ISSN 2278-3091
Volume 9, No.5, September - October 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse57952020.pdf
https://doi.org/10.30534/ijatcse/2020/57952020

Data Mining as Tools to Improve Marketing Campaign

Nilo Legowo[1], Kevin HWibawa[2]
[1]Information Systems Management Department, BINUS Graduate Program-Master of Information Systems Management, Bina Nusantara University, Jakarta, Indonesia, nlegowo@binus.edu
[2]Information Systems Management Department, BINUS Graduate Program-Master of Information Systems Management, Bina Nusantara University, Jakarta, Indonesia, kevin.wibawa001@binus.ac.id

ABSTRACT

Role of data has big ipact and many uses for companies. Data is commonly used to help management make better decisions. Data can portray what customers a company is having, what customers want, what are customers characteristics, and so on. At the insurance companies, the use of data can help marketing to create a campaign that right on targeted segment. Insurance companies can create a campaign that aim to target potential customers with certain demographics labeled on them, such as income, educational background, age, etc. This work describes data mining approaches aim to build a predictive model. CRISP-DM framework is used to define the processes and tasks in data mining projects. This study will use K-means clustering technique.

Key words: CRISP-DM, Data Mining, K-means, Insurance, Marketing.

1. INTRODUCTION

Insurance industry in Indonesia is facing challenge. With the huge number of populations, there are only 15.8 out of 100 persons have been had insurance literacy; and 12 out of 15.8 persons have been covered by insurance [1]. It sums up that there are still many citizens in Indonesia do not have insurance[2]. This can bring insurance companies growth opportunities.

The next discussion is how insurance companies respond to those growth opportunities. Acquiring new customers has become one of many major business problems. Many companies put extra efforts focusing on sales department that is not worth the cost. Conventional sales approach is to targeting potential customers who are assessed can meet policy constraints. Studies showed that this effort yield below expectations. In contrast, doing sales effort with the guidance of quantitative data mining approach can lead better return and successful results[3].

Data is increasing rapidly day to day due to rapid development of information technology and its usage by the public,Useful information can be obtained when these raw data's are studied properly[4]. Data mining can be defined as extraction process of previously unknown information from a large datasets and analysis process of hidden patterns of data based on different perspectives into useful information, which is collected and assembled in common areas, such as data mining algorithms, providing efficient analysis, and so on[5].Widely perceived that the right use of data in data mining should cut costs and increase revenue at many companies. The use of data mining in insurance industry can help to identify population segments among existing customers through which potential customers could be targeted[3]. By using cluster analysis, target groups can be identified based on the existing available data of policy holders. After that, companies can create customer-centric products that fit target population. Hence, marketing campaign will be more efficient.

2. LITERATURE REVIEW

2.1 Insurance

Insurance concept can be defined as transferring risks that can arise from certain activities to third party using velocity of money to compensate economic losses [6],[7]. In this global world, insurance companies offer various type of products, such as life insurance, auto insurance, home insurance, recreational insurance, and so on. Many aspects of life can be insured since whatever tasks humans do, those will carry risks as well. In addition, due to great competition in insurance sectors, insurance companies compete in offering best insurance products with affordable and reasonable price which are expected could fulfill needs of potential customers[6].

Several studies show that many factors influencing insurance demand, such as occupation, income, lifestyle, education level, and so on[6], [8], [9]. These attributes can be value to insurance companies if they were used to giving clear prediction. In this work, available customers data from insurance company is used for data mining process to give insights of what products fit certain potential customers.

2.2 Data Mining

Data mining a process extracting of dredging or gathering important data that has been known before but can be understood and useful form large database, It is useful to make important business decision [10]. Data mining are activity of analyzing large data to look data patern. Data

mining is grouping data into small group and prediction the aim to value of a continuous variable [10].

2.3 Clustering Technique: K-means

K-means is a typical clustering algorithm widely used for clustering large set of data in data mining. Mac Queen firstly propose that K-means was one of the most simple in 1967, it was applied to solve unknown cluster[10].
The algorithm consists of two sperate phases there are select k center randomly and take each data object to the nearest center[11]. Distance between each data object and the cluster center is determine by euclidean distance[11]. The first step when all the data object is included in some cluster is completed and early grouping is done. Recalculating the average of the early formed cluster. This iterative process continues repeatedly until the criterion function become minimum.

Supposing that the target object is x, xi indicates the average of cluster Ci, criterion function is defined as follows:

$$E=\sum_{i=1}^{k}\sum_{x\in C_i}\left|x-x_i\right|^2$$

E is the sum of the squared error of all objects in database. The distance of criterion function is Euclidean distance, which is used for determining the nearest distance between each data object and cluster centre. The Euclidean distance between one vector x=(x1 ,x2 ,...xn) and another vector y=(y1 ,y2 ,...yn),

The process of k-means algorithm as follow:

Input: Number of desired clusters, k, and a database D={d1, d2,...dn} containing n data objects.

$$d(x_i,\ y_i)=\left[\sum_{i=1}^{n}(x_i-y_i)^2\right]^{1/2}$$

Output: A set of k clusters

Steps:

1) Randomly pick k data objects from dataset D as initial cluster centers.
2) Repeat the process to randomly select;
3) Count the distance between each data object di (1 <= i<=n) and all k cluster centers cj(1<=j<=k) and assign data object di to the nearest cluster.
4) For each cluster j (1<=j<=k), recalculate the cluster centeruntil no changing in the center of clusters

2.4 CRISP-DM Framework

CRISP-DM is most widely used methodology for data developing data mining project. Many leading companies using CRSIP-DM for data mining project. CRISP-DM is a vendor independent methodology so it can be used with any data mining tools and can be applied to solve any data mining problem. CRIS-DM is divided into 6 phases. This phase is described in figure below

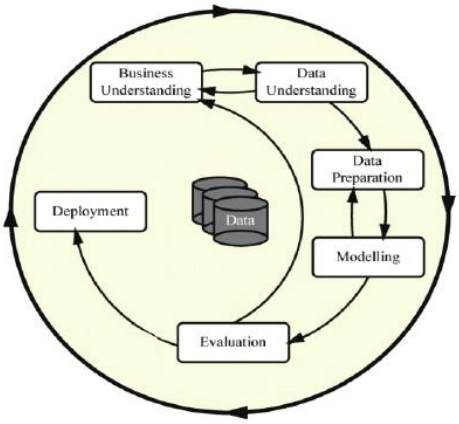

Figure 1: CRISP-DM Framework

- Business Understanding: In this stage, we identify business problem, user needs, and what kind of solution we can offer.
- Data Understanding: In this stage, we identify what kind of data we will be using to solve the problem.
- Data Preparation: In this stage, we select the data and do features to the data, such as selection, extraction, and engineering.
- Modelling: In this stage, we select and create the model and choose values for the various model parameters that fits the model, so that we can optimize the model.
- Evaluation: In this stage, we evaluate the model and to see if our business objective was successful.
- Deployment: In this stage, we deploy the model.

2.5 Davies-Boulding Index

Davies- Boulding Index (DBI) is one of methods for measuring and evaluating clustering method[12], [13]. By using DBI, we can maximize inter-cluster distance and minimize distance between point in a cluster[12]. In a simple way, DBI computes the quality of clustering that has been conducted.

2.6 Rapid Miner

In this work, we use Rapid Miner to do our works. Rapid Miner is one of software platforms is used to do data loading, data transformation, data preprocessing and evaluation, predictive analytics and statistical modeling data deployment. With the basis of Java programming language, we can do data mining experiments from interchange format and automating large-scale experiments[14]. In addition, transparent and efficient data handling are provided by Rapid Miner.

3. RESEARCH METHODOLOGY

In this paper, we use the CRISP-DM Framework. CRISP-DM has been the most used methodology in data mining. Therefore, we choose it as our model. with crisp-dm

we can understand the process of data mining starting from business understanding, data understanding to the deployment stage. We could say CRISP-DM is a methodology that covers the process of data mining from data collection to implementation phase. The six phases of CRISP-DM are shown in Fig.6

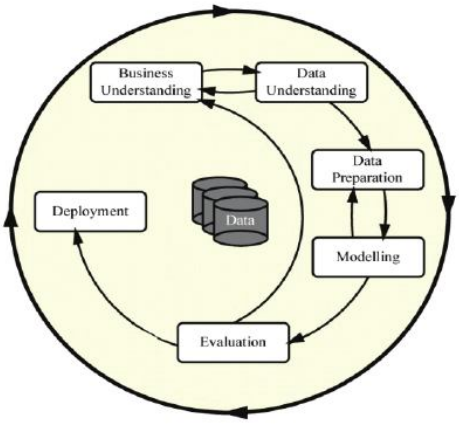

Figure 2: CRISP-DM Approach

1) Business Understanding

In this work, data mining apporaches use to help company create a customer-centric product targeting new customer that have not been reached or small customer segment. this approach is also expected to help companies to make the right marketing campaigns.

2) Data Understanding

This stage begins with the collection of data from the company database. Production data that contains customer information and total premiums will be processed in this study. The files provided are:

- DataProduction.csv is a dataset containing customer data details such as name, age, gender, age, and city. The dataset consists of 15,541 records.

3) Data Preparation

After understanding the data used, in this stage the data will be explored and adjusted as needed. There are four steps in the data pre-processing which are data i) consolidation, ii) data cleaning, iii) data transformation and iv)reduction[15].This step is done to improve the quality of the data itself by avoiding missing data, data errors (noisy data), and inconsistent data in the dataset.The steps taken are data cleaning and data transformation. In this paper gender will be transform into numeric. Male = 1, Female =2.

4) Modeling

The algorithm used in this training data is k-means clustering. The RapidMiner workflow for k-means clustering for this study case is as follows :

1. Normalize Data

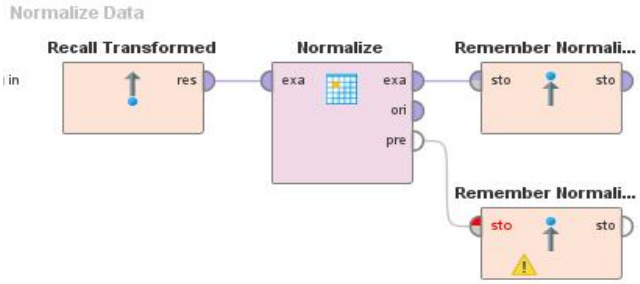

Figure 3: Normalize Process

Table 1: Normalize Nodes and Description

Node	Description
Recall Transformed	Recall transformed data
Normalize	Standardize all columns
Remember Normalized	Remember normalized data that can be use later
Remember Normalization Model	Remember normalized model that can be use later

Before we do custering, first we normalize the dataset using linear transformation to convert data which improve the accuracy of k-means clustering algorithms . Parameter use in Normalize node is Z-transformation method.

2. Clustering

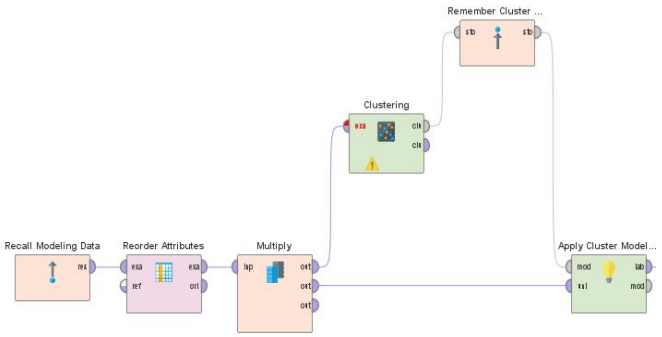

Figure 4: Clustering Normalize Data

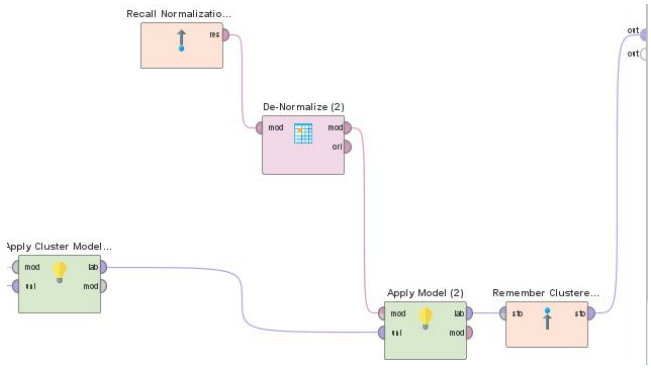

Figure 5: Clustering De-Normalize Data

Table 2 : Clustering Nodes and Description

Node	Description
Recall Modelling Data	Recall data that has been normalized
Reorder Attribute	Order columns alphabetically
Multiply	Copy data to be clustered later
Clustering	Create cluster model
Remember Cluster Model	Remember the cluster model to use later
Apply Cluster Model on Data	Use k-means model to data provided
Recall Normalization Data	recall data that has been normalized
De-Normalize	Turn the model into de-normalization model
Apply Model	Apply de-normalization model to clustered data
Remember Clustered Model	Remember the clustered data

In clustering the first process is to create cluster model based on normalized data and then turn the model into de-normalized mode. Parameter use in the cluster are k min: 2 and k max: 10 and it will runs 10 times till the best number of cluster is obtained.

5) Evaluation

At the evaluation phase, based on the algorithm (k-means clustering) the numbe of cluster obtained will be measured by cluster distance performance using Davies Bouldin Index.

6) Deployment

After conducting training data from the phases above, in this deployment phase the results of the training data in the form of clusters will be presented to the business development manager and board of director of the company so this result can benefit the company and it can be can be taken into consideration in making customer-centric product which is the purpose of making this paper.

4. EXPERIMENT AND RESULT

4.1 Data Understanding

Life insurance customer data can be obtained from the database of PT XYZ insurance product sales in 2018 2019, amounting to 15541 data. consists of 5 attributes. attributes can be seen in the following table:

Table 3: Attribute

Name	Description and Values
Customer Name	The name of the customer who bought a life insurance product (varchar)
Age	Age of the customer (numeric)
Gender	Gender of the customer (binary: “Male”,”Female”
Premi	The total premium paid by the customer for an insurance (numeric)
City	The city where the customer lives (categorical :”Jakarta”,”Bandung”,”Medan”,etc)

4.2 Data Preprocessing

At this stage several stages are carried out on the data before processing. The RapidMiner workflow for preprocessing is as follows:

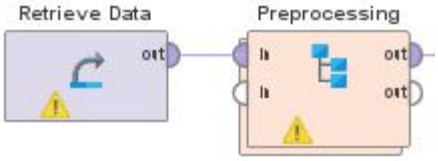

Figure 6: Preprocessing

Table 4. Preprocessing Nodes and Description

Node	Description
Retrieve Data	Load dataset
Preprocessing	All preprocessing steps happening inside this operator

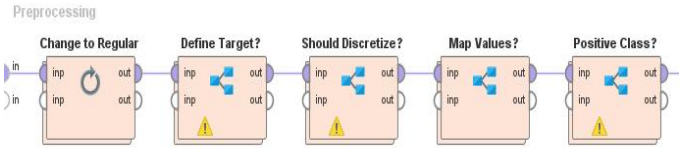

Figure 7: Preprocessing Steps(1)

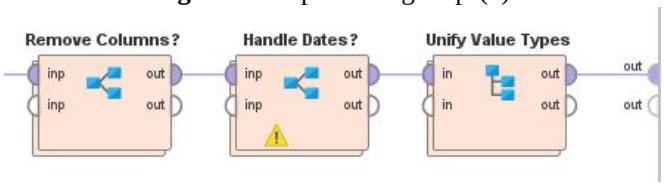

Figure 8: Preprocessing Steps(2)

Table 5. Preprocessing Steps Nodes and Description

Node	Description
Change to Regular	Change role of all columns to ‘regular’
Define Target?	To define a target column
Should Discretize	To discretize numerical target column
Map Values	Should we map nominal values?
Positive Class	Do we need to define positive class?
Remove Columns	To Remove Columns that not needed
Handle Dates	Should we handle dates?
Unify Value Types	Unify all value types

4.3 Result and Visualitation

After the data has been understood and preprocessed, clustering is done using the k-means algorithm. the results of the clustering are then visualized into a cluster model visualizer. The RapidMiner workflow for Visualization as follows :

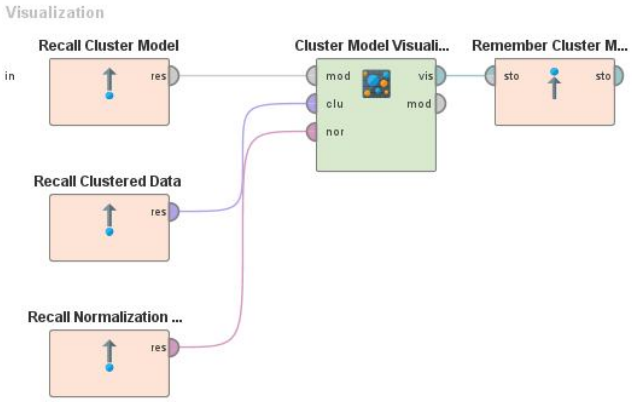

Figure 9: Visualization

As for the node seen in the workflow and the description is as follows:

Table 6. Visualization Nodes and Description

Node	Description
Recall Cluster Model	Recall cluster model created
Recall Cluster Data	Recall data clustered
Recall Normalization Model	Recal Normalized model
Cluster Model Visualizer	Creates the cluster model visualizer
Remember Cluster Model Visualizer	Remember cluster model visualizer

The result shown taking K=4 :

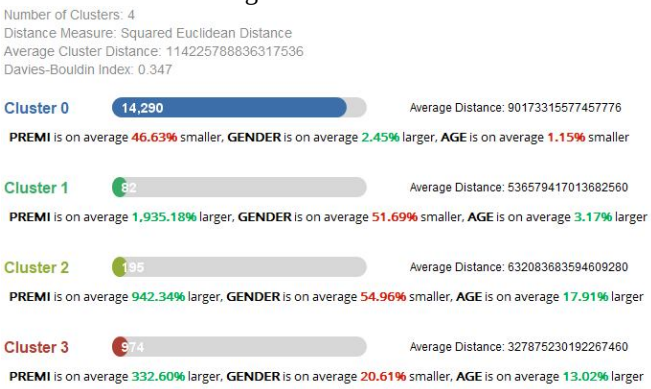

Figure 10: Clusters Overview in Rapid Miner (K=4)

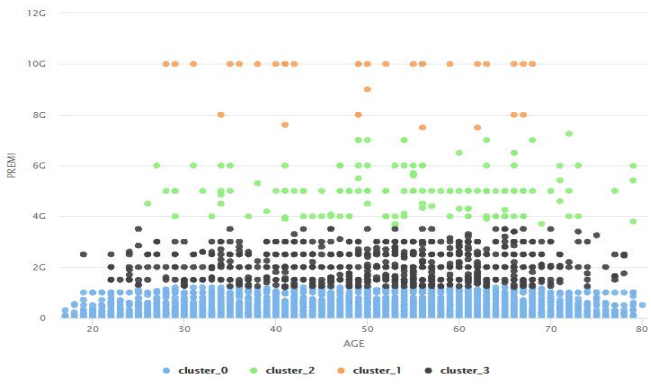

Figure 11: Clusters formed by K-means in Rapid Miner

Table 7: Centroid Table

Cluster	Age	Gender	Premi
Cluster_0	47.507	1.595	303104522.268
Cluster_1	48.841	1.280	9702439024.390
Cluster_2	53.390	1.262	4993589487.179
Cluster_3	51.880	1.461	2101742582.401

Based on the results obtained, we can see that the average age of a customer who buys a life insurance product is 45+. This can be said to be dangerous for the condition of the company where older people have a greater frequency of claims compared to young people. from figure 9 it can be seen that cluster 0 has the most number of customers which is 14290 customers. with an average premium they pay is 300 million. this can be said to be good for the company because 300 million is a relatively large premium in life insurance. Moreover, from cluster 2 we see that there are 82 customers who pay a very large premium, which is an average of 9 billion. Clusters 2 and 3 totaling 1169 customers also look positive, they want to pay a premium above 1 billion. But the thing that must be a concern is the average age above 50 in clusters 2 and 3 will greatly affect the frequency of claims. for gender itself, it can be said that this insurance reaches men and women evenly where there is no striking difference between the two genders who buy life insurance.

4.4 Evaluation

In this paper, Davies Bouldin Index is used to evaluate and determine the best number of clusters that can be obtained. In RapidMiner this operation can be performed on a cluster distance performance segmentation. The overflow in RapidMiner is as follows:

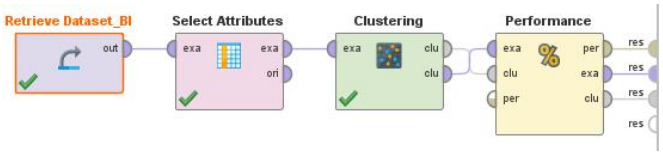

Figure 11: Cluster Distance Performance
T

Table 8: Evaluation Nodes and Description

Nodes	Desccription
Retrieve Dataset	To retrieve dataset use in the process
Select Attributes	Select only attributes needed to perform clustering
Clustering	Create cluster
Performance	Evaluate cluster number created based on Davies Bouldin Index

With k = 4 score obtained is 0,347. This score can be said to be quite good because the minimum score that must be obtained is 0 and a lower score indicates better clustering. when compared if k = 3 the score obtained is 0.389. because

the score k = 4 is smaller than the score k = 3 it can be said that the cluster in the dataset used in this case study is better when divided into 4.

5. CONCLUSION

The results of the data mining approach carried out in this paper give a clear idea that the premiums obtained by insurance companies are relatively large and profitable for the company in the long run. But the average age of a customer who buys a company insurance product is classified as having a negative impact because the claim ratio will increase as the customer agesFor this reason, the results obtained from this paper suggest that life insurance companies make a product that can reach young people or millennials, where currently millennial is the main focus of various markets. this is also encouraged by the growing number of millennial populations. Gender is not a problem in the formation of products because of the results that can be said that insurance products can reach men and women. The results of this case study can be developed in the future by including city variables to help the company design a more targeted marketing campaign. Applying data mining approach , insurance companies can now translate large amounts of sales data into effective business insight. They can avoid risks, analyze their customer patterns, reduce costs incurred, even make a product that meets the needs of the market or follow the existing trends that will ultimately add to the company's revenue.

REFERENCES

[1] A. Kusumawardhani, DIni; Sagala, Saut; Hafiz, Ichsan; Ramadhani, "Insurance for low-income families as an earthquake risk financing instrument," 2019.

[2] N. Laoli, "Kesadaran masih rendah, Akrindo gencar melakukan literasi asuransi," *2019*, Oct-.

[3] A. B. Devale, "Applications of Data Mining Techniques in Life Insurance," vol. 2, no. 4, pp. 31–40, 2012.

[4] K. Singh Rawat and C. Author, "Comparative Analysis of Data Mining Techniques, Tools and Machine Learning Algorithms for Efficient Data Analytics," vol. 19, no. 4, pp. 56–61, 2017.

[5] K. Wisaeng, "A Comparison of Different Classification Techniques for Bank Direct Marketing," *Int. J. Soft Comput. Eng.*, 2013.

[6] P. K. Paudel and A. Silwal, "General public awareness in life insurance."

[7] J. Zhou, "Social relations and insurance awareness of residents," *Anthropologist*, vol. 24, no. 2, pp. 551–560, 2016.

[8] Waseem-Ul-Hameed, M. Azeem, M. Ali, S. Nadeem, and T. Amjad, "The role of distribution channels and educational level towards insurance awareness among the general public," *Int. J. Supply Chain Manag.*, vol. 6, no. 4, pp. 308–318, 2017.

[9] O. H. S. H.Sanjeewa, W.S.; Hongbing, "DETERMINANTS OF LIFE INSURANCE CONSUMPTION IN EMERGING INSURANCE MARKETS OF SOUTH-ASIA," *Int. J. Information, Bus. Manag.*, vol. 11, pp. 109–129, 2019.

[10] N. Shi, X. Liu, and Y. Guan, "Research on k-means clustering algorithm: An improved k-means clustering algorithm," *3rd Int. Symp. Intell. Inf. Technol. Secur. Informatics, IITSI 2010*, pp. 63–67, 2010.

[11] A. Fahim, A.-B. M.Salem, F. Torkey, and M. Ramadan, "Efficient enhanced k-means clustering algorithm," *J. Zhejiang Univ. Sci. A*, vol. 7, pp. 1626–1633, 2006.

[12] A. K. Singh, S. Mittal, P. Malhotra, and Y. V. Srivastava, "Clustering Evaluation by Davies-Bouldin Index(DBI) in Cereal data using K-Means," *Proc. 4th Int. Conf. Comput. Methodol. Commun. ICCMC 2020*, no. Iccmc, pp. 306–310, 2020.

[13] I. M. K. Karo, A. F. Huda, and K. MaulanaAdhinugraha, "A cluster validity for spatial clustering based on davies bouldin index and Polygon Dissimilarity function," *Proc. 2nd Int. Conf. Informatics Comput. ICIC 2017*, vol. 2018-Janua, pp. 1–6, 2018.

[14] A. Pandey, "Study and Analysis of K-Means Clustering Algorithm Using Rapidminer A CASE STUDY ON STUDENTS ' EXAM RESULT Abhinn Pandey," vol. 4, no. 12, pp. 60–64, 2014.

[15] V. R. Sayoc, T. K. Dolores, M. C. Lim, L. Sophia, and S. Miguel, "International Journal of Advanced Trends in Computer Science and Engineering Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse68832019.pdf Computer Systems in Analytical Applications," vol. 8, no. 3, pp. 195–200, 2019.