RESEARCH ARTICLE   OPEN ACCESS

The Prediction of Best-Selling Product Using Naïve Bayes Algorithm (A Case Study at PT Putradabo Perkasa)

Arif Budiyanto[1], Saruni Dwiasnati[2]

[12]Mercu Buana University, Faculty of Computer Science,
Jl. Meruya Selatan 1 No. 1, Kembangan, 11650, Indonesia
Email : 41514110148@mercubuana.ac.id[1], Saruni.Dwiasnati@mercubuana.ac.id[2]

--*************************--------------------------------

Abstract:

Data Mining is a technique for processing and extracting large data into information which can form new data. The technique is applied by using Knowledge Discovery Process in Database (KDD). The objective of the research is determining which product is the best-selling in 2018, so that the increase of customer's demand can be anticipated in the following year. In this research, the authors employ classification method in producing the best-selling product information by using algorithms naïve Bayes. There are some variables involved, such as type of goods, brand of goods, quality of goods, price of goods, and Target. Rapid minner Studio 9.0 is a tool for assessing data which calculated to produce a model. The analysis results are expected to be used by the company for preparation supply of the best-selling products. The findings reveal that the level of data accuracy is is 78,33% and the best-selling product based on sales is the IP Camera product with type Infinity I-993V.

Keywords **— Data Mining, Classification, Naïve Bayes , Sales Transaction, Rapidminer Studio 9.0.**

--*************************--------------------------------

1. INTRODUCTION

PT Putradabo Perkasa is a company engaged in the procurement and goods services of CCTV systems along with Access Control systems. It has a few problems related to process data for goods supply in the warehouse. The current system is still conventional with books and sometimes the admin who records the amount of inventory is often not in accordance with the existing stock. In general, Data Mining is a process or activity to collect large-sized data then extract the data into information that can form new data. Due to the rapid growth of technology, data mining era is increasingly needed in the world of work. It can reduce level of errors and make work more efficient and effective [1]. By using data mining technique, the processed data will be useful for company. The method is classification using algoritma naïve bayes. It is started from a problem, namely the increase of selling transaction which cause the admin does not know that there are certain items that have a high level of selling and must be purchased. In this case, the researchers will classify sales transaction data to make the analysis process become easier, and result of analysis can be used to predict which items are in demand and which items are not in demand [2].

In processing data, the researchers apply Knowledge Discovery in Database (KDD). This process serves to discard inconsistent data and missing value data. In general, Naive Bayes is a simple probabilistic classification that calculates a set of probabilities by summing the frequency and combination of values from the given dataset [3].

The data mining process framework is composed of 3 stages, namely data collection, data transformation, and data analysis. The process of data mining begins with preprocessing which consists of collecting data to produce raw data which needed by data mining, then it is followed by data transformation in order to convert raw data into formats that can be processed by data mining, for example through filtration or aggregation. The results of data transformation will be used by data analysis to generate new knowledge using several techniques, such as statistical analysis, machine learning, and information visualization[4].

2. BASIC THEORY DAN METHODS

2.1 Classification

Classification is a method for grouping data in a system according to predetermined rules. Classification can also be interpreted as grouping new data or new objects based on observed variables with the aim of predicting an object from an unknown class or unknown category. Classification is a data mining technique that looks at the attributes of a group of data that has been previously defined. So, it can provide a *classification* of new data by manipulating the data that has been classified and use the results to provide some rules. These rules are implemented on new data in order to be further classified. The purpose of the *classification* is grouping class, so that unknown records in the previous category can be classed accurately[5].

2.2 Data mining

Data Mining is a technique for finding information from large amounts of data which stored in database, warehouse data, or other storage media. Data mining methods is various. The best methods for data mining are C4.5, k-means, supporting vector, priori, EM, PageRank, AdaBoost, KNN (Nearest Neighbors), Naive Bayes, CART (Classification and Regression Tree) methods. While for each method adjusts to the data mining function, namely the characterization function, discrimination function, association function, prediction function, classification function, and cluster function[6].

2.3 Naïve Bayes Algorithm

The Bayes theorem invented by Thomas Bayes in the 18th century is a pattern in Bayesian learning that is very familiar and has a directed viewpoint to produce the required probability results. This problem stems from Bayes's theorem and freedom hypothesis, gives rise to a classification of statistics based on the occurring possibilities. This algorithm is a conventional technique and can be used to try more complex patterns. The Naive Bayes algorithm can be interpreted as a simple probabilistic classification that calculates a set of probabilities by summing the frequencies and combinations of values from the given dataset as a result of probability which has certain characters or not from the interaction category with other category characteristics [7].

The equation of the Bayes theorem is:

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)}$$

Note:
X : Data with unknown classes
H : X data hypothesis is a specific class
P(H|X) : Probability of hypothesis based on condition (posteriori probability)
P(H) : Probability of hypothesis H (prior probability)
P(X|H) : Probability of X is based on conditions in hypothesis H
P(X) : Probability X

2.4 KDD

KDD stands for Knowledge Discovery from Data. KDD began to be developed in the early 1990s. Fayyad in 1996 initiated the KDD model process and set the steps for the Data Mining project[8].
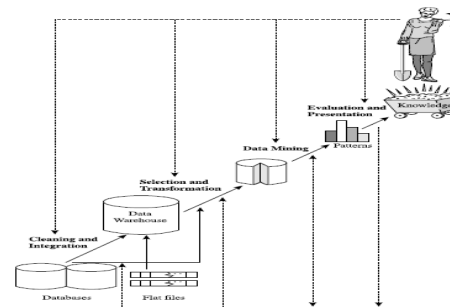

Fig.1 Stages of Methodology KDD

Based on figure 1, the stages of KDD methodology can be explained as follows.

1) **Data Cleaning**
   At this stage, select the relevant data from the database by separating inconsistent data and irrelevant data.
2) **Data Integration**
   At this stage, integration of existing data is done by combining various data sources into one source.
3) **Data Selection**
   At this stage, the selection of relevant data is applied by analyzing database.
4) **Data Transformation**
   At this stage, changes to the existing data format are made into the appropriate data format to be processed in data mining.
5) **Data Mining**
   At this stage, data mining process is carried out, by applying certain methods to obtain hidden information from existing data.
6) **Pattern Evaluation**
   At this stage, identification of interesting patterns obtained from the results of data mining, and it will be represented later.
7) **Knowledge Presentation**
   At this stage, visualization and presentation of knowledge about the techniques used to obtain knowledge.

2.5. Rapidminer Studio 9.0

RapidMiner is a data science software platform developed by a company which has the same name as that which provides an integrated environment for data preparation, machine learning, deep learning, text mining, and predictive analysis. It is used for business and commercial purposes, also for research, education, training, rapid prototyping, and application development, and supports all steps in the machine learning process including data preparation, visualization results, model validation, and optimization. RapidMiner is developed in open core models, which is limited to 1 logic processor and 10,000 data lines, it is available under the AGPL license[9].

2.6 ROC Curve

some basic notions about the ROC (Receiver Operating Characteristic) curves, widely used in assessing the results of predictions (fore- casts). We schematically present the most important aspects regarding the ROC curves Just to familiarize the reader with this technique.

1) The ROC curves were first developed by electrical and radar engineers during World War II for detecting enemy objects in battlefields (e.g., differentiating en emy aircrafts from flocks of birds -see, for instance, the štory of Pearl Harbor attack in 1941, or problems of British radar receiver operators)
2) ROC curves have long been used in signal detection theory, to depict the trade-off between hit rates (signal) and false alarm rates (noise) of classifiers. Soon after that, they have been introduced in psychology to account for perceptual detection of signals.
3) ROC curves are also commonly used in medical research.
4) ROC curves are also usually used in machine learning and data mining research. One of the earliest adopters of ROC curves in machine learning was Spackman,who demonstrated the value of ROC curves in evaluating and comparing algorithms. Recent years have seen an increase in the use of ROC curves in the machine learning community.
5) In classification problems, the ROC curve is a technique for visualizing, organizing and selecting classifiers, based on their performance.

The calculation results are visualized by the Receiver Operating Characteristic (ROC) or Area Under Curve (AUC). ROC has a level of diagnostic value, namely (Gorunescu, 2011):

a. Accuracy up to 0.90 – 1.00 = *excellent classification*

b. Accuracy up to 0.80 – 0.90 = *good classification*

c. Accuracy up to 0.70 – 0.80 = *fair classification*

d. Accuracy up to 0.60 – 0.70 = *poor classification*

e. Accuracy up to 0.50 – 0.60 = *failure*

3. RESULT AND DISCUSSION

3.1. Research Methodology

This type of research is a qualitative method using a case study approach. The case study approach is a type of approach used to investigate and understand an event or problem that has occurred by gathering various kinds of information and processing it to obtain a solution, so that the problems revealed can be resolved.

3.2 Data Collection Method

The researchers conduct a case study directly to PT Putradabo Perkasa to analyze sales transaction data. The author also takes sample data for later analysis. Identification of valid patterns, potential benefits can be easily understood. The availability of data can be generated for the use of information technology in almost all fields of life, this can utilize the information and knowledge contained in the data overflow, which is then called data mining. Data mining is a process of discovering knowledge (knowledge discovery) that is mined from a large collection of data. The application of data mining is in the management of educational world, business management, production control, and market analysis, for example, enable the acquisition of patterns and relationships that can be used to improve learning methods, increase sales of a product, or manage existing resources well[10].

3.3 Initial Process

Initial process is process of excel readability (Read Excel). Data training and data testing is unified in the form of Excel 2016. The amount of data transaction is 219 in the year of 2018. There are two classes in the data testing, namely selling and buying class.

Fig.2 Data Training

Fig.3 Data Training

3.4 Validation Process

The process of validation is the process of analyzing and looking for a model based on the best performance. Where read excel is connected to the validation block. To see the process of training data and testing data, it can be done by double clicking the validation block.
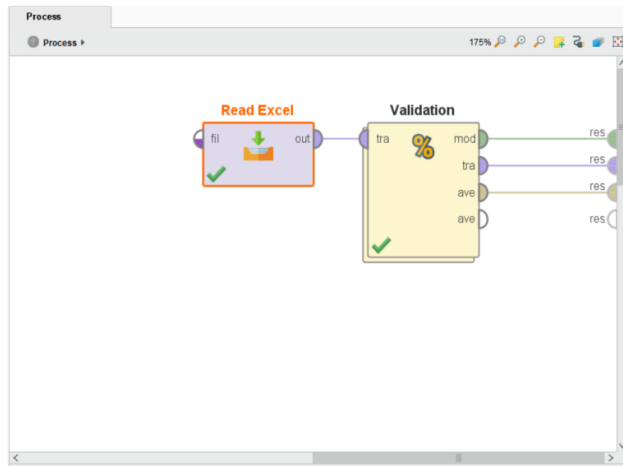
Fig.4 Validation Process

3.5 Training and Testing Process

Training Process is training data model (Naïve Bayes). Testing process is a testing process that will produce a graph or pattern. In figure 4.3, it is explained in the training data section by using naïve bayes blocks, whereas models and performance is applied in the block testing section.
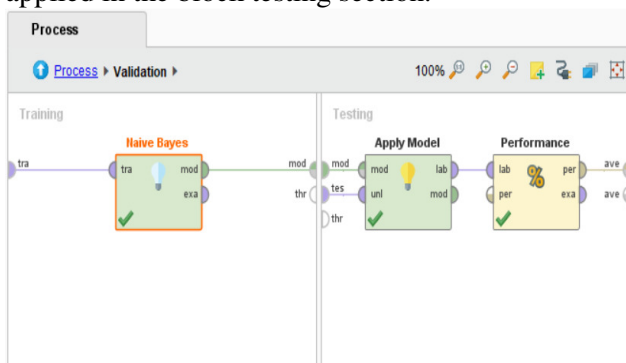
Fig.5 Training and Testing

3.6 Performance Vector Result

By using Rapidminer Studio 9.0, classification process with naïve Bayes algorithm aim to find out the products that are often sold based on sales transactions which will result accuracy, precision, recall, and AUC.

PerformanceVector

```
PerformanceVector:
accuracy: 78.33%
ConfusionMatrix:
True:     Ya        Tidak
Ya:       41        6
Tidak:    7         6
precision: 46.15% (positive class: Tidak)
ConfusionMatrix:
True:     Ya        Tidak
Ya:       41        6
Tidak:    7         6
recall: 50.00% (positive class: Tidak)
ConfusionMatrix:
True:     Ya        Tidak
Ya:       41        6
Tidak:    7         6
AUC (optimistic): 0.731 (positive class: Tidak)
AUC: 0.723 (positive class: Tidak)
AUC (pessimistic): 0.715 (positive class: Tidak)
```

Fig.6 PerformanceVektor View Text

3.8 Accuracy

From the testing result of Naïve Bayes algorithm, the level of accuracy data up to 78,33%.

accuracy: 78.33%	true Ya	true Tidak	class precision
pred. Ya	41	6	87.23%
pred. Tidak	7	6	46.15%
class recall	85.42%	50.00%	

Fig.7 Accuracy

3.9 Simple Distribution Results

The results of simple distribution based on kind of transaction are yes class up to 0,808 and no class up to 0, 196.

SimpleDistribution

Distribution model for label attribute Laris

Class Ya (0.804)
4 distributions

Class Tidak (0.196)
4 distributions

Fig.8simpeldistribution view Text

Fig. 9 simpeldistribution graph

Based on figure 9, it can be concluded that the best-selling product based on sales is IP camera product with type infinity I-993V. By knowing the best-selling product, the preparation for the supply product in the next year can be anticipated.

3.10AUC (Area Under Curve)

The testing results produce a Receiver Operating Characteristic (ROC) curve which can be seen in the picture below:

Fig. 10Curve ROC

In Figure 10, the AUC value is 0.723. With the obtained value, it can be said that the classification belongs to the Fair classification category.

4. CONCLUSIONS

From the results of the analysis conducted by the researchers, there are several conclusions obtained including:

1) Based on the data testing that has been done using the naïve bayes algorithm, the level of data accuracy is 81.36%. The AUC value obtained by using the naïve bayes algorithm is 0.723. The AUC classification level
2) with the value obtained has entered the fair classification category.
3) Based on the simple distribution chart, the best-selling product is IP camera with type Infinity I-993V.
4) For further research, it is expected that the results of the analysis can be implemented into an application.

REFERENCES

[1] P. Protoview, B. Vol, and B. Ringgold, "Predictive Analytics and Data Mining : Concepts and Practice With RapidMiner," vol. 2, pp. 1–2, 2015.

[2] P. O. L. Júnior, L. G. C. Júnior, and A. L. Zambalde, "Applying Textmining to Classify News About Supply and Demand in the Coffee Market," vol. 14, no. 12, pp. 4768–4774, 2016.

[3] J. Sistem, I. Fakultas, I. Komputer, and U. Sriwijaya, "PENERAPAN KNOWLEDGE MANAGEMENT SYSTEM (KMS) MENGGUNAKAN TEKNIK KNOWLEDGE DATA DISCOVERY (KDD) PADA PT PLN (PERSERO) WS2JB RAYON KAYU AGUNG," J. Sist. Inf. (JSI), VOL. 8, NO. 2, Oktober 2016, vol. 8, no. 2, pp. 1038–1055, 2016.

[4] B. Studi, K. Cv, M. Abadi, A. Yunaldi, and C. V Mahkota, "METODE NAIVE BAYES DALAM MEMPREDIKSI STOK," pp. 46–54.

[5] Y. An, S. Sun, and S. Wang, "Naive Bayes Classifiers for Music Emotion Classification Based on Lyrics," no. 1, pp. 635–638, 2017.

[6] R. Avrizal, A. Wibowo, A. S. Yuniarti, D. A. Sandy, and K. Prihandani, "Analysis Comparison of The Classification Data Mining Method to Predictthe Decisions of Potential Customer Insurance Abstract :," vol. 5, no. 5, pp. 15–20, 2018.

[7] Saruni Dwiasnati, Yudo Devianto " Naive Bayes Optimization Based On Particle Swarm Optimization to Predict the Decision of Insurance Customer Candidate " Vol. 5 - Issue 5 (08 - 14) - September - October 2018 , International Journal of Computer Techniques (IJCT) , ISSN: 2394 - 2231 , www.ijctjournal.org

[8] P. Ristoski and H. Paulheim, "Web Semantics : Science , Services and Agents on the World Wide Web Semantic Web in data mining and knowledge discovery : A comprehensive survey," vol. 36, pp. 1–22, 2016.

[9] A. Kori, "Comparative Study of Data Classifiers Using Rapidminer," vol. 5, no. 2, pp. 1041–1043, 2017.

[10] Firdaus Diky " Penggunaan Data Mining dalam Kegiatan Sistem Pembelajaran Berbantuan Komputer" Volume 6 Nomor 2 Tahun 2017 (Hal 91-97), Jurnal Format , ISSN : 2089 -5615.

[11] U. S. Fed, N. Service, I. Us, S. News, and D. C. Washington, "OKLAHOMA STATE UNIVERSITY AWARDS DATA MINING , BUSINESS ANALYTICS CERTIFICATES TO MAY GRADUATES," no. June, pp. 1–6, 2015.

[12] V. K. Goyal, "A Comparative Study of Classification Methods in Data Mining using RapidMiner Studio."

[13] N. P. Astuti and M. R. Arief, "ANALISIS PREDIKSI TINGKAT KETIDAKDIPLINAN SISWA MENGGUNAKAN ALGORITMA NAÏVE BAYES CLASSIFIER (STUDI KASUS : SMK NEGERI 1 PACITAN)," pp. 6–8, 2015.

[14] P. Bidang, K. Sains, Y. Mardi, J. Gajah, M. No, and S. Barat, "Jurnal Edik Informatika Data Mining : Klasifikasi Menggunakan Algoritma C4 . 5 Data mining merupakan bagian dari tahapan proses Knowledge Discovery in Database (KDD) . Jurnal Edik Informatika."

[15] E. D. Rahmawati and M. Susilowati, "Sistem Informasi Persediaan Dan Penjualan Barang Berbasis Web Pada Koperasi Kosma 15," pp. 9–10, 2015.