487 | Terrell | XXXXXXXX | Home Office | PR 187 |
| 9 | TRANSFER | 2 | 1 | 68.43 | 285.14 | Late deliv | 1 | 73 | Sporting G | Miami | EE. UU. | XXXXXXXX | Erica | 19486 | Stevens | XXXXXXXX | Corporate | FL 759 |
| 10 | CASH | 3 | 2 | 133.72 | 278.59 | Late deliv | 1 | 73 | Sporting G | Caguas | Puerto Ric | XXXXXXXX | Nichole | 19485 | Olsen | XXXXXXXX | Corporate | PR 205 |
| 11 | CASH | 2 | 1 | 132.15 | 275.31 | Late deliv | 1 | 73 | Sporting G | San Ramo | EE. UU. | XXXXXXXX | Oprah | 19484 | Delacruz | XXXXXXXX | Corporate | CA 913 |
| 12 | TRANSFER | 6 | 2 | 130.58 | 272.03 | Shipping c | 0 | 73 | Sporting G | Caguas | Puerto Ric | XXXXXXXX | Germane | 19483 | Short | XXXXXXXX | Corporate | PR 405 |
| 13 | TRANSFER | 5 | 2 | 45.69 | 268.76 | Late deliv | 1 | 73 | Sporting G | Freeport | EE. UU. | XXXXXXXX | Freya | 19482 | Robbins | XXXXXXXX | Consumer | NY 324 |
| 14 | TRANSFER | 4 | 2 | 21.76 | 262.2 | Late deliv | 1 | 73 | Sporting G | Salinas | EE. UU. | XXXXXXXX | Cassandra | 19481 | Jensen | XXXXXXXX | Corporate | CA 131 |
| 15 | DEBIT | 2 | 1 | 24.58 | 245.81 | Late deliv | 1 | 73 | Sporting G | Caguas | Puerto Ric | XXXXXXXX | Natalie | 19480 | Mcfadden | XXXXXXXX | Corporate | PR 253 |
| 16 | TRANSFER | 2 | 1 | 16.39 | 327.75 | Late deliv | 1 | 73 | Sporting G | Peabody | EE. UU. | XXXXXXXX | Kimberley | 19479 | Sharpe | XXXXXXXX | Corporate | MA 641 |
| 17 | DEBIT | 2 | 1 | -259.58 | 324.47 | Late deliv | 1 | 73 | Sporting G | Caguas | Puerto Ric | XXXXXXXX | Sade | 19478 | Lancaster | XXXXXXXX | Corporate | PR 257 |
| 18 | PAYMENT | 5 | 2 | -246.36 | 321.2 | Late deliv | 1 | 73 | Sporting G | Canovana | Puerto Ric | XXXXXXXX | Brynne | 19477 | Giles | XXXXXXXX | Corporate | PR 734 |
| 19 | CASH | 2 | 1 | 23.84 | 317.92 | Late deliv | 1 | 73 | Sporting G | Paramour | EE. UU. | XXXXXXXX | Ciara | 19476 | Bird | XXXXXXXX | Corporate | CA 778 |
| 20 | DEBIT | 2 | 1 | 102.26 | 314.64 | Late deliv | 1 | 73 | Sporting G | Caguas | Puerto Ric | XXXXXXXX | Bo | 19475 | Griffin | XXXXXXXX | Consumer | PR 513 |
| 21 | PAYMENT | 0 | 0 | 87.18 | 311.36 | Shipping c | 0 | 73 | Sporting G | Mount Pro | EE. UU. | XXXXXXXX | Kim | 19474 | Simon | XXXXXXXX | Consumer | IL 172 |

Fig. 2. Supply Chain Dataset

| Row No. | Customer Segment | Type | Days for shi... | Days for shi... | Benefit per ... | Sales per cu... | Delivery Stat... | Lat |
|---|---|---|---|---|---|---|---|---|
| 1 | Consumer | DEBIT | 3 | 4 | 91.250 | 314.640 | Advance ship... | 0 |
| 2 | Consumer | TRANSFER | 5 | 4 | -249.090 | 311.360 | Late delivery | 1 |
| 3 | Consumer | CASH | 4 | 4 | -247.780 | 309.720 | Shipping on ti... | 0 |
| 4 | Home Office | DEBIT | 3 | 4 | 22.860 | 304.810 | Advance ship... | 0 |
| 5 | Corporate | PAYMENT | 2 | 4 | 134.210 | 298.250 | Advance ship... | 0 |
| 6 | Consumer | TRANSFER | 6 | 4 | 18.580 | 294.980 | Shipping can... | 0 |
| 7 | Home Office | DEBIT | 2 | 1 | 95.180 | 288.420 | Late delivery | 1 |
| 8 | Corporate | TRANSFER | 2 | 1 | 68.430 | 285.140 | Late delivery | 1 |
| 9 | Corporate | CASH | 3 | 2 | 133.720 | 278.590 | Late delivery | 1 |
| 10 | Corporate | CASH | 2 | 1 | 132.150 | 275.310 | Late delivery | 1 |
| 11 | Corporate | TRANSFER | 6 | 2 | 130.580 | 272.030 | Shipping can... | 0 |

ExampleSet (180,519 examples, 1 special attribute, 52 regular attributes)

Fig. 3. Dataset Training

| Row No. | Customer S... | Type | Days for shi... | Days for shi... | Benefit per ... | Sales per cu... | Delivery Stat... | Late_deli |
|---|---|---|---|---|---|---|---|---|
| 1 | Consumer | DEBIT | 3 | 4 | 91.250 | 314.640 | Advance ship... | 0 |
| 2 | Consumer | TRANSFER | 5 | 4 | -249.090 | 311.360 | Late delivery | 1 |
| 3 | Consumer | CASH | 4 | 4 | -247.780 | 309.720 | Shipping on ti... | 0 |
| 4 | Home Office | DEBIT | 3 | 4 | 22.860 | 304.810 | Advance ship... | 0 |
| 5 | Corporate | PAYMENT | 2 | 4 | 134.210 | 298.250 | Advance ship... | 0 |
| 6 | Consumer | TRANSFER | 6 | 4 | 18.580 | 294.980 | Shipping can... | 0 |
| 7 | Home Office | DEBIT | 2 | 1 | 95.180 | 288.420 | Late delivery | 1 |
| 8 | Corporate | TRANSFER | 2 | 1 | 68.430 | 285.140 | Late delivery | 1 |
| 9 | Corporate | CASH | 3 | 2 | 133.720 | 278.590 | Late delivery | 1 |
| 10 | Corporate | CASH | 2 | 1 | 132.150 | 275.310 | Late delivery | 1 |

ExampleSet (500 examples, 1 special attribute, 52 regular attributes)

Fig. 4. Dataset Training

Figure 1. Dataset Training

*B. Implementation of the Naïve Bayes Algorithm and K-Nearest Neighbors (K-NN)*

In this study, two classification algorithms, Naïve Bayes and k-Nearest Neighbors (K-NN), were applied to analyze the classification performance of supply chain data. Both algorithms were chosen because they are equally popular in data processing, but have different approaches and characteristics. Naive Bayes relies on probabilistic theory


ISSN: XXXX-XXXX; DOI:XXXX

with the assumption of feature independence, while K-NN is based on the principle of proximity between data. This fundamental difference allows the research to gain a broader perspective on the effectiveness of classification models under various supply chain dataset conditions, which tend to be complex and heterogeneous.

Naive Bayes Algorithm functionality relies on calculating the likelihood of a piece of data falling under a particular category based on the distribution of data used in training. The first step was to choose the type of algorithm, which depends on the nature of the data, i.e., Gaussian Naive Bayes to work with numbers (profit and delivery time) and Multinomial Naive Bayes to work with categorical variables (payment method, delivery status, etc.). Having undergone such training data, the model produces prediction probabilities over training data and ultimately determines the one that has the greatest probability of being the most appropriate. The key strengths of this algorithm are simplicity, speed of computation, and low memory usage. Naive Bayes is highly effective this way and can be adopted on data sets that contain a multitude of samples and a sufficient amount of features. The assumption of feature independence is not always correct, but nevertheless, this algorithm can provide good and consistent results.

On the other hand, k-Nearest Neighbors (K-NN) is a non-parametric algorithm that does not create a model but instead simply stores all the training data and predicts by determining how similar the test data is to the training data. The first step involves finding the best value of k, say k=3,5, or 7, to establish the number of nearest neighbors used in the classification algorithm. Euclidean Distance is typically used as the method of measuring the distance between data points, but other measures can be applied, such as Manhattan and Minkowski, based on the situation at hand. The measured test data are then assigned the majority category of nearest neighbours. Its weakness lies in the fact that K-NN is inadequate when dealing with data that has non-linear distribution patterns, whereas its advantage lies in the fact that large datasets, like the predicted item, may demand high computational costs.

To evaluate the performance of both algorithms, this study uses the metrics of accuracy, precision, recall, and F1-score [28]. Accuracy measures the proportion of correct predictions out of all the data, precision assesses the accuracy of predictions for a specific class, recall indicates the model's ability to capture all positive data within a class, while the F1-score provides a balance between precision and recall. The evaluation results show that Naïve Bayes tends to provide stable performance on datasets that align with the assumptions of probabilistic distribution, with a very high accuracy rate. Conversely, K-NN is more competitive in data conditions with non-linear patterns and high complexity, although it takes longer to make predictions.

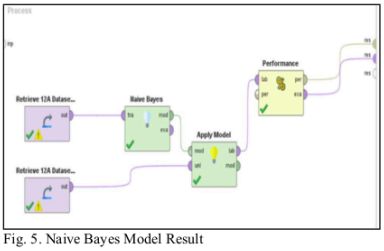



Fig. 5. Naive Bayes Model Result

Figure 5 shows the process of implementation of Naive Bayes Model used in the study. The datset of supply chain (which is acquired in the UCI dataset) is first retrieved and then passed to the Naive Bayes algorithm to train the model. The next step after model building is model application which is the model used to test the data. Performance module was used to measure the classification results on the basis of which quantitative value metrics such as accuracy, precision, recall, and F1 score were found. The workflow indicates that the data that can be used in the direct Naive Bayes method is processed with high efficiency without having complex additional parameters and is simplified further to be implemented [14], [15].

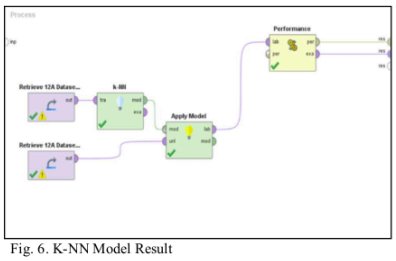



Fig. 6. K-NN Model Result

The process of the k -Nearest Neighbors (K-NN) model is presented in Figure 6. The first steps are identical to Naive Bayes since the first one is to retrieve the data (Retrieve UCI Dataset) and then process the data in the K-NN algorithm. It is at this point that the issue of establishing the value of the k parameter is important because it will influence the quality of the classification findings. This is followed by the Apply Model process of the test data to obtained class predictions on the basis of similarity with the training data. The Performance module is used to carry out model performance evaluation in which the results of the evaluation metrics are shown in detail. The K-NN algorithm has more computational requirements when making predictions as compared to the Naive Bayes because the threshold distance of each test data set needs to be calculated against all the training data [16], [17].

Overall, both images show similar classification modelling stages, starting from dataset acquisition, processing with the selected algorithm, applying the model to test data, and finally evaluating performance using performance metrics. The difference lies in the way the core algorithms used work, where Naïve Bayes relies on a

probabilistic approach, while K-NN relies on a distance (similarity-based) approach. Through this visualisation, it can be understood that although the modelling process seems simple, the choice of the correct algorithm significantly affects the effectiveness of classification results in the context of supply chain management.

TABLE 1. PERFORMANCE ACCURACY PER ALGORITHM

| | Naïve Bayes | | | | k-Nearest Neighbours (K-NN) | | | |
|---|---|---|---|---|---|---|---|---|
| | **true Customer** | **true Home Office** | **true Corporate** | **class precision** | **true Customer** | **true Home Office** | **true Corporate** | **class precision** |
| Accuracy | | 99,75% | | | | 73,75% | | |
| pred Customer | 211 | 0 | 0 | 100,00% | 171 | 16 | 31 | 78,44% |
| pred Home Office | 0 | 54 | 0 | 100,00% | 12 | 31 | 10 | 58,49% |
| pred Corporate | 0 | 1 | 134 | 96,26% | 28 | 8 | 93 | 72,09% |
| class recall | 100,00% | 98,18% | 100,00% | | 81,04% | 56,36% | 69,40% | |

Table 1 shows the outcome of the performance analysis of the Naive Bayes and the k-Nearest Neighbors (K-NN) algorithms to classify three target classes in the supply chain data, i.e., Customer, Home Office, and Corporate. The performance is measured in terms of total accuracy, the class precision and class recall which is obtained after calculating the confusion matrix of each model. The analysis findings indicate that Naive Bayes is much better in terms of the accuracy of 99.75, which is significantly more than that of K-NN (73.75). There were 211 Customer data, 54 Home Office data and 134 Corporate data that were classified with a great success rate and very high accuracy in the Naive Bayes model. The mistake was made in only one point of the Corporate class, which was falsely categorised as Home Office. Class precision Customer and Home Office were 100 and Corporate 96.26. The recall value was also quite high, Customer and Corporate had a 100% and Home Office had a 98.18% recall value. It implies that the Naive Bayes model is able to identify nearly all the data and reduce the misclassification errors.

In its turn, the K-NN algorithm performed with lower satisfaction. Of 211 customer data objectives, 171 were rightly categorized and the rest were wrongly categorized as Home Office (16) and Corporate (31). This was also done with the Home Office and Corporate classes where K-NN tended to commit cross classification errors. The best K-NN precision was recorded with Customer class (78.44%), whereas the worst precision was with Home Office (58.49%). The values of the recall were low, 81.04 Customer, 56.36 Home Office, and 69.40 Corporate. This means that K-NN cannot easily differentiate patterns across classes especially when the classes have data closely distributed together.

From this comparison, it can be concluded that the characteristics of the supply chain dataset used are more suitable for the feature independence assumption in Naïve Bayes, resulting in more accurate, consistent, and efficient predictions. Conversely, the performance of K-NN is influenced by the proximity between data and the selection of the parameter k, making it less optimal for datasets with complex variations. Therefore, for the case of customer segment classification in the supply chain, Naïve Bayes proved to be more effective than K-NN in supporting data analysis and decision-making based on historical data.

Based on the findings, it should be stated that the use of Naive Bayes, in its case, is rather recommended for the data classification system within a supply chain management when the use of this algorithm is necessary to compete with customers or determine the delivery status. It has a high potential to be used within data-driven systems because it provides quick, precise, and effective models; thus, it is highly relevant in this domain because it can deliver real-time prediction results. In the meantime, it is possible that K-NN can also be applied as an alternative, but to enhance the result, one should adjust parameters (size of k) and distance minimization to improve the work of K-NN. The implications of the findings of this research paper are as follows: machine learning algorithms ought to be applied in the current state of supply chains.

Considering that, in addition to the automatization of decision-making processes, the application of classification models also contributes to digital transformation and increasing resilience to external disruptions, the supply chains become more resilient. This study can be improved in the future by using more and more diverse algorithms, including Decision Trees, Random Forests, or Support Vector Machines (SVMs), and also by analyzing much larger datasets. Other technologies that can be integrated to develop more intelligent, more transparent, and more adaptable to changes in market supply chain systems can be generated, including Internet of Things (IoT) and blockchain.

## IV. CONCLUSION

According to the results of the conducted research, one can conclude that the Naive Bayes algorithm outperforms k-Nearest Neighbors (K-NN) when it comes to customer segmentation as part of the supply chain management. Naive Bayes was able to attain a 99.75% accuracy and high precision and recall on virtually every class, while K-NN was able to attain only 73.75% accuracy with relatively low rates of precision and recall. It implies that the traits of the adopted dataset are conducive to the probability-heavy assumptions of Naive Bayes over the vicinity-based K-NN methodology. Thus, Naive Bayes is more precise, efficient, and effective in aiding the data analysis and decision-making in the supply chain systems.

One might offer that to compare the two approaches further, the additional algorithms such as Decision Tree, random forest, or Support Vector machine (SVM) can be


ISSN: XXXX-XXXX; DOI:XXXX

added and researched. In addition, to increase the validity of the research, it would be more appropriate to test them using broader and more diversified data sources induced by real-life settings. This is also achieved by optimization of K-NN over its parameters, i.e. computation of k-value of data, distance measure and alternative measures of distances etc. In addition, they can create technology smarter, more transparent, and adaptable to market dynamics supply chain systems when used together with other technology platforms, such as the Internet of Things (IoT) and blockchain, to contribute to firms becoming more competitive in the digital world and more resilt against market fluctuations.

## REFERENCES


[1] A. Belhadi, S. Kamble, A. Gunasekaran, and V. Mani, "Analyzing the Mediating Role of Organizational Ambidexterity and Digital Business Transformation on Industry 4.0 Capabilities and Sustainable Supply Chain Performance," *Supply Chain Management: An International Journal*, vol. 27, no. 6, pp. 696–711, Nov. 2022, doi: 10.1108/SCM-04-2021-0152.

[2] L. V. Lerman, G. B. Benitez, J. M. Müller, P. R. de Sousa, and A. G. Frank, "Smart Green Supply Chain Management: A Configurational Approach to Enhance Green Performance through Digital Transformation," *Supply Chain Management: An International Journal*, vol. 27, no. 7, pp. 147–176, Dec. 2022, doi: 10.1108/SCM-02-2022-0059.

[3] D. H. Semnasti, "Perencanaan Supply Chain Management pada Seneca Coffe Studio," in *WALUYO JATMIKO PROCEEDING*, Feb. 2025, pp. 1–7. doi: 10.33005/wj.v17i1.107.

[4] G. Wilson, O. Johnson, and W. Brown, "The Role of Machine Learning in Predictive Analytics for Supply Chain Management," Aug. 05, 2024. doi: 10.20944/preprints202408.0343.v1.

[5] B. Susanto, "Pengaruh Digitalisasi Operasional dan Integrasi Rantai Pasokan (SCI) terhadap Kinerja Berkelanjutan pada UMKM di Yogyakarta," Universitas Islam Indonesia, Yogyakarta, 2025. Accessed: Sep. 19, 2025. [Online]. Available: https://dspace.uii.ac.id/bitstream/handle/123456789/55902/21311577.pdf?sequence=1&isAllowed=y

[6] D. S. Manik, Nazaruddin, and N. Panjaitan, "Literatur Review: Penggunaan Algoritma Machine Learning untuk Optimalisasi Rantai Pasokan," in *Talenta Conference Series: Energy and Engineering (EE)*, 2025, pp. 1038–1047. doi: https://doi.org/10.32734/ee.v8i1.2671.

[7] Hartatik *et al.*, *Data Science for Business: Pengantar & Penerapan Berbagai Sektor*. Jambi: PT Sonpedia Publishing Indonesia, 2023.

[8] A. Asrul *et al.*, "Pemanfaatan Big Data Analytics dalam Proses Manajemen Teknologi untuk Prediksi Permintaan Pasar," *Jurnal Minfo Polgan*, vol. 13, no. 2, pp. 2433–2438, Jan. 2025, doi: 10.33395/jmp.v13i2.14516.

[9] A. A. Hapsari, *Manajemen Risiko Keuangan : Strategi Proteksi dan Pengambilan Keputusan*. Padang: Takaza Innovatix Labs, 2025.

[10] L. Columbus, "10 Ways Enterprises Are Getting Results From AI Strategies," Forbes. Accessed: Sep. 19, 2025. [Online]. Available: https://www.forbes.com/sites/louiscolumbus/2020/06/04/10-ways-enterprises-are-getting-results-from-ai-strategies/?sh=3cc144776fdb

[11] M. Mulders and M. Haarman, "Predictive Maintenance 4.0 Predict the unpredictable," KJ Dordrecht, 2017.

[12] D. Sari, N. Harahap, Y. Sasmita, S. A. Sitepu, and Arsyadona, "Manajemen Risiko dalam Rantai Pasok Global pada Masa Pandemi: Studi Kasus Ekspor Produk UMKM di Indonesia," *Jurnal Rumpun Manajemen dan Ekonomi*, vol. 2, no. 4, pp. 115–123, 2025, doi: https://doi.org/10.61722/jrme.v2i4.5331.

[13] G. Ambrogio, L. Filice, F. Longo, and A. Padovano, "Workforce and supply chain disruption as a digital and technological innovation opportunity for resilient manufacturing systems in the COVID-19 pandemic," *Comput Ind Eng*, vol. 169, p. 108158, Jul. 2022, doi: 10.1016/j.cie.2022.108158.

[14] S. M. Hawari, Semin, and S. Adiyono, *Transformasi Digital pada Manajemen Pelabuhan*. Pekalongan: Penerbit NEM, 2024.

[15] IDC, "IDC Predicts: By 2026, 60% of Asia/Pacific Enterprises Will Require Professional Services for Readiness and Health Checks on Applications and Infrastructure," Needham, 2025. Accessed: Sep. 19, 2025. [Online]. Available: https://my.idc.com/getdoc.jsp?containerId=prAP53268825

[16] E. Nurmiati and A. Nashikha, "OPTIMALISASI E-CRM PADA STARTUP DIGITAL UNTUK MENINGKATKAN RETENSI PELANGGAN: SYSTEMATIC LITERATURE REVIEW," *JURNAL PERANGKAT LUNAK*, vol. 7, no. 2, pp. 133–143, Jun. 2025, doi: 10.32520/jupel.v7i2.4126.

[17] D. Nasien *et al.*, "Perbandingan Implementasi Machine Learning Menggunakan Metode KNN, Naive Bayes, dan Logistik Regression Untuk Mengklasifikasi Penyakit Diabetes," *JEKIN - Jurnal Teknik Informatika*, vol. 4, no. 1, pp. 10–17, Feb. 2024, doi: 10.58794/jekin.v4i1.640.

[18] S. Sahar, "Analisis Perbandingan Metode K-Nearest Neighbor dan Naïve Bayes Clasiffier Pada Dataset Penyakit Jantung," *Indonesian Journal of Data and Science*, vol. 1, no. 3, Dec. 2020, doi: 10.33096/ijodas.v1i3.20.

[19] P. D. Rinanda, B. Delvika, S. Nurhidayarnis, N. Abror, and A. Hidayat, "Perbandingan Klasifikasi Antara Naive Bayes dan K-Nearest Neighbor Terhadap Resiko Diabetes pada Ibu Hamil," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 2, no. 2, pp. 68–75, Sep. 2022, doi: 10.57152/malcom.v2i2.432.

[20] S. Baadel, K. Attai, B. Bassey, E. Attai, A. Majeed, and F.-M. Uzoka, "Comparative Analysis of Machine Learning Classifiers for Yellow Fever Diagnosis Using Causative Data: Evaluating Naïve Bayes, KNN, RIPPER, and PART," 2025, pp. 160–169. doi: 10.1007/978-3-031-92178-0_13.

[21] S. Ganesh and R. Baskar, "Comparative analysis of early detection of lung cancer through breath analysis by comparing K-nearest neighbours (KNN) with Naive Bayes," 2025, p. 020169. doi: 10.1063/5.0262284.

[22] S. Saberi, M. Kouhizadeh, J. Sarkis, and L. Shen, "Blockchain technology and its relationships to sustainable supply chain management," *Int J Prod Res*, vol. 57, no. 7, pp. 2117–2135, Apr. 2019, doi: 10.1080/00207543.2018.1533261.

[23] M. M. Queiroz, D. Ivanov, A. Dolgui, and S. F. Wamba, "Impacts of epidemic outbreaks on supply chains: mapping a research agenda amid the COVID-19 pandemic through a structured literature review," *Ann Oper Res*, vol. 319, no. 1, pp. 1159–1196, Dec. 2022, doi: 10.1007/s10479-020-03685-7.

[24] S. S. A and K. K. T. G, "Comparative Study of Naive Bayes, Gaussian Naive Bayes Classifier and Decision Tree Algorithms for Prediction of Heart Diseases," *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, vol. 9, no. 3, pp. 475–486, 2021, Accessed: Sep. 19, 2025. [Online]. Available: https://www.researchgate.net/profile/Sushma-Sa-2/publication/350525962_Comparative_Study_of_Naive_Bayes_Gaussian_Naive_Bayes_Classifier_and_Decision_Tree_Algorithms_for_Prediction_of_Heart_Diseases/links/6363dce8431b1f5300689ddc/Comparative-Study-of-Naive-Bayes-Gaussian-Naive-Bayes-Classifier-and-Decision-Tree-Algorithms-for-Prediction-of-Heart-Diseases.pdf

[25] P. Cunningham and S. J. Delany, "k-Nearest Neighbour Classifiers - A Tutorial," *ACM Comput Surv*, vol. 54, no. 6, pp. 1–25, Jul. 2022, doi: 10.1145/3459665.

[26] M. I. C. Rachmatullah, "Proposed Modification of K-Means Clustering Algorithm with Distance Calculation Based on Correlation," *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika*, vol. 8, no. 1, p. 136, May 2022, doi: 10.26555/jiteki.v8i1.23696.

[27] L. M. Sinaga, Sawaluddin, and S. Suwilo, "Analysis of classification and Naïve Bayes algorithm k-nearest neighbor in data mining," *IOP Conf Ser Mater Sci Eng*, vol. 725, no. 1, p. 012106, Jan. 2020, doi: 10.1088/1757-899X/725/1/012106.

[28] G. Naidu, T. Zuva, and E. M. Sibanda, "A Review of Evaluation Metrics in Machine Learning Algorithms," 2023, pp. 15–25. doi: 10.1007/978-3-031-35314-7_2.

We suggest that you use a text box to insert a graphic (which is ideally a 300 dpi TIFF or EPS file, with all fonts embedded) because, in an MSW document, this method is somewhat more stable than directly inserting a picture.

To have non-visible rules on your frame, use the MSWord "Format" pull-down menu, select Text Box > Colors and Lines to choose No Fill and No Line.


ISSN: XXXX-XXXX; DOI:XXXX

# Manuscript

ORIGINALITY REPORT

| 8% | 8% | % | % |
|---|---|---|---|
| SIMILARITY INDEX | INTERNET SOURCES | PUBLICATIONS | STUDENT PAPERS |

PRIMARY SOURCES

| | | |
|---|---|---|
| 1 | journal.mediadigitalpublikasi.com<br>Internet Source | 1% |
| 2 | www.mdpi.com<br>Internet Source | 1% |
| 3 | digitalcommons.usu.edu<br>Internet Source | 1% |
| 4 | ijctjournal.org<br>Internet Source | 1% |
| 5 | tunasbangsa.ac.id<br>Internet Source | 1% |
| 6 | www.saiie.co.za<br>Internet Source | <1% |
| 7 | www.icmhi.org<br>Internet Source | <1% |
| 8 | eprint.innovativepublication.org<br>Internet Source | <1% |
| 9 | journal.abhinaya.co.id<br>Internet Source | <1% |
| 10 | www.aasmr.org<br>Internet Source | <1% |
| 11 | ajomc.asianpubs.org<br>Internet Source | <1% |
| 12 | aircconline.com<br>Internet Source | <1% |
| 13 | journal.irpi.or.id<br>Internet Source | <1% |

| | | |
|---|---|---|
| 14 | jurnal.polgan.ac.id<br>Internet Source | <1 % |
| 15 | nano-ntp.com<br>Internet Source | <1 % |
| 16 | www-emerald-com-443.webvpn.sxu.edu.cn<br>Internet Source | <1 % |
| 17 | ebin.pub<br>Internet Source | <1 % |
| 18 | ejournal.unitomo.ac.id<br>Internet Source | <1 % |
| 19 | newsletter.x-mol.com<br>Internet Source | <1 % |
| 20 | www.thapar.edu<br>Internet Source | <1 % |

| | | | |
|---|---|---|---|
| Exclude quotes | On | Exclude matches | Off |
| Exclude bibliography | On | | |