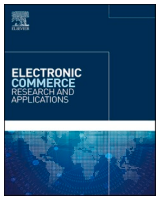
Electronic Commerce Research and Applications 62 (2023) 101340



Contents lists available at ScienceDirect

Electronic Commerce Research and Applications

journal homepage: www.elsevier.com/locate/elerap





Feature selection for Turkish Crowdfunding projects with using filtering and wrapping methods

Murat Kilinc [a,*], Can Aydin [b]

[a] Manisa Celal Bayar University, Computer Research and Application Center, 45140 Manisa, Turkey
[b] Dokuz Eylul University, Department of Management Information Systems, 35390 Izmir, Turkey

ARTICLE INFO

Keywords:
Feature selection
Crowdfunding
Machine learning
Success classification

ABSTRACT

Crowdfunding (CF) platforms host an increasing number of projects, where financial support from backers plays a vital role in project realization. Unfortunately, CF projects have experienced a downward trend in success rates. To address this issue, it is crucial to identify the factors that influence success by analyzing project characteristics. In our study, we collected project data from Turkey's Fongogo CF platform, performed feature selection, and rigorously tested the results. We employed various methods such as Pearson correlation, mutual information statistics, chi-square, Fisher's score from filtering methods, and recursive feature elimination from wrapper methods to understand feature relationships. We proposed a cross-validated recursive feature elimination method for feature selection. The identified success factors were classified using diverse machine learning algorithms, with the Gradient Boosting algorithm achieving the highest result of 84.28%. The results obtained with wrapper methods highlight the potential of utilizing features in decision support processes to enhance CF success.

1. Introduction

Crowdfunding refers to raising funds from the masses for the purpose of raising financial resources, often used by entrepreneurs or project creators. CF has become a fast-growing phenomenon, especially in the last 10 years, with the financial power that the masses can create more easily on the internet (Borrero-Domínguez et al., 2020; Short et al., 2017). Crowdfunding refers to raising funds from the masses for the purpose of raising financial resources, often used by entrepreneurs or project creators (Fanea-Ivanovici and Baber, 2021). CF has become a fast-growing phenomenon, especially in the last 10 years, with the financial power that the masses can create more easily on the internet (Baber, 2021; Sakarya and Bezirgan, 2018). Today, the fund collected for all projects is around 35 billion dollars (Crowdcrux, 2020). When evaluated in terms of human capital and collected funds, the dynamics of this growing ecosystem should be analyzed in detail.

Although the market for CF has grown rapidly in recent years, projects on platforms have not been successfully financed. For example, according to a study conducted in 2019, the success rate of projects on Kickstarter was 37.44% (Yeh and Chen, 2022). In Turkey, Crowdfunding platforms started to operate fully after 2011, and there was a full mobilization after 2013 (Çubukçu, 2017). With the launch of platforms such as Fongogo, CrowdFon, FonlaBeni, Arıkovanı, Buluşum, the number of projects has gradually increased. Today, the number of platforms has also increased with the establishment of FonBulucu, Startupfon, Startupmarket and Ideanest platforms. Moreover, while Turkey's characteristics, such as the social security system and competition on CF platforms, create a unique context for CF projects, they can also create complexity for fundraisers. In this context, the lack of social security means that fundraisers need to navigate their campaigns more carefully and effectively (Peprah and Shneor, 2022). The important role that trust plays in campaign marketing strategies holds the potential for greater success in this context. As a result of this potential, interest in CF platforms is growing in Turkey.

However, even though the interest in CF platforms is increasing, the success rate shows a noticeable downward trend. According to a study including Fongogo, the most popular reward-based CF platform in Turkey, the success rate of Fongogo was 30% in 2017. In a similar study, the success rate of CF platforms in Turkey was 29.18% in 2018 (Çonkar and Canbaz, 2018). By 2021, the number of projects increased approximately 5 times, but the success rate on the Fongogo platform decreased to 26.6%. This decline in the success of projects on CF platforms can be explained by the features of the projects. Because the success of each project is directly related to the features such as category, past

* Corresponding author.
E-mail addresses: kilinc.murat@cbu.edu.tr (M. Kilinc), can.aydin@deu.edu.tr (C. Aydin).

https://doi.org/10.1016/j.elerap.2023.101340
Received 22 June 2023; Received in revised form 24 October 2023; Accepted 27 November 2023
Available online 3 December 2023
1567-4223/© 2023 Elsevier B.V. All rights reserved.

experience of the project creator, gender, social media, goal amount, number of words in the project text, number of images, number of updates, etc. In order to clearly establish the connections and patterns between these relationships, it is necessary to collect all kinds of data that can be obtained from reward-based CF platforms and analyze and identify only the factors that most affect success. Because one of the main reasons for the decline in success in CF platforms is poorly prepared projects. To solve this problem, a machine learning-based approach can be used to develop a mechanism for predicting success based on CF features. In order to develop the relevant mechanism, project quality should be conveyed to users according to the CF success signaling theory and funding processes should be carried out accordingly (Chen et al., 2023).

Accordingly, the research question "What are the factors that most affect quality signaling and project success in the Turkish CF ecosystem?" was posed. Based on the related question, firstly, the features that affect success in CF platforms were identified through literature research in Chapter 2. Then, the dataset was prepared using data scraping techniques in the methodology section, and the specified feature selection methods and machine learning algorithms were chosen for binary classification-based success prediction. In the findings section, the results were tested using performance evaluation criteria after success classification with the selected features and algorithms. The findings were compared with other studies in the literature, the results were discussed, and the groundwork was laid for web-based business intelligence platforms that can predict success based on project characteristics. When the findings were summarized, it was seen that the features (category of the project, number of projects supported by the project creator, number of campaign days, length of the promotional video, number of images, number of social media followers, number of content words, and target amount) selected with the RFECV (Recursive Feature Elimination with Cross-Validation) method, in response to the research question put forward, performed the best classification with 84.28%. In other words, the findings show that the features selected with RFECV are of great importance for the success of a project.

2. Related works

Within the scope of the research question posed, many different studies have been found in the literature on feature selection and the factors affecting success in CF platforms. Since each CF platform can contain different features, the differences in the studies examined are striking. However, the reward based Fongogo platform we used for our study is inclusive because it contains many features.

2.1. Success factors in Crowdfunding

Success factors in CF platforms can vary greatly between platforms. However, the factors affecting success in reward-based CF platforms are largely created with a similar structure. For this reason, studies in the literature generally focus on the same features. In this direction, in the study conducted by Shneor and Vik in 2020, the factors affecting the success of CF were determined by examining the academic studies presented only in English between the years 2010–2017. It is emphasized in the study that the number of content words, the gender of the project owner, the number of social media followers, the length of the video, past experiences, etc., directly affect the success of CF (Shneor and Vik, 2020). In another study by Alazazi et al., it is seen that the amount of targets, the number of campaign comments, the impact on social media and the features of project images affect the success of the projects added to the CF platforms. In the study, success prediction was also made with machine learning algorithms and the Support Vector Machines (SVM) algorithm gave the best result (Alazazi et al., 2020). In the study by Borrero et al., it is emphasized that the geographical location variable, human capital, gender and experience play a major role in CF success. The study also stated that the supporters in Spain attach importance to the past experiences of the project owner (Borrero-Domínguez et al., 2020). In another study conducted by Wei et al. in 2022, the tension between novelty and similarity in DFI projects was addressed and patterns for successful projects were revealed. The research measured the degree of similarity of DFI projects and determined that the optimal level target for funding and success is the amount of funds raised in previous projects (Wei et al., 2022). In a study conducted by Zhong et al. in 2022, the relationship between the relationships between CF projects and funding performance is analyzed. The research extends the concept of project network and shows that models with network-based features perform better with experimental results. (Zhong et al., 2022). In another study by Shi et al., the difficulty of video and audio analysis was mentioned among the factors used to predict the success of the project on the CF platform. However, it was stated that these factors should not be ignored in order to analyze the projects better. Because these factors provide good advertising for promoters (Shi et al., 2021). In another study by Jhaveri et al. in 2019, it was stated that the factors that most affected the success of the project were the number of campaign days, the target amount, the category of the project, location and media content. Also, success was predicted with Random Forest, CatBoost, XGBoost and AdaBoost algorithms and AdaBoost gave the best result with 84% (Jhaveri et al., 2019). In the study carried out by Li et al. in 2016, it was mentioned that the classification-based approach in CF projects is not correct, and it was explained that the balance of the number of successful and unsuccessful projects in the data set should be well established. One of the other elements of the study is to enable the project creators to have an idea of where they stand compared to other projects in terms of reaching the target amount before the project starts. This actually reflects business analytics and decision support processes (Li et al., 2016). In the research conducted by Koch and Siering in 2015, it was stated that the comprehensive information in the form of text, pictures and videos included in the project description positively affects the success of financing. Contrary to other studies, as a result of the research, it was concluded that the past experiences of the project owner didn't affect the success factor. In addition, the study suggests that the number of words be included in the dataset by going deeper into features such as the project description for future studies (Koch and Siering, 2015).

3. Feature selection

In the research of Mhanovic and Kevric in 2020, WEKA application was used for feature selection. The Gain Attribute Evaluator algorithm was chosen to reduce the number of features in the data set. According to the findings, in addition to obtaining higher accuracy values thanks to feature selection, classification time is also reduced. If an application is to be developed, it means performance for the UI (Mehanović and Kevrić 2020). In the study of Ryoba et al. in 2021, feature selection was made with the Whale Optimization Algorithm (WOA) method and a success rate of 90.28% was achieved with the K-Nearest Neighbors (KNN) algorithm. It is recommended to measure the effect of all features on success by using the Pearson correlation coefficient during feature selection in the study. In addition, it was also stated in the study that campaigns with more facebook interaction number and frequently asked questions (FAQ) features were more popular and more likely to be successful (Ryoba et al., 2021). In a study conducted by Kaynar et al. in 2018, chi-square, information gain, gain rate, gini coefficient, oneR, genetic algorithm, relief, forward and backward feature selection algorithms were used to determine attack detection. Classification was made with the features obtained as a result of the selection, and the results were tested with evaluation criteria. Also, in the study, the use of extreme learning machines is recommended because the training phase is easier for attack detection (Kaynar et al., 2018). The results of 10 feature selection methods used in another study by Kou et al. in 2020 were evaluated according to performance, stability and efficiency with 3 different classification algorithms. In the study, it is recommended to use

Table 1
Features of the created dataset.

Success Factors	Data Type	Explanation	Supporting Literature
Project Category	Categorical	Project categories within the CF platform. There may be category differences between platforms.	(Greenberg et al., 2013; Shi et al., 2021; Wang et al., 2020)
Funding Type	Categorical	Projects created according to the "all or nothing" or "all left" model	(Ryoba et al., 2021; Zvilichovsky,Danziger,and Steinhart, 2018)
PO Gender	Categorical	Gender of the project owner	(Shneor and Vik, 2020), (Borrero-Domínguez et al., 2020)
PO Supported Projects	Numeric	Number of projects supported by the project owner on the CF platform	(Koch and Siering, 2015; Ryoba et al., 2021; Shi et al., 2021; Yeh and Chen, 2022)
PO Subscribed Projects	Numeric	Number of projects subscribed by the project owner on the CF platform	(Borrero-Domínguez et al., 2020; Zhou et al., 2018)
PO Project Count	Numeric	Total number of projects on the CF platform of the project owner	(Borrero-Domínguez et al., 2020; Greenberg et al., 2013; Ryoba et al., 2021)
PO Team Count	Numeric	The number of times the project owner took part in the team in other projects within the CF platform	(Borrero-Domínguez et al., 2020; Orçun and Sezgin, 2020; Ryoba et al., 2021; Shi et al., 2021)
Location	Categorical	The location of the project or the location of the owner of the project	(Jhaveri et al., 2019; Mollick, 2014; Wang et al., 2020; Yuan et al., 2016)
Started Date	Numeric	The date the project was presented to the supporters	(Ryoba et al., 2021; Zhou et al., 2018)
Campaign Day	Numeric	Number of active days that the project expects support on the CF platform	(Mollick, 2014; Ryoba et al., 2021; Shi et al., 2021; Wang et al., 2020; Zhou et al., 2018)
Promo Video and Length of Video in Seconds	Categorical, Numeric	Whether there is a video prepared for the promotion of the project and the length of this video in seconds	(Greenberg et al., 2013; Koch and Siering, 2015; Shneor and Vik, 2020)
Images Count	Numeric	Number of images used for project promotion	(Alazazi et al., 2020; Koch and Siering, 2015; Yeh and Chen, 2022)
Updates and FAQs Count	Numeric	Number of updates and FAQs until the end of the campaign days of the project	(Kuppuswamy and Bayus, 2013; Mollick, 2014; Shneor and Vik, 2020; Yeh and Chen, 2022)
Comment Count	Numeric	Number of comments from supporters about the project	(Ryoba et al., 2021; Shi et al., 2021)
Supporters Count	Numeric	The number of people who support the project by purchasing from the rewards in the campaign	(Mollick, 2014; Yeh and Chen, 2022)
Reward Count	Numeric	The number of rewards that the project owner has prepared for the supporters	(Greenberg et al., 2013; Shneor and Vik, 2020; Zhou et al., 2018)
Team Member	Numeric	Number of people in the project team	(Ralcheva and Roosenboom, 2020; Yeh and Chen, 2022)
Web Site	Categorical	Whether the project has a website	(Borrero-Domínguez et al., 2020; Shi et al., 2021)
Social Media and Number of Followers	Categorical, Numeric	Whether the project has Facebook, Instagram, Twitter or Linkedin accounts and the number of followers in these accounts	(Atsan and Erdoğan, 2015; Koch and Siering, 2015; Mollick, 2014; Shneor and Vik, 2020; Yuan et al., 2016; Greenberg et al., 2013)
Tags Count	Numeric	Number of descriptive tags created by the project creator	(Yeh and Chen, 2022)
Description Text and Content Word Count	Numeric	The text in which the project was introduced to the supporters and the number of words in it	(Koch and Siering, 2015; Ryoba et al., 2021; Shneor and Vik, 2020; Yuan et al., 2016; Zhou et al., 2018)
Target Amount	Numeric	The amount of funds requested by the project owner from the supporters for the realization of the project	(Etter,Grossglauser,and Thiran, 2013; Greenberg et al., 2013; Lukkarinen et al., 2016)
Pledged and Support Rate	Numeric	The amount of money and support rate collected in the number of days of the campaign together with the target amount offered	(Kuppuswamy and Bayus, 2013)
Success State	Numeric	The project that reaches the target amount within the number of days of the campaign is considered successful and the others are considered unsuccessful. Category of the project according to this situation	(Greenberg et al., 2013; Wang et al., 2020)

PO: Project Owner

more than one performance measure to evaluate the methods. (Kou et al., 2020). Another study by Yan and Zhang in 2015 used recursive feature elimination (RFE) one of the main wrapper methods implemented with SVM. It is stated that the SVM-RFE combination, which is stated to outperform typical selection algorithms, is suitable for size reduction and data interpretation (Yan and Zhang, 2015). In the research conducted by Misra and Yadav in 2020, feature reduction classification was made for type 2 diabetes by recursive feature selection with cross validation (RFECV). In the study, which reached 84% accuracy with Logistic Regression (LR), the problem of over-fitting in wrapper feature selection methods arises (Misra and Yadav, 2020).

4. Signaling theory and project quality signaling

In CF, fundraisers request funds for different needs and communicate these requests to donors through online campaigns. In this case, fundraisers are considered as signal generators, sending specific signals through their online campaigns. In this way, donors make funding decisions based on quality signals and can change their decisions based on the reliability of the signals. Therefore, CF is consistent with signaling theory. CF platforms allow fundraisers to send specific signals and donors to use these signals to make funding decisions. This process can help reduce information asymmetry and influence decision-making. Signaling theory is related to the way users invest in or support projects in CF. In crowdfunding, a project creator or fundraiser has information about a specific project or product. However, potential investors or backers may not have this information and may use signals to make their decisions (Chen et al., 2023). Signals may consist of elements such as the project description provided by the fundraiser, previous achievements or other social proof. The credibility of the signals depends on factors such as the project creator's reputation, previous achievements or reliable references (Steigenberger and Wilhelm, 2018). Donors or investors evaluate these signals and make a decision about the project. Where signals are strong, investors or backers are more likely to fund or support the project (Kromidha and Robson, 2016). Therefore, signaling theory emphasizes the importance of the use and reliability of signals in the decision-making process of investors or backers in crowdfunding.

To summarize the studies in the literature, the factors that most affect project success were revealed by using meta-fields from CF platforms operating in different parts of the world (Table 1). In other words, the features in the dataset of the study were created by analyzing the studies in the literature and the CF platforms in Turkey. The features on the table that indicate past experience and are related to the project creator are not present in every CF platform. However, since the impact of past experiences mentioned in the literature on success cannot be

ignored, they are included in our study. In addition, since filtering and wrapper methods have been used in the literature to solve similar problems with good results, the appropriate methods were applied in the dataset. In this direction, this study contributes to the lack of regional success factor and project quality signal analysis in the literature. Because it is seen that people in geographically different cultures have different tendencies to support or invest in the CF. For this reason, it is necessary to reveal the success factors of the CF ecosystem in Turkey and to create success prediction models.

5. Method

Within the scope of our study, the CF data set obtained by data scraping and collection methods was made ready for feature selection after incomplete data completion, data preprocessing and normalization processes. First, in the feature selection process with filtering methods, individual features were determined with Pearson Correlation (PCC), mutual information statistics (MI), chi-square (CHI), and Fisher score (FS). Afterwards, the selection phase of the factors affecting the CF success was completed with RFE and RFECV methods, which are wrapper methods.

Evaluation of accuracy scores using machine learning (ML) algorithms, SVM, Gaussian Naive Bayes (GNB) and Logistic Regression (LR), Random Forest (RF) and Gradient Boosting (GBM) algorithms for binary classification after feature selection tested with criteria. Also, the best parameters for the use of ML algorithms were applied within the method, ensuring that they were found.

For the test phase, the F1 Score, precision, recall, area under the curve (AUC) and Kappa Scores, which are among the performance evaluation criteria, were used (Fig. 1). Also, the values revealed by the AUC scores were also evaluated and analyzed within the scope of the ROC curve. The scores obtained are calculated with the values in the confusion matrix.

5.1. Filter methods

The common feature of filtering methods is to reveal the relationship between any success factor in the data set and the target variable (Marono et al., 2007; Öğündür, 2019). Accordingly, a subset from the entire feature group is selected by filtering. Then, the selected features are applied with learning algorithms and their performance is observed.

5.2. Pearson correlation

Pearson Correlation is one of the most used methods to measure the relationships between the variables in the data set. The correlation coefficient obtained in the method takes values between −1 and +1 (Liu et al., 2020). In other words, it is the covariance of the two available variables divided by the standard deviations of these two variables. If the coefficient approaches the minimum or maximum values, it shows that the relationship between the features is high. In addition, it should be analyzed whether the variables in the data set conform to the normal distribution during the use phase. When two variables (features) x and y are accepted respectively; $\sum x$ = sum of values in all x variable column, $\sum x^2$ = the sum of the squares of the values in the entire x variable column, $\sum y$ = sum of values in all y variable column, $\sum y^2$ = the sum of the squares of the values in the entire y-variable column, $\sum xy$ = the sum of the products of the values in the x and y columns, with n = number of data, the correlation coefficient is calculated with the following formula (1).

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{\left[n\sum x^2 - (\sum x)^2\right]\left[n\sum y^2 - (\sum y)^2\right]}} \quad (1)$$

In terms of its use in the study, the use of PCC for determining success factors in CF projects emerges as a necessity. Because these projects usually involve a large number of variables and understanding the relationships between these variables can affect the success of the project. Therefore, by identifying the relationships between these factors, PCC can help to understand which factors affect project success more.

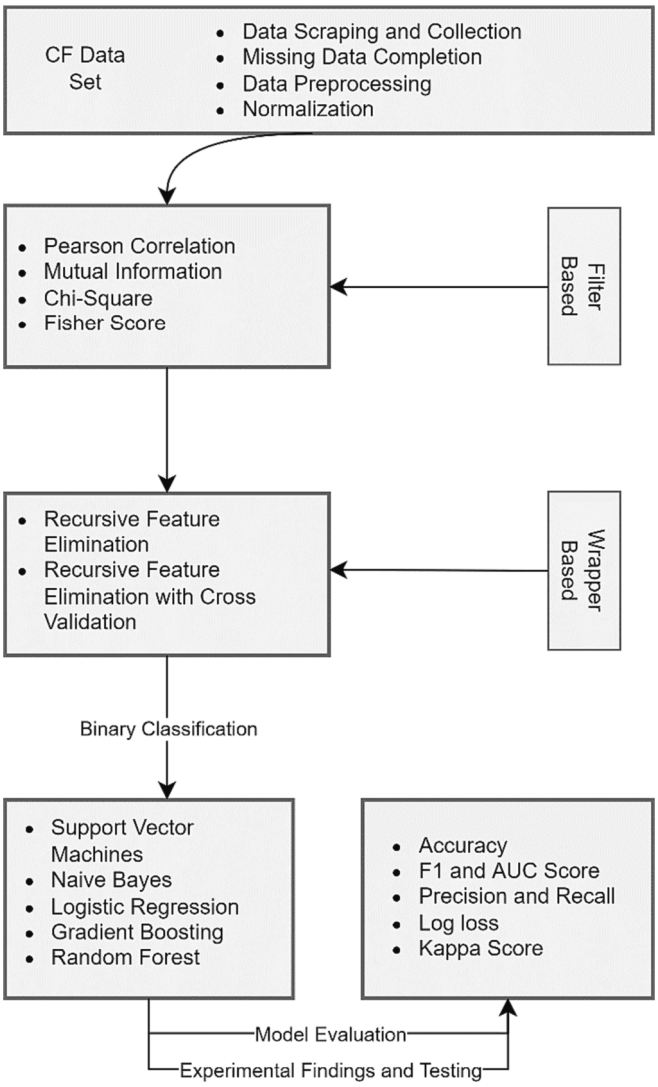


Fig. 1. Feature selection methods and performance measures.

5.3. Mutual information

Because it is computational, it is one of the most popular feature selection methods for simple interpretation and efficiency (Tang et al., 2014). While analyzing the mutual information statistics of the two selected random variables, the dependency between the two variables is measured (Witten et al., 2017). That is, in this method, in which the amount of information contained in the variables in the data set is measured, a random variable is selected, and observations are made according to other variables. The resulting measure is symmetrical. The mutual information statistics for the selected random variables x and y are calculated with the following formula (2). According to the formula, H(x) is an entropy for x and H(x|y) is a conditional entropy for x given y. In addition, a reduction operation can be performed for a k value determined by using the SelectKBest property on the features selected by the method.

$$I(x,y) = H(x) - H(x|y)$$
$$= H(x) + H(y) - H(x,y),$$

$$MI = \sum\sum P(x,y)*\log\frac{P(x,y)}{P(x)*P(y)} \quad (2)$$

In terms of usage, the MI method is a method that should be used to determine success factors, especially in CF projects. This method helps to identify the most important features by measuring the relationship with the target variable during feature selection. Therefore, MI method can help to increase the success of the project by helping to select the right features in CF projects.

5.4. Chi-Square

It is another filtering method used to measure the relationship between variables. (Kocakafa, 2020). It is applied for the variables in the data set numerically, not categorically or negative. The relationship between the target variable and n features is measured using the chi-square test (3). The resulting result can be reduced by specifying a value of k with the SelectKBest property.

$$X_c^2 = \sum_{i=1}^{n}\frac{(O_i - E_i)^2}{E_i} \quad (3)$$

The CHI method is one of the methods that should be used to select the right features in CF projects. Because CHI helps to select the right features by determining the importance ranking of the features. In addition, identifying unnecessary features can contribute to increasing the success of the project by increasing efficiency.

5.5. Fisher's score

The Fisher Score, which calculates the relationship between each non-negative feature and class, is frequently preferred for filter-based feature selection and is used to evaluate categorical and numerical variables in classification tasks (Tang et al., 2014). The aim of the method is to find a subset of features in the data space covered by the selected features, such that the distances between data points of different classes are as large as possible (Gu et al., 2012). Fisher Score of S_i, the i th feature in the data set; μ_{ij} is calculated by the mean and variance of the i th feature in the j th grade for ρ_{ij} (4).

$$S_i = \frac{\sum_{k=1}^{K} n_j(\mu_{ij} - \mu_i)^2}{\sum_{k=1}^{K} n_j {\rho_{ij}}^2} \quad (4)$$

When considered for feature selection in CF projects, the FS method is used to measure the relationship between the project features and the target variable. In this way, a feature importance ranking can be created by determining the importance score for each feature. This importance ranking helps to determine the most important features to be used in the project. The fact that it is an effective method in classification problems and that it can select the right features to increase classification accuracy has enabled the use of the FS method in the study.

5.6. Wrapper methods

In wrapper methods where performance measurements are made according to the applied model score, the selected method is trained with machine learning models for the subsets of the features, and the appropriate features are selected for the data set. The parameters on the applied methods can have a great effect on the results. With these parameters, it is possible to select n features that give the best results or to make consecutive forward–backward selections.

5.7. Recursive feature elimination

A popularly used wrapper-based selection algorithm for feature selection. The main reason for its frequent use is the convenience it provides in configuration and use. Because it is very effective in predicting the target variable or selecting the most relevant features in a training dataset. The working logic of the method is that when the model is fully created, the feature importance levels are calculated by ordering and the least important features are eliminated during the model reconstruction at each stage of the method. (Kuhn and Johnson, 2013). That is, starting with all the features in the dataset, a subset of features is searched for by removing the desired number of features. There are two important choices made when using RFE. These preferences are, is the selection of the number of features to be used and the learning algorithm to be used to ensure the selection of these features. The performance of the method depends on the good configuration of the parameters to be used. For this reason, the parameters that give the best results are found by making changes in the RFE parameters in the feature selection process. In our study, SVM and Random Forest (RF) algorithms were preferred for prediction parameter selection (Guyon et al., 2002). In addition, feature deletion can be performed by applying cross validation with the RFE method (RFECV). In this way, the optimal number of features that should be used in learning algorithms can be determined in the method applied. However, since there may be excessive learning situations, data interpretation processes should be done together with evaluation criteria.

When evaluated for use in our study, RFE and RFECV are considered as methods with a high potential to provide a solution to the feature selection problem of CF projects. Because the preferred methods can help to select the right features by determining the importance ranking of the features. In addition, since they are also effective in high-dimensional data sets, the model performance can also be increased.

5.8. Machine learning algorithms and evaluation metrics

SVM, GNB, LR, RF and GBM algorithms were used in our study to determine the classification capabilities of feature subsets created by feature selection methods. SVM is a supervised ML algorithm that allows to find a boundary between the two classes that are furthest from any point in the training data. The vector space-based method is suitable for complex but small and medium-sized datasets (Akca, 2020). GNB is a probability-based supervised ML algorithm that can work with little data and can be used in real-time systems due to its speed. Binary classification, text classification and recommendation systems are common application areas with GNB. LR is an ML algorithm that is frequently used in binary classification problems and gives successful results. It enables the analysis to be performed statistically when there are one or more independent variables in the data set. RF is a popular ML algorithm that consists of a large number of decision trees, combines them at the decision stage, and makes classification or regression estimation for various problems (Sağbaş and Ballı, 2016). The solution of overfitting, which is one of the biggest problems in decision trees, can be provided with RF. An algorithm with a high classifier capacity can be obtained by adjusting the parameter settings when using RF according to the characteristics of the data set. It can also be used to rank the importance of factors in wrapper methods in feature selection. GBM, on the other hand, is similar to RF in its structure and works with decision trees (Aydın, 2018). Basically, it provides to obtain strong learners from weak learners obtained for each iteration. Therefore, its difference from RF is that the residual errors created by the previous decision tree are minimized. For this reason, it is considered one of the best classification algorithms. As a result, the above-mentioned algorithms have been preferred in the study because they provide the criteria for compatibility with binary classification problems, working in accordance with the characteristics of the data set, and providing high performance in solving the problem in question.

On the other hand, the accuracy value revealed by the ML algorithms mentioned above and evaluation criteria are used for the performance measurement of these accuracy values. Accuracy (5), recall (6), precision (7), F1 score (8) according to true positive (TP), true negative (TN), false positive (FP) and false negative (FN) values obtained according to

the confusion matrix, true positive rate (TPR) (10) and false positive rate (FPR) (11). TPR and FPR values are used to create the ROC curve with the AUC score. In addition, the formula (9) is used to find the Kappa value (KPP), where P_0 = accepted rate, P_c = expected rate. If the KPP value is above 0.4, it means that the agreement between the variables is sufficient.

$$A_{CC} = \frac{TP + TN}{TP + FP + FN + TN} \tag{5}$$

$$Rec = \frac{TP}{TP + FN} \tag{6}$$

$$Prec = \frac{TP}{TP + FP} \tag{7}$$

$$F1 = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \tag{8}$$

$$KPP\,(\kappa) = \frac{(P_0 - P_c)}{(1 - P_0)} \tag{9}$$

$$TPR = \frac{TP}{TP + FN} \tag{10}$$

$$FPR = \frac{FP}{FN + TP} \tag{11}$$

5.9. Dataset

Reward-based projects on the Fongogo CF platform are divided into 14 different categories and presented to supporters. Since the gender of the project owners is in the data set, the gender of the projects created by the associations/institutions or organizations is included in the data set as uncertain. Since it is emphasized in some studies in the literature that the past experience and gender of the project owner affect the success, how many projects the project creators have supported before, how many projects they have subscribed to, how many projects they have realized, how many project teams they are in, and the gender of the project creator have been added to the dataset for analysis. The location data where the project was carried out was created by taking the city information of the projects realized in 64 provinces in Turkey. The year in which the project was started is also included in the data set, along with the number of days it stayed in the campaign. Also, whether there is a promotional video, the length of the video in seconds, if there is, the total number of images in the gallery of the project, the number of questions and answers in the FAQ tab, the number of updates, the number of comments made on the project are included in the data set. Apart from this, the number of supporters, the number of rewards given by the project owner to the supporters, the number of people in the project team, whether there is a website belonging to the project, the number of social media belonging to the project, the number of followers in the social media are included in the data set. In addition, the number of tags created by the project owner and the number of words in the text in the project description are included in the dataset so that the project can be found more easily. The 1–2 sentence summary of the project that users encounter when they first open the project is included in the data set for word analysis that can be done. Finally, the target amount determined by the project owner to realize the campaign, the amount collected in support of this target, the rate of this support and whether the project was successful as a result of the campaign period are included in the data set of our study. The data set consists of 36 columns and 1075 rows in total, with all meta fields (project names, descriptions, etc.). Since the foundation of Fongogo, all projects on the CF platform have been added to the dataset (Table 1). Also, the dataset used was obtained from publicly available meta domains. In the study, the data that users can access through the relevant CF platform without any privacy or membership has been analyzed. In our study, to avoid the problem of over-learning by using all our data, missing data was supplemented with average values with known classes. Since the binary class values in the dataset are distributed as 30%-70%, the number of features was reduced to avoid high variance and low bias. In this way, a successful classification model was created by identifying the factors that affect the success rate the most. Since the number of projects in the Turkish CF ecosystem is limited, missing data were used in model building instead of removing them. Since there were very few missing features in the missing data used, imputation was performed in accordance with the literature.

Data Scraping and Collect

Data scraping and collection were primarily performed using the BeautifulSoup4 and Scrapy libraries in Python. In cases where errors are encountered, the "Instant Data Scraper" data scraping bot developed by webrobots.io and running on Google Chrome was also used. The word counts in the data set were collected with a separate Chrome extension, "Word Counter". Meta fields such as the length in seconds of the promotional video, for which data scraping methods did not work, and the number of social media followers were collected manually and included in the data set.

Data Preprocessing

Filling (Imputation) and deletion methods are used for missing or missing data in the data set. Data scaling method was used to give better results of prediction algorithms and to avoid overfitting. The data set we created with 36 meta fields consists of 7 categorical and 20 numeric-continuous features. The remaining 9 features consist of data containing text. For this reason, the results were analyzed using 27 success factors in our study. In this direction, 7 categorical features were converted into numerical form and made ready for analysis before feature selection processes.

Missing Data Completion

The rate of missing/missing data in the data set is 4.83%. In the literature, the rate of missing data is expected to be 15% or more to exclude any variable from the data set. (AnalyticHouse, 2019). For this reason, the averaging method, which is often preferred in general models, has been preferred for numerical columns. Categorical data were filled with the constant value assignment method, with the highest frequency variable being found.

Data Scaling

It is necessary to normalize the distribution in the data set in order to avoid over-learning and the algorithms used in performing the estimation and feature selection operations to give better results. For this reason, reducing each feature to equal intervals standardizes the data set by converting it to a normal distribution (Kılınç et al., 2020).

$$X'_i = \frac{X_i - \min(X)}{\max(X) - \min(X)} \tag{12}$$

For scaling carried out in this direction, the min–max scaler method of Python's Sklearn Library was used (12).

Findings

For the feature selection for the research question put forward within the scope of the study, Pearson correlation was first used to measure the positive or negative relationship between the variables in the data set and shown in matrix form (Table 2). Pearson correlation was preferred because the variables conformed to the normal distribution. Also, the features used for success prediction; the definition table consisting of values such as mean, standard deviation, minimum, maximum, data type and number of data is also included in the study (Table 3).

The coefficient obtained when interpreting the correlation matrix has a value between −1 and 1. As the coefficient approaches 1, the linear relationship between the two variables increases, while as it approaches −1, an inverse linear relationship increases. As it approaches 0, there is no relationship between the two variables. The larger the absolute value of the Pearson coefficient, the stronger the relationship. Coefficients with absolute values in the range of $r > 0.1$ / $r < 0.3$ and $r < -0.1$ / $r > -0.3$ indicate a very low relationship, coefficients in the range of $r > 0.3$ /

$r < 0.5$ and $r < -0.3$ / $r > -0.5$ indicate a moderately strong relationship, and coefficients in the range of $r > 0.5$ / $r < 1.0$ and $r < -0.5$ / $r > -1.0$ indicate a strong relationship (Hair et al., 2014; Xiang et al., 2021). Accordingly, the prominent variable relationships according to Table 2 are summarized as follows;

- The number of projects supported by the project owner and the number of projects created are positively related to the number of campaign supporters.
- There is a positive relationship between the number of campaign supporters and the number of projects supported by the project owner, the number of updates and the number of comments.
- Project success is positively related to the number of projects supported by the project owner, the number of updates to the project, and the number of supporters. Also, there is a high correlation between the number of project supporters and the number of projects supported by the project creator. That is, as the creator of the project supports other projects, the support rate of his own project increases. However, the positive relationship with the project success status and the number of supporters has no semantic value. The number of supporters cannot be used in the success classification of projects. Because together with the collected amount, the number of supporters is determined exactly at the end of the project.
- There is a relationship between funding type and success. According to the PCC, for projects funded as "all or nothing" funding, the type of funding has an impact on backers' funding of the project.
- There is a relationship between location data and success for 64 different provinces in Turkey. According to the data set, it is seen that the CF projects in the provinces of the Marmara Region are more successful. In the projects initiated in the Eastern Anatolia Region as a location, there is a failure situation. However, due to the limited number of data, a definite interpretation about the location cannot be made. Because location data is unevenly distributed in 64 provinces.
- The start year of the projects is negatively related to the number of promotional videos and images, and positively related to the number of tags. Therefore, the number of video-visual materials in the projects presented recently decreased, while the number of tags increased.
- Categories have a correlation with project success. In our study, the failure of the projects in the technology category and the success of the film-photography category stand out.
- There is a positive relationship between the number of content words and the number of rewards. Also, the number of team members is also related to the content word count. It has been observed that the text content in the project increases as the number of people in the team increases.
- There is a positive relationship between the number of social media, website status, funding type, number of images, number of team members and content words of the project.
- The number of rewards is negatively related to the funding method and positively related to the number of images. In other words, while the number of rewards is high in the "all or nothing" model, the number of rewards decreases in the "all or nothing" model, and the number of images increases as the number of rewards increases.
- The past experiences of the project creator are positively related to the success status. In this context, on the Fongogo CF platform, how many projects the project creator has previously supported, how many projects he has subscribed to, how many projects he owns and how many project teams he is in are taken into consideration by the supporters. Also, the gender of the project creator is also related to the success status.

In the table created by PCC (Table 2), some of the factors that positively or negatively affect the success of CF are shown below (Fig. 2).

In this direction, the 5 features that most affect the success of CF were selected by PCC among the filtering methods. The selected features were determined as the location of the project, the total number of projects on the CF platform of the project creator, the number of updates of the project, the number of projects supported by the project owner and the promotional video (Table 4).

In MI and CHI, which are KBest applicable feature selection algorithms, feature selection results were analyzed to take $k = 5$ intuitively.

According to MI, the factors affecting the success the most in the CF dataset were the category of the CF project, the number of projects supported by the project creator, the length of the promotional video in seconds, the number of updates of the project and the number of people in the project team. According to CHI, the selected features were found to be the number of the project team with the project creator, the number of tags of the project, the number of words in the content in the project meta field, and the length of the promotional video in seconds (Table 4).

Finally, among the filtering methods, the order of importance of the features in the feature selection made with the FS method was found and added to the study (Fig. 3). According to FS, the most influential factors on CF success were found to be the number of projects supported by the project creator, the amount of target set for funding the project, the length of the promotional video in seconds, the number of social media followers of the project and the number of updates.

In wrapper methods, the RFE method assigns weights to each feature using one of the classification algorithms. Accordingly, the features with the smallest absolute weights are removed from the data set. This process is repeated until the desired number of features is obtained. The data set we created was analyzed by selecting 5 features in the RFE method, as in other filtering methods. According to the results of the RFE, the features that most affect the success of CF are the number of projects supported by the project owner, the length of the video, the number of social media followers, the number of content words and the target amount. Finally, the RFECV method is considered as a recursive feature elimination method that allows the number of selected features to be optimally adjusted by cross validation. In this direction, firstly, the cross-validation scores of the features were determined. According to the cross-validation scores, the optimal result can be obtained with 8 features (Fig. 4).

The 8 most important features selected by RFECV are shown in the overall table (Table 4). In order to understand the validity of the selected features, classification analysis was performed with ML methods. In this context, the features selected by filter-based and wrapper-based methods were classified with LR, SVM, GNB, RF and GBM algorithms and tested with ACC, F1, Prec, Rec, AUC and KPP performance evaluation criteria (Table 5). The specified test process covers the feature selection flow required for an ideal classification. The raw data obtained by flow scraping techniques, together with the distinctive features of their features, were pre-processed, prepared for classification algorithms and enabled the establishment of prediction models (Fig. 5). The prediction models created were compared and analyzed according to the algorithm used. (Table 5).

According to the analysis results, the GBM algorithm gave the best binary classification result for the filtering and wrapper-based methods used in all methods (Table 5).

In this direction, GBM reveals the best accuracy value with 80.18% when classification is made with the features selected with PCC. However, when looking at other algorithms, it is seen that the classification capacity is close to each other. At this point, the model evaluation criteria were decisive for the selection (Table 5a). When the ML algorithms used according to the features selected with MI are compared, GNB gives the highest accuracy score with 80.18%. However, the F1, Prec, Rec, AUC and KPP scores obtained when tested with the evaluation criteria show that the GBM algorithm provides a better classification with 80.07% (Table 5b). When classification is made according to the features selected with CHI, GBM provides the highest accuracy score with 78.94 %. When tested with evaluation metrics, the scores generated

Table 2
Correlation matrix of success variables.

Variables	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
1 Project Category	1.00																										
2 Funding Type	−0.020	1.00																									
3 PO Gender	−0.067	−0.042	1.00																								
4 PO Supported Projects	−0.066	−0.016	0.112	1.00																							
5 PO Subscribed Projects	−0.009	−0.010	0.048	0.197	1.00																						
6 PO Project Count	0.025	−0.016	0.045	0.359	−0.008	1.00																					
7 PO Team Count	−0.044	0.005	0.040	0.168	−0.004	0.053	1.00																				
8 Location	0.087	−0.044	−0.083	−0.038	−0.018	−0.067	−0.021	1.00																			
9 Started Date	0.057	0.098	−0.006	−0.112	−0.055	−0.040	−0.074	0.017	1.00																		
10 Campaign Day	0.042	−0.034	−0.018	−0.021	−0.013	−0.060	−0.041	0.079	−0.034	1.00																	
11 Promotional Video	0.008	−0.079	−0.002	−0.101	−0.019	−0.026	0.015	0.043	−0.238	0.001	1.00																
12 Video Lenght	−0.055	0.051	0.008	0.056	−0.002	0.029	−0.005	0.005	0.010	0.031	0.453	1.00															
13 Images Count	−0.029	0.024	0.021	0.074	0.020	0.004	0.039	−0.014	−0.233	0.043	−0.057	0.069	1.00														
14 FAQ Count	0.041	−0.007	−0.054	−0.024	−0.005	−0.026	−0.010	0.013	0.144	0.010	−0.034	0.039	0.008	1.00													
15 Updates Count	−0.049	−0.024	0.010	0.140	0.083	0.094	0.104	−0.034	−0.107	0.002	−0.064	0.049	0.664	0.005	1.00												
16 Comments Count	0.012	−0.008	−0.014	−0.009	−0.001	−0.010	−0.005	−0.020	0.038	−0.005	−0.024	0.007	0.002	0.085	0.018	1.00											
17 Supporters Count	−0.075	0.039	0.049	0.327	0.107	0.150	0.072	−0.113	−0.011	−0.018	−0.172	0.072	0.145	0.040	0.243	0.507	1.00										
18 Reward Count	−0.104	−0.188	0.051	0.117	0.038	0.003	0.060	−0.012	−0.022	0.036	0.012	0.001	0.131	−0.030	0.102	−0.026	0.156	1.00									
19 Team Member	−0.065	−0.011	0.010	0.009	0	0.121	0.063	−0.006	0.004	0.030	−0.065	0.015	0.104	−0.015	0.046	−0.004	0.076	0.111	1.00								
20 Web Site	0.092	0.113	0.003	0.070	0.044	0.059	0.018	−0.013	−0.052	0.004	−0.078	0.122	0.152	0.053	0.101	−0.019	0.102	0.069	0.048	1.00							
21 Social Media	0.003	0.169	0.001	0.072	0.016	0.097	−0.001	−0.064	0.041	−0.007	0.182	0.125	0.160	0.027	0.086	−0.011	0.117	0.079	0.156	0.442	1.00						
22 Followers	0.002	0.136	−0.080	0.002	−0.004	−0.014	−0.007	−0.063	−0.018	−0.014	−0.046	0.021	0.031	−0.007	0.016	0.028	0.178	0.031	−0.031	0.144	0.141	1.00					
23 Tags Count	0.024	−0.044	0.031	−0.030	−0.029	0.008	0.036	−0.022	0.241	−0.001	−0.066	0.025	−0.038	−0.002	−0.008	−0.030	−0.022	0.110	0.118	0.111	0.146	−0.009	1.00				
24 Content Word	−0.077	−0.138	0.104	0.025	0.003	0.051	0.022	−0.045	0.015	−0.013	−0.069	0.108	0.142	0.028	0.074	−0.032	0.076	0.260	0.141	0.101	0.170	−0.005	0.089	1.00			
25 Target Amount	−0.010	−0.003	−0.008	−0.007	0	−0.009	−0.022	0.032	−0.054	0.011	0.047	−0.021	0.079	−0.001	−0.005	0	−0.005	−0.008	−0.016	−0.019	−0.032	−0.002	−0.030	0.033	1.00		
26 Pledged	−0.049	0.115	−0.051	−0.016	−0.003	−0.022	−0.008	−0.037	0.062	−0.008	−0.034	0.001	0.024	0.064	−0.014	−0.001	0.038	0.038	−0.016	−0.027	0.020	−0.010	−0.014	0.023	−0.001	1.00	
27 State	0.240	0.196	0.135	0.275	0.051	0.169	0.090	0.264	−0.032	−0.048	0.280	0.096	0.111	−0.049	0.207	0.056	0.472	0.141	0.129	0.107	0.136	0.102	0.038	0.086	0.119	−0.048	1.00

PO: Project Owner

Table 3
Descriptive statistics of success variables.

Identifiers	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
Average	4.941	0.096	1.156	0.667	0.009	1.225	0.026	7.42	2018	47.19	0.27	84.89	6.212	0.223	0.581	0.094	25.16	6.6	2.010	0.328	1.371	3.770	2.145	412.4	23.175	6.128	0.735
Standard deviation	3.123	0.2957	0.542	2.685	0.03	0.850	0.2	12.94	1.83	17.20	0.444	113.6	8.934	1.471	2.716	2.367	60.58	1.937	2.035	0.469	1.219	23.582	2.291	283.4	38.860	16.285	0.441
Minimum	0	0	0	0	0	1	0	0	2013	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0
Maximum	13	1	2	31	1	7	4	63	2021	118	1	1651	221	30	71	77	976	10	28	1	4	1449	12	3473	600.000	264.150	1
Data Type	int	bool	int	int	int	int	int	int	int	int	bool	int	int	int	int	int	int	int	int	bool	int	int	int	int	float	float	bool
Number of Data	1075	1075	1075	1075	1075	1075	1075	1075	1075	1075	1075	1075	1075	1075	1075	1075	1075	1075	1075	1075	1075	1075	1075	1075	1075	1075	1075

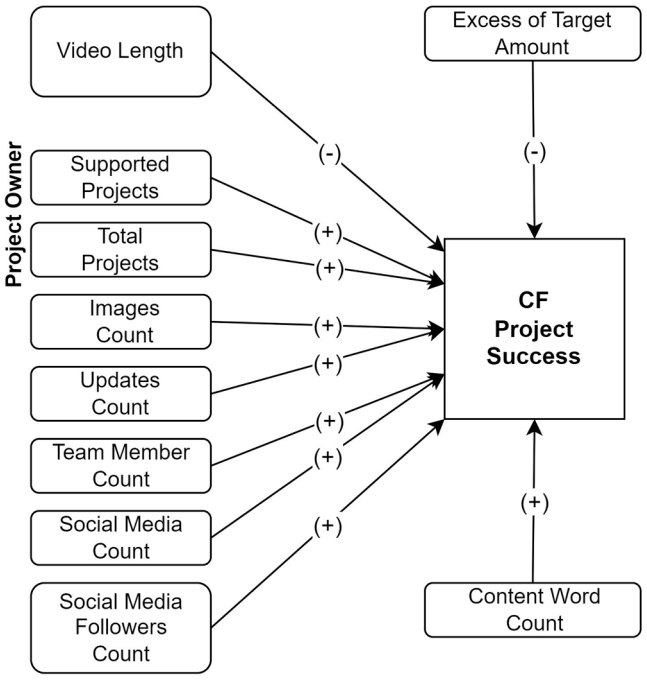


Fig. 2. Non-categorical factors affecting project success according to PCC.

with the confusion matrix confirm this (Table 5c). When a classification was made according to the selected features with FS, which is the last of the filtering methods, GBM obtained the highest accuracy score of 81.42%. Performance evaluation criteria show that the resulting score can classify more consistently than other ML algorithms (Table 5d).

In the wrapper methods, the experimental classification results with the features selected by RFE indicate that the GBM algorithm outperforms LR, SVM, GNB, RF with an accuracy score of 80.80%. The evaluation criteria obtained confirm this (Table 5e). Finally, the features selected with the RFECV method applied with cross validation reveal the highest accuracy rate in the study with 84.28%. The F1, Prec, Rec and AUC scores obtained with the complexity matrix for RFECV show that the classification performance is high. The KPP score of 0.55 explains that there is a significant agreement between the variables (Table 5f).

The results were also tested with the ROC curve, which shows the performance of the ML algorithms (Fig. 6). According to the ROC, the higher the area under the curve (AUC) obtained from the false positive rate to true positive rate, the better the classification and the better the model performance. In this direction, the AUC score obtained in the classification process with the 8 features optimally proposed by the RFECV method is at the highest level with 0.85. However, when the results are examined, it is seen that the criteria obtained with RF are close to GBM. For this reason, log loss is also included in the evaluation metrics obtained with the RFECV features, where the highest results are obtained. Logarithmic loss is a metric used in classification models that shows how close a prediction probability can be to the truth. In comparison with RFECV features in terms of logarithmic loss; The RF had a value of 0.51, while the GBM showed a better result with a value of 0.38. When the results are interpreted, GBM applied with RFECV was accepted as the most successful model in the study with an accuracy rate of 84.28%, since the lower logarithmic loss means better classification performance.

Therefore, the findings of our study reduce 36 success factors to 8 with RFECV in order to predict CF success. The findings obtained in this context suggest the category of the project, the number of projects supported by the project creator, the number of days of the campaign, the length of the promotional video, the number of images, the number of social media followers, the number of content words and the target

Table 4
Features selected by methods.

Method		Selected Features
Filter Based	PCC	Location, total number of projects on the platform (PC), number of updates, number of supported projects (PC), promotional video
	MI	Category, number of supported projects (PC), length of promotional video, number of updates, number of people on the project team
	CHI	Number of project teams (PC), number of tags, amount of targets, number of images, number of supported projects (PC)
	FS	Target amount, number of projects supported (PC), number of social media followers, number of updates, length of promotional video
Wrapper Based	RFE	The number of projects supported by the project owner, the length of the promotional video, the number of social media followers, the number of content words, the target amount
	RFECV	Category, number of supported projects (PC), number of days of campaign, length of promotional video, number of images, number of social media followers, number of content words, target amount

PC: Project Creator.

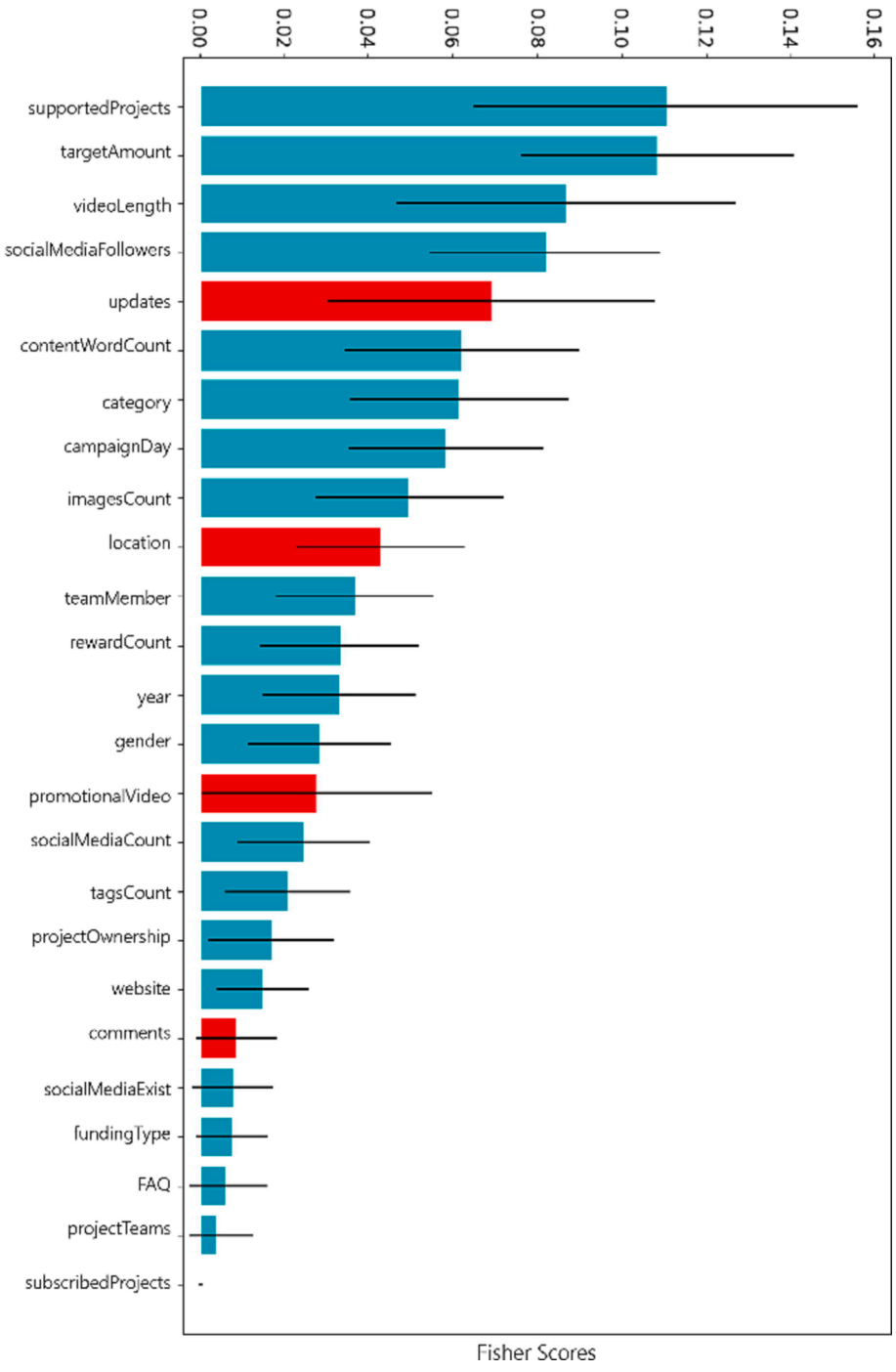


Fig. 3. Feature importances according to FS.

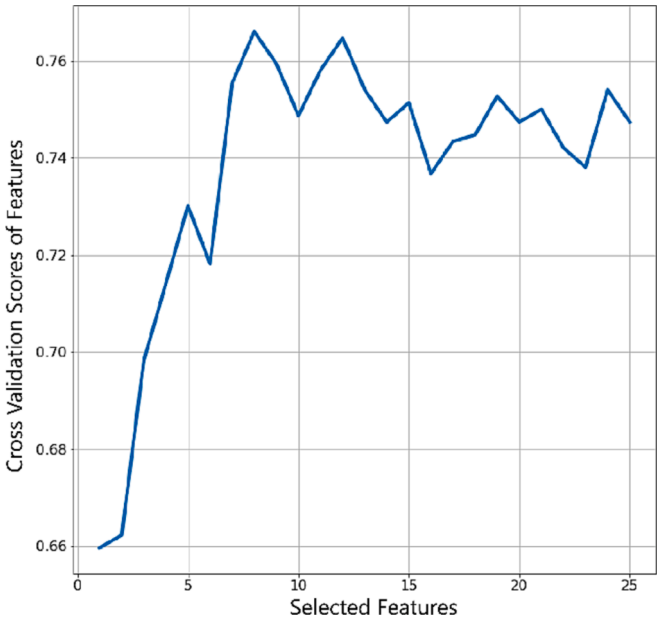


Fig. 4. Cross validation scores of features according to RFECV.

amount for classification with ML methods. Within the scope of the findings of the research question in our study, it was ensured to reveal the quality signal in CF projects and to determine the factors affecting project success. When the findings are evaluated within the signaling theory, it is revealed that the success factors and project quality signals obtained as a result of feature selection directly affect project success in the Fongogo CF ecosystem in Turkey. For this reason, the features obtained are used for success prediction with machine learning algorithms, resulting in models with high success rates that can be accessed by end users.

6. Discussion and conclusion

Feature selection processes begin with data collection. Because in classification-based approaches, it is recommended that the number of data belonging to each class be equal in order to avoid overfitting (Ryoba et al., 2021). The class balance in the data set in our study is approximately 30%-70%. The main reason for this situation is the small number of data in the CF platform we analyzed. Data deletion can be done easily in order to maintain balance in the data sets of CF platforms with thousands of projects such as Kickstarter and Indiegogo. However, the number of CF projects in Turkey is not enough to delete data. For this reason, all the data collected within the scope of our study were used and the missing data with a certain class were completed with average values in order to avoid overfitting. In this context, the number of features was reduced for the case of high variance low bias, which causes overfitting, and 8 factors that affect the success the most and ensure successful classification were identified. It is also seen in other examples in the literature that the amount of data in the dataset used in the study is sufficient for machine learning and other analytical stages (Balki et al., 2019). However, in general, the limitations of the study can be listed as not evaluating the quality of text, images or videos, not evaluating the rhetorical style of the project (emotional and utilitarian/logical cues), use of humor or irony, etc., and focusing only on reward-based CF platforms in Turkey. In this context, a structure for future studies is provided.

On the other hand, similar results were obtained with our study in the research of Jhaveri in 2019. (Jhaveri et al., 2019). In the study, it was stated that the factors that most affect the success in CF platforms are the number of campaign days, the target amount, the category of the project, the location and the media content. Only the location factor does not match the results of our study. The main reason for this is that the projects in the CF ecosystem in Turkey show an uneven distribution in terms of location. This situation has the potential to change over time. In addition, 84% success rate was obtained with the ML algorithms in the study. The boosting algorithms used provide the opportunity to obtain good results as in our research. In our study, it is also seen that unlike the research of Koch and Siering, the past experiences of the project creator affect the success of the CF project, just like Borrero's research in Spain (Koch and Siering, 2015; Borrero-Domínguez et al., 2020). In particular, the number of previously supported projects on the CF platform is one of the factors that most affect the success of CF.

As seen in the examples in the literature, the support and investment behaviors of users in the CF ecosystem vary regionally, economically, legally, culturally and sociologically, and the factors affecting success also vary. Therefore, the features to be selected in order to measure the success of the CF in advance and to create a decision support system according to the shortcomings of the project have the potential to vary according to the applications to be developed. Below we list the conditions that cause regional differences in the factors affecting the success of the CF.

Table 5
Comparison of ML algorithms using evaluation metrics.

(a) Experimental PCC results by dataset

Alg.	ACC	F1	Prec	Rec	AUC	KPP
LR	79.87	87.61	80.13	96.63	84.82	0.36
SVM	79.87	87.66	79.93	97.05	78.86	0.35
GNB	78.94	87.12	79.31	96.63	82.31	0.32
RF	79.56	87.10	81.38	93.69	81.15	0.39
GBM	**80.18**	**87.59**	**81.29**	**94.95**	**81.95**	**0.39**

(b) Experimental MI results by dataset

Alg.	ACC	F1	Prec	Rec	AUC	KPP
LR	78.01	86.62	78.49	96.58	78.15	0.28
SVM	79.25	87.56	78.40	97.15	75.82	0.29
GNB	80.18	87.83	80.20	96.05	79.14	0.36
RF	76.47	84.42	82.40	86.55	77.86	0.36
GBM	**80.07**	**87.59**	**81.29**	**94.95**	**83.38**	**0.39**

(c) Experimental CHI results by dataset

Alg.	ACC	F1	Prec	Rec	AUC	KPP
LR	76.16	85.66	76.92	96.63	73.10	0.20
SVM	77.08	86.19	77.51	96.05	70.64	0.23
GNB	75.23	84.96	76.87	94.95	71.80	0.18
RF	74.61	83.46	80.23	86.97	73.06	0.29
GBM	**78.94**	**86.92**	**80.14**	**94.01**	**77.60**	**0.34**

(d) Experimental FS results by dataset

Alg.	ACC	F1	Prec	Rec	AUC	KPP
LR	77.39	86.30	77.96	96.63	77.91	0.25
SVM	78.94	87.12	79.31	96.63	74.94	0.32
GNB	77.39	86.14	78.54	95.37	76.68	0.27
RF	81.11	87.52	85.25	89.91	80.41	0.48
GBM	**81.42**	**88.14**	**83.20**	**93.69**	**80.70**	**0.45**

(e) Experimental RFE results by dataset

Alg.	ACC	F1	Prec	Rec	AUC	KPP
LR	77.39	86.35	77.76	97.05	75.62	0.24
SVM	77.08	86.29	77.15	97.89	74.23	0.22
GNB	76.16	85.55	77.28	95.79	73.94	0.21
RF	79.56	86.74	83.07	90.75	77.86	0.42
GBM	**80.80**	**87.93**	**81.88**	**94.58**	**80.20**	**0.42**

(f) Experimental RFECV results by dataset

Alg.	ACC	F1	Prec	Rec	AUC	KPP
LR	79.25	87.23	79.78	96.21	82.81	0.34
SVM	80.84	88.12	80.98	96.63	77.19	0.40
GNB	80.18	87.64	81.07	95.37	80.20	0.39
RF	83.59	89.33	85.71	93.27	84.32	0.54
GBM	**84.28**	**89.11**	**85.65**	**91.85**	**84.53**	**0.55**

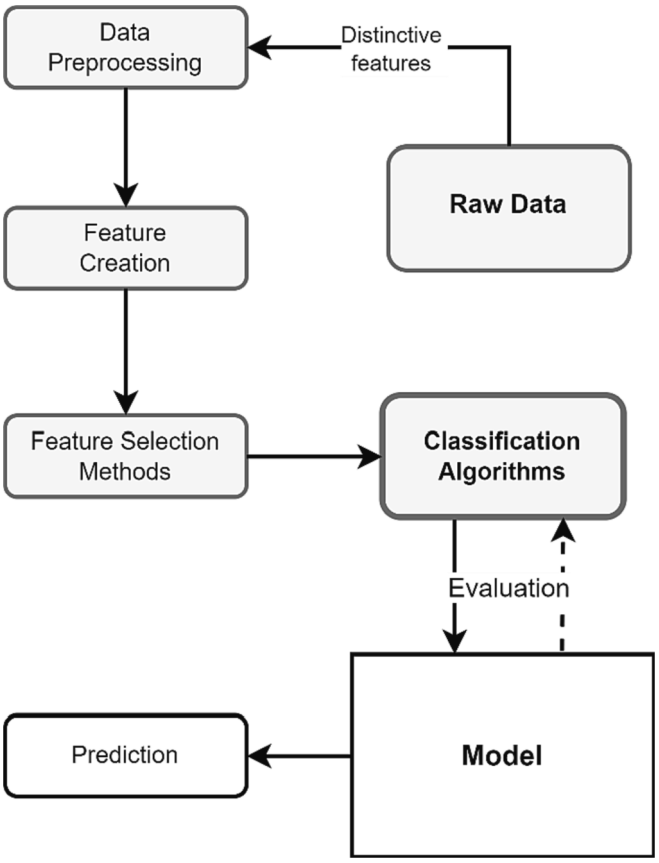


Fig. 5. General feature selection flow for ideal classification.

- Economic conditions: Economic differences between countries can affect the success of CF. In a wealthier country, people have more spending power and are more likely to spend more money on such projects.
- Cultural differences: Cultural differences can change people's inclination to contribute to CF projects. In some countries, there is more solidarity and community spirit in society, while in others individualism is more dominant. Therefore, a successful CF campaign in one country may not be as successful in other countries. Also, in some cultures it may be less common to donate money or support a foreign project.
- Legal regulations: CF may be subject to different legal regulations in many countries. In some countries, crowdfunding is prohibited or restricted, while in others it is less regulated. Therefore, a crowdfunding campaign in one country may be successful, while in another it may fail.
- Marketing strategies: The success of CF campaigns can depend on the right marketing strategies. People in different countries may respond to different marketing techniques. Therefore, marketing strategies used in one country may be successful, while in another country the same strategies may fail. However, online CF platforms have the advantage of being able to reach people in different countries by providing global reach. However, the use of platforms can also vary across countries. In some countries, internet access is not widespread, while in others online platforms are more popular.
- The nature of the projects: The nature of CF projects can influence people's inclination to engage and contribute. For example, in some countries there may be more interest in issues such as the environment, social justice or civil rights, while in others there may be more interest in technological innovation or education projects. Therefore, a successful CF project in one country may not have the same success in another.

With the situations listed above, it is necessary to determine the success factors of the CF ecosystems offered to users in countries and to

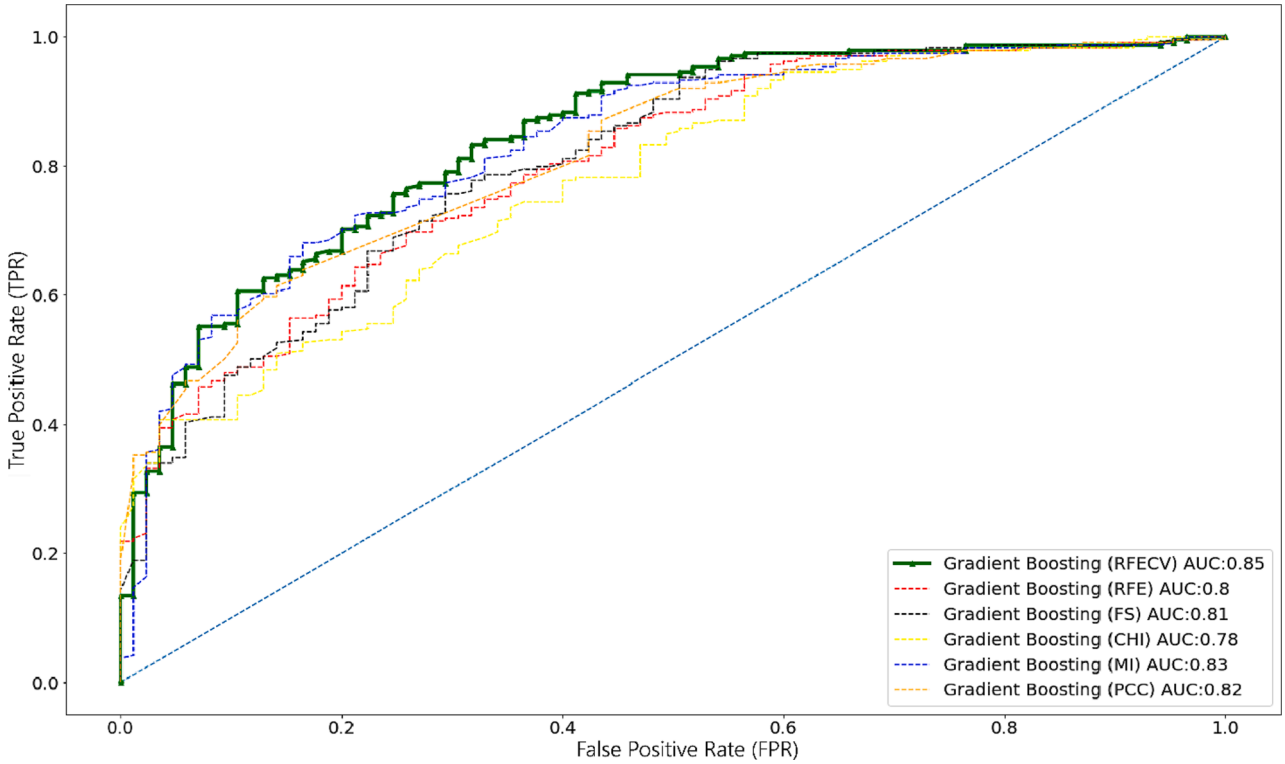


Fig. 6. Display of classification results on the ROC curve.

create decision support systems by activating success prediction tools in accordance with the relevant factors.

As a result, our study is the first study in which the features that affect the success of CF platforms in Turkey are determined and tested. In this direction, after the data set was prepared with the projects created in Turkey, filtering and wrapping methods were used. Afterwards, the features that provide the best success classification were selected and tested with ML algorithms. The results show that the features selected with RFECV (the category of the project, the number of projects supported by the project creator, the number of days of the campaign, the length of the promotional video, the number of images, the number of social media followers, the number of content words and the target amount) make the best classification. With the application of the GBM algorithm, an accuracy rate of 84.28% was obtained in the classification. In other words, the findings show that the features selected with RFECV are of great importance for the success of a project. Therefore, users who create projects on CF platforms should create these features more carefully in their projects. Also, as the number of data on CF platforms in Turkey increases, the features that most affect success may change in the future. For this reason, feature selection should be done dynamically as the data set expands.

As a result, feature selection for CF projects provides many benefits for the active use of the data set. Because feature selection directly affects the results of machine learning algorithms in success prediction processes and supports the establishment of a better structure. In this direction, feature selection firstly provides to avoid overfitting in the data set. It is seen that the test data cannot be classified well as a result of overfitting, especially in data sets with many features. Secondly, the data set often contains features that will not affect the relevant research. Variables that are especially related to names can be given as an example. These variables cause a low-quality input as well as a low-quality output. Also, a large number of features may cause performance degradation at the point of application development, as it will cause the established model to run slowly and time-consuming. Finally, the data set should be simple and explainable. Therefore, it loses its explainability when the number of features in a dataset is too large. For these reasons, feature selection should be applied meticulously in multidimensional data sets.

This study explains the key factors affecting the success of crowdfunding platforms operating in Turkey and supports them with empirical data based on machine learning. Understanding the factors affecting the success of these platforms is of great importance for entrepreneurs, investors and policymakers. The factors derived based on feature selection factors are central to the study in the context of the growth of the crowdfunding ecosystem in Turkey. Thus, it aims to contribute to the development of data-driven strategies that will shape the future development of CF platforms in Turkey. Based on the results of this research, entrepreneurs, investors and platform operators can take steps to build a more effective and sustainable CF ecosystem.

In future works, the revealed structure can be further strengthened by increasing the number of data by including other CF platforms. Also, embedded and hybrid methods for feature selection methods can be tried for the data set and the results can be compared. Thus, the success factors suggested by the methods used will be able to create a prediction structure with machine learning and artificial neural network methods. When this structure is presented to CF project creators on a web application, it has the potential to gain decision support for project success and increase project success.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Murat Kilinc reports financial support was provided by Scientific and Technological Research Council of Turkey.

Data availability

Data will be made available on request.

Acknowledgement

The study was supported by the project number 121E363 within the scope of the TUBITAK (The Scientific and Technological Research Council of Turkey) 1002 project. Also, the related study is derived from the doctoral thesis named “Designing and Researching Effects Business Intelligence Environment for Crowdfunding Projects with Machine Learning Methods.”

References

AnalyticHouse. 2019. “Eksik Değer Analizleri.” Retrieved March 23, 2021 (https://www.analytichouse.com/eksik-deger-analizleri/).

Alazazi, M., Wang, B., Allan, T.Z., 2020. Success Factors of Donation-Based Crowdfunding Campaigns. A Machine Learning Approach.

Atsan, N., Erdoğan, E.O., 2015. Girişimciler İçin Alternatif Bir Finansman Yöntemi: Kitlesel Fonlama (Crowdfunding). Eskişehir Osmangazi Üniversitesi İktisadi Ve İdari Bilimler Dergisi 10 (1), 297–320.

Aydın, C., 2018. Makine Öğrenmesi Algoritmaları Kullanılarak İtfaiye İstasyonu İhtiyacının Sınıflandırılması. European Journal of Science and Technology 169–175. https://doi.org/10.31590/ejosat.458613.

Baber, H., 2021. Secrets of Successful Crowdfunding Campaigns. Adv. Busin.-Relat. Sci. Res. J. 11 (2).

Balki, I., Amirabadi, A., Levman, J., Martel, A.L., Emersic, Z., Meden, B., Garcia-Pedrero, A., Ramirez, S.C., Kong, D., Moody, A.R., Tyrrell, P.N., 2019. Sample-Size Determination Methodologies for Machine Learning in Medical Imaging Research: A Systematic Review. Can. Assoc. Radiol. J. 70 (4), 344–353.

Borrero-Domínguez, C., Cordón-Lagares, E., Hernández-Garrido, R., 2020. Analysis of success factors in crowdfunding projects based on rewards: A way to obtain financing for socially committed projects. Heliyon 6 (4), e03744.

Chen, Y., Zhou, S., Jin, W., Chen, S., 2023. Investigating the Determinants of Medical Crowdfunding Performance: A Signaling Theory Perspective. Internet Res. 33 (3), 1134–1156, 71701180.

Çonkar, M.K., Canbaz, M.F., 2018. Kitle Fonlaması Finansman Yöntemi: Türkiye'de Sistemin Geliştirilmesine Yönelik Öneriler. Optimum Ekonomi Ve Yönetim Bilimleri Dergisi 5 (2), 119–132.

Çubukçu, C., 2017. Kitlesel Fonlama : Türkiye'deki Kitlesel Fonlama Platformları Üzerinden Bir Değerlendirme. Girişimcilik Ve İnovasyon Yönetimi Dergisi 6 (2), 155–172.

Etter, V., Grossglauser, M., Thiran, P., 2013. Launch Hard or Go Home! Predicting the Success of Kickstarter Campaigns. In: COSN 2013 - Proceedings of the 2013 Conference on Online Social Networks 177–82. https://doi.org/10.1145/2512938.2512957.

Fanea-Ivanovici, M., Baber, H., 2021. Predicting Entrepreneurial and Crowdfunding Intentions? A Study of Romania and South Korea. AMFITEATRU Econ. J. 23 (Special15), 1003.

Akca, M. Fatih, 2020. “Nedir Bu Destek Vektör Makineleri?” Retrieved May 9, 2021 (https://medium.com/deep-learning-turkiye/nedir-bu-destek-vektör-makineleri-makine-öğrenmesi-serisi-2-94e576e4223e).

Crowdcrux, 2020. “Crowdfunding Statistics in 2020.” Retrieved March 1, 2021 (https://www.crowdcrux.com/crowdfunding-statistics-in-2020/).

Greenberg, M.D., Pardo B., Hariharan, K., Gerber, E., 2013. “Crowdfunding Support Tools: Predicting Success & Failure.” Pp. 1815–20 in *CHI '13 Extended Abstracts on Human Factors in Computing Systems*.

Gu, Q., Li, Z., Han, J., 2012. Generalized Fisher Score for Feature Selection ArXiv Preprint 2012 ArXiv:1202.3725.

Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene Selection for Cancer Classification Using Support Vector Machines. Mach. Learn. 46, 389–422.

Hair, J., Black, W., Babin, B., Anderson, R. (Eds.), 2014. Multivariate Data Analysis. Pearson Education.

Jhaveri, S., Khedkar, I., Kantharia, Y., Jaswal, S., 2019. Success Prediction Using Random Forest, Catboost, Xgboost and Adaboost for Kickstarter Campaigns. In: Proceedings of the 3rd International Conference on Computing Methodologies and Communication, ICCMC 2019. Institute of Electrical and Electronics Engineers Inc, pp. 1170–1173.

Kaynar, O., Arslan, H., Görmez, Y., Işik, Y.E., 2018. Makine Öğrenmesi ve Öznitelik Seçim Yöntemleriyle Saldırı TespitiIntrusion Detection with Machine Learning and Feature Selection Methods. Bilişim Teknolojileri Dergisi 11 (2), 175–185.

Kılınç, M., Tarhan, Ç., Aydın, C., 2020. Could mobile applications' success be increased via machine learning and business intelligence methods? Eur. J. Sci. Technol. 2020 (20), 805–814. https://doi.org/10.31590/ejosat.793069.

Kocakafa, T., 2020. “Özellik Oluşumu ve Özellik Seçimi.” *Veri Bilimi Okulu*. Retrieved May 9, 2021 (https://www.veribilimiokulu.com/ozellik-olusumu-ve-ozellik-secimi feature-selection-1/).

Koch, J.-A., Siering, M., 2015. Crowdfunding success factors: The characteristics of successfully. Eur. Conf. Inform. Syst. Proc. 1–15.

Kou, G., Yang, P., Peng, Y.i., Xiao, F., Chen, Y., Alsaadi, F.E., 2020. Evaluation of Feature Selection Methods for Text Classification with Small Datasets Using Multiple Criteria

Decision-Making Methods. Appl. Soft Comput. J. 86, 105836 https://doi.org/10.1016/j.asoc.2019.105836.
Kromidha, E., Robson, P., 2016. Social identity and signalling success factors in online crowdfunding. Entrep. Reg. Dev. 28 (9–10), 605–629. https://doi.org/10.1080/08985626.2016.1198425.
Kuhn, M., Johnson, K. (Eds.), 2013. Applied Predictive Modeling. Springer New York, New York, NY.
Li, Y., Rakesh, V., Reddy, C.K., 2016. Project Success Prediction in Crowdfunding Environments. In: WSDM 2016 - Proceedings of the 9th ACM International Conference on Web Search and Data Mining. Association for Computing Machinery, Inc, New York, NY, USA, pp. 247–256.
Kuppuswamy, V., Bayus, B.L., 2013. Crowdfunding Creative Ideas: The Dynamics of Project. Backers in Kickstarter.
Liu, Y., Yong, M., Chen, K., Li, Y., Guo, J., 2020. Daily activity feature selection in smart homes based on pearson correlation coefficient. Neural Process. Lett. 51 (2), 1771–1787. https://doi.org/10.1007/s11063-019-10185-8.
Lukkarinen, A., Teich, J.E., Wallenius, H., Wallenius, J., 2016. Success drivers of online equity crowdfunding campaigns. Decis. Support Syst. 87, 26–38. https://doi.org/10.1016/j.dss.2016.04.006.
Marono, N.S., Betanzos, A.A., Sanroman, M.T., 2007. Filter methods for feature selection. A comparative study?. In: Intelligent Data Engineering and Automated Learning - IDEAL 2007, pp. 790–799. Vol. 4881.
Mehanović, D., Kevrić, J., 2020. Phishing website detection using machine learning classifiers optimized by feature selection. Traitement Du Signal 37 (4), 563–659. https://doi.org/10.18280/TS.370403.
Misra, P., Yadav, A.S., 2020. Improving the classification accuracy using recursive feature elimination with cross-validation. Internat. J. Emerg. Technol. 11 (3), 659–665.
Mollick, E., 2014. The dynamics of crowdfunding: an exploratory study. J. Bus. Ventur. 29 (1), 1–16. https://doi.org/10.1016/j.jbusvent.2013.06.005.
G. Öğündür Öznitelik Seçimi (Feature Selection) Teknikleri Medium 2019 Retrieved September 3, 2021 (https://medium.com/@gulcanogundur/öznitelik-seçimi-feature-selection-teknikleri-5cd8cbab7706.
Orçun, Ç., Sezgin, O.B., 2020. Ödül Temelli Kitle Fonlama Projesi İçin Kritik Başarı Faktörleri. 19th International Business Congress.
Peprah, P.B., Shneor, R., 2022. A trust-based crowdfunding campaign marketing framework: theoretical underpinnings and big-data analytics practice. Internat. J. Big Data Manage. 2 (1), 1. https://doi.org/10.1504/ijbdm.2022.119453.
Ralcheva, A., Roosenboom, P., 2020. Forecasting success in equity crowdfunding. Small Bus. Econ. 55 (1), 39–56. https://doi.org/10.1007/s11187-019-00144-x.
Ryoba, M.J., Qu, S., Zhou, Y., 2021. Feature subset selection for predicting the success of crowdfunding project campaigns. Electron. Mark. 31 (3), 671–684.
Sağbaş, E.A., Ballı, S., 2016. Akıllı Telefon Algılayıcıları ve Makine Öğrenmesi Kullanılarak Ulaşım Türü Tespiti. Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi 22 (5), 376–383. https://doi.org/10.5505/pajes.2015.63308.
Sakarya, Ş., Bezirgan, E., 2018. Kitlesel Fonlama Platformları: Türkiye ve Yurtdışı Karşılaştırması. Düzce Üniversitesi Sosyal Bilimler Dergisi 8 (2), 18–33.
Shi, J., Yang, K., Xu, W., Wang, M., 2021. Leveraging deep learning with audio analytics to predict the success of crowdfunding projects. J. Supercomput. 77 (7), 7833–7853.
Shneor, R., Vik, A.A., 2020. Crowdfunding success: A systematic literature review 2010–2017. Balt. J. Manag. 15 (2), 149–182.
Short, J.C., Ketchen, D.J., McKenny, A.F., Allison, T.H., Duane Ireland, R., 2017. "Research on crowdfunding: reviewing the (very recent) past and celebrating the present. Entrepreneurship: Theory Practice 41 (2), 149–160. https://doi.org/10.1111/etap.12270.
Steigenberger, N., Wilhelm, H., 2018. Extending signaling theory to rhetorical signals: evidence from crowdfunding. Organ. Sci. 29 (3), 529–546. https://doi.org/10.1287/orsc.2017.1195.
Tang, J., Alelyani, S., Liu, H., 2014. Feature selection for classification: A review. In: Data Classification: Algorithms and Applications, pp. 37–64.
Wang, W., Zheng, H., Yenchun, J.W., 2020. Prediction of fundraising outcomes for crowdfunding projects based on deep learning: A multimodel comparative study. Soft. Comput. 24 (11), 8323–8341. https://doi.org/10.1007/s00500-020-04822-x.
Wei, Y., Hong, J., Tellis, G.J., 2022. Machine learning for creativity: using similarity networks to design better crowdfunding projects. J. Mark. 86 (2), 87–104.
Witten, H., I. Frank, M. Hall, and C. Pal. 2017. *Data Mining: Practical Machine Learning Tools and Techniques*.
Xiang, W., Zhang, H., Yang, S., Zhou, M., Lin, W., Wen, J., 2021. A differential pilot protection scheme for MMC-based DC grid resilient to communication failure. IEEE J. Emerg. Selected Top. Power Electron. 9 (5), 5631–5645. https://doi.org/10.1109/JESTPE.2020.3035672.
Yan, K.e., Zhang, D., 2015. Feature selection and analysis on correlated gas sensor data with recursive feature elimination. Sens. Actuators B 212, 353–363. https://doi.org/10.1016/j.snb.2015.02.025.
Yeh, J.-Y., Chen, C.-H., 2022. A machine learning approach to predict the success of crowdfunding fintech project. J. Enterp. Inf. Manag. (107) 35 (6), 1678–1696.
Yuan, H., Lau, R.Y.K., Xu, W., 2016. The determinants of crowdfunding success: A semantic text analytics approach. Decis. Support Syst. 91, 67–76. https://doi.org/10.1016/j.dss.2016.08.001.
Zhong, C., Xu, W., Du, W., 2022. Success prediction of crowdfunding campaigns with project network: A machine learning approach. J. Electron. Commer. Res. 23 (2), 99–114.
Zhou, M.J., Baozhou, L., Fan, W.P., Alan Wang, G., 2018. Project description and crowdfunding success: an exploratory study. Inf. Syst. Front. 20 (2), 259–274. https://doi.org/10.1007/s10796-016-9723-1.
Zvilichovsky, D., Danziger, S., Steinhart, Y., 2018. Making-the-product-happen: A driver of crowdfunding participation. J. Interact. Mark. 41, 81–93. https://doi.org/10.1016/j.intmar.2017.10.002.