The International Conference on Natural Sciences, Mathematics, Application, and Technology (ICON-SMART) Proceedings
e-ISSN: 2963-3915

Vol. 2, December 2022
pp. 1-9

Clustering of Travel Insurance Cases with K-Modes Algorithm

Sheilta Alphenia[a] and Kariyam[b]

Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Islam Indonesia, Yogyakarta, Indonesia

[a)] Corresponding author: 18611109@students.uii.ac.id
[b)] kariyam@uii.ac.id

Abstract. Travel insurance is protection against risks that may occur when a person travels, including tourist trips. Information about the characteristics of tourists helps insurance companies in creating new products. In this study, travelers will be grouped based on the attributes of age category, income category, the number of families category, education level, type of work, history of health, frequency of travel, and frequency of going abroad. Clustering for mixed data nominal and ordinal usually pay less attention to ordinal attribute information. We use the k-modes algorithm with a measure of proximity that positions the existence of the essence of the sequence on the ordinal attribute. We classified 710 tourists who already have travel insurance into two clusters based on this method. At the same time, as many as 1277 travelers who do not have travel insurance were into four groups. Based on the profiles of each group, we conclude that there are similarities in the characteristics of 290 travelers who do not have travel insurance with 332 travelers who already have insurance. This group is private workers who graduated from college, are 30 years old, have no history of chronic disease, have a family of four, and are upper-middle-income. This group also rarely travels and never abroad. Insurance companies can target prospective tourists with these characteristics in offering their products.

Keywords: clustering, insurance, travelers, k-modes algorithm

INTRODUCTION

In today's digital era, data can be found anywhere. The amount of data is in line with the needs of the data itself which is now also widely used. The existing data is not only numeric but also categorical. From the data, we can generate new information to make the right decision based on reality with existing data. One of the methods for gathering information is cluster analysis. Cluster analysis is a method for partitioning data into several groups with homogeneity within the group and heterogeneity between groups. The grouping of data is based on the calculation of the minimum distance of the data to the center of the cluster. Each group will have its own characteristics that will provide new information. Cluster analysis is widely applied in various sectors such as marketing, economics, health, government, society, and culture.

Cluster analysis with categorical data was introduced by Huang [1], called K-Modes Clustering. The treatment for categorical data cluster analysis was different from the treatment for numerical data. The determination of the cluster center initiated by Huang for categorical data is based on mode measurement. However, the treatment for calculating the distance of data to the center of the cluster on categorical data in the form of nominal and ordinal from the algorithm initiated by Huang is still the same. Nominal data is different from ordinal data because nominal data only produce information in the form of labels while ordinal data produces information in the form of labels and has a sequence meaning so that the treatment between these two types of data scales needs to be distinguished so as not to reduce the information contained in the data. Yuan et al. [2] propose a different distance measurement method for ordinal data so that the order label information from the ordinal data is not omitted. The implementation of Yuan's proposed method is still rarely found, while categorical data in the form of nominal and ordinal are often found in datasets of various fields, one of which is insurance.

According to the news from Kontan [3] during the pandemic, the need for insurance in the health sector continues to grow in line with the need for protection during the pandemic. People try to mitigate risk by buying health insurance products. However, as the number of positive cases of COVID-19 begins to decrease, it is predicted will have an impact on the travel insurance business. This was stated by the Executive Director of the Indonesian General Insurance Association (AAUI) Dody Dalimunthe to Kompas [4] that this prediction was based on the start of many tourist destinations and the allowed traveling activities. Dody also added that insurance companies need to pay attention to the needs of travelers, such as providing attractive insurance products according to the needs of travelers.

Based on these needs to obtain products that are suitable for travelers, it is necessary to further investigate the characteristics of travelers who register (users) with the insurance company. Therefore, it is necessary to study travel insurance data using the cluster analysis method which can be used as a method to determine the characteristics of the user. The characteristics of the user can be used as a reference to make attractive insurance products according to user needs.

LITERATURE REVIEW

K Modes Clustering Algorithm

Cluster analysis is a method for grouping data that have similarities in variables into one of several groups. While the data that lacks the same variables will be placed in different groups [5].

The stages of the K-Modes clustering algorithm in this research are detailed as follows:

1. Determine the nominal and ordinal variables.
2. Determine the number of clusters of *k*.
3. Select the initial cluster center randomly from *k* objects.
4. Calculate the nominal and ordinal distance from the object to the center of the cluster based on the calculation of the Yuan dissimilarity.
5. Group objects to the center of the cluster with the closest distance.
6. Calculate the mode value of each cluster.
7. The mode is then used as the center of the new cluster.
8. Repeat steps 4 to 7 until the center of the cluster remains and no objects move the cluster.

Nominal Dissimilarity Measurement

The equation for similarity nominal data defined in equation 1 below.

$$sim_{nom^k}(x,y) = \frac{f(x,nom^k) \equiv f(y,nom^k)}{\sum_{z \in U} f(x,nom^k) \equiv f(z,nom^k)} \quad 1)$$

where $x, y \in U$ and $nom^k \in \{nom_i^1, nom_i^2, \dots, nom_i^p\}$ also

$$f(x,nom^k) \equiv f(y,nom^k) = \begin{cases} 1\ jika\ f(x,nom^k) = f(y,nom^k) \\ 0\ lainnya \end{cases} \quad 2)$$

so that the calculation of the dissimilarity or distance from the nominal data is defined as $Ndis(x,y) = \sum_{nom^k \in U} Ndis_{nom^k}(x,y)$, where

$$Ndis_{nom^k}(x,y) = 1 - sim_{nom^k}(x,y)\ dan\ k = 1,2,\dots,p \quad 3)$$

Ordinal Dissimilarity Measurement

he equation for similarity ordinal data defined in equation 4 below.

$$Odis_{ord^l}(x_i,x_j) = \begin{cases} 1 - sim_{ord^l}(x_i,x_j) & 1 \le \delta_l \le 2, \\ \left|ord_i^l - ord_j^l\right| Dod_1 & 2 < \delta_l \le 2\eta_{max,} \\ \left|ord_i^l - ord_j^l\right| Dod_2 & \delta_l > 2\eta_{max} \end{cases} \quad 4)$$

Where:

1. $1 \leq \delta_l \leq 2$
   When $1 \leq \delta_l \leq 2$ is true, that means ord^l only have two different values. The inequality between two different values is similar to a nominal attribute. Therefore, the measurement of the inequality of nominal attributes is used to calculate the inequality of the two attribute values in this situation. $sim_{ord^l}(x_i, x_j)$ calculate according to equation 3.
2. $2 < \delta_l \leq 2\eta_{max,}$
   As mentioned above, $\eta_k(k = 1,2, \dots, p)$ is domain size of nom^k. $\eta_{max} = \max\{\eta_1, \eta_2, \dots, \eta_p\}$ is the maximum value of domain size from nominal atrribute. In order to calculate Dod_1, we used $MSOA(ord^l)$. $MSOA(ord^l)$ is a set thath contain nominal attribute nom^t which adjust the conditions of $|\eta_t - \delta_l| = \min\{|\eta_i - \delta_l|\}(i = 1,2, \dots, p)$ where η_t is domain size of $nom^t, \eta_i(i = 1,2, \dots, p)$ is domain size which the nominal attibute is related. After we got $MSOA(ord^l)$, calculated all non-zero $sim_{nom^t}(x, y)$ foreach nominal attribute nom^t from $MSOA(ord^l)$. $sim_{nom^t}(x, y) \neq 0$ when $f(x, nom^t) = f(y, nom^t)$, the required calculation is $sim_{nom^t}(x, y)$ from a pair of the same attribute values for each $nom^t \in MSOA(ord^t)$. After that, $Dod_1 = \min\{sim_{nom^t}\}$ where $nom^t \in MSOA(ord^t)$.
3. $\delta_l > 2\eta_{max}$
   In this condition, we used calculation of Dod_2 below.
   $$Dod_2 = \frac{Dod_1}{2|\delta_l - 2\eta_{max}|} \quad (5)$$
   where $\eta_{max} = \max\{\eta_1, \eta_2, \dots, \eta_p\}$ and Dod_1 calculted using point 2.

Silhouette Coefficient

The quality and strength of the cluster can be seen by using the Silhouette Coefficient method. Silhouette method is based on calculation of distance pair matrix from all data[6]. The silhouette coefficient score is obtained from the average value of the Global Silhouette Index. The steps to calculate the silhouette coefficient score are as follows.

1. Calculate the value of cohesion, can be symbolized by a_i^j, is the average of the i-th distance to all other data that are still in one cluster.
   $$a_i^j = \frac{1}{m_j - 1} \sum_{r=1;\ r \neq i}^{m_j} d(x_i^j, x_r^j) \quad (6)$$
2. Calculate the value of separation, can be symbolized by b_i^j, is the minimum value of the average distance of the i-th data in the cluster j-th in all data other than j .
   $$b_i^j = min_{n=1,2,\dots,k;\, n \neq k} \left\{ \frac{1}{m_n} \sum_{r=1; r \neq i}^{m_n} d(x_i^j, x_r^n) \right\} \quad (7)$$
3. Calculate silhouette coefficient score using equation 8.
   $$s_i^j = \frac{b_i^j - a_i^j}{\max\{a_i^j, b_i^j\}} \quad (8)$$

where:

j = *cluster*
i = data index $(i = 1,2, \dots, m_j)$
m_j = the amount of data in j-th *cluster*
$d(x_i^j, x_r^j)$ = the distance of i-th object with r-th object in the j cluster
m_n = the amount of data in n-th *cluster*
$d(x_i^j, x_r^n)$ = the distance of i-th object with r-th object in the n cluster
a_i^j = average distance of i-th object to all data in same cluster
b_i^j = average distance of i-th object to all data that not in the same cluster with i-th object
s_i^j = *silhouette coefficient score of* i-th object in

RESEARCH METHOD

Data and Source of Data

The population used in this research are customers of a tour and travel service provider in India. The sample used is as many as 1987 registered customers of these service providers divided to 710 travellers who already have insurance and 1277 travellers who do not have insurance. The data used in this study is secondary data taken from the Kaggle website, with the address https://www.kaggle.com/tejashvi14/travel-insurance-prediction-data accessed on December 02, 2021.

Variable Research

The variables used in this study were 8 categorical variables. Six of them have nominal scale and 2 are ordinal scale. Table 1 is a description of the variables used in this research.

TABLE 1. Variables Research and Description.

No	Variable	Notation	Scale
1	**Age** (Age of travellers)	• 0 : Age ≤ 30 • 1 : Age > 30	Ordinal
2	**Employment Type** (The sector in which travellers is employed)	• 0 : Government sector • 1 : Private sector	Nominal
3	**Graduate or Not** (Whether the travellers is college graduate or not)	• 0 : No • 1 : Yes	Nominal
4	**Annual Income** (Type of annual income from travellers)	• 0 : Poor • 1 : Low income • 2 : Middle income • 3 : Upper-Middle income • 4 : High income	Ordinal
5	**Family Members** (Number of members in traveller's family)	• 0 : < 4 • 1 : 4 • 2 : > 4	Nominal
6	**Chronic Disease** (Whether the traveller suffers from any major disease or conditions like diabetes/high BP or asthma)	• 0 : No • 1 : Yes	Nominal
7	**Frequent Flyer** (Derived data based on traveller's history of booking air tickets on atleast 4 different instances in the last 2 years[2017-2019]	• 0 : No • 1 : Yes	Nominal
8	**Ever Travelled Abroad** (Has the customer ever travelled to a foreign country [not necessarily using the company's services])	• 0 : No • 1 : Yes	Nominal

Step of Research

The data analyzed using descriptive statistics to gain surface information. After that, we used K-Modes clustering algorithm to form several clusters which are homogeneus within clusters and heterogeneus between clusters. Silhouette coefficient then used to be based of choosing number of cluster that optimum.

RESULT AND DISCUSSION

Descriptive Statistics

The data used is the data of travel agency service travellers with a total of 1987 travellers. The age of service travellers/users is divided into two categories, namely the age of users above 30 years and the age of service users a maximum of 30 years with most agency service users aged less than 30 years. The type of work of travel agency service users is divided into two types, namely the type of government workers and private and private sector workers. Users of this travel agency service are divided into three types of income, namely middle income, upper-middle income, and high income.

The users of this travel agency service are mostly found among private and private sector workers with most of this income being high income, which is as many as 796 people. Upper-middle income among private and private sector workers consists of 600 people, while middle-income users only consist of 21 people. For service users with the type of government workers, the majority are focused on upper-middle income, which is as many as 293 people. Users with high and middle income in this group have a relatively equal number of comparisons, namely 155 people and 122 people, respectively.

Many service users from agencies in this study have never traveled abroad and have not traveled to at least 4 different destinations within 2 years (2017-2019). Meanwhile, a minority of users have traveled abroad and have traveled to at least 4 different destinations within 2 years (2017-2019). Travel agency service users who have a family of 4 members account for a quarter of the total number of users or to be more precise 25.42%. Service users with <4 family members are almost as many as 4 families, which is 23.65% of the total users. Meanwhile, family members with more than 4 are the majority category, which is 50.93%.

The majority of service users from travel agencies in this study are university graduates, amounting to 85.15% of the total service users of the agency. Meanwhile, only 14.85% of service users are not graduates from universities. Most travel agency service users do not have a chronic disease so they do not sign up for travel insurance. However, users who have chronic diseases are also more likely to not register for travel insurance than those who do.

K-Modes Clustering Algorithm

In this research, K-Modes clustering analysis was used using data on service users who do not have travel insurance are1277 users and already had insurance are 710. Then the two data will be compared to the clusters formed.

Initial Cluster

Before using clustering analysis, the first group center is determined by randomly selecting several k (number of clusters) from the data. In this study, clusters of 2, 3, 4, 5, 6, and 7.

TABLE 2. Inital Clusters

Number of Cluster	Status	Initial Cluster
2	Do not have insurance	{X166, X735}
	Already have insurance	{X464, X264}
3	Do not have insurance	{X362, X1148, X78}
	Already have insurance	{X535, X581, X396}
4	Do not have insurance	{X457, X437, X1007, X172}
	Already have insurance	{X354, X304, X326, X606}
5	Do not have insurance	{X432, X114, X72, X1133, X153}
	Already have insurance	{X329, X546, X378, X77, X671}
6	Do not have insurance	{X1170, X1125, X89, X1002, X289, X1135}
	Already have insurance	{X679, X234, X584, X263, X195, X564}
7	Do not have insurance	{X929, X1117, X775, X1063, X607, X293, X356}
	Already have insurance	{X544, X13, X61, X140, X188, X387, X593}

Nominal Distance

The output of the calculation for nominal distance of the data to the initial cluster for number of clusters 2, 3, 4, 5, 6, and 7. The following shows the calculation of the nominal distance for a cluster of 2 for data users who do not have insurance.

$$Ndis(x_i, x_j) = \begin{pmatrix} 5{,}9973 & 5{,}995 \\ 5{,}9951 & 5{,}9963 \\ 5{,}9929 & 5{,}9941 \\ \vdots & \vdots \\ 5{,}997 & 5{,}9982 \\ 5{,}9961 & 5{,}9973 \\ 5{,}9951 & 5{,}9963 \end{pmatrix}$$

Ordinal Distance

In this research, there are two variables that have an ordinal scale, namely the Age and Annual income variables. Before calculating the ordinal distance on the two variables, first determine the domain size of all the variables used.

TABLE 3. Initial Clusters

No	Variabel	Domain	Domain Size
1	Age	{<=30, 30}	2
2	Employment Type	{ Government Sector, Private Sector/Self employed }	2
3	Graduate Or Not	{Yes, No.}	2
4	Annual Income	{Middle Income, Upper-Middle Income , High Income}	3
5	Family Members	{<4, =4, >4}	3
6	Chronic Diseases	{Yes, No.}	2
7	Frequent Flyer	{Yes, No.}	2
8	Ever Travelled Abroad	{Yes, No.}	2

Age as the first ordinal variable, has a domain size of 2, so the calculation of the distance Age meets the first condition of ordinal dissimilarity. The following shows the distance matrix for clusters of 2 for data of users who do not have insurance.

$$Odis_1(x_i, x_j) = \begin{pmatrix} 1 & 0{,}9978 \\ 1 & 0{,}9978 \\ 0{,}9988 & 1 \\ \vdots & \vdots \\ 0{,}9988 & 1 \\ 0{,}9988 & 1 \\ 1 & 0{,}9978 \end{pmatrix}$$

Annual Income as the second ordinal variable, has a domain size of 3 with the maximum domain size for nominal variables is 3, namely Family Members. This second ordinal variable satisfies the second condition of the calculation of the ordinal dissimilarity. The MSOA value is obtained from finding the difference between domain size from nominal variables and the Annual Income from ordinal variable. The variable with the smallest difference in domain size to the Annual Income variable is Family Members with a difference of 0. The similarity matrix between the data on the Family Members variable produces a 1277x1277 matrix with rounding as follows.

$$Sim_{FamilyMembers}(x_i, x_j) = \begin{pmatrix} 1 & 0{,}0016 & 0 & \dots & 0{,}0016 & 0{,}0016 & 0 \\ 0{,}0016 & 1 & 0 & \dots & 0{,}0016 & 0{,}0016 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0{,}0016 & 0{,}0016 & 0 & \cdots & 1 & 0{,}0016 & 0 \\ 0{,}0016 & 0{,}0016 & 0 & \cdots & 0{,}0016 & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & 1 \end{pmatrix}$$

Dod_1 is a non-zero minimum value of the similarity of variables in MSOA so that the value of Dod_1 is 0.0016. The second ordinal distance matrix that is formed based on the second condition is shown as follows for the number of clusters as much as 2 in the data of users who do not have insurance.

$$Odis_2(x_i, x_j) = \begin{pmatrix} 0{,}0016 & 0 \\ 0 & 0{,}0016 \\ 0{,}0016 & 0 \\ \vdots & \vdots \\ 0 & 0{,}0016 \\ 0 & 0{,}0016 \\ 0{,}0016 & 0 \end{pmatrix}$$

Meanwhile, the data of users who already have insurance obtained a Dod1 value of 0.0024.

Combined Distance and Cluster Placement

The combined distance is the sum of the distances on the nominal and ordinal variables. Then each data is placed in one cluster with the initial cluster having the minimum distance between other initial clusters. The following shows the distance matrix for users who do not have insurance.

$$d(x_i, x_j) = \begin{pmatrix} 6{,}9989 & 6{,}9928 \\ 6{,}9951 & 6{,}9957 \\ 6{,}9932 & 6{,}9941 \\ \vdots & \vdots \\ 6{,}9957 & 6{,}9998 \\ 6{,}995 & 6{,}9989 \\ 6{,}9967 & 6{,}9941 \end{pmatrix} = \begin{pmatrix} Cluster\ 2 \\ Cluster\ 1 \\ Cluster\ 1 \\ \vdots \\ Cluster\ 1 \\ Cluster\ 1 \\ Cluster\ 2 \end{pmatrix}$$

New Initial Clusters and Iteration

The new initial cluster is determined from each cluster mode. The following is the mode of cluster on users data who do not have insurance.

TABLE 4. Mode of the first two clusters

Cluster	**Age**	*Employment Type*	*Graduate Or Not*	*Family Members*	*Annual Income*	*Chronic Diseases*	*Frequent Flyer*	*Ever Travelled Abroad*
1	0	1	1	>4	4	0	0	0
2	1	0	1	>4	3	0	0	0

The mode of each cluster that has been obtained is then used as a new initial cluster and the distance of the data from the new initial cluster is calculated. The distance of the data to each new initial cluster is shown as follows.

$$d(x_i, x_j) = \begin{pmatrix} 6{,}9973 & 6{,}9912 \\ 6{,}9935 & 6{,}9941 \\ 6{,}9965 & 6{,}9973 \\ \vdots & \vdots \\ 6{,}9941 & 6{,}9982 \\ 6{,}9933 & 6{,}9973 \\ 6{,}9968 & 6{,}9941 \end{pmatrix} = \begin{pmatrix} Cluster\ 2 \\ Cluster\ 1 \\ Cluster\ 1 \\ \vdots \\ Cluster\ 1 \\ Cluster\ 1 \\ Cluster\ 2 \end{pmatrix}$$

The mode value of each of these clusters is shown in Table 5.

TABLE 5. Mode of the second iteration two clusters

Cluster	**Age**	*Employment Type*	*Graduate Or Not*	*Family Members*	*Annual Income*	*Chronic Diseases*	*Frequent Flyer*	*Ever Travelled Abroad*
1	0	1	1	>4	4	0	0	0
2	1	0	1	>4	3	0	0	0

Because the mode of the new cluster is the same as the mode of the previous cluster, the iteration is stopped.

Silhouette Coefficient

In this research, the k-modes cluster grouping was calculated for the number of clusters 2, 3, 4, 5, 6 and 7 which was then searched for the optimal number of clusters with the silhouette coefficient. The silhouette coefficient values for each resulting cluster are described in Table 6.

TABLE 6. Silhouette Coefficient Score for Each Cluster

Number of Cluster	Status	Silhouette Coefficient Score
2	Do not have insurance	0.0002460585
	Already have insurance	**0.000449372**
3	Do not have insurance	0.0002759532
	Already have insurance	0.000378861
4	Do not have insurance	**0.0002947690**
	Already have insurance	0.000363028
5	Do not have insurance	0.0002748279
	Already have insurance	0.000285797
6	Do not have insurance	0.0002624325
	Already have insurance	0.000252003
7	Do not have insurance	0.0002290883
	Already have insurance	0.000417435

According to the silhouette cluster score, the highest value is used so that in the case study the data of travellers who do not have insurance are grouped into 4 clusters, which is 0.00029477, while the data of travellers who already have insurance are grouped into 2 clusters with a silhouette value. of 0.000449372.

Profilization

The groups formed from data of travellers who do not have insurance are grouped into four groups with details:

1. The first group is a group that consists of the majority of the age group of 30 years and under or the group of early workers and does not have a chronic disease. The sector of work in this group is workers in the private sector with a college graduate background. This group has the majority of family members numbering more than 4 members. In this group, most of them have never been abroad and are not service users who frequently fly. Service users in this group are mostly people with upper middle income. This group consists of as many as 570 users.
2. The second group is a group that consists of the majority of the age group of 30 years and under or the group of early workers and does not have a chronic disease. The sector of work in this group is workers in the private sector with a non-college graduate background. This group has the majority of family members numbering more than 4 members. In this group, most of them have never been abroad and are not service users who frequently fly. Service users in this group are mostly people with upper middle income. This group consists of as many as 196 users.
3. The third group is a group consisting of the majority of the age group of 30 years and under or groups of early workers and have chronic diseases. The sector of work in this group is workers in the private sector with a college graduate background. The majority of this group has a family of more than 4 members. In this group, most of them have never been abroad and are not service users who frequently fly. Service users in this group are mostly people with upper middle income. This group consists of as many as 221 users.
4. The fourth group is a group that consists of the majority of the age group of 30 years and under or the group of early workers and does not have a chronic disease. The sector of work in this group is workers in the private sector with a college graduate background. The majority of this group has a family of 4 members. In this group, most of them have never been abroad and are not service users who frequently fly. Service users in this group are mostly people with upper middle income. This group consists of as many as 290 users.

While the data of users who have insurance are grouped into two groups with the following details.

1. The first group is a group consisting of 332 which the maximum age is 30 years and is a private worker. In this group, the majority of service users are college graduates with 4 families and upper-middle income. Users in this group rarely travel or have never been abroad. In this group, the majority of users do not have a history of chronic disease.
2. The second group is a group consisting of 378 which are more than 30 years old and are private workers. In this group, the majority of service users are college graduates with a family of 4 and high income. Users in this group rarely take different flights but have been abroad. In this group, the majority of users do not have a history of chronic disease.

When compared between the characteristics of the cluster formed in the fourth group of service users who do not have insurance, they are the same as the characteristics of the first group of service users who have insurance. The similarity of these characteristics can make it possible for service users who do not have insurance in the fourth group to have insurance. This can be achieved by providing special offers.

CONCLUSION

We already classified 710 tourists who already have travel insurance into two clusters based on this method. At the same time, as many as 1277 travellers who do not have travel insurance into four groups. Based on the profiles of each group, we conclude that there are similarities in the characteristics of 290 travellers who do not have travel insurance with 332 travellers who already have insurance. This group is private workers who graduated from college, are 30 years old, have no history of chronic disease, have a family of four and are upper-middle-income. This group also rarely travel and never abroad. Insurance companies can target prospective tourists with these characteristics in offering their products.

ACKNOWLEDGMENTS

Processed of making this reseach, the authors had the pleasure of working with Direktorat Penelitian dan Pengabdian Masyarakat Universitas Islam Indonesia.

REFERENCES

1. Z. Huang, *Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values* (Kluwer Academic Publishers, Netherlands, 1998), pp. 283-304.
2. F. Yuan, Y. Yang and T. Yuan, *A dissimilarity measure for mixed nominal and ordinal attribute data in k-Modes algorithm*, (Springer, Berlin, 2020), pp. 1498-1509.
3. Kontan, *Ada pandemi, bisnis asuransi kesehatan terus melaju*, 30 Augst 2021. [Online]. Available: https://keuangan.kontan.co.id/news/ada-pandemi-bisnis-asuransi-kesehatan-terus-melaju.
4. Kompas, *Bisnis Asuransi Perjalanan Diprediksi Bakal Mulai Cerah*, 27 October 2021. [Online]. Available: https://money.kompas.com/read/2021/10/27/203000826/bisnis-asuransi-perjalanan-diprediksi-bakal-mulai-cerah.
5. S. Susanto and D. Suryadi, *Pengantar Data Mining Menggali Pengetahuan dari Bongkahan Data*, (ANDI, Yogyakarta, 2010).
6. D. A. I. C. Dewi and D. A. K. Pramita, Matrix: Jurnal Manajemen Teknologi dan Informatika, 9(3): 102-109 (2019).