



Application of Data Mining to Predict Procurement of Office Writing in Al-Ikhwon Middle School using Naïve Bayes Method

Aida Sopia, Muhammad Syahrizal

Department of Computer Science, Universitas Budi Darma, North Sumatra, Indonesia
Email: aidasop12@gmail.com

Abstract—Office stationery (ATK) is now a necessity for almost everyone, especially for corporate instances or educational institutions. The need for office stationery is often an unexpected need to buy, this is what makes some educational institutions overwhelmed in doing their work, when they find out their office stationery is out of stock, so it is not uncommon to make work in company or institutional instances education is not completed on time, one of the ways to be more efficient is by implementing data mining to predict the purchase of office stationery (ATK) at the company or educational institution's intents, especially at the AL IKHWAN Middle School in Tanjung Morawa. The Naive Bayes method is used to analyze data in pattern recognition and predict purchase of office stationery (ATK) at AL IKHWAN Middle School in Tanjung Morawa. The data needed is the data for the purchase of office stationery (ATK) last month as test data, calculated from the date of the first purchase until the expiry date of office stationery (ATK) at AL IKHWAN Middle School in Tanjung Morawa. The results of this study are to be able to predict whether the office stationery (ATK) at AL IKHWAN Middle School in Tanjung Morawa can be bought back, or it can still be used for a long time, and if more than four types of stationery at the AL IKHWAN Middle School in Tanjung Morawa the lack of writing instruments from the two then the purchase of new stationery is feasible to do, from the amount of data occurring out of stationery.

Keywords: Prediction, Office Stationery, Naïve Bayes

1. INTRODUCTION

Data mining is a series of processes to get useful information from large database warehouses. Data mining can also be interpreted as extracting new information taken from large chunks of data that helps in decision making. In data mining there are many techniques in the process, including the naïve bayes algorithm, decision tree, artificial neural networks and many others. Prediction is a process of systematically estimating about something that is most likely to occur in the future based on past and present information that is owned, so that the error (the difference between something that happens with the estimated results) can be minimized. Predictions do not have to provide definitive answers to events that will occur, but rather try to find answers as close as possible to occur. Utilization of existing data in information systems to support decision making activities, it is not enough to rely solely on operational data, a data analysis is needed to explore the potential of existing information. Decision makers try to utilize data warehouse that has been owned to dig up information that is useful to help make decisions. Office stationery (ATK) has now become the need of almost everyone, especially for corporate agencies or educational institutions. the use of existing data in the information system to support decision-making activities, it is not enough just to rely on operational data alone, we need a data analysis to explore the potential of existing information. Decision makers try to use the data warehouse that is already owned to dig up information that is useful to help make decisions. The need for office stationery is often an unexpected need to buy, this makes some educational institutions overwhelmed in doing their work, when in know their office stationery turns out to have run out, so it is not uncommon to make work in company agencies or educational institutions not completed on time, one way to be more efficient is by implementing data mining to predict the purchase of office stationery (ATK) on company intansi or educational institutions, especially in the Middle School AL-IKHWAN Tanjung Morawa that is using data mining techniques. So that a lot of data can be used optimally. In previous research is still not optimal and still needs to be improved, as there is no determination of the time available in the warehouse with what is recorded on the computer so that it can result in a difference in the amount of inventory, besides remembering stationery the office is a small object and easily transferred, it is not easy to ascertain how much expenditure must be borne within a certain period of time. Then it is necessary to determine the time in checking the inventory of goods in the computer with those available in the warehouse, so as to minimize the difference in recording the amount Inventory. Naive Bayes algorithm is one of the algorithms found in classification techniques. Naive Bayes is a classification with probability and statistical methods raised by the British scientist Thomas Bayes, which predicts future opportunities based on past experience so that it is known as the Bayes Theorem. The theorem is combined with Naive where it is assumed that conditions between attributes are mutually independent. The Naive Bayes classification is assumed that the presence or absence of certain characteristics of a class has nothing to do with the characteristics of other classes [1].



2. THEORY

2.1 Data Mining

Data mining as the process of obtaining useful information from large database warehouses. Data mining can also be interpreted as the extraction of new information taken from large chunks of data that helps in making decisions. The term data mining is sometimes also called knowledge discovery. One technique in data mining is how to trace existing data to build a model, then use the model so that it can recognize other data patterns that are not in a stored database. The need for predictions can also utilize this technique. In data mining, data grouping can also be done. The aim is so that we can know the universal pattern of existing data. Anomaly of transaction data also needs to be detected to find out the next follow-up that can be taken. All of these are aimed at supporting the company's operational activities so that the company's ultimate goal is expected to be achieved.

2.2 Naïve Bayes

Bayes is a simple probabilistic-based prediction technique based on the application of the Bayes theorem / Bayes rules with a strong assumption of independence (non-dependence) (Naive). In other words, in Naive Bayes the model used is the independent feature model. In Bayes (especially Naive Bayes), the strong intention to refer to features is that a feature in a data is not related to the presence or absence of other features in the same data. The relationship between Naïve Bayes with classification, correlation of hypothesis, and evidence with classification is that the hypothesis in the Bayes theorem is the class label that is the target of mapping in the classification, while the proof is the features that are input in the classification model. If X is an input vector that contains features and Y is a class label, Naïve Bayes is written as $P(Y / X)$. The notation means that the Y class label probability is obtained after the X features are observed. This notation is also called the final probability (posterior probability) for Y , while $P(Y)$ is called the initial probability (prior probability) Y .

During the training process the final probability learning ($P / Y / X$) of the model for each combination of X and Y must be made based on the information obtained from the training data by building the model, an X test data can be classified by finding the value of Y by maximizing the value ($P(Y / X)$) obtained.

$$P(Y|X) = \frac{P(Y) \prod_{i=1}^n P(X_i|Y)}{P(X)} \quad (1)$$

However, for features with numeric (continuous) type there are special treatments before they are included in Naïve Bayes. The trick is Perform discretization on each continuous feature and replace the value of the continuous feature with a discrete interval value. This approach is carried out by transforming continuous features into ordinal features.

Analyze specific forms of probability distributions for continuous features and estimate distribution parameters with training data. The gaussian distribution is usually chosen to represent the conditional probability of a continuous feature in a class $P(X_i | Y)$, while the gaussian distribution is characterized by two parameters: mean, μ , and variance, σ . For each class y_j , the conditional probability y for the X_i feature is:

$$P(X_i = x_i | Y = y_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}}} \exp \left[-\frac{(x_i - \mu_{ij})^2}{2\sigma_{ij}^2} \right] \quad (2)$$

2.3 Prediction

Prediction is predicting, predictions about things that have not yet happened. While according to some experts argue that, predicting is the process of scientific forecasts to gain systematic knowledge based on physical evidence. These words or terms can give rise to various perceptions. It can even mean multiple or multiple interpretations. No wonder that there is overlapping in the community in interpreting it. According to the large Indonesian dictionary, predictions are the result of predicting or predicting or estimating activities. Prediction can be based on mere scientific or subjective methods. Take for example, weather predictions are always based on the latest data and information based on observations including satellites. Likewise, earthquake predictions, volcanoes erupt or disasters in general. However, predictions such as soccer matches, sports, etc. are generally based on subjective views with their own perspective predicting them.

3. RESULT AND DISCUSSION

Analysis can be defined as the decomposition of a complete information system into its component parts with the aim of identifying and evaluating problems, opportunities, obstacles that occur in the expected needs so that



improvements can be proposed. The study was conducted at Al-Ikhwan Middle School and based on research conducted data processing is still done manually. Analysis in the procurement of office stationery (ATK) really needs to be done at Al-Ikhwan Middle School. This analysis is useful to get the main problems that actually are the core problems that often occur in Al-Ikhwan Middle School. In this case it is necessary to analyze the procurement of office stationery (ATK), this aims to make the school know the steps that must be taken in recording ATK supplies in the coming month.

Office Stationery (ATK) procurement data needs to be known because to find out what office equipment has been used so far, this data will be useful to find out what equipment has been used up and needs to be added next month. As for the procurement of ATK data the last 3 months can seen in the picture below.

Office Stationery (ATK) data usage needs to be known because to find out what office equipment has been used so far, this data will be useful to find out what equipment has been used up and needs to be added next month. As for the use of ATK data the last 3 months can seen in the picture below.

Table 1. ATK Usage Data for February-April 2017

No	Pengadaan		Penggunaan		Jmlh Sisa Barang	Keterangan
	Jenis Barang	Jumlah barang	Jmlh Barang Terpakai	Jmlh Barang		
1	Ballpoint Boxy Uniball /001	48 Pcs	24 Pcs	24 Pcs	24 Pcs	Tidak Perlu Ditambah
2	Kertas HVS	402 L	393 L	9 L	9 L	Perlu ditambah
3	Penggaris besi 100 cm	12 Pcs	10 Pcs	2 Pcs	2 Pcs	Tidak Perlu ditambah
4	Spidol besar Snowman	48 Pcs	24 Pcs	24 Pcs	24 Pcs	Tidak Perlu ditambah
5	Tinta spidol Snowman Maker	12 Pcs	10 Pcs	2 Pcs	2 Pcs	Perlu ditambah
6	Kapur	252 Pcs	190 Pcs	60 Pcs	60 Pcs	Tidak Perlu ditambah
7	Penghapus tulis/Whiteboard papan	12 Pcs	9 Pcs	3 Pcs	3 Pcs	Tidak Perlu ditambah
8	Buku Absensi	24 Pcs	20 Pcs	4 Pcs	4 Pcs	Tidak Perlu ditambah
9	Buku Double Folio	9 Pcs	4 Pcs	5 Pcs	5 Pcs	Tidak Perlu ditambah
10	Lem Kertas	6 Pcs	3 Pcs	3 Pcs	3 Pcs	Tidak Perlu ditambah
11	Pulpen Pilot	60 Pcs	28 Pcs	32 Pcs	32 Pcs	Tidak Perlu ditambah
12	Amplop	24 Pcs	22 Pcs	2 Pcs	2 Pcs	Perlu ditambah
13	Buku Tulis	24 Pcs	22 Pcs	2 Pcs	2 Pcs	Tidak Perlu ditambah

Then from the table above can be calculated using the Naive Bayes Classification formula, as for how it works as follows:

$$X \text{ needs to be added} = \frac{9 + 2 + 2}{3} = \frac{13}{3} = 4,333$$

$$X \text{ tidak perlu ditambah} = \frac{24 + 2 + 24 + 60 + 3 + 4 + 5 + 3 + 32 + 2}{10} = \frac{159}{10} = 15,9$$

$$s^2 \text{ perlu ditambah} = \frac{(9 - 4,333)^2 + (2 - 4,333)^2 + (2 - 4,333)^2}{3 - 1}$$

$$s^2 = \frac{21,780 + 5,443 + 5,443}{2} = \frac{32,666}{2} = 16,333$$

$$S \text{ perlu ditambah} = \sqrt{16,333} = 4,041$$

$$s^2 \text{ tidak perlu ditambah} = \frac{(24 - 15,9)^2 + (2 - 15,9)^2 + (24 - 15,9)^2 + (60 - 15,9)^2 + (3 - 15,9)^2 + (4 - 15,9)^2 + (5 - 15,9)^2 + (3 - 15,9)^2 + (32 - 15,9)^2 + (2 - 15,9)^2}{10 - 1}$$

$$s^2 = \frac{65,61 + 193,21 + 65,61 + 1,944 + 166,41 + 141,61 + 118,81 + 166,41 + 259,21 + 193,21}{9} = \frac{1,372034}{9} = 152,448$$

$$S \text{ tidak perlu ditambah} = \sqrt{152,448} = 12,347$$

Because the final probability value does not need to be added, office stationery does not need to be added.



4. IMPLEMENTATION

To run Weka application, you must first prepare the needs of the program to be implemented both in terms of hardware and computer software. Below are the testing steps in the Naïve Bayes algorithm using WeKa tools. Create the xlsx format on a Microsoft Excel worksheet

No	Jenis Barang	Jumlah barang	Jmlh Barang Terpakai	Jmlh Sisa Barang	Keterangan
1	Balpoint Boxy Unball /001	48 Pcs	24 Pcs	24 Pcs	Tidak Perlu Ditambah
2	Kertas HVS	402 L	393 L	9 L	Perlu Ditambah
3	Penggaris besi 100 cm	12 Pcs	10 Pcs	2 Pcs	Tidak Perlu Ditambah
4	Spidol besar Snowman	48 Pcs	24 Pcs	24 Pcs	Tidak Perlu ditambah
5	Tinta spidol Snowman M&Ker	12 Pcs	10 Pcs	2 Pcs	Perlu Ditambah
6	Kapur	252 Pcs	190 Pcs	60 Pcs	Tidak Perlu Ditambah
7	Penghapus papan tulis/Whiteboard	12 Pcs	9 Pcs	3 Pcs	Tidak Perlu Ditambah
8	Buku Abstraksi	24 Pcs	20 Pcs	4 Pcs	Tidak Perlu Ditambah
9	Buku Double Folio	9 Pcs	4 Pcs	5 Pcs	Tidak Perlu Ditambah
10	Lembar Kertas	6 Pcs	3 Pcs	3 Pcs	Tidak Perlu Ditambah
11	Philips B&K	60 Pcs	28 Pcs	32 Pcs	Tidak Perlu ditambah
12	Amplong	24 Pcs	22 Pcs	2 Pcs	Perlu Ditambah
13	Buku Tulis	24 Pcs	22 Pcs	2 Pcs	Tidak Perlu Ditambah

Figure 1. Data Format xlsx

The data format is stored on a Microsoft Excel worksheet. Microsoft Excel is the xlsx data storage database, with save as type being CSV (Comma Delimited). Microsoft Excel will be connected to Weka Tools. Then Data appears with 5 Attributes

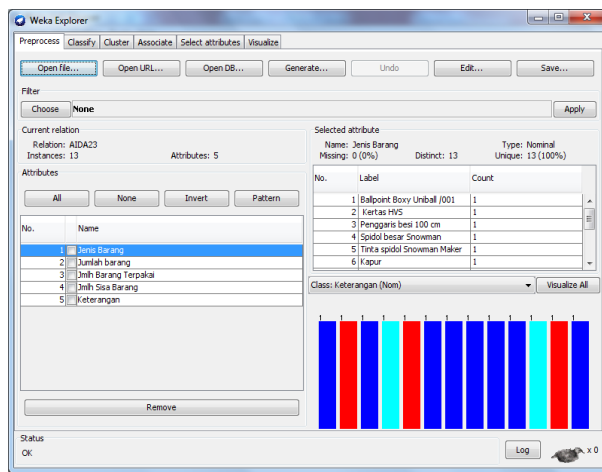


Figure 2. Attribute Item Type

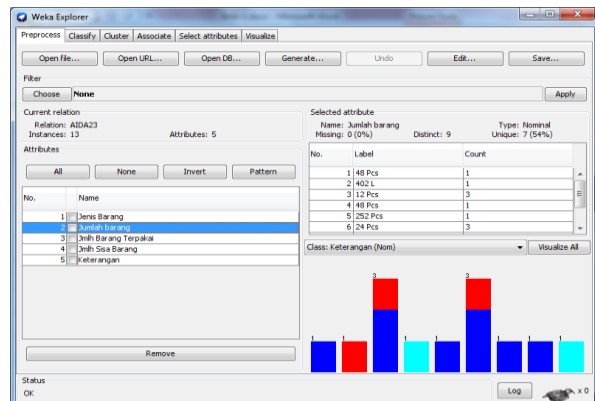


Figure 3. Attribute Amount of Goods

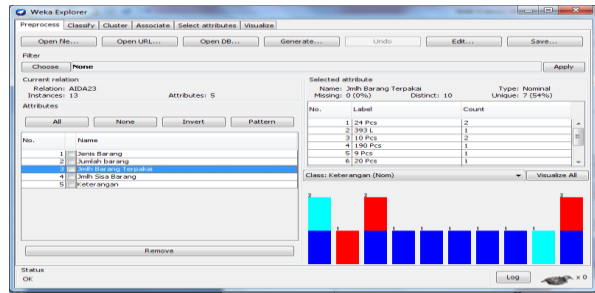


Figure 4. Attribute Amount of Goods Used

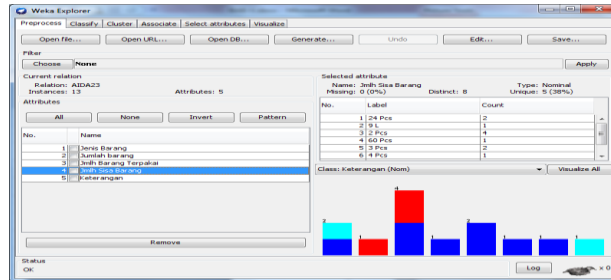


Figure 5. Amount of remaining goods

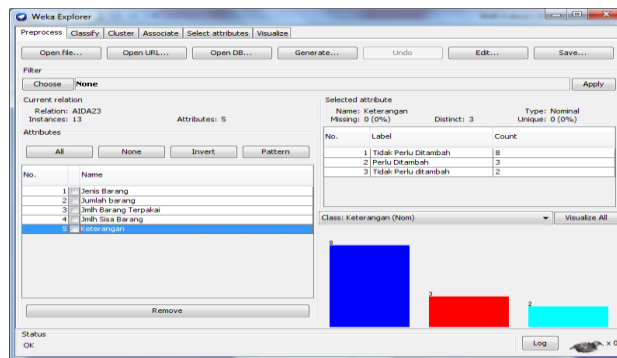


Figure 6. Information

Naive Bayes Method, Select the "Classify" menu then click "Choice" to choose the Naive Bayes method

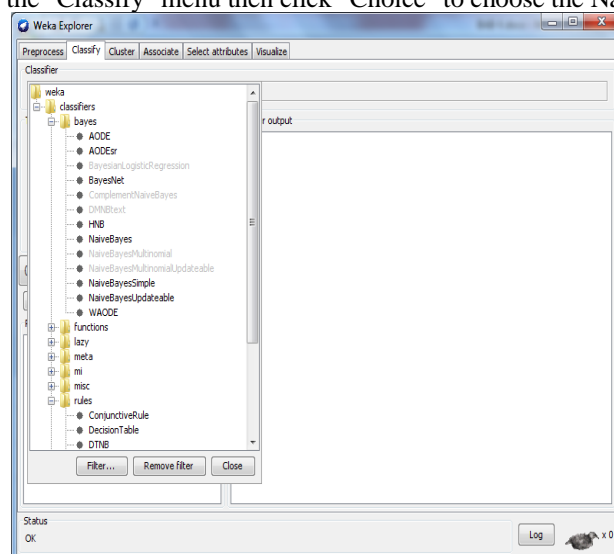


Figure 7. Display of the Naive Bayes Method

Results of classification using the Naïve Bayes Algorithm



Select the "Use Training Set" menu. Then select the "Start" test using the Naïve Bayes algorithm successfully displayed.

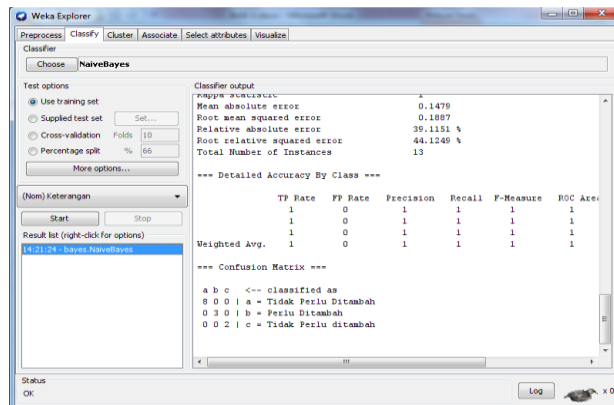


Figure 8 Naive Bayes Algorithm Classification Results

Based on figure 8 above, it can be seen that the percentage for Correctly Classified Instance is 100% while the percentage for Incorrectly Classified Instance is 0%.

5. CONCLUSION

Based on the results of the author's analysis, it can be concluded in outline in making this system are as follows.

1. The process of procuring office stationery with a total of 13 items based on testing with the Weka application, the results do not need to be added.
2. By implementing the Naïve Bayes method it can be easier to predict office stationery.
3. Test results with Weka application in office stationery prediction are less than optimal.

REFERENCES

- [1] Bustami, "PENERAPAN ALGORITMA NAIVE BAYES," jurnal informatika(journal of informatics) ISSN 1978-0524, vol. vol 8 No 1, no. <http://journal.uad.ac.id/index.php/JIFO/article/view/2086/1335>, 2014.
- [2] Siti Rohimah, Identitas Sekolah. Medan, 2008.
- [3] KBBI. <http://kbbi.web.id/terap-2>.
- [4] Eko Prasetyo, Data Mining. Yogyakarta, 2012.
- [5] Dicky Nofriansyah, "Konsep Data Mining vs Sistem Pendukung Keputusan". Yogyakarta, 2014.
- [6] Efori Buulolo, "Implementasi Algoritma Apriori Pada Sistem Persediaan Obat(Studi Kasus:Apotik Estomihi Rumah Sakit Medan)," Pelita Informatika Budi Darma, vol. vol:lv,nomor:1, agustus 2013.
- [7] Emha dan Taufik Luthfi, Algoritma Data Mining. Yogyakarta, 2009.
- [8] KBBI. <http://Kbbi.web.id/prediksi>.
- [9] kbbi. <http://kbbi.web.id>.
- [10] kbbi. <http://kkbi.web.id/atK>.
- [11] Suwanto Raharjoi, Ema Utami, Logika Algoritma dan Implementasinya dalam Bahasa Phyton di GNU/Linux. Yogyakarta, 2004.
- [12] Feri Sulianta & Dominikus Juju, Data mining. Jakarta, 2010.