IJAIDM TURNITIN

by Rachma Oktari

Submission date: 11-Feb-2022 05:50PM (UTC-0800)

Submission ID: 1760470677

File name: IJAIDM_-_Rachma_Oktari_-_turnitin.pdf (186.52K)

Word count: 2412

Character count: 11800

Optimization of the Naïve Bayes Classifier (NBC) Algorithm Using the Sparrow Search (SSA) Algorithm to Predict the Distribution of Goods Receipts

ABSTRACT

Distribution must be able to meet all needs based on sales orders from consumers, be responsible for the delivery order process running optimally, and ensure the good receipt process is in accordance with consumer sales order requests. PT. Diamond Cold Storage currently uses Enterprise Resource Planning (ERP) to record all reports from production to sales. But in reality there are still some obstacles in the distribution section. In the good receipt process, several items were found that did not match the sales order, such as: the item did not match the order request or the item did not match the order request. The process of mismatching the good receipt with the sales order will be met with the completion of the good receipt process or the bad thing is that there is a cancellation, so this causes a loss for the company. This study uses data mining techniques with the Naïve Bayes Classifier algorithm to predict the distribution of goods receipts based on distribution data, and uses the Sparrow Search Algorithm (SSA) algorithm to optimize the Nave Bayes Classifier by selecting features to improve accuracy. In this study, the results obtained that the SSA algorithm can improve the performance of NBC from 95.05% to 97.95%.

1. INTRODUCTION

The aspect of product distribution is a crucial spotlight because most of the production costs are spent by the distribution process to ordering agents. Distribution must be able to meet all needs based on sales orders from consumers, be responsible for the delivery order process running optimally, and ensure the good receipt process is in accordance with consumer sales order requests. PT. Diamond Cold Storage currently uses Enterprise Resource Planning (ERP) to record all reports from production to sales. But in reality there are still some obstacles in the distribution section. In the good receipt process, several items were found that did not match the sales order, such as: the item did not match the request order or the item did not match the request order.

The settlement process takes quite a long time, between the distribution party and the customer, while in the case of cancellation it is influenced by several factors, namely: from internal (company) and external (customer) parties. Internal factors, including: deliveries that exceed the rules or time limits from the customer, items that are rejected by the customer on the grounds that they do not match the quality desired by the customer. External factors, among others: a customer warehouse that is already full, input errors in the business to business customer system, and this causes losses for the company.

Based on the description above, this study uses data mining techniques to find the accuracy value which is used as a reference for predicting the distribution of goods receipts by looking at customer data, items, item quality, delivery distance, vehicle temperature, vehicle conditions, weather, traffic conditions, delivery time, sales orders, and delivery orders. The basic approach in data mining is to summarize data and extract useful issues that were previously unknown. Data Mining can find hidden trends and patterns that don't arise in simple query analysis and as a result can have a crucial part in finding knowledge and making decisions. Such tasks can be predictive such as classification and regression or descriptive such as clustering and association [1].

Several studies have been carried out using data mining techniques to explore various issues from a database, such as the research conducted by Erwina Nurul Azizah, et al using Web History data and the number of interactions of students' web pages with the Naïve Bayes algorithm and the C4.5 algorithm to predict academic performance. students in the Learning Environment. the origin of the data that has been processed using two algorithms. The results obtained, both algorithms have almost the same level of accuracy. Naive Bayes accuracy is superior to 63.8% of C4.5 only 0.2% different from Naive Bayes accuracy [2]. Subsequent research conducted by Meylan Wongkar, et al in analyzing sentiment with twitter data regarding the 2019 presidential candidate of the Republic of Indonesia using the python programming language. In this study, a comparison was made using NBC, SVM and K-NN methods using RapidMiner, and resulted in an NBC accuracy value of 75.58%, an SVM accuracy value of 63.99% and a K-NN accuracy value of 73.34%. NBC outperforms other methods [3].

Naïve Bayes has several advantages, namely fast in calculation, simple algorithm and high accuracy. Naïve Bayes Classifier is better applied in large data and can handle incomplete data (missing values). and can handle irrelevant attributes and noise data. However, the Naïve Bayes Classifier also has drawbacks, namely the selection of attributes that affect the accuracy value. So the Nave Bayes Classifier needs to be optimized by weighting the attributes so that the Nave Bayes Classifier can work more effectively [4].

In solving these problems, this study uses one of the meta-heuristic optimizations, namely the Sparrow Search algorithm proposed by Xue and Shen in 2020 to increase the accuracy of the Naïve Bayes Classifier [5]. Meta-heuristic optimization techniques have become very popular over the last 2 periods, such as Particle Swarm Optimization (PSO) which was first developed by Kennedy and Eberhart in 1995, Ant-based techniques were first developed by Dorigo in 1996 using Ant Colony Optimization (ACO) to complete the Traveling Salesman, Genetic Algorithm was first developed by John Holland in the 1970s, Harmony Search Algorithm (HSA) was introduced by Zong Woo Geem, Joon Hoon Kim, and GV Loganathan in 2001, Bee Algorithm was developed by Pham,

There are several reasons why meta-heuristics are so commonly used, namely: simplicity, flexibility, derivation-free procedures, and avoidance of local optima. so that many theoretical works use optimization techniques like this in various fields of learning [6]. In the Sparrow Search Algorithm (SSA) technique, there are several ways to optimize, including reducing the error rate and feature selection. The data used in this study is the distribution of goods which aims to predict the distribution of goods receipts. With the Sparrow Search Algorithm (SSA) algorithm can optimize the accuracy value of the Naïve Bayes Classifier by means of feature selection.

2. RESEARCH METHOD

The following is a schematic of the research stages regarding the prediction of the distribution of goods receipts, which can be seen in Figure 1.

2.1. Dataset Collection

The process of collecting data, the data used is distribution data originating from delivery order and good receipt data from 2020 to 2021.

.

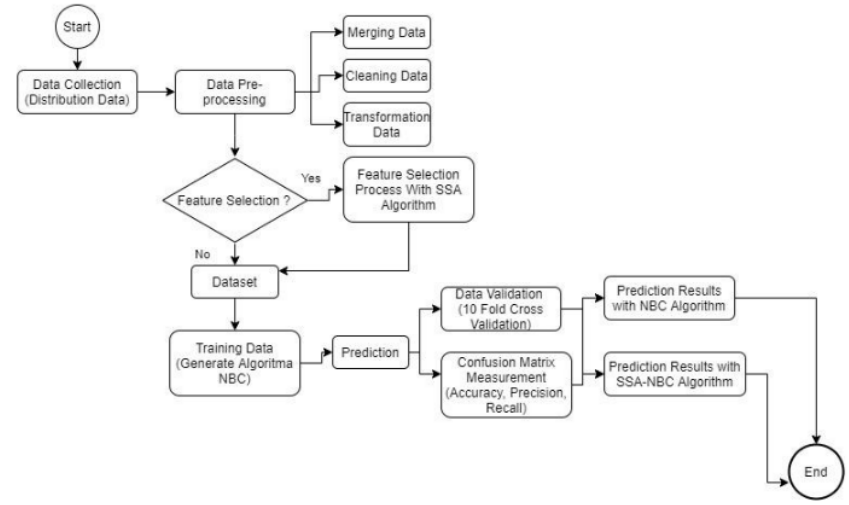
Start
Data Collection (Distribution Data)
Data Pre-processing
Merging Data
Cleaning Data
Transformation Data
Feature Selection ?
Yes
Feature Selection Process With SSA Algorithm
No
Dataset
Training Data (Generate Algoritma NBC)
Prediction
Data Validation (10 Fold Cross Validation)
Confusion Matrix Measurement (Accuracy, Precision, Recall)
Prediction Results with NBC Algorithm
Prediction Results with SSA-NBC Algorithm
End

Figure 1. Research Stage

2.2. Dataset Preprocessing

The process of converting raw data or known as raw data, such as: merging data (data from several files such as data per month of delivery orders and good can be combined into one distribution file), then discarding noise data, duplicate data, empty value data, then the transformation process data from nominal to numeric, and then perform the feature selection process with the SSA algorithm for SSA-NBC testing. The model is built using the NBC algorithm and the feature selection method using the SSA algorithm.

In this stage, the distribution data will be reprocessed using data preprocessing techniques, then the training process will be carried out using the NBC and SSA-NBC algorithms and then a comparison will be made based on the values of accuracy, precision, and recall. The training model is shown in Figure 4. After preprocessing the data, then the data is divided into training data and testing data which will be entered into the NBC model architecture.

2.3. Training Data

Data training will be carried out using the NBC and SSA algorithms. The training conducted using google colab platform. The experiment in the first step will be trained using the NBC algorithm and then using NBC-SSA to get a comparison of the results of accuracy, precision and recall values. The following is an architectural model using a comparison between the NBC and SSA-NBC algorithms in finding the best accuracy, precision and recall values from the two models presented in Figure 2 and Figure 3.

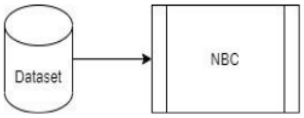

Figure 2. Model Architecture Without SSA

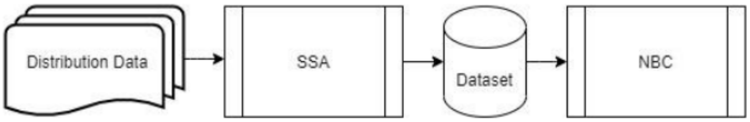

Figure 3. Model Architecture with SSA

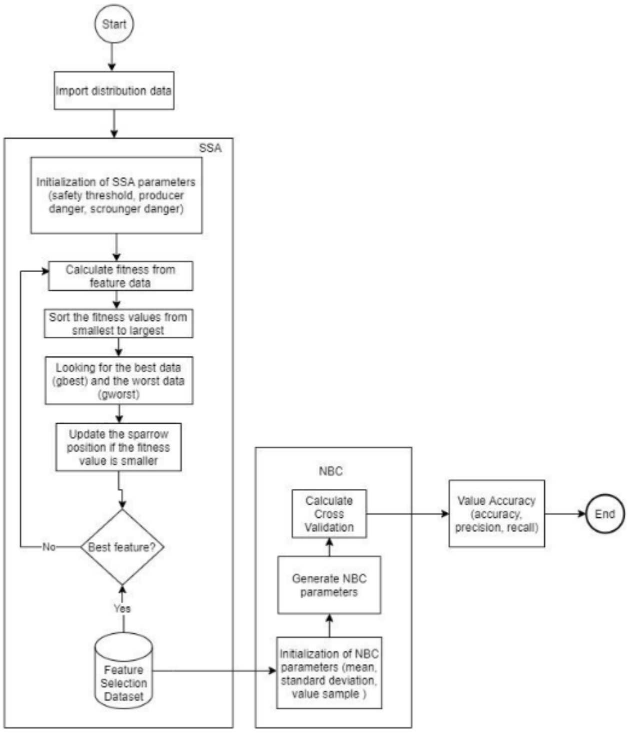
Start
Import distribution data
SSA
Initialization of SSA parameters (safety threshold, producer danger, scrounger danger)
Calculate fitness from feature data
Sort the fitness values from smallest to largest
Looking for the best data (gbest) and the worst data (gworst)
Update the sparrow position if the fitness value is smaller
Best feature?
No
Yes
Feature Selection Dataset
NBC
Initialization of NBC parameters (mean, standard deviation, value sample)
Generate NBC parameters
Calculate Cross Validation
Value Accuracy (accuracy, precision, recall)
End

Figure 4. Training Model

2.4. Feature Selection

The feature selection process in this study is to find data attributes that have an effect on improving NBC performance. The feature selection process in this study is to determine the safety threshold value, producer danger, scrounger danger, lower limit value, upper limit value and features compared to other features through a 5 epoch process, the results obtained are the accuracy results from the comparison of each feature, can be seen in Table 1. Furthermore, the selected feature is the feature that has the highest accuracy from each epoch. The features before being selected were 11 features, after going through the feature selection process, the number of features became 6. The data from the feature selection are presented in Table 2.

Table 1. Best Feature Results in Process 5 Epoch

Epoch	Maximum Accuracy	Maximum Accuracy Column	Maximum Precision	Maximum Precision Column	Maximum Recall	Maximum Recall Column
0	95.04%	[Full Data]	82.82%	[Full Data]	73.97%	[Full Data]
1	96.51%	[X3, X7, X9 and X11]	84.09%	[X3, X7, X10 and X11]	85.44%	[X3, X7, X9, X10 and X11]
2	96.51%	[X3, X7, X9 and X11, X1, X3, X5, X7, X9, X10 and X11]	85.27%	[X3, X7, X10 and X11]	85.87%	[X2, X3, X7, X9, X10 and X11]
3	96.56%	[X3, X7, X9 and X11]	84.86%	[X3, X7, X9, X10 and X11]	85.41%	[X3, X7, X9, X10 and X11]

.

4	97.00%	[X3, X5, X7, X9, X10 and X11]	85.68%	[X3, X7, X10 and X11]	85.23%	[X3, X7, X9, X10 and X11, X3, X5, X7, X9, X10 and X11]
5	96.78%	[X3, X5, X7, X9, X10 and X11]	84.95%	[X3, X7, X9, X10 and X11]	84.67%	[X2, X3, X7, X9, X10 and X11]

Table 2. Feature Selection Results

Actual string data with the highest accuracy after optimization and feature selection processing

['X3', 'X5', 'X7', 'X9', 'X10', 'X11']

	QUANTITY DO	**TEMPERATURE**	**WEATHER**	**ITEM QUALITY**	**VEHICLE**	**OTHER PROBLEMS**	**GOOD RECEIPT**
0	1.0	FROZEN	BRIGHT	GOOD	OK	THERE IS NOT ANY	FULL
1	1.0	FROZEN	BRIGHT	GOOD	OK	THERE IS NOT ANY	FULL
2	1.0	FROZEN	BRIGHT	GOOD	OK	THERE IS NOT ANY	FULL
3	1.0	FROZEN	BRIGHT	GOOD	OK	THERE IS NOT ANY	FULL
4	1.0	FROZEN	BRIGHT	GOOD	OK	THERE IS NOT ANY	FULL
...	...	...	...	...	...	...	...
7996	1.0	DRY	BRIGHT	GOOD	NOT OK	THERE IS NOT ANY	FULL
7997	1.0	DRY	BRIGHT	GOOD	NOT OK	THERE IS NOT ANY	FULL
7998	1.0	DRY	BRIGHT	GOOD	OK	THERE IS NOT ANY	FULL
7999	1.0	DRY	BRIGHT	GOOD	OK	THERE IS NOT ANY	FULL
8000 rows × 7 columns							

2.5. Validation

In 10 fold Cross Validation, the amount of data used is 8000 data, the data is divided into 10 folds of the same size, so it has 10 subsets of data, 9 folds (7200 data) for training, 1 fold (800 data) for testing . Cross Validation K-fold is used because it can reduce computational time while maintaining the accuracy of the estimate [11]. Model architecture for the use of the NBC algorithm without SSA and using NBC with SSA.

2.6. Measurement Stage

The model architecture for using the NBC algorithm without SSA and NBC using SSA can be seen in Figure 2 and Figure 3. The 10 cross validation process before and after feature selection with 3x3 confusion matrix measurements, because there are 3 good receipt targets or labels used, namely: cancel, partial, and full.

3. RESULT AND ANALYSIS

After the data training process with the NBC algorithm and the SSA-NBC algorithm, using 10 fold cross validation data validation and measurement of the confusion matrix, the following are the experimental results of the model that has been trained, which can be seen in Table 4.

The NBC model without SSA uses 11 data attributes, including customer, item, quantity do (delivery order), distance, vehicle temperature, traffic, weather, delivery time, item quality, vehicle condition, other problems, and targets. The results of the average value of 95.05% accuracy, 81.01% average precision and 74.27% recall average. For the NBC model using SSA, it uses features that have been selected into 6 features, namely quantity do, vehicle temperature, weather, item quality, vehicle condition, other problems, and targets. The results of the average value of accuracy 97.95%, the average precision 84.40% and the average recall 83.32%. The comparison graph of training results is available in Figure 6.

Table 3. Training Results

CV K-Fold	*NBC*			*SSA-NBC*		
	Accuracy	*Precision*	*Recall*	*Accuracy*	*Precision*	*Recall*
1	96,13%	81,87%	74,17%	97,63%	95,09%	78,89%
2	96,75%	83,65%	76,58%	97,75%	93,55%	77,78%
3	97,63%	84,17%	81,77%	99,88%	70,37%	94,27%
4	95,63%	84,85%	77,12%	98,75%	85,19%	74,83%
5	95,50%	88,99%	69,17%	99,00%	85,19%	83,16%
6	96,13%	86,58%	72,07%	97,10%	93,18%	68,42%
7	94,88%	85,76%	72,67%	97,88%	77,69%	66,67%
8	93,50%	82,67%	71,58%	96,38%	78,63%	93,85%
9	95,88%	80,70%	73,81%	98,75%	82,55%	97,70%
10	88,50%	60,88%	73,72%	96,38%	82,55%	97,67%
Average	**95,05%**	**82,01%**	**74,27%**	**97,95%**	**84,40%**	**83,32%**

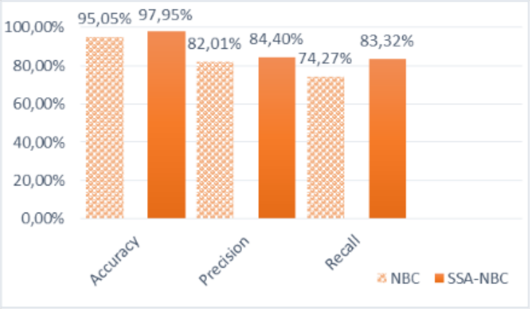

Figure 5. Comparison of Accuracy, Precision and Recall Values

Based on the research that has been done, SSA has increased the performance of NBC by 2.90%, from 95.05% to 97.95%. Based on the selection of features, SSA produces features that can increase the accuracy of the prediction of the distribution of goods received, namely: quantity do, vehicle temperature, weather, item quality, vehicle condition, other problems, and targets.

4. CONCLUSION

Based on research that has been done in predicting the distribution of goods receipts by selecting features using SSA to improve NBC performance, it can be concluded that SSA can improve NBC performance. So that the SSA-NBC method can predict the distribution of goods receipts more precisely on the distribution dataset of PT. Diamond Cold Storage in 2020 to 2021.

IJAIDM TURNITIN

ORIGINALITY REPORT

12%	4%	11%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Ikhsan Romli, Toga Pardamean, Sufajar Butsianto, Tri Ngudi Wiyatno, Effendi bin Mohamad. "Naive Bayes Algorithm Implementation Based on Particle Swarm Optimization in Analyzing the Defect Product", Journal of Physics: Conference Series, 2021 Publication	3%
2	Sultan Zayid Al-Hinai. "On the Security of the LILI Family of Stream Ciphers Against Algebraic Attacks", Lecture Notes in Computer Science, 2007 Publication	3%
3	www.researchgate.net Internet Source	2%
4	dinastipub.org Internet Source	2%
5	Erwina Nurul Azizah, Utomo Pujianto, Eki Nugraha, Darusalam. "Comparative performance between C4.5 and Naive Bayes classifiers in predicting student academic performance in a Virtual Learning	1%

	Environment", 2018 4th International Conference on Education and Technology (ICET), 2018 Publication	
6	Yannis Marinakis. "Intelligent and nature inspired optimization methods in medicine: the Pap smear cell classification problem", Expert Systems, 11/2009 Publication	1%
7	N L W S R Ginantra, I G A D Indradewi, E Hartono. "Machine learning approach for Acute Respiratory Infections (ISPA) prediction: Case study Indonesia", Journal of Physics: Conference Series, 2020 Publication	1%
8	"AI 2013: Advances in Artificial Intelligence", Springer Science and Business Media LLC, 2013 Publication	1%
9	link.springer.com Internet Source	1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		