4thICITB

IMPLEMENTATION OF NAIVE BAYES CLASSIFIER ALGORITHM FOR DETERMINING CUSTOMERS' BUYING INTEREST ON SOFA (A CASE STUDY IN BAYAU KELUMER FURNITURE)

Ketut Artaye[1)], Kevin Moniaga Suharta[2)]

Institute of Informatics and Business
Email: artajaya@darmajaya.ac.id, ouma.vinshu@gmail.com

ABSTRACT

Interest is the personal intention and related to attitude.Individuals who are interested in objects have strength and desires to approach or posses the objects. Consumers' buying interest can be seen from many aspects. One of the aspects is sales. Sales are the important element in marketing divisions of a company to obtain more profits in order to continue the businesses. Kelumer Bayau is the company engaged in marketing and selling sofas with 8-year experience. Kelumer Bayau has a high quality with very competitive prices. The formulation of this research was how Kelumer Bayau predicted or determined the customers' buying interest on the sales of furniture based on recorded data. The prediction was based on the owners' decision to determine a number of goods provided by the company and to avoid the a large quantities of goods. Therefore, the objective of this research was predicting and determining the customers' buying interest using Naive Bayes Clasification algorithm through data mining techniques. This research used training and testing data about sofa sales for the last 3 years with the several variables i.e., types, colors, sizes, motifs and prices.

Keywords: Prediction, Naive Bayes Classification, Sofa

INTRODUCTION

Interest is the personal intention and related to attitude.Individuals who are interested in objects have strength and desires to approach or posses the objects. Buying interest is the consumers' tendency to buy a brand measured by the level of the possibility of consumers who buy something.

Consumers' buying interest can be seen from many aspects. One of the aspects

is sales. Sales are the important element in marketing divisions of a company to obtain more profits in order to continue the businesses. The goals of company are producing goods and services in order to meet the consumers' needs and to reduce unemployment.

The tighter business competition in this globalization era makes companies to rearrange their business strategies and tactics. The essence of competition lies in a company that can implement the production process better and faster than its business competitors or create different or unique products that cannot be produced by competitors. Therefore, the implementation of information technology and communication is needed in the field of business as a tool to win the competition, especially in the sale of products or services.

The problem faced by Bayau Kelumer Furniture is the formulation of this research was how Kelumer Bayau predicted or determined the customers' buying interest on the sales of furniture based on recorded data. The prediction was based on the owners' decision to determine a number of goods provided by the company and to avoid the a large quantities of goods. The effective long and short term planning depends on the prediction of demand for the company's products. The company was more facilitated in production scheduling on condition that this prediction was implemented in the part of the production planning process. It was because this prediction was able to provide the best output so that the expected risk of errors caused by planning errors was able to be minimized.

Kelumer Bayau Furniture produced sofas during 8-year experience with a high quality at very competitive prices. Kelumer Bayau Furniture needed standards, innovation, and ability to move quickly so that Kelumer Bayau Furniture was expected to be one of the best Furniture in Indonesia in producing Indoor Sofa. This research proposed the use of the Naive Bayes method to predict customers' buying interest on Sofa at Kelumer Bayau Furniture. The Naive Bayes Classifier algorithm had accuracy and speed in extracting knowledge in the database.

The previous research was done by Dicky (2016). This research used Naïve Bayes Clasifier with the aim at finding out customers' buying interests on XL sim card at Sumber Utama Telecommunikasi Ltd.).This research only used 3 attributes and 19 data sets. The added attributes and data were able to improve the level of accuracy to the system.

Another research was conducted by Felida (2017) using Naïve Bayes for predicting Furniture Production Time at Gajendra Furniture with the aim at producing accuracy by 79%. A further research was expected to optimize to produce higher accuracy.

In the previous studies, there were still weaknesses or lack of an implementation used in the certain attributes so that the system was still very weak to achieve maximum data accuracy. Due to the problem of this research, the researcher proposed "IMPLEMENTATION OF NAIVE BAYES CLASSIFIER ALGORITHM FOR DETERMINING CUSTOMERS' BUYING INTEREST ON SOFA. It was expected that this research was able optimize

the income and the advancement of Kelumer Bayau Furniture.

LITERATURE REVIEW

Data Mining

Data mining is the process to get useful information from the database warehouse. Data mining techniques are to trace existing data to build a model. This model is used to recognize the other data patterns that are not in the stored database. These patterns are recognized by certain devices that can provide accurate data information and provide insight which can also be used or studied more by the users.

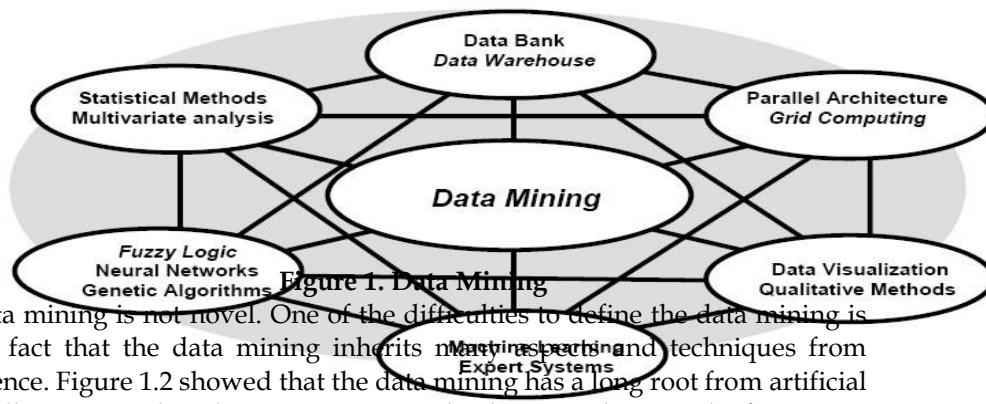

Figure 1. Data Mining

Data mining is not novel. One of the difficulties to define the data mining is the fact that the data mining inherits many aspects and techniques from science. Figure 1.2 showed that the data mining has a long root from artificial intelligence, machine learning, statistics, databases, and retrieval information. The data mining can be divided into several stages. These stages are correlated each other in which the user is directly involved with the knowledge base. These stages include:

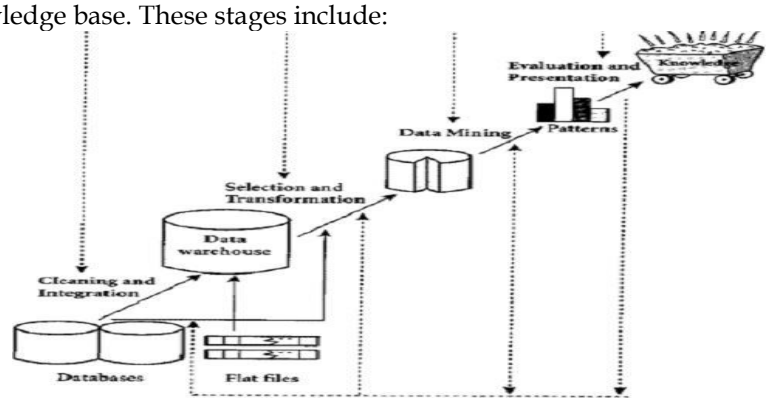

Figure 1.2 Stages of Data Mining

Stages of the data mining process are as follows:

a. Data Cleaning

The data obtained from company database and the result of experiments has imperfect entries i.e., missing data, invalid data, or erroneous typing. In addition, there are also data that are not relevant to the data mining hypothesis. The irrelevant data is also disposed because its existence can reduce the quality or accuracy of the data mining results later. Garbage is the term that is often used to describe this stage. Data cleaning also affects the data mining system performance because the obtained data can reduce the number and complexity.

b. Intergrating Data

Data integration is the step to conduct attributes that identify unique entities i.e., name, product type, customer number, etc. Data integration needs to be done carefully because errors in data integration can produce distorted results and mislead action. Data integration is based on a type of products but the data actually are combined from different categories so that there is a correlation between products that actually do not exist. In this data integration, the transformation and the data cleaning are needed and because the data of two databases are different or the writing is not the same or the data of a database is not in another database.

c. Data Transformation

Some data mining techniques need special formats of data before they are implemented. Some standard techniques such as association analysis and cluster can only accept categorical data input. Therefore, the data in the form of a numerical number that continues needs to be divided into several intervals. This process is often called binning. Here is also the selection of data needed by the data mining technique. This data transformation and selection also determine the quality of data mining results because there are some characteristics of certain data mining techniques that depend on this stage.

d. Data mining

Data mining is the process of finding interesting patterns or information in selected data using certain techniques or methods. The techniques, methods, or algorithms in data mining vary greatly. The choice of the right method or algorithm greatly depends on the overall KDD goals and processes.

e. Evaluation

In this stage the results of data mining techniques in the form of typical patterns and prediction models are evaluated to assess whether the existing hypothesis is indeed achieved. If it turns out that the results obtained are not according to the hypothesis there are several alternatives that can be taken such as: making feedback to improve the data mining process, trying other data mining techniques that are more appropriate, or accepting these results as unexpected results that might be useful.

The Naïve Bayes Classifier Algorithm

The Naive Bayes Classifier algorithm is based on probabilistic calculations with the assumption that each feature is independent. Naive Bayes Classifier is the text classification method that is quite popular to use and this algorithm has advantages in terms of learning speed and tolerance to the value lost from features. To handle numerical data, this algorithm uses a probability density function, meaning that the data is considered to follow the normal distribution to then calculate the average value and the standard deviation.

To represent a class, there are characteristics of the instructions needed to do a classification that is useful to explain that the probability of entering certain characteristics into the posterior class. Opportunities for the emergence of a class (before the entry of the sample, often called prior), multiplied by the probability of the emergence of global sample characteristics are also called evidence. The evidence value is always fixed for each class in one sample. The posterior value is compared to the other class's posterior values to determine what class a sample is.

The Naive Bayes classification is assumed that the existence or not of a particular feature of a class has nothing to do with the characteristics of other classes. The equation of Bayes's theorem is as follows:

By:

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

X = Data with unknown classes;

H = The data hypothesis X is a certain class label;

P (H | X) = Probabilistic hypothesis H based on the condition of X (posteriori probability);

P (H) = Probabilistic H hypothesis (prior probability);

P (X | H) = Probability X based on conditions in hypothesis H;

P (X) = Probabilistic X;

RESEARCH METHOD

This stages were explained about the data sources used and the data pre-processing stage.

1. Data Sources

The source of data used in this study was taken from the data of Kelumer Bayau Furniture sales in the last three years, namely from 2015 to 2017 and the chart of sales data for the last 3 years were seen in Figure 3.1.

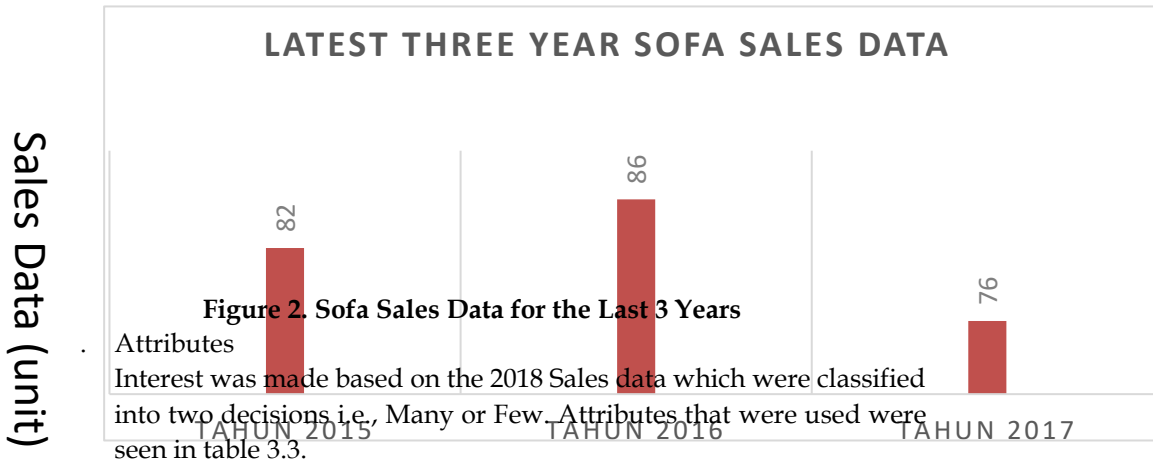

Figure 2. Sofa Sales Data for the Last 3 Years

. Attributes

Interest was made based on the 2018 Sales data which were classified into two decisions i.e., Many or Few. Attributes that were used were seen in table 3.3.

Table 1. Buyer Interest Attributes

Atribut	Keterangan
Sofa Type	Is a sofa model produced or sold by Bayau Kelumer Furniture.
Color	Is the color of the sofa itself.
Motif	Is a motif used by a sofa or not patterned (plain).
Size	Is the size of the sofa made, this attribute can be of normal or large value.
Price	Is the price of the sofa itself, these attributes are grouped into two, namely <= 3000000 and> 3000000
Interest	It is a statement of the customer's buying interest in the sofa, this class is grouped into two, which are many and few.

RESULTS AND DISCUSSION

Based on training data, data classification was calculated by managing the attributes or input data that had been predetermined, i.e., the type of sofa, color, motif, size, price using the naïve Bayes algorithm. The following was an example of a sofa test data that was not yet known about Customer Buying Interests on the sofa.Examples of test data were seen in table 1.2.

Tabel 2. Test Data

Type of sofa	Color	Motif	Size	Price	Interest

3 SeaterSofa	Black	None	Large	> 3000000	?

Based on the test data above, the results were determined by the following steps:

1. Calculating the Number of Classes

P (Interest | Lots) = 36/73"The amount of data for selling couches with buying interest was divided by total data".P (Little Interest) = 37/73. "The amount of data Sales of sofa with little buying interest was divided by the total data".

2. Calculating the Number of the Same Classes

P(type of sofa = 3 SeaterSofa | Interest = Big) = 12/36
P(color = black| Interests = Big) = 19/36
P(motif = None | Interest = Big) = 20/36
P(size = Large | Interest = Big) = 19/36
P(price = > 3000000 | Interest = Big) = 14/36
P(type of sofa = 3 SeaterSofa | Interest = Little) = 11/37
P(color = Black| Interest=Little) = 3/37
P(motif = None | Interest =Little) = 14/37
P(size = Large| Interest =little) = 17/37

P(price = > 3000000 | Interest =Little) = 24/37

3. Multiplying All Big and Little Attribute Results

P(type of sofa = 3 Seater Sofa | Interest =Big) x P(color = Black | Interest =Big) x P(motif = None | Interest =Big) x P(size = Large| Interest =Big) x P(price = > 3000000 | Interest = Big) x P(Interest|Big)
Big = 12/36 * 19/36 * 20/36 * 19/36 * 14/36 * 36/73
= 0.33333*0.52777*0.55555*0.52777*0.38888*0.50684
= 0.01016

P(type of sofa = 3 Seater Sofa | Interest = Little) x P(color = Black| Interest = Little) x P(motif = None | Interest= Little) x P(size = Large | Interest= Little) x P(price = > 3000000 | Interest= Little) x P(Interest|Little)
Little= 11/37 * 3/37 * 14/37 * 17/37 * 24/37 * 37/73
= 0.29729*0.08108*0.37837*0.45945*0.64864*0.49315
= 0.00134

Based on the results of the above calculation, it was seen that the probability value between (Lots of Interest) = 0.01016 and P (Little Interest) = 0.00134 was greater (P | Lots of Interest), so it was concluded that the product with 3 Seater Sofa, black , not patterned, large size, and the price of more than Rp. 3,000,000 that became in great demand by customers.

CONCLUSIONS AND RECOMMENDATIONS

Conclusion

After doing all the analysis, design, implementation, and evaluation of the system, some conclusions can be drawn as follows:

1. The Naïve Bayes Classifier algorithm can be used to determine the customer buying interest in the sofa.
2. The Naïve Bayes Classifier algorithm can be used as a basis in the decision-making process by Kelumer Bayau Furniture in terms of producing sofas.
3. Each variable / attribute in the Naïve Bayes Algorithm has the same opportunities in the classification and prediction process.
4. Based on the testing that has been done, the Naïve Bayes Classifier algorithm is accurate enough to be implemented in the case of prediction or classification because it has an accuracy rate of 82%.

Suggestion

Based on the conclusions, there are some suggestions as follows:

1. The algorithm used can be developed again by comparing other algorithms such as K-Nearst Neighor, Neural Network, and so on.
2. Increase the amount of data training to increase the level of accuracy.

REFERENCES

[1] Anhar. 2010. *PHP & MySql Secara Otodidak*. Jakarta: PT TransMedia.

[2] Arief M Rudianto. 2011. *Pemrograman Web Dinamis menggunakan PHP dan MySQL.* C.V ANDI OFFSET. Yogyakarta.

[3] Artaye, Ketut (2015). *Implementation of Naïve Bayes Classification method to predict graduation time of IBI Darmajaya Scholar.*

[4] Felida, Naifiri Novitasari (2017). *Prediksi waktu Produksi Mebel menggunakan Metode Naïve Bayes.*

[6] Kusrini, & Luthfi, E. T. (2009). *Algoritma Data Mining*. Yogyakarta: C.V Andi Offset.

[7] Nurjoko (2016). *Aplikasi Data Mining Untuk Memprediksi Tingkat Kelulusan Mahasiswa Menggunakan Algoritma Apriori Di IBI Darmajaya.*

[8] Prasetyo, Eko (2014). *Data Mining Mengolah Data Menjadi Informasi Menggunakan Matlab.*