

BANK CUSTOMER SEGMENTATION MODEL USING MACHINE LEARNING

Vira Bunga Tiara¹, Amril Mutoi Siregar², Dwi Sulistya Kusumaningrum³, Tatang Rohana⁴

^{1, 2, 3, 4}Department of Informatics, Faculty of Computer Science, Buana Perjuangan University, Karawang, Indonesia

email: if20.viratiara@mhs.ubpkarawang.ac.id¹, amrilmutoi@ubpkarawang.ac.id², dwi.sulistya@ubpkarawang.ac.id³, tatang.rohana@ubpkarawang.ac.id⁴

Abstract

Banks generally carry out marketing strategies by offering deposit products directly to customers. However, this method is less effective because it requires individualized communication without considering the customer's interest in the product offered. Therefore, this research aims to categorize the classification of bank customers into Yes and No. This research uses a dataset of bank deposits taken from KTM. This research uses a bank deposit dataset taken from Kaggle, the data consists of 11162 rows with 17 attributes. PCA technique was used for feature selection which was optimized by reducing the dimensionality of the dataset before modeling. It was found that the best model accuracy was SVM RBF kernel with C parameters achieving 80.51% accuracy and ANN 80.78%, but ANN showed a higher ROC graph than SVM because ANN performance results were faster than SVM. Thus, the overall performance measurement of ANN is much better.

Keywords : Classification, Deposit Bank, Machine Learning, PCA, Supervised Algorithm

Received: 05-02-2024 | **Revised:** 19-03-2024 | **Accepted:** 25-03-2024

DOI: <https://doi.org/10.23887/janapati.v13i1.75233>

INTRODUCTION

In an era where data availability is increasingly abundant, the banking sector is one of the most affected by technological advances, especially in data management and analysis [1]. Data is a valuable asset that can provide deep insights to improve marketing strategies and customer management. By leveraging marketing data, banks can analyze customer transaction patterns, product preferences and financial habits to develop more effective and targeted marketing strategies [2]. The combination of sophisticated marketing techniques and intelligent marketing data analysis enables banks to gain competitive advantage, reach customers more effectively, optimize product sales and maintain market share. One of the banking products that is the focus of this research is deposits. Bank deposits are becoming quite a popular and necessary financial instrument for many individuals and business entities [3].

In carrying out deposit marketing campaigns conducted by banks over the phone and sending messages to customers for marketing purposes, there are situations where officers need to contact clients more than once

to ascertain whether they are interested in subscribing to a deposit or not. This process is not only ineffective, but also requires significant expenditure. This ineffectiveness in the marketing process occurs because bank officers do not understand the characteristics of clients who have the potential to subscribe to deposits. The marketing department should be able to determine potential customers by considering factors such as trustworthiness, time limit, risk level, and credit objectives [4]. This is very important as the marketing department is responsible for keeping the customer from facing difficulties in repaying the loan, which is often the main risk in assessing any loan grant. With the large amount of customer data available, banks need a method or system that can ensure effectiveness, time efficiency, and cost in conducting the marketing campaign process. Data analysis can be done by using customer classification to evaluate the extent of the success of marketing campaigns towards customers which would be important data in improving the deposit marketing process.

Classification is a part of Machine Learning where the process involves finding a model or function that describes and

distinguishes a class or concept in the data. This model is generated through analysis of training data and then used to determine the class label of an unknown object [5]. Various techniques in Machine Learning such as Logistic Regression, Support Vector Machine, K-Nearest Neighbor, Random Forest, Artificial Neural Networks, and Decision Tree can be applied to solve various classification problems [6]. Several previous studies have discussed the use of PCA for various Machine Learning algorithms. This research uses Data Mining method and Feature Selection Method (PCA) with the proposed model is k-means to determine the potential customer group which gives 87.76% accuracy score with UCI Machine Learning Repository dataset of actual data about bank telemarketing [7]. This research successfully predicts telemarketing bank customers for deposits, using the correlation-based feature selection method combined with the Multilayer Perceptron Neural Networks classification method. Classification using MLPNN with varying Mean Squared Error (MSE) values has an accuracy rate of 80.5%, recall 29.0%, and precision 77.4% with the Portuguese banking institution dataset in UCI Machine Learning Repository [8]. Similarly, [9] Conducted research to predict Banking Customer Term Deposits, using Random Forest, Logistic Regression, SVC, and XGBoost methods. The results showed that the random forest and xgboost models were the most effective, with accuracy reaching 91.7% using data taken from Hugging Face. Another study developed a model to classify potential deposit customers using the Ensemble Least Square Support Vector Machine model with AdaBoost. The results showed that this method achieved an accuracy of 95.15%. The authors used the bank direct marketing dataset which can be

accessed through the University of California at Irvine (UCI) Machine Learning Repository [10].

Based on the explanation of the previous problem and with the support of previous research findings related to the application of feature selection to optimize the performance of machine learning algorithms in classifying marketing banks, this research develops various Machine Learning methods, and applies 6 Machine Learning algorithms namely, Logistic Regression, Support Vector Machine, K-Nearest Neighbor, Random Forest, Artificial Neural Network, and Decision Tree. The contribution to this research lies in the application of the PCA technique for each different classifier. Each classifier is assessed based on performance metrics, mainly using ROC (Receiver Operating Characteristic) and Confusion Matrix. With the classification results using supervised algorithms, customers with existing parameters can be classified between Yes and No. Therefore, this pattern can be used as a benchmark. Thus, this pattern can be used to benchmark customers who take time deposits (Yes) and customers who do not take time deposits (No). It is hoped that this model can help banks to support sustainable growth in the banking sector and increase the effectiveness of segmentation with various Machine Learning methods.

METHOD

Research Procedure

This research process begins with the literature review stage which aims to find the theoretical basis used and search for relevant scientific literature to support the research. In this overall research, the steps or stages of research are as follows:

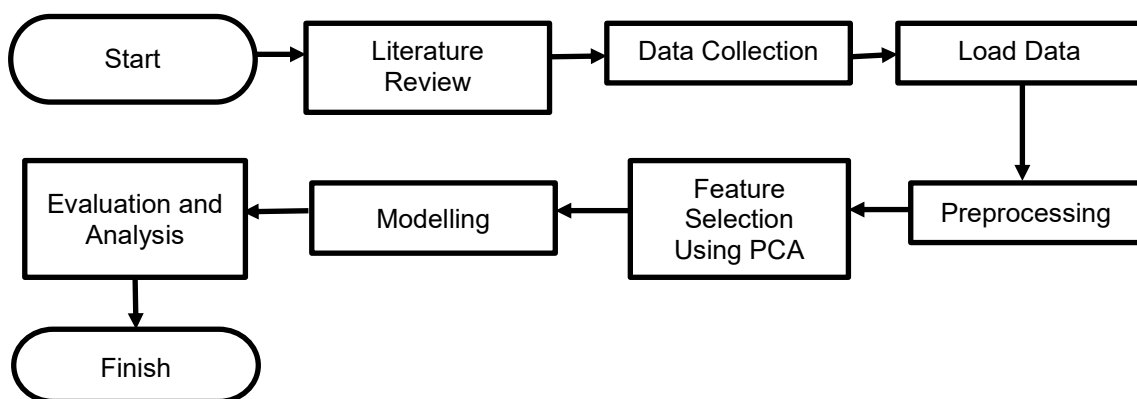


Figure 1. Flowchart of research stage

Data Collection

The dataset used in this study comes from the bank marketing dataset obtained from the Kaggle website, which can be accessed via the link (<https://www.kaggle.com/datasets/janiobachman/bank-marketing-dataset>). This dataset consists of 11162 rows with 17 attributes, this dataset consists of a number of attributes as seen in Table 1.

Then, this research conducted Exploratory Data Analysis (EDA). In this dataset, there are two final categories that are deposit attributes, namely class Yes and class No. These two classes will be the target of the classification process, as shown in the Figure 2.

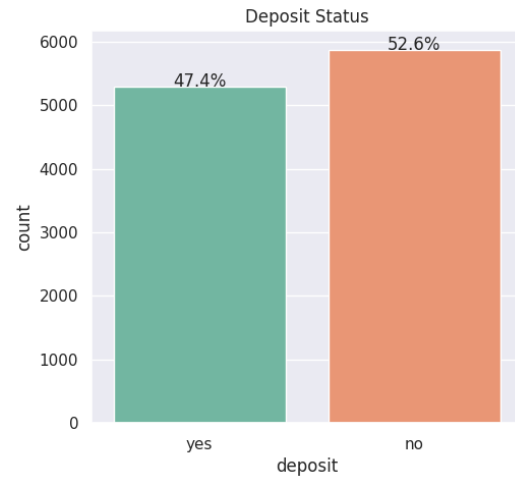


Figure 2. EDA distribution class in dataset

Table 1. Kaggle Bank Dataset Description

Attribute Name	Type	Description
<i>Age</i>	Numeric	Age
<i>Job</i>	Category	Employment Category
<i>Marital</i>	Categories	Marital Status Category
<i>Education</i>	Categories	Last Education Category
<i>Default</i>	Categories	Category Having Debt in arrears
<i>Balance</i>	Numeric	Customer Balance (in Euro currency)
<i>Housing</i>	Category	Category Having a Home Loan
<i>Loan</i>	Category	Category Is there a loan?
<i>Contact</i>	Categories	Category Type of Communication
<i>Day</i>	Numeric	Date of Last Contact within the Year
<i>Month</i>	Categories	Categorical Month of Last Contact within the Year
<i>Duration</i>	Numeric	Time of Last Call
<i>Campaign</i>	Numeric	Number of Contacts During Credit Offer
<i>Pdays</i>	Numeric	Number of Days Customer Contacted from Previous Offer
<i>Previous</i>	Numeric	Number of Calls Before Offer to Customer
<i>Poutcome</i>	Category	Past Offer Result Categories
<i>Deposito</i>	Categories	Current Offer Result Categories

Preprocessing

In the process of obtaining high-quality data, several techniques were used, including:

- 1) The initial stage involves the data cleaning process, where the dataset goes through a thorough cleaning process to ensure optimal quality. In




Figure 3. General look for missing values and duplicate data

this step, observed data values are carefully checked to ensure proper presentation without any duplicate prefixes for each feature [11]. It was found that the dataset used in this study had no missing values or duplicate data, as described in the information.

- 2) In this research, the process of transforming the data is done to improve the accuracy and efficiency of the algorithm. This process involves encoding categories into numerical values and categorical values and then separating them. For example, we assigned a value of 1 for deposit "Yes" and a value of 0 for deposit label "No". It is important to note that other features besides deposit, consist of both numerical and categorical data.

Categorical data describes information related to a particular category or group [12]. In this dataset, variables such as job, marital, education, default, housing, loan, contact, month, and outcome are categorical data. For example, the job column provides information about the customer's job type, such as admin, technician, or services. Marital indicates marital status such as married. Education describes the level of education such as secondary or tertiary. Meanwhile, default, housing, and loan represent yes/no categories, and contact can contain the type of contact used. This categorical data provides a complete picture of the non-numerical attributes of the individuals in the dataset. Information

such as employment type, marital status, education level, and more help in understanding the characteristics or profile of the individual represented by each row in this dataset.

- 3) Furthermore, data scaling and normalization are also carried out to produce standardized data using the Z-Score scaling method. The purpose of this normalization is to achieve standards in the data [13]. The formula applied for Z-Score scaling used in this data normalization process, is as follows:

$$Z = \frac{X - \mu}{\sigma} \quad (1)$$

Z is the resulting Z-Score value, X is an important value for considering feature standardization in data analysis. The symbol μ represents the mean value of the entire dataset, while σ signifies the standard deviation.

The use of Z-score was chosen because of its effectiveness in scaling data that has a normal distribution as well as its resistance to outliers [14]. The data consisting of 11162 rows has been divided into training and test data. The ratio of the two datasets was set at 80:20, with the training data consisting of 80% and the test data consisting of 20%. With a limited number of outliers in this dataset, the 20% set aside for test data is considered sufficient for the purpose of evaluating the final model. With a total of 8929 training data, while the test data is 2233 data. After the data is divided, the dataset is then feature selected using Principal Component Analysis (PCA). The next step is to enter the dataset into the model and run a series of testing processes.

Feature Selection Using PCA

PCA is a method that aims to reduce the dimensionality of data by processing component variables to lower dimensions, with the aim of maximizing the variance of the represented data [15]. The dataset applied in this study has a high number of rows consisting of 17 rows. The large number of rows is an obstacle in achieving optimal results and can cause overfitting. Therefore, PCA was applied to this dataset with the aim of converting 17 rows into only 11 rows, in an attempt to improve the performance results. The main benefits of PCA are to reduce the risk of overfitting, removal of correlated

features, and improvement of Machine

Learning algorithm performance [16].

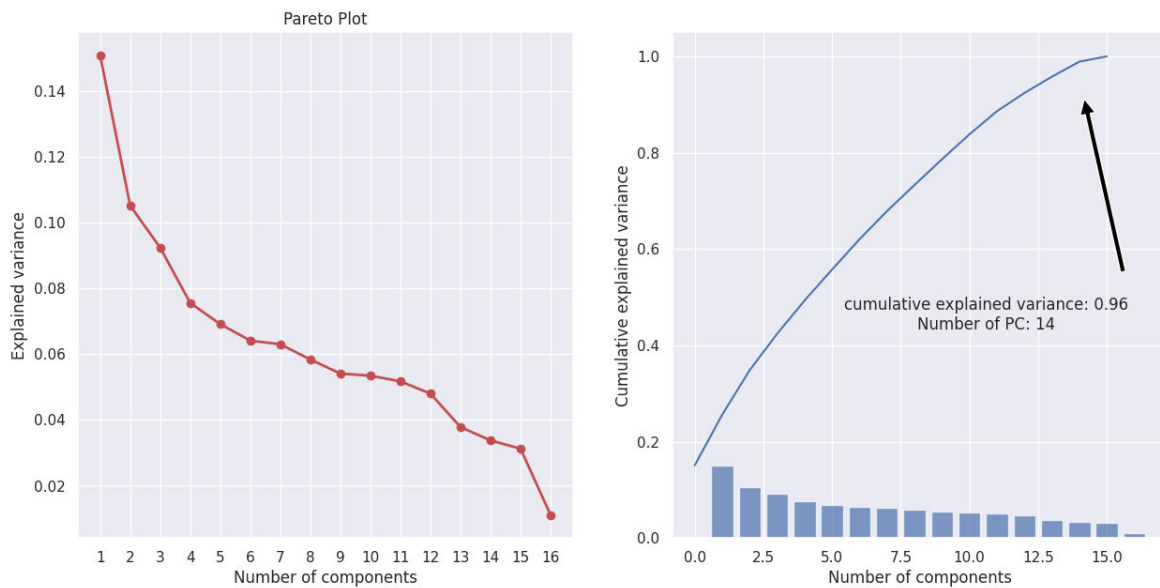


Figure 4. Pareto plot of eigenvalue in PCA

In this study, we used the pareto plot, as seen in Figure 5, to evaluate the eigenvalues and determine the optimal number of principal components used in the next modeling step. In a pareto plot graph on a PCA pareto plot curve, the x-axis generally shows the number of components, while the y-axis shows the single value or variance explained by each component. This curve provides a visual representation of the contribution of each component to the total variance in the dataset. The pareto plot graph indicates that the optimal number of principal components (PC) is 14, which explains a cumulative variance of 0.96. The concept of cumulative variance explained is the essence of PCA, a dimension reduction technique often used in multivariate data analysis. Afterwards, we used the dimensionality of the training and testing data for the next step in the modeling process.

Model Overview

1) Logistic Regression

Logistic Regression is a statistical technique that is often used to describe the relationship between binary dependent variables and one or more independent variables. This method is a type of supervised learning algorithm and is used to predict the probability of the target variable [17]. The target variable is constructed on the basis of linear regression to evaluate the output and minimize the error. Furthermore, a complex

estimation function is used which is either a sigmoid function or a logistic function. In a bank data environment, logistic regression can be applied to analyze variables such as decisions related to opening deposits, loans, or other financial decisions. The analysis is based on information such as age, employment type, balance, marital status, and other relevant variables contained in the dataset. The formula of logistic regression is used to model the probability of an outcome that is binary 0 or 1 based on the data set, where the model projects the probability using a logistic regression function [18]. The logistic regression function is described as follows:

$$P(Y = 1) = \frac{1}{1 + e - (b_0 - b_1x_1 + b_2x_2 + \dots + b_kx_k)} \quad (2)$$

$P(Y = 1)$ is the conditional probability that the dependent variable (y) is 1 given a set of values of the independent variable (x).

e is the power of the Euler number.

b_0 is a constant or bias.

b_1, b_2, \dots, b_k are the coefficients for each independent variable x_1, x_2, \dots, x_k . respectively.

x_1, x_2, \dots, x_k are the values of the independent variables.

It expresses the relationship between a binary dependent variable and a number of associated independent variables in the form of probabilities. The goal of logistic regression

is to find coefficient values $b_0, b_1, b_2, \dots, b_k$ that can predict the probability of a particular class of the dependent variable based on given values of the independent variables.

2) Support Vector Machine

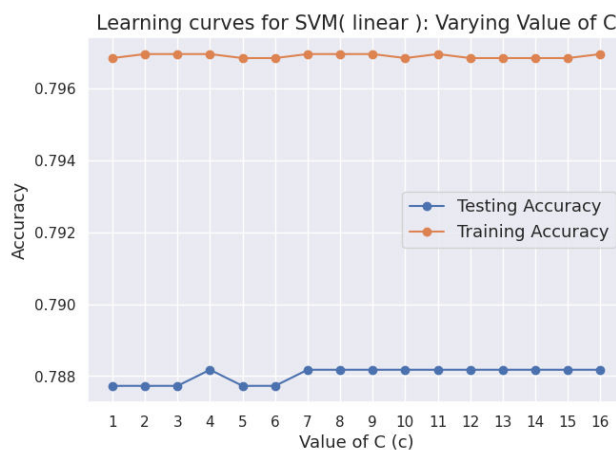
The Support Vector Machine (SVM) method can be used to classify target variables in bank datasets by utilizing relevant features [19]. It aims to predict or analyze certain target variables, such as decisions related to deposit opening or other financial aspects. It is important to select an appropriate kernel such as linear, rbf, or poly based on the specific characteristics of the data to ensure optimal results. It has proven its success in the fields of classification, regression, time series prediction, and estimation. The main goal of this approach is to find an optimal separating hyperplane that is able to correctly classify the data points as efficiently as possible and maximally distinguish between points from two classes [20]. The SVM formula is as follows:

$$f(x) = \sum_{i=1}^n a_i y_i k(x_i, x_j) + b \quad (3)$$

N is the number of training samples, a_i denotes the weights calculated in the training process, y_i is the class label of the i training sample, $K(x_i, x_j)$ stands for the kernel, and b is the bias value of the term. In this study, we compare the linear and RBF kernels to find out which kernel produces the best C value. the formulas of the Linear and RBF kernels are:

$$K(x_i, x_j) = x_i \cdot x_j \quad (4)$$

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \quad (5)$$



In the context of formula (4), x_i represents the feature vector of an example, while x_j refers to the feature vector of an example in the training set. Furthermore, in formula (5), $\|x_i, x_j\|^2$ represents the square of the Euclidean distance between x_i and x_j , where \exp denotes the exponential value of the number x , which is Euler's constant (approximately 2.71828), and σ is a parameter that controls the width of the RBF kernel. Linear and RBF kernel functions are the most commonly used in this context, the selection of alternative parameters is crucial to achieve a more optimal fit, hence this study uses the C parameter.

The optimal value used in this study lies in the RBF kernel with a C value of 4, because it provides similar accuracy in the training and testing stages, as shown in Figure 6. The C parameter serves as a control for misclassification of training data in SVM. The C value regulates how much punishment is given to misclassified data points against the separating hyperplane. A higher C value results in a larger determinant, making the model more rigorous in handling misclassifications in the training data. By choosing the right C value, we can optimize the balance between fitting the model to the training data and preventing overfitting.

Based on Figure 6. why RBF is much better than linear because this dataset is a non-linear dataset which contains various features that include demographic information, customer behavior, and marketing campaign details, and the target is whether the client accepts or does not accept the marketing offer.

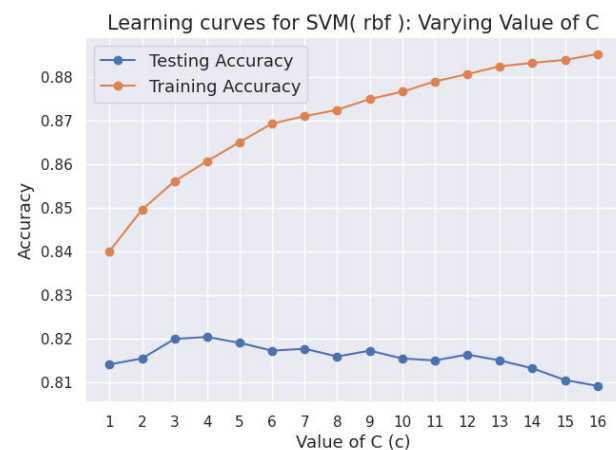


Figure 5. SVM linear and RBF kernel for best C value

The bank dataset used evaluates two types of approaches in SVM analysis: linear kernel and RBF kernel. It can be seen in the performance of the linear kernel, although this model is able to handle linear relationships between features and targets, but in the case of the bank dataset used, the performance of this model may be limited and has complex and non-linear relationships. Then, the RBF kernel approach is used to find a more complex non-linear mapping of the data to a higher dimension. Thus, it is able to capture more complicated relationships between features and targets in the dataset [20].

Comparing the results between linear kernel and RBF kernel, it can be concluded that SVM with RBF kernel gives much better performance than linear kernel to predict the decision of acceptance or rejection of marketing offers in this dataset. With higher accuracy the approach of using RBF kernel is a superior choice and obtains optimal results.

3) K-Nearest Neighbor

The KNN method is a simple and efficient non-parametric classification approach based on supervised learning [21]. KNN operates by identifying the k nearest samples from an existing dataset. When a new unknown sample appears, the algorithm classifies the sample into the most similar class. In other words, in the classification process, the algorithm determines the group of test samples with k training samples that are the closest class to the test samples, then attributes them to the highest class [22]. There are various distance measurement techniques that can be applied in this algorithm, and one of the commonly used techniques is the 'Euclidean' distance method, which is calculated using the formula :

$$d(x, y) = \sqrt{(x - y)^2} \quad (6)$$

In this research, a value of k of 5 was chosen which is the Euclidean distance (d) between two points x and y . The distance is measured using the previously mentioned formula for each neighbor.

4) Random Forest

Random Forest presents as an ensemble classifier for decision tree learners. This method uses multiple decision trees so that each decision tree depends on a random vector value chosen separately with the same

distribution for all decision trees. Random forest is actually a way to combine many decision trees learned on different sets of the same data with the target to reduce variance [23]. The advantage of using RF is that it comes with high dimensional data, without the need for dimensionality reduction and feature selection. The training rate is also higher and it is easy to use in parallel models. Random Forest uses the Gini coefficient to build the decision tree [24]. The training set has n features in the Gini index coefficient taken from the CART (Classification and Regression Trees) learning system to create a decision tree. The Gini coefficient measures the non-uniformity between values in a frequency distribution. The Gini coefficient has the following formula:

$$Gini(T) = 1 - \sum_{j=1}^n (P_j)^2 \quad (7)$$

In formula (7), a Gini coefficient of zero represents perfect similarity while a coefficient of 1 represents maximum dissimilarity between the values. If a dataset T contains examples from n classes and P_j is the relative frequency.

5) Artificial Neural Network

ANN is an artificial intelligence computational network designed based on the biological structure of the human brain [25]. Artificial Neural Network (ANN) has been widely adopted in a wide array of research, making it a significant research subject. In particular, the use of these networks has provided remarkable achievements, especially in marketing bank classification and early stage prediction. Typically, the structure of an ANN model consists of three layers: Input, Hidden, and Output. Each layer consists of a network of interconnected neurons, with a non-linear activation function that increases the network's capacity to understand non-linear patterns. The initial stage starts from the Input layer which receives data and passes it to the Hidden layer for processing. The result of the analysis is then sent to the Output layer, where the final result is obtained [26]. However, due to these limitations, training an Artificial Neural Network (ANN) is likely to involve a series of computationally complex processes. The activation process of an ANN is described in equation (8).

$$J_1 = \sum_k V_{1k} X_k + b_1 \quad (8)$$

$$K_1 = g_1(J_1)$$

The activation function in equation (8) is embodied in g_1 , while V_{lk} is defined as the weight connecting the input layer and the hidden layer. The term b_1 refers to the difference between the input layer and the hidden layer at each connection. In addition, X_k represents the input at the input layer, J_1 is the sum of the inputs weighted by the bias, and K_1 represents the output of the hidden layer generated by the activation function.

$$J_m = \sum_l V_{ml}X_l + b_m \quad (9)$$

$$K_m = g_m(J_m)$$

In formula (9), g_m reflects the activation function, V_{ml} indicates the weights connecting the output layer m with the hidden layer l , and b_m is the difference between the hidden layer and the output layer at each connection. X_l refers to the output of the hidden layer at each node, J_m is the sum of the weights at the output layer, and K_m represents the final output of the output layer. In the context of (2) and (3), m

indicates the output layer, l refers to the hidden layer, and k represents the input layer.

In this study, three dense hidden layers were implemented using ReLU and Sigmoid activation functions as shown in Figure 7. ReLU was used to replace negative values with zero while keeping positive values unchanged. Next, two levels of dropout and Sigmoid activation were applied with batch size = 100 and epoch = 500. The use of dropout layers aims to overcome the overfitting problem that occurs when the validation value of the loss is high, and the training loss is low. Therefore, this study uses early stopping with parameters verbose=1 and patience=40 to overcome it, as shown in Figure 8.

Early Stopping is a strategy in the model training process that stops the training process if there are indications of overfitting as measured through metrics on the validation dataset. This approach aims to prevent the model from memorizing excessive training data in order to improve the model's ability to perform better generalization [27].

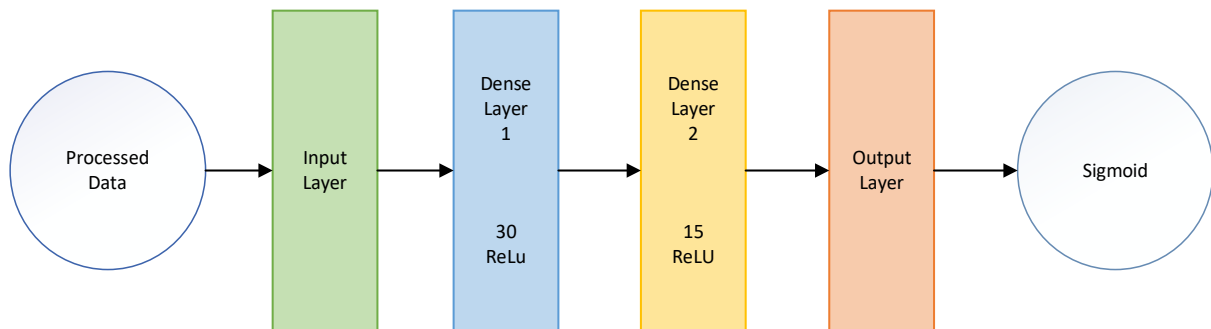


Figure 7. Proposed ANN flowchart method

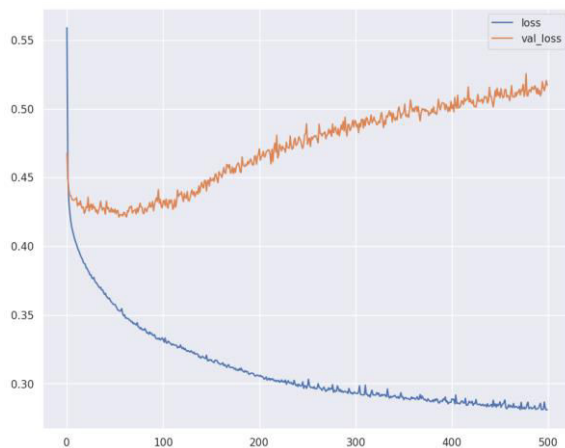


Figure 8. ANNs visualization before using Early Stopping

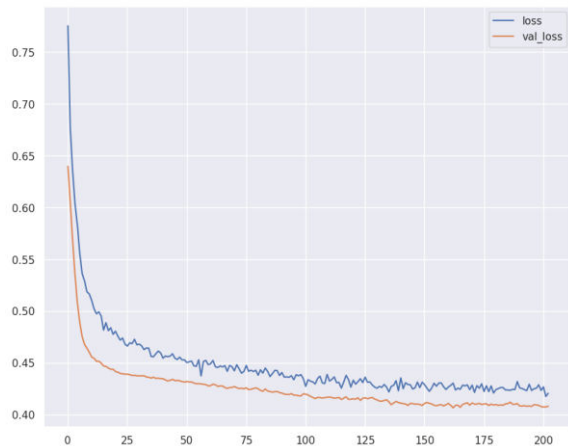


Figure 9. ANNs visualization after using Early Stopping

6) Decision Tree

Decision tree is an algorithm that uses a tree structure as a prediction model that is generally used for decision making. Each tree has branches that represent attributes that must be met to continue to the next branch until it reaches the leaves [28]. In data processing, the decision tree determine the attributes of the decision root by using a gain ratio calculation. The calculation illustrates that gain refers to how much information is obtained by knowing the attribute value, while split information is used for attributes that have more than two variations [29]. The rules applied to this decision tree model are defined using the conjunction 'IF'. The formula for Gain Ratio of an attribute *A* against a data set *S* in a classifier algorithm is as follows:

$$Gainratio(S, A) = \frac{Gain(S, A)}{SplitInformation(S, A)} \quad (10)$$

Gain(S, A) is the Gain value of attribute *A* against data set *S*.

Split Information(S, A) is the Split Information value of attribute *A* against data set *S*.

RESULT AND DISCUSSION

After data pre-processing, the classification performance is visually represented using a number of confusion matrices. Data pre-processing involves steps such as replacement of missing values and extraction of minimum and maximum values from the dataset. Subsequently, data scaling

and dimensionality reduction using Principal Component Analysis (PCA) was applied to all Machine Learning algorithms used in this study. Performance evaluation is done using confusion matrix and ROC curve.

Performance Measurement with confusion matrix and ROC curve

This research applies 6 methods to the dataset and each method measures the performance of the model using Accuracy, Precision, and Recall metrics. The formula used to measure performance is:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (11)$$

$$Precision = \frac{TP}{TP + FP} \quad (12)$$

$$Recall = \frac{TP}{TP + FN} \quad (13)$$

The following is the meaning of each formula in (11), (12), and (13) mentioned [3]:

- TP (True Positive) = (Positive results that are correctly classified as customers taking deposits).
- TN (True Negative) = (A negative result that is correctly classified as not taking a deposit).
- FP (False Positive) = (A negative result that is incorrectly classified as a customer taking a deposit, when in fact it is not: Type I error).
- FN (False Negative) = (A positive result that is misclassified as not taking a deposit, when in fact it did: Type II error).

Accuracy is the total amount of data that is correctly predicted by Machine Learning, compared to the total of all data which can be calculated in formula (11). Precision is the percentage of relevant elements that can express how often the model is able to make correct predictions, and can be calculated based on the formula in (12). Meanwhile, Recall is the percentage of relevant elements that are successfully classified correctly by the model against all relevant data, so it can be calculated by the formula that can be seen in (13). Below are the performance results of each algorithm shown in Table 2. as well as the confusion matrix with x labels representing the test data and y labels representing the model predictions, shown in Figure 9.

Table 2. Performance each algorithm based on confusion matrix

Model and	No (0)	Deposit (1)
-----------	--------	-------------

Accuracy Score	Accuracy	Precision	Recall	Accuracy	Precision	Recall
ANN (82.08%)	82%	88%	76%	82%	77%	89%
SVM:RBF (82.04%)	82%	84%	81%	82%	80%	83%
DT (72.77%)	72%	74%	75%	72%	72%	71%
RF (79.71%)	79%	82%	81%	79%	79%	80%
KNN (77.78)	76%	75%	82%	76%	78%	71%
LR (78.45%)	78%	78%	82%	78%	79%	75%

Table 3. Performance each algorithm without PCA

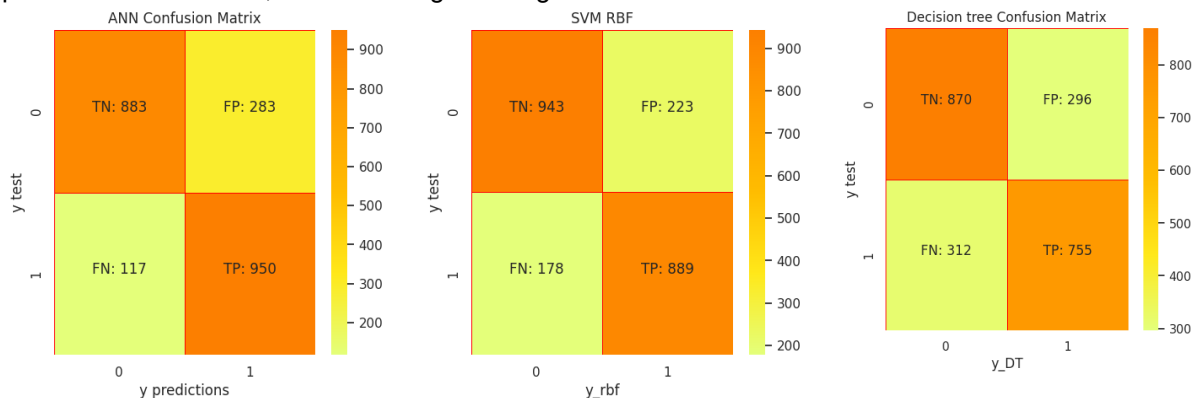
Model and Accuracy Score	No (0)			Deposit (1)		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
ANN (80.51%)	80%	86%	75%	80%	76%	86%
SVM:RBF (78.68%)	78%	83%	74%	78%	75%	84%
DT (74.74%)	74%	78%	72%	74%	72%	78%
RF (81.32%)	81%	90%	72%	81%	75%	92%
KNN (72.81)	72%	71%	80%	72%	75%	65%
LR (79.08%)	79%	82%	76%	79%	76%	82%

The findings from the experiments conducted show that the use of PCA significantly improves the performance of the model in this study. By using PCA, the model successfully overcomes the overfitting problem and provides more accurate predictions. In contrast, if PCA is not used, the model tends to give less satisfactory results. This is mainly due to the high-dimensional complexity of the dataset, which complicates the model training process and increases the risk of overfitting. Thus, it is concluded from the results conducted that using PCA is much more effective than not using PCA in analyzing this bank dataset. PCA helps to improve model performance, reduce complexity, and produce more accurate predictions. Therefore, when dealing with high-

dimensional datasets, the use of PCA is a more advisable option to optimize model performance.

Comparison of Method Performance with ROC Curve

This research compares the performance of the proposed method using the ROC (Receiver Operating Characteristic) Curve, which serves as a method in model selection and evaluation in two-class classification problems [30]. The ROC curve can be generated from the True Positive Rate (TPR) and False Positive Rate (FPR) results calculated from the confusion matrix, as shown in formulas (15) and (16).



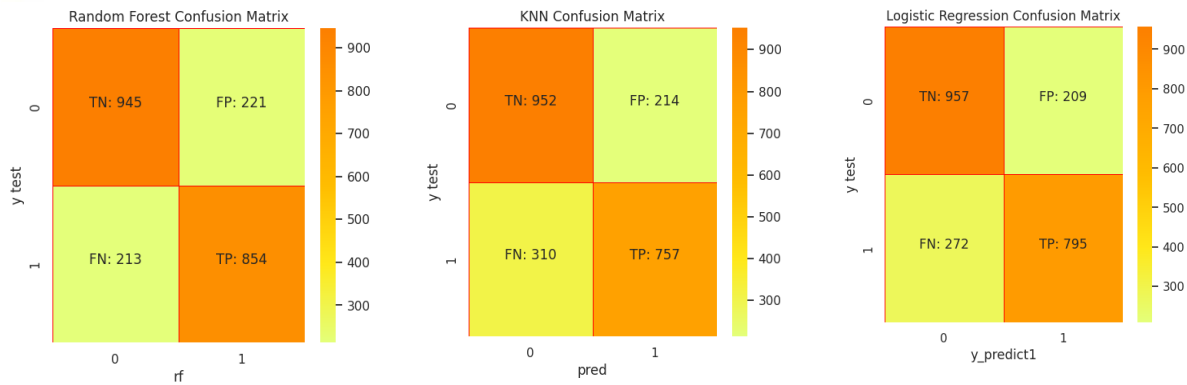


Figure 10. Confusion matrix of all models with PCA

The TPR and FPR values of each model are presented in Table 3. while the visualization is shown in Figure 10. with TPR (True Positive Rate) as the x-label and FPR (False Positive Rate) as the y-label of each algorithm. The formula is:

$$TPR = \frac{TP}{Actual\ Positive} = \frac{TP}{TP + FN} \quad (14)$$

$$FPR = \frac{FP}{Actual\ Negative} = \frac{FP}{TN + FP} \quad (15)$$

The classification results of all Machine Learning models can be seen from the highest accuracy resulting from the SVM algorithm 82.02% and ANN 82.08%, while the lowest accuracy is found in the Decision Tree algorithm with a value of 72.77%. Although SVM and ANN have accuracy with slightly different results, the ROC (Receiver Operating Characteristic) curve in ANN is wider and the False Positive Rate (FPR) value is also better than SVM.

CONCLUSION

Based on research conducted with a dataset obtained from the Kaggle website regarding bank deposits, the use of feature selection with the PCA method was applied to

various Supervised Machine Learning algorithms. Based on research conducted with a dataset obtained from the Kaggle website regarding bank deposits, the use of feature selection with the PCA method was applied to various Supervised Machine Learning algorithms. The results showed that the highest accuracy was achieved by the ANN algorithm, which reached 82.08%. Evaluation through the ROC curve shows that the ANN graph has a higher performance due to the confusion matrix calculation results with FP=0 and FPR=0 values. When FP and FPR are 0, this is considered a good result as it shows that the classification model does not make a mistake by predicting something as Yes when it is not. From the overall analysis, the performance of SVM with RBF kernel and using c-value selection approach shows better results compared to all Machine Learning algorithms tested in this study. For future research, it is recommended to perform several approaches to improve the accuracy value and classification quality in classifying bank marketing data to improve marketing. One of them is by using other feature selection techniques such as forward selection to get the best set of attribute

Table 4. Result of FPR and TPR each algorithm

Method	FPR	TPR
ANN	0.0,0.22469983,1.0	0.0,0.86597938,1.0
SVM RBF	0.0,0.18696398,1.0	0.0,0.76101218,1.0
DT	0.0,0.24957118,1.0	0.0,0.70290534,1.0
RF	0.0,0.18782161,1.0	0.0,0.81068416,1.0
KNN	0.0,0.18353345,1.0	0.0,0.70946579,1.0
LR	0.0,0.17924528,1.0	0.0,0.74507966,1.0

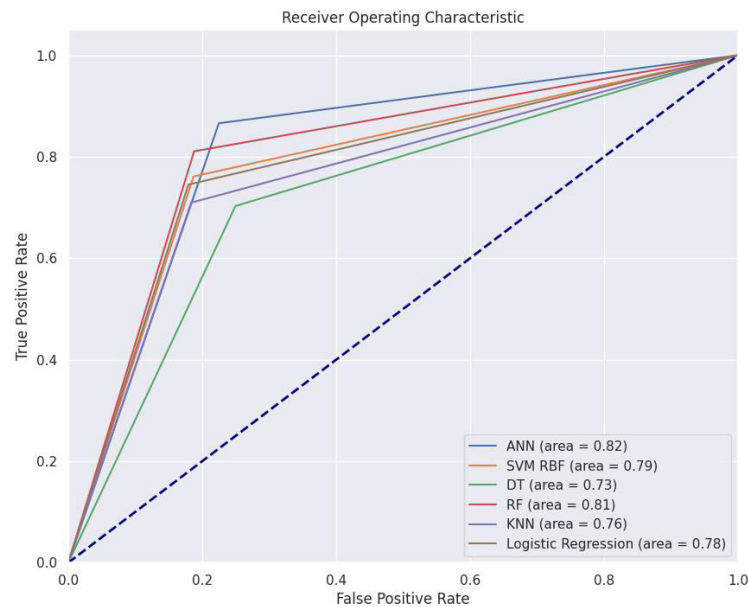


Figure 10. ROC of all models

REFERENCES

- [1] F. Izhari, "Teknik Machine Learning untuk Bank Marketing Dataset," *Semin. Nas. Inform.*, 2021, [Online]. Available: <https://www.ejournal.pelitaindonesia.ac.id/ojs32/index.php/SENATIKA/article/view/1187>
- [2] A. D. R. Prabowo and M. Muljono, "Prediksi Nasabah Yang Berpotensi Membuka Simpanan Deposito Menggunakan Naive Bayes Berbasis Particle Swarm Optimization," *Techno.Com*, vol. 17, no. 2, pp. 208–219, 2018, doi: 10.33633/tc.v17i2.1648.
- [3] K. N. Abd Halim*, A. S. Mohd Jaya, and A. F. A. Fadzil, "Data Pre-Processing Algorithm for Neural Network Binary Classification Model in Bank Tele-Marketing," *Int. J. Innov. Technol. Explor. Eng.*, vol. 9, no. 3, pp. 272–277, 2020, doi: 10.35940/ijitee.c8472.019320.
- [4] M. A. Sheikh, A. K. Goel, and T. Kumar, "An Approach for Prediction of Loan Approval using Machine Learning Algorithm," *Proc. Int. Conf. Electron. Sustain. Commun. Syst. ICESC 2020*, no. Icesc, pp. 490–494, 2020, doi: 10.1109/ICESC48915.2020.9155614.
- [5] D. Wang, "Research on bank marketing behavior based on machine learning," *ACM Int. Conf. Proceeding Ser.*, pp. 150–154, 2020, doi: 10.1145/3421766.3421800.
- [6] E. Villamosm, I. Kar, and D. Tansz, "Machine Learning Based Customer Decision Support," pp. 1196–1201, 2020.
- [7] M. Al Hammadi, "Identifying Prospective Clients for Long-Term Bank Deposit," 2022, [Online]. Available: <https://scholarworks.rit.edu/theses>
- [8] A. N. Puteri, A. Arizal, and A. D. Achmad, "Feature Selection Correlation-Based pada Prediksi Nasabah Bank Telemarketing untuk Deposito," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 20, no. 2, pp. 335–342, 2021, doi: 10.30812/matrik.v20i2.1183.
- [9] M. Rasikh, A. Riyyasy, W. N. Aghniya, and H. Tantyoko, "Penerapan Algoritma Machine Learning Untuk Memprediksi Term Deposit Nasabah Perbankan," vol. 8798, 2023.
- [10] F. Aziz and B. L. E. Panggabean, "Klasifikasi Nasabah Potensial menggunakan Algoritma Ensemble Least Square Support Vector Machine dengan AdaBoost," *J. Inform. J. Pengemb. IT*, vol. 8, no. 3, pp. 269–274, 2023, doi: 10.30591/jpit.v8i3.5675.
- [11] Z. Song, "A Data Mining Based Fraud Detection Hybrid Algorithm in E-bank," *Proc. - 2020 Int. Conf. Big Data, Artif. Intell. Internet Things Eng. ICBAIE 2020*, pp. 44–47, 2020, doi: 10.1109/ICBAIE49996.2020.00016.
- [12] A. F. R. N. L. N. S. Z. N. Taufik Hidayatulloh 1, "Algoritma C4.5 Untuk Menentukan Kelayakan Pemberian Kredit," *J. Larik*, vol. Vol.2No.2, no. 2, pp.

- 1–11, 2022.
- [13] L. Liu, "A Self-Learning BP Neural Network Assessment Algorithm for Credit Risk of Commercial Bank," *Wirel. Commun. Mob. Comput.*, vol. 2022, 2022, doi: 10.1155/2022/9650934.
- [14] K. Alkhatib and S. Abualigah, "Predictive Model for Cutting Customers Migration from banks: Based on machine learning classification algorithms," *2020 11th Int. Conf. Inf. Commun. Syst. ICICS 2020*, pp. 303–307, 2020, doi: 10.1109/ICICS49469.2020.239544.
- [15] M. Z. Khan et al., "The Performance Analysis of Machine Learning Algorithms for Credit Card Fraud Detection," *Int. J. online Biomed. Eng.*, vol. 19, no. 3, pp. 82–98, 2023, doi: 10.3991/ijoe.v19i03.35331.
- [16] A. Abdulhafedh, "Incorporating K-means, Hierarchical Clustering and PCA in Customer Segmentation," *J. City Dev.*, vol. 3, no. 1, pp. 12–30, 2021, doi: 10.12691/jcd-3-1-3.
- [17] B. Adamyk, A. Skirka, K. Snihur, and O. Adamyk, "Analysis of Trust in Ukrainian banks based on Machine Learning Algorithms," *2019 9th Int. Conf. Adv. Comput. Inf. Technol. ACIT 2019 - Proc.*, pp. 234–239, 2019, doi: 10.1109/ACITT.2019.8779974.
- [18] P. Hemalatha and G. M. Amalanathan, "A Hybrid Classification Approach for Customer Churn Prediction using Supervised Learning Methods: Banking Sector," *Proc. - Int. Conf. Vis. Towar. Emerg. Trends Commun. Networking, ViTECoN 2019*, pp. 1–6, 2019, doi: 10.1109/ViTECoN.2019.8899692.
- [19] M. A. Rahim, M. Mushafiq, S. Khan, and Z. A. Arain, "RFM-based repurchase behavior for customer classification and segmentation," *J. Retail. Consum. Serv.*, vol. 61, no. September 2020, p. 102566, 2021, doi: 10.1016/j.jretconser.2021.102566.
- [20] V. Djuricic, L. Kascelan, S. Rogic, and B. Melovic, "Bank CRM Optimization Using Predictive Classification Based on the Support Vector Machine Method," *Appl. Artif. Intell.*, vol. 34, no. 12, pp. 941–955, 2020, doi: 10.1080/08839514.2020.1790248.
- [21] M. H. Effendy, D. Anggraeni, Y. S. Dewi, and A. F. Hadi, "Classification of Bank Deposit Using Naïve Bayes Classifier (NBC) and K–Nearest Neighbor (K–NN)," *Proc. Int. Conf. Math. Geom. Stat. Comput. (IC-MaGeStiC 2021)*, vol. 96, pp. 163–166, 2022, doi: 10.2991/acsr.k.220202.031.
- [22] F. Itoo, Meenakshi, and S. Singh, "Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection," *Int. J. Inf. Technol.*, vol. 13, no. 4, pp. 1503–1511, 2021, doi: 10.1007/s41870-020-00430-y.
- [23] A. Hidayanti, A. M. Siregar, S. Arum, P. Lestari, and Y. Cahyana, "Model Analisis Kasus Covid-19 di Indonesia Menggunakan," *J. Pengkaj. dan Penerapan Tek. Inform.*, vol. 15, no. 1, pp. 91–101, 2022.
- [24] R. Sistem, "JURNAL RESTI Comparison of the Accuracy of Drug User Classification Models Using," vol. 5, no. 158, pp. 1348–1353, 2023.
- [25] I. E. Tsolas, V. Charles, and T. Gherman, "Supporting better practice benchmarking: A DEA-ANN approach to bank branch performance assessment," *Expert Syst. Appl.*, vol. 160, p. 113599, 2020, doi: 10.1016/j.eswa.2020.113599.
- [26] N. Ghatasheh, H. Faris, I. AlTaharwa, Y. Harb, and A. Harb, "Business analytics in telemarketing: Cost-sensitive analysis of bank campaigns using artificial neural networks," *Appl. Sci.*, vol. 10, no. 7, pp. 8–13, 2020, doi: 10.3390/app10072581.
- [27] Y. Deng, D. Li, L. Yang, J. Tang, and J. Zhao, "Analysis and prediction of bank user churn based on ensemble learning algorithm," *Proc. 2021 IEEE Int. Conf. Power Electron. Comput. Appl. ICPECA 2021*, pp. 288–291, 2021, doi: 10.1109/ICPECA51329.2021.9362520.
- [28] P. Appiahene, Y. M. Missah, and U. Najim, "Predicting Bank Operational Efficiency Using Machine Learning Algorithm: Comparative Study of Decision Tree, Random Forest, and Neural Networks," *Adv. Fuzzy Syst.*, vol. 2020, 2020, doi: 10.1155/2020/8581202.
- [29] C. Chen, L. Geng, and S. Zhou, "Design and implementation of bank CRM system based on decision tree algorithm," *Neural Comput. Appl.*, vol. 33, no. 14, pp. 8237–8247, 2021, doi: 10.1007/s00521-020-04959-8.
- [30] D. Boughaci and A. A. K. Alkhalwaldeh, "Appropriate machine learning techniques for credit scoring and bankruptcy prediction in banking and finance: A comparative study," *Risk Decis. Anal.*, vol. 8, no. 1–2, pp. 15–24, 2020, doi: 10.3233/RDA-180051.

