Data Mining: The Classification Method to Predict the Types of Motorcycle Spare Parts to be Restocked

S L B Ginting¹, Y R Ginting², Sutono¹ and A Rakhman¹

¹Computer System Study Program, Faculty of Engineering and Computer Science, Universitas Komputer Indonesia, Jl. Dipati Ukur 112-116, Bandung 40132, Indonesia.
²Department of Mechanical Engineering, Universitas Riau, Kampus Bina Widya KM 12.5, Simpang Baru, Pekanbaru 28293, Indonesia.

Email: selvia.lorena@email.unikom.ac.id

Abstract. The research intends to create an application which is able to analyse sales data in a motorcycle company to predict the types of spare parts which should be stocked. This prediction is crucial since problems are often encountered while restocking. For instance, when there have been some imprecisions occurring in deciding regarding the types of spare parts to restock, the spare parts accumulate. It can cause inefficiency in terms of storage, the products quality deteriorates due to having been stored for too long, and sometimes the best-selling products are not available in the warehouse. This application is developed with Naive Bayes Classifier (NBC) method which has a high accuracy in predicting future occurrences. This method works by calculating the probability value in each attribute class and determining the optimal probability value. From the test results, 4500 training data with 200 sample test data has 90% similarity with the results of the restock decision without application. For 500 test data, the similarity was 96%. It is proven that this method has a high accuracy so that it can help the decision makers solved the company problem in predicting the types of motorcycle parts to be restocked.

1. Introduction
The increase of motorbike users in Indonesia every year is directly proportional to the demand for motorcycle parts [1]. The need for motorcycle parts is influenced by many factors. One of the most influential factors is demand when performing a motorcycle service [2], [3]. The higher the demand for spare parts is, the more stock of spare parts must be available in the company's warehouse. Nevertheless, they should not be excessive [4], [5]. In reality, however, there are problems caused by errors in the restock process, namely the accumulated parts that are not sold out in a relatively long time and on the other hand, lack of the number of certain parts needed by the customer [6]. This condition can cause the company loss since the products can be stored too long (deadstock). Deadstock can result in reduced product quality, reduced selling value and efficiency in terms of the use of storage space in the warehouse [7], [8]. Based on these conditions, it requires an application that is able to provide decisions automatically and accurately in predicting the restock of the types of motorcycle parts, using data and information that already exist in the company.

One of the data mining methods will be applied to build this application, namely the Naive Bayes Classifier (NBC) method [9]. NBC in this application will be easier to analyze large-scale data compared to manual analysis. This method can also simplify the analysis processes [10]. The NBC method will be carried out by utilizing motorcycle parts sales data (company historical data) which has thousands of records, analyze and then automatically predict the types of spare parts that will be restocked in the
future. This method will produce two final decisions, namely restock and not restock, so that with
the help of this output, decision making can be performed with less difficulty by the one responsible. When
the decision generated by the application is a restock for certain types of motorcycle parts, the number
of spare parts of the required type will be increased so that they can meet the demand. Conversely, if the
decision is not restocked, then certain types of parts will not be added.

A number of researchers from various fields have succeeded in applying this NBC method. Meir et al., applying the NBC method in processing high-throughput screening (HTS) data to determine
the active and inactive compounds and the level of accuracy was 91% [11]. Dewan, Nouria and Mohammad
combined the NBC and Decision Tree methods in the Intrusion Detection System (IDS) to detect
suspicious activities in a system or a network and obtained 99% accuracy [12]. Furthermore, Wengyuan
Dai et al who used the Naïve Bayes Classifier method to classify text on documents with the output of
97.3% [13].

Based on the results of previous studies above, it can be clearly said that the NBC method has a high
enough accuracy value in terms of predicting what will happen in the future. This is due to the fact that
the way it works is collecting existing data, studying the data and the results can provide assistance to a
precise and accurate decision making. The intention of this study is that decision-makers in motorcycle
companies can be assisted in making a final decision regarding the type of spare parts to be restocked.
Decisions taken must be precise and accurate so that the accumulation of certain types of spare parts
does not occur and conversely, the need for certain types of spare parts is not lacking.

2. Method

2.1. Data Mining

Data mining is a process that aims to extract and identify useful information and knowledge
from large-scale databases using statistical techniques, mathematics, artificial intelligence and machine
learning. The ability of data mining in searching for valuable business information from a large number
of databases can be analogous to mining precious metals from the source below the ground.

2.2. Classification

One of the basic functions of data mining is classification. The classification method is a
process of finding a model or a function to describe the class or the concept of data. The results of the
process can be used to support in describing important data and to predict the trends of the data in the
future [15].

2.3. Naïve Bayes Classifier (NBC)

NBC is a classification method that applies Bayes theory that works with a statistical approach
in pattern recognition. By recognizing patterns that have been studied, NBC is able to predict
probabilities that will occur in the future based on previous experiences. The equations of Bayes' theory
are as follows: [12], [13], [15], [16]

\[ P(A|B) = \frac{P(A)}{P(B)} P(A|B) \]  

By using assumption that is very high independence from each information (F_1, F_2, F_3, ..., F_n), then the
following equation can be applied:

\[ P(F_i|F_j) = \frac{P(F_i \cap F_j)}{P(F_j)} = \frac{P(F_i) P(F_j)}{P(F_j)} = P(F_i) \]

For i ≠ j, it becomes:

\[ P(F_i|C, F_j) = P(F_i|C) \]
It can also be written into the expression as follows:

\[ P(F_1, F_2, F_3, \ldots, F_n | C) = P(C) \prod_{i=1}^{n} P(F_i | C) \]

Then it can be described as:

\[ P(F_1, F_2, F_3, \ldots, F_n | C) = P(F_1 | C)P(F_2 | C), P(F_3 | C) \ldots P(F_n | C). P(C) \] (3)

From the above equation it can be written as (Bayes theorem):

\[ P(F_1, F_2, F_3, \ldots, F_n | C) = \frac{1}{P(F_1, F_2, \ldots, F_n)} \prod_{i=1}^{n} P(F_i | C) \]

\[ P(C | F_1, F_2, F_3 \ldots, F_n) = \frac{P(C)}{Z} \prod_{i=1}^{n} P(F_i | C) \] (4)

With the above equation, a model is acquired from the Naïve Bayes theorem which can be used in the classification process. The steps that must be done in this NBC method are [12], [13], [15], [16]:

1. Calculate the total number of classes/labels.
2. Calculate the number of cases in each class.
3. Multiply all of the class variables.
4. Compare the results of the entire class.

2.4. Data Source

The data used in this study were obtained from the database of PT. Nafriya Abadi Motor as the official dealer of PT. Honda in Bandung, Indonesia. The data source consists of 5000 data and then the data are divided into two parts – 500 data used as testing data and 4500 data as training data.

3. Results and Discussion

3.1. Data Analysis

Data obtained from the field needs cannot be used directly. Data pre-processing and data cleaning has to be done first, by adding several labels and deleting unnecessary data. In addition, an encoding process is needed that aims to convert quantitative data into qualitative data. After carrying out the process, the new label is generated, namely, the label Description (“Tidak Laku”, “Cukup Laku”, “Sangat Laku”).

There are a number of labels with their respective functions. The Part Code Label functions are to provide a unique code so that the data will not be confused with other data. The Part Name Label is used for the name of the motorcycle parts. The Quantity Label contains the sale of auto parts within one month. Label Type (Scooter, Sport, Cubs) is used to classify motorcycle types. The Month and Year Label provides information regarding the time when the motorcycle parts are sold. The Description Label contains information about the number of motorcycle parts that have been sold, then the Decision Label is used as a reference whether the parts must be restocked or not restocked.

3.2. Implementation Calculation Using NBC

The application performs calculations by using the NBC method and the results of these calculations will be divided into two different classes, namely the restock and not restock classes. The calculation is carried out by using the Bayes theory formula that has been described previously, which will find the
highest probability value to make the right decision in terms of restocking. If the value of \( P(\text{Restock} \mid X) > P(\text{Not Restock} \mid X) \) then the result generated is restocked and if \( P(\text{Restock} \mid X) < P(\text{Not Restock} \mid X) \) then the result is not restocked.

The following is a table of test data that have not had a decision result yet. Immediately after calculation using the NBC method, the decision will be generated. The calculation process in this test data is by reading the data pattern from the training data that has been done before.

1. The first stage is to count the classes number

   \[ P(Y \mid \text{Restocked}) = \frac{11}{25}, \text{the amount of data “Restocked” is divided by the amount of data available} \]

   \[ P(Y \mid \text{Not Restocked}) = \frac{14}{25}, \text{the amount of “Not Restocked” data is divided by the amount of data available} \]

2. The second stage calculates the same number of cases in the same class

   \[ P(\text{Type} = \text{Scooter} \mid Y = \text{Restocked}) = \frac{9}{11}, \text{the amount of data “Scooter” is divided by the amount of data restocked} \]

   \[ P(\text{Type} = \text{Scooter} \mid Y = \text{Not Restocked}) = \frac{9}{14}, \text{the amount of data “Scooter” divided by the number of data not restocked} \]

   \[ P(\text{Keterangan} = \text{Cukup Laku} \mid Y = \text{Restocked}) = \frac{3}{11}, \text{the amount of “Cukup Laku” data is divided by the amount of data restocked} \]

   \[ P(\text{Keterangan} = \text{Cukup Laku} \mid Y = \text{Not Restocked}) = \frac{0}{14}, \text{the amount of “Cukup Laku” data is divided by the amount of data not restocked} \]

3. The third stage is to multiply all variables

   \[ P(\text{Type} = \text{Scooter} \mid Y = \text{Restocked}) \times P(\text{Keterangan} = \text{Cukup Laku} \mid Y = \text{Restocked}) \times P(Y = \text{Restocked}) = \frac{9}{11} \times \frac{3}{11} \times \frac{11}{25} = 0,098 \]

   \[ P(\text{Type} = \text{Scooter} \mid Y = \text{Not Restocked}) \times P(\text{Keterangan} = \text{Cukup Laku} \mid Y = \text{Not Restocked}) \times P(Y = \text{Not Restocked}) = \frac{9}{14} \times \frac{0}{114} \times \frac{14}{25} = 0 \]

4. The fourth stage is the conclusion

   Due to the results obtained are \( P(\text{Restocked}) \) greater than \( P(\text{Not Restocked}) \) then the final decision to look for is "Restocked".

3.3. Test Result

   The following are the results of the training process and classification for 200 and 500 test data using the application that has been built. There is a match between the results of prediction using data mining application compared to historical sales data. There are 180 data are compatible, therefore it can be calculated the success percentage is:

   \[
   \text{success rate} = \frac{180}{200} \times 100\% = 90\%
   \]

   There are 480 out of 500 compatible data and 20 data are incompatible. The success percentage is:

   \[
   \text{success rate} = \frac{480}{500} \times 100\% = 96\%
   \]
The outcome of the application is Not Restocked. It is obtained due to the results of multiplication Not Restocked class are greater than that of Restocked class. On the other hand, in figure 3(b) the outcome is Restocked because the results of multiplication Restocked class are greater than that of Not Restocked class. The calculations and explanations are in the previous section.

The decision results are influenced by the results of multiplying the probabilities of each class itself. If the probability value for restocking is greater than not restock, then the final decision is restocked and vice versa -- if the probability value of not restock is greater than the restock then the result generated, the decision is not restocked. From the experiments that have been carried out, it can be said that to obtain the maximum results, the historical data used as training data should be propagated in order to produce a better and more accurate decision in restocking spare parts [11],[12],[16].

4. Conclusion

Based on the tests carried out, the system has a success rate of 96% which proves that this application can help the decision maker in deciding the type of parts to be restocked. Accuracy in this final decision is greatly influenced by the number of data training and data testing. The more varied and frequent data training is carried out, the higher the similarity between the results of testing using the system with the ones without using the system. Moreover, the amount of data testing must be far less than the amount of data training – for example, 500 data testing is carried out with 4500 data training. From these results, it can be ascertained that the application built using the NBC Method succeeds in providing an accurate final decision in determining whether a type of spare parts should be restocked or not. With the use of this application in a motorcycle company, the need to restock a certain spare part can be coordinated with the demand. Thus, the accumulation of certain types of spare parts stored can be avoided so as to minimize the loss of the company.

5. Acknowledgement

We give many thanks to Ms Hj. Nurlaela Aziz, the President Director of PT. Nafriya Abadi Motor, who has allowed us to conduct research within the company. We also give thanks to Mr H. Dulyamin, the PLC dealer of PT. Nafriya Abadi Motor who has assisted us with the process of collecting data from the beginning until this application was successfully built to completion. And we also thank all the company staffs who has facilitated this research.

References
[1] Lin R J, Tan K H Tan and Geng Y.2013. Market Demand, Green Product Innovation, and Firm Performance: Evidence From Vietnam Motorcycle Industry J. Clean. Prod. 40(4), pp.101–107
[2] Syntetos A A, Keyes M and Babai M Z 2009 Demand Categorisation in a European Spare Parts Logistics Network Int. J. Oper. Prod. Manag. 29(3), pp.292–316
[3] Huiskonen J.2001.Maintenance Spare Parts Logistics: Special Characteristics and Strategic Choices Int. J. Prod. Econ. 71(3), pp.125–133
[4] Gunasekara A, Pisedtasalasai A and Power D M 2004 “Macroeconomic Influence on the Stock Market: Evidence from an Emerging Market in South Asia J. Emerg. Mark. Financ. 1(1) pp.285–304
[5] LeBaron B, Arthur W B and Palmer R 2002. Time Series Properties of an Artificial Stock Market J. Econ. Dyn. Control. 23(2), pp.1487–1516
[6] Fortuin L 2006. Stocking Strategy for Service Parts ± a Case Study Int. J. Oper. Prod. Manag. 20(3), pp. 656–674
[7] Kennedy W J, Wayne Patterson J and Fredendall L D.2002. An Overview of Recent Literature on Spare Parts Inventories Int. J. Prod. Econ. 76 (3), pp 201–215
[8] Wagner S M and Lindemann E.2008. A Case Study-Based Analysis of Spare Parts Management in the Engineering Industry Prod. Plan. Control 19 (1), pp 397–407
[9] Lewis D D. 1998. Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval. European conference on machine learning (USA: AT&T Labs - Research). 3 (1), pp. 4–15

[10] Chai X et al. 2005. Test-Cost Sensitive Naive Bayes Classification. Fourth IEEE International. 7(3), pp. 51–58

[11] Glick M et al. 2004. Enrichment of Extremely Noisy High-Throughput Screening Data Using a Naïve Bayes Classifier. J. Biomol. Screen 9 (3), pp. 32–36

[12] Singh D M, Harbi N and Zahidur Rahman M. 2010. Combining Naive Bayes and Decision Tree for Adaptive Intrusion Detection. Int. J. Netw. Secur. Its Appl. 2 (3), pp. 12–25

[13] Dai W et al. 2007. Transferring Naive Bayes Classifiers for Text Classification. Sci. Technol. 4 (1), pp. 540–545

[14] Patil T R and Sherekar S S. 2013. Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification. Int. J. Comput. Sci. Appl. 6 (2), pp. 256–261

[15] Ginting S L B et al. 2018. The Development of Bank Application for Debtors Selection by Using Naïve Bayes Classifier Technique. IOP Conference Series: Materials Science and Engineering, 407 (1), pp. 1–26

[16] Rozi I F, Pramono S H and Dahlan E A. 2012. Implementasi Opinion Mining (Analisis Sentimen) Untuk Ekstraksi Data Opini Publik pada Perguruan Tinggi (Implementation of Opinion Mining for the Extraction of Public Opinion Data at Higher Education). J. EECCIS Electrics, Electron. Commun. Control. Informatics, Syst. 6(1), pp. 37–43