Future Computer Science Journal (FCSJ)
Vol. 1, No. 1, June 2023: xxx-xxx
http://asasijournal.id/index.php/fcsj
http://doi.org/10.xxxxx/fcsj.xxxx.xxx

Improving the Accuracy of the Logistic Regression Algorithm Model using SelectKBest in Customer Prediction Based on Purchasing Behavior Patterns

Rofik[1*], Nurul Hidayat[2]
[1]Computer Science Department, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Indonesia
[2]Informatics Department, Faculty of Engineering, Universitas Jenderal Soedirman, Indonesia

Abstract
The development of increasingly sophisticated science and technology allows anyone to easily create and run a business. This provides convenience to consumers with a variety of shopping options that are more numerous in this era but also poses challenges in increasingly fierce business competition. Therefore, companies need to develop effective marketing strategies to achieve profitability and sustainable growth. The right marketing strategy should be aimed at meeting customer needs. Several studies have been conducted to classify customers based on their purchasing patterns, but have not applied a fixed combination of features so the accuracy obtained is still not optimal. The purpose of this research is to improve the accuracy of the logistic regression model in predicting customers based on their purchasing behavior patterns with SelectKBest. The proposed new algorithm model is Logistic Regression using feature selection in the form of chi-square scores to improve the combination of the use of features to better fit the characteristics of the predicted model. The first research process is pre-processing, namely performing feature selection with chi-square scores and normalizing data with a standard scaler. The second process is split data, dividing the data into training and testing data. The third process is modeling. Modeling is done with 7 algorithms, namely KNN, Gradient Boosting, Logistic Regression, Decision Tree, Naïve Bayes, SVM, and Random Forest to compare performance. And the fourth is model evaluation. The model is tested using datasets from the UCI Machine Learning repository platform. The evaluation results show that the Logistic Regression algorithm can produce the greatest accuracy of 93.18% with precision, recall, and f1-score of 95% each. This research shows that optimizing the Logistic Regression model with SelectKBest can improve the accuracy of predicting customers based on their purchasing patterns.

Keywords:
Prediction;
Logistic Regression; Comparison;
Customer Behavior, Patterns ;

Article History:
Received: June 22, 2023
Revised: June 22, 2023
Accepted: June 22, 2023
Published: June 22, 2023

Corresponding Author:
Rofik
Computer Science Department,
Universitas Negeri Semarang,
Indonesia
Email:
rofikn4291@students.unnes.ac.id

INTRODUCTION

The development of technology to date has made it easy for everyone to do whatever they want, such as shopping whenever and wherever they are. This of course will also increase the competition between companies. They are competing to make strategies so that in this fierce era their business is always successful and profitable. In the business world, customers are one of the most important assets. The success of a business is directly proportional to how successful it is in dealing with and treating customers. Most startups and large companies invest heavily in building good relationships with their customers [1], [2]. In this digital era, customers have the privilege to be able to choose many choices in terms of products, prices, offers, and others through various online marketplaces. So the soul to handle this problem is that companies need to understand customer purchasing behavior to develop effective marketing strategies and increase product sales.

The application of Machine Learning has been proven to be able to present broader economic opportunities and be able to create and maintain a competitive business advantage [3]. Along with technological advances, techniques have been found that make it easier to understand customer purchasing behavior compared to less effective, traditional methods [4]. Machine Learning has proven effective in analyzing big data in customer segmentation and decision-making [5], one of which is by analyzing data to predict customers through their purchasing behavior. The company's marketing strategy is carried out by classifying customers, because companies can't have different strategies per individual customer, and customers have similarities in purchasing

behavior [6]. Customers who tend to buy a product, tend to buy again in that period. Not only that, but customer experience in a store also affects customer purchasing behavior and shopping decisions at this time and in the future [7], [8]. Thus, by understanding customer purchasing behavior, companies will be greatly helped in making strategic planning and decision-making for their business because it is based on customer data and a deep understanding of customer preferences itself which has an impact on the results obtained more accurately, so it is also in line with business success [9], [10], [11]. Target decisions from customer classification help decide between customers who should be pursued or not wang [12].

The company's task in addition to attracting new customers, must also be able to maintain previous customers who benefit the company a lot because in general, the cost of trying to get new customers is greater than maintaining existing customers [13]. One way to reach loyal customers and benefit the company a lot can be done by studying the customer's purchasing behavior, buying time habits, types of products purchased, and other factors. Because of this understanding, companies can also quickly identify dissatisfaction from their customers when shopping, so that improvements can be made quickly [14].

Unfortunately, previous research conducted by Chaubey et al in 2022 which predicted customer purchasing behavior with several machine learning algorithms such as KNN, Naïve Bayes, XGboost, and others obtained the maximum accuracy obtained was 92.42% [1]. Research in understanding customer behavior was also conducted by Youngjung Suh, in this study, it was found that understanding customer behavior can be used to predict customer churn itself [15]. Related research was also conducted by Srivastava PR et al to analyze the psychology of customer purchases with the highest accuracy of 91% using a neural network algorithm [16]. While research with a systematic review conducted by Byrne A et al is to predict consumer preferences with DFCM with grouping results that have a fairly high accuracy [17]. This author's research uses the Logistic Regression algorithm to classify customers based on their purchasing behavior, based on the dataset available at UCI. Modeling is also done with other algorithms, which are then compared for accuracy. The results of research with this algorithm can be used in making the company's marketing strategy because the algorithm used produces high accuracy.

METHOD

The research method used in this research is divided into several stages, where the stages are carried out sequentially, and other stages are carried out if the previous stages have been carried out. The stages of method carried out are divided into 4 stages, namely Pre-processing, Split Data, Modeling, and Model Evaluation. The stages of the research can be seen in the Figure 1 below.

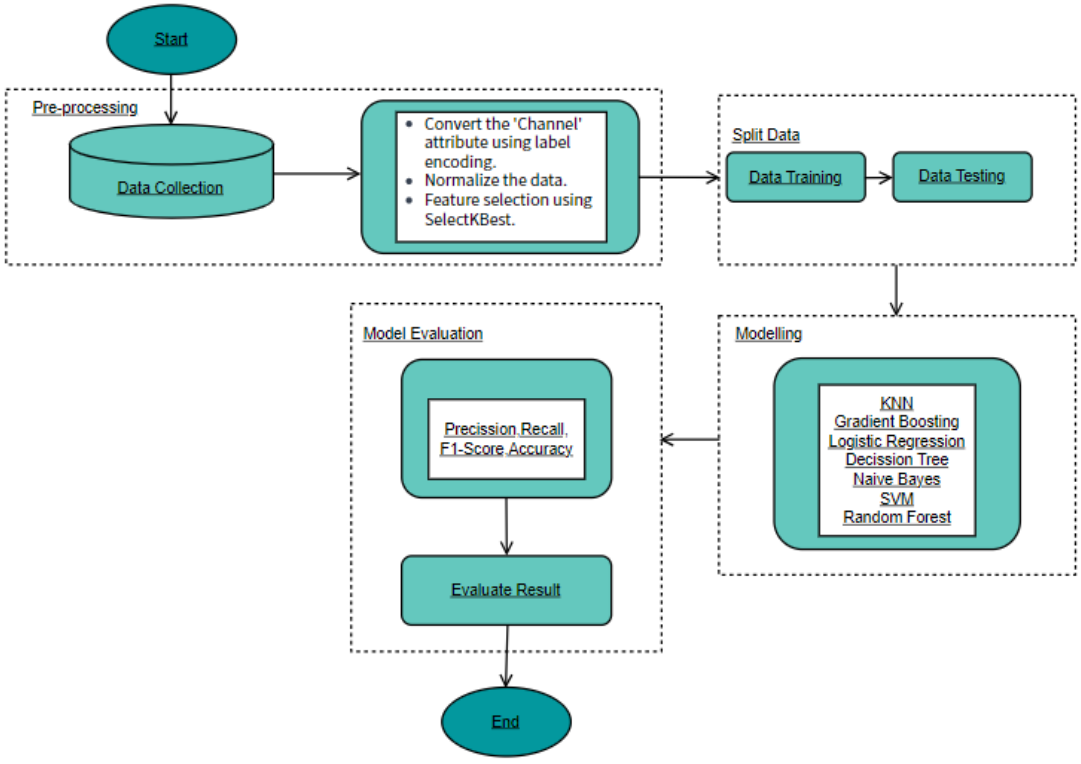

Figure 1. Research Stage

1. Pre-Processing Stage

The initial stage is Pre-processing, which consists of data collection and data cleaning. Data collection is done by taking data that can be accessed publicly at UCI on an account titled Wholesale Customer [1]. The dataset is public and can be downloaded easily through the link: https://archive-beta.ics.uci.edu/datasets?search=Wholesale%20customers . The dataset has 440 records, which consist of 8 attributes, namely Channel, Region, Fresh, Milk, Grocery, Frozen, Detergents_Paper, and Delicatessen. Which is the Channel column which consists of HoReCa and Retail data is used as the target class. The 298 datasets are data from HoReCa customers and 142 from Retail customers.

Then after data collection, the data preprocessing stage is carried out so that the data used is truly accurate in supporting the customer classification process. Cleaning is done in the form of deleting empty columns. The dataset used is quite clean and consistently contains numerical numbers for each column. Attribute conversion is done with label encoding on the 'Channel' column as the class. Data normalization was also performed using StandardScaler to normalize the dataset by changing each feature to the same scale.

Feature selection was performed using SelectKBest with a chi-square score of 2 (chi2). This aims to select the features that are most relevant and have a significant influence on the target variable in the dataset. The chi-square score (chi2) is used as a scoring function to evaluate the relevance of features to the target variable. The get_support method is used to get the best feature index. The selected feature index is used to select the columns corresponding to the dataset. By selecting only relevant features, this method is able to improve model performance and reduce overfitting.

2. Split Data

Split the data into training and training data sets. The data division was done randomly using the train-test split technique. Most of the data is used as model training, then the rest will be used to test the model that has been built. Because customer predictions are based on their purchasing behavior, the size of the data division is also adjusted. The data is divided based on the algorithms that are being applied, in order to achieve the greatest accuracy.

3. Data Modeling

After the data is clean and ready to use, modeling is done using several algorithms including KNN, Gradient Boosting, Logistic Regression, Decision Tree, Naïve Bayes, SVM, and Random Forest. Each algorithm is applied using the same dataset, 7 algorithms are applied because they have different ways of modeling. Thus, it will certainly have an impact on the final result in the form of accuracy generated from each algorithm. Feature selection is also carried out in model building because feature engineering can be done to strengthen algorithm predictions [18]. The excess of features in knowledge exploration is not only able to increase the complexity of the problem but can also result in the knowledge gained being inconsistent because it depends on unnecessary attributes [19]. The modeling uses 6 features namely Channel, Fresh, Milk, Grocery, Frozen, Detergents_Paper, Delicassen, and Channel as the class. Researchers consider these features irrelevant in predicting customers based on their purchasing behavior.The data is modeled to be able to predict customers based on their purchasing behavior into the following algorithm.

3.1 Classification using KNN

The K-Nearest Neighbor (KNN) algorithm is applied to predict customers based on their purchasing behavior. As the name suggests, the concept of KNN is to find the nearest neighbor in the training set and score the candidate category based on the nearest neighbor's class. The highest score is assigned to the class [20]. In this case, this algorithm calculates the distance between the new customer data and the old customer data (training data), based on the calculated distance the KNN algorithm will find the K customers that are most similar to the new customer [21]. This will then determine the label of the new customer, based on the decision of the majority of labels from K's old customers. In the application of this algorithm, the dataset is divided into training data and testing data where the ratio is 80% and 20%.

3.2 Classification using Gradient Boosting

Gradient Boosting is a machine learning algorithm that is also applied in customer classification based on their purchasing behavior. This algorithm is suitable for customer classification because it improves model performance gradually and is able to overcome overfitting problems that may also occur in complex models. The algorithm works by dividing the data into small subsets and building a simple model on the first subset for purchase prediction, then the model is used to predict customer behavior on the next subset. The prediction error of the

second subset will be used to improve the first subset model, and this is done iteratively [22]. This process repeats until all subsets are used. In the application of this algorithm, the dataset is divided into training data and testing data where the ratio is 70% and 30%.

3.3 Classification using Logistic Regression

Logistic models use mathematical equations to map input variables into output probabilities that are in the range of 0 to 1. Regression does not provide an estimate of the dependent variable but rather provides a probability that the dependent variable will fall into a certain category based on the value of the independent variable [23]. This algorithm is used as one of the reasons for its ability to evaluate each input variable separately so that it can provide useful information for companies in making business decisions. In the case of customer prediction, the logistic regression algorithm models the relationship between the features that influence purchasing decisions and the probability of purchasing a customer. In implementing this algorithm, the dataset is divided into data training and data testing where the ratio is 80% and 20%.

3.4 Classification using Decision Tree

The Decision Tree algorithm is used in the application of models. After all, it has wide applications for prediction-related problems because it is easy to use and the results can be interpreted properly [24]. The decision tree model will divide the dataset into smaller subsets based on the most influential variables in predicting customer buying behavior. Each subset is further divided into smaller subsets with increasingly specific rules. In implementing this algorithm, the dataset is divided into data training and data testing where the ratio is 70% and 30%.

3.5 Classification using Naive Bayes

The Naive Bayes algorithm is applied because it is suitable for use in customer classification, this algorithm can overcome the problem of unbalanced data and is easy to classify new customer data. The way this algorithm works is by calculating the probability of each feature in each target class and then calculating the probability of the target class based on all the features that exist. This model will later determine the most likely target class for each customer based on the highest probability value. In implementing this algorithm, the dataset is divided into data training and data testing where the ratio is 80% and 20%.

3.6 Classification using SVM

SVM is applied in customer classification based on their purchasing behavior because of its ability to overcome overlapping problems between classes and is also able to overcome noise in the data. SVM can also be used in multi-class classification and can produce models with high accuracy. The SVM algorithm operates by using an optimal hyperplane that separates the datasets that have been defined as an optimization problem [25]. In implementing this algorithm, the dataset is divided into data training and data testing where the ratio is 80% and 20%. In the case of customer prediction, SVM makes use of the optimal hyperplane to segregate customers based on their buying behavior. SVM looks for a hyperplane that has the maximum distance to data samples from each class and can maximize the separation of each class.

3.7 Classification using Random Forest

Random Forest is applied in customer classification modeling because of its advantages in overcoming overfitting problems and has better performance in predicting customer behavior compared to other algorithms such as Decision Trees. Random Forest works by dividing the data into several subsets that are used to build a decision tree, then each tree in the Random Forest generates customer predictions based on their buying behavior on a subset of data that has never been used before. Predictive accuracy in random forest modeling is by taking the average response from each decision tree prediction [26]. By using independent data subsets to build a decision tree, this algorithm can reduce overfitting that may occur. In implementing this algorithm, the dataset is divided into data training and data testing where the ratio is 80% and 20%.

4. Model evaluation

After modeling each algorithm, performance testing is carried out to test how well the models that have been made in classifying customers. Performance testing is done with a confusion matrix. The confusion matrix is able to present the value in the prediction of the results that have been with this model evaluation, the accuracy, precision, recall, and F1-Score of each algorithm used with the same dataset can be determined. The model evaluation calculation uses accuracy, precision, recall, and f1-score based on the values in:

TP (True Positives) is the number of HoReCa or Retail customers that are correctly classified as Horeca or Retail.

TN (True Negatives) is the number of Horeca or Retail customers that are correctly classified as not Horeca or Retail.
FP (False Positives) is the number of non-Horeca or Retail customers that were incorrectly classified as Horeca or Retail.
FN (False Negatives) is the number of Horeca or Retail customers that were incorrectly classified as not Horeca or Retail.

1. Accuracy

Accuracy is a widely used metric in research as a standard way to measure the ratio between correct and correctly predicted [27]. The working concept of accuracy involves comparing the number of correct predictions (TP and TN) to the total number of data samples evaluated. The following is the formula equation for calculating accuracy.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (1)$$

2. Precision

Precision measures how well the model identifies customers with specific buying behavior patterns. The equation for calculating precision is as follows.

$$Precision = \frac{TP}{(TP + FP)} \quad (2)$$

3. Recall

Recall measures the extent to which the model is able to find all instances of actual Horeca or Retail customers in the dataset. The calculation formula is as follows.

$$Recall = \frac{TP}{(TP + FN)} \quad (3)$$

4. F1-Score

F1-Score is the average of precision and recall which gives an idea of the balance between the two metrics. The calculation formula is as follows.

$$F1 - Score = \frac{2 * (Presisi * Recall)}{(Presisi + Recall)} \quad (4)$$

Visualization of the performance of each algorithm model using the ROC Curve was also carried out. The ROC Curve can provide an in-depth view of the algorithm's performance in the classification of customer detection based on their purchasing behavior. The ROC curve depicts the relationship between the rate of true positives (sensitivity) and the rate of false positives (specificity-1) at various thresholds. Each ROC curve represents the sensitivity of the algorithm in distinguishing between positive and negative classes. The closer the ROC curve is to the upper left corner, the higher the sensitivity of the algorithm. In addition, the comparison of ROC curve positions between algorithms can also give an idea of the relative performance of each algorithm in predicting customer behavior.

RESULTS AND DISCUSSION

From the results of feature selection using SelectKBest with a chi-square score (chi2). It is found that the features support modeling to achieve optimal performance there are 6 best features from 8 features contained in the dataset. These features include Fresh, Milk, Grocery, Frozen, and Detergents_paper. The correlation between the features selected from the SelectkBest results can be seen in the Figure 2 below.

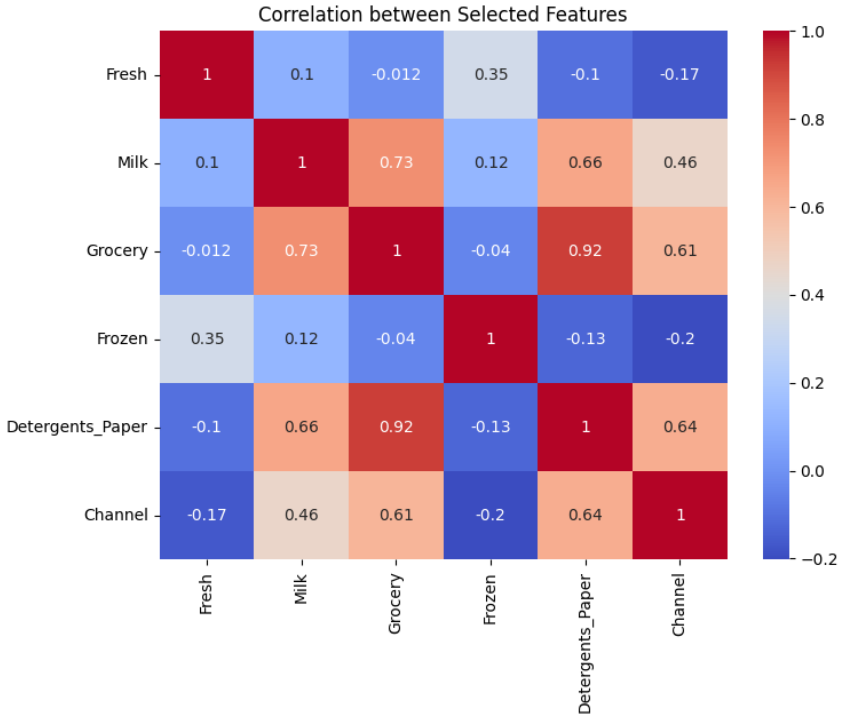

Figure 2. Correlation Between Selected Features

From the modeling of each algorithm used, namely KNN, Gradient Boosting, Logistic Regression, Decision Tree, Naïve Bayes, SVM, and Random Forest, the model performance can be seen in Table 1.

Table 1. Machine Learning Algorithm Test Results

Algorithms	Precision	Recall	F1-Score	Accuracy
KNN	97.00%	91.00%	94.00%	90.90%
Gradient Boosting	83.33%	92.10%	87.50%	92.42%
Logistic Regression	95.00%	95.00%	95.00%	93.18%
Decission Tree	95.00%	91.00%	93.00%	89.77%
Naïve Bayes	94.00%	94.00%	94.00%	90.90%
SVM	97.00%	89.00%	93.00%	89.77%
Random Forest	97.00%	91.00%	94.00%	90.90%

Based on the table above, Logistic Regression shows excellent performance in customer prediction, as shown by evaluation metrics such as precision, recall, F1-Score, and accuracy. It proves a proven accuracy of 93.18%, which is the highest among the accuracies obtained by other algorithms. The Random Forest algorithm also showed good performance in prediction, although the recall was slightly lower than that obtained from the Logistic Regression algorithm. The high precision and balanced F1-Score show that this algorithm also excels in classifying customers.

K-Nearest Neighbors (KNN), Gradient Boosting, Decision Tree, Naïve Bayes, and Support Vector Machine (SVM): These algorithms show variations in performance in customer prediction. KNN, despite achieving high accuracy, has lower precision and F1-Score compared to the other algorithms. Gradient Boosting showed significant improvement in recall and F1-Score, but precision could still be improved. Decision Tree algorithm achieved high recall, but relatively lower precision. Naïve Bayes showed good performance in general, but not as good as Logistic Regression and Random Forest. Meanwhile, the SVM algorithm has lower performance in terms of precision and F1-Score. Therefore, although some of these algorithms have high accuracy and are quite superior in performance, they are less effective in terms of precision and F1-Score.

Overall, Logistic Regression was the best-performing algorithm in terms of precision, recall, F1-Score, and accuracy. Random Forest also showed promising performance. The rest of the algorithms showed different advantages and disadvantages in different evaluation metrics, which highlights the importance of selecting the appropriate algorithm based on specific performance requirements.

ROC Curve for each algorithm can be seen in the image below.

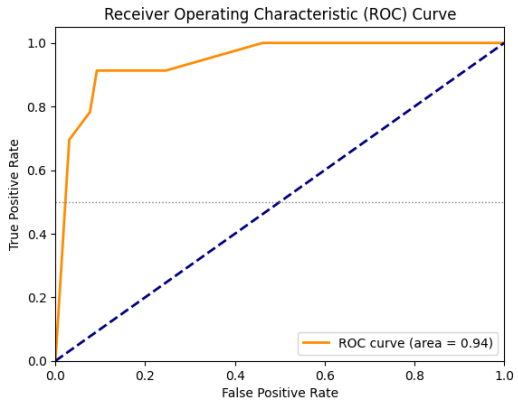

Figure 3. ROC Curve KNN Algorithm

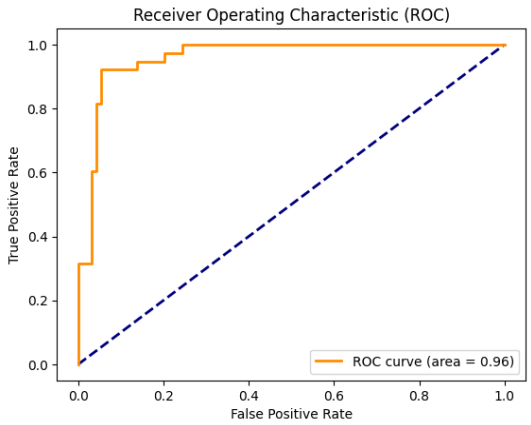

Figure 4. ROC Curve Gradient Boost Algorithm

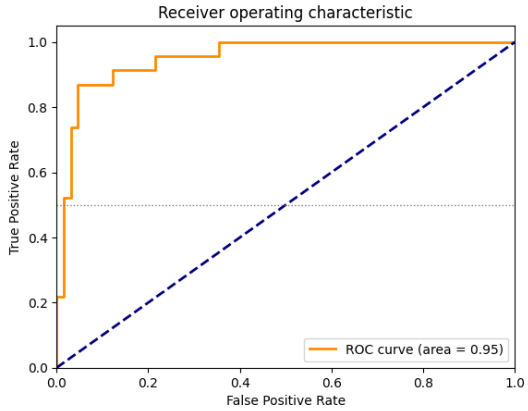

Figure 5. ROC Curve Logistic Regression Algorithm

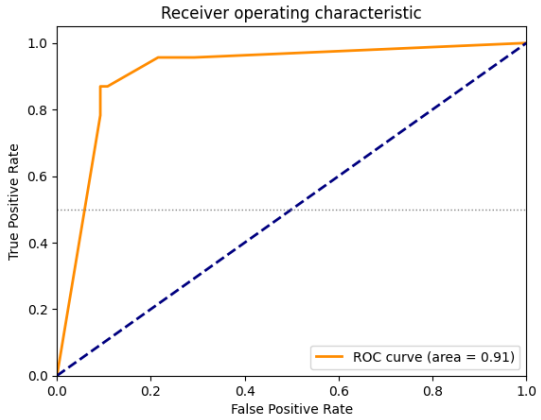

Figure 6. ROC Curve Decision Tree

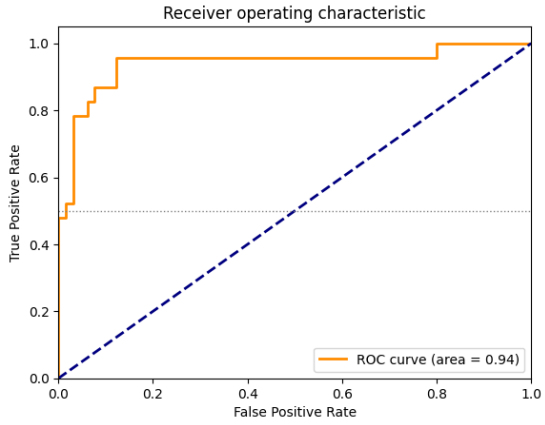

Figure 7. ROC Curve Naive Bayes Algorithm

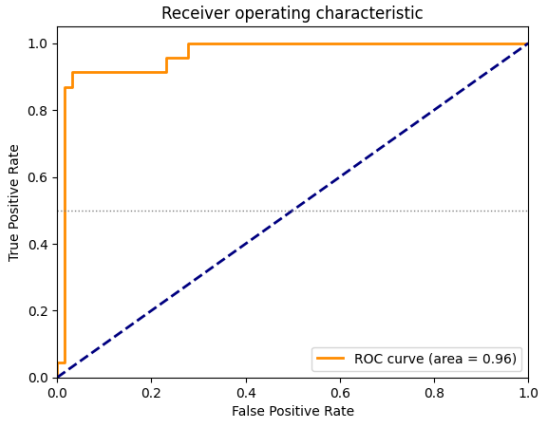

Figure 8. ROC Curve SVM Algorithm

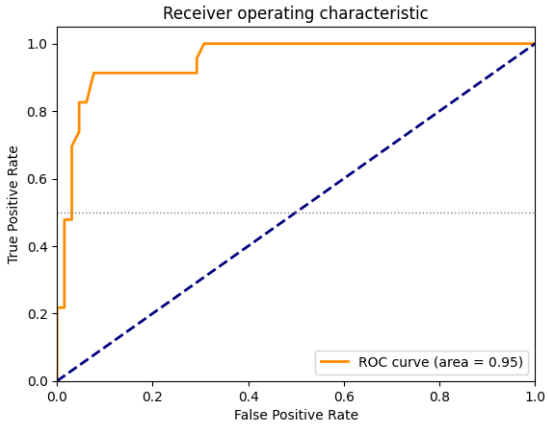

Figure 9. ROC Curve Random Forest Algorithm

Models in algorithms with high ROC Curve values tend to perform better in distinguishing between HoReCa and retail customers. Such is the case with Logistic Regression and Random Forest, which have ROC Curve values of 0.943 and 0.966. However, model accuracy is not always proportional to the ROC Curve value, as seen in the performance of KNN and Gradient Boosting. This shows that although the accuracy is almost the same, Gradient Boosting has a slightly better ability to distinguish classes.

CONCLUSION

In this research, customer classification is carried out based on their purchasing behavior patterns using 7 machine learning algorithms. These algorithms are KNN, Gradient Boosting, Logistic Regression, Decision Tree, Naïve Bayes, SVM, and Random Forest. This research shows the effectiveness of the Logistic Regression algorithm in classifying customers based on their purchasing behavior patterns. Data normalization is done with Satndard Scaller. Feature extraction is performed using SelectKBest with chi-square (chi2) scores to select the better combination of features to maximize accuracy. The results of the evaluation trials showed success in improving the accuracy of the Logistic Regression modeling algorithm after feature selection. The Logistic Regression algorithm model proved to be the best model by obtaining an accuracy of 93.18%, with precision, recall, and f1-score of 95% each.

REFERENCES

[1] G. Chaubey, P. R. Gavhane, D. Bisen, and S. K. Arjaria, "Customer purchasing behavior prediction using machine learning classification techniques," *J. Ambient Intell. Humaniz. Comput.*, no. 0123456789, 2022, doi: 10.1007/s12652-022-03837-6.

[2] C. O. Sakar, S. O. Polat, M. Katircioglu, and Y. Kastro, "Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks," *Neural Comput. Appl.*, vol. 31, no. 10, pp. 6893–6908, 2019, doi: 10.1007/s00521-018-3523-0.

[3] A. De Mauro, A. Sestino, and A. Bacconi, "Machine learning and artificial intelligence use in marketing: a general taxonomy," *Ital. J. Mark.*, vol. 2022, no. 4, pp. 439–457, 2022, doi: 10.1007/s43039-022-00057-w.

[4] L. Fan, "Research on Precision Marketing Strategy of Commercial Consumer Products Based on Big Data Mining of Customer Consumption," *J. Inst. Eng. Ser. C*, vol. 104, no. 1, pp. 163–168, 2023, doi: 10.1007/s40032-022-00908-7.

[5] A. Alghamdi, "A Hybrid Method for Big Data Analysis Using Fuzzy Clustering, Feature Selection and Adaptive Neuro-Fuzzy Inferences System Techniques: Case of Mecca and Medina Hotels in Saudi Arabia," *Arab. J. Sci. Eng.*, vol. 48, no. 2, pp. 1693–1714, 2023, doi: 10.1007/s13369-022-06978-0.

[6] S. Shamshoddin, J. Khader, and S. Gani, "Predicting consumer preferences in electronic market based on IoT and Social Networks using deep learning based collaborative filtering techniques," *Electron. Commer. Res.*, vol. 20, no. 2, pp. 241–258, 2020, doi: 10.1007/s10660-019-09377-0.

[7] L. Zhao, Y. Zuo, and K. Yada, "Sequential classification of customer behavior based on sequence-to-sequence learning with gated-attention neural networks," *Adv. Data Anal. Classif.*, 2022, doi: 10.1007/s11634-022-00517-3.

[8] Z. Gharibshah, X. Zhu, A. Hainline, and M. Conway, "Deep Learning for User Interest and Response

Prediction in Online Display Advertising," *Data Sci. Eng.*, vol. 5, no. 1, pp. 12–26, 2020, doi: 10.1007/s41019-019-00115-y.
[9] A. Alsayat, "Customer decision-making analysis based on big social data using machine learning: a case study of hotels in Mecca," *Neural Comput. Appl.*, vol. 35, no. 6, pp. 4701–4722, 2023, doi: 10.1007/s00521-022-07992-x.
[10] A. Mitra, A. Jain, A. Kishore, and P. Kumar, "A Comparative Study of Demand Forecasting Models for a Multi-Channel Retail Company: A Novel Hybrid Machine Learning Approach," *Oper. Res. Forum*, vol. 3, no. 4, pp. 1–22, 2022, doi: 10.1007/s43069-022-00166-4.
[11] S. xia Chen, X. kang Wang, H. yu Zhang, and J. qiang Wang, "Customer purchase prediction from the perspective of imbalanced data: A machine learning framework based on factorization machine," *Expert Syst. Appl.*, vol. 173, no. January, p. 114756, 2021, doi: 10.1016/j.eswa.2021.114756.
[12] C. Wang, "Efficient customer segmentation in digital marketing using deep learning with swarm intelligence approach," *Inf. Process. Manag.*, vol. 59, no. 6, p. 103085, 2022, doi: 10.1016/j.ipm.2022.103085.
[13] Y. Zhao, Z. Shao, W. Zhao, J. Han, Q. Zheng, and R. Jing, "Combining unsupervised and supervised classification for customer value discovery in the telecom industry: a deep learning approach," *Computing*, 2023, doi: 10.1007/s00607-023-01150-4.
[14] J. Joung and H. Kim, "Interpretable machine learning-based approach for customer segmentation for new product development from online product reviews," *Int. J. Inf. Manage.*, vol. 70, no. February, p. 102641, 2023, doi: 10.1016/j.ijinfomgt.2023.102641.
[15] Y. Suh, *Machine learning based customer churn prediction in home appliance rental business*, vol. 10, no. 1. Springer International Publishing, 2023. doi: 10.1186/s40537-023-00721-8.
[16] P. R. Srivastava, P. Eachempati, R. Panigrahi, A. Behl, and V. Pereira, "Analyzing online consumer purchase psychology through hybrid machine learning," *Ann. Oper. Res.*, 2022, doi: 10.1007/s10479-022-05023-5.
[17] A. Byrne, E. Bonfiglio, C. Rigby, and N. Edelstyn, "A systematic review of the prediction of consumer preference using EEG measures and machine-learning in neuromarketing research," *Brain Informatics*, vol. 9, no. 1, 2022, doi: 10.1186/s40708-022-00175-3.
[18] R. A. de Lima Lemos, T. C. Silva, and B. M. Tabak, "Propension to customer churn in a financial institution: a machine learning approach," *Neural Comput. Appl.*, vol. 34, no. 14, pp. 11751–11768, 2022, doi: 10.1007/s00521-022-07067-x.
[19] D. T. Tran and J. H. Huh, *Building a model to exploit association rules and analyze purchasing behavior based on rough set theory*, vol. 78, no. 8. Springer US, 2022. doi: 10.1007/s11227-021-04275-5.
[20] N. Hidayat, M. F. Al Hakim, and J. Jumanto, "Halal Food Restaurant Classification Based on Restaurant Review in Indonesian Language Using Machine Learning," *Sci. J. Informatics*, vol. 8, no. 2, pp. 314–319, 2021, doi: 10.15294/sji.v8i2.33395.
[21] J. Nagaraju and J. Vijaya, "Boost customer churn prediction in the insurance industry using meta-heuristic models," *Int. J. Inf. Technol.*, vol. 14, no. 5, pp. 2619–2631, 2022, doi: 10.1007/s41870-022-01017-5.
[22] L. Zhou, H. Fujita, H. Ding, and R. Ma, "Credit risk modeling on data with two timestamps in peer-to-peer lending by gradient boosting," *Appl. Soft Comput.*, vol. 110, p. 107672, 2021, doi: 10.1016/j.asoc.2021.107672.
[23] S. Isak-Zatega, A. Lipovac, and V. Lipovac, "Logistic regression based in-service assessment of mobile web browsing service quality acceptability," *Eurasip J. Wirel. Commun. Netw.*, vol. 2020, no. 1, 2020, doi: 10.1186/s13638-020-01708-2.
[24] N. Chaudhuri, G. Gupta, V. Vamsi, and I. Bose, "On the platform but will they buy? Predicting customers' purchase behavior using deep learning," *Decis. Support Syst.*, vol. 149, no. May, p. 113622, 2021, doi: 10.1016/j.dss.2021.113622.
[25] Abdullah-All-Tanvir, I. Ali Khandokar, A. K. M. Muzahidul Islam, S. Islam, and S. Shatabda, "A gradient boosting classifier for purchase intention prediction of online shoppers," *Heliyon*, vol. 9, no. 4, p. e15163, 2023, doi: 10.1016/j.heliyon.2023.e15163.
[26] S. Baghla and G. Gupta, "Performance Evaluation of Various Classification Techniques for Customer Churn Prediction in E-commerce," *Microprocess. Microsyst.*, vol. 94, no. September, p. 104680, 2022, doi: 10.1016/j.micpro.2022.104680.
[27] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, pp. 1–13, 2020, doi: 10.1186/s12864-019-6413-7.