International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

Implementation of Random Forest in Online Shoppers Purchasing Intention Analysis

Farisha Irwayu[1], Lintang Yuniar Banowosari[2], Karmilasari[3]
[1,2,3]Management of Information System, Gunadarma University, Depok, West Java, Indonesia-+62

Abstract**—** *Online Shoppers are currently a trend in recent years which is done by almost everyone. Most online stores have a large number of visitors each day. However, only a small percentage of these visitors will make a purchase. This has an effect on sales data. As a result, a system is required that can analyse visitor behavior in order to determine and classify online shoppers' intentions regarding purchase transactions in order to produce profits for the company. This study approach employs machine learning methods, namely using the Random Forest algorithm. Data normalization, dataset undersampling, one hot encoding attribute, and feature selection are all part of the dataset preprocessing process. After preprocessing, the classification model uses 5723 datasets and produces 13 features as input data. The Random Forest model achieved an accuracy of 87.15% using a dataset distribution of 80% training data and 20% test data. The Random Forest model predicts Online shoppers' purchasing intentions fairly well.*

Keywords**—** *Classification, Machine learning, Online shoppers, Random forest, Revenue.*

I. INTRODUCTION

Nowadays, Online Shopper is currently a trend in recent years which is done by almost everyone, so that potential visitors are identified at the time they are browsing the website [1]. Compared to going to the store in person, customers prefer to shop online. The pandemic that has occurred in Indonesia since 2020, namely Covid-19, has accelerated digital adoption of UMKM (Usaha Micro, Kecil dan Menengah) who are shifting their business from offline to online.

Online shopping which is becoming increasingly popular all over the world makes more and more consumers tend to buy products online. So for e-commerce websites, it is important to study the consumption habits of customers and the factors that influence consumption intentions to run a business and increase sales [2]. A survey from the Katadata Insight Center (KIC) shows that on average, SMEs use two to three marketplaces to sell. There have been many online store platforms that provide many products to meet consumer needs [3]. Most online stores receive countless visitors every day, but only a small percentage of those visitors will make a purchase [4].

Machine learning is one method that can predict whether online shoppers will make a purchase transaction or just visit and browse online stores [5]. This study analyzes the purchase intention data set of online shoppers to find the factors that influence purchase intention and develops a classification model that predicts the purchase intention of online shoppers based on machine learning algorithms. The model in this study uses the Random Forest algorithm. It can be assumed that the prediction problem is classified into two classes, where the aim is to predict the "buy" and "don't buy" labels for visitors to the online store [1].

Understanding the behaviour and intention of online customers has become immensely important for marketing, improving customer's experience which, in return, increases sales [6]. Predicting the purchase intention of online shoppers can be a step for online stores to better understand their customers. Make predictive models possible to infer the factors that influence the buying behavior of customers. Thus predicting buyer intentions can help online stores target customers with the right products, at the right time and thus can take the right steps towards automating marketing decision making [7].

In this study, predictions of Online Shoppers Purchasing Intention were made using Random Forest algorithm model, which then the results of the model were then implemented on a simple website with the Flask framework. It is hoped that this research will be able to classify and predict the intentional behavior of online buyers who will make a purchase or not by using machine learning methods. So that it can be a reference for decision making in making business strategies that can help increase sales to consumers and generate income.

II. METHODOLOGY

Research Methods

The research consists of several stages of the process of making concepts in predicting Online Shoppers Purchasing Intention behavior using the Random Forest algorithm. The series of stages carried out in making the system in this study are described in a general chart which can be seen in Figure 1.

The initial stage of the research started from collecting the Online Shoppers Purchasing Intention (OSPI) dataset taken from the UCI Machine Learning Repository website page. The OSPI dataset is then imported into Google Colab for further processing. The second stage is to do preprocessing on the dataset. Preprocessing is the process of manipulating and preparing the OSPI dataset before being processed into the model. The purpose of preprocessing is to adjust the data to be more compatible with the library that will be used. The preprocessing stage for OSPI data consists of several processes, namely dataset normalization, record elimination, attribute encoding, and attribute selection.

Then in the third stage, namely split dataset. Before conducting the training process, the dataset was resampled by dividing the dataset into two parts, namely training data and testing data. The fourth stage is the process of making a Random Forest model for the classification of Online Shoppers Purchasing Intentions. The design of the model uses

Farisha Irwayu, Lintang Yuniar Banowosari, and Karmilasari, "Implementation of Random Forest in Online Shoppers Purchasing Intention Analysis," *International Research Journal of Advanced Engineering and Science*, Volume 9, Issue 1, pp. 23-29, 2024.

International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

the Random Forest Classifier method. The fifth stage is to do a model evaluation on the Random Forest model. The model that has been made is then evaluated to see the results of the model's performance.

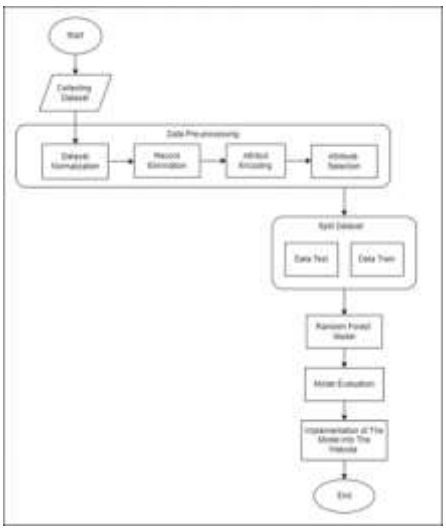

Fig. 1. Research methods scheme

The process of evaluating the model is done by looking at the results of the prediction of the model with the calculation of the confusion matrix which is calculated automatically and visualized. The last stage is implementing the model into the website. After the model is evaluated, then the website design process is carried out so that the model can be implemented on the website. Then the website testing process is carried out when the website creation has been completed. Based on the results of the analysis, conclusions and suggestions are then made.

Dataset Collection

The source data used in this study is the "Online Shoppers Purchasing Intention Dataset" dataset, which is public data provided from the University of California Irvine (UCI) Machine Learning Online Repository website page [8]. The main purpose of this dataset is to predict the purchase intention of website visitors on online stores. This dataset has little missing value and all features of the dataset are relevant to the online visitor's purchase intention based on inference.

The dataset consists of feature vectors belonging to 12,330 sessions. It was formed so that each session would belong to a different user in a 1-year period to avoid any tendency to a specific campaign, special day, user profile, or period. Of the 12,330 sessions in the dataset, 85% (10,422) were negative class samples that did not end finalize a transaction with shopping, and the rest 15% (1908) were positive class samples ending with shopping or purchasing [9].

The dataset has 18 attributes which are divided into two parts, namely the numerical features used in the user actor analysis model and the categorical features used in the user behavior analysis model [9]. The dataset consists of 10 numerical and 8 categorical attributes, with detailed information listed in Table 1 and Table 2 on the Dataset Description. The Revenue attribute will be used as the class label.

TABLE 1. Numerical Features

Feature name	Feature description	Min. Value	Max. Value
Administrative	Number of pages visited by the visitor aboutaccount management	0	27
Administrativeduration	Total amount of time (in seconds) spent bythe visitor on account management related pages	0	3398
Informational	Number of pages visited by the visitor aboutWeb site, communication and address information of the shopping site	0	24
Informationalduration	Total amount of time (in seconds) spent bythe visitor on informational pages	0	2549
Product related	Number of pages visited by visitor aboutproduct related pages	0	705
Product relatedDuration	Total amount of time (in seconds) spent bythe visitor on product related pages	0	63973
Bounce rate	Average bounce rate value of the pagesvisited by the visitor	0	0.2
Exit rate	Average exit rate value of the pages visitedby the visitor	0	0.2
Page value	Average page value of the pages visited bythe visitor	0	361
Special day	Closeness of the site visiting time to a specialday	0	1.0

TABLE 2. Categorical Features

Feature name	Feature description	Number of Categorical Values
OperatingSystem	Operating system of the visitor	8
Browser	Browser of the visitor	13
Region	Geographic region from which the sessionhas been started by the visitor	9
TrafficType	Traffic source by which the visitor has arrived at the Web site (e.g., banner, SMS,direct)	20
VisitorType	Visitor type as ‘‘New Visitor,’’ ‘‘ReturningVisitor,’’ and ‘‘Other’’	3
Weekend	Boolean value indicating whether the date ofthe visit is weekend	2
Month	Month value of the visit date	12
Revenue	Class label indicating whether the visit hasbeen finalized with a transaction	2

Data Preprocessing

Data preprocessing is carried out at an early stage to process and prepare data before the classification process is carried out. Data preprocessing is carried out with the aim of producing more structured data, getting more accurate results, and making data easier to understand when making predictions. The preprocessing of the data in this study

Farisha Irwayu, Lintang Yuniar Banowosari, and Karmilasari, “Implementation of Random Forest in Online Shoppers Purchasing Intention Analysis,” *International Research Journal of Advanced Engineering and Science*, Volume 9, Issue 1, pp. 23-29, 2024.

consisted of several stages, namely dataset normalization, record elimination, attribute encoding and attribute selection.

A. Dataset Normalization

The normalization process is carried out on the dataset so that the data is ready for use. At this stage, checks are carried out to check whether the dataset has missing values and ensure there are no missing values left, because if it is not handled properly it can cause errors in the next process and can affect the algorithm model and cause inaccurate predictions. In this study, the OSPI dataset was not found to have a missing value so there was no need to make adjustments and delete data, so the dataset can be continued to the next preprocessing process.

B. Record Elimination

After the data normalization process, the next step in preprocessing is the elimination of records by performing an undersampling technique to equalize the amount of data and use it to make the distribution of the same data. This is done because the amount of data in the online shoppers purchasing intention dataset is not balanced. It is known that the amount of data from the OSPI dataset is 12,330 where 85% of the data, namely 10,422 data, is a negative class or class that does not make a purchase transaction, and 15% of the data, which is 1908, is a positive class or class that does a purchase transaction [9].

Therefore, in order to avoid bias, an undersampling process is carried out where the amount of data for the negative class is randomly selected and the sample is removed from the majority class. So the number of datasets for positive and negative classes is balanced with the total data being 3815.

C. Attribute Encoding

In the next process, attribute encoding is carried out, namely one hot encoding categorical feature. Since machine learning algorithms assume and require data to be numeric, categorical data must be processed first in order to be accepted. The dataset consists of 10 numeric attributes and 8 categorical attributes. Because all the operations in machine learning-based predictors are mathematical in nature, so they cannot provide input that has categories such as Months: 'New Visitor', 'Returning Visitor' and 'Other Visitor' into the model.

The easiest way to handle this type of data is with label encoding, where each category in a given attribute is encoded with a unique number. However, predictor models can also be biased towards some categories that have been coded with higher numerical values. To avoid this effect, one hot encoding is used for the OSPI dataset. One hot encoding is a process where categorical data is converted into numeric data for use in machine learning.

The category features are converted into binary features called 'one-hot' encoding, meaning that if the feature is represented by that column it will receive a value of 1, otherwise it will receive a value of 0. The features carried out by the one hot encoding process are the Visitor Type and Month features. After encoding, the initial 18 features increased to 28 features. An example of one hot encoding can be seen in Figure 2.

Visitor Type
New Visitor
Returning Visitor
Other Visitor

→

New Visitor	Returning Visitor	Other Visitor
1	0	0
0	0	1
0	1	0

Fig. 2. Example of one hot encoding

Furthermore, the deletion of attributes is also carried out which aims when the attribute selection process is carried out, only attributes that have an important role will be selected and used for the classification process.

D. Attribute Selection

The next preprocessing process is attribute selection. Because the features generated from one hot encoding have 28 input features, it is necessary to select the features that have the greatest influence and have an important role in revenue prediction (revenue), and remove features that do not have much effect on revenue. This step aims to enable faster model training and to avoid model complexity on unnecessary input. This study uses SelectKbest SKlearn.ensemble to find out the best features from the highest score, and choose the best 13 features from 28 features. SelectKbest is a method provided by sklearn to rank features from a dataset based on their "importance" with respect to the target variable.

Split Dataset

The data used to train the algorithm, in order to achieve the desired model. After getting the desired model, data testing is needed. Test data is data used to determine the performance and correctness of the algorithm that has been trained in the model. The process of dividing the dataset is carried out with the train_test_split function from the Sklearn library, with a presentation of 80% for train data and 20% for test data which is also used as validation data.

Random Forest Model

The Schematic of the Random Forest algorithm as shown in Figure 3. First is a random selection of N samples by making bootstrap sampling or replacement from the original dataset, then obtaining N subsamples to construct each decision tree. Then select a feature to construct a decision tree node from a random subset of all features, and create a decision tree. Repeat both Steps and create a random forest with each tree. The random forest algorithm builds a decision tree from each random sample that has been previously created so that it gets a predicted outcome from each decision tree. Then perform calculations from the prediction results for each decision tree that has been made.

After each tree has its own prediction results, then voting is done, namely the selection of the most chosen prediction results. The final prediction is obtained by taking the most predictions from the predictions of all trees in the forest [10]. The classification process uses the random forest algorithm by utilizing a library in python, namely sklearn with imported

Farisha Irwayu, Lintang Yuniar Banowosari, and Karmilasari, "Implementation of Random Forest in Online Shoppers Purchasing Intention Analysis," *International Research Journal of Advanced Engineering and Science*, Volume 9, Issue 1, pp. 23-29, 2024.

International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

Random Forest Classifier to retrieve packages that have been provided by python.

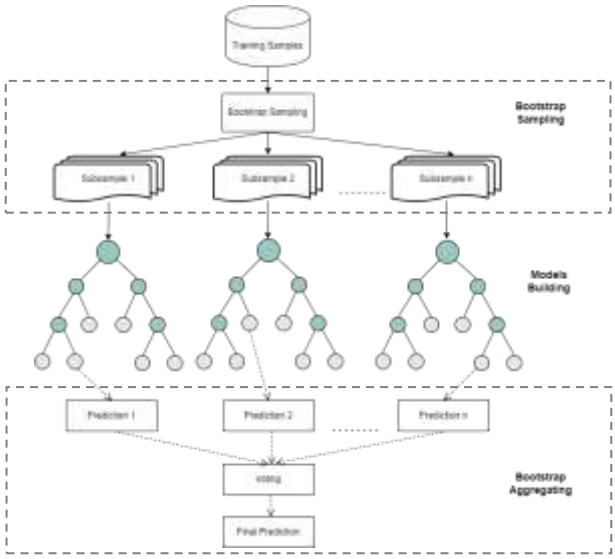

Fig. 3. The scheme of random forest algorithm [10]

Model Evaluation

The classification model is then trained using train data and tested using test data to see the accuracy of the classification model in predicting online shoppers. The training and testing are carried out with different percentages of datasets, namely 70% 30%, 80% 20%, and 90% 10%, so that it can be seen in what percentage the classification model produces the best accuracy value. The model that has been trained must then be saved, but before saving the model and implementing it into the website, an evaluation process is carried out to see how accurate the model has been formed. The evaluation of the Random Forest model was carried out using the calculation of the confusion matrix and the Classification Report with the help of the Sklearn Library [11]. The Classification Report displays the precision, recall, and F1-score values of the model. The higher or closer to 1 the resulting F1-score value, the better the model that has been trained [12].

Confusion matrix is one method that can be used to measure the performance of a classification model. This study uses the confusion matrix parameter to measure the accuracy of the model, so that it can be known by making a confusion matrix table to find out the values of True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN) [13].

- TP (True Positive): Total data with positive true value and positive predictive value
- FP (False Positive): Total data with negative true value and positive predictive value
- TN (True Negative): Total data with negative true value and negative predictive value
- FN (False Negative): Total data with positive true value and negative predictive value

Some of the performance metrics that are common and often used in the form of classification reports are Accuracy, Precision, Recall and F1-score. The classification report can be calculated using the value of the confusion matrix [6].

- Accuracy

  Accuracy is the ratio of correct predictions (positive and negative) to the overall data. Accuracy can be determined using equation (1).

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

- Precision

  Precision is the ratio of true positive predictions to the overall positive predicted results. Precision can be determined using the equation (2).

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

- Recall

  Recall is the ratio of true positive predictions compared to the total number of true positive data. Recall can be determined using the equation (3).

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

- F1-score

  F1-score is the average value of precision and recall. The F1-score can be determined using the equation (4).

$$\text{F1-Score} = 2 \times \{(\text{Recall} \times \text{Precision})/(\text{Recall} + \text{Precision})\} \quad (4)$$

Prediction Result Evaluation

Data visualization is a technique of communicating data or information by describing it into visual objects. The results of the prediction of the Online Shoppers Purchasing Intention model using the Random Forest method will be visualized in the form of a plot diagram. From the results of the visualization, it can be seen a visual description of the prediction results of the model with its actual value and visualization of the importance of the model's feature.

III. RESULT AND DISCUSSION

Dataset Collection Results

Online Shoppers Purchasing Intention (OSPI) data obtained from the UCI Repository. The result of importing the dataset is shows the size of the OSPI dataset which has 12330 records and 18 columns [9]. Furthermore, OSPI dataset creates a new dataframe. The process of creating a dataframe produces a new dataframe called df. The dataframe contains 18 attributes. The results of the df dataframe is attached in figure 4.

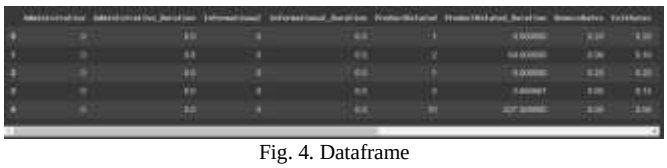

Fig. 4. Dataframe

Pre-Processing Results

Datasets that have been collected and imported still need to be preprocessed with the aim of obtaining more structured data according to classification needs. The data preprocessing process in this study consists of several stages, namely dataset normalization, record elimination, attribute encoding and

Farisha Irwayu, Lintang Yuniar Banowosari, and Karmilasari, "Implementation of Random Forest in Online Shoppers Purchasing Intention Analysis," *International Research Journal of Advanced Engineering and Science*, Volume 9, Issue 1, pp. 23-29, 2024.

International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

attribute selection. Dataset normalization of the dataset shows that the dataset does not have a missing value, the results of checking for missing values show in figure 5.

OSPI record elimination dataset with undersampling method produced 3815 data from the initial number of 12330 data. the results after the elimination of the OSPI dataset records show in figure 6.

From 10422 False data (not making purchases) to 1907 false data, so that the OSPI dataset has the same amount of data and there will be no dominant data. And accuracy can be more accurate. The results of processing False data on the OSPI dataset can be seen by comparing the amount of data before and after which is attached in figure 7.

The result of attribute encoding is one hot encoding categorical features process, by converting categorical features to numerical. The features that are carried out for the one hot encoding process are the Month and VisitorType features. The results of One hot encoding can be seen in figure 8. which shows the number of attributes from 18 attributes to 28 attributes.

```
df.isnull().sum()

Administrative             0
Administrative_Duration    0
Informational              0
Informational_Duration     0
ProductRelated             0
ProductRelated_Duration    0
BounceRates                0
ExitRates                  0
PageValues                 0
SpecialDay                 0
Month                      0
OperatingSystems           0
Browser                    0
Region                     0
TrafficType                0
VisitorType                0
Weekend                    0
Revenue                    0
dtype: int64
```

Fig. 5. Dataset normalization

	Administrative	Administrative_Duration	Informational	Informational_Duration	ProductRelated	ProductRelated_Duration	BounceRates	ExitRates	PageValues	SpecialDay	Month	OperatingSystems	Browser	Region	TrafficType	VisitorType	Weekend	Revenue
0	3	87.833333	0	0.0	27	798.333333	0.000000	0.012644	22.916036	0.8	Feb	2	2	3	1	Returning_Visitor	False	True
1	10	1005.666667	0	0.0	36	2111.341667	0.004348	0.014493	11.439412	0.0	Feb	2	6	1	2	Returning_Visitor	False	True
2	4	61.000000	0	0.0	19	607.000000	0.000000	0.026984	17.535969	1.0	Feb	1	1	7	4	Returning_Visitor	True	True
3	9	111.500000	1	48.5	49	1868.819697	0.000000	0.020709	1.709015	0.0	Mar	2	2	7	2	Returning_Visitor	False	True
4	2	56.000000	1	144.0	67	2563.783333	0.000000	0.006797	19.342650	0.0	Mar	2	2	4	2	New_Visitor	False	True
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3810	0	0.000000	0	0.0	45	2650.416667	0.013333	0.034074	0.000000	0.0	Dec	3	7	2	1	Returning_Visitor	True	False
3811	1	146.000000	0	0.0	11	267.833333	0.018182	0.036364	0.000000	0.0	Mar	1	1	4	2	Returning_Visitor	True	False
3812	2	67.500000	0	0.0	45	1860.000000	0.022727	0.037500	0.000000	0.0	Nov	2	5	1	1	Returning_Visitor	False	False
3813	1	73.000000	1	740.0	50	1434.266128	0.000000	0.019231	22.032769	0.0	Mar	2	2	3	2	Returning_Visitor	True	False
3814	0	0.000000	0	0.0	6	74.000000	0.000000	0.028000	0.000000	0.0	Mar	2	4	6	8	Returning_Visitor	False	False

3815 rows × 18 columns

Fig. 6. Record elimination

Revenue	Count	Proportion
False	10422	0.845255
True	1908	0.154745

Before

Revenue	Count	Proportion
True	1908	0.500131
False	1907	0.499869

After

Fig. 7. Comparison of the number of datasets after record elimination

```
X.columns

Index(['Administrative', 'Administrative_Duration', 'Informational',
       'Informational_Duration', 'ProductRelated', 'ProductRelated_Duration',
       'BounceRates', 'ExitRates', 'PageValues', 'SpecialDay',
       'OperatingSystems', 'Browser', 'Region', 'TrafficType', 'Weekend',
       'Month_Aug', 'Month_Dec', 'Month_Feb', 'Month_Jul', 'Month_June',
       'Month_Mar', 'Month_May', 'Month_Nov', 'Month_Oct', 'Month_Sep',
       'VisitorType_New_Visitor', 'VisitorType_Other',
       'VisitorType_Returning_Visitor'],
      dtype='object')
```

Fig. 8. Comparison of the number of datasets after record elimination

On attribute elimination, namely the removal of columns in order to get the best features based on needs. Attributes that are removed are VisitorType_Other and Month_Aug, because at the attribute selection stage the requiredfeatures are the VisitorType_Returning_Visitor, VisitorType_New_Visitor and Page Values features which have a better effect on the model based on the amount of data. The result of the elimination of attributes is the column in the OSPI dataset to 26 columns which show in Figure 9.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3815 entries, 0 to 3814
Data columns (total 26 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Administrative             3815 non-null   int64
 1   Administrative_Duration    3815 non-null   float64
 2   Informational              3815 non-null   int64
 3   Informational_Duration     3815 non-null   float64
 4   ProductRelated             3815 non-null   int64
 5   ProductRelated_Duration    3815 non-null   float64
 6   BounceRates                3815 non-null   float64
 7   ExitRates                  3815 non-null   float64
 8   PageValues                 3815 non-null   float64
 9   SpecialDay                 3815 non-null   float64
 10  OperatingSystems           3815 non-null   int64
 11  Browser                    3815 non-null   int64
 12  Region                     3815 non-null   int64
 13  TrafficType                3815 non-null   int64
 14  Weekend                    3815 non-null   int64
 15  Month_Dec                  3815 non-null   uint8
 16  Month_Feb                  3815 non-null   uint8
 17  Month_Jul                  3815 non-null   uint8
 18  Month_June                 3815 non-null   uint8
 19  Month_Mar                  3815 non-null   uint8
 20  Month_May                  3815 non-null   uint8
 21  Month_Nov                  3815 non-null   uint8
 22  Month_Oct                  3815 non-null   uint8
 23  Month_Sep                  3815 non-null   uint8
 24  VisitorType_New_Visitor    3815 non-null   uint8
 25  VisitorType_Returning_Visitor  3815 non-null  uint8
dtypes: float64(7), int64(6), uint8(13)
memory usage: 468.2 KB
```

Fig. 9. Attribute elimination

Attribute selection is performed on the dataset before doing the dataset split process. the features of the one hot encoding process produce 26 input features, so it is necessary to select the best features that have the highest score to select the features to be used in the classification model. Feature

Farisha Irwayu, Lintang Yuniar Banowosari, and Karmilasari, “Implementation of Random Forest in Online Shoppers Purchasing Intention Analysis,” *International Research Journal of Advanced Engineering and Science*, Volume 9, Issue 1, pp. 23-29, 2024.

selection using the SelectKBest function from the sklearn library. There are 11 features with the highest score. From the results of the best feature selection, two more features were added, namely Bouncerates and exitrates. And the final result of the best feature selection is the best feature that becomes the model parameter, namely there are 13 features which show in Figure 10.

```
['Administrative',
 'Administrative_Duration',
 'Informational',
 'Informational_Duration',
 'ProductRelated',
 'ProductRelated_Duration',
 'BounceRates',
 'ExitRates',
 'PageValues',
 'SpecialDay',
 'Month_May',
 'Month_Nov',
 'VisitorType_New_Visitor']
```

Fig. 10. Best feature selection

Split Dataset Results

The results of the distribution of the dataset with a percentage of 80% 20% produces 3052 train data and 763 test data which show in figure 11. The dataset that has been distributed will then be processed to a classification model making.

```
print(y_train.count())
print(y_test.count())

3052
763
```

Fig. 11. Split dataset

Result of Random Forest Model Evaluation

The distribution of the percentage of 80% train data and 20% test data has highest accuracy results in the random forest model. The models identified early purchase intention with good performance in terms of Area Under Curve (AUC) score [2]; [12]. The result of the AUC score generated from the model is 93%. Accuracy results of the random forest model are shown in figure 12.

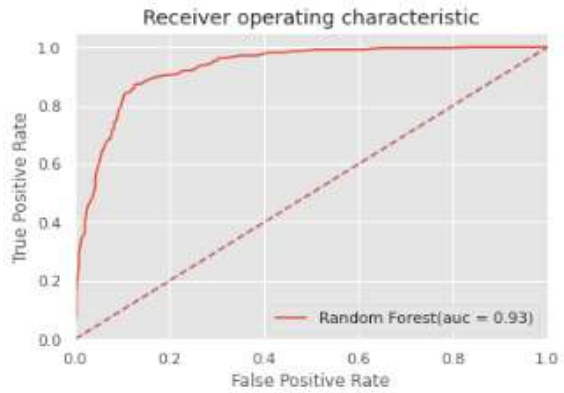

Fig. 12. AUC Score

The resulted in a train accuracy value of 99.93% and a test accuracy of 87.15% that shown in figure 13.

```
print('Train score: ' + str(model_rf.score(X_train, y_train)))
print('Test score: ' + str(model_rf.score(X_test, y_test)))

Train score: 0.9993446920052425
Test score: 0.8715596330275229
```

Fig. 13. The results of percentage of train and test

The validation of the classification model is carried out using the calculation of the confusion matrix. The results of the best proportion presentation on the split dataset are 80% train data and 20% test data, so the validation process is only carried out on that percentage. The results of the Classification Model Evaluation can be seen based on the calculation of the confusion matrix as indicated by the calculation of the Accuracy model, Recall, Precision, and F1-score values which are attached in figure 14 and the table of confusion matrix is show in figure 15.

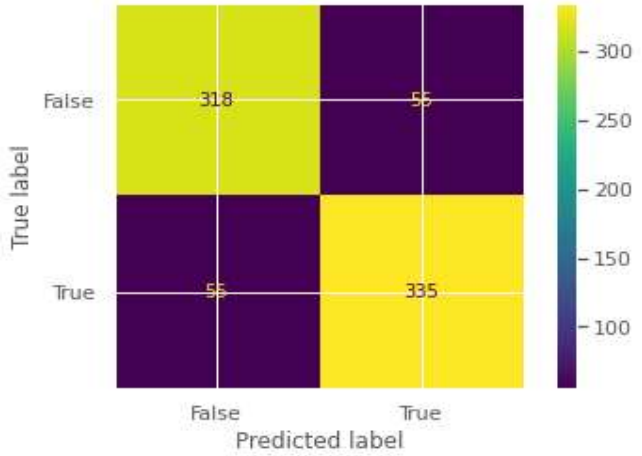

Fig. 14. Confusion matrix

```
Accuracy: 0.8715596330275229
Recall: 0.8743589743589744
Precision: 0.8743589743589744
F1-Score: 0.8743589743589744
```

Fig. 15. Classification report

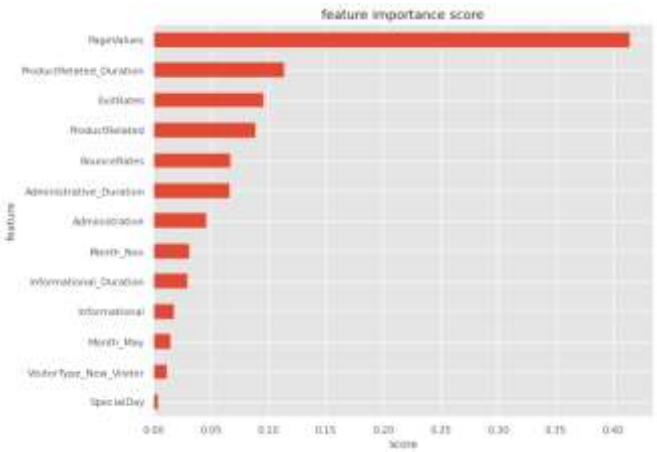

Fig. 16. Feature importance score

Farisha Irwayu, Lintang Yuniar Banowosari, and Karmilasari, "Implementation of Random Forest in Online Shoppers Purchasing Intention Analysis," *International Research Journal of Advanced Engineering and Science*, Volume 9, Issue 1, pp. 23-29, 2024.

The feature importance score shows what features are the most influential in the Online Shoppers Purchasing Intention classification process. The results of the Feature Importance Visualization of the Random Forest model are in the form of a graph that can be seen in Figure 16.

IV. CONCLUSION

This study can generate predictions of online shoppers' purchase intentions by applying machine learning algorithms, namely the Random Forest method. The study produces a classification model of online shoppers purchasing intention using 5723 datasets after preprocessing. The resulting classification model uses 13 features as input data. The research succeeded in producing an accuracy value of the random forest model, which is 87.15% with the proportion of distribution of datasets being 80% (3052) training data and 20% (763) test data.

This study also produced a recall value of 87.43%, a precision value of 87.43% and an f1-score value of 87.43%. It can be seen that the Random Forest model provides a fairly good level of accuracy because the average classification report value tends to be close to 1 or 100%. Based on trials, the model implemented into a website-based application with the Flask framework has been able to predict the interest of customer visits to websites that lead to purchase transactions or not.

In the future work, further development to perfect the application for predicting online shoppers purchase intention is by collecting data with more samples and features, as well as a balanced proportion of data, therefore the classification model can obtain more accurate prediction results and the accuracy value will be higher. Other than that, multiple random forests might useful to improve the accuracy, and the best way is to split the features by considering their technical meaning with regards to the field.

ACKNOWLEDGMENT

We would like to express my appreciation to all those who have supported me during our research and study. I would like to express my deep gratitude to Dr. Lintang Yuniar Banowosari and Dr. Karmilasari for they patient guidance, advice and assistance in keeping my progress on schedule We would also like to extend our thanks to all the previous person and team who published their research online. Without their proceeding and journal, this study wouldn't be done. Finally, we wish to thank our parents and peers for their support and encouragement throughout the study.

REFERENCES

[1] K. Baati and M. Mohsil, "Real-time prediction of online shoppers' purchasing intention using random forest," in *IFIP Advances in Information and Communication Technology*, 2020, vol. 583 IFIP, pp. 43–51, doi: 10.1007/978-3-030-49161-1_4.

[2] M. S. Satu and S. F. Islam, "Modeling Online Customer Purchase Intention Behavior Applying Different Feature Engineering and Classification Techniques," 2023, p. 13, doi: https://doi.org/10.21203/rs.3.rs-3185752/v1.

[3] I. Kurniawan, Abdussomad, M. F. Akbar, D. F. Saepudin, M. S. Azis, and M. Tabrani, "Improving the Effectiveness of Classification Using the Data Level Approach and Feature Selection Techniques in Online Shoppers Purchasing Intention Prediction," *J. Phys. Conf. Ser.*, vol. 1641, no. 1, pp. 1–7, 2020, doi: 10.1088/1742-6596/1641/1/012083.

[4] M. T. Teye and Y. M. Missah, "Investigating Factors that Affect Purchase Intention of Visitors of E-commerce Websites Using a High Scoring Random Forest Algorithm," *Int. J. Eng. Res. Technol.*, vol. 13, no. 12, pp. 5105–5112, 2020, [Online]. Available: http://www.irphouse.com.

[5] X. Shi, "The Application of Machine Learning in Online Purchasing Intention Prediction," in *International Conference on Big Data and Computing*, 2021, pp. 21–29, doi: 10.1145/3469968.3469972.

[6] M. R. Kabir, F. Bin Ashraf, and R. Ajwad, "Analysis of different predicting model for online shoppers' purchase intention from empirical data," in *22nd International Conference on Computer and Information Technology, ICCIT 2019*, 2019, pp. 1–6, doi: 10.1109/ICCIT48885.2019.9038521.

[7] S. Kumar and Chandrakala, "A Survey on Customer Churn Prediction using Machine Learning Techniques," *Int. J. Comput. Appl.*, vol. 154, no. 10, pp. 13–16, 2016, doi: 10.1109/CSDE53843.2021.9718460.

[8] C. Sakar and Y. Kastro, "Online Shoppers Purchasing Intention Dataset," 2018. http://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset.

[9] 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.

[10] H. Wang, M. Lei, Y. Chen, M. Li, and L. Zou, "Intelligent identification of maceral components of coal based on image segmentation and classification," *Appl. Sci.*, vol. 9, no. 16, pp. 2–15, 2019, doi: 10.3390/app9163245.

[11] C. I. Agustyaningrum, W. Gata, R. Nurfalah, U. Radiyah, and M. Maulidah, "Komparasi Algoritma Naive Bayes, Random Forest Dan Svm Untuk Memprediksi Niat Pembelanja Online," *J. Inform.*, vol. 20, no. 2, pp. 164–173, 2020, doi: 10.30873/ji.v20i2.2402.

[12] C. I. Agustyaningrum, M. Haris, R. Aryanti, and T. Misriati, "Online Shopper Intention Analysis Using Conventional Machine Learning And Deep Neural Network Classification Algorithm," *J. Penelit. Pos dan Inform.*, vol. 11, no. 1, pp. 89–100, 2021, doi: 10.17933/jppi.v11i1.341.

[13] A. Purnama, A. Maulana Yusup, A. Wibowo, and D. Susilawati, "Uji Algoritma Random Forest Pada Dataset Online Shoppers Purchasing Intention," *IKRA-ITH Inform. J. Komput. dan Inform.*, vol. 5, no. 1, pp. 101–108, 2021, [Online]. Available: https://journals.upi-yai.ac.id/index.php/ikraith-informatika/article/view/920.