Customer Profile Prediction model based on classification through approach Support Vector Machine (SVM)

Herman Mawengkang [1)*] **Ogiana**[2)], **Elviawaty Muisa Zamzami** [3)]
[1,2,3)] Universitas Sumatera Utara, Medan, Indonesia
[1)] mawengkang@usu.ac.id

Submitted : July 2, 2022 | **Accepted** : July 30, 2022 | **Published** : July 31, 2022

Abstract: Nowadays the market is characterized globally, products and services are almost identical and there are many suppliers. The most important aspect in classifying data in data mining is classification. Classification techniques have been widely used in many problems in research. The purpose of this research is to build a model that can predict behavior based on the information of each customer. This research was conducted by making a Prediction Model of Customer Profile Based on Classification Through the Support Vector Machine Approach which aims to obtain a package prediction accuracy value that is suitable for WO (Wedding Organizer) customers in classifying based on the profile of prospective customers. In the optimization results on the SVM model kernel function, the linear and polynomial kernels get the same accuracy value on the training data of 99.29% and the testing data of 94.92%. The lowest accuracy value was obtained in the RBF kernel function of 97.16% on training data and 96.61% on testing data. the best precision class value in the data testing was obtained in the basic package at 100%. The total value of the appropriate prediction on the training data was obtained by 56 samples from a total of 59 samples, and 3 samples that did not match the prediction with an accuracy of 94.92% on the data testing.

Keywords: Classification; kernel; Optimization; prediction; data; SVM

INTRODUCTION

Nowadays the market is characterized globally, products and services are almost identical and there are many suppliers. Due to the size and complexity of the market, mass marketing becomes expensive and the return on investment is often questionable. So that the difficulties experienced by potential consumers become confused about choosing the best service, then data mining is used to implement and overcome problems related to classification through certain attributes.. (Somantri, Wiyono, and Dairoh 2016)

The most important aspect in classifying data in data mining is classification. Classification techniques have been widely used in many problems in a research. The method in classifying data groupings that will study training data can use a classification algorithm such as SVM (Support Vector Machine). (Pratama, Wihandika, and Ratnawati 2018)

SVM is a classification method using machine learning (supervised learning) methods that can predict criteria based on patterns from the results of the training process created by Vladimir Vapnik. Classification is done with a dividing line (hyperlane) that separates the positive and negative opinion classes. (Sari and Haranto 2019)

In the classification company can only select customers who meet certain profitability criteria based on their individual needs and purchasing patterns. This is achieved by building a model to predict the future value of an individual customer based on demographic characteristics, lifestyle, and previous behavior. The model generates information that will focus customer retention and recruitment programs on building and retaining the most profitable customer base. This is called customer behavior modeling (CBM) or customer profiling. (Made adi Pranata and Gede sri Darma 2018)

To get the most out of predicting customer desires is customer profile data in order to help marketers to their targets because they better understand the characteristics of their customer base. The long-term motivation of customer profiling is to turn this understanding into automated interactions with their customers.

*name of corresponding author

Customer profiles describe customers by their attributes, such as age, income, and lifestyle. This is done by building a model of customer behavior and estimating its parameters. Customer profiling is a way of applying external data to a population of potential customers. Depending on the available data, the model can be used to find new customers or to "stop" existing bad customers. (Akinyelu and Ezugwu 2019)

The classification method used in this study is the Support Vector Machine (SVM). SVM is a technique that is included in the supervised class so that its implementation requires a target class. This SVM can find a global optimal solution. Furthermore, this research will study the Prediction Model of Customer Profile Based on Classification Through the Support Vector Machine Approach

LITERATURE REVIEW

Support Vector Machine (SVM)

The support vector machine uses nonlinear mapping to convert the original training data to higher dimensions. In this new dimension, we will look for a linear optimal separator hyperplane (that is, a "decision boundary" separating data from one class to another). With proper nonlinear mapping for sufficiently high dimensions, the data of the two classes are separated by a hyperplane. SVM finds this hyperplane using support vectors (class boundaries) and margins (defined by support vectors) (Han and Kamber 2006).

SVM simultaneously minimizes empirical misclassification and maximizes geometric boundaries. So SVM is called Classifier Maximum Margin. (Srivastava and Bhambhu 2010)

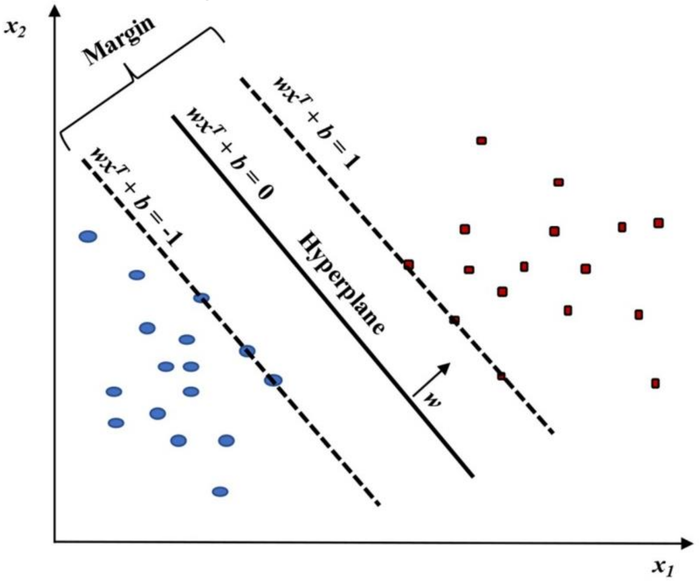

Fig. 1 Linear SVM model. Two classes (red versus blue) are classified

In the image above you can start with a simple SVM and then experiment with various 'standard' kernel functions. Depending on the nature of the problem, it is possible that one kernel is better than another. Optimal kernel functions can be selected from fixed kernel assets in a statistically strict manner using cross-validation

There are many kernel functions in SVM, function selection is also a matter of research. However, for general purposes, there are several popular kernels as follows::

- Linear kernel:

$$\mathrm{K}\ (x_i, x_j) = \ x_i^T x_j \qquad (2.1)$$

- Polynomial kernel:

$$K\ (x_i, x_j) = \ (y\ x_i^T x_j + r) d, y > 0 \qquad (2.2)$$

- RBF kernel:

$$K\ (x_i, x_j) = \ exp(-y \,||\, x_i x_j \,||^2), \qquad y > 0 \qquad (2.3)$$

- Sigmoid kernel:

$$K\ (x_i, x_j) = \ tanh(y\ x_i^T x_j + r)\ \ y > 0 \qquad (2.4)$$

*name of corresponding author

Decision Trees

A decision tree is a tool that uses classification or regression to predict responses to data. Classification is used when features are grouped, and regression is used when data is continuous. Decision trees are one of the main data mining methods. A decision tree is made up of root nodes, branches, and leaf nodes. To evaluate the data, follow the path from the root node to reach the leaf node.

The decision tree must be created using the purity index that will divide the nodes. For CHDD, each of the 297 tuples was evaluated based on the decision tree and arrived at a positive or negative evaluation for liver disease. This is compared to the initial decision parameters in the CHDD to check for false positives or false negatives, giving us the accuracy, specificity, and sensitivity of the model. The separation criteria used also indicate the importance of each attribute..

METHOD

This research was conducted by making a Prediction Model of Customer Profile Based on Classification Through the Support Vector Machine Approach which aims to obtain a package prediction accuracy value that is suitable for WO (Wedding Organizer) customers in classifying based on the profile of prospective customers. In this study, an increase in the value of accuracy in classifying demographic characteristics, lifestyle, and previous behavior will be carried out using the SVM Model technique which aims to get the right prediction for customers who want to take a package at WO (Wedding Organizer).

3.1 Input

The dataset taken is real data from Best Wedding (WO) customers taken through the history of customers who have made transactions.

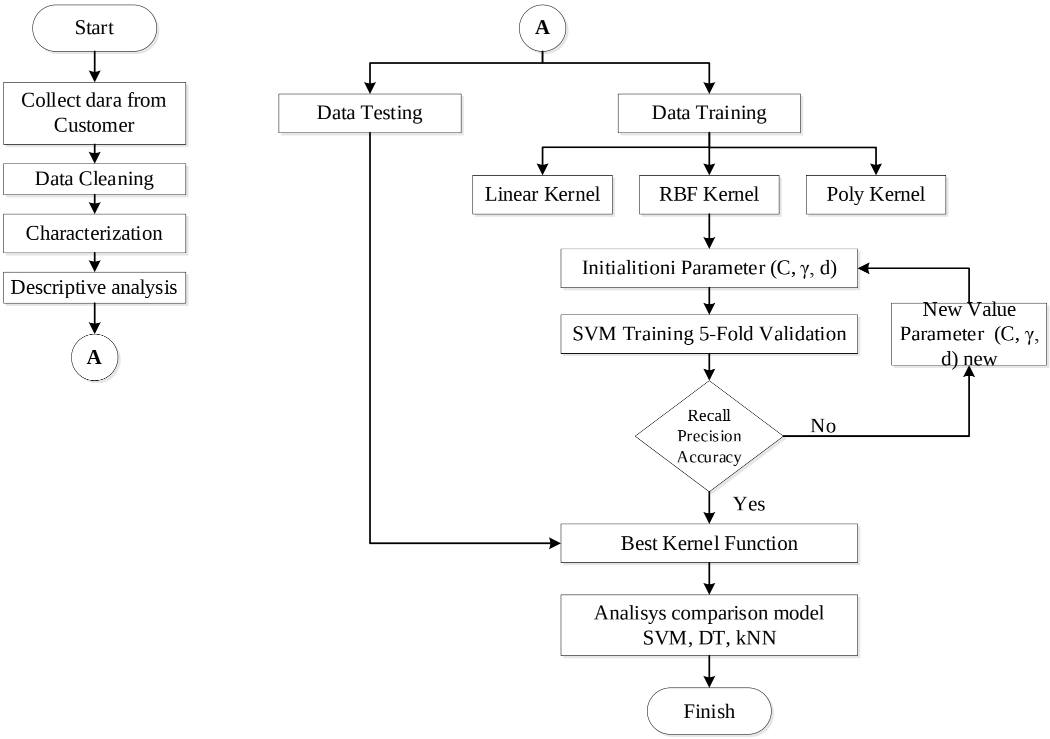

Fig. 2 General Architecture Research

Process

At this stage is the SVM modeling process. The process consists of preprocessing, segmentation, post processing, and classification.

Preprocessing

At the preprocessing stage, the dataset will be processed in order to extract good features. This stage includes processing data.

Segmentation

*name of corresponding author

The stage after preprocessing is segmentation which aims to produce a classification using the SVM Classification technique.

Post processing

The results of the segmentation stage still contain unwanted attributes or data, so it is necessary to eliminate these data using connected component analysis.

Classification

In this study, the method used for the classification process is SVM Modeling.

Output

Process *Training*

Before entering this training stage, the SVM Model is made. All prediction results will be presented in order of value. which is the result of processing the input that has been received by the input data and classified by SVM. And next, the calculation of the output weight will be carried out. The result of this process is the highest score to become the recommendation.

Process Testing

This stage is carried out to determine the effectiveness of the SVM Modeling method in the classification of wedding packages at WO Best Wedding.

.

RESULT

This chapter will explain the results of testing the training dataset and classification tasks using the Support Vector Machine (SVM) algorithm with the Decision Tree and KNN algorithms as comparisons. Testing of training data and task data is carried out using Rapidminer software. The dataset classification process in the form of customer needs is carried out using the Support Vector Machine (SVM) algorithm that has been optimized to get the best accuracy on multiclass data, with labels in the form of selected packages with a total of 3 classes. The research has 15 attributes based on demographics/profiles, attributes and packages. The research trials were carried out using a computer with the following hardware and software specifications:

1. Processor Intel Pentium Core i3
2. RAM 4GB
3. Microsoft Excel 2010
4. Rapidminer for Academic 9.1.

The analytical method used in this study uses the LibSVM tools on Rapidminer for multiclass label data. Test analysis is measured using indicators of accuracy in the training process and testing on the dataset. The total dataset obtained is 200 samples consisting of 15 attributes and 3 classes on labels in the form of packages selected by customers. The total dataset is then divided into two using split data tools, for training data and testing data with a ratio of 70:30 using the stratified sampling method. To get the best accuracy value, a comparison of three types of kernels in SVM will be used, namely linear, RBF and polynomial. For each type of kernel, the parameters will be optimized in the form of C, and d. The kernel type with the highest accuracy value will be compared with the DT and KNN models to get the appropriate model in predicting customer needs.

Table.1 Atribut Dataset Penelitian

Demographic Dataset	Age	0 = 20-25 1=26-30 2=>30 year	
	Last education	0 = Senior High School 1=Diploma 2=Bachelor 3=Postgraduate	
	ethnics	0 = Jawa 1 = Mandailing	7 = Acehnese 8 = *Chinese*

*name of corresponding author

		2 = Padang 3 = Karo 4 = Batak 5 = Melayu 6 = Sunda	9 = Banten 10 = Jambak 11 = Simalungun 12 = Banjar
	Job	0 = PNS 1 = Private Employee 2 = BUMN Employee 3 = Enterpreneur 4 = Housewife	
	Domicile	0 = Medan 1 = Outside Medan	
	Venue of the event	0 = Hall 1 = Hotel 2 = House	
	Event	0 = Akad 1 = Resepsi 2 = Adat-Resepsi 3 = Akad-Resepsi 4 = Akad-Adat-Resepsi 5 = Lamaran- Akad-Adat-Resepsi	
Behavior Dataset	Theme	0 = Elegan 1 = Classic 2 = Rustic 3 = Glamour 4 = White 5 = Vintage 6 = Garden 7 = Etnic 8 = Combine	
	Tone Flower	0 = Light 1 = Pastel	
	Gown Type	0 = Akad, Tradisional, Nasional 1 = Akad, Nasional 3 = Nasional	
	Gown Colour	Integer data based on the color combination used	
	Hena	0 = White 1 = Red	
	Theme Make UP	0 = Natural 1 = Bold 2 = Barbie Look	
	Band	0 = Full Band 1 = Acoustic 2 = Traditional	
	Dancer	0 = Traditional 1 = Modern	
Dataset Label	Package	0 = Basic Package 1 = Intermediate Package 2 = Full Package	

*name of corresponding author

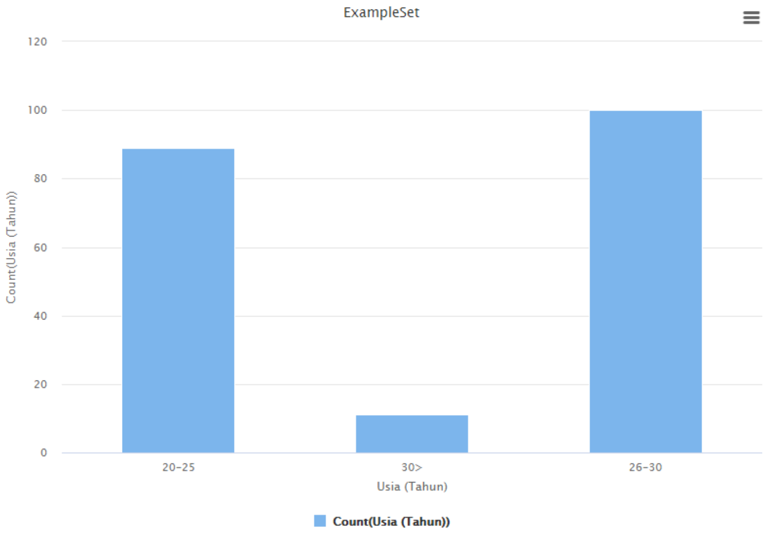

Fig 3 Customer Description by Age

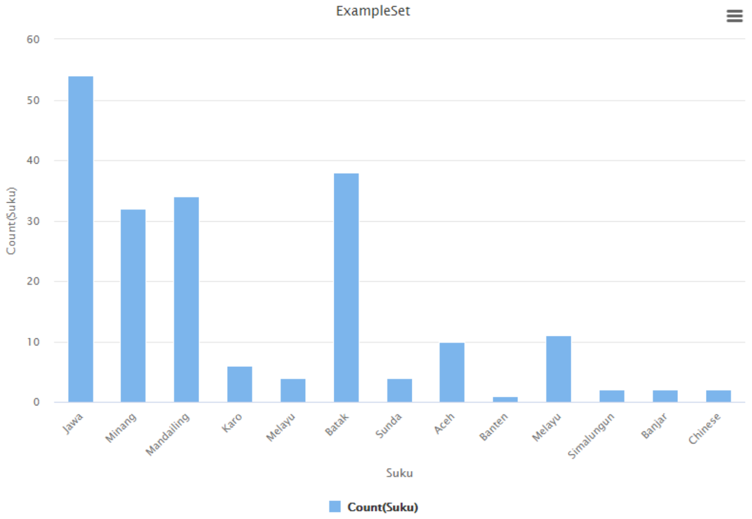

Fig 4 Customer Description by Tribe

Table 2 Confusion Matrix Data Testing SVM Polynomial

Kernel	Data Training	Data testing
	Accuracy	Accuracy
Linear	99,29%	94,92%
RBF	97,16%	96,61%
Polynomial	99,29%	94,92%

Table 3 Comparison of Accuracy Between Models

Model	Accuracy	
	Data Training	Data Testing
SVM – Kernel Polynomial	99,29%	94,92%
Decision Tree	98,58%	96,61%
KNN	91,88%	83,05%

Table.4 Parameter Model SVM Polynomial

Parameter	
SVM-Type	C-SVC

Kernel Type	Polynomial
C	39,136
d	1
Support Vector	129

Table.5 Confusion Matrix Data Training SVM Polynomial

		True			Class Precision
		Basic Package	Full Package	Intermediate Package	
Prediction	Basic Package	19	0	0	100.00%
	Full Package	0	52	0	100.00%
	Intermediate Package	0	1	69	98.57%
Class Recall		100.00%	100.00%	98.11%	

Table 6 Confusion Matrix Data Testing SVM Polynomial

		True			Class Precision
		Basic Package	Full Package	Intermediate Package	
Prediction	Basic Package	8	0	0	100.00%
	Full Package	0	21	2	91.30%
	Intermediate Package	0	1	27	96.43%
Class Recall		100.00%	95.45%	93.10%	

DISCUSSIONS

The data used in this study is the best wedding organizer user data which is divided into three main datasets, namely demographic datasets, behavioral datasets and selected package datasets. A total of 200 samples of customers were collected which were expected to show the characteristics.

Before conducting further discussion, the first thing to do is to do a descriptive analysis on the variables used. Descriptive analysis aims to get an overview of the data both the number and type of attributes, frequency of data and characteristics of customers. Based on Fig 2, it can be seen the comparison of customers by age. The best WO customers are dominated by the age of 26-30 with a total sample of 100 samples, while the category of customers with the least age is found at the age of >30 with a total of 11 samples.

Based on Fig 3, it can be seen the comparison of customers by ethnicity. The category of best wedding customers is dominated by Javanese with a total of 54 samples. Meanwhile, the lowest frequency of customers based on tribes is obtained from customers with offerings of 1 sample. According to Damanik, (2018) the Batak tribe is a tribe originating and living in North Sumatra. This is what makes the tribe that takes the second most is the Batak. Besides the Javanese, the majority have spread throughout Indonesia.

Polynomial Kernel SVM

The kernel polynomial function in SVM is a form of non-linear kernel, where its use is very suitable to solve problems in non-normalized training data. According to (Diani 2017), there are two parameters that affect the accuracy of the polynomial kernel, namely C and Degree(d). In this study, optimization of parameters C and d was carried out with a range of 0.01-10 and 1-5. The best accuracy on the training data of the SVM model using a polynomial kernel is obtained at the values of C = 39.136 and d = 1 with an accuracy rate of 99.3%. after obtaining the optimum parameter values, then a model is formed using the best value for each parameter to obtain a confusion matrix on data training and data testing. The following are the parameters of the SVM model using the Polynomial kernel.

In table 5 it can be seen that the best precision class value in the training data is obtained in the basic package at 100%, and the best recall class value is obtained in the basic package at 100%. From the data training, the model predicts 19 samples of customers who choose the basic package with a total data of 19 customers who choose the basic package. With a total of 140 correct answers out of a total of 141 samples, the accuracy of the model is 99.29%. In table 6, it can be seen that the best precision class value in the data testing was obtained in the basic package at 100%, and the best recall class value was obtained in the basic package at 100%%. The total value of the appropriate prediction on the training data was obtained by 56 samples from a total of 59 samples, and 3 samples that did not match the prediction with an accuracy of 94.92% on the testing data.

SVM Kernel Function Comparison Results

After testing the effect of the type of kernel on the accuracy of the model, then the accuracy level of each kernel will be compared. This comparison aims to determine the appropriate kernel function in determining the accuracy of the package classification that will be selected by the customer.

From table 2, it can be seen that the linear kernel function obtained a value of 99.29%, and the polynomial kernel of 99.29% in the training data. Both functions provide the same level of accuracy both on training data and testing data. In the next model comparison test, the kernel polynomial will be selected as a kernel function in the model. This is based on data that is polynomial. Because Polynomial function is the best function of SVM, Therefore, the prediction model to estimate the customer profile will use the polynomial function

After optimizing the parameters for each model, then a comparison of the level of accuracy between models is carried out. In table 3 it can be seen that the comparison of the accuracy of the model using training data and testing data, it is found that the SVM model using the polynomial kernel function provides the best accuracy of 99.92% on training data and 94.92% on testing data, compared to the 98.58 DT model. % on training data and 96.61% on testing data. The lowest model accuracy is obtained in the KNN model of 91.88% on training data and 83.05% on testing data. From these results, it can be concluded that the SVM model is more appropriate to use in predicting the package that will be taken by the best wedding organizer customers.

CONCLUSION

In the optimization results on the SVM model kernel function, the linear and polynomial kernels get the same accuracy value on the training data of 99.29% and the testing data of 94.92%. The lowest accuracy value was obtained in the RBF kernel function of 97.16% on training data and 96.61% on testing data. Of the three models used, namely the SVM model, decision tree and KNN. The SVM model provides the best accuracy rate of 99.29% on training data and 94.92% on testing data. This shows that the SVM model is most suitable to be used to predict the package that will be taken from Best Wedding Organizer customers based on demographic information and customer behavior.

REFERENCES

Akinyelu, Andronicus A,, & Absalom E, Ezugwu, (2019), "Nature Inspired Instance Selection Techniques for Support Vector Machine Speed Optimization," *IEEE Access* 7: 154581–99,

Feng, Rui, Chunlin Song, & Huihe Shao, (2004), "Drifting Model Approach to Modeling Based on Weighted Support Vector Machines," *Journal of Systems Engineering and Electronics* 15(4): 610–14,

Han, Jiawei, & Micheline Kamber, (2006), "Data Mining : Concepts and Techniques (2nd Edition) Bibliographic Notes for Chapter 2 Data Preprocessing," : 1–6,

Huang, Shujun et al, (2018), "Applications of Support Vector Machine (SVM) Learning in Cancer Genomics," Cancer Genomics and Proteomics 15(1): 41–51,

Omara, Ibrahim et al, (2020), "A Hybrid Approach Combining Learning Distance Metric and DAG Support Vector Machine for Multimodal Biometric System," IEEE Access 9,

Srivastava, Durgesh K,, & Lekha Bhambhu, (2010), "Data Classification Using Support Vector Machine," *Journal of Theoretical and Applied Information Technology* 12(1): 1–7,

Janusz, G, (2003), Data mining and complex telecommunications problems modeling, *J, Telecommun, Inform, Technol,, no, 3,* pp, 115-120,

Bounsaythip, C, & Rinta-Runsala, E, (2001), Overview of Data Mining for Customer Behavior Modeling, Research report TTE1-2001-18, VTT Information Technol- ogy.

Damanik, Erond, L, (2018), Menolak Evasive Identity: Memahami Dinamika Kelompok Etnik di Sumatera Utara. *Journal of Social and Cultural Anthropology* 4 (1): 9-22

Somantri, Oman, Slamet Wiyono, & Dairoh Dairoh, 2016, "Metode K-Means Untuk Optimasi Klasifikasi Tema Tugas Akhir Mahasiswa Menggunakan Support Vector Machine (SVM)," *Scientific Journal of Informatics* 3(1): 34–45.

*name of corresponding author

Pratama, Arif, Randy Cahya Wihandika, & Dian Eka Ratnawati, 2018, “Implementasi Algoritme Support Vector Machine (SVM) Untuk Prediksi Ketepatan Waktu Kelulusan Mahasiswa,” *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer* 2(April): 1704–8.

Sari, Bety Wulan, & Fadholi Fat Haranto, 2019, “Implementasi Support Vector Machine Untuk Analisis Sentimen Pengguna Twitter Terhadap Pelayanan Telkom Dan Biznet,” *Jurnal Pilar Nusa Mandiri* 15(2): 171–76.

made adi Pranata, i, and gede sri Darma, 2018, “Jurnal Manajemen Dan Bisnis,” *Jurnal Manajemen dan Bisnis* 15(1): 15–18,

*name of corresponding author