Journal of Life Economics
Volume / Cilt: 8, Issue / Sayı: 1, 2021, pp. 59-68
E - ISSN: 2148-4139
URL: https://www.journals.gen.tr/jlecon
DOI: https://doi.org/10.15637/jlecon.8.1.06

Geliş : 07 Ocak 2021
Kabul: 29 Ocak 2021

Received: 07 January 2021
Acccepted: 29 January 2021

ARAŞTIRMA MAKALESİ / RESEARCH ARTICLE

Comparison of machine learning classification algorithms for purchasing forecast

Rabia Özdemir[1] **Münevver Turanlı**[2]

1 İstanbul Ticaret University, Institute of Science, Department Head of Statistics, İstanbul, TURKEY, E-mail: rabiaozdemir1@gmail.com,

2 Prof. Dr., İstanbul Ticaret University, Faculty of Humanities and Social Sciences, İstanbul, TURKEY. E-mail: mturanli@ticaret.edu.tr

Abstract

With the development of computer technologies and invention of internet, many concepts have entered our lives. With the starting of wide usage of globalized internet network, concept of machine learning has emerged in time for smarter management of data flow in big dimensions. In line with technological developments, all activities began to be carried to digital environment and as a result of this, concept of e-commerce has entered our lives. E-commerce is one of the areas where machine learning is used most widely. By examining product purchasing situations in accordance with data available at the enterprises, various researches have been made for selection of most appropriate model in order to predict future data. In the study it was mentioned about concepts of e-commerce and machine learning and by applying Logistic Regression, Naïve Bayes and Support Vector Machines being machine learning classification algorithms, it has been aimed to determine the model having best accuracy ratio.

Keywords: E-commerce, Logistic Regression, Naïve Bayes, Support Vector Machines, Classification

JEL codes: L81, C11, C38, C39, C53

Citation/Atıf: ÖZDEMİR, R. & TURANLI, M. (2021). Comparison of Machine Learning Classification Algorithms for Purchasing Forecast. Journal of Life Economics. 8(1), 59-68, DOI: 10.15637/jlecon.8.1.06

Corresponding Author/ Sorumlu Yazar:
Rabia Özdemir
E-mail: rabiaozdemir1@gmail.com

1. INTRODUCTION

With the invention and development of internet, world has underwent a big change and internet which was used by big institutions at the beginning and which had certain limits, has begun to be used by all the institutions and people. On the other hand, while e-commerce has attained an important place in our lives in the last 10 years, historically its application leans on the past. Without requiring for people and institutions to exist physically, e-commerce has turned into a multi-billion-dollar worth industry. Number of people purchasing commodities and services online increase significantly worldwide in each passing day.

As people live as being dependent on digital devices and internet in our current world, a new user behavior has emerged. New user approach has affected majority of markets and sectors and as people tend to realize their transactions online, shopping which is one of the biggest economical values was carried to digital devices and hence, e-commerce began to be applied (Çelik, Ertemel, 2016). As a result of application of e-commerce, each user becomes a data source. For example, when a person buys a product or a service, when he makes research on a subject, when he shares his likes, opinions, skills, interests in social communication platforms, and shortly in each activity he does he continuously produces data.

E-commerce that is considered as non-ending sector of future, provides a wide scale application area with machine learning and artificial intelligence and for this reason there is need for new contents and algorithms. Companies use these algorithms to evaluate their current positions and to direct them in order to meet customers' requests and provide a preferable service. These algorithms are important for businesses and anyone who wants to trade on e-commerce, whether their products are purchased or not, in order to predict this situation and develop strategies.

In this study, first of all the concept of e-commerce will be tried to be explained and then, models will be explained theoretically, and in application by using data of customers purchasing products through e-commerce website three different classification algorithms will be established and by comparing these established algorithms, model providing the most correct result will be determined.

As a result of model that is determined, by using the correct algorithm about whether the newly coming customer will purchase the product or not, targets can be set by making prediction on factors such as revenue and performance relating with the future.

2. LITERATURE REVIEW

In the field of e-commerce, survey studies were conducted to investigate demographic information on online shoppers and to make determinations about purchasing habits and the results were analyzed with statistical tests. Machine learning classification methods for purchasing in the field of e-commerce are used in many enterprises, but there is no detailed academic study.

Due to the wide application areas of machine learning, many studies have been done on different subjects. Credit risk assessments, text classification, mapping and disease diagnoses are the most commonly used topics in terms of classification. There are various studies on the comparison of classification methods related to many subjects with each other in the literature. Some current and comprehensive classification studies to predict two classes, including Naive Bayes (NB), Logistic Regression (LR) and Support Vector Machines (SVM) methods, are as follows.

In 2019, Wuday conducted a credit scoring study for corporate applications for opening of a corporate franchise branch opening using the UCI German Credit data. Considering the approval for branch opening, the most successful models is Support Vector Machines (SVM) and Multiplayer Perceptron (MLP), the Random Forest (RF) method for the German Credit data set has given the most successful results. Küçük (2019) used the Gauss Process, Naive Bayes, Support Vector Machines, Random Forest, K-Nearest Neighbour, Logistic Regression, Desicion Tree classification methods and concluded that the most accurate prediction model for diagnosing Parkinson's disease is the Gauss Process. Günay (2018) has made a customer loss analysis on whether customers in the telecommunications sector will be out of membership with machine learning techniques. Logistics Regression, Support Vector Machine, Naïve Bayes, Artificial Neural Networks and New Approach: Multiplied Possibilities Method and the most successful model is Artificial Neural Networks. Bagui et al. (2017) used machine learning Support Vector Machine, Logistics Regression, Naives Bayes, K-Nearest Neighbour, Random Forest, Gradient Boosting Tree (GBT) methods to classigate VPN network traffic flow and investigated the model that gave the most accurate result. They have determined that the models that give the most accuracy are RF, GBT.

3. E-COMMERCE

In many sources relating with e-commerce, definitions made by institutions of different countries and international organizations can be seen. According to these definitions, the most general definition of e-commerce is specified as "carrying out commercial business, transactions and acts in electronic environment". E-commerce covers transactions relating to all commercial activities at the organizational and individual level without physical communication and it is based on the application of all digital data, sounds, texts and visual images which are produced, processed and transferred in this area. At this point, e-commerce covers consumer protection, competition, finance and payment systems, taxation, intellectual, industrial and commercial property rights, security, legal regulations, ending of disputes, etc. (Kantarcı, et al., 2017)

E-commerce covers all the platforms that enable meeting of businesses or individuals who want to sell their products and services anywhere in the world and customers who need the products sold anywhere in the world. With the help of e-commerce platforms, many companies and individuals can easily proceed with international growth rather than local growing. Most significant features of e-commerce are such that it can be used in every environment where internet access is available, and that it can be accessed by means of personal use technological devices such as mobile phones, tablets, laptops and desktop computers.

It is possible to summarize electronic commerce with below specified items (Kaplancan Güler, 2017):

- It is realization of product, data, service or collection processes by connecting with digital devices or online networks,
- It is the usage of technology in work processing and automation of commercial transactions,
- It is a place for establishing a stakeholder environment for group members to realize activities, establish cooperation, and to learn,
- It is revealing of shopping skill of service and information by means of online applications and virtual networks,
- It is a tool that shows the desire of companies, customers and administration to reduce service costs while increasing the quality of products and shortening the delivery dates of the products.

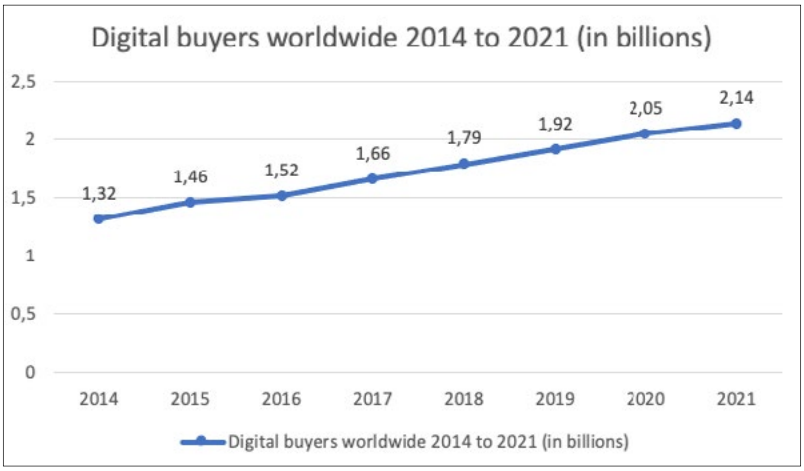
E-Commerce has many advantages besides these characteristics. If it is explained in general terms, we can count below particulars:

- Convenience; It can be accessed to e-commerce sites 24 hours a day, all days of the week. Users have the easiness of accessing the product and service they want on 24/7 basis. In order to be able to shop from a physical store or request service from any institution, it is necessary to be physically present in stores or shops within the determined working hours. E-commerce provides a lot of easiness regarding this time aspect.
- Access speed; Time is reduced significantly while using e-commerce sites. Customer does not experience problems arising from crowd in the store, and he does not wait for any queue, etc. Due to the bandwidth and information processing support specified in the e-commerce platforms access is enabled quickly at any time. Regarding this, users do not have to wait for any time. Customers need time to reach physical stores.
- Availability of lots of options; Customers can see and evaluate all of the many options of store, product and services and they can obtain the desired product from different vendors. They can have access to products in different segments, and in summary they can reach to a wide product range. These products do not have a certain limit, they can reach all of the products that cannot fit in a store due to its size. For example, Amazon's first slogan was "The World's Largest Bookstore". This is one of the first initiatives to be made as a store that can be reached by people all over the world via e-commerce of books that cannot fit in a library (Pedersen, 1995).
- Easy accessibility; Customers who shop in a physical store may have difficult time in finding out which section and shelf of the store a particular product is on. In e-commerce, visitors can browse the product category pages and they can immediately access to the product using the product search feature.
- International access is one of the main reasons of being preferred. It is a big opportunity for sellers to reach to a wide audience, to discover different customer potentials, and to increase their earnings. It also enables customers to meet with vendors and products that are in different regions.

In addition to these advantages, e-commerce also has some disadvantages such as having limited customer service, inability to see or touch a product before purchasing, waiting time for product shipment, and security problems as a result of problems during product purchase.

As technology develop in time, advantages and disadvantages may change and their number may increase or decrease. Worldwide and Turkey use of e-commerce numbers are shown in Figure 1 and Figure 2.

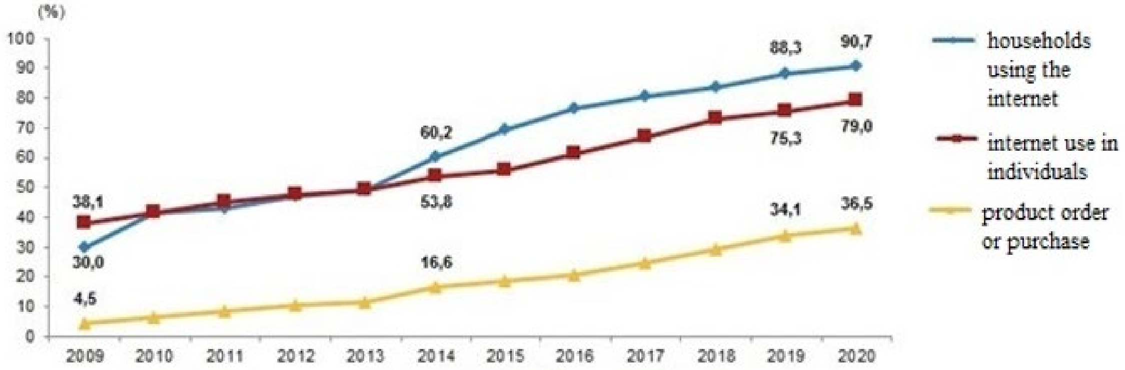
Figure 1. Digital Users Realizing Purchasing Throughout the World 2014-2021.



Source: Statista

As it is seen in Figure 1, 2.05 billion people in the world make purchases over the internet in 2020. 41% of internet users and 26% of the world population make purchases on the internet.

Figure 2. Basic Indicators About Usage of Information Technologies by Households, 2009-2020.



Source: TÜİK

As seen in Figure 2, the proportion of individuals using the Internet in 2020, according to TUIK data, Turkey is 79%. The rate of ordering or purchasing products over the Internet was 36.5%.

Types of e-commerce after these increases seen in internet usage and purchases of individuals; Business-to-Business E-commerce, Business-to-Consumer E-commerce, Consumer to Consumer E-commerce will be examined under three headings.

Business to business e-commerce (Business to Business-B2B): Ordering from business to business, obtaining invoices and paying prices in B2B electronic environment. B2B is the form of e-commerce that is mostly used worldwide. This form of trade, defined as trade between businesses, is carried out both through sectoral portals and through the firm's own site. While B2B enables businesses to be stronger in the sector, it provides the opportunity to acquire new export markets and new customers in the domestic market, especially in crisis environments (Sarısakal and Aydın, 2003).

Business to consumer e-commerce (Business to Customer-B2C): E-Retailing, from business to consumer, is examined in two dimensions by taking into account the type of product sold and the way the product is sold. The type of product sold can be physical (such as clothing, electronics, etc.) or digital (such as software, music, books), while the medium where the product is sold can be either the seller's own website or shopping platform of a third party. Shopping platforms offer customers diversity and price advantages by collecting the products of more than one seller on the same website. Amazon, eBay and shopping platforms such as Gittigidiyor in Turkey assume the role of mediator between the consumer and the seller in the e-retailing operations (TCKB, 2013).

Customer to customer e-commerce (Customer to Customer-C2C): E-commerce from consumer to consumer; It is usually provided by websites that offer free classifieds, forums, auctions, and pages for startup entrepreneurs. eBay, Sahibinden.com, Gittigidiyor.com Amazon offers pages to create C2C applications and activities. Said websites earn their revenues from the small commissions they get from the seller. On the other hand, these websites offer a secure and comfortable payment method (Statista, December, 2020).

3.1. History of E-Commerce

Although the formation of the concept of e-commerce has come out more in the last two decades, the first formations of this concept extend to earlier times. E-commerce started to make great progress with the spread of the internet after 1995. Due to this reason, it is better to evaluate the emergence of the concept of e-commerce before and after 1995.

When the studies conducted before 1995 are examined, it is found out that Jamie Bartlett writes in his book titled The Dark Net that the first online transaction was a drug deal. In 1972, long before eBay or Amazon were established, students from Stanford University in California and MIT (Massachusetts Institute of Technology) in Massachusetts realized the first online transaction. Using the Arpanet account in AI labs, Stanford students secretly sold small quantities of marijuana. In his 2005 book titled as What the Dormouse Said: How the Sixties Counterculture Shaped the Personal Computer Industry, John Markoff says that the first online transaction was a drug deal. Information Technology Engineer Michael Aldrich and his colleague Peter Champion pioneered the birth of online shopping in 1979 with their inventions to connect a computer to a television. (Coleman & Ganong, 2014). Later on a private network called Minitel was launched in 1982 by France Telecom, enabling users to call phone numbers, set up travel reservations, make financial transactions, and shop online (Cornelius et al., 2002). This first tele-shopping is referred to as electronic shopping according to some sources (Çelik, 2015).

In 1995, when Jeff Bezos established Amazon.com, e-commerce began to move to different dimensions. First internet radio stations that broadcast 24 hours a day without any promotional activity were established and started broadcasting under the names NetRadio and Radyo HK. Dell and Cisco have embarked on using the internet for commercial activities. eBay was founded as AuctionWeb by computer programmer Pierre Omidyar. eBay is one of the companies that follow Paypal. I hepsiburada.co the entry into operation in 2000, the sector was made in Turkey.

4. MACHINE LEARNING

Following the emergence and development of the concept of artificial intelligence in the 1950s, artificial intelligence learning processes were started in the 1980s and the concept of machine learning developed in this context. Together with these technologies, the concept of deep learning has emerged in the 2010s. Machine learning enables a computer system to make predictions using historical data and accordingly make decisions. Machine learning has different definitions in the literature by different people and organizations with similar meaning.

Machine learning is the process of solving situations being previously observed and problems being defined with machine learning techniques depending on a specific data set (Aydın, 2018) Main purpose of machine learning is to create models that can improve themselves, detect complex patterns and find solutions to new problems using previous data (Tantuğ ve Türkmenoğlu, 2015).

Model which is using machine learning can be an explanatory model that is used to extract information, since it can be a model with ability to predict for making decisions about the future. The main purpose of machine learning can be expressed as finding the parameter values that provide the best performance at the end of the model (Alpaydın, 2010). Machine learning is programming computers to optimize a performance metric using sample data or past experiences. They use statistical theory in establishing mathematical models, because the important thing is to infer from a sample. On the other hand, computer science has two important roles in machine learning. First one of these is the need for efficient algorithms to solve the optimization problem and to store and process the large amount of data generally possessed during the training phase. Second, after a model is learned, its representation and algorithmic solution must also be effective for its inference (Alpaydin, 2016).

Machine learning is the concept of completing the learning action by observing real data within certain limits, making use of these data and developing them over time, so that machines can learn, think, act like humans with the support of certain programs.

Machine learning can be examined in four different ways as shown below:

- Supervised learning: This method is a machine learning estimation method that is used to reach the result / output variable by using the existing data as input. The purpose of supervised learning is to generate functions to predict output values based on a set of input values and to explain the relationship between output values from data sets in which both input and output values are included. Classification and Regression analysis are among these methods.
- Unsupervised learning: This method is a prediction method applied by using the available data as input in the absence of output / result value. In addition, this method tries to create a relationship and pattern with the available data. As it is known, data has a unique structure and to make arrangements within the data itself, clustering is done according to whether the samples are more or less. On the other hand, clustering is done according to the density and similarity of the data. Shortly, by looking at a few key attributes of a data point, one usually tries to guess other attributes with which they are associated.
- Semi-supervised learning: This method is a learning between supervised learning and unsupervised learning. It is a type of learning which uses untagged data and some tagged data together.
- Reinforcement learning: In this learning method, a goal is specified to the learning system to fulfill it and it is expected for it to achieve this goal through trial and error. The system organizes trial-and-error processes according to the most rewarding actions and learns how to achieve the given goal (Mohri, Rostamizadeh, Talwalkar, 2012).

4.1. Logistic Regression

Logistic Regression analysis is one of the traditional statistical analyzes that is mostly used to examine the cause-effect relationship between dependent and independent variables or variables. These analyzes differ, especially because the variable is qualitative or quantitative.

Logistic regression model differs from linear regression models as its dependent variable is categorical and is used for classification operations. Logistic regression analysis takes its name from the logit transformation applied to the dependent variable. As it is the case in linear regression models, the aim is to establish a model that can define the relationship between dependent and independent variables. Since none of the assumptions sought in linear regression is sought in logistic regression, it has more flexible usability (Mertler, Vannatta, 2005). Generally logistic regression model can be defined as follows:

$$L = \left[\frac{P_i}{1-P_i}\right] = Z_i = b_0 + b_1 X_i + e_i \qquad (1)$$

The probability that it is here is P_i or the probability of not being 1-P_i is calculated with the equivalence 1 1/(1+e^{-z}). Being here is written in the form of: $Z=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\cdots+\beta_nX_n$ β's seen in the expression above show the regression coefficient. Therefore, p values can be obtained by taking the antelope of Z values (Özer, 2004).

Logistic regression analysis according to the number of categories included in the dependent variable; It is divided into three parts as Binary, Multiple (more than two categories) Classifier (Multinomial), Multiple (more than two categories) Sorter (Multiordinal). The dependent variable consists of two categories, since it is desired to make estimates about whether customers buy or not, therefore, Binary Logistic Regression Analysis was used in the study.

4.2. Naive Bayes

Naïve Bayes method which is used as a statistical method is an algorithm created by using the bayes theorem. The reason for its being most frequently used algorithms in the literature is its easy understanding and applicability. Using this method, it is possible to find the probability that a data belongs to the class value of the targeted attribute. In other words, it is a method that calculates the probability of new incoming data being one of the existing class tags by using existing, classified data (Bozkır, Sezer, 2009). In short, this theorem calculates the conditional probability of the class to which the data belongs and it uses the Bayes theorem to anticipate the probability of the class to which the data belongs. Bayes theorem is defined as:

$$P(H\backslash X) = \frac{P(X\backslash H)P(H)}{P(X)} \qquad (2)$$

The expression number 2 above is X; feature vector, H is the hypothesis expressing the probability of an attribute vector belonging to a class such as C. P (H ׀ X) represents the successor probability. Given the Bayes theorem, the algorithm of the Naive Bayes classifier is as follows;

Assume that the class label of each X in a sample that is thought to represent the data set is specific. X is a vector of n features and represented as X=(x_1, x_2, … x_n). According to these explanations below formula has been obtained:

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \qquad (3)$$

Let's assume that it shows C_1, C_2,… C_m m grain classes as shown in the expression number 3 above. The Naive Bayes classifier tries to find the value with the highest succession probability P (C_i ׀ X) of all classes to find out whether a vector X belongs to the class C_i.

Since the value of P (X) is the same for all classes, only the expression P (X ׀ C_i) P (C_i) should be made to the maximum. P (C_i) expression is the ratio of the number of elements in class C_i to the number of elements. P (X ׀ C_i) expression is calculated with the following equation number 4, assuming X is an attribute vector containing n values.

$$P(H|C_i) = \Pi_{k=1}^{n} P(X_k|C_i) \qquad (4)$$

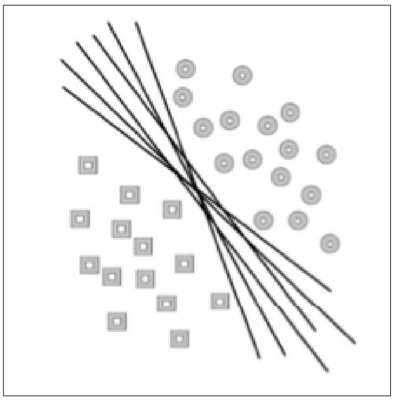
As a result, as seen in Equation 4, the classifier C_i class with the largest expression P (X ׀ C_i) P (C_i) is selected as the class of the X vector (Kaynar and others, 2017).

4.3. Support Vector Machines

Even though the support vector machines method is used for both regression and classification purposes, it is mostly used for classification in machine learning. This method can classify both linear and nonlinear data, however it mostly tries to classify the data linearly. On the other hand, this method is a highly preferred method because it produces significant accuracy with less calculations. Support Vector Machines method is mainly used to distinguish the data belonging to two classes in the most appropriate way. For this, decision boundaries or in other words, hyper planes are determined and so they try to find the best line separating the class.

This method, which separates the data of two classes from one another, aims to determine the best decision function (hyperplane) to be able to realize an efficient classification (Vapnik, 2000). In other words, the purpose of the method is to determine the hyperplanes that make the most efficient separation among the infinite grain hyperplanes that provide classification, which is important in classifying the data of two classes. These explanations are as seen in Figure 3.

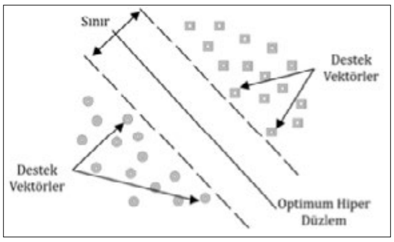
Figure 3: Hyper Planes for Linearly Separable Data



Source: Kavzoğlu & Çölkesen, 2009

When determining these hyperplanes, the boundary distance between two hyperplanes is maximized by using the principle of structural risk minimization.

Figure 4: Optimum Hyperplane and Support Vectors



Source: Kavzoğlu and Çölkesen; 2009

As a result, as seen in Figure 4, the optimum hyperplane is obtained, where the distance between the plane and the closest points (support vectors) of the classes to the plane is maximized (Vapnik, 1995).

5. APPLICATION

In this study, data of a Turkey-based e-commerce website (incikcincik.co) is used. The data includes information about 101,763 visitors who entered the site between dates of March 2019 and March 2020 via Google Analytics. A unique customer number is assigned to each visitor, and data covers the number of times these customers enter the site, the time they stay on the site, the rate of bounce off the site, the number of transactions and whether they have made purchases.

Within the context of the study, the dataset is divided into two parts as training and test sets with the hold-out method. While the training set consists of the data on which the model was trained, the test set; it is the data which is used to see how well the model performed over untrained data. Separation of data as training and test varies according to the size of the data set. In the study, the data set was divided into 70% education and 30% test data. Machine learning classification algorithms on confusion matrix data were tested with Python 3.7 programming language in Jupyter Notebook 6.1.4 environment.

Some evaluation metrics are used to evaluate the models created with classification algorithms and to determine which classification model produces more accurate results. These are based on a table which is commonly referred to as the confusion matrix. It is a table layout developed to visualize the performance of a classifier in machine learning and statistical classification problems (Han, Kamber, 2012).

Table 1: Confusion Matrix with Two Classes

		Predicted class	
		0	1
Actual class	0	True Negative (TN)	False Negative (FN)
	1	False Positive (FP)	True Positive (TP)

As seen in Table 1, in confusion matrix with two classes;

- True Positive (TP); represents how many of the data belonging to the positive class were correctly classified by the classifier.
- True Negative (TN); represents how many of the negative-class data are correctly classified by the classifier.
- False Negative (FN); It is the labeling of data that actually belongs to the positive class as a negative class as a result of the classification.
- False Positive (FP); it is the labeling of data belonging to the negative class as positive class as a result of classification).

As per results of confusion matrix, below values are used for comparison purpose.

Accuracy: This is the scale providing ratio of correctly classified inputs to total inputs.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (5)$$

Precision: This scale is the ratio of correctly classified positive inputs to total positive values.

$$Precision = \frac{TP}{TP+FP} \qquad (6)$$

Recall: This scale is the ratio of correctly classified positive inputs to actual positive values.

$$Recall = \frac{TP}{TP+FN} \qquad (7)$$

F1 scale: It is a scale calculated as harmonic average of precision and recall to ensure comparison.

$$F1\ scale = \frac{2*P*R}{P+R} \qquad (8)$$

Following above explanations with LR, NB, SVM methods prediction purchasing status of customers is made and results shown in table 2 have been obtained.

Table 2: Comparison of Classification Methods

	Logistic regression		Naive Bayes		Support Vector Machines	
	Not Purchasing	Purchasing	Not Purchasing	Purchasing	Not Purchasing	Purchasing
Precision	0.85	0.94	0.62	0.98	0.83	0.98
Recall	0.94	0.84	0.99	0.43	0.94	0.43
F1 scale	0.89	0.89	0.76	0.60	0.88	0.60
Accuracy	0.89		0.70		0.88	

Table 2 shows the results of three classification methods that predict the purchasing status of customers. As a result of the Logistic Regression Analysis 85% of the people who do not buy and 94% of the customers who make purchases have been correctly estimated. The ratio of customers who do not make a correctly predicted purchase to total estimates is 94%, while the ratio of customers who make anticipated purchases to total estimates is 84%. The classification algorithm made with the Logistic Regression model has generally reached an accuracy rate of 89%.

As a result of the Naïve Bayes analysis, 62% of the people who do not buy and 98% of the customers who make purchases have been correctly estimated. The ratio of customers who do not make a correctly predicted purchase to total estimates is 99%, and the ratio of customers who make predicted purchases to total estimates is 43%. Classification algorithm with Naive Bayes has reached 70% accuracy in general.

When it is looked at the Support Vector Machine analysis; 83% of non-purchasers and 98% of purchasers have been correctly estimated. It is seen that the ratio of customers who do not make a correctly predicted purchase to the total estimates is 94%, and the ratio of the customers who have made the estimated purchases to the total estimates is 43%. Classification algorithm made with Support Vector Machine has reached 88% accuracy in general.

6. RESULT

As a result of technological developments, e-commerce has reached an endless development level. Basic technologies form the structure of e-commerce and they use machine learning and artificial intelligence algorithms to catch up with the developing era. When the data obtained by e-commerce is used with the right algorithms, many useful results can be obtained both with regards to the seller and the customer.

People who produce goods and services dealing with e-commerce must determine a path for themselves using their existing data in order to make strategic progress towards their customers. Using machine learning algorithms; Customer profiles can be determined, customer-oriented campaigns, discounts, basket options and many personal privileges can be provided. In addition by means of correct machine learning algorithms in relation to data they have in their hands companies can realize various actions that promote purchasing such as application of periodical campaign, extension of product range that is liked, application of different purchasing options and cargo options, with the aim to benefit from opportunities both relating with customers and the market. On the other hand, businesses can make analyzes for their competitors and as a result of all these analyzes, they can determine their strategies, choose the right decisions, increase their income and volume, and provide access to more buyers in larger markets.

In this study, Logistic Regression, Naïve Bayes and Support Vector Machine methods have been applied to find the right algorithm for whether customers make purchases in the business-to-consumer e-commerce type, and the method that gives the best result is investigated. The methods by which the machine learning model is established try to reach the correct result with factors such as the redundancy and distribution of the data. While creating the machine learning model, the most important particular is to what extent the learning takes place, meaning how accurate the model makes predictions.

Examining the accuracy rates of the methods, it is seen that Logistic Regression has rate of 89%, Naïve Bayes has rate of 70%, and Support Vector Machine has rate of 88%. Naïve Bayes is observed to be the method with the lowest results with 70%. If the dependent variable is categorical, Logistic Regression Analysis being frequently used, has been the classification method that gives the best accuracy in this study, but the customers who make purchases are of great importance for the company. It is equally important to correctly anticipate customers who make purchases. Logistic Regression Analysis correctly predicted purchasing customers with rate of 94%, Naïve Bayes with rate of 98% and Support Vector Machines with rate of 98%. Taking into account the correct prediction success rates of both general model and purchasing customers, it has been seen that Support Vector Machines give successful results.

This study guides people who want to work in this field to have an idea. Each data set has its own features. It is seen in the literature that different classification models for different subjects make high accuracy. Other classification algorithms can be tried to find the right model for whether customers make purchases or not, and more detailed studies can be done by using different size data sets and different variables.

REFERENCES

- ALAN, A. & KARABATAK, M. (2020). Veri Seti - Sınıflandırma İlişkisinde Performansa Etki Eden Faktörlerin Değerlendirilmesi. *Fırat Üniversitesi Müh. Bil. Dergisi*. 32(2), 531-540.
- ALPAYDIN, E. (2010). *Introduction to Machine Learning, United States of America*. Massachusetts Institute of Technology. Second Edition, ISBN-13: 978-0-262-01243-0.
- ALPAYDIN, E. (2016). *Machine learning: The New AI, United States of America*. Massachusetts Institute of Technology. Fist Edition, ISBN-13: 978-0262529518.
- AYDIN, C. (2018). Makine Öğrenmesi Algoritmaları Kullanılarak İtfaiye İstasyonu İhtiyacının Sınıflandırılması. *European Journal of Science and Technology*. 14, 169-175.
- BAGUI, S., FANG, X., KALAIMANNAN, E., BAGUI, S.C. & SHEEHAN, J., (2017). Comparison of Machine-Learning Algorithms for Classification of VPN Network Traffic Flow Using Time-Related Features. *Journal of Cyber Security Technology*, https://doi.org/10.1080/23742917.2017.1321891, 1(2), 108-126.
- BARTLETT, J. (2015). *The Dark Net: Inside the Digital Underworld, United States of America*. Melville House, ISBN: 978-1-61219-489-9.
- BOZKIR, A.S, SEZER, E. & GÖK, B. (2009). Öğrenci Seçme Sınavında (ÖSS) Öğrenci Başarımını Etkileyen Faktörlerin Veri Madenciliği Yöntemleriyle Tespiti. *5th International Advanced Technologies Symposium,* 13-15 May, Karabük, 1-7.
- COLEMAN, M.J. & GANONG, L.H. (2014). *The Social History of the American Family: An Encyclopedia.* United States of America: Sage Publications.
- COLLEY, W., (2019), *Comparison of Machine Learning Algorithms For Financial Evaluations*. Thesis (MS), Kocaeli: Gebze Technical University.
- CORNELIUS, P., KIRKMAN, G., SACHS, J. & SCWAB, K. (2002). *Country Profiles. The Global Information Technology Report 2001-2002: Readiness for the Networked World*, New York, Oxfod: Oxford University Press.
- ÇELIK, B. (2015). *An Exploratory Analysis of Online Shopping Behavior in Turkey.* Thesis (MS), Bahçeşehir University, Istanbul.
- ÇELİK, B. & ERTEMEL, A.V. (2016). An Exploratory Analysis of Online Shopping Behavior in Turkey. *International Journal of Commerce and Finance*, 2(1), 67-80.
- GÜNAY, M. (2018). *Makine Öğrenmesiyle Müşteri Kayıplarının Tahmini.* Thesis (MS), Istanbul University, İstanbul.
- HAN, J., KAMBER, M. & PEI, J. (2012). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers.
- KANTARCI, Ö., ÖZALP, M., SEZGİNSOY, C., ÖZAŞKINLI, O. & CAVLA, C. (2017). *Dijitalleşen Dünyada Ekonominin İtici Gücü: E-Ticaret*, TÜSİAD Publication, ISBN: 978-605-165-022-7, April, 04-587.
- KAPLANCAN, G.V. (2017). *Türkiye'de ve Dünya'da E-Ticaret, Sanal İşletme ve Sanal Mağazacılığın Gelişimi ve Karşılaşılan Sorunlar Üzerine Bir Vaka İncelemesi*. Thesis (MS), Nisantasi University, Istanbul.
- KAVZOGLU, T. & COLKESEN, I. (2009). A Kernel Functions Analysis for Support Vector Machines for Land Cover Classification, *International Journal of Applied Earth Observation and Geoinformation,* 11, 352-359.
- KAYNAR, O., TUNA M.F., GÖRMEZ Y. & DEVECI, M.A. (2017). Makine Öğrenmesi Yöntemleriyle Müşteri Kaybi Analizi. *C.Ü. İktisadi ve İdari Bilimler Dergisi*, 18(1).
- KÜÇÜK, R.G. (2019). *Makine Öğrenmesi Yöntemleri ile Parkinson Hastalığının Teşhis Edilmesi.* Thesis (MS), Istanbul Aydın University, Istanbul.
- MARKOFF, J. (2005). *What the Dormouse Said: How the Sixties Counterculture Shaped the Personal Computer Industry*. Penguin Books. ISBN:9780670033829.
- MERTLER, C. A. & VANNATTA, R. A. (2005). *Advanced and Multivariate Statistical Methods: Practical Application and Interpretation.* CA: Pyrczak, Glendale.
- MOHRI, M., ROSTAMIZADEH, A. & TALWALKAR, A. (2012). *Foundations of Machine Learning.* UK, London: The MIT Press Cambridge.
- ÖZER, H. (2004). *Nitel Değişkenli Ekonometrik Modeller.* Ankara: Nobel Yayın Dağıtım.
- PEDERSEN, P. (1995). *World's Largest Bookseller Opens on the Web, Amazon,* https://press.aboutamazon.com/news-releases/news-release-details/worlds-largest-bookseller-opens-web/ [Accessed Date: 22/01/2021]
- REPUBLIC OF TURKEY MINISTRY OF DEVELOPMENT. (2013). Internet and E-Commerce Entrepreneurship Axis Current Situation Report. *Information Society Strategy Project Renewal of April* 10.
- SARISAKAL, M., NUSRET & AYDIN M. ALI. (2003). New Face of E-Commerce Mobile Commerce. *Aviation and Space Technologies Magazine*, 1(2), 83-90.
- STATISTA, COPPOLA D. (NOV 27, 2020). *Number of digital buyers worldwide from 2014 to 2021* https://www.statista.com/statistics/251666/number-of-digital-buyers-worldwide/ C2C E-Commerce, https://www.statista.com/markets/413/topic/983/c2c-e-commerce/
- TANTUĞ, A. C., & TÜRKMENOĞLU, C. (2015). *Türkçe Metinlerde Duygu Analizi*, Thesis (MS), Istanbul Technical University, Istanbul.
- TÜİK, DOĞAN A. (AGUST 25 2020). Hanehalkı Bilişim Teknolojileri (BT) Kullanım Araştırması, https://data.tuik.gov.tr/Bulten/Index?p=Hanehalki-Bilisim-Teknolojileri-(BT)-Kullanim-Arastirmasi-2020-33679
- VAPNIK, V.N. (2000). *The Nature of Statistical Learning Theory*. Second Edition, USA, New York: Springer-Verlag, ISBN:0387987800.
- VAPNIK, V.N. (1995). *The Nature of Statistical Learning Theory*. USA, New York: Springer-Verlag, ISBN:9780387945590.