Printer User Data Classification Uses the Naïve Bayes Algorithm and the C4.5 Algorithm

Rahmat Rian Hidayat[1], Saruni Dwiasnati[2]
[1,2]Faculty of Computer Science, Universitas Mercu Buana, Jl. Meruya Selatan, RT.4/RW.1, Meruya Sel., Kec. Kembangan, Kota Jakarta Barat, Daerah Khusus Ibukota Jakarta 11650
Email address: rahmat.rian @ mercubuana.ac.id[1], saruni.dwiasnati @ mercubuana.ac.id[2]

Abstract— *The printer is a hardware device used for printing. In the world of computers, printers include peripheral output devices that present text or graphical representations on paper or similar media. Peripheral devices can indeed be in the form of internal or external, but usually the term is more often addressed to an external device that is connected to a computer directly such as a printer. Even so, some devices inside a computer are still referred to as peripherals. The background of this research is due to the buildup of a printer in the division that actually does not require a lot of printers stored in the room. This research was conducted to classify printer user data using Data Mining Techniques that serve to help a division to determine the number of printers that must be available in a division to minimize the buildup of printers that are not used. Data mining techniques used in this study are the Naïve Bayes algorithm and the C4.5 algorithm to get which algorithm has an accuracy rate better than other algorithms. C4.5 algorithm is a group of algorithms using decision trees. The decision tree is a very powerful and well-known classification and prediction method. The richer the information or knowledge contained by the training data, the accuracy will increase. Naïve Bayes Bayesian classification algorithm is a statistical classification that can be used to predict the probability of membership of a class. The data used for this study were collected from January to December in 2018.*

Keywords— *Printer, Naïve Bayes Algorithm, Decision Tree, C4.5 Algorithm.*

I. Introduction

Computers are composed by several devices that are used to make computers better and function more optimally. One of them is called a peripheral device. Peripheral device is a device that is connected to a computer that is external or internal. The printer is a device used for printing, print that is taken from word print which means print. In the electronic world, especially relating to computer printers, it is classified as peripherals output that is useful for printing documents, images, and other things. The electric printer was first made in 1968 in Japan by Epson with the initial name EP-101 while the printer design was first made during the 19th century mechanically. During the 1980s the need for a higher speed printer performance when printing more documents was needed so that several printer renewal systems emerged including the line printer, dot matrix, and daishy wheel. The concept of line printers is to have the same output as a typewriter but process faster. Daishy wheel has almost the same concept as a typewriter. Whereas the dot matrix concept is to produce a combination of graphics and writing output but the quality of output or printouts is low. To produce good prints, usually people use plotter instead of printer. Over time, precisely in 1984 came the laser printer which was originally named the HP LaserJet. This printer has additional PostScript features made by Apple LaserWritter which is then better known as desktop publishing. The PostScript concept combines graphics and writing the same as the dot matrix concept but prints better.

Printer supplies in each division on each floor need to be seen every day, whether the use of printers in the division and the floor in accordance with the procurement of the number of printers available. Is the printer used as needed or just put the printer on a desk and not used. Prediction errors in the placement of printers in a division or floor result in stacking printers that should be used optimally in accordance with the needs of each division and floor. This hoarding causes waste of printer devices which are only placed on the table. In this study the data used to determine whether the division needs to be added or not the printer device for each department on each floor.

Data mining is a term used to describe the discovery of knowledge in a database or often called Knowledge Discovery in a database [1]. Data mining is used for rule calculation, the data serves to find the feasibility of the available data sets. Data mining has several parts one of the methods is the classification method. The classification method can work by putting data that has been compiled into better data into one of several classes that have been previously defined. For example classifying data on areas affected by a storm or not affected by a storm to find out the category of the area. Then from the data will be classified into different classes according to predetermined categories. From the classification data, classes are obtained from the decision tree (decission tree). The advantage of the decision tree is that it can produce rules that are easily implemented quickly. Decision tree can generated by a variety of methods, one of which is the C4.5 algorithm so that data processing is computerized using data mining techniques. The algorithm that will be used in this research is Naïve Bayes algorithm and C4.5 algorithm. This algorithm will determine the factors that indicate the item needs to be added to the stock or not. In this study using 2 algorithm C4.5 algorithm and Naïve Bayes algorithm, C 4.5 algorithm has been explained in the previous discussion, Naïve Bayes algorithm is one of classification techniques in data mining[2]. Where the analysis will be carried out to obtain information on old data on the willingness of printers in a department in the company.

It is expected that from the research conducted on the printer data samples can be obtained information that can help the company in making decisions whether the department for the availability of printers needs to be added or whether the printer that is available needs to be transferred to the department that loads it. The problem formulation in this research is to determine the decision making method in placing a printer in a department to classify printer data using the Naive Bayes Algorithm and C 4.5 Algorithm to produce accuracy values which are better than the two algorithms that can be used by companies.

II. Literature Review

The problem raised from this research has two problems that can be identified, namely how the process of grouping data to produce printer data accuracy values within a year backward and needs to be done to check which divisions require printer addition or subtraction. The method used is the classification method with the Naïve Bayes algorithm and the C4.5 algorithm.

A. Definition of Data Mining

Data mining or data mining is between branches of computer science and statistics, by utilizing the progress of both disciplines to extract information from large databases [3]. Data mining is a series of processes to explore the added value of knowledge that has not been known from a data set. Data mining is a process that uses statistical techniques, mathematics, artificial intelligence, and machine learning to extract and identify useful information and related knowledge from various large databases [4].

Systematically, there are three main steps in data mining:

a. Data exploration / initial processing

Data exploration or initial processing consists of data cleaning, data normalization, data transformation, incorrect data handling, dimension reduction, selection of feature subsets, and so on.

b. Build models and validate

Building a model and validating it means analyzing various models and selecting the model with the best predictive performance. In this step methods such as calcification, regression, cluster analysis, anomaly detection, association analysis, sequential pattern analysis are used.

c. Application

Implementation means determining the model in the data so as to produce a prediction of the problem under investigation.

B. Definition of Decision Tree

Decision Tree is an algorithm that is included in supervised learning. It is said supervised learning because in the process of grouping this algorithm uses initial data to form decision rules. Thus, the data that has been collected must already have a group label.

The decision tree algorithm is as follows [5]:

1. Select the attribute as the root
2. Create a branch for each value
3. Divide cases in branches
4. Repeat the process for each branch until all cases in the branch have the appropriate class

To choose an attribute as the root, based on the highest gain value of the existing attributes. The formula to calculate the gain is as follows:

$$\text{Gain(S,A)} = \text{Entropy(S)}\text{-}\sum_{i=1}^{n}\frac{|S_i|}{|S|}Entropy\ (S_i)$$

Where:
S = Data set universe
A = Attribute
N = Number of partition attribute A
| Si | = Number of cases in the i-th partition
| S | = Number of cases in S

C. Naïve Bayes Algorithm

Naïve Bayes is a statistical classification model that can be used to predict the probability of membership of a class. Naïve Bayes is based on the Bayes theorem which has classification capabilities similar to decision trees and neural networks. The Naive Bayes algorithm predicts future opportunities based on past experience so it is known as the Bayes Theorem. The main characteristic of the Naïve Bayes Classifier is a very strong assumption (naive) of independence from each condition / event. The advantage of using this algorithm is that this method only requires a small amount of training data to determine the estimated parameters needed in the classification process to get a decision.

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

Where:
X = data with an unknown class
H = data hypothesis X is a specific class
P (H | X) = probability of hypothesis H based on condition X (probabilistic post theory)
P (H) = probability of hypothesis H (prior probabilistic)
P (X | H) = probability of X based on these conditions
P (X) = probability of X

B. C 4.5 Algorithm

C4.5 algorithm is a group of Decision Tree algorithms. This algorithm has input in the form of training samples and samples. Training samples are sample data that will be used to build a tree that has been tested for correctness. While samples are data fields which will be used as parameters in classifying data [3]. Some The development carried out in C4.5 is, among others, able to overcome the missing value, can overcome the continuing data, and pruning.

III. Method and Materials

A. Identification

This research is motivated because there is a buildup of a printer in the division that actually does not require a lot of printers there. This research was conducted to classify printer user data using Data Mining Techniques that serve to help a division to determine the number of printers. Data mining techniques used in this study are Naïve Bayes algorithm and C4.5 algorithm to get which algorithm has better accuracy than other algorithms. The data used for this study were collected from January to December in 2018. According to research conducted by Fajar Edi Prabowo and Achmad Kodar [4] conducting research to find prediction results on the graduation data of Mercu Buana University Informatics Engineering students using the Naïve Bayes classification method. Sunjana conducts research to find patterns of customer status to be used as material for company analysis in determining prospective customers in the future [5]. Other studies on prediction of preterm births conducted by Ari using the C4.5 algorithm based on Particle Swarm Optimization. This study uses 250 record data records and produces an accuracy rate of 93.60% with the C4.5 algorithm, while the PS4 based C4.5 produces an accuracy of 96.00%. Optimization of C4.5. C4.5 algorithm based on Praticle Swarm Optimization (PSO) has the highest accuracy rate of 96% compared to the other two algorithms [6]. While other studies aim to optimize features in the classification of data mining with the C4.5 algorithm using Particle Swarm Optimization (PSO) to detect blood sugar levels in patients. The dataset used is the Effect of Physical Activity on Blood Sugar Levels in H.Abdul Manan Simatupang Hospital Range. The dataset used was 42 records with 10 attributes. The results of this study found that Particle Swarm Optimization (PSO) can improve C4.5 accuracy performance from 86% to 95% [7].

There are several steps applied in this research.

3.1. Data retrieval

Retrieval of data in this study using data from related divisions in the object of this study. The use of Data Mining is very easy for researchers who want to be obtained. The services provided by Data Miner also allow data retrieval to be exported to the desired file format.

3.2. Data Cleaning

After retrieving data from the relevant division, the data cannot be directly entered in processing to predict the increase in the number of printers in a division, then proceed with the Data Cleaning stage.

3.3 Data Mining

The modeling for this study is illustrated in full as shown in Figure 1. Which is the overall operator used in research on rapidminer software.

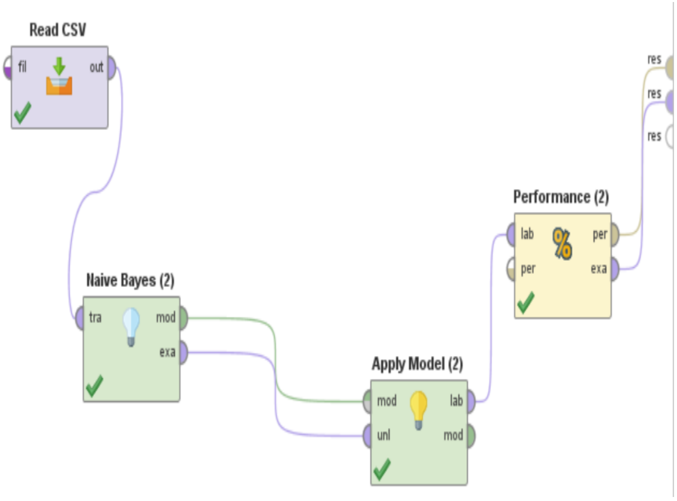
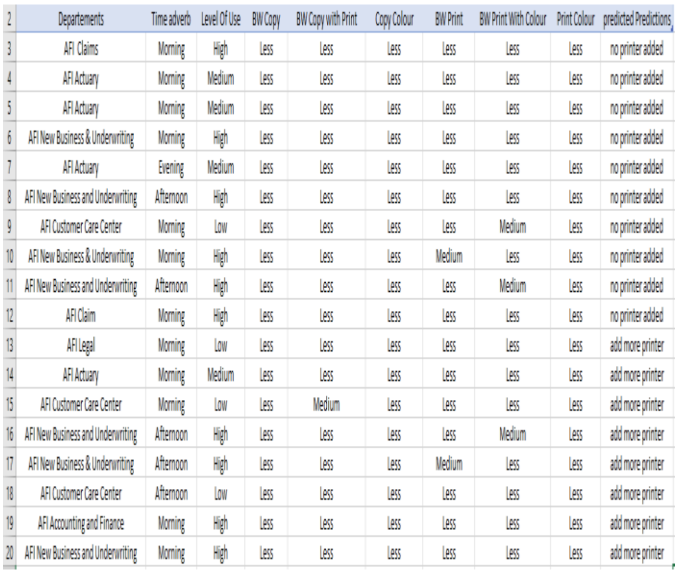


Figure 1. Modeling Research in Rapidminer

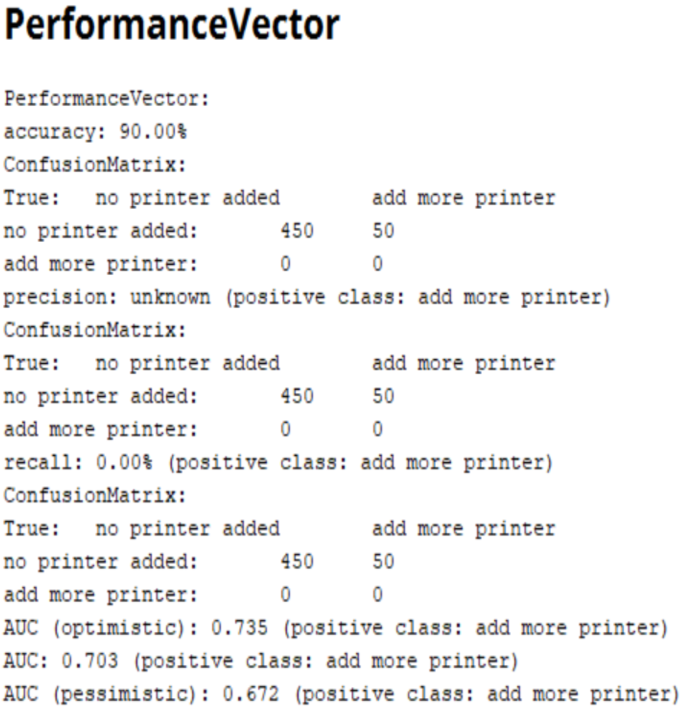
In the above modeling obtained from a data set of 500 records, and several variables that support to produce it, as below:



2	Departements	Time adverb	Level Of Use	BW Copy	BW Copy with Print	Copy Colour	BW Print	BW Print With Colour	Print Colour	predicted Predictions
3	AFI Claims	Morning	High	Less	Less	Less	Less	Less	Less	no printer added
4	AFI Actuary	Morning	Medium	Less	Less	Less	Less	Less	Less	no printer added
5	AFI Actuary	Morning	Medium	Less	Less	Less	Less	Less	Less	no printer added
6	AFI New Business & Underwriting	Morning	High	Less	Less	Less	Less	Less	Less	no printer added
7	AFI Actuary	Evening	Medium	Less	Less	Less	Less	Less	Less	no printer added
8	AFI New Business and Underwriting	Afternoon	High	Less	Less	Less	Less	Less	Less	no printer added
9	AFI Customer Care Center	Morning	Low	Less	Less	Less	Less	Medium	Less	no printer added
10	AFI New Business & Underwriting	Morning	High	Less	Less	Less	Medium	Less	Less	no printer added
11	AFI New Business and Underwriting	Afternoon	High	Less	Less	Less	Less	Medium	Less	no printer added
12	AFI Claim	Morning	High	Less	Less	Less	Less	Less	Less	no printer added
13	AFI Legal	Morning	Low	Less	Less	Less	Less	Less	Less	add more printer
14	AFI Actuary	Morning	Medium	Less	Less	Less	Less	Less	Less	add more printer
15	AFI Customer Care Center	Morning	Low	Less	Medium	Less	Less	Less	Less	add more printer
16	AFI New Business and Underwriting	Afternoon	High	Less	Less	Less	Less	Medium	Less	add more printer
17	AFI New Business & Underwriting	Afternoon	High	Less	Less	Less	Medium	Less	Less	add more printer
18	AFI Customer Care Center	Afternoon	Low	Less	Less	Less	Less	Less	Less	add more printer
19	AFI Accounting and Finance	Morning	High	Less	Less	Less	Less	Less	Less	add more printer
20	AFI New Business and Underwriting	Morning	High	Less	Less	Less	Less	Less	Less	add more printer

Figure 2. Data Set

By using Rapidminner studio from data processing, the results are obtained, as follows:



PerformanceVector

```
PerformanceVector:
accuracy: 90.00%
ConfusionMatrix:
True:   no printer added        add more printer
no printer added:       450     50
add more printer:       0       0
precision: unknown (positive class: add more printer)
ConfusionMatrix:
True:   no printer added        add more printer
no printer added:       450     50
add more printer:       0       0
recall: 0.00% (positive class: add more printer)
ConfusionMatrix:
True:   no printer added        add more printer
no printer added:       450     50
add more printer:       0       0
AUC (optimistic): 0.735 (positive class: add more printer)
AUC: 0.703 (positive class: add more printer)
AUC (pessimistic): 0.672 (positive class: add more printer)
```

Figure 3. Performance Vector

From the test data above, the AUC curve will be produced by producing the results as explained below:

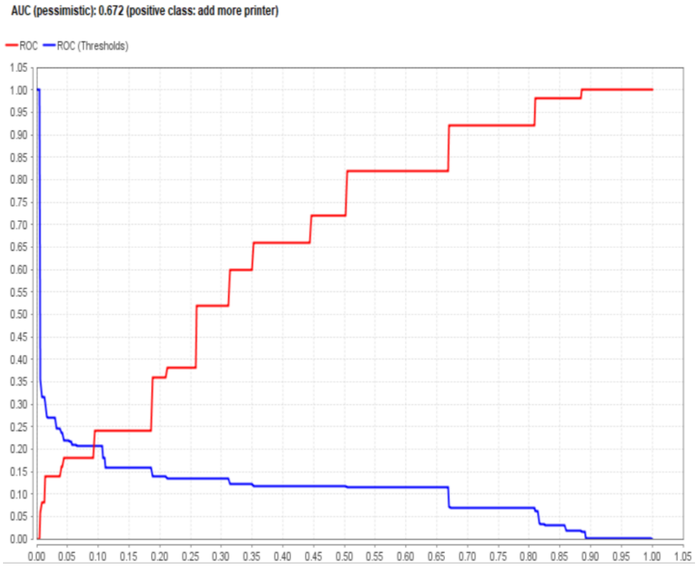
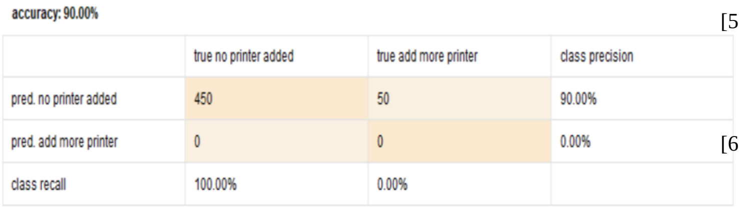


Figure 4. AUC curve

IV. Result and Discussion



accuracy: 90.00%

	true no printer added	true add more printer	class precision
pred. no printer added	450	50	90.00%
pred. add more printer	0	0	0.00%
class recall	100.00%	0.00%	

Figure 5. Accuracy uses the Naïve Bayes Algorithm

The picture above is the calculation of accuracy data using the Naïve Bayes algorithm. It is known that the training data consists of 500 data records, with the resulting accuracy value of 90% with a maximum prediction of the printer does not need to be added.

Classification is a data processing technique that divides objects into several classes according to the number of classes desired [8].

V. Conclusion

Based on the research that has been done, it can be concluded that this research is going well, in the search whether the division needs to add a printer or not using the Naïve Bayes algorithm.

References

[1] Gupta, G. dan H. Anggarwal, “Improving Customer Relationship Management Using Data Mining”, International Journal of Machine Learning and Computing, 2 (6), 874-877, 2012

[2] Alvino Dwi Rachman Prabowo, Muljono,” Prediksi Nasabah Yang Berpotensi Membuka Simpanan Deposito Menggunakan Naive Bayes Berbasis Particle Swarm Optimization”, Techno.COM, Vol. 17, No. 2, 208-219, 2018

[3] Data mining Terapan. Surabaya: Graha Ilmu. Sunjana, 2010. Seminar Nasional Aplikasi Teknologi Informasi 2010. Snati 2010

[4] Fajar Edi Prabowo, Achmad Kodar,” Analisis Prediksi Masa Studi Mahasiswa Menggunakan Algoritma Naïve Bayes”, Jurnal Ilmu Teknik dan Komputer, Vol. 3 No. 2 Juli, 2019, ISSN 2548-740X E-ISSN 2621-1491

[5] Rama Aji Pangestu, Sabar Rudiarto, Devi Fitrianah, “Aplikasi Web Berbasis Algoritma K-Nearest Neighbour Untuk Menentukan Klasifikasi Barang Studi Kasus: Perum Perur”, Jurnal Ilmu Teknik dan Komputer, Vol. 2 No. 1 Januari, ISSN 2548-740X E-ISSN 2621-1491,9-19,2018

[6] Puspita Ari., 2016. Prediksi Kelahiran Bayi Secara Prematur dengan Menggunakan Algoritma C4.5 Berbasis Particle Swarm Optimization. Jurnal Teknik Informatika STMIK Antar Bangsa.Vol. II, pp.11-16.

[7] Dwi Meylitasari Br.Tarigan, Dian Palupi Rini, Samsuryadi,” Seleksi Fitur pada Klasifikasi Penyakit Gula Darah Menggunakan Particle Swarm Optimization (PSO) padaAlgoritma C4.5”, Jurnal Resti (Rekayasa Sistem dan Teknologi Informasi) Vl. 4 No. 3, 569-575 ISSN Media Elektronik: 2580-076, Hal 569 -575. 2020

[8] Muhammad Fauzul Arifin, Devi Fitrianah,”Penerapan Algoritma Klasifikasi C4.5 dalam Rekomendasi Penerimaan Mitra Penjualan Studi Kasus : PT Atria Artha Persada”, IncomTech, Jurnal Telekomunikasi dan Komputer, vol.8, no.2, 2018.