Comparative Analysis of Machine Learning Algorithms on E-commerce data sets for online sales prediction

*Estera Wisniewska*

*National College of Ireland*

*Higher Diploma in Science in Data Analytics  
Dublin, Ireland*

Abstract – The project presents a comparative analysis of five Machine Learning Algorithms used for classification problems in the E-Commerce field. Classifiers (K-Nearest Neighbor, Naïve Bayes, Logistic Regression, Decision Tree, Random Forest) were tested on three different-sized data sets, at different stages of data preparation: raw, normalized, and class balanced. Different classification problems: binary or multiclass classification, and lastly, its execution time. For each test, classifiers present changes in the confusion matrix to specific characteristics of data. The results show that Naïve Bayes is the weakest but the fastest to apply, K-NN works well in most classification tasks if the data set is properly prepared. Logistic Regression shall be used in binary classification as multiclass is very problematic. The strongest Classifiers are Decision Tree and Random Forest which produced very high scores in all of the tested examples.

Keywords — Classifier, KNN, NB, Logistic Regression, Decision tree, Random Forest

# Introduction

Electronic Commerce, also known as E-commerce or e-business has gained tremendous popularity in the past decade. There are two types of e-commerce, business-to-business (B2B), where companies conduct their business with other businesses, like suppliers, and distributions. The second part is business-to-consumer (B2C) [1] [2]. This project will use the second type.

## Literature Review on Machine Learning Application in E-commerce

With the rapid growth of e-commerce, there is a need to identify potential clients, understand and distinguish their behavior and create proper methods to convince an undecided potential buyer to make a purchase.

There are many approaches to the task. First of all, it is important to distinguish the level of purchasing intentions [3], [4], next to discover patterns in users’ navigation path [5] and create strongly matching recommendations [6], [7].

The project will analyze the user’s navigation path and Machine Learning classifiers will predict if a user purchased an item or not.

## Machine Learning brief background

Machine Learning is a branch of Artificial Intelligence where machines are trained to learn how to process data and make use of it. They can perform predictive analysis much faster than humans, with a higher level of accuracy.

The Machine Learning field is divided into three categories:

* Supervised learning, where the training stage is required. The model uses labeled input (train set) and learns the relationship between the input and output data. After the training session, the model is ready to predict the outcome of new, unseen data [8].
* Unsupervised learning, is the opposite of Supervised learning. Models use unlabelled data and work to identify patterns in the raw dataset on their own. It is most suitable for discovering and understanding the relationship within the data set or in scenarios where trends and patterns are not known in general [8].
* Reinforcement learning, where the model learns the optimal behavior in an environment to maximize its reward [8].

## Definitions

* Training set: is a significantly larger part of the data set that the model can use to learn patterns and relations between features
* Test set: is the remaining part of the data set that which model did not see yet and can test its prediction on.
* Confusion Matrix: it is an n-dimensional square matrix - where n is the number of distinct target values
* Precision: is the ratio of true positives (TP) to the sum of true positives and false positives (FP)
* Recall: is the ratio of true positives to the sum of true positives and false negatives (FN)
* F-1 score: is an evaluation metric that assesses the predictive skills of the model.

*F1 Score = 2\*(Precision \* Recal)/(Precision+Recal)*

* Chart

  Description automatically generated**Synthetic Minority Oversample Technique (SMOTE): oversample technique which synthetically produces samples of minority class to match the majority class size**

## Classification Algorithms

Five classification algorithms were selected for this project:

* K-Nearest Neighbour (KNN) is a non-parametric classification method that uses distance to classify an object. The algorithm relies on finding the K closest data points to the given test point and using the class label of those points to predict the class of the new point [9].
* Gaussian Naïve Bayes (NB) is based on the principle of Bayes’ Theorem, the probability of an event occurring based on prior knowledge of conditions that might be related to the event.
* Logistic Regression (LOG) is a linear model that uses input features to predict the probability of an event occurring.
* Decision Tree (DT) uses the tree-like model to make a prediction based on the values of different features in the data.
* Random Forest (RF) is an ensemble learning method that combines the predictions of multiple decision trees to create a more accurate model.

# RESEARCH METHODOLOGY

Project compared classifiers from a few angles. Five selected classifiers were tested on:

* Different size data sets (small, medium, large).
* Different stages of data preparation (raw, normalized, balanced).
* Chart, treemap chart

  Description automatically generatedDifferent classification problems (binary classification and multiclass classification).
* Execution time.

## Data Understanding

### ‘Online shopper’s intentions’

The dataset contains 12330 instances and 18 features. The label column, ‘Revenue’, has only two possible classes, ‘True’ and ‘False’ which is understandable as to whether the sale was finalized or not.

Data is a mainly numeric type with high variance and outliers.

Figure ‘Online shopper’s intentions’ outliers

Distributions of the data in the column are not normal and skewed.

A picture containing diagram

Description automatically generated

Figure 'Online shopper's intentions' features distribution

There is a very low correlation between variables and the label column ‘Revenue’. Figure 3 below shows that only two features are correlated higher than 0.6.

Figure 'Online shopper's intention' correlation

### ‘Online Shop’

The data set contains 41694 instances and 26 features, 18 of which present information if the browsing person clicks on a specific area of the website or not. For example, if a person checked the shopping basket or not.

The label column, ‘Item’, is a binary class with a very simple understanding if a product was ordered or not.

Timeline

Description automatically generatedAfter preparation, the data set had only discrete numbers with the majority of values 0 and 1. There are no outliers.

Figure 'Online shop' correlation

Figure 4 presents the correlation between features in the data set. Most of the are not correlated with each other. There are 3 higher correlated features to the label column ‘Item’. Columns: ‘checked delivery detail’, ‘sign in’, and ‘saw checkout’ have correlations higher than 0.6.

### ‘Online Shop USA’

The data set is large and contains 286 392 instances and 10 features. The label column, ‘Status’, is multiclass and represents the status of the online purchase. There are 3 different classes, 1-‘completed’ which means that the item was ordered and the customer decided to keep the item.

Class 2 – ‘refund’. After ordering and receiving the item, the customer decided to return an item and asked for a refund.

Class 3 – ‘cancelled’. The customer initially ordered the item but cancelled before paying any money.

The data set has mixed types of features. Three of them are numerical–intervals. One is categorical–ordinal, and the remaining are categorical–nominal.

After data preparation, the correlation was investigated.

Figure 'Online Shop USA' correlationChart, table, treemap chart

Description automatically generated

Figure 5 shows that there is a very low correlation between variables. Only one column, ‘Gender’ has a higher negative correlation with ‘Prefix name’. No significant correlation between the label and the remaining features was discovered.

## Data Preparation

### ‘Online shopper’s intentions’

Dataset preparation consisted of:

- removing null values, 279 instances were removed.

- mapping the ordinal features into integers. The month column was transformed and values from 2 to 12 were assigned.

- mapping the nominal features into integers. Bool columns were converted to 0 and 1, and ‘Visitor Type’ into 1, 2, and 3.

After the initial Machine Learning model, the dataset was normalized using the following formula:

*(df-df.min()) / (df.max() - df.min())*

This helped to reduce noise in the dataset and improve the score of the machine learning model.

Lastly, to create a third ML model comparison, classes were balanced.

Dataset is binary. Class 1 had 7106 instances and class 2, had 1329. Data was highly imbalanced which may affect the overall model score. Using the oversampling technique, SMOTE, samples for the minority class were generated.

### ‘Online Shop’

The data set does not have any missing values therefore the preparation of the data set started with deleting 3 unnecessary columns: ‘User ID’, ‘date’, and ‘time(s)’.

The next step consisted of applying one-hot encoding to categorical features (’device’ column) and mapping the remaining features to integers.

After the first initial attempt for label prediction, data was balanced using SMOTE. This allowed for further comparison of the classifiers.

### ‘Online Shop USA’

Data has no missing values. One unnecessary column (‘State’) was deleted. The next step required mapping features to integers.

The label is multiclass and imbalanced. Class 1 has 64274 instances, class 2 – 57447 instances, and class 3 – 78753, therefore SMOTE technique was applied after the initial run with Classifiers.

# Machine Learning model comparison

Five models were tested on three different data sets, different stages of data preparation, execution time

### ‘Online shopper’s intentions’

This is a small dataset and was used multiple times for ML algorithm comparison.

Comparing process started by creating ML models just after transforming all data into integers. At this point, data was not normalized, with outliers and imbalanced.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ML classifier | Test Accuracy | Precision | Recall | F1 | AUC |
| KNN | 0.884 | 0.690 | 0.407 | 0.512 | 0.688 |
| NB | 0.926 | 0.713 | 0.844 | 0.773 | 0.892 |
| LOG | 0.956 | 0.894 | 0.794 | 0.841 | 0.889 |
| DT | 0.965 | 0.875 | 0.893 | 0.884 | 0.935 |
| RF | 0.977 | 0.951 | 0.889 | 0.919 | 0.940 |

Table 'Online shoppers intentions' ML Classifiers scores on a basic dataset

Moving forward, data were scaled and outliers were removed. Table 2 shows Classifier’s results built on scaled data.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ML classifier - Normalized | Test Accuracy | Precision | Recall | F1 | AUC |
| KNN | 0.952 | 0.953 | 0.730 | 0.829 | 0.862 |
| NB | 0.902 | 0.635 | 0.916 | 0.750 | 0.907 |
| LOG | 0.970 | 0.980 | 0.826 | 0.896 | 0.911 |
| DT | 0.971 | 0.897 | 0.926 | 0.911 | 0.853 |
| RF | 0.981 | 0.955 | 0.923 | 0.939 | 0.957 |

Table 'Online shoppers intentions' ML Classifiers scores on normalized data

Lastly, classes were balanced. Class 1 – purchase was done, and class 2 – no purchase, were equal.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ML classifier - SMOTE | Test Accuracy | Precision | Recall | F1 | AUC |
| KNN | 0.911 | 0.652 | 0.863 | 0.743 | 0.981 |
| NB | 0.827 | 0.462 | 0.957 | 0.623 | 0.881 |
| LOG | 0.947 | 0.761 | 0.935 | 0.839 | 0.942 |
| DT | 0.962 | 0.838 | 0.920 | 0.877 | 0.945 |
| RF | 0.978 | 0.909 | 0.946 | 0.927 | 0.965 |

Table 'Online shoppers intentions' ML Classifiers scores on SMOTE data

Chart, line chart

Description automatically generated

Figure F-1 Score for different data preparation stages – full data set

Figure 6 presents F-1 scores for all three tests performed on a full ‘online shoppers’ intentions’ data set.

Almost all classifiers, except NB, score the highest just after normalization. SMOTE technique produced similar results as classifiers created on raw data with the exception of KNN whose scores significantly improved after balancing classes. Once again NB classifier score worsens after SMOTE.

Figure 7 below presents the F-1 scores for three tests performed on 3 selected columns only. This selection was performed in order to answer the research question.

Chart, line chart

Description automatically generated

Figure F-1 Score for different data preparation stages - selected features

Scores of all classifiers are very low in comparison to classifiers’ scores based on a full data set. The normalization process was detrimental to almost all classifiers, except NB, which benefited this time.

SMOTE technique was advantageous for all models, especially for Logistic Regression. The model improved the F-1 score from 0.038 to 0.314.

### ‘Online shop’

The medium size data set, ‘Online shop’ was used for the comparison of 5 classifiers.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ML classifier -imbalanced data | Test Accuracy | Precision | Recall | F1 | AUC |
| KNN | 0.989 | 0.816 | 0.950 | 0.878 | 0.970 |
| NB | 0.982 | 0.709 | 0.978 | 0.822 | 0.980 |
| LOG | 0.991 | 0.838 | 0.981 | 0.904 | 0.986 |
| DT | 0.987 | 0.828 | 0.898 | 0.861 | 0.945 |
| RF | 0.990 | 0.833 | 0.950 | 0.888 | 0.971 |

Table 'Online shop' Classifiers scores

Data were discrete from 0 to 1, 0 to 2 in one feature, and 0 to 3 in one feature. There was no need to scale data.

Label classes were balanced using SMOTE.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ML classifier - SMOTE | Test Accuracy | Precision | Recall | F1 | AUC |
| KNN | 0.983 | 0.719 | 0.981 | 0.830 | 0.982 |
| NB | 0.980 | 0.692 | 0.959 | 0.804 | 0.970 |
| LOG | 0.988 | 0.793 | 0.981 | 0.877 | 0.985 |
| DT | 0.987 | 0.798 | 0.934 | 0.860 | 0.962 |
| RF | 0.988 | 0.803 | 0.967 | 0.878 | 0.978 |

Table 'Online shop' Classifiers score – SMOTE

Chart, bar chart

Description automatically generated

Figure Models scores

Figure 8 shows that there are no strong differences in scores between imbalanced and balanced label classes.

### ‘Online Shop USA’

The large-size data set, ‘Online Shop USA’ was used for the comparison of 5 classifiers. It was a multiclass classification and every classifier was tested separately to analyse how accurately the model predict each of the classes.

|  |  |
| --- | --- |
| ML classifier - imbalanced | Test Accuracy |
| KNN | 0.6166 |
| NB | 0.4334 |
| LOG | 0.5028 |
| DT | 0.6563 |
| RF | 0.6653 |

Table Test Accuracy scores

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ML classifier - imbalanced | Class 1 | Class 2 | Class 3 | Average |
| KNN | 0.33 | 0.29 | 0.38 | 0.33 |
| NB | 0.26 | 0.50 | 0.48 | 0.41 |
| LOG | 0.14 | 0.54 | 0.64 | 0.44 |
| DT | 0.60 | 0.56 | 0.77 | 0.64 |
| RF | 0.59 | 0.57 | 0.78 | 0.65 |

Table F-1 Classifiers score for each class - imbalanced

Chart, bar chart

Description automatically generated

Figure Models scores for imbalanced data - multiclass

Data were oversampled and all classes had 78753 instances. Classifiers run through data separately to predict label classes.

|  |  |
| --- | --- |
| ML classifier - balanced | Test Accuracy |
| KNN | 0.6152 |
| NB | 0.4155 |
| LOG | 0.5048 |
| DT | 0.6563 |
| RF | 0.6632 |

Table Test Accuracy score - balanced

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ML classifier - balanced | Class 1 | Class 2 | Class 3 | Average |
| KNN | 0.59 | 0.52 | 0.70 | 0.60 |
| NB | 0.20 | 0.71 | 0.43 | 0.45 |
| LOG | 0.19 | 0.54 | 0.64 | 0.46 |
| DT | 0.60 | 0.57 | 0.76 | 0.64 |
| RF | 0.59 | 0.58 | 0.78 | 0.65 |

Table F-1 Classifiers score for each class - balanced

Figure Models scores on balanced data - multiclass

Figure 10 shows that SMOTE balancing does not improve the model score almost at all for Classifiers: Logistic Regression, Decision Tee, and Random Forest. Their scores improved at best 5%, while K-Nearest Neighbour Classifier improved by 26% for Class 1, 23% for class 2, and 32% for Class 3.

Lat comparison performed to all classifiers was based on execution time with a comparison to the F-1 Score.

Execution time of all classifiers was performed after balancing the classes and then results were scored from 1 to 5, where 1 is the shortest time, and 5 is the longest execution time.

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedSimilarly, scores were scaled. The lowest score among the 5 classifiers gets a score of 1, and the highest is 5.Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Figure 12 13 Execution time vs F-1 Score

Figures 11, 12, and 13 show that in all tests KNN Classifier results are the weakest. It took almost the longest time of all classifiers and produced very weak results.

NB scores are very low in all data sets, but at the same time, worked the fastest. Fandom Forest was the slowest but produced the highest scores.

# Statistical analysis

Models were used to answer 3 questions.

1. Can an online purchase be predicted only by analyzing the time that person spent browsing an online shop website?

From ‘Online shoppers intention’ duration columns (‘Administrative\_Duration’, ‘Informational\_Duration’, ‘ProductRelated\_Duration’, and ‘Revenue’) were selected. They represent the time that a person spent on the selected website of the online shop before making a purchase or leaving the shop.

Classifiers run through data three times. First where data was ‘raw’ – just after transformation into a numeric value. Second, after the normalization process, and third after balancing classes.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ML classifier - Raw | Test Accuracy | Precision | Recall | F1 | AUC |
| KNN | 0.826 | 0.300 | 0.101 | 0.151 | 0.529 |
| NB | 0.818 | 0.283 | 0.125 | 0.173 | 0.534 |
| LOG | 0.845 | 0.393 | 0.020 | 0.038 | 0.507 |
| DT | 0.754 | 0.225 | 0.247 | 0.236 | 0.547 |
| RF | 0.800 | 0.217 | 0.117 | 0.152 | 0.520 |

Table Classifiers scores on raw data set.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ML classifier - Normalized | Test Accuracy | Precision | Recall | F1 | AUC |
| KNN | 0.810 | 0.207 | 0.054 | 0.085 | 0.507 |
| NB | 0.816 | 0.360 | 0.153 | 0.214 | 0.550 |
| LOG | 0.834 | 0.357 | 0.008 | 0.016 | 0.503 |
| DT | 0.741 | 0.208 | 0.203 | 0.205 | 0.525 |
| RF | 0.786 | 0.185 | 0.087 | 0.119 | 0.506 |

Table Classifiers scores on normalized data set.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ML classifier - SMOTE | Test Accuracy | Precision | Recall | F1 | AUC |
| KNN | 0.636 | 0.207 | 0.473 | 0.288 | 0.569 |
| NB | 0.772 | 0.257 | 0.247 | 0.242 | 0.558 |
| LOG | 0.698 | 0.243 | 0.443 | 0.314 | 0.594 |
| DT | 0.647 | 0.196 | 0.408 | 0.265 | 0.550 |
| RF | 0.644 | 0.199 | 0.424 | 0.271 | 0.554 |

Chart, line chart

Description automatically generatedTable Classifiers scores on balanced data set.

Figure Models scores on different stages of data set preparation - selected features

All models scored very badly from small feature selection. SMOTE significantly improved the results of all classifiers. Currently selected features did not bring satisfactory results to be certain that only selected features (‘duration’) can be used for sale prediction.

It is advisable to perform feature importance prior to selecting features for small data set.

1. Can the purchase of an item be predicted only by analysing a person’s behaviour/clicks at different places on a shop website?

Using the ‘Online shop’ dataset, relevant features were selected. The new data set consisted of:

'basket\_icon\_click','basket\_add\_list','basket\_add\_detail','image\_picker','account\_page\_click','promo\_banner\_click','detail\_wishlist\_add','list\_size\_dropdown','list\_size\_dropdown','checked\_delivery\_detail','checked\_returns\_detail','saw checkout','saw\_sizecharts','saw\_delivery','saw\_homepage', ‘Item’.

A new data set was tested for all 5 Classifiers. Table 9 below shows classifiers scores of selected features on imbalanced data. Class 0 counts 31879 and class 1 counts 1470.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ML classifier - imbalanced | Test Accuracy | Precision | Recall | F1 | AUC |
| KNN | 0.992 | 0.856 | 0.995 | 0.920 | 0.993 |
| NB | 0.986 | 0.762 | 0.995 | 0.863 | 0.990 |
| LOG | 0.992 | 0.854 | 0.989 | 0.917 | 0.991 |
| DT | 0.991 | 0.853 | 0.966 | 0.901 | 0.979 |
| RF | 0.992 | 0.851 | 0.984 | 0.913 | 0.988 |

Table 'Online shop’' ML Classifiers scores on imbalanced data for selected features only.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ML classifier - SMOTE | Test Accuracy | Precision | Recall | F1 | AUC |
| KNN | 0.992 | 0.852 | 0.995 | 0.918 | 0.993 |
| NB | 0.986 | 0.772 | 0.971 | 0.860 | 0.979 |
| LOG | 0.989 | 0.803 | 0.992 | 0.887 | 0.990 |
| DT | 0.988 | 0.797 | 0.981 | 0.880 | 0.985 |
| RF | 0.990 | 0.801 | 0.995 | 0.888 | 0.992 |

Table 'Online shop’' ML Classifiers scores on balanced data for selected features only.

Chart, bar chart

Description automatically generated

Figure Models score for selected features - medium on balanced and imbalanced classes

Figure 13 shows that there is no significant difference in Classifiers scores. Both data sets, balanced and imbalanced, have very similar scores.

Lastly, scores were compared to Classifier scores based on the full data set. It was necessary to understand if the question can be answered by smaller data set.

Chart, line chart

Description automatically generated

Figure Models scores - full data set vs selected features

F-1 scores for Classifiers are higher on selected features that on the full data set.

Chart, histogram

Description automatically generated

Figure Improvement of models scores between full data set and selected features only

KNN and NB have the highest improvement on the F-1 Score when only selected features were used.

Logistic Regression, Decision Tree, and Random Forest Classifiers did not profit at the same rate.

1. Which model can best predict the outcome of a possible sale, the fastest?

Five classifiers were tested on all three data sets. All of them were balanced before starting the timer. Execution time in milliseconds are presented in the Table below.

|  |  |  |  |
| --- | --- | --- | --- |
| ML classifier - SMOTE | Small | Medium | Large |
| KNN | 667.69 | 67910.69 | 5554.5 |
| NB | 24.20 | 70.89 | 135.38 |
| LOG | 55.28 | 279.57 | 2981.37 |
| DT | 55.31 | 129.37 | 655.26 |
| RF | 1297.49 | 2270.64 | 21203.82 |

Table Classifiers Execution time in milliseconds

Results for time and F-1 vary strongly, therefore scaling was applied. The model with the highest F-1 score received 5, and the lowest 1. Similarly, with time, a model which took the longest time to produce results received 5, and the shortest,1.

Figures 11,12, and 13 show the scaled results, which clearly present that the Decision tree is one of the fastest options and still can produce very high prediction scores.

# CONCLUSION

Models were tested on various different data sets, types, and problems.

Classifiers were tested on three different sizes of data sets, at different stages of data preparation, with a various number of features and classes to predict.

Results of each test produced the following results:

K-Nearest Neighbour Classifier:

* Tables 1 and 2 show that KNN Classifier scores the best just after data normalization. The model is sensitive to outliers. These data points can have a disproportionately large effect on the distance calculations, and therefore, a much larger influence on the output predictions made by the algorithm.
* KNN algorithm is sensitive to irrelevant features. Scores from Table 5 and 14 shows that the algorithm works better on a smaller number of relevant features.
* Classifier works quite well on multiclass classification while classes are balanced, producing similar scores as Random Forest, Table 9.

Gaussian Naïve Bayes Classifier:

* Algorithm performs better when features in the data are independent of each other. Figures 3 and 4 show that the overall proportion of correlated features in the data set ‘Online Shop’ is lower than in ‘Online shopper’s intention’
* Classifier benefits from removing outliers as Tables 1 and 2 present.
* During testing, the Classifier could not make a prediction about a category that was not present in the training set and assigned zero probability to it. The ‘Zero frequency’ problem was solved by changing the ratio of training and test set to 80:20 instead of 70:30.

Logistic Regression:

* Figures 3 and 4, and results from Tables 2 and 4 bring the conclusion that the performance of the Logistic Regression classifier strongly depends on the relevance of the features in the data.
* Among all selected algorithms, the Logistic Regression algorithm worked very badly on imbalanced, small data sets with only a few features (Table 12).

Decision Tree:

* Tables 1 and 2, as well as Tables 10 and 11 show that the Decision Tree Classifier does not necessarily need normalizing data before applying the algorithm. The difference in scores between raw data and normalized is between 2 and 3 percent.
* Due to the nature of this Classifier, the used time to receive results was significantly longer than in previous models.
* Table 4 and Table 5 show that the algorithm is not very affected by class imbalance.

Random Forest:

* Classifier scores very high on the raw data set, Table 1, the algorithm can handle the outliers.
* Balancing classes is not always necessary. Tables 4 and 5, and Tables 7 and 9 show a very small impact of SMOTE technique on the F-1 score in the Random Forest Classifier.

Classifiers were tested for execution time (in milliseconds) based on three, different-size data sets. Data set 1, ‘Online shopper’s intention’ after SMOTE had 14 200 instances, data set 2, ‘Online shop’ after SMOTE had 36 764, and data set 3, ‘Online Shop USA’, after SMOTE 234 726.

Even though Random Forest takes a much longer time to produce classification time, it scores always very high in many different data sets, and therefore, if time or expenses do not matter, and the system can process a large amount of data, this classifier would be the best choice.

# Bibliography

|  |  |
| --- | --- |
| [1] | Y. Tian, "Reaserchgate - History of e-commerce," 2007. [Online]. Available: https://www.researchgate.net/publication/314408412\_History\_of\_E-Commerce. [Accessed 2022]. |
| [2] | Shahid Amin, Keshav Kansana, Jenifur Majid, "Reaserchgate - a review paper on e-commerce," 2016. [Online]. Available: https://www.researchgate.net/publication/304703920\_A\_Review\_Paper\_on\_E-Commerce. [Accessed 2022]. |
| [3] | Grazyna Suchacka, Magdalena Skolimowska, Aneta Potempa, "reaserchgate - Classification of e-customer session based on Suport Vector Machine," 2015. [Online]. Available: https://www.researchgate.net/publication/282794728\_Classification\_Of\_E-Customer\_Sessions\_Based\_On\_Support\_Vector\_Machine. [Accessed 2022]. |
| [4] | Grazyna Suchacka, Grzegorz Chodak, "reaserchgate - Practical aspects of LOG file analysis for e-commerce," 2013. [Online]. Available: https://www.researchgate.net/publication/289068998\_Practical\_Aspects\_of\_Log\_File\_Analysis\_for\_E-Commerce. [Accessed 2022]. |
| [5] | Ching-Huang Yun, Ming-Sayan Chen, "Reaserchgate - Mining web transaction patterns in a electronic comerce environment," 2000. [Online]. Available: https://www.researchgate.net/publication/2617977\_Mining\_Web\_Transaction\_Patterns\_in\_an\_Electronic\_Commerce\_Environment. [Accessed 2022]. |
| [6] | T. I. J. o. E. E. 12(2), "Reaserchgate - Using fuzzy association rules to design e-commerce personalized recommendation system," 2014. [Online]. Available: https://www.researchgate.net/publication/274477783\_Using\_Fuzzy\_Association\_Rules\_to\_Design\_E-commerce\_Personalized\_Recommendation\_System. [Accessed 2022]. |
| [7] | Qing Duan, Jian Li, Yu Wang, "Reaserchgate - The application of fuzzy association rule mining in e-commerce information system mining," 2012. [Online]. Available: https://www.researchgate.net/publication/288339749\_The\_Application\_of\_Fuzzy\_Association\_Rule\_Mining\_in\_E-Commerce\_Information\_System\_Mining. [Accessed 2022]. |
| [8] | K. Lakhwani, "Reaserchgate - Machine Learning," 2021. [Online]. Available: https://www.researchgate.net/publication/353141813\_Machine\_Learning. [Accessed 2022]. |
| [9] | D.T.Larose, Discovering knowledge in Data: an introduction to Data Mining, Wiley, 2005. |
| [10] | Guofang Kuang, li yuanchen, "Reaserchgate - Using fuzzy association rule to design e-commerce personalized recommendation system," 2014. [Online]. Available: https://www.researchgate.net/publication/274477783\_Using\_Fuzzy\_Association\_Rules\_to\_Design\_E-commerce\_Personalized\_Recommendation\_System. [Accessed 2022]. |