Journal of Computer Networks, Architecture and High Performance Computing
Volume 4, Number 2, July 2022
https://doi.org/10.47709/cnahpc.v4i2.1488

Submitted : May 9, 2022
Accepted : July 10, 2022
Published : July 18, 2022

Performance Comparison Supervised Machine Learning Models to Predict Customer Transaction Through Social Media Ads

Afandi Nur Aziz Thohari[1)*], Rima Dias Ramadhani[2)]
[1)] Politeknik Negeri Semarang, Indonesia
[2)] Institut Teknologi Telkom Purwokerto, Indonesia
[1)]afandi@polines.ac.id, [2)]rima@ittelkom-pwt.ac.id

ABSTRACT

The application of machine learning has been used in various sectors, one of which is digital marketing. This research compares the performance of six machine learning algorithms to predict customer transaction decisions. The six algorithms used for comparison are Perceptron, Linear Regression, K-Nearest Neighbors, Naïve Bayes, Decision Tree, and Random Forest. The dataset is obtained from Facebook ads transaction data in 2020. The goal is to get a model that has the best performance so that it can be deployed to the web. The method that is used to compare the results is a confusion matrix and also uses visualization of the model to get the prediction error that occurred. Based on the test results, the random forest algorithm has the highest accuracy, recall, and f1-score values, with scores of 96.35%, 95.45%, and 93.32%. The logistic regression algorithm generated the highest precision value, which was 94.44%. Based on the data visualization presented by the random forest algorithm, it has minor prediction errors; there are four data. Therefore, it can be concluded that the random forest algorithm has the best performance because it has the highest value in the three confusion matrix measurements and the smallest data prediction error. The model of the random forest algorithm is deployed to the web platform and can be accessed at the link iklan-sosmed.herokuapp.com.

Keywords: Customer Transaction; Machine Learning; Prediction; Performance Comparison; Random Forest

INTRODUCTION

Currently, Indonesia has entered the industrial revolution 4.0. The entry of the industrial revolution 4.0 is marked by automation in all fields. The impact of industrial revolution 4.0 is replacing human workers with machines. Machines will replace repetitive work because devices or computers can be more intelligent than humans in a specific area (Maguire, Moser, and Maguire 2020). The intelligence that machines acquire results from repeated training, even thousands or millions of times. The machine training process will produce a learning model that can be used to predict or associate a group of data. The method of processing data into a model form is called machine learning.

Machine learning is part of artificial intelligence, the main pillar of the industrial revolution 4.0. There are two machine learning approaches in terms of data usage: supervised learning and unsupervised learning (Berry, Mohamed, and Yap 2019). Supervised learning uses labeled data, while unsupervised learning uses unlabeled data. Supervised learning is used for classification and regression, while unsupervised learning is used for clustering and association.

Currently, a lot of research on machine learning has been done. One topic of machine learning research that is often carried out is regression or prediction. Several studies regarding the implementation of machine learning for prediction are research conducted by (Fitrianah et al. 2021) which predicts house prices using linear regression. Then paper by (Fitrianah et al. 2021) researched expect potential customers using the naïve Bayes algorithm. The previous research from (Edric and Tamba 2022) used a random forest algorithm to predict heart failure. These three studies can make predictions with an accuracy rate above 85%. Based on the prediction results, machine learning can indirectly help in decision-making.

Machine learning model generated from data processing using machine learning algorithms. The research problem is to create a machine learning model that can predict customer decisions to buy goods on social media. There are many machine learning algorithms currently being developed. This study will compare the performance of each machine learning model to find out the Gap in the resulting accuracy.

Based on a literature study (Hindrayani, Anjani, and Nurlaili 2021), six machine learning algorithms can be

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 4, Number 2, July 2022
https://doi.org/10.47709/cnahpc.v4i2.1488

Submitted : May 9, 2022
Accepted : July 10, 2022
Published : July 18, 2022

used to make accurate predictions. The six algorithms are Linear Regression, Perceptron, Naive Bayes, K-Nearest Neighbors, Decision Tree, and Random Forest. The purpose of this research is to find out which machine learning algorithm has the best performance in predicting transaction decisions. The dataset is taken from the customer's transaction recap on Facebook ads. The benefits of this research can help sellers to determine target customers. If ads appear on the right target, the number of sales will increase. The machine learning algorithm that has the best performance will be deployed to the web, making it easier for sellers to make predictions.

LITERATURE REVIEW

Perceptron

Perceptron is a machine learning algorithm that is used for linear classification. This algorithm belongs to Supervised Machine Learning, which is used for Binary Classification problems. The binary type will separate the input data into two classes, for example, class C1 and class C2. The two classes that have been formed must be linearly separable, as shown in Fig.1.

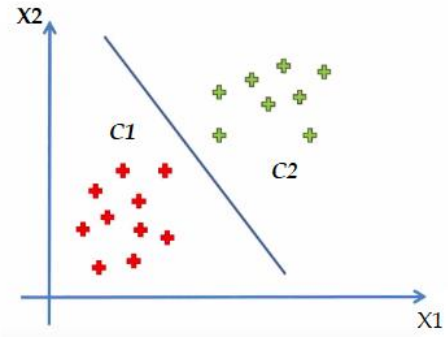


Fig.1 Lineary Separable

The perceptron network architecture is similar to the Hebb network architecture. Fig. 2 is a perceptron network diagram consisting of several input units and one output unit (Yudhistiro 2017).

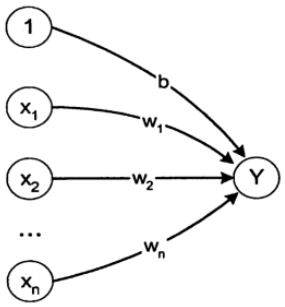


Fig. 2 Perceptron Network Diagram

X is the input data, while W is the weight. The output value (Y) is obtained from the result of multiplying the input data (X) with the weight (W). The perceptron network calculation process is shown in equation 1.

$$W_1X_1 + W_2X_2 + \ldots\ldots + W_nX_n + b = \square \qquad (1)$$

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 4, Number 2, July 2022
https://doi.org/10.47709/cnahpc.v4i2.1488

Submitted : May 9, 2022
Accepted : July 10, 2022
Published : July 18, 2022

Bias (b) is a hyperparameter used during the training process and has a positive value in the range of not more than 0.0–1.0. Bias controls how quickly model changes are made in response to the estimated error each time the 'weight' is updated.

Logistic Regression

Logistic regression is used to find the relationship between features (inputs) and the probability of output results. This classification algorithm was developed to overcome the problem of data outliers that cannot be solved using a linear regression algorithm. There are 3 important processes that are used to get high accuracy using the logistic regression algorithm (Saiful 2021).

1. Determination of coefficient with Maximum Likelihood+R-squared (R^2),
2. Determination of coefficients with Gradient Descent
3. Data Preparation on Logistic Regression

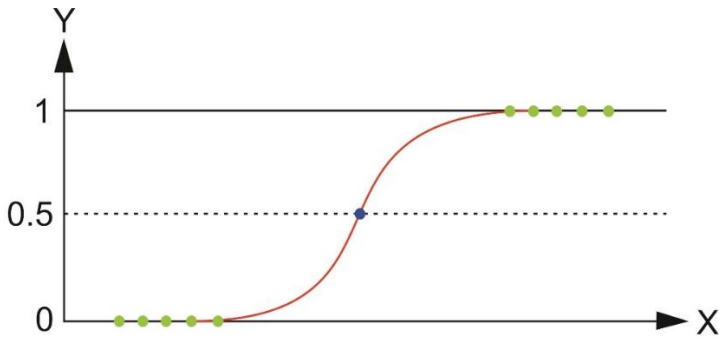


Fig. 3 Logistic Curve Function

The Logistic Function described in Fig.3 is a function formed by equating the Y value in the Linear Function with the Y value in the Sigmoid Function. The purpose of the Logistic Function is to represent the data that we have in the form of a Sigmoid function. The equation of the Logistic Function is shown in equation 1 (Kudryashov 2015).

$$\frac{1}{1 + w^{-(b0 + b1 * X)}} \qquad (2)$$

The graph of the logistics function in Fig. 3 can reach all data outliers that cannot be reached with a linear function. The logistic function is a combination of inverse sigmoid function and linear function. In equation 1 there is a coefficient b0 + b1 * X, which is used for line changes. The way to get the highest likelihood is by filling in the values of b0 and b1, converting them into sigmoid form, and calculating the likelihood value. Keep repeating the process to fill in the values of b0 and b1 until you get the highest likelihood value. The higher the likelihood value, the higher the accuracy of the resulting model.

K-Nearest Neighbors

K-Nearest Neighbors, abbreviated as KNN, is a supervised learning algorithm for classification. K is the number of nearest neighbors of the new data we want to predict. The classification process is carried out by calculating the closest distance to the neighboring data. The most common technique to find the nearest neighbor is calculating each point's Euclidean distance. Euclidean distance will see the spread between 2 points in two-dimensional space. Equation 2 calculates the euclidean distance (Dokmanic et al. 2015).

$$d = \sqrt{(X2 - X1)^2 - (Y2 - Y1)^2} \qquad (3)$$

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 4, Number 2, July 2022
https://doi.org/10.47709/cnahpc.v4i2.1488

Submitted : May 9, 2022
Accepted : July 10, 2022
Published : July 18, 2022

Calculating the Euclidean distance in two dimensions (d), requires the Cartesian coordinates represented by the x and y axes. The x1 and y1 coordinates indicate the location of the data that is the nearest neighbor. At the same time, the coordinates x2 and y2 indicate the location of the new data. Using equation 2, the closest distance from two location points can be seen, namely the location of the new data and the location of the neighboring data. In addition to the Euclidean distance, several formulas are used to calculate 2-point locations: Hamming distance, Manhattan Distance, and Minkowski distance.

Naïve Bayes

The Naive Bayes algorithm is a type of supervised learning that studies the probability of an object with specific characteristics being grouped into a particular class. Using Naive Bayes requires past data because this classifier predicts future possibilities based on past experience (Han, Kamber, and Pei 2011). The advantage of Naive Bayes requires small training data to determine the parameter estimates needed in the classification process. Naive Bayes uses Bayes' theorem to calculate predictive probabilities. The equation of Bayes' theorem is shown in equation 3 (Vembandasamy, Sasipriya, and Deepa 2015).

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)} \qquad (4)$$

Bayes' theorem in equation 2 produces the probability value of hypothesis C based on condition X (posterior probability). X is data with an unknown class, while C is data hypothesis. P(C) is an introductory probability class, while P(X) is a previous probability predictor. P(X|C) is the likelihood which is the probability of predictor given class.

Decision Tree

A decision tree is a popular machine learning algorithm used for classification and prediction. The structure of this algorithm is described in the form of a tree with several branches. Making a decision tree starts with making various choices and investigating the possible outcomes of those choices. There are three elements in making a decision tree, as follows (Charbuty and Abdulazeez 2021):

1. root node : ultimate goal or major decision to be taken
2. branches : various action options
3. leaf node : possible outcome of each action

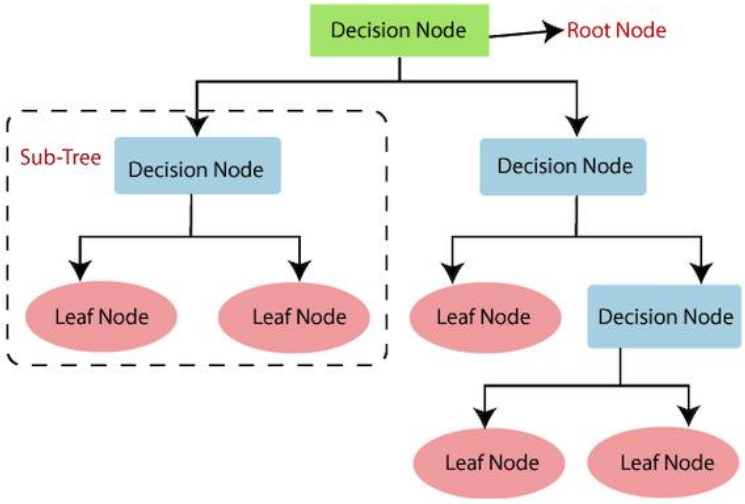


Fig. 4 Decision Tree Visualization

Fig. 4 is a visualization of the use of a decision tree. Generally, a decision tree starts with a single node or node. Then the node branches to state the available options. Furthermore, each unit will have new additions. Therefore, this method is called a 'tree' because its shape resembles a tree with many branches. Some of the advantages of

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 4, Number 2, July 2022
https://doi.org/10.47709/cnahpc.v4i2.1488

Submitted : May 9, 2022
Accepted : July 10, 2022
Published : July 18, 2022

decision trees are that they are easy to understand and analyze, numerical or categorical, require little processing, and are easy to conclude.

Random Forest

Random Forest is the development of the decision tree algorithm. This supervision algorithm is used for the classification of large amounts of data. Random forest classification is done through tree merging by conducting training on the sample data held. Using more trees will affect the accuracy that will be obtained for the better. Determination of the classification by random forest is taken based on the voting results of the formed tree (Kullarni and Sinha 2013). The winner of the tree created is determined by the most votes. Fig. 5 is an illustration of using the random forest algorithm.

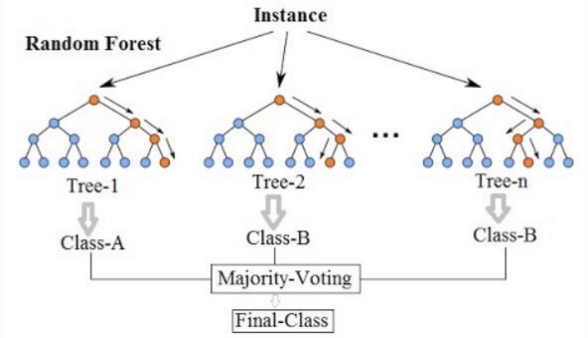


Fig. 5 Random Forest Structure

The classification process in the random forest begins with breaking the existing sample data into a random decision tree. After the tree is formed, voting will be carried out on each class from the sample data. Then, combine the votes from each category and take the most votes. Random forest in data classification will produce the best vote (Paul et al. 2018). The advantage of using a random forest is that it is able to classify data that has incomplete attributes, can be used for classification and regression, and can be used to handle large sample data.

METHOD

The method we use to predict customer transactions is shown in Fig. 6. There are several stages that must be passed to get an accurate prediction model, namely the sets of data collection, data preprocessing, data splitting, creating a classifier, and model evaluation. The explanation of each stage is as follows.

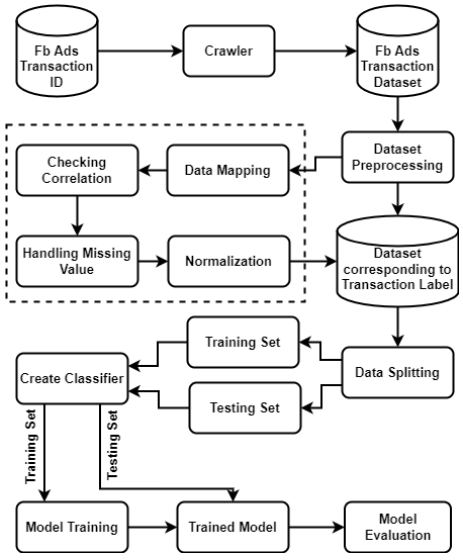


Fig. 6 Our Purposed Methodology

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 4, Number 2, July 2022
https://doi.org/10.47709/cnahpc.v4i2.1488

Submitted : May 9, 2022
Accepted : July 10, 2022
Published : July 18, 2022

Data Retrieval

The data comes from transactions that occur through Facebook ads during 2020. Data retrieval is carried out using a web crawler technique. The results of the scan or crawler are customer transaction data consisting of 5 fields, namely ID, Gender, Age, Salary, and Transaction. The number of datasets used in this research is 400. This data will be processed to get a machine learning model.

1. Preprocessing Data

After getting the dataset, the dataset preprocessing process is carried out which consists of several processes.

a. Data Mapping

The data contained in the transaction field is categorical data. During the age and salary fields are numerical data. Therefore, it is necessary to do a mapping to convert categorical transaction data into numerical data. The resulting labels are 0 and 1. A value of 0 means no, while a value of 1 means yes

b. Checking Correlation

The process to determine the correlation or relationship of each field. This process is essential to know what inputs are used to predict customer transactions. In this research, the information included is age and salary because they have the most significant correlation value to the transaction. The correlation data for each field is shown in Table 1.

Table 1. Correlation of Each Field

	ID	Gender	Age	Salary	Transaction
ID	1.000000	0.025249	-0.000721	0.071097	0.007120
Gender	0.025249	1.000000	0.073741	0.060435	0.042469
Age	-0.000721	0.073741	1.000000	0.155238	**0.622454**
Salary	0.071097	0.060435	0.155238	1.000000	**0.362083**
Transaction	0.007120	0.042469	**0.622454**	**0.362083**	1.000000

c. Handling Missing Value

There are some missing or unfilled data during the data retrieval process using web crawling techniques. If the data is left blank, it will affect the model's accuracy performance. Therefore, the missing values are handled using the imputation technique. The imputation type implemented on the dataset is mean.

d. Normalisasi data

The dataset used has a different range of values. For example, age has a value range in the tens, while salary has a value range in the millions. Therefore, it is necessary to normalize the data so that the two fields have the same degree of value. By having the same range of values, the model's accuracy will be higher.

Data Splitting

After normalizing the data, divide the dataset into training data and test data. In this study, the percentage of training data is 80%, while the test data is 20%. The distribution of the dataset is done randomly using the python library, namely scikit-learn.

Create Classifier

There are 6 classifiers: perceptron, logistic regression, k-nearest neighbors, nave Bayes, decision tree, and random forest. After the classifier is created, then the training process is carried out. Then each of these classifiers will produce a model with various levels of accuracy.

Model Evaluation

After getting the model, the next step is to evaluate the model of each classifier. The evaluation process is carried out by comparing the Performance Metrics values; there are precision, recall, f1-score, and accuracy (Sokolova and Lapalme 2009). To calculate the value of performance metrics equations (5) to (8).

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 4, Number 2, July 2022
https://doi.org/10.47709/cnahpc.v4i2.1488

Submitted : May 9, 2022
Accepted : July 10, 2022
Published : July 18, 2022

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (5)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (6)$$

$$\text{F1} - \text{Score} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}} \quad (7)$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \quad (8)$$

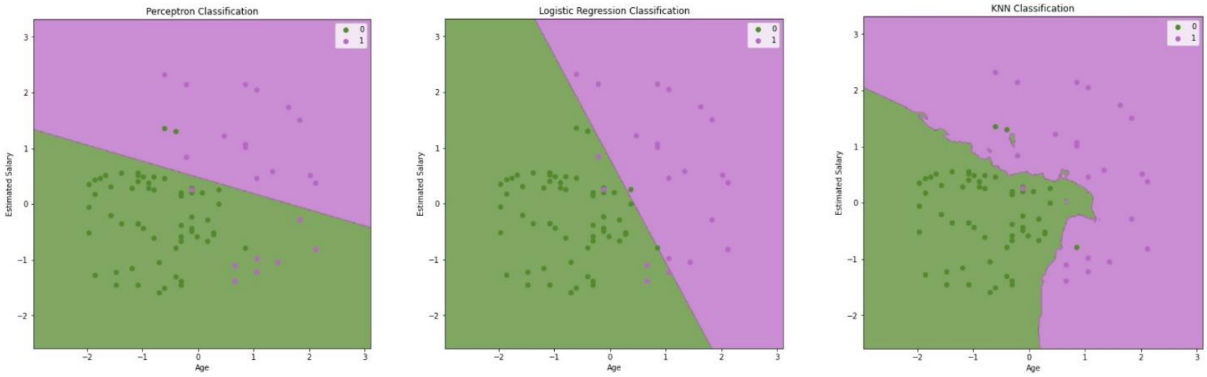
RESULT

The results of the methodology implementation in Fig. 1 is a machine learning model that can predict customer transaction decisions. There are six models with various prediction results. Predicted data is presented in the form of a confusion matrix, as shown in Table 2. True Negative (TN) means that the model correctly predicts customers who do not make transactions. In contrast, Trus Positive (TP) implies that the model correctly predicts customers who make transactions. False Negative (FN) means that the model is wrong in predicting customers who do not make transactions. Meanwhile, False Positive (FP) implies that the model is wrong in predicting customers who make transactions.

Table 2. Confusion Matrix of Each Model

Model	True Negative	True Positive	False Negative	False Positive
Perceptron	56	14	8	2
Logistic Regession	57	17	5	1
K-Nearest Neighbors	55	21	1	3
Naïve Bayes	55	18	4	3
Decision Tree	56	20	2	2
Random Forest	56	21	1	2

The data confusion matrix Table 2 can be visualized into a scatter plot to find out the results of the data distribution. Each classifier has a different visualization pattern. The results of the visualization of the six machine learning models can be shown in Fig. 7.

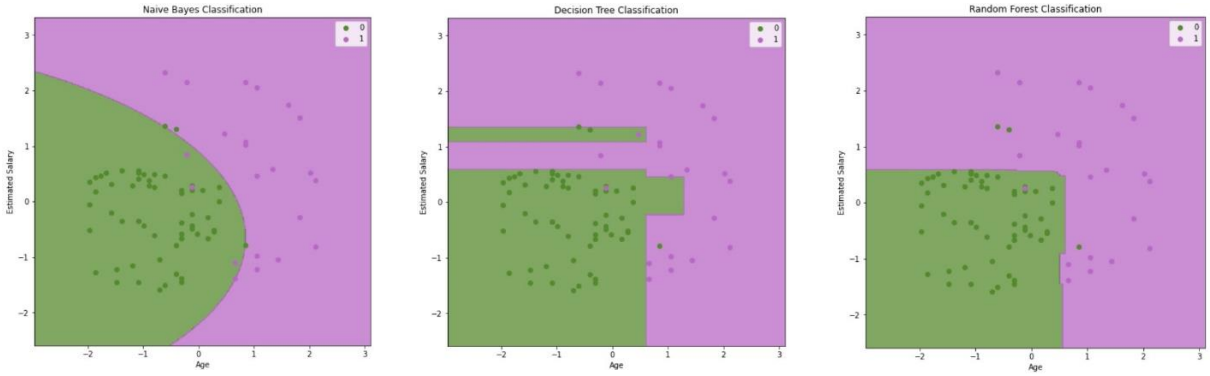


(a) (b) (c)

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 4, Number 2, July 2022
https://doi.org/10.47709/cnahpc.v4i2.1488

Submitted : May 9, 2022
Accepted : July 10, 2022
Published : July 18, 2022



(d) (e) (f)

Fig. 7 Model Visualization (a) Perceptron, (b) Logistic Regression, (c) K-Nearest Neighbors, (d) Naïve Bayes, (e) Decision Tree, (f) Random Forest

The data visualized in Fig. 7 is the test data. Based on the visualization of Fig. 7 there is still data that should be 0 (no transaction), predicted to make a transaction. On the other hand, there is data 1 (performing a transaction), which is expected not to complete a transaction. However, only a few of my data were mispredicted, so the model's accuracy is still high. In addition to accuracy, several other parameters are used to determine the performance of machine learning models. The parameters are Precision, Recall, and F1-Score. The performance metric parameters are calculated through equations (5) to (8). The results of the calculation of performance metrics are shown in Table 3.

Table 3. Performance Metrics of Each Model

Model	Precision	Recall	F1-Score	Accuracy
Perceptron	87,50%	63,63%	73,67%	87,50%
Logistic Regession	**94,44%**	77,27%	84,93%	92,50%
K-Nearest Neighbors	87,50%	**95,45%**	91,27%	95,00%
Naïve Bayes	85,71%	81,81%	83,71%	91,25%
Decision Tree	90,90%	90,90%	90,90%	95,00%
Random Forest	91,30%	**95,45%**	**93,32 %**	**96,25%**

Based on the data presented in Table 3, it is known that the model generated by the Logistic Regression algorithm has the highest Precision value. Meanwhile, the highest recall value is generated from 2 models: K-Nearest Neighbors and Random Forest. Besides having the highest recall value, the Random Forest model also has the highest value in the F1-Score and Accuracy result. Random forest excels in three types of tests, so it can be concluded that the random forest algorithm has the best performance.

The Random Forest algorithm is suitable for large amounts of data. The algorithm that is formed from this decision tree can make accurate decisions. The more trees used, the higher the accuracy of the model. This research uses a large amount of dataset, amounting to 400 data. The random forest algorithm has a customizable root depth and leaf count. However, in this study, the default values for the parameters of root depth and the number of roots were used.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 4, Number 2, July 2022
https://doi.org/10.47709/cnahpc.v4i2.1488

Submitted : May 9, 2022
Accepted : July 10, 2022
Published : July 18, 2022

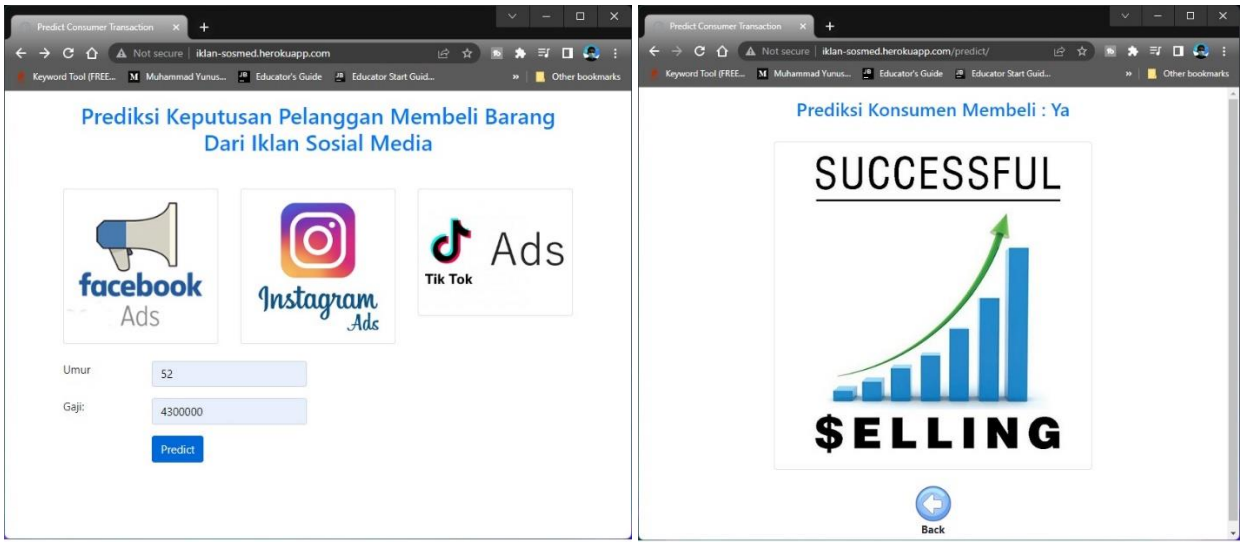


Fig. 8 Deployment Model in Website

Fig. 8 shows the deployment model's result from the random forest algorithm. Deployment results are presented on a web platform. Users can enter age and salary to determine the model's prediction results. The tool used in the deployment process is the flask framework. For the results to be publicly accessible, heroku is used as a cloud platform that provides free web hosting. The final results of the deployment process can be seen on access to the website iklan-sosmed.herokuapp.com.

DISCUSSIONS

In this research, the model generated by the random forest algorithm has the best performance among 5 other machine learning models. However, the other five models also have a high level of accuracy and only have a slight difference from the model generated by the random forest algorithm. Fig. 9 compares the accuracy of the model presented in the bar chart. The random forest has the highest accuracy because this algorithm is a development of a decision tree algorithm that has many tree combinations. The random forest algorithm has Bagging, which is used for random feature selection (Lin, Lin, and Lane 2021). This bagging can improve accuracy when the random feature is used and provide error estimates when combining trees.

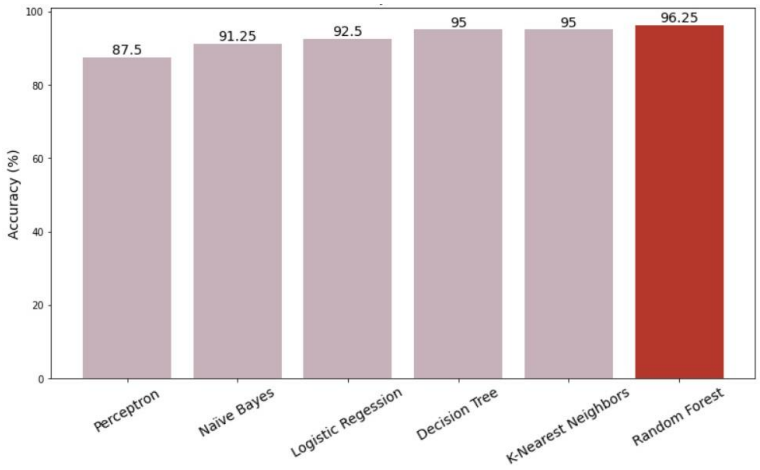


Fig. 9 Accuracy of Each Model

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 4, Number 2, July 2022
https://doi.org/10.47709/cnahpc.v4i2.1488

Submitted : May 9, 2022
Accepted : July 10, 2022
Published : July 18, 2022

Other machine learning models can be further improved in performance if they use appropriate hyperparameter tuning. In this study, each machine learning algorithm's default parameters are used so that each model's performance is standard before getting the best hyperparameter value.

Several studies aim to predict customer transaction decisions, but different case studies. The first research was conducted by (Leonardo, Pratama, and Chrisnatalis 2020), who made sales predictions using telemarketing. The result is the random forest algorithm has the highest accuracy of 90%. Then the second study from (Dhankhad, Mohammed, and Far 2018) predicts credit card fraud using supervised machine learning algorithms. The result is that the model generated by random forest, stacking, and XGB has a high accuracy of 95%. Based on the comparison of the accuracy of the two studies, it is known that our research has a better accuracy value of 96.25%.

CONCLUSION

One of the machine learning implementations in the digital marketing sector is to determine potential customers on social media. Several supervised learning algorithms can predict customer decisions to make transactions. This research compares six supervised learning algorithms for predicting transactions made by customers. The goal is to determine which algorithm has the best performance value to deploy on a web platform. Based on the testing results, the random forest algorithm has the best performance for predicting customer transactions on social media. The model generated by random forest produced the highest Accuracy, Recall, and F1-Score with percentages of 96.35%, 95.45%, and 93.32%.

Meanwhile, the highest precision value was generated by the logistic regression algorithm with a value of 94.44%. Based on the performance matrix data, random forest excels in three measurements, so it can be concluded that the model of the random forest algorithm has the best performance. The results of the random forest model deployment can be accessed at the link iklan-sosmed.herokuapp.com. When implementing the six algorithms, some parameters can be changed. Suggestions for the next determination, please add or change the hyperparameters contained in each algorithm. Do this repeatedly until you get a high accuracy value, which is close to 1 (one). Then visualize the results in a bar chart to find out the amount of data that was mispredicted.

REFERENCES

Berry, M. W., Mohamed, A., & Yap, B. W. (Eds.). (2019). *Supervised and unsupervised learning for data science*. Springer Nature.

Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, *2*(01), 20-28.

Dhankhad, S., Mohammed, E., & Far, B. (2018, July). Supervised machine learning algorithms for credit card fraudulent transaction detection: a comparative study. In *2018 IEEE international conference on information reuse and integration (IRI)* (pp. 122-125). IEEE.

Dokmanic, I., Parhizkar, R., Ranieri, J., & Vetterli, M. (2015). Euclidean distance matrices: essential theory, algorithms, and applications. *IEEE Signal Processing Magazine*, *32*(6), 12-30.

Fitrianah, D., Dwiasnati, S., & Baihaqi, K. A. (2021). Penerapan Metode Machine Learning untuk Prediksi Nasabah Potensial menggunakan Algoritma Klasifikasi Naïve Bayes. *Faktor Exacta*, *14*(2), 92-99.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Hindrayani, K. M., Anjani, A., & Nurlaili, A. L. (2021). Penerapan Machine Learning pada Penjualan Produk UMKM: Studi Literatur. *SENADA*, *1*(01), 19-23.

Tamba, S. P. (2022). PREDIKSI PENYAKIT GAGAL JANTUNG DENGAN MENGGUNAKAN RANDOM FOREST. *Jurnal Sistem Informasi dan Ilmu Komputer Prima (JUSIKOM PRIMA)*, *5*(2), 176-181.

Kudryashov, N. A. (2015). Logistic function as solution of many nonlinear differential equations. *Applied Mathematical Modelling*, *39*(18), 5733-5742.

Kullarni, V. Y., & Sinha, P. K. (2013). Random Forest classifier: a survey and future research directions. *Int. J. Adv. Comput*, *36*(1), 1144-1156.

Leonardo, R., Pratama, J., & Chrisnatalis, C. (2020). Perbandingan Metode Random Forest Dan Naïve Bayes Dalam Prediksi Keberhasilan Klien Telemarketing. *Jurnal Teknologi Dan Ilmu Komputer Prima (Jutikomp)*, *3*(2), 455-459.

Lin, E., Lin, C. H., & Lane, H. Y. (2021). Applying a bagging ensemble machine learning approach to predict functional outcome of schizophrenia with clinical symptoms and cognitive functions. *Scientific Reports*, *11*(1), 1-9.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 4, Number 2, July 2022
https://doi.org/10.47709/cnahpc.v4i2.1488

Submitted : May 9, 2022
Accepted : July 10, 2022
Published : July 18, 2022

Maguire, P., Moser, P., & Maguire, R. (2020). Are people smarter than machines?. *Croatian Journal of Philosophy*, *20*(1 (58)), 103-124.

Paul, A., Mukherjee, D. P., Das, P., Gangopadhyay, A., Chintha, A. R., & Kundu, S. (2018). Improved random forest for classification. *IEEE Transactions on Image Processing*, *27*(8), 4012-4024.

Saiful, A. (2021). Prediksi Harga Rumah Menggunakan Web Scrapping dan Machine Learning Dengan Algoritma Linear Regression. *JATISI (Jurnal Teknik Informatika dan Sistem Informasi)*, *8*(1), 41-50.

Sokolova, M., & Lapalme, G. (2009, September). Classification of opinions with non-affective adverbs and adjectives. In *Proceedings of the International Conference RANLP-2009* (pp. 421-427).

Vembandasamy, K., Sasipriya, R., & Deepa, E. (2015). Heart diseases detection using Naive Bayes algorithm. *International Journal of Innovative Science, Engineering & Technology*, *2*(9), 441-444.

Yudhistiro, K. (2017). Pemanfaatan Neural Network Perceptron pada Pengenalan Pola Karakter. *SMATIKA JURNAL*, *7*(02), 21-25.

* Corresponding author