

Hardware sales forecasting using clustering and machine learning approach

Rani Puspita, Lili Ayu Wulandhari

Department of Computer Science, BINUS Graduate Program-Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia

Article Info

Article history:

Received Nov 2, 2021

Revised Jun 1, 2022

Accepted Jun 20, 2022

Keywords:

Clustering
Evaluation
Forecasting
Hardware sales
Machine learning

ABSTRACT

This research is a case study of an information technology (IT) solution company. There is a problem that is quite crucial in the hardware sales strategy which makes it difficult for the company to predict the number of various items that will be sold and also causes the excess or shortage in hardware stocking. This research focuses on clustering to group various of items and forecast the number of items in each cluster using a machine learning approach. The methods used in clustering are k-means clustering, agglomerative hierarchical clustering (AHC), and gaussian mixture models (GMM), and the methods used in forecasting are autoregressive integrated moving average (ARIMA) and recurrent neural network-long short-term memory (RNN-LSTM). For clustering, k-means uses two attributes, namely "Quantity and Stock" as the best feature in this case study. Using these features the k-means obtain silhouette results of 0.91 and davies bouldin index (DBI) values of 0.34 consisting of 3 clusters. While for forecasting, RNN-LSTM is the best method, where it produces more cost savings than the ARIMA method. The percentage of the difference in saving costs between ARIMA and RNN-LSTM to the actual cost is 83%.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Rani Puspita

Department of Computer Science, BINUS Graduate Program-Master of Computer Science

Bina Nusantara University, Jakarta, Indonesia 11480

Email: rani.puspita@binus.ac.id

1. INTRODUCTION

Business lately has become very fast and growing. This makes companies compete with each other [1]. Many companies are engaged in information technology (IT) Solutions. One of the businesses whose development will always increase is the hardware sales business. With the help of sales in marketing products, both hardware and software, it really helps the company financially. But so far, the company has had quite a crucial problem in its hardware sales strategy which makes it difficult for the company to predict the number of items to be sold and also sometimes causes the company to experience excess or shortage in hardware inventory.

According to [2] sales forecasting can be used for companies or other to anticipate things that will come. If the company does wrong in forecasting sales, things can happen that are not desirable. For example, the company cannot meet the sudden increase in consumer demand. Or maybe consumer demand is not in accordance with the company's estimates so that the existing goods are not sold. In other words, the company may experience excess stock of goods. This of course can bring losses to the company. This is in line with the opinion [3]. According to him, sales forecasting is very important to do. Sales forecasting refers to predicting the future by assuming factors in the past (in this case it can be data from the past) that will have an influence in the future. Clustering also very much needed to see hardware cluster which one is classified

as high, medium, or low according to the previously determined characteristics that become the reference for doing forecasting. In addition to using data on goods sold and existing stock, other characteristics in the data can also be used to analyze clusters of the hardware data sold. It aims for a promotional strategy that can be used by the company [4].

The clustering method that is most widely used is k-means clustering method [5]. K-means is an algorithm used in grouping which separates data into different clusters. Basically the use of this algorithm depends on the data obtained and the conclusions to be reached at the end of the process. K-means clustering is a method for performing clusters that are affected by the selection of the initial cluster centroid [6]. So that in the use of the k-means clustering algorithm, there are two rules. The first rule is to determine the number of clusters that need to be included and the second is to have attributes of type numeric because clustering can only process numerical data [7]. Besides k-means clustering, there is also agglomerative hierarchical clustering (AHC). AHC is a clustering technique that forms a hierarchy so as to form a tree structure. Thus, the grouping process is carried out in stages or stages. There are 2 methods in the hierarchical clustering algorithm, namely agglomerative (bottom-up) and divisive (top-down) [8]. In addition there are also gaussian mixture models (GMM). GMM is a model consisting of components of gaussian functions [9].

Then one method that is very popular and can be used for forecasting is using the recurrent neural network-long short-term memory (RNN-LSTM) method. LSTM is a method that can be used to study a pattern in time series data. LSTM is a type of RNN [10]. This is in line with the opinion [11] which states that the LSTM performs better in practice. LSTM is universal in other words LSTM provides enough network units to be able to calculate whether a conventional computer can calculate, as long as it has the right weight matrix, which can be viewed as a program. There is also another algorithm with basic time series data, namely ARIMA. Autoregressive Integrated Moving Average (ARIMA) is a forecasting technique that uses a correlation technique between a time series. Then, the model finds patterns of correlation between series of observations [12]. Based on the background, a research focuses on the implementation of data mining with machine learning methods as a solution to sales problems that often experience excess or lack of stock in warehouses and to find out hardware sales strategies and increase turnover in the company.

2. RELATED WORKS

Researchers read and conduct literature studies in journals related to the chosen research topic. It aims to learn all things related to research and to assist researchers in identifying research problems so that this will support the course of research. The summary of the related works to data mining in sales is shown in Table 1, the summary of the related works to clustering is shown in Table 2 and the summary of the related works related to forecasting is shown in Table 3.

Table 1. Related works for data mining in sales

References	Method	Process	Result
Fithri and Wardhana [13]	k-means clustering	Data collection, data mining, perform analysis with k-Means clustering, testing with davies bouldin index (DBI) values	Successfully implemented the k-means clustering algorithm for the sales cluster. The results of testing using DBI values produces a value of 0.2.
Johannes and Alamsyah [14]	Decision tree	Cross-industry standard process for data mining (CRISP-DM).	Managed to determine the prediction of the number of items sold by the viewers, the price, and the type of shoes.
Soepriyanto <i>et al.</i> [15]	k-nearest neighbor (k-NN) and Naïve Bayes	Data collection, analysis, implementation, testing.	Successfully predict stock prices. Naïve Bayes produces an accuracy value of 69.38, and the k-NN method produces an accuracy value of 67.25%.
Nadeak and Ali [16]	Apriori	Data mining, association, Apriori algorithm, testing.	Successfully utilize artificial intelligence techniques in drug sales.
Edastama <i>et al.</i> [17]	Apriori	Data cleaning, data integration, data selection.	Succeeded in obtaining information on the most in-demand items so that it can be used to increase sales growth and marketing of eyewear.

Based on Table 1, it can be concluded that in the sale of goods, analysis needs to be carried out to assist the company in managing sales strategies and of course helping the company in increasing revenue. Researchers will use clustering and forecasting. In this study, researchers will use k-means clustering, AHC and GMM in clustering and use ARIMA and RNN-LSTM methods for forecasting. This research will certainly be able to increase profits for the company and can also help the company to find out the sales strategy for the following year.

Table 2. The summary of the related works in clustering

References	Method	Process	Result
Puspita and Sasmita [18]	k-means clustering	Determine the number of clusters, determine the centroid value of each cluster, calculate the distance between data, and calculate the minimum object distance.	Successfully implemented the k-means algorithm in classifying tourist visits to the city of pagar alam to increase visitors.
Rani <i>et al.</i> [19]	k-means clustering, and frequent pattern (FP) growth	Data collection, k-means algorithm and FP Growth algorithm, analyzing data, implementing data.	Succeeded in grouping student score data to make it easier for students to take expertise courses in the next semester.
Irawan [20]	k-means clustering	CRISP-DM.	Successfully applied data mining techniques with the k-means clustering method which aims to help students determine the correct course according to the established criteria.
Shen <i>et al.</i> [21]	GMM	Description of the operation dataset, analysis of heating load patterns, GMM clustering for heating load patterns, prediction model, dan evaluation of the proposed models.	It can be seen that GMM can help analyze the timing and energy signals of each sub-pattern.
Rashid <i>et al.</i> [22]	k-means Clustering, and GMM	Adding peripheral cluster, dataset description, conditioning on previous frames, adding constraints, combinatorial clustering (k-means and GMM), determine K, data processing k-means and GMM.	Successfully applied the k-means and GMM methods and tested the method using sparse multidimensional data obtained from the use of video game sales all around the world.

Table 3. The summary of the related works in forecasting

References	Method	Process	Result
Xu <i>et al.</i> [23]	ARIMA, and deep belief network (DBN)	Data collection, training data, prediction, testing.	Successfully demonstrated that the model has a high predictive accuracy and may be a useful tool for time series forecasting.
Gupta <i>et al.</i> [24]	Support vector machine (SVM), prophet forecasting model, and linear regression	Data collection, perform SVM, perform linear regression, perform prophet forecasting model, train and test models.	Successfully predict active rate, death rate, and cured rate in India by analyzing COVID-19 data.
Malki <i>et al.</i> [25]	ARIMA	Dataset description, perform ARIMA models, model selection, data normalization, experimental result, and evaluation.	The study predicts that there could be a second rebound of the pandemic within one year. Based on this research, this helps the government to act quickly.
Alabdulrazzaq <i>et al.</i> [26]	ARIMA	Analysis, ARIMA parameters optimization, ARIMA model validation.	Managed to apply the Arima model for the prediction of Covid 19 in Kuwait and get precise and good accuracy.

3. THEORY AND METHODS

In clustering, researchers will use the k-means clustering, AHC and GMM methods. Then for forecasting, researchers will use the ARIMA and RNN-LSTM methods. The following is an explanation of the theory and methods for clustering and forecasting.

3.1. K-means clustering

K-means clustering is one of the techniques of clustering in the data mining modeling process without supervision and method of grouping data by partition. The data are grouped into several groups and each group has characteristics that are similar to or the same as the others but with other groups having different characteristics [27]. In other words, k-means clustering is a similar container of objects. If objects whose behavior is closer, they will be grouped in one class and those that are far or not similar are grouped in clusters different [27]. The clustering steps with the k-means algorithm are: i) Define K/N clusters; ii) Initialize the centroid randomly; iii) Find the nearest object using Euclidean distance; iv) Recalculate the data in each cluster to get the mean; and v) Restore data and put it back it to centroid. If the data in the cluster does not change, then the step cluster stops but if the center cluster is still changing, then it must return to number 3 until the cluster does not change anymore. Steps of k-means clustering can be seen in the process flow Figure 1. The formula Euclidean distance according to [28] is:

$$d(x_i, y_i) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (1)$$

Where:

d = Distance
 i = Number of data
 y = Centroid
 x = Data

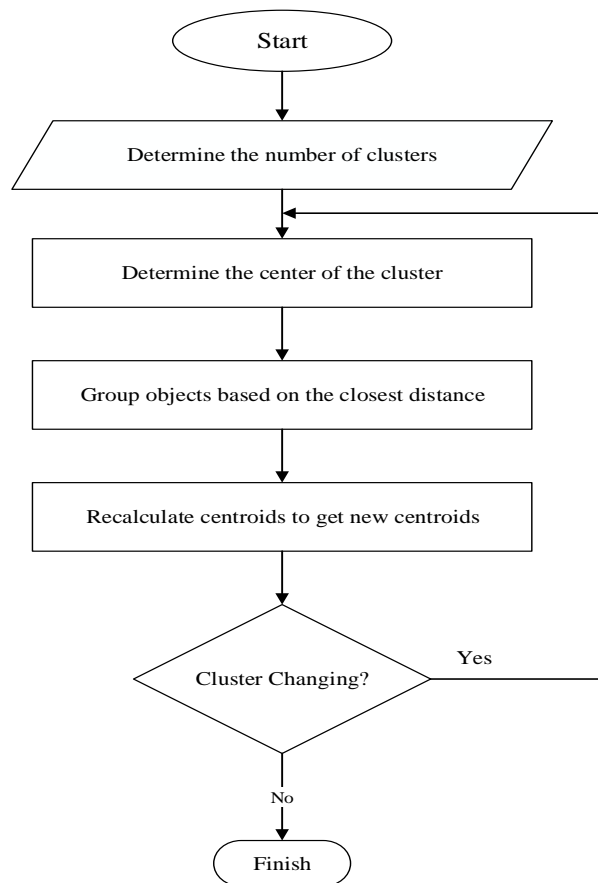


Figure 1. Flow process k-means clustering

3.2. AHC

Agglomerative hierarchical grouping is a hierarchical grouping method with approach bottom-up. The grouping process starts from each data as a group, then recursively looks for the closest group as a pair to join as a large group. This is in line with the opinion of Krisman et al. that hierarchical clustering is a technique of clustering that forms a hierarchy so as to form a tree structure. Thus, the grouping process is carried out in stages or stages. There are 2 methods in the algorithm, hierarchical clustering namely Agglomerative (bottom-up) and Divisive (top-down) [29]. The steps of the AHC method are: i) Calculate euclidian distance; ii) Merge the two closest clusters; iii) Update the distance matrix according to the agglomerative clustering method. For example, using single linkage, average linkage, or complete linkage; iv) Repeat steps 2 and 3 to get define number of clusters; and v) Output of clustering.

3.3. GMM

Gaussian mixture models is a type of density model consisting of components of functions Gaussian [9]. GMM is applied to study the distribution parameters based on the optimal threshold that corresponds to the minimum calculated error probability [21]. GMM is an accurate method and the number of clusters is predetermined [30]. This is in line with the opinion [31] that GMM is a method that can be used for data clustering. GMM is a mathematical model that attempts to estimate the probability density of a data

distribution using a mixed finite distribution gaussian. Gaussian is the most widely used distribution. When GMM is used as a method clustering, then it will determine the number of clusters [32].

Expectation-maximization (EM) algorithm is a method used for maximum likelihood to estimate distribution parameters. EM algorithm is an iterative method. EM consists of two steps: expectation (E-step) and maximization (M-step) [33]. The following is the formula for EM Algorithm:

$$p(x|\theta) = \sum_{j=1}^M \pi_j p(x|\theta_j) \quad (2)$$

The steps are:

- Input: Training dataset
- Initialize: Π_j, μ_j, \sum_j for each j distribution function
- Repeat:
 - E-Step:

$$w_{ij} = P(j|x_i) = \frac{P(j)p(x_i|j)}{p(x_i)} \quad (3)$$

- M-Step (update parameter):

$$p_j = \frac{1}{n} \sum_{i=1}^N w_{ij}, \mu_j = \frac{\sum_{i=1}^N w_{ij} x_i}{\sum_{i=1}^N w_{ij}}, \sum_j = \frac{\sum_{i=1}^N w_{ij} (x_i - \mu_j)(x_i - \mu_j)^T}{\sum_{i=1}^N w_{ij}} \quad (4)$$

- Do this until the parameters do not change

Below is an explanation of the symbols:

- Σ = Variance-covariance matrix
- μ_j = Mean
- w_{ij} = Mixture ratio
- $p(x|\theta_j)$ = Mixture components
- x_i = Random variable
- π = Mixture proportion

3.4. RNN-LSTM

RNN is a machine learning architecture that has a combination of networks in loops. Networks loop allow information to remain [34]. The method RNN method is known to be able to process sequential text data. RNN has three layers, namely input layer, output layer and hidden layer [35]. Figure 2 is an illustration of the RNN architecture. The LSTM method was first proven by Hochreiter and Schmidhuber in 1997. LSTM is a method that can be used to study a pattern in time series data. LSTM is a type of Recurrent Neural Network [34]. In LSTM architecture, content cells are more complex than RNN. In LSTM there are three gates, including input, forget and output gates. The input gate aims to enter new data, while erasing unimportant information contained in the forget gate and affecting the output at the same time is the task of the output gate. Figure 3 is an illustration of the LSTM architecture.

3.5. ARIMA

Autoregressive Integrated moving average is a model that ignores the variables independent as a whole in forecasting [36]. The ARMA model is a combination of the autoregressive (AR) and moving average (MA). The AR model is a method to see the movement of a variable through the variable itself while the MA model is used to find out the movement of a variable with its residuals in the past [37]. ARIMA is also known as the time series method Box Jenkins. ARIMA is well known in forecasting time series [38]. The following is the formula for ARIMA:

Autoregressive (AR):

$$Y_t = \theta_0 + \theta_1 Y_{t-1} + \theta_2 Y_{t-2} + \dots + \theta_p Y_{t-p} + \epsilon_t \quad (5)$$

Moving Average (MA):

$$Y_t = \mu + \epsilon_t - \omega_1 \epsilon_{t-1} - \omega_2 \epsilon_{t-2} - \dots - \omega_q \epsilon_{t-q} \text{ or Y-AR} \quad (6)$$

Autoregressive and moving average:

$$Y_t = \theta_0 + \theta_1 Y_{t-1} + \theta_2 Y_{t-2} + \dots + \theta_p Y_{t-p} + \epsilon_t - \omega_1 \epsilon_{t-1} - \omega_2 \epsilon_{t-2} - \dots - \omega_q \epsilon_{t-q} \quad (7)$$

The symbols are described:

Y_{t-1} = Time series data in the time period $(t-1)$

$\theta_{1,2,p}$ = Coefficient

Y = True value in period t

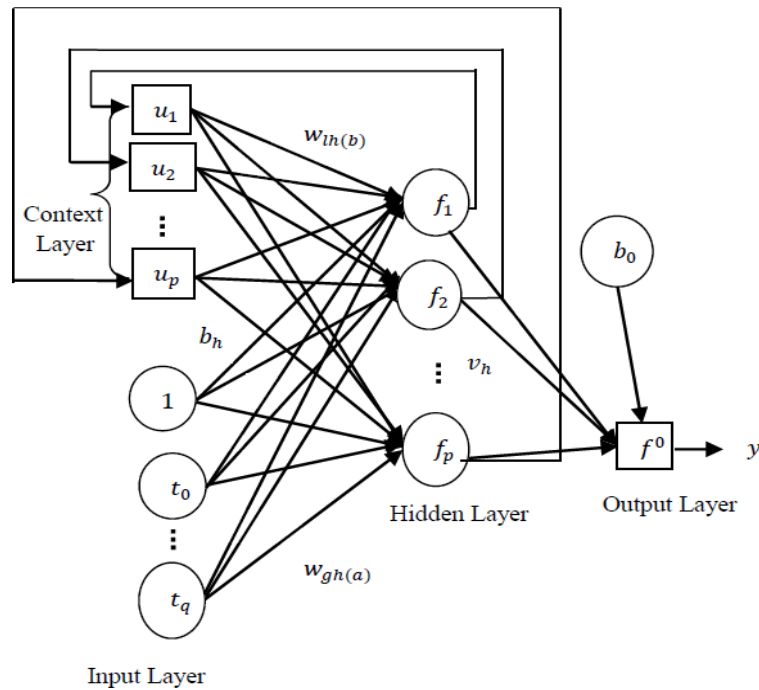


Figure 2. RNN architecture

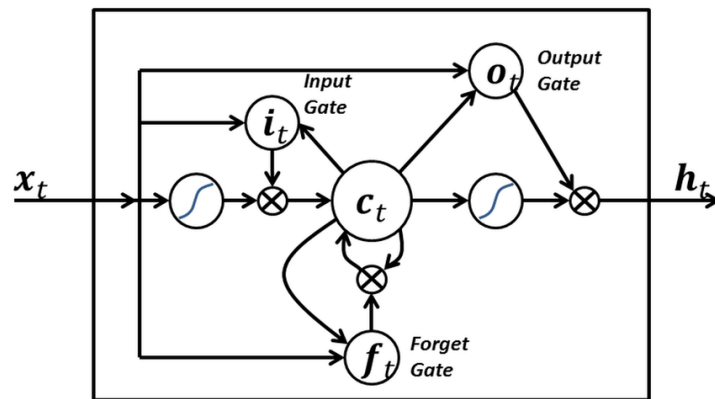


Figure 3. LSTM architecture

4. RESEARCH METHODOLOGY

The framework is a logical sequence to solve a research problem as outlined in a diagram flow from beginning to end, so that research can run systematically according to the concepts that have been made. The research framework for implementing data mining in clustering and forecasting will be outlined in Figure 4. Based on the framework in Figure 4, the first stage in the research that must be carried out is a literature review. Literature review is very important to identify problems and determine the objectives of the research.

Literature review was conducted to collect previous research journals on sales and to collect journals on methods of machine learning in the form of clustering and forecasting. From the literature review, the researcher concludes that for clustering, the researcher will use the k-means clustering, AHC and GMM methods where these methods are frequently used methods and produce an accurate evaluation in terms of clustering. Then for forecasting, the methods used are ARIMA and RNN-LSTM methods. Where the method is a method that is very often used and produces an accurate evaluation in terms of forecasting. After conducting a literature review, it is continued by identifying problems in the research and proceeding to the stage of data collection and analysis.

Then proceed with modeling. In modeling, the first thing to do is clustering using k-means clustering, AHC and GMM. After that, evaluate with each method. After evaluating clustering, the next step is to determine model clustering which is the best. After that, it is continued by doing forecasting using ARIMA and RNN-LSTM based on the results of clustering. Then evaluate the forecasting and determine the model forecasting best based on the evaluation that has been done. When the best method for has been selected clustering and forecasting, and the research objectives have been achieved, the process is complete.

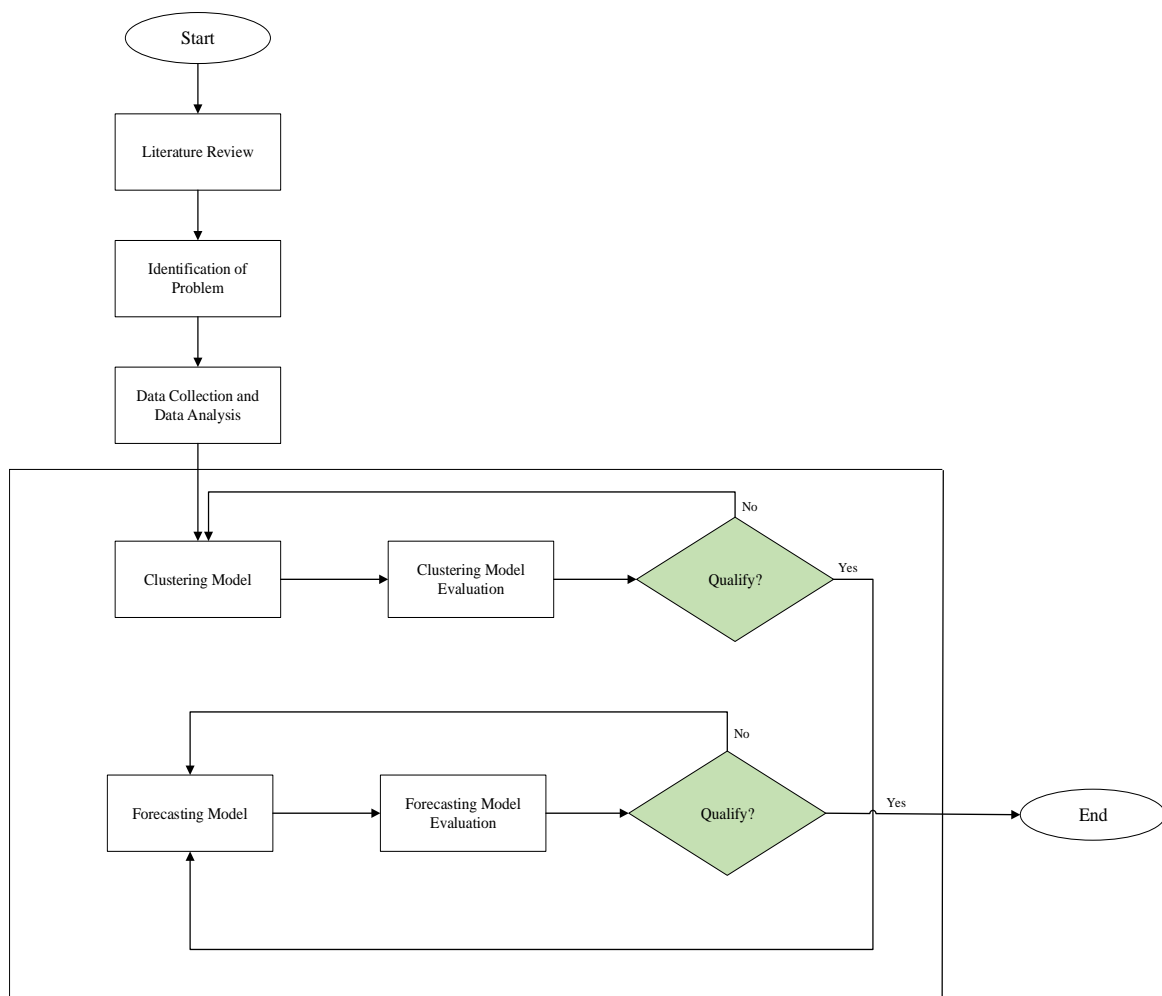


Figure 4. Research methodology

5. PROPOSED METHODS

Proposed methods aims to describe what solutions are proposed to the problems that have been described in the background section. There are five activities in this sub-chapter. For more details, in this session, the design and manufacture of solutions will be carried out which is illustrated in Figure 5.

The first step to do is to prepare the dataset. Then do pre-processing, for example by normalizing the data. After the data is deemed sufficient for modeling, the next step is to do clustering using the methods

k-means clustering, AHC and GMM and then train the data. After that, the results will be obtained clustering from each method and the next step is to evaluate the model of each method. The evaluation will use silhouettes and DBI values. Silhouette refers to a method of validation of consistency within clusters of data. Then the results of the evaluation of the two methods are compared and seen which method clustering is better. When the results are obtained clustering best, the next step is forecasting with ARIMA and RNN-LSTM. The first stage in forecasting is to train the data. After that, the results will be obtained forecasting sales of hardware from each method. The RNN-LSTM method uses Python. The first step is to define a library. In this method, Scikit-learn is used. Then there is the sequential class which is part of the Keras library which aims to connect between layers. This method activates the LSTM and dense layer. The dense layer output is 1 neuron. The hidden layer is in the form of a 3D input layer using numpy reshape. The activation function uses ReLU, optimization uses Adam, and the number of epochs is 50. Then evaluate the model of each method and compare the evaluation results. Evaluation forecasting using root mean square error (RMSE), and calculate the amount of saving cost from each forecasting model. RMSE is a calculation between actual and predicted. RMSE which has a small value is more accurate than RMSE which has a large value. Then when the evaluation results are visible, the next step is to compare which method is better for forecasting. After comparing and choosing the method, the researcher will know which method is the best for clustering and forecasting. In addition, after determining the best method from the evaluation results, the main objective in this research will also be achieved, namely developing and implementing data mining sales of hardware with methods of machine learning to determine clustering of hardware sold and developing and implementing data mining for forecasting sales of Hardware based on clustering that has been made with methods of machine learning to determine stock hardware as a sales strategy for the company.

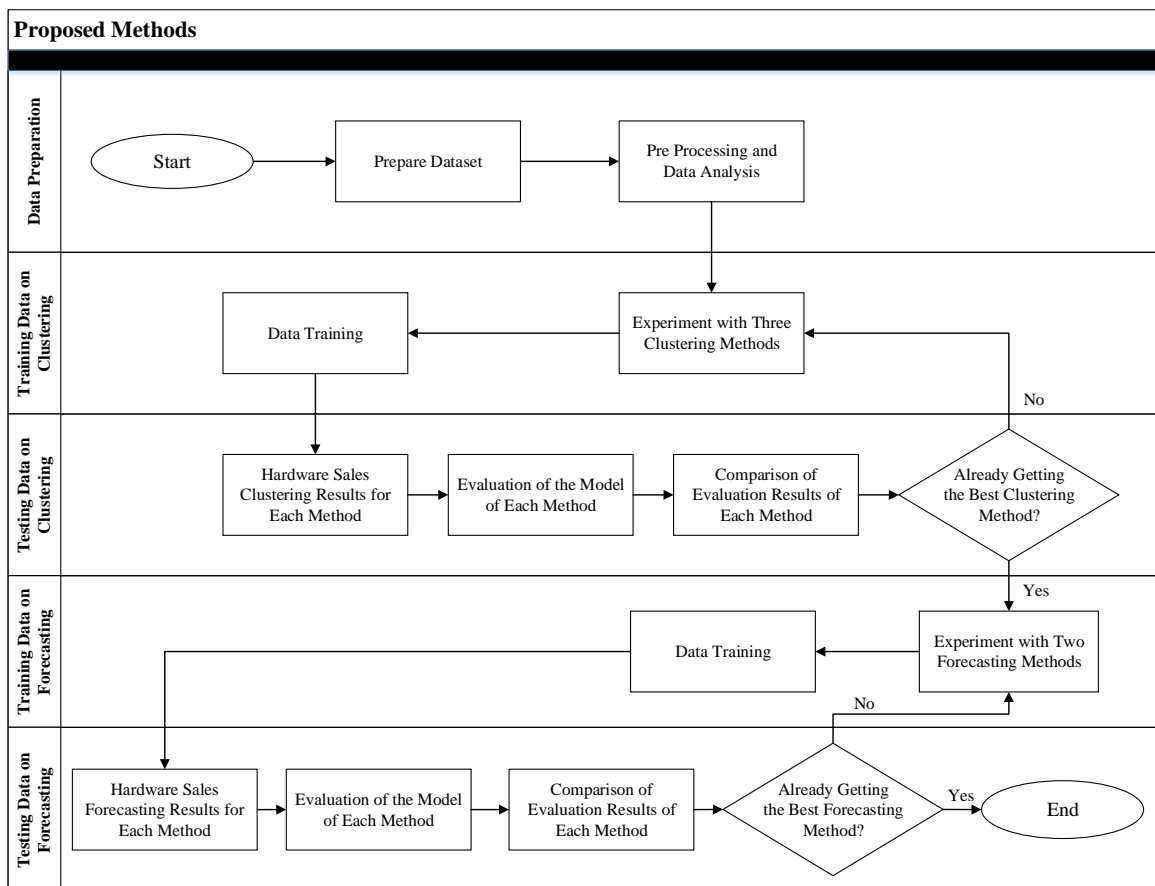


Figure 5. Flow diagram proposed methods

6. RESULTS AND DISCUSSION

The Following is a summary obtained from several clustering scenarios that have been carried out. There is some information such as Method, Number of Clustering, Attribute, Silhouette, and DBI Values. A summary of clustering is in Table 4.

Based on the results of the experiment clustering and based on the summary of Table 4, k-means clustering using two attributes, namely "Quantity and Stock", is the best method for the case study in this study. Evaluation for the k-means clustering method obtained results silhouette of 0.91 and DBI values of 0.34 consisting of 3 clusters, namely 146 data for cluster 1, 3 data for cluster 2 and 3 data for cluster 3. And Table 5 is a summary of forecasting.

Table 4. Evaluation result for clustering

Method	Number of Clustering	Attribute	Silhouette	DBI Values
k-means clustering	3	Quantity, Stock	0.91	0.34
k-means clustering	3	Quantity, Stok, Price	0.80	0.37
k-means clustering	3	Quantity, Stok, Price, Customers	0.79	0.32
AHC	3	Quantity, Stock	0.87	0.50
AHC	3	Quantity, Stok, Price	0.80	0.37
AHC	3	Quantity, Stok, Price, Customers	0.79	0.28
GMM	3	Quantity, Stock	0.68	0.72
GMM	3	Quantity, Stok, Price	0.001	1
GMM	3	Quantity, Stok, Price, Customers	0.74	0.55

Table 5. Evaluation result for forecasting

Method	Detail	Attribute	Saving Cost
ARIMA	Forecasting Cluster 1 (above 10)	Date, Quantity, Stock	-100%
ARIMA	Forecasting Cluster 1 (below 10)	Date, Quantity, Stock	85%
ARIMA	Forecasting Cluster 2 (above 10)	Date, Quantity, Stock	64%
ARIMA	Forecasting Cluster 2 (below 10)	Date, Quantity, Stock	99%
ARIMA	Forecasting Cluster 3 (above 10)	Date, Quantity, Stock	88%
ARIMA	Forecasting Cluster 3 (below 10)	Date, Quantity, Stock	100%
RNN-LSTM	Forecasting Cluster 1 (above 10)	Date, Quantity, Stock	-100%
RNN-LSTM	Forecasting Cluster 1 (below 10)	Date, Quantity, Stock	86%
RNN-LSTM	Forecasting Cluster 2 (above 10)	Date, Quantity, Stock	64%
RNN-LSTM	Forecasting Cluster 2 (below 10)	Date, Quantity, Stock	99%
RNN-LSTM	Forecasting Cluster 3 (above 10)	Date, Quantity, Stock	80%
RNN-LSTM	Forecasting Cluster 3 (below 10)	Date, Quantity, Stock	100%

Based on Table 5, it can be seen the overall results of the forecasting evaluation. In addition, it is also known the amount of saving cost based on experiments using the ARIMA and RNN-LSTM methods, it can be seen that the RNN-LSTM method is better because it produces more cost savings than the ARIMA method. Percentage of saving cost against actual cost based on these two methods is 83%.

7. CONCLUSION

Clustering is done using three methods, namely k-means clustering, AHC, and GMM. In each method three experiments were carried out. The first experiment uses the "Quantity and Stock" attribute, the second experiment uses the "Quantity, Stock and Price" attribute, then the third experiment uses the "Quantity, Stock, Price, and Customer" attribute. The best method for clustering is k-means clustering using 2 attributes. with a silhouette of 0.91 and DBI values of 0.34. Then, forecasting is done using two methods including ARIMA and RNN-LSTM. In each method, six experiments were carried out. The first experiment using training data and testing cluster 1 above 10, the second experiment uses training data and testing cluster 1 below 10, the third experiment uses training data and testing cluster 2 above 10, the fourth experiment uses training data and testing cluster 2 under 10, the fifth experiment uses training data and testing cluster 3 above 10 and the sixth experiment using training data and testing cluster 3 below 10. The best method for forecasting is RNN-LSTM because it produces more cost savings than the ARIMA method. Percentage of saving cost for ARIMA against actual cost is 83% and percentage of saving cost for RNN-LSTM against actual cost is 84%. The percentage of the difference in saving costs between ARIMA and RNN-LSTM to the actual cost is 83%.

ACKNOWLEDGEMENTS

The authors thank to Bina Nusantara University for the research grant and supporting this research.




REFERENCES

- [1] O. Saritas, P. Bakhtin, I. Kuzminov, and E. Khabirova, "Big data augmented business trend identification: the case of mobile commerce," *Scientometrics*, vol. 126, no. 2, pp. 1553–1579, 2021, doi: 10.1007/s11192-020-03807-9.
- [2] V. Sohrabpour, P. Oghazi, R. Toorajipour, and A. Nazarpour, "Export sales forecasting using artificial intelligence," *Technol. Forecast. Soc. Change*, vol. 163, no. November, p. 120480, 2021, doi: 10.1016/j.techfore.2020.120480.
- [3] H. Wei and Q. Zeng, "Research on sales Forecast based on XGBoost-LSTM algorithm Model," *J. Phys. Conf. Ser.*, vol. 1754, no. 1, 2021, doi: 10.1088/1742-6596/1754/1/012191.
- [4] M. Iqbal, J. Ma, N. Ahmad, K. Hussain, and M. S. Usmani, "Promoting sustainable construction through energy-efficient technologies: an analysis of promotional strategies using interpretive structural modeling," *Int. J. Environ. Sci. Technol.*, vol. 18, no. 11, pp. 3479–3502, 2021, doi: 10.1007/s13762-020-03082-4.
- [5] S. Huang, Z. Kang, Z. Xu, and Q. Liu, "Robust deep k-means: An effective and simple method for data clustering," *Pattern Recognit.*, vol. 117, p. 107996, 2021, doi: 10.1016/j.patcog.2021.107996.
- [6] D. A. N. Wulandari, R. Annisa, and L. Yusuf, "an Educational Data Mining for Student Academic Prediction Using K-Means Clustering and Naïve Bayes Classifier," *Semin. Nas. Apl. Teknol. Inf.*, pp. 155–160, 2020, doi: 10.33480/pilar.v16i2.1432.
- [7] R. D. Dana, D. Soilihudin, R. H. Silalahi, D. Kurnia, and U. Hayati, "Competency test clustering through the application of Principal Component Analysis (PCA) and the K-Means algorithm," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1088, no. 1, p. 012038, 2021, doi: 10.1088/1757-899x/1088/1/012038.
- [8] R. T. Adek, R. K. Dinata, and A. Ditha, "Online Newspaper Clustering in Aceh using the Agglomerative Hierarchical Clustering Method," *Int. J. Eng. Sci. Inf. Technol.*, vol. 2, no. 1, pp. 70–75, 2021, doi: 10.52088/ijesty.v2i1.206.
- [9] J. Gu, J. Chen, Qiming Zhou, and H. Zhang, "Gaussian mixture model of texture for extracting residential area from high-resolution remotely sensed imagery," *ISPRS Work. Updat. Geo-spatial Databases with Imag. 5th ISPRS Work. DMGISs*, no. 1973, pp. 157–162, 2007.
- [10] M. Elsaraiti and A. Merabet, "A Comparative Analysis of the ARIMA and LSTM Predictive Models and Their Effectiveness for Predicting Wind Speed," *Energies*, vol. 14, no. 20, 2021, doi: 10.3390/en14206782.
- [11] J. Zhao, D. Zeng, S. Liang, H. Kang, and Q. Liu, "Prediction model for stock price trend based on recurrent neural network," *J. Ambient Intell. Humaniz. Comput.*, vol. 12, no. 1, pp. 745–753, 2021, doi: 10.1007/s12652-020-02057-0.
- [12] N. Talkhi, N. Akhavan Fatemi, Z. Ataei, and M. Jabbari Nooghabi, "Modeling and forecasting number of confirmed and death caused COVID-19 in IRAN: A comparison of time series forecasting methods," *Biomed. Signal Process. Control*, vol. 66, no. February, p. 102494, 2021, doi: 10.1016/j.bspc.2021.102494.
- [13] F. A. Fithri and S. Wardhana, "Cluster Analysis of Sales Transaction Data Using K-Means Clustering At Toko Usaha Mandiri," *J. PILAR Nusa Mandiri*, vol. 17, no. 2, pp. 113–118, 2021.
- [14] R. Johannes and A. Alamsyah, "Sales prediction model using classification decision tree approach for small medium enterprise based on Indonesian e-commerce data," Mar. 2021, doi: 10.48550/arXiv.2103.03117.
- [15] B. Soepriyanto, P. Studi, and S. Informasi, "Comparative analysis of K-NN and naïve bayes methods to predict stock prices," *Int. J. Comput. Inf. Syst.*, vol. 2, no. 2, pp. 49–53, May 2021, doi: 10.29040/ijcis.v2i2.32.
- [16] S. I. Nadeak and Y. Ali, "Analysis of Data Mining Associations on Drug Sales at Pharmacies with APRIORI Techniques," *IJISTECH (International J. Inf. Syst. Technol.)*, vol. 5, no. 1, p. 38, 2021, doi: 10.30645/ijistech.v5i1.113.
- [17] P. Edastama, A. S. Bist, and A. Prambudi, "Implementation Of Data Mining On Glasses Sales Using The Apriori Algorithm," *Int. J. Cyber IT Serv. Manag.*, vol. 1, no. 2, pp. 159–172, 2021, doi: 10.34306/ijcitsm.v1i2.46.
- [18] D. Puspita and S. Sasmita, "Application of K-Means Algorithm in Grouping of City Tourism City Pagar Alam," *Sinkron*, vol. 7, no. 1, pp. 28–32, 2022, doi: 10.33395/sinkron.v7i1.11220.
- [19] L. N. Rani, S. Defit, and L. J. Muhammad, "Determination of Student Subjects in Higher Education Using Hybrid Data Mining Method with the K-Means Algorithm and FP Growth," *Int. J. Artif. Intell. Res.*, vol. 5, no. 1, pp. 91–101, 2021, doi: 10.29099/ijair.v5i1.223.
- [20] Y. Irawan, "Implementation Of Data Mining For Determining Majors Using K-Means Algorithm In Students Of SMA Negeri 1 Pangkalan Kerinci," *J. Appl. Eng. Technol. Sci.*, vol. 1, no. 1, pp. 17–29, 2019, doi: 10.37385/jaets.v1i1.18.
- [21] X. Shen, Y. Zhang, K. Sata, and T. Shen, "Gaussian Mixture Model Clustering-Based Knock Threshold Learning in Automotive Engines," *IEEE/ASME Trans. Mechatronics*, vol. 25, no. 6, pp. 2981–2991, 2020, doi: 10.1109/TMECH.2020.3000732.
- [22] S. Rashid, A. Ahmed, I. Al Barazanchi, and Z. A. Jaaz, "Clustering Algorithms Subjected to K-Mean and Gaussian Mixture Model on Multidimensional Data Set," *Period. Eng. Nat. Sci.*, vol. 7, no. 2, pp. 448–457, 2019.
- [23] W. Xu, H. Peng, X. Zeng, F. Zhou, X. Tian, and X. Peng, "A hybrid modelling method for time series forecasting based on a linear regression model and deep learning," *Appl. Intell.*, vol. 49, no. 8, pp. 3002–3015, 2019, doi: 10.1007/s10489-019-01426-3.
- [24] A. K. Gupta, V. Singh, P. Mathur, and C. M. Travieso-Gonzalez, "Prediction of COVID-19 pandemic measuring criteria using support vector machine, prophet and linear regression models in Indian scenario," *J. Interdiscip. Math.*, vol. 24, no. 1, pp. 89–108, 2021, doi: 10.1080/09720502.2020.1833458.
- [25] Z. Malki *et al.*, "ARIMA models for predicting the end of COVID-19 pandemic and the risk of second rebound," *Neural Comput. Appl.*, vol. 33, no. 7, pp. 2929–2948, 2021, doi: 10.1007/s00521-020-05434-0.
- [26] H. Alabdulrazzaq, M. N. Alenezi, Y. Rawajfih, B. A. Alghannam, A. A. Al-Hassan, and F. S. Al-Anzi, "On the accuracy of ARIMA based prediction of COVID-19 spread," *Results Phys.*, vol. 27, p. 104509, 2021, doi: 10.1016/j.rinp.2021.104509.
- [27] S. A. Fahad and M. M. Alam, "A modified K-means algorithm for big data clustering," *Int. J. Sci. Eng. Comput. Technol.*, vol. 6, no. 4, pp. 129–132, Apr. 2016.
- [28] M. Faisal, E. M. Zamzami, and Sutarman, "Comparative Analysis of Inter-Centroid K-Means Performance using Euclidean Distance, Canberra Distance and Manhattan Distance," *J. Phys. Conf. Ser.*, vol. 1566, no. 1, 2020, doi: 10.1088/1742-6596/1566/1/012112.
- [29] M. Roux, "A Comparative Study of Divisive and Agglomerative Hierarchical Clustering Algorithms," *J. Classif.*, vol. 35, no. 2, pp. 345–366, 2018, doi: 10.1007/s00357-018-9259-9.
- [30] S. Kannan, "Intelligent object recognition in underwater images using evolutionary-based Gaussian mixture model and shape matching," *Signal, Image Video Process.*, vol. 14, no. 5, pp. 877–885, 2020, doi: 10.1007/s11760-019-01619-w.
- [31] Z. Wang, C. Da Cunha, M. Ritou, and B. Furet, "Comparison of K-means and GMM methods for contextual clustering in HSM," *Procedia Manuf.*, vol. 28, pp. 154–159, 2019, doi: 10.1016/j.promfg.2018.12.025.
- [32] A. Mirzal, "Statistical Analysis of Microarray Data Clustering using NMF, Spectral Clustering, Kmeans, and GMM," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 5963, no. c, pp. 1–1, 2020, doi: 10.1109/tcbb.2020.3025486.
- [33] S. Srivastava, G. DePalma, and C. Liu, "An Asynchronous Distributed Expectation Maximization Algorithm for Massive Data: The DEM Algorithm," *J. Comput. Graph. Stat.*, vol. 28, no. 2, pp. 233–243, 2019, doi: 10.1080/10618600.2018.1497512.




- [34] A. A. Ningrum, I. Syarif, A. I. Gunawan, E. Satriyanto, and R. Muchtar, "Deep learning-lstm algorithm for power transformer lifetime," *JTIHK*, vol. 8, no. 3, pp. 593–548, 2021, doi: 10.25126/jtiik.202184587.
- [35] D. Li and J. Qian, "Text sentiment analysis based on long short-term memory," *IEEE Int. Conf. Comput. Commun. Internet*, 2016, doi: 10.1109/CCL.2016.7778967.
- [36] S. Noureen, S. Atique, V. Roy, and S. Bayne, "Analysis and application of seasonal ARIMA model in Energy Demand Forecasting: A case study of small scale agricultural load," *Midwest Symp. Circuits Syst.*, vol. 2019-Augus, pp. 521–524, 2019, doi: 10.1109/MWSCAS.2019.8885349.
- [37] A. Saikhu, C. V. Hudiyanti, J. L. Buliali, and V. Hariadi, "Predicting COVID-19 Confirmed Case in Surabaya using Autoregressive Integrated Moving Average, Bivariate and Multivariate Transfer Function," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1077, no. 1, p. 012055, 2021, doi: 10.1088/1757-899x/1077/1/012055.
- [38] Ü. Ç. Büyüksahin and Ş. Ertekin, "Improving forecasting accuracy of time series data using a new ARIMA-ANN hybrid method and empirical mode decomposition," *Neurocomputing*, vol. 361, pp. 151–163, 2019, doi: 10.1016/j.neucom.2019.05.099.

BIOGRAPHIES OF AUTHORS



Rani Puspita    is a master's student at BINUS Graduate Program-Master of Computer Science, Bina Nusantara University with a focus on data science. Her undergraduate education background is Informatics Engineering at UIN Syarif Hidayatullah Jakarta. She is also a system analyst. She can be contacted at email: rani.puspita@binus.ac.id.



Lili Ayu Wulandhari    is a lecturer at BINUS Graduate Program-Master of Computer Science, Bina Nusantara University. She is also a data scientist. She can be contacted at email: lili.wulandhari@binus.ac.id.