Online Shoppers Purchasing Intention Dataset

Cansu AYTEN – 171180010 – cansuayten1@gmail.com

Data set Link: https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset

1. Abstract

a. Summary of the Project

The main purpose of the project, which was developed using the Online Shoppers Purchasing Intention dataset, is to predict the purchase intention of the users who will shop online with the help of the data we have. While doing this, some machine learning algorithms were used. The data until the point of applying the algorithm was visualized and analyzed. After the preprocessing was completed, the models were trained and then compared according to some evaluation criteria.

b. Used Methods

Different machine learning algorithms were used to predict customers' purchase intention. Among these algorithms are KNN (K-Nearest Neighbors), Support Vector Machine, Random Forest Classifier and Decision Tree Classifier algorithms. It has been tried to reach the best results by using different metrics in these four algorithms.

c. Findings and Conclusions

After the implementation of the algorithms, the trained models were evaluated. While evaluating, some metrics were used. These are Accuracy, Precision, F1-Score and Recall metrics. Using these metrics, the models are ranked relative to each other. Although they are very close to each other in terms of accuracy, the model with the highest value in general is the model in which the Random Forest Classifier algorithm is used. In the same way, the Random Forest Classifier algorithm has the highest overall score in the F1 – Score metric. In Precision and Recall metrics, the Support Vector Machine algorithm is mostly in the top rank**.**

2. Introduction

A. Explain the Problem

As the internet occupies a great place in people's lives as a result of the extraordinary rapid development of technology, e-commerce has started to grow day by day. Because nowadays, people have started to prefer online shopping instead of traditional shopping methods. For example, AliExpress, an online shopping service, had 528 million visitors in January 2021. The traffic type of these visitors varies as mobile or desktop, and the average visit time is six minutes [1]. Online shopping has saved people's time and made their lives easier. During the pandemic, which is one of the fastest growing times in online shopping, people have met their needs by using online shopping platforms instead of leaving their homes. For example, Amazon, one of the largest e-commerce companies, reached the highest revenue in 2020, partly due to the coronavirus. While it had a revenue of 280.5 billion dollars in 2019, this amount increased to 386 billion dollars in 2020 [2]. With the popularization of online shopping, the curiosity of what customers will or should not buy has arisen in sellers. It can be said that the reason is to increase customer satisfaction and therefore increase sales. In this project, it has been tried to predict a customer's purchase or non-purchase action by using data on online shopping. The Online Shoppers Purchasing Intention dataset has 12,330 data. Each session in this data belongs to a different user. Data were collected over a one-year period. 10,422 of the data show that the shopping result is negative and 1908 is positive. 84.5% of the data show that users do not complete the transaction with a purchase, and 15.5% show that users complete their transaction with a purchase [3].

The data set consists of 18 features, 10 numerical and 8 categorical. Here the 10 numerical attributes are Administrative, Administrative_Duration, Informational, Informational _Duration, ProductRelated, Product_RelatedDuration, BounceRate, ExitRate, PageValus, and Special Day. Eight categorical attributes are OperatingSystems, Browser, Region, TrafficType, VisitorType, Weekend, Month and Revenue. Revenue is used as the class label. Revenue is a boolean that indicates whether a user has made a purchase [3].

B. Outline Logic of the Paper

In this report, the purchase intention of online shoppers is analyzed using the Online Shoppers Purchasing Intention dataset. This can be achieved by using and analyzing historical data of customers. The most basic purpose is to make a prediction whether a customer will buy or not. In the project, firstly, the available data was discovered. There are no missing values in the data. Afterwards, visualizations were made with the data. After these visualizations, some preprocessing was applied to the data. Some categorical attributes have undergone appropriate processing and have been rendered as 0-1. After the data were separated as train

and test, performances were measured by giving them to the algorithms. Here, KNN (K-Nearest Neighbors), Support Vector Machine, Random Forest Classifier and Decision Tree Classifier classification algorithms are used and the results are reported.

C. Provide General Conclusions

Visualizations were made for each attribute in the project. The relationship of each attribute to Revenue is examined.

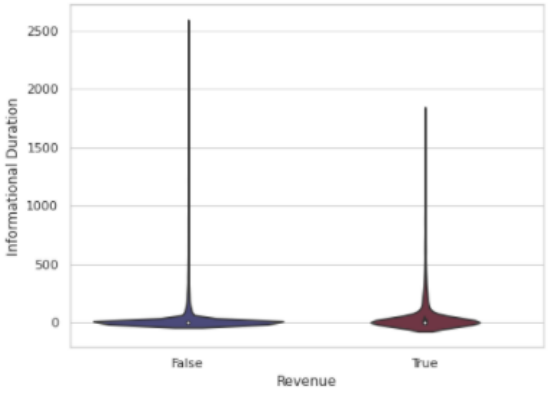

Figure 1: Relationship between informational duration and revenue

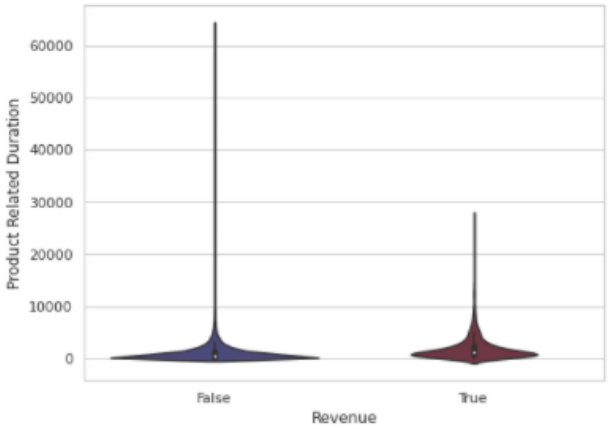

Figure 2:Relationship between product related duration and revenue

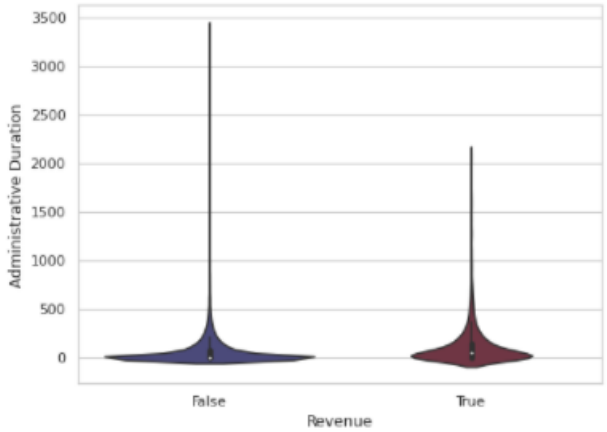

Figure 3:Relationship between administrative duration and revenue

When Figure 1, Figure 2 and Figure 3 are examined, it is seen that customers spend less time on the pages in general. Customers spend more time on pages containing product and information if they are going to shop. Customers who do not shop spend less time.

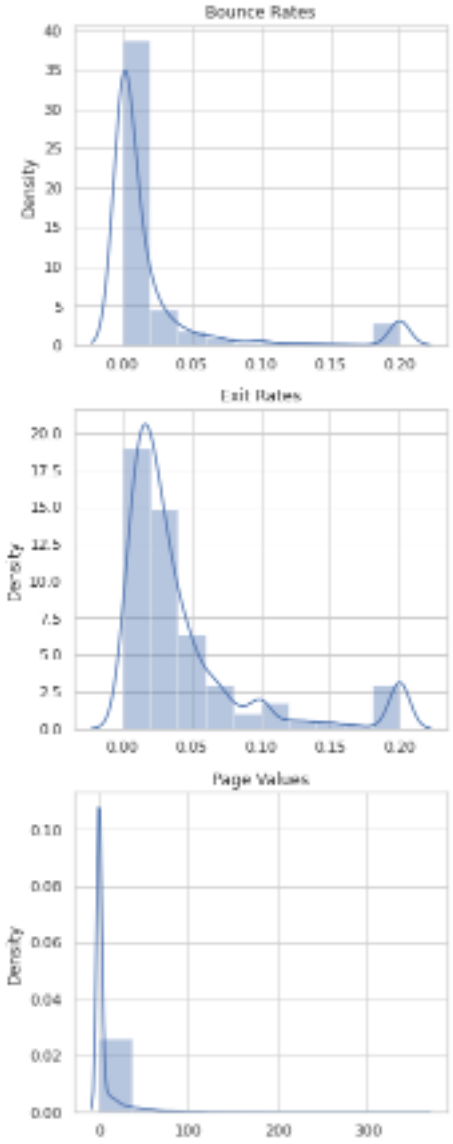

Figure 4:Charts of Bounce Rate, Exit Rate and Page Value

When these graphs are examined, it shows that BounceRate and ExitRate are more common at low values, therefore customers interact with the page more.

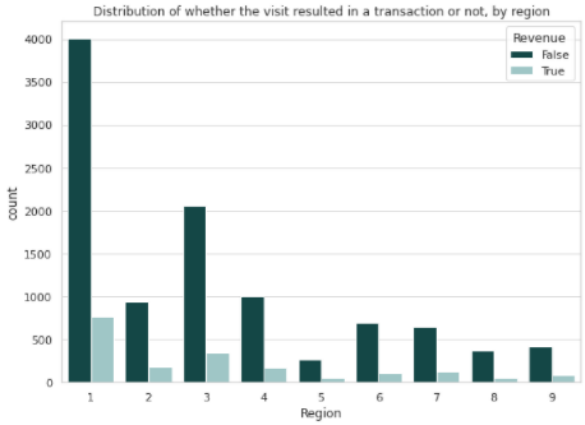

Figure 5:Bar chart showing the Revenue - Region distribution

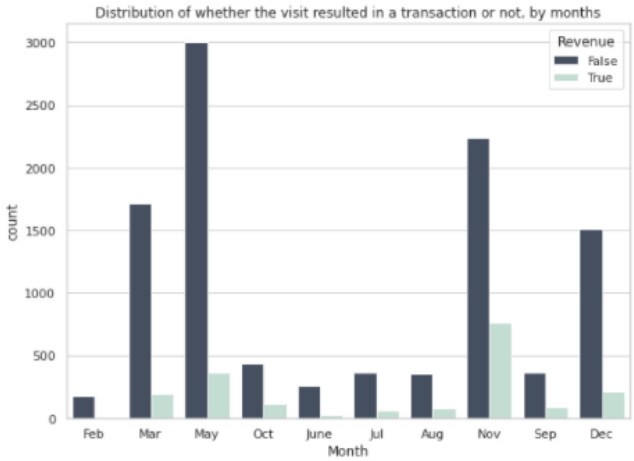

Figure 6: Bar chart showing the Revenue - Month distribution

Figure 7: Bar chart showing the Revenue - Weekend distribution

Looking at some of the graphs based on income (Figure 5-6-7), more purchases are made on weekdays than on weekends. Likewise, the act of not buying takes place more often. In other words, transactions made on weekdays are generally more. When the months are examined, the number of non-purchasing events in May is higher than the others. November is the month with the most shopping, as shopping increases towards the end of the year. When the regions are examined, Region 1 is the region with the most transactions compared to the others.

After these, appropriate pre-processes were carried out in order to express the categorical data numerically. For example, the boolean data Revenue and Weekend are converted to numeric values as 0-1. LabelEncoding technique was used for this process. Month and Visitor Types are similarly separated from each other and displayed as 0-1. Here, the get_dummies() function is used. Then the data was scaled. After that, the data is divided into 70 by 30 as training and test data. After these processes applied to the data, the classification process was started. Four types of algorithms were used. The first of these is the KNN (K-Nearest Neighbors) algorithm. In this algorithm, manhattan, minkowski and euclidean values are used for distance and 1,3,5 as the number of neighbors. KNN makes predictions according to which class the neighbors of the value to be estimated are in the most. The second is the Support Vector Machine algorithm. In this algorithm, rbf, poly, sigmoid and linear values are given to the kernel hyper parameter. The third is Random Forest Classifier. Finally, the Decision Tree Algorithm was used. In both algorithms, the criterion took two different values -gini and entropy- and the results were reported separately accordingly. These reported results are Accuracy, Precision, Recall and F1 Score. In general, the algorithm with the highest accuracy and F1 Score was Random Forest Classifier.

3. Purpose For This Paper

Nowadays, people have mostly started to prefer online shopping. The reason for this can be explained as preventing waste of time by finding what they are looking for in a short time, online shopping offers more options, everything can be handled without going to a store, online shopping is often more discounted than stores, and the comments of other users who buy a product can be read. As a result of this increase in online shopping, retailers want to know what customers will or will not buy. This enables the collection and accumulation of data from many companies, and using this information to find things like what people would prefer or not. Large companies are trying to maximize customer satisfaction by using this experience. The aim of this study is to predict the purchase intention of users. In this paper, information about the dataset used in this study is given, the methods used are explained, the outputs are evaluated and past studies are examined.

4. Literature Review

In the study named “Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks”, a user analysis system was created using Online Shoppers Purchasing Intention Dataset. In this system, both the user's intention to buy or not and the user's intention to leave the site are estimated. This system consists of two modules. In the first module, the most distinguishing features for the forecasting process were found using filter feature selection techniques, and then the

purchase intention of the users was estimated using machine learning algorithms. Support Vector Machines (SVM), Multilayer Perceptron (MLP), and Decision Tree algorithms were used as machine learning algorithms. The K-Nearest Neighbor algorithm was not preferred in this study because it is thought to be not suitable for real-time estimation. While evaluating the success of the algorithms used, Accuracy, F1 Score, and True-Positive/True-Negative Rate metrics were used. Training and test data were randomly selected 100 times and new results were obtained. A t-test was applied to see how different the algorithm accuracies were. It is stated that the highest accuracy belongs to Decision Tree. But in the data set, the negative class label is more than the positive class label. This situation creates class imbalance and affects the success rate. Therefore, the F1 Score is used instead of the accuracy score. Because it is stated that the F1 Score is a metric that takes into account the unbalanced class situation. It was said that the highest F1 score with 0.58 points belonged to MLP, and the lowest F1 score with 0.52 belonged to SVM (linear kernel). Some techniques have been applied to improve the outputs, namely the success rates. These techniques are oversampling and feature selection preprocessing techniques. It has been mentioned that with filter-based feature selection techniques, algorithms can perform better with fewer features. In this study, correlation, mutual information (MI), and Minimum Redundancy Maximum Relevance (mRMR) filter feature selection techniques were used. In the results obtained after oversampling -it was stated that it was applied only to the training data- MLP again gave the highest accuracy score since class imbalance was prevented. In addition, the F1 score increased to 0.86 and thus the selected algorithm was the MLP algorithm. Then, feature selection techniques were applied. Six features were selected with the MRMR technique and the chosen technique was the MRMR technique. The reason for this is that the MRMR technique achieves higher success with fewer features than other techniques. The data set used in the second module consists of 9,800 sessions of 3,500 visitors. During these sessions, 185,000 web pages were visited. In this section, it is tried to predict the customer's intention to leave the site without completing a transaction, using only click data. To achieve this, LSTM networks, a type of RNN, were used. It is stated that the LSTM - RNN model predicts that the customer will leave the site after one transaction, with an accuracy rate of 74.3%. If the second module gives a value higher than the specified threshold, the first module is triggered. In this way, it is aimed to identify these customers so that content can be presented to customers who intend to purchase and are likely to leave the site. In the study, it was stated that the increase in the threshold value caused negative results [4].

5. Research Questions

1. What is the distribution of whether the visit results in a transaction or not, by regions?

When Figure 5 is examined, the number of visits to region 1 is much higher than in other regions. Likewise, the number of completing the process for region 1 is quite high compared to other regions.

2. What is the distribution of whether the visit results in a transaction or not, by months?

When Figure 6 is examined, it is seen that there is data for 10 months and May is the most visited month. But the month with the most transactions is November. The number of visitors in February is quite low compared to other months.

3. What is the distribution of whether the visit results in a transaction or not, by weekend?

When Figure 7 is examined, the number of visits on weekdays is higher than the number of visits on weekends. Likewise, purchases on weekdays are higher than on weekends.

4. What is the distribution of whether the visit results in a transaction or not, by visitor type?

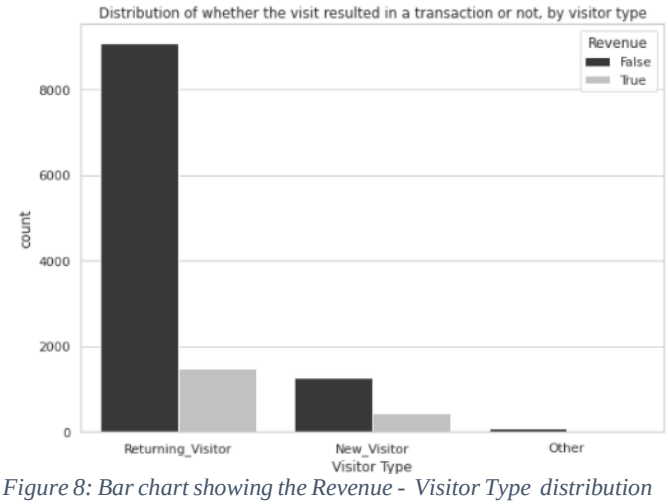

Figure 8: Bar chart showing the Revenue - Visitor Type distribution

When Figure 8 is examined, New_Visitor has a higher transaction completion rate than Returning_Visitor. But the number of Returning_Visitor is higher than both new_visitor and other

5. What is the distribution of whether the visit results in a transaction or not, by traffic type?

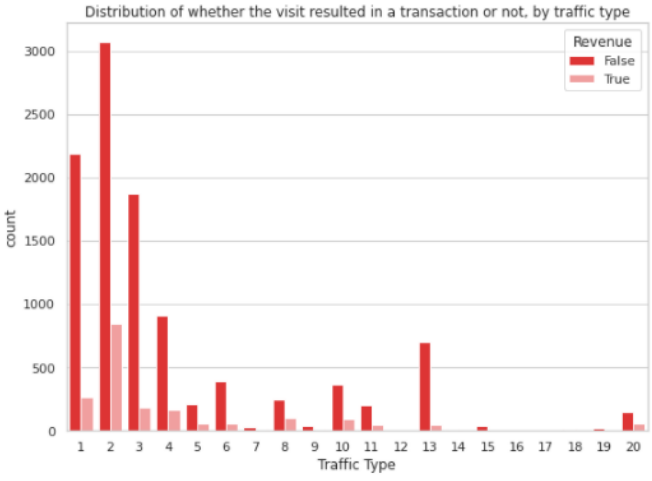

Figure 9:Bar chart showing the Revenue - Traffic Type distribution

5, 7, 9, 11, 12, 15, 16, 17, 18, 19, 20 traffic types have very few transactions or not. It can be said by examining Figure 9 that 2 traffic types are used at most.

6. What is the distribution of whether the visit results in a transaction or not, by browser?

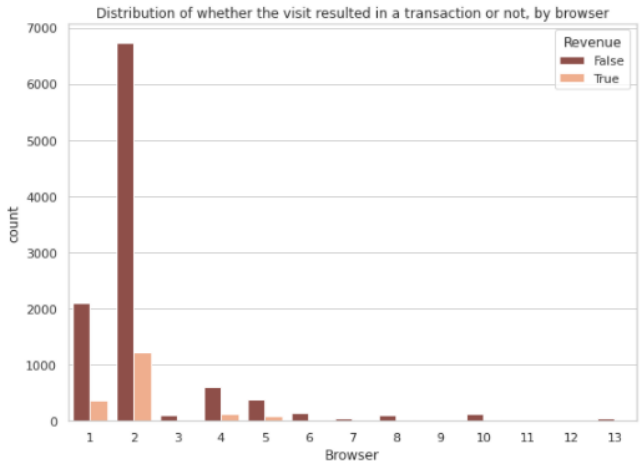

Figure 10:Bar chart showing the Revenue - Browser distribution

It can be seen from the graph in Figure 10 that 3, 6, 7, 8, 9, 10, 11, 12, 13 browsers are used less by users. It can be said that browser number 2 is used the most among other browsers.

7. What is the distribution of whether the visit results in a transaction or not, by operating systems?

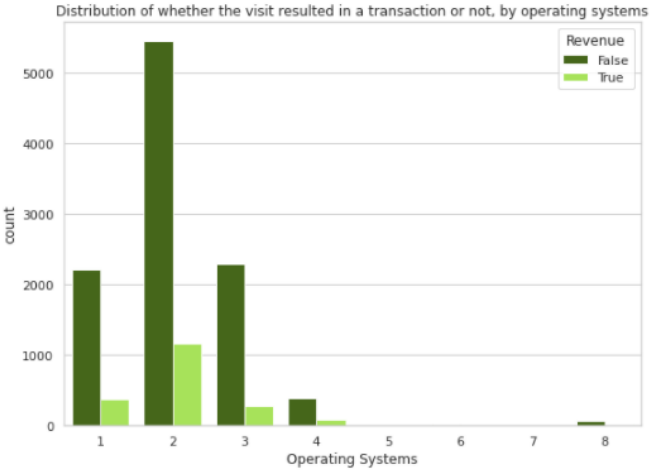

Figure 11:Bar chart showing the Revenue - Operating System distribution

Looking at Figure 11, it can be concluded that operating systems 5, 6, 7, 8 are used less by users than other operating systems. This means that users do not prefer these operating systems too much. Another result that can be reached from the graph is that the operating system number 2 is preferred more than the others, regardless of whether the operation is performed or not.

6. METHODOLOGY

In order to prepare the Online Shoppers Purchasing Intention Dataset for algorithms, it is necessary to apply some preprocessing to the data. The data set does not contain missing data. For this reason, there is no need for the operations to be done for the missing data in this study. Then, Weekend and Revenue attributes with Boolean data type need to be converted to numeric data. LabelEncoder was used during this process. Thus, all the values in the columns were converted to numbers such as 0-1. Likewise, Month and Visitor Type categorical data should be converted to numeric data. For this, the get_dummies() function in the Pandas library is used. After the preprocessing is applied to the data, different algorithms are applied.

a. K-Nearest Neighbors Algorithm

This algorithm is a supervised learning algorithm. An inference is made about the class of the data by looking at the k neighbors of a data to be known which class it belongs to. In this algorithm, the data is placed in the class in which the neighbors of the data are the most. Nearest neighbors are also found according to some metrics. Examples of these metrics are Minkowski, Manhattan, Euclidean, and Chebyshev.

Within the project, three different values were given to the distance criterion and the number of neighbors parameters, which are important criteria of the performance of the KNN algorithm, and nine different outputs were obtained. While 1,3,5 was given as the number of neighbors, Manhattan, Euclidean and Minkowski were used as the distance criterion. In this algorithm with RandomSearchCV, 1, 3, 5, 10, 15 values as the number of neighbors and Euclidean, Minkowski, and Manhattan criteria were preferred as distance criteria. After running this process, the best parameter results were obtained and these parameters were given to the algorithm and new results were obtained. The KNN results considered within the scope of the study were the results that were not obtained by the RandomSearchCV process.

b. Support Vector Machine Algorithm

This algorithm is a supervised learning algorithm. It can be explained as a line drawing that separates the two classes. Here it is necessary to find the most accurate line. There are infinitely many straight lines separating the two classes. While finding the line, the distance of the line to be drawn between the two classes to the elements should be maximized. In this way, the two classes are best separated from each other.

In the project, the kernel parameter, which is one of the parameters of the Support Vector Machine algorithm, is given linear, poly, rbf, and sigmoid, and the results obtained by these methods are reported. Changes in these parameters affect the success rate. With RandomSearchCV, the best result was tried to be found by giving different values to the C and kernel parameters in this algorithm. The C parameter has a different value range from 1 to 10, the best parameter output is obtained by giving linear, poly, rbf and sigmoid values for the kernel parameter. The SVM results considered within the scope of the study were the results that were not obtained with the RandomSearchCV process.

c. Decision Tree Classifier Algorithm

This algorithm is a supervised learning algorithm. The first node of the Decision Tree is called the root, the ones below the root are called the nodes, and the lowest nodes are called the leaves. Decision trees can consist of categorical and numerical data. By applying some rules to a dataset, it is provided to divide the dataset into smaller pieces. One of the important points is how to divide the data.

In the project, different criterion values were given and different results were tried to be obtained. These values are gini and entropy. In this algorithm with RandomSearchCV, 3, 5, 7, 8 values are given to the max_features parameter, 2, 5, 10 values are given to the min_samples_split, gini and entropy values are given to the criterion, and the parameter that gives the best result among these parameters has been tried to be obtained. The Decision Tree Classifier results considered within the scope of the study were the results that were not obtained by the RandomSearchCV process.

d. Random Forest Classifier Algorithm

By dividing the data into parts, different decision trees are created from these parts. The predictions of these decision trees are aggregated and evaluated. As the number of trees increases, the accuracy increases. Leaf nodes contain classes.

Within the scope of the project, different values were given to the criterion parameter as it was first applied in the Decision Tree Classifier. These are gini and entropy. Different results were obtained by giving the value of 100 to the n_estimators parameter. As applied to all algorithms, RandomSearchCV operation was also performed for this algorithm. In this algorithm, the values of 100, 200, 500 are given to the n_estimators parameter, the values of 3, 5, 7, 8 are given to the max_features parameter, the values of 2, 5, and 10 are given to the min_samples_split parameter, and the gini and entropy values are given to the criterion parameter. As a result of testing these parameters, the best parameter output was obtained and these parameters were used in the algorithm. The Random Forest Classifier results considered within the scope of the study were the results that were not obtained by the RandomSearchCV process.

7. FINDINGS

Algorithm	**Accuracy**
Random Forest Classifier	0.898892
Support Vector Machine	0.888348
K Nearest Neighbors	0.872668
Decision Tree Classifier	0.861314

Algorithm	**F1 -Score**
Random Forest Classifier	0.634051
Decision Tree Classifier	0.561913
Support Vector Machine	0.553514
K Nearest Neighbors	0.471338

Algorithm	**Precision**
Support Vector Machine	0.771318
Random Forest Classifier	0.727689
K Nearest Neighbors	0.701550
Decision Tree Classifier	0.552013

Algorithm	**Recall**
Decision Tree Classifier	0.572174
Random Forest Classifier	0.563478
Support Vector Machine	0.445217
K Nearest Neighbors	0.426087

```
K Nearest Neighbors
Accuracy:  0.8718572587185726
Recall:  0.3426086956521739
Precision:  0.6723549488054608
F1-Score:  0.4539170506912442
Confusion Matrix:
 [[3028   96]
 [ 378  197]]
```

Figure 12:One of the KNN results obtained with RandomSearchCV

```
Support Vector Machine
Accuracy:  0.8907812922411462
Recall:  0.5078260869565218
Precision:  0.7070217917675545
F1-Score:  0.5910931174089069
Confusion Matrix:
 [[3003  121]
 [ 283  292]]
```

Figure 13:One of the SVM results obtained with RandomSearchCV

```
Decision Tree Classifier
Accuracy:  0.8972695323060287
Recall:  0.5791304347826087
Precision:  0.7070063694267515
F1-Score:  0.6367112810707456
Confusion Matrix:
 [[2986  138]
 [ 242  333]]
```

Figure 14:One of the Decision Tree Classifier results obtained with RandomSearchCV

```
Random Forest Classifier
Accuracy:  0.8951067856177345
Recall:  0.5565217391304348
Precision:  0.7064017660044151
F1-Score:  0.6225680933852141
Confusion Matrix:
 [[2991  133]
 [ 255  320]]
```

Figure 15:One of the Random Forest Classifier results obtained with RandomSearchCV

Many different outputs were produced using different parameters within the four classification algorithms. When the Accuracy, F1 Score and Precision values obtained after running the program are examined, it is seen that the Random Forest Classifier algorithm gives a higher value than the others. In the Recall score, the Support Vector Machine algorithm gave the highest value. In order to analyze these scores with other parameters and to make hyperparameter optimization, the GridSearchCV method was researched and applied, but since the process took too long even for an algorithm, it was removed from the project and an alternative method, RandomSearchCV, was applied. It was easy to adapt the code to this method because their implementations are the same. RandomSearchCV method gives results in a shorter time by trying fewer combinations than GridSearchCV. But values as good as those obtained from GridSearchCV may not be obtained. With this method, the accuracy value for the Decision Tree Classifier is higher than the previous one. It even surpassed Random Forest by a very small margin to have the highest accuracy. Recall score is greatly increased for Decision Tree Classifier and Support Vector Machine. The Precision score increased only in the Decision Tree and decreased in the rest. Finally, when the F1-Score was examined, an increase was observed in the Support Vector Machine and Decision Tree. The results considered were those not obtained with RandomSearchCV.

7. Conclusion And Recommendations

In this project, it has been tried to create models that can predict a customer's purchase intention by using the information obtained in the past of customers who shop online. Various classification algorithms were used while training these models. Since there is an unbalanced distribution in the dataset, a conclusion should be reached by evaluating other scores instead of just looking at the accuracy value. When the results were examined, it was understood that while the Accuracy, F1 Score, Recall, and Precision scores were evaluated, the Random Forest Classifier algorithm was generally at the top, therefore this algorithm could be preferred. It has been found that methods such as GridSearchCV and RandomSearchCV can be applied to increase the overall evaluation scores. With these methods, it was concluded that parameters that can produce better results can be found by giving different values to the parameters and trying different combinations, thus improving the performance of the algorithms.

References

[1] G. Deyan, 40 Remarkable AliExpress Market Share Statistics You Need to Know in 2021, December 2021. https://techjury.net/blog/aliexpress-market-share/#:~:text=528%20million%20is%20the%20number%20of%20visitors%20on%20AliExpress%20in%20January%202021.&text=As%20of%20the%20first%20month,shoppers%20visited%20about%20seven%20pages.

[2] C. David, Amazon Statistics (2022) , January 2021. https://www.businessofapps.com/data/amazon-statistics/

[3] Sakar, C.O., Polat, S.O., Katircioglu, M. et al. Neural Comput & Applic (2018).

[4] Sakar, C.O., Polat, S.O., Katircioglu, M. et al. Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks. Neural Comput & Applic 31, 6893–6908 (2019). https://doi.org/10.1007/s00521-018-3523-0