Journal of
Information Systems Engineering
and Business Intelligence

Vol.8, No.2, October 2022
Available online at: http://e-journal.unair.ac.id/index.php/JISEBI

Data Mining Techniques in Handling Personality Analysis for Ideal Customers

Nur Ghaniaviyanto Ramadhan[1)*], Adiwijaya[2)]

[1)]*Software Engineering, Institut Teknologi Telkom Purwokerto, Indonesia*
Jl. DI Panjaitan No.128, Karangreja, Purwokerto Kidul, Purwokerto Selatan, Kabupaten Banyumas
[1)]ghani@ittelkom-pwt.ac.id

[2)] *Informatics, Telkom University, Indonesia*
Jl. Telekomunikasi. 1, Terusan Buahbatu - Bojongsoang, Dayeuhkolot, Kabupaten Bandung
[2)]adiwijaya@telkomuniversity.ac.id

Abstract

Background: Personality distinguishes individuals from one another, guides their actions and reactions, and dictates their preferences in many aspects of life, including shopping.
Objective: This study determines the characteristics of an ideal customer based on individual personality.
Methods: Data mining techniques used in this study are K-nearest neighbour (KNN), linear support vector machine (SVM), and random forest. This study also applies the synthetic minority oversampling technique (SMOTE) to overcome the imbalance in the amount of data.
Results: This study shows that the application of the SMOTE and random forest models resulted in 88% accuracy, 79% precision, and 70% recall, which are the highest compared to other models.
Conclusion: SMOTE in this research is unsuitable for use in the KNN and linear SVM classification models. Ensemble-based models such as random forest can produce high accuracy when SMOTE is applied for data pre-processing.

Keywords: Data Mining, Personality, Random Forest, SMOTE

Article history: Received 16 June 2022, first decision 23 August 2022, accepted 8 September 2022, available online 28 October 2022

I. INTRODUCTION

Personality is what distinguishes human beings from one another. It also guides actions, preferences, and behaviours [1] in many aspects of life, including shopping and consumption. Different consumers' behaviours are reflected in shopping lists from supermarkets, online shops, or souvenir outlets. These behaviours could represent consumers' personalities. Machines can be trained to recognise and assess individual personalities [2], including from their shopping lists. Past research has done this through music preferences [3], Twitter data [4], Facebook, and YouTube [5]. The methods used include XGBoost [2], an ensembled way [2], linear regression [4], decision tree [5], support vector machine [5], and robust model [6]. The practical implications of personality analysis using machine learning include social network analysis [7], recommendation systems [8], fraud detection [9], authorship attribution [10], and sentiment analysis or opinion mining [11]. Machine learning is better than questionnaire investigations or expert interviews because traditional personality assessment methods are expensive and less practical [12].

User personality analysis has been done in past research using supervised learning. Mavis et al. seek to model the association between social media messages and the big five personality traits [4]. They used Twitter posts and user statistics for analysis [4]. Several methods were used to capture user profiles, and the supervised learning techniques were then compared [4]. Meanwhile, Shumanov et al. aimed for two goals: first, to show how an algorithm may be used to determine consumer personality traits using contextual data, and second, to investigate the variables that affect the relationship between personality traits and advertising persuasiveness [13]. Vargas et al. discover market segments

[*] Corresponding author

doi: http://dx.doi.org/10.20473/jisebi.8.2.175-182

using impartial logic [14]. They used neutrosophic logic, a decision-making method using a neural network. It provides a nuanced interpretation of language phrases for qualitative information analysis in market segmentation.

To enhance the effectiveness of sentiment categorisation for customer evaluations, Zhang et al. suggested a model that considers all characteristics and explores interactive relationships within and across features [15]. Maheswari et al. classified clients using the SVM algorithm based on their purchasing patterns [16]. They used inventory and sales datasets available on the Internet and evaluated them using algorithms. Tandera et al. developed a system to predict personality using data from Facebook [17]. The personality model used was the big five personality model. The algorithm used is deep learning, with an accuracy of 74.17%.

Majumder et al. developed a deep learning method to identify personality from a text [18]. The big five psychological profiles can identify features based on the text. Distinct binary classifiers can be trained with a similar design for each feature. The implementation used a deep convolutional neural network (CNN) with a specific design. M. Rahman et al. provided an empirical method for determining the optimum personality recognition performance [19] by contrasting numerous activation functions, including sigmoid, tanh, and leaky ReLU. Convolutional neural networks receive pre-processed and vectorised text as input. The length of each word, sentence, record, and the feature vector is multiplied by the input size. The experimental analysis used five personality traits named EXT, NEU, AGR, CON, and OPN. Likewise, Pratama et al. constructed a working framework to predict identity from content composed by Twitter users [20]. Using deep learning, Mehta et al. examined models applied to personality recognition [21]. Elmitwally et al. propose identifying identity and behaviour based on minimal input [22]. This has also been done using the big five personalities [23]. Singh et al. investigate the use of Twitter profiles to predict users' personalities [24], involving 450 profiles and over 1 million tweets.

Based on the previous studies, this study focuses on the personality of shoppers, using several data mining classification methods. Therefore, this study will focus on the data balancing techniques used. In previous studies, using data balancing techniques for personality analysis problems is still rare.

II. METHODS

Fig. 1 shows the proposed system in this study. The initial step starts with obtaining data related to customer personality. The second step is pre-processing, namely identifying missing values and balancing data. The third step is classification using several data mining techniques, such as random forest, KNN, and SVM. The last step is to analyse the classification results based on accuracy, precision, and recall.

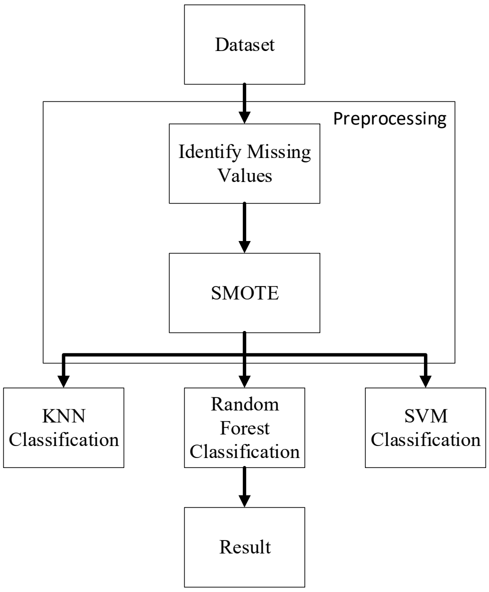

Fig. 1 proposed system

A. Dataset

This research dataset uses a dataset about customer personalities [25], containing 2240 rows of data and 29 columns. Table 1 shows the characteristics of the dataset in this study. This dataset is obtained from customers' transactions at a store. The data underwent pre-processing before data values were identified, and the amount of data was balanced.

TABLE 1
CHARACTERISTICS DATASET

No	Features	Type
1	ID	Numeric
2	Year Birth	Numeric
3	Education	Text
4	Marital Status	Text
5	Income	Numeric
6	Kid Home	Numeric
7	Teen Home	Numeric
8	Dt Customer	Date
9	Recency	Numeric
10	Spent Wine	Numeric
11	Spent Fruits	Numeric
12	Spent Meat Products	Numeric
13	Spent Fish Products	Numeric
14	Spent Sweet Products	Numeric
15	Spent Gold Products	Numeric
16	Num Deals Purchases	Numeric
17	Num Web Purchases	Numeric
18	Num Catalog Purchases	Numeric
19	Num Store Purchases	Numeric
20	Num Web Visits Month	Numeric
21	AcceptedCmp3	Numeric
22	AcceptedCmp4	Numeric
23	AcceptedCmp5	Numeric
24	AcceptedCmp1	Numeric
25	AcceptedCmp2	Numeric
26	Complain	Numeric
27	Cost Contract	Numeric
28	Revenue	Numeric
29	Response	Numeric

B. Pre-processing

This process is carried out by selecting the features used in this study. Features that are not used in this study are ID, Year, and Dt Customer because they are not relevant. Education and marital status features are changed into numeric form.

C. Identifying the Missing Values

In this process, the identification of null values in the dataset was carried out. The dataset of this study has an invalid value of 24 rows in the income feature. This can disrupt the process of balancing data classes and classifications. Therefore, the null values were deleted, so the total row of data in the dataset was 2216.

D. SMOTE

This step is done by balancing the data class, namely the response. There are two types of reactions to the dataset: accepting an offer is written with the number 1, or rejecting the proposal is registered with the number 0. The method used in this process is the SMOTE oversampling model. In this process, for the data amounting far below the dominant class, additional data will be added so that the amount is close to balance with the prevailing style.

TABLE 2
APPLY SMOTE

	Response class 0	Response class 1
Before applying SMOTE	1883	333
After applying SMOTE	1883	1428

Table 2 shows the results obtained after applying SMOTE. For minor class 1, after the SMOTE technique was carried out, there was a significant addition of about 1000 data. The class was almost balanced with the amount of data in the substantial class 0.

E. Classification

The final step is classification is to compare 30% of testing data with 70% of training data to determine accuracy, precision, and recall. This study uses several data mining classification methods: random forest, SVM, and KNN.

- SVM

The linear SVM algorithm with the chosen kernel is used for the classification procedure. An essential component of machine learning theory is the SVM, which is effective in many scientific and engineering applications, particularly in classification [26]. The SVM classification idea can be described in (1): suppose there are m samples of observations (training set), (xi, yi), *i = 1, 2, ..., m* where:

$$x_i^T = (xi1, \dots xid) \in R^d \#(1)$$

Where x_i^T is the d-dimensional feature of the sample, i and $y \in$ *{-1, +1}* is the coded label class. If sample *xi* is assigned to a positive class, so *yi* is +1, and if assigned to a negative class, so *yi* is -1. This training set can be separated by a hyperplane $w^T xi + b = 0$, where *w* is the weight vector, and *b* is bias. The linear kernel formula can be seen in (2).

$$K(x, z) = x^T z \#(2)$$

Using linear kernels on data with various properties, such as text data, is appropriate. The kernel functions significantly influence the accuracy obtained in an SVM analysis and the parameters employed [26].

- Random Forest

The classification or regression ensemble learning technique builds many decision trees throughout the training process and outputs. The results as the class mode (classification) or average prediction (regression) of individual trees [27]. Random forest is a well-known supervised learning classification model in several classification fields [27].

$$Entropy = \sum_{i}^{c} \quad Pi\ log\ log\ 2\ \ Pi \#(3)$$

The entropy (information gain) in (3) is used in this detection process, where c is the number of values in the target attribute (number of classification classes), Pi is the sample portion for type i.

- KNN

The most effective K value, or one that comes the closest to the outcome, is found by this algorithm. Use the following equations (4) - (6), and to calculate the similarity efficiency in KNN (7) [28].

$$Sim(d1, d2) = \frac{d1. d2}{||d1||_2 ||d2||_2} \#(4)$$

where d1 and d2 are the vector documents used, each neighbour is assigned a weight using the similarity in each neighbour up to d0, as shown in formula (5).

$$score(d0, C_i) = \sum_{dj \in KNN(d0)} \quad Sim(d0, dj)\delta(dj, Ci)\#(5)$$

where KNN (d) is the closest set of K-neighbours of the d0 document. δ(dj,Ci) stands for classification of dj documents related to class Ci. Formula (6) is a derivative of the formula δ(dj,Ci).

$$\delta(dj, Ci) = \{1\ dj\ \in Ci\ 0\ dj\ \notin Ci\ \} \#(6)$$

Finally, to decide on the KNN, Equation (7) is used.

$$C = arg\ arg\ max_{ci}\left(score(d0, C_i)\right)\ \#(7)$$

To see the results of accuracy, precision, and recall, calculations using the confusion matrix (8), (9), and (10) are used. The experiments carried out in this study were to apply three data mining classification methods (random forest, SVM, and KNN). In addition, the experiments are conducted by applying data class balancing techniques using SMOTE. Equations (8) - (10) are formulas for calculating accuracy, recall, and precision [22].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}\#(8)$$

$$Recall = \left(\frac{TP}{TP\ + FN}\right)\ \#(9)$$

$$Precision = \left(\frac{TP}{TP\ + FP}\right)\ \#(10)$$

III. Results

Table 3 displays the accuracy, precision, and recall results in each method conducted by experiments in this study. The highest measurement results in terms of accuracy, precision, and recall were obtained in the proposed method of this research, namely SMOTE + Random Forest, at 88%. In comparison, the lowest accuracy was the SMOTE + KNN method at 66%.

TABLE 3
Result

No	Methods	Accuracy %	Precision %	Recall %
1	Random Forest	88	72	58
2	Linear SVM	84	65	56
3	KNN	84	65	56
4	**SMOTE + Random Forest**	**88**	**79**	**70**
5	SMOTE + Linear SVM	73	63	73
6	SMOTE + KNN	66	58	64

Table 4 shows the comparison with previous research on personality classification. This study also compared the accuracy results with several previous studies. The results show that the SMOTE + Random Forest method is more effective than previous studies. The farthest difference in accuracy reached 22%. The results on Random Forest with SMOTE + Random Forest have the same accuracy, but for precision and recall, the results are different. After applying the sampling method, the precision and recall results increased.

TABLE 4
Comparison to Other Study

No	Results	Accuracy %
1	Utami et al [1]	56
2	Kunte et al [2]	82.5
3	Farnadi et al [5]	78.25
4	Mehta et al [21]	58
5	Elmitwally et al [22]	73
6	**This Research**	**88**

Table 3 shows the SMOTE and random forest models obtained the highest results of 88% for accuracy, 79% for precision, and 70% for recall. These results were compared using only a single random forest classification model, which resulted in lower precision and recall by 7% and 12%, respectively. This study also tested other methods: linear SVM and KNN. The single linear SVM method results are compared to SMOTE for accuracy, precision, and recall. The results obtained for the single KNN method compared with SMOTE and KNN are the same: decreasing accuracy and precision results but increasing recall results. After applying the SMOTE model, the decrease in accuracy is the most significant at 18%, while accuracy declined by 11%. Based on the results obtained, applying the SMOTE model to imbalanced data problems for several models does not improve the accuracy results. Still, there is an increase in precision and recall.

Table 4 also shows that this study yielded a higher accuracy, even with a difference of up to 20%. Previous research only used supervised learning data mining methods without data balancing. Even with data balancing, a previous study could only produce an accuracy of 56%. This was due to the selection of a supervised learning model that did not match the form of the data used.

IV. DISCUSSION

The problem of decreasing yield after applying the SMOTE method is because, at the time of oversampling, minority class data was balanced out forcibly to match the majority class. These problems cannot be appropriately classified for some models, such as linear SVM and KNN. However, the problem can be organised well by the random forest model. The three methods used for classification have different characteristics. For example, for linear SVM, good accuracy results are obtained based on the type of kernel used. Meanwhile, for KNN, good accuracy results are determined by the right K point. In contrast, accuracy results depend on the number of resulting decision trees and information gained for a random forest. Therefore, in applying the SMOTE method, only the random forest model can improve the accuracy, precision, and recall results.

TABLE 5
RESEARCH CONTRIBUTION

Personality Identification	Balancing Data	Methods
Utami et al [1]	Yes	SVC
Kunte et al [2]	No	XGBoost
Farnadi et al [5]	No	Regression
Mehta et al [21]	No	Deep Learning
Elmitwally et al [22]	No	Decision Tree
This research	Yes	Random Forest

This study had higher results than previous studies. Table 4 and 5 is a comparison of the contributions of previous studies with the research of this paper. For example, Kunte et al. [2] used the XGBoost classification model without performing a dataset balancing technique. It only resulted in an accuracy of 82.5%. Meanwhile, Utami et al. [1] carried out a dataset balancing technique and used the SVC classification model. However, the accuracy tended to be low at 56%. From the previous two studies, the accuracy results can still be improved. The results of this proposed study are 88% with balancing techniques and random forest classification models. Using classification models and dataset balancing techniques can result in significantly different accuracies of 6%-22%. Nonetheless, the limitation of this study is that the data balancing dataset is used only with limited data mining classification models.

V. CONCLUSIONS

Based on the results obtained in this study, the best data mining methods are SMOTE and random forest, with 88% accuracy. The lowest accuracy is by SMOTE and linear SVM models at 73%. Meanwhile, the lowest precision is 58% at the application of SMOTE and KNN. The lowest recall is 56%, applying a single KNN and linear SVM model. The SMOTE method in this research is unsuitable for use in the KNN and linear SVM classification models. Ensemble-based models such as random forest can produce high accuracy even though the SMOTE model is applied for data pre-processing. We can combine under-sampling imbalanced data methods with deep learning classification models for future problems.

Author Contributions: *Nur Ghaniaviyanto Ramadhan*: Software, Conceptualization, Methodology, Writing - Original Draft, Writing - Review & Editing, Supervision. *Adiwijaya*: Methodology, Writing - Original Draft, Review & Editing.

Funding: Research received no specific grant from any funding agency.

Conflicts of Interest: The authors declare no conflict of interest.

REFERENCES

[1] E. Utami, I. Oyong, S. Raharjo, A. Dwi Hartanto, and S. Adi, “Supervised learning and resampling techniques on DISC personality classification using Twitter information in Bahasa Indonesia,” *Appl. Comput. Informatics*, 2021.

[2] A. Kunte and S. Panicker, “Personality Prediction of Social Network Users Using Ensemble and XGBoost,” *Adv. Intell. Syst. Comput.*, vol. 1119, pp. 133–140, 2020.

[3] T. Krismayer, M. Schedl, P. Knees, and R. Rabiser, “Predicting user demographics from music listening information,” *Multimed. Tools Appl.*, vol. 78, no. 3, pp. 2897–2920, 2019.

[4] G. Mavis, I. H. Toroslu, and P. Karagoz, “Personality Analysis Using Classification on Turkish Tweets,” *Int. J. Cogn. Informatics Nat. Intell.*, vol. 15, no. 4, pp. 1–18, 2021.

[5] G. Farnadi *et al.*, “Computational personality recognition in social media,” *User Model. User-adapt. Interact.*, vol. 26, no. 2–3, pp. 109–142, 2016.

[6] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, “Personality Predictions Based on User Behavior on the Facebook Social Media Platform,” *IEEE Access*, vol. 6, pp. 61959–61969, 2018.

[7] F. Celli and L. Rossi, “The role of Emotional Stability in Twitter Conversations,” *Proc. Work. Semant. Anal. Soc. Media, conjunction with EACL 2012*, pp. 10–17, 2012.

[8] A. Roshchina, J. Cardiff, and P. Rosso, “A comparative evaluation of personality estimation algorithms for the twin recommender system,” *Int. Conf. Inf. Knowl. Manag. Proc.*, pp. 11–17, 2011.

[9] F. Enos, S. Benus, R. L. Cautin, M. Graciarena, J. Hirschberg, and E. Shriberg, “Personality factors in human deception detection: Comparing human to machine performance,” *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2, pp. 813–816, 2006.

[10] K. Luyckx and W. Daelemans, “Personae: a corpus for author and personality prediction from text,” *Proc. Sixth Int. Conf. Lang. Resour. Eval.*, pp. 2981–2987, 2008.

[11] J. Golbeck and D. L. Hansen, “Computing political preference among Twitter followers,” *Conf. Hum. Factors Comput. Syst. - Proc.*, pp. 1105–1108, 2011.

[12] D. Nie, Z. Guan, B. Hao, S. Bai, and T. Zhu, “Predicting personality on social media with semi-supervised learning,” *Proc. - 2014 IEEE/WIC/ACM Int. Jt. Conf. Web Intell. Intell. Agent Technol. - Work. WI-IAT 2014*, vol. 2, pp. 158–165, 2014.

[13] M. Shumanov, H. Cooper, and M. Ewing, “Using AI predicted personality to enhance advertising effectiveness,” *Eur. J. Mark.*, 2021.

[14] J. Vargas, N. ALberto, and O. Arevalo, “Algorithms for Decision Making Through Customer Classification,” *Int. Conf. Intell. Comput. Inf. Control Syst. Adv. Intell. Syst. Comput.*, vol. 1272, 2021.

[15] Y. Zhang, J. Wang, and X. Zhang, “Personalized sentiment classification of customer reviews via an interactive attributes attention model,” *Knowledge-Based Syst.*, vol. 226, p. 107135, 2021.

[16] K. Maheswari, P. Packia, and A. Priya, “Predicting Customer Behavior in Online Shopping Using SVM Classifier,” *IEEE Int. Conf. Intell. Tech. Control. Optim. Signal Process.*, 2017.

[17] T. Tandera, Hendro, D. Suhartono, R. Wongso, and Y. L. Prasetio, “Personality Prediction System from Facebook Users,” *Procedia Comput. Sci.*, vol. 116, pp. 604–611, 2017.

[18] N. Majumder, S. Pouria, A. Gelbukh, and E. Cambria, “Deep Learning-Based Document Modeling for Personality Detection from Text,” *IEEE Intell. Syst.*, pp. 74–79, 2017.

[19] M. A. Rahman, A. Al Faisal, T. Khanam, M. Amjad, and M. S. Siddik, “Personality Detection from Text using Convolutional Neural Network,” *1st Int. Conf. Adv. Sci. Eng. Robot. Technol. 2019, ICASERT 2019*, vol. 2019, no. Icasert, pp. 1–6, 2019.

[20] B. Y. Pratama and R. Sarno, “Personality classification based on Twitter text using Naive Bayes, KNN and SVM,” *Proc. 2015 Int. Conf. Data Softw. Eng. ICODSE 2015*, pp. 170–174, 2016.

[21] Y. Mehta, N. Majumder, A. Gelbukh, and E. Cambria, “Recent trends in deep learning based personality detection,” *Artif. Intell. Rev.*, vol. 53, no. 4, pp. 2313–2339, 2020.

[22] S. Elmitwally, Nouh Sabri; Kanwal, Asma; Abbas, Sagheer; Khan, Muhammad A.; Khan, Muhammad Adnan; Ahmad, Munir; Alanazi, “Personality Detection Using Context Based Emotions in Cognitive Agents,” *C. Mater. Contin.*, vol. 70, no. 3, pp. 4947–4964, 2022.

[23] A. Mamta, Bhamare; K, “Prediction of Personality Traits in Facebook Users,” *Lect. Notes Data Eng. Commun. Technol.*, vol. 86, 2022.

[24] M. S. R. B and A. Singh, “Personality Detection by Analysis of Twitter Profiles,” *8th Int. Conf. Soft Comput. Pattern Recognition, SoCPaR 2016*, no. 2194–5357, 2018.

[25] Kaggle, “Customer Personality Analysis,” *https://www.kaggle.com/imakash3011/customer-personality-analysis*, vol. November, 2021.

[26] N. G. Ramadhan and T. I. Ramadhan, “Analysis Sentiment Based on IMDB Aspects from Movie Reviews using SVM,” *Sinkron*, vol. 7, no. 1, pp. 39–45, 2022.

[27] N. G. Ramadhan, F. D. Adhinata, A. Jala, T. Segara, and D. Putra, “Deteksi Berita Palsu Menggunakan Metode Random Forest dan Logistic Regression,” *J. Ris. Komput.*, vol. 9, no. 2, pp. 251–256, 2022.

[28] N. G. Ramadhan, “Indonesian Online News Topics Classification using Word2Vec and,” *J. Resti*, no. 158, pp. 7–10, 2021.

Publisher’s Note: Publisher stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.