Design of expert system for owning motorcycle with Naive Bayes classifier

M Irfan, W B Zulfikar, C N Alam and M A Ramdhani*
Department of Informatics, UIN Sunan Gunung Djati Bandung, Bandung, Indonesia

*m_ali_ramdhani@uinsgd.ac.id

Abstract. Nowadays, motorcycle is widely used by many people. There are several major categories of motorcycles in Indonesia such as automatic gear (matic), moped, and sport. Each motorcycle has a lot of types; even each type may be divided into several sub types. Generally, a sub-type of motorcycle is distinguished by features and spare parts that determine the sales price of motorcycle. Therefore, customers have trouble to determine the type of motorcycle that suits their needs and wishes. This article proposes a model that helps customer decide the motorcycle in accordance with the following criteria: gender, occupation, income, and age. This model compares and calculates a new customer with many previous customers with Naive Bayes Classifier. In evaluation phase, this model shows that 89.12% of customers feel satisfied. It means that this method valuable implement on this model. In further work, customer’s advance criteria like color, durability, and resale price should increase the customer’s satisfaction.

1. Introduction
Nowadays, we discover several major categories of motorcycles in Indonesia such as automatic gear (matic), moped, and sport. Each motorcycle has a lot of types; even each type may be divided into several sub types. Generally, a sub-type of motorcycle is distinguished by features and spare parts that determine the sales price of motorcycle. Therefore, customers have trouble to select the type of motorcycle that suits their needs and wishes [1][2][4].

Technological devices are designed to enhance a quality of human’s life [5][6], one of those which are enable efficiency and effectiveness in business process within a field of decision making is information systems [7]. This systems is a combination of information technology utilizations and human activity upon a set of agreed procedures, IS has a high level of flexibilities to develop and scalable [8], expert systems is generally known as a branch of IS applications.

Expert Systems, is defined as a system that organized data process [9], it adopts several knowledge of experts into a computational code so the computer enable to solve a specific problem and acts as an expert [10]. Refers to several research, an expert system has a high capability in decision making, the system has an accurate data accessibility and efficient run-time [11], high accuracy [12][13], and to support a proper decision [14], low cost [15], extended accessibility [16], intensify user knowledge [17], increase productivity [18], provide a better data and information [19], and in the certain cases are potentially used as data storage [20].

This article proposes a model that helps customer decide the motorcycle in accordance with the following criteria: gender, occupation, income, and age. This model compares and calculates a new
customer with many previous customers with Naive Bayes Classifier. In previous work, this method have a good performance in order to classify several condition and situation [21], [22]. Technically, this work try to cluster several determinat factor as follow income and age in order to fulfill requirement model. Income values as well as age, has a diverse value. Therefore, this work classify income into 3 class as follow low, medium, and high. Beside it, this work classify age into 2 class as follow young and old.

2. Literature review

2.1. Knowledge discovery in database
Pattern recognition is a discipline that studies the ways in which objects are classified into classes or categories and recognizes data trends. Depending on the application of these objects may be patients, students, credit applicants, images or signals or other measures that need to be classified or sought for regression. Generally, this subject is called pattern recognition or pattern recognition [23]–[26][27].

Data mining, also known as knowledge discovery in database (KDD), is an activity that includes collecting, using historical data to find regularities, patterns or relationships in large data sets. One of outputs from data mining can be used to improve future decision making. Therefore, the term pattern recognition is now rarely used because it includes part of the data mining. Language intended for clustering, classification, regression, variable selection and market basket analysis or association rules [23][25][28][29], [30].

2.2. Naive Bayes classifier
Naive Bayes Classifier is one of the classification algorithms. This algorithm has been implemented in various fields and gives good results [21], [22]. The advantage is that this method requires only a small amount of training data to determine the estimated parameters required in the especially classification process. Since it is assumed to be an independent variable, only the variance of a variable in a class is required to determine the classification, not the whole of the covariance matrix [11].

\[
P(H | X) = \frac{P(X | H)P(H)}{P(X)} \tag{1}
\]

\begin{align*}
X & = \text{data with unknown class} \\
H & = \text{hypothesis data } X, \text{ one specific class} \\
P(H|X) & = \text{probability hypothesis } H \text{ based on } X \text{ condition (posteriori probability)} \\
P(H) & = \text{probability hypothesis } H \text{ (prior probability)}
\end{align*}

Naive bayes classifier has a common form as in (1) as on a bayesian basis. This algorithm has a similar capability to the decision tree and neural network [31].

3. Related Works

In previous works, naive bayes classifier was implementing on several scope like literature and health[21], [22][32]. This method have a good performance and high number of accuration. Therefore, this method also implement into several text mining work text especially detect swear word, classify hate speech and sentiment analysis [33][34][35][36].

4. Method

4.1. Business understanding phase
According to observation result, there are several major categories of motorcycles in Indonesia such as automatic gear, moped, and sport. Each motorcycle has a lot of types; even each type may be divided into several sub types. Generally, a sub-type of motorcycle is distinguished by features and spare parts that determine the sales price of motorcycle. Therefore, customers have trouble to determine the type of
motorcycle that suits their needs and wishes. Based on data training, gender, occupation, income, and age affected to class label. Sample of data training show on following table.

Table 1. Sample of data training.

| No. | Name   | Gender | Occupation     | Income  | Age | Owning                  |
|-----|--------|--------|----------------|---------|-----|-------------------------|
| 1   | Marna  | M      | Traders        | 3.000k  | 26  | Scoopy Sporty FI        |
| 2   | Subai  | M      | Traders        | 3.000k  | 58  | Beat All new FI CW      |
| 3   | Deni   | M      | Private employees | 1.300k | 32  | New Supra X 125 CW      |
| 4   | Juanti | F      | Housewife      | 5.000k  | 41  | Beat All New FI CW      |
| 5   | Suganda| M      | Traders        | 4.000k  | 49  | New Supra X 125 FI CW   |

Table 1 describe 5 of all data training with two value gender as follow : male and female, several occupation, various income, age, and their choiced motorcycle.

4.2. Data understanding and preparation phase
In data understanding and preparation phase, this work try to cluster several determinant factor as follow income and age in order to fullfill requirement model. Income values as well as age, has a diverse value. Therefore, this work classify income into 3 class as follow low, medium, and high. Beside it, this work classify age into 2 class as follow young and old.

Table 2. Clustering of Income.

| No. | Range                  | Class |
|-----|------------------------|-------|
| 1   | Rp. 1,700k - Rp. 2,300k | Low   |
| 2   | Rp. 2,400k - Rp. 4,000k | Medium|
| 3   | > Rp. 4,000,000        | High  |

Table 3. Clustering of age.

| No. | Range | Class |
|-----|-------|-------|
| 1   | 17 - 32 | Young |
| 2   | > 33   | Old   |

In data understanding and preparation phase this work present the statistics of data training including all determinant factor as follow gender, occupation, income, and age that show on following figure.

Figure 1. Statistics of gender.
Figure 2. Statistics of age.
4.3. Modeling phase

In modeling phase, this model implement 4 determinant as follow gender, age, occupation, and salary. Based on data training, this model obtained 3 class as follow moped, automatic, and sport. Each of class contains several sub class or variant. This model proposes 17 class label as describe on following table.

![Figure 3. Statistics of income.](image)

![Figure 4. Statistics of occupation.](image)

| No. | Class                | Variant                          |
|-----|----------------------|----------------------------------|
| 1   | Moped                | New Supra X 125 CW               |
| 2   | Moped                | Revo Fit 110                     |
| 3   | Moped                | Blade                            |
| 4   | Automatic            | Revo New Absolute CW             |
| 5   | Automatic            | New Supra X 125 MMC              |
| 6   | Automatic            | Scoopy Sporty FI                 |
| 7   | Automatic            | Beat All new FI CW              |
| 8   | Automatic            | New Vario ABS Matic              |
| 9   | Automatic            | Beat All New CW                 |
| 10  | Automatic            | New Vario CW                    |
| 11  | Automatic            | Vario Techno 125 CBS             |
| 12  | Automatic            | CBR 150 R STD                   |
| 13  | Sport                | Verza 150 CW                     |
| 14  | Sport                | CB 150 Street Fire MMC           |
| 15  | Sport                | CBR 150 Repsol                   |
| 16  | Sport                | CBR 150 R STD                   |
As example, we have a customer with following attribute: a male, 36 years old (classified old), buruh as his occupation, with income per month Rp. 3,000,000 (classified as medium). According to (1), this work calculate prior probability each class label.

\[
P(\text{Purchase} = \text{scoopy sporty fit}) = \frac{4}{51} = 0.07
\]

\[
P(\text{Purchase} = \text{beat all new fi cw}) = \frac{9}{51} = 0.17
\]

\[
P(\text{Purchase} = \text{new supra x 125 cw}) = \frac{2}{51} = 0.03
\]

\[
P(\text{Purchase} = \text{revo fit 110}) = \frac{8}{51} = 0.15
\]

\[
P(\text{Purchase} = \text{beat all new fi sporty}) = \frac{4}{51} = 0.07
\]

\[
P(\text{Purchase} = \text{cb 150 R streetfire}) = \frac{1}{51} = 0.01
\]

\[
P(\text{Purchase} = \text{vario techno 125 cbs}) = \frac{1}{51} = 0.01
\]

\[
P(\text{Purchase} = \text{new vario ABS matic}) = \frac{3}{51} = 0.05
\]

\[
P(\text{Purchase} = \text{beat all new fi}) = \frac{1}{51} = 0.01
\]

\[
P(\text{Purchase} = \text{new vario cw}) = \frac{2}{51} = 0.03
\]

\[
P(\text{Purchase} = \text{verza 150 cw}) = \frac{3}{51} = 0.05
\]

\[
P(\text{Purchase} = \text{beat cw fi sporty}) = \frac{1}{51} = 0.01
\]

\[
P(\text{Purchase} = \text{CB 150 streetfire MMC}) = \frac{1}{51} = 0.01
\]

\[
P(\text{Purchase} = \text{new supra x 125 MMC}) = \frac{1}{51} = 0.01
\]

\[
P(\text{Purchase} = \text{revo new absolut cw}) = \frac{1}{51} = 0.01
\]
\[
P(Purchase = CB\ 150\ STD) = \frac{1}{51} = 0.01
\]

\[
P(Purchase = blade\ cw) = \frac{1}{51} = 0.17
\]

Next step, this work calculate the same value of each determinant with the same class label. For example, we used income and the formula describe on (19)-(35).

\[
P(I = Medium\ and\ P = scoopy\ sporty\ fit) = \frac{2}{4} = 0.50
\]

\[
P(I = Medium\ and\ P = beat\ all\ new\ fi\ cw) = \frac{2}{4} = 0.55
\]

\[
P(I = Medium\ and\ P = new\ supra\ x\ 125\ cw) = \frac{0}{2} = 0
\]

\[
P(I = Medium\ and\ P = revo\ fit\ 110) = \frac{1}{8} = 0.12
\]

\[
P(I = Medium\ and\ P = beat\ all\ new\ fi\ sporty) = \frac{5}{9} = 0.55
\]

\[
P(I = Medium\ and\ P = cb\ 150\ R\ streetfire) = \frac{1}{1} = 1.00
\]

\[
P(I = Medium\ and\ P = vario\ techno\ 125\ cbs) = \frac{0}{1} = 0
\]

\[
P(I = Medium\ and\ P = new\ vario\ ABS\ matic) = \frac{1}{3} = 0.30
\]

\[
P(I = Medium\ and\ P = beat\ all\ new\ fi) = \frac{1}{1} = 1.00
\]

\[
P(I = Medium\ and\ P = new\ vario\ cw) = \frac{0}{2} = 0
\]

\[
P(I = Medium\ and\ P = verza\ 150\ cw) = \frac{0}{3} = 0
\]

\[
P(I = Medium\ and\ P = beat\ cw\ fi\ sporty) = \frac{1}{1} = 1
\]

\[
P(I = Medium\ and\ P = CB\ 150\ streetfire\ MMC) = \frac{0}{1} = 0
\]

\[
P(I = Medium\ and\ P = new\ supra\ x\ 125\ MMC) = \frac{0}{1} = 0
\]
Next step, permutation value of each class label will accumulate by each class label, the largest value means that this case has same class. The table below shows that beat all new fi cw obtained the largest result, it means that the current customer will owning beat all new fi cw.

| No. | Class Label          | Value       |
|-----|----------------------|-------------|
| 1   | Beat all new fi cw   | 0.002715    |
| 2   | Scoopy sporty fi     | 0.002188    |
| 3   | Revo fit 110         | 0.000563    |
| 4   | CB 150 R Streetfire  | 0           |
| 5   | Beat cw fi sporty   | 0           |

5. Result and discussion
This model compares and calculates a new customer with many previous customers with Naive Bayes Classifier. In this phase, we applied 2529 new customer and we obtained 2254 of 2529 or 89.12% of customers agree and feel satisfied.

6. Conclusion
This model compares and calculates a new customer with many previous customers with Naive Bayes Classifier. In evaluating phase, this model shows that 89.12% of customers agree and feel satisfied. We conclude that this model is not optimal yet. According to fig 4., there are many kinds of occupation that may affected to model’s accurate value. In further work, customer’s advance criteria like color, durability, and resale price should increase the customer’s satisfaction.

References
[1] T Ryanthi, B Suharjo and I Muflikhati 2016 perilaku konsumen terhadap keputusan pemilihan merk sepeda motor J. Apl. Bisnis dan Manaj. 2(1) p. 23
[2] Y Sudarno and S W Utomo 2014 Pengaruh Kualitas Produk, Harga, Iklan Terhadap Keputusan Pembelian Pada Dealer Mpm Motor Madiun Assets J. Akunt. dan Pendidik. 3(1) p. 35
[3] A Asna and A Asna Analisis Perilaku Konsumen Dalam Keputusan Pembelian Sepeda Motor Pada Mahasiswa Se-Malang Raya J. Ekon. Mod. 6(2) pp. 130–139
[4] N K L Putri, N M Asih and D P E Nilakusmawati 2015 Faktor-Faktor Yang Menentukan Kepuasan Pelanggan Sepeda Motor Matic Honda Di Kota Denpasar 2015 E-Jurnal Mat. 4(1) p.1
[5] M A Ramdhani, H Aulawi, A Ikhwana and Y Mauluddin 2017 Model of green technology adaptation in small and medium-sized tannery industry J. Eng. Appl. Sci. 12(4) pp. 954–962
[6] W B Zulfikar, Jumadi, P K Prasetyo and M A Ramdhan 2018 Implementation of Mamdani Fuzzy Method in Employee Promotion System IOP Conf. Ser. Mater. Sci. Eng. 288(1) p. 12147
[7] A Pamoragung, K Suryadi and M A Ramdhan 2006 Enhancing the implementation of e-Government in Indonesia through the high-quality of virtual community and knowledge portal
in *Proceedings of the European Conference on e-Government ECEG* pp. 341–348

[8] H Aulawi, M A Ramdhani, C Slamet, H Ainissyifa and W Darmalaksana 2017 Functional Need Analysis of Knowledge Portal Design in Higher Education Institution *Int. Soft Comput.* **12**(2) pp. 132–141

[9] D S Maylawati, W Darmalaksana and M A Ramdhani 2018 Systematic Design of Expert System Using Unified Modelling Language *IOP Conf. Ser. Mater. Sci. Eng.* **288**(1) p. 12047

[10] M A Ramdhani 2013 Metodologi Penelitian untuk Riset Teknologi Informasi (Bandung: UIN Sunan Gunung Djati Bandung)

[11] C Slamet, A Rahman, A Sutedi, W Darmalaksana, M A Ramdhani and D S Maylawati 2018 Social Media-Based Identifier for Natural Disaster *IOP Conf. Ser. Mater. Sci. Eng.* **288**(1) p. 12039

[12] C Slamet, R Andrian, D S Maylawati, W Darmalaksana and M A Ramdhani 2018 Web Scraping and Naive Bayes Classification for Job Search Engine *IOP Conference Series: Materials Science and Engineering* **288**, no. 1, pp. 1–7, 2018.

[13] W B Zulfikar, A Wahana, W Uriawan and N Lukman 2016 Implementation of association rules with apriori algorithm for increasing the quality of promotion in *Proceedings of 2016 4th International Conference on Cyber and IT Service Management CITSM*

[14] Y A Gerhana, W B Zulfikar, A H Ramdani and M A Ramdhani 2018 Implementation of Nearest Neighbor using HSV to Identify Skin Disease in *IOP Conference Series: Materials Science and Engineering* **288**(1)

[15] A Rahman, C Slamet, W Darmalaksana, Y A Gerhana and M A Ramdhani 2018 Expert System for Deciding a Solution of Mechanical Failure in a Car using Case-based Reasoning *IOP Conf. Ser. Mater. Sci. Eng.* **288**(1) p. 12011

[16] C Slamet, A Rahman, M A Ramdhani and W Darmalaksana 2016 Clustering the Verses of the Holy Qur’an Using K-Means Algorithm *Asian J. Inf. Technol.* **15**(24) pp. 5159–5162

[17] D S Maylawati, M A Ramdhani, W B Zulfikar, I Taufik and W Darmalaksana 2017 Expert system for predicting the early pregnancy with disorders using artificial neural network in *2017 5th International Conference on Cyber and IT Service Management CITSM*

[18] W B Zulfikar, Jumadi, P K Prasetyo and M A Ramdhani 2018 Implementation of Mamdani Fuzzy Method in Employee Promotion System *IOP Conf. Ser. Mater. Sci. Eng.* **288**(1) p. 12147

[19] D S A Maylawati, M A Ramdhani, A Rahman and W Darmalaksana 2017 Incremental technique with set of frequent word item sets for mining large Indonesian text data 2017 *5th Int. Conf. Cyber IT Serv. Manag. CITSM* pp. 1–6

[20] A Taofik, N Ismail, Y A Gerhana, K Komarujaman and M A Ramdhani 2018 Design of Smart System to Detect Ripeness of Tomato and Chili with New Approach in Data Acquisition in *IOP Conference Series: Materials Science and Engineering* **288**(1) p. 12018

[21] W B Zulfikar and N Lukman Perbandingan Naive Bayes Classifier Dengan Nearest Neighbor Untuk Identifikasi Penyakit Mata *J. Online Inform.* **1**(2) pp. 82–86

[22] W B Zulfikar, M Irfan, C N Alam and M Indra 2017 The comparation of text mining with Naive Bayes classifier, nearest neighbor, and decision tree to detect Indonesian swear words on Twitter in 2017 *5th International Conference on Cyber and IT Service Management (CITSM)* pp. 1–5

[23] Larose D T and Larose C D 2014 *An Introduction to Data Mining,* “in Discovering Knowledge in Data an introduction to data mining” (Hoboken NJ USA: John Wiley & Sons, Inc.) pp. 1–15.

[24] F J Martinez López and J Casillas 2015 *KDD/Data Mining* (Chichester UK: John Wiley & Sons, Ltd)

[25] T Pang Ning 2006 *Introduction to Data Mining* (Boston: Pearson Addison Wesley)

[26] Kusriini and T Luthfi 2009 *Algoritma Data Mining* (Yogyakarta: Andi Offset)

[27] D Suryani, M Irfan, W Uriawan and W B Zulfikar 2016 Implementasi Algoritma Divide And Conquer Pada Aplikasi Belajar Ilmu Tajwi *J. Online Inform.* **1**(1) pp. 13–19
[28] D Setiawati, I Taufik, Jumadi and W Z Budiawan 2016 Klasifikasi Terjemahan Ayat Al-Quran Tentang Ilmu Sains Menggunakan Algoritma Decision Tree Berbasis Mobile J. Online Inform. 1(1) pp. 24–27

[29] E Elisa 2017 Analisa dan Penerapan Algoritma C4.5 Dalam Data Mining Untuk Mengidentifikasi Faktor-Faktor Penyebab Kecelakaan Kerja Kontruksi PT. Arupadhatu Adisesanti JOIN (Jurnal Online Inform.) 2(1) pp. 36–41

[30] D Syahid, Jumadi and D Nursantika 2016 Sistem Klasifikasi Jenis Tanaman Hias Daun Philodendron Menggunakan Metode K-Nearest Neighbour (Knn) Berdasarkan Nilai Hue, Saturation, Value (Hsv) J. Online Inform. 1(1) pp. 20–23

[31] Kusrini and E T Luthfi 2009 Algoritma Data Mining (Yogyakarta: Andi)

[32] K Nirmala, N Venkateswaran and C V Kumar 2017 HoG based Naive Bayes classifier for glaucoma detection in TENCON 2017 - 2017 IEEE Region 10 Conference pp. 2331–2336

[33] N R Fatahillah, P Suryati and C Haryawan 2017 Implementation of Naive Bayes classifier algorithm on social media (Twitter) to the teaching of Indonesian hate speech in 2017 International Conference on Sustainable Information Engineering and Technology (SIEET) pp. 128–131

[34] R Mehra, M K Bedi, G Singh, R Arora, T Bala and S Saxena 2017 Sentimental analysis using fuzzy and naive bayes in 2017 International Conference on Computing Methodologies and Communication (ICCMC) pp. 945–950

[35] M Granik and V Mesyura 2017 Fake news detection using naive Bayes classifier in 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON) pp. 900–903

[36] S Das and A K Kolya 2017 Sense GST: Text mining & sentiment analysis of GST tweets by Naive Bayes algorithm in 2017 Third International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN) pp. 239–244