Factors that Affect Customer Credit Payments During COVID-19 Pandemic: An Application of Light Gradient Boosting Machine (LightGBM) and Classification and Regression Tree (CART)

Imas Wihdah Misshuari, Ratna Herdiana, Farikhin
Department of Mathematics,
Faculty of Science and Mathematics
Universitas Diponegoro
Tembalang, Semarang, Jawa Tengah 50275, Indonesia
imasswmisshuari@gmail.com, ratnaherdiana@lecturer.undip.ac.id
farikhin.math.undip@gmail.com

Teuku Afrizal
Department of Public Administration
Faculty of Social Science and Political Science
Universitas Diponegoro
Tembalang, Semarang, Jawa Tengah 50275, Indonesia
teukurian@lecturer.undip.ac.id

Jumadil Saputra
Faculty of Business, Economics and Social Development
Universiti Malaysia Terengganu
21030 Kuala Nerus, Terengganu, Malaysia
jumadil.saputra@umt.edu.my

Abstract

The COVID-19 pandemic affects whole segments of the world, including the banking and credit sectors. The banking and credit sector plays an essential role in the Indonesian economy structure due to the function as a collector and channel of funds by creating products offered to people who need to use the credit services. We compare the accuracy of two different machine learning methods - Light Gradient Boosting Machine (LightGBM) and Classification and Regression Tree (CART). We apply this method to classify data and determine factors that affect credit payments in debtor data at PT BPR Syariah Gebu Prima Medan, which contains current credit and bad credit debtors. Based on the evaluation conducted on the classification of factors that affect credit payments, findings showed that the factors that influence the CART method are customers who have a maximum total income of IDR 27,750,000 with a ceiling of more than IDR 57,500,000 are customers who have maximum family members. 3.5 and the LightGBM method is the ceiling, total income, family members, age, and gender with importance values of 65200, 65100,13000,9800, and 4200. However, the CART method has a higher accuracy rate of 85.9% than the LightGBM method is 81%.

Keywords
Machine Learning, Light Gradient Boosting Machine (LightGBM), Classification and Regression Tree (CART).

1. Introduction

Covid 19 (Corona Virus Disease 2019) is a disease caused by the SARS-Cov 2 virus (Severe Acute Respiratory Syndrome Coronavirus 2) or better known as the Corona Virus. The Covid - 19 pandemic affects all economic sectors in the world. The economy has experienced sluggish growth, many trading activities have stopped, so there is no income and economic activity. If this continues and not knowing when it will stop, the impact will be even greater and spread to the banking and credit sectors. The banking sector is one of the factors that play an important role in Indonesia's economic structure because it functions as a collector and channel of funds by creating various products to be offered to people who wish to use credit services such as vehicle crediting. Credit is giving a loan to a borrower based on an agreement between the borrower or Bank. The borrower needs to pay off his debt at a predetermined time with interest. When a bank lends money to a customer, the Bank naturally expects the money to come back. To reduce the loan's risk not returning (bad credit), several credit requirements at banks that applicants must meet are namely the willingness to pay and the ability to pay loans and interest (Indriana et al. 2018, Kartini 2018).

The Bank is a creditor institution that often experiences credit problems so that the Bank must determine which debtors are eligible for credit. The problem faced by BPRS MEDAN is that the Bank checks credit data with current and non-current status manually so that it is inefficient and errors may occur. Therefore, the need for a method to automatically classify data so that the Bank in checking the data does not make mistakes and finds the factors that affect credit payments by customers. It can be solved using classification methods to determine the factors that affect credit payments by customers at BPRS Medan. Classification is a method of grouping objects based on the characteristics of the object of classification and functions to obtain hidden information from huge piles of data. The results of this processing can be used as predictions or decision assistants in solving a problem.

Factors that affect credit payments by customers at BPRS MEDAN can also be determined using machine learning methods. Machine learning is a method capable of handling big data by developing computer algorithms. Fundamentally, how machine learning works is to learn like humans by using examples and getting each statement's results. This method produces information, and then it can be used as knowledge to solve a problem as an input-output process (Nurhayati and Iswara 2019). Several methods that have been developed in machine learning are Naive Bayes, Support Vector Machine, XGBOOST, Light Gradient Boosting Machine, and Classification and Regression Tree.

1.1 Objectives

In this paper, we apply the LightGBM and CART methods to the classification problem of factors that affect credit payments and the accuracy of these methods. Both methods have similarities in the advantages of handling large-scale data and having good accuracy values. The aim is that the LightGBM and CART methods can uncover the factors that affect the credit payment and produce the best accuracy score.

2. Literature Review

Pandemic affected the people's economy and caused debtors who had borrowed from banking (Bidani et al. 2020). Therefore, in this case, the government stimulated the banking sector for the debtor community affected by Covid-19. So far, banking performance can be categorised as still in good condition and still maintained despite the Covid-19 outbreak. However, suppose the epidemic has not been resolved for a long time. In that case, the banking system's economic performance is likely to decline or worsen the banking performance for the next several months, and even years will depend on how Covid-19 in the present. Problems at the BPRS MEDAN bank can be alternatively resolved by machine learning methods, including Light Gradient Boosting Machine (LightGBM) and Classification and Regression Tree (CART). According to (Trivedi. 2020, Provenzano et al. 2020), machine learning can be implemented on credit scoring problems. Tanjung and Kartiko (2017) show that the amount of income influences credit payments' status, age and ceiling with the accuracy of the classification results formed by 84.2%. Prabawati (2019) shows that the CART algorithm has been able to classify the length of the student's study period who joined the Jakarta State University organisation, which produces an average accuracy of 80%.

Machado et al. (2019) research on LightGBM shows that compared to the XGBoosting method, the accuracy of LightGBM is higher than ordinary regression and other decision tree models and Minastireanu and Mesnita (2019) able to obtain an accuracy of 98%. Zhang et al. (2019) showed that LightGBM could overcome overfitting and has good stability with high accuracy. To classify the factors that affect credit payments, we use the LightGBM algorithm. LightGBM is an upgrade from Gradient Boosting Decision Trees. Although conventional GBDT is widely used because of its accuracy and efficiency, several improvements can be made to obtain a more efficient and faster algorithm without making a trade-off between accuracy and efficiency. The larger the data instance, the more training

time the algorithm will take. It gives rise to data sampling in which only a portion of the data is used to draw satisfactory conclusions from the data set.

3. Methods

3.1 Classification and Regression Tree (CART)

CART (Classification and Regression Trees) is a method developed by Leo Breimann, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. CART, one of the data exploration techniques, is the decision tree technique. This method describes the relationship between the response variable and one or more predictor variables. CART was developed for the topic of classification analysis, both for categorical and continuous response variables. CART can generate a classification tree, if the response variables are categorical and generate a regression tree if the variables are continuous. In this study, the response variable has a categorical scale, so the tree classification method is used. According to Maulana et al., (2015), Alkanomi, (2015), CART is a nonparametric statistic that can be used for classification analysis, and the decision tree technique can solve the problems in this study. Implementation of CART consists of the following four steps, namely:

1. Splitting Nodes.
   The splitting process starts from the main node, which consists of the data to be sorted. Sorting is done to sort the data into two groups, namely groups that enter the left node and those that enter the right node.
2. Class Assignment.
   Class assignment is carried out from the beginning of sorting the vertices until the end node is formed because each node that is formed has the opportunity to become the final node. Assignment of each end node based on the rule for the highest number of class members, namely if:

$$P(J_0|t) = max_j\, P(j|t) = \; max_j\, \frac{N_j(t)}{N(t)} \quad (1)$$

3. Stop The Splitting.
   A node will become the final node or will not be re-split if there is only one observation in each child node, all observations in each child node have an identical response variable distribution, and there is a limit to the maximum number of tree depths determined by the researcher. If this is fulfilled, then tree development is stopped and a maximum classification tree is obtained.
4. Pruning.
   The maximum classification tree that is formed is possible to be very large. The more sorting is done, the higher the accuracy rate, but with a very large size it will be difficult to understand, causing overfitting (very complex value matching) for new data. This problem is overcome by pruning the maximum classification tree to obtain the optimal classification tree. Pruning measures used to obtain a proper tree size are:

$$R(T)\sum_{teT} r(t)P(t) = \sum_{teT} R(t) \quad (2)$$

   with a resubstitution estimate is a tree misclassification cost or tree resubstitution cost (probability of misclassification caused by the formed classification tree), is the proportion of observations that enter a node, is a set of final nodes, whereas resubstitution estimate is the probability of misclassification in a node-specific t which is defined as follows:

$$r(t) = 1 - \; max_j\, P(j|t) \quad (3)$$

   The initial step of the pruning process is carried out on $r(t)$, namely the subtree by taking which is the left child node and which is the right child node as a result of sorting the parent node. If two child nodes are obtained, and the parent node satisfies equation (4), then the child nodes are truncated. This process is repeated until no more pruning is possible (Timoveef, 2014).

$$R(t) = R(t_L) + R(t_R) \quad (4)$$

3.2 Light Gradient Boosting Machine (LightGBM)

Ma (2018), LightGBM (Light Gradient Boosting Machine) is a fast and efficient Gradient Boosting Decision Tree algorithm or method in a framework designed by Microsoft in 2016. LightGBM can be used in classification,

sorting, regression etc. as well as supports efficient parallel training. LightGBM is a learning method gradients based on decision tree techniques, improvement ideas and relatively new methods. (Fan et al. 2019, Zeng, 2019). LightGBM uses leaf-sage techniques to produce trees and finds the leaves with the greatest variance to perform division. LightGBM can be distinguished from the different Gradient Boosting Decision Tree methods by calculating the gain of variation. In LightGBM, the splits occur considering weak and strong learners (small and big gradients, g_i) (Machado et al. 2019, Sousa, 2019).

$$V_j^*(d) = \frac{1}{n}\left(\frac{\left(\sum_{x_i \in A_l} g_i + \frac{1-a}{b}\sum_{x_i \in B_l} g_i\right)^2}{n_l^j(d)} + \frac{\left(\sum_{x_i \in A_r} g_i + \frac{1-a}{b}\sum_{x_i \in B_r} g_i\right)^2}{n_r^j(d)} \right) \qquad (5)$$

where $A_l = \{x_i \in A : x_{ij} \leq d\}$, $A_r = \{x_i \in A : x_{ij} > d\}$, $B_l = \{x_i \in B : x_{ij} \leq d\}$, $B_r = \{x_i \in B : x_{ij} > d\}$, d is the point in data where the split is calculated in order to find the optimal gain in variance and the coefficient $\frac{1-a}{b}$ is used to normalise the sum of the gradients over B back to the size of A^c.
LightGBM uses a leaf-wise growth strategy with a depth limit to find a leaf node with the largest split gain in all of the current leaf nodes, then splits, and so on (Wang and Wang 2020), as shown in Figure 1:

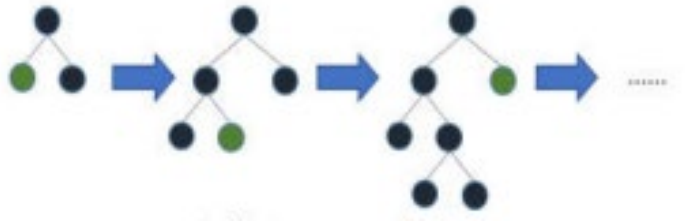

Figure 1. Schematic diagram of leaf – wise strategy growth tree

Minastireanu and Mesnita (2019), LightGBM will achieve optimal with the main parameters, including:

- Number of leaves - number of leaves in each tree.
- max_depth - maximum depth in tree algorithm to reduce over fitting, if maximum depth < 0 then there is no depth limit.
- min_data - the minimum amount of data in a tree.
- bagging_fraction - like feature fraction which distinguishes it, data is selected randomly and cannot be repeated and must be greater than 0.
- Lambda_l1 and Lambda_l2 - which are used to overcome over fitting by controlling l1 / l2.

3.1 Measurement of Classification Accuracy Test

A system that performs classification is expected to classify all data sets correctly, but it cannot be denied that the performance of a system cannot be 100% correct (Prasetyo, 2012). So that a classification system must be measured for the accuracy of its classification using a confusion matrix, Table 1. A confusion matrix is a tool that has a function to analyse whether the classifier is good at recognising tuples from different classes. The values of True Positive and True Negative provide information when the classifier classifies data is true. In contrast, False Positive and False Negative provide information when the classifier is incorrect in classifying data (Fibrianda and Bhawiyuga 2018).

Table 1. Confusion Matrix

	Predicted No	Predicted Yes
Actual No	True Negative (TN)	False Positive (FP)
Actual Yes	False Negative (FN)	True Positive (TP)

Where TP (True Positive) is the number of data with a positive true value, a positive predictive value FP (False Positive) is the number of data with a negative true value, a positive FN (False Negative) predictive value is the

number of data with a positive true value and a negative TN predictive value (True Negative) is the amount of data with a negative true value and a negative predictive value. Based on the content of the confusion matrix, it can be seen that the number of data from each class that was predicted correctly, namely (TN + TP) and the correctly classified data, namely (FN + FP). The quantity of the confusion matrix can be summarised into two values, namely accuracy and error rate. By knowing the amount of data that is classified correctly, the accuracy of the prediction results can be seen, and the data by knowing the amount of data that is classified incorrectly can determine the error rate of wrong predictions. These two quantities are used as the classification accuracy matrix.

To find out the classification accuracy with formulas:

$$Accuracy = \frac{TN + TP}{Total} \qquad (6)$$

To find out the error rate (prediction error) a formula is used :

$$Error\ Rate = \frac{FN + FP}{Total} \qquad (7)$$

4. Data Collection

The problem faced by PT BPRS Gebu Prima is that the Bank is manually checking credit data with current and non-current status, so it is inefficient, and errors can occur. Therefore, we need a method to automatically classify data so that the Bank in checking the data does not make any more mistakes and can determine the factors that affect credit payments by customers. The Light Gradient Boosting Machine (LightGBM) method and Classification and Regression Trees (CART) must be made based on the problems that occur at PT BPRS Gebu Prima to obtain the factors that affect credit payments by customers. In the LightGBM and CART methods, the factors that need to be considered are the dependent and independent variables. The dependent variable is the credit payment status, and the independent variable is gender, age, number of dependents, total income and ceiling. In order to determine the factors that affect credit payments by customers at PT BPRS Gebu Prima using the Classification and Regression Trees method, several data are needed relating to the data being faced. Sources of data used in this study come from debtor data at PT. BPR Syariah Gebu Prima Medan, which is debtor data for the years 2018-2020.

5. Results and Discussion

5.1 Classification and Regression Tree

Analysis conducted on PT BPR Syariah Gebu Prima Medan's debtor data using CART classification resulted in 13 nodes consisting of 1 main node, 2 inner nodes, and 7 terminal nodes. The independent variables entered into the classification tree. According to the amount of improvement, they sorted are age, sex, and the number of dependents with a tree depth of 3 as described in the Classification Tree Figure.

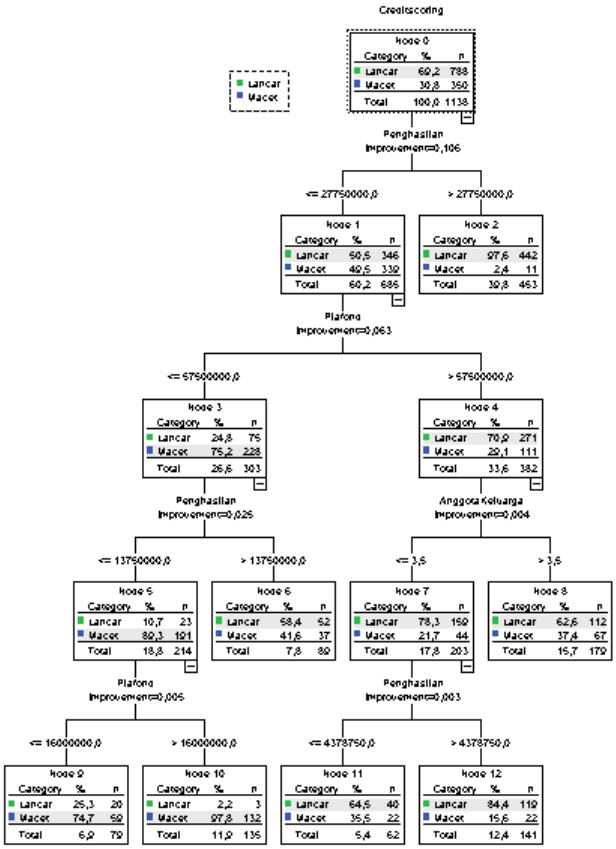

Figure. 2 Classification Tree

Based on the classification tree that has been formed using the CART method, the sorter with the highest improvement value is used as the initial sort in the classification tree. It can be explained that the initial sorting at the parent node (node 0) is a maximum total income of IDR 27,750,000 with a ceiling of more than IDR 57,500,000 followed by maximum family members. 3.5. This CART classification analysis has a risk estimate of 0.160 which can be seen in the Risk Table. It can be explained that the prediction results for categories (current / non-current) based on the classification tree formed have an error rate of 16%.

Table 2. Risk estimated value.

Risk	
Estimate	Std. Error
0.160	0.011

5.2 Light Gradient Boosting Machine (LightGBM)

Analysis carried out on PT BPR Syariah Gebu Prima Medan's debtor data using a Light Gradient Boosting Machine (LightGBM) resulted in ceiling, total income, family members, age, and gender with importance values of 65200, 65100, 13000, 9800 and 4200 as shown in Table 3.

Table 3. Importance results for original features

Feature	Importance
Ceiling	65200
Total Income	65100
Family Members	13000

Age	9800
Gender	4200

In this way, we have to perform a lgb parameter tuning. In the tuning process, we have set up the following parameters:

- Min_child_weight = 0.6715
- Max_depth = 12
- Num_leaves = 20
- Min_Child_Sample = 24
- Bagging_fraction = 0.8538
- Lambda_l1 = 0.7467
- Lambda_l2 = 0.6911

5.3 Accuracy

From the results of confusion matrix calculations carried out in the classification process in the Classification and Regression Tree method, Table 4 shows a summary of the values is generated:

Table 4. Confusion Matrix CART

	Predicted No	Predicted Yes	Accuracy (%)	Error Rate (%)
Actual No	740	48	85.9	14
Actual Yes	113	237		

From the results of confusion matrix calculations carried out in the classification process in the Light Gradient Boosting Machine method, a summary of the values is generated in Table 5 as follows:

Table 5. Confusion Matrix LightGBM

	Predicted No	Predicted Yes	Accuracy (%)	Error Rate (%)
Actual No	193	33	81	19
Actual Yes	32	84		

6. Conclusion

Analysis conducted on PT BPR Syariah Gebu Prima Medan's debtor data using CART classification resulted in 13 nodes consisting of 1 main node, 2 inner nodes, and 7 terminal nodes. Factors that influence the Bank's decision to pay credit by customers based on the CART classification tree are total income, ceiling and family members, which are sorted based on the amount of improvement value originating from the Gini value. While Lightgbm, the influencing factors are ceiling, total income, family member, age and gender, which are sorted based on the improvement value. However, the CART method has a higher accuracy rate of 84% than the LightGBM method, which is 81%.

References

Bidani, A. S., Mangunsong, F., and Siska, K., Banking sector in Covid 19, *Journal of the Law*, vol.9, 2020

Fan, J., Ma, X., Wu, L., Zhang, F., Yu, X., and Zeng, W., Light Gradient Boosting Machine: An efficient soft computing model for estimating daily reference evapotranspiration with local and external metrological data, *Journal of The Agriculture Water Management*, 2019.

Fibrianda, M.F., and Bhawiyuga, A., Comparative analysis of attack detection accuracy on computer networks with the Naive Bayes method and the support vector machine, *Journal of information technology and technology development*, vol.2, pp.3112-3123, 2018.

Indriani, F., and Kartini, D., The classification pattern of the UMKM business sector with CART uses the information gain feature selection, *Journal of the electrical engineering and informatics*, 2018.

Ma, X., Sha, J., Wang, D., Yu, Y., Yang, Q., and Niu, X., Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGBoost algorithms according to different high dimensional data cleaning, *Journal of the Electronica Commerce Research Applications*, vol.31, pp.24-39,2018

Machado, M. R., Karray, S., and Sousa, I. T. D., LightGBM: An effective decision tree gradient boosting method to predict customer loyalty in the finance industry, *Proceedings of the 14th International conference science and education,*2019.

Maulana, M., Rifqi, and Al-kanomi., M. A., Information Gain to determine the effect of attributes on credit approval classification, *Journal of the Litbang,* vol.9, 2015

Minastireanu, E. A., and Mesnita, G., LightGBM machine learning algorithm to one click fraud detection, *Journal of Information Assurance and Cybersecurity*, 2019.

Nurhayati, B., and Iswara, I. P., Development of unsupervised learning technique algorithms on big data analysis on social media as an online promotional media for the community, *Journal of the Informatics Engineering*, vol.12, 2019.

Prabawati, N. I., The performance of the Classification and Regression Tree (CART) algorithm in classifying the length of study of students participating in organisations at the State University of Jakarta*, Journal of informatics and computer engineering education*,2019.

Prasetyo,E., *Data mining concepts and applications using MATLAB*, ANDI, Yogyakarta, 2012.

Provenzano,A.R., Trifiro, D., Datteo, A., Giada, L., Jean, N., Riciputi, A., Pera, G. L., Spadaccino, M., Massaron, L., and Nordio, C., Machine learning approach for credit scoring, *Journal of arXiv Cornell University*, 2020.

Tanjung, R. H. and Kartiko, Application of the CART method to determine the factors that affect credit payments by customers, *Journal of the Industrial and Computational Statistic*, vol.2, pp 78-83,2017.

Timoveef, R., Classification and Regression Trees (CART) Theory and Applications*, Center of Applied Statistics and Economics Humboldt University,* 2014.

Trivedi, S. K., A study credit scoring modeling with different feature selection and machine learning approach, *Journal technology in society*,vol.63,2020.

Wang, Y., and Wang, T., Application of improved LightGBM model in blood glucose prediction, *Applied Science*, 2020

Zhang, S., Hu, Y., Tan, Z., Research on borrowers credit classification of P2P network loan based on LightGBM algorithm, *Journal of Embedded System*, vol.11,2019.

Biography

Imas Wihdah Misshuari is a master's student in mathematics at the Department of Mathematics, Faculty of Science and Mathematics, Universitas Diponegoro Semarang, Central Java, Indonesia. Her research interests include statistics, financial mathematics, and applied mathematics.

Ratna Herdiana is a senior lecturer at the Department of Mathematics, Faculty of Science and Mathematics, Universitas Diponegoro Semarang, Central Java, Indonesia. She holds a PhD degree in mathematics from the University of Queensland. She has taught in several universities including University of Indonesia, Universiti Teknologi Petronas, and Hail University-KSA. Her research interests include numerical methods, optimisation, and applied mathematics.

Farikhin is vice dean of Faculty of Science and Mathematics, Diponegoro University, Associate Professor at the Department of Mathematics, Faculty of Science and Mathematics, Universitas Diponegoro Semarang, Central Java, Indonesia. He holds a PhD degree in mathematics from Universiti Malaysia Terengganu. His research interests include analysis, scientific computation, and applied mathematics.

Teuku Afrizal is a senior lecturer in the Department of Public Administration, Universitas Diponegoro, Indonesia, and a researcher at the university's SDGs Center. He is an expert on Development Studies, especially on environmental sociology and anthropology, anthropology of development, community development studies, social theory, and demography. He gained research experience in several projects funded by Ford Foundation, Toyota Foundation, UNDP, Scalabrini Migration Center, IKMAS, UKM, National Grant (Malaysia Government), BorIIS, UMS, Sabah

Forestry Department, Jhon Hopkins University, Boston University, Sultan Mizan Royal Foundation (YDSM) and Malay Strategy Foundation. He has published several articles in international refereed journals, chapter in books and book. He is a member of the Malaysia Social Science Association

Jumadil Saputra is a PhD holder and works as a senior lecturer in the Department of Economics, Faculty of Business, Economics and Social Development, Universiti Malaysia Terengganu, Malaysia. He has published 125 articles Scopus/ WoS indexed. As a lecturer, he has invited as a speaker in numerous universities, the examiner (internal and external), the reviewer for article journal and proceeding, the conference committee, journal editorial board, and others. He is a professional member of the International Business Information Management Association (IBIMA), Ocean Expert: A Directory of Marine and Freshwater Professional, and Academy for Global Business Advancement (AGBA). His research areas are Quantitative Economics (Microeconomics, Macroeconomics, and Economic Development), Econometrics (Theory, Analysis, and Applied), Islamic Banking and Finance, Risk and Insurance, Takaful, i.e., financial economics (Islamic), mathematics and modelling of finance (Actuarial). His full profile can be accessed from https://jumadilsaputra.wordpress.com/home-2/.