İSTANBUL UNIVERSITY PRESS

Istanbul Business Research, 50(2), 255-274
DOI: 10.26650/ibr.2021.50.0117
http://ibr.istanbul.edu.tr/
http://dergipark.gov.tr/ibr

Istanbul Business Research

Submitted: 29.09.2020
Revision Requested: 07.01.2021
Last Revision Received: 11.03.2021
Accepted: 26.05.2021
Published Online: 07.10.2021

RESEARCH ARTICLE

Do Machine Learning and Business Analytics Approaches Answer the Question of 'Will Your Kickstarter Project be Successful?

Murat Kılınç[1], Can Aydın[2], Çiğdem Tarhan[3]

Abstract

Kickstarter is one of the popular crowdfunding platforms used to implement business ideas on the web. The success of crowdfunding projects such as Kickstarter is realized with future financial support. However, there is no platform where users can get decision support before presenting their projects to supporters. To solve this problem, a platform where users can test their projects is required. Within this scope, a business intelligence model that works on the web has been developed by combining business analytics and machine learning methods. The data used for business analytics has been brought to a state that can provide inferences through visualization, reporting and query processes. Within the scope of machine learning, various algorithms were applied for success classification and the best results were given by 91% Random Forest, 85% Decision Tree, 84% K-Nearest Neighbors (KNN) algorithms. F1-Score, Recall, Precision, Mean Squared Error (MSE), Kappa and AUC values were analyzed to determine the most successful models. Thus, Kickstarter users will be able to see their shortcomings and have a prediction about success before presenting their projects to their backers.

Keywords

Machine Learning, Business Analytics, Decision Support System, Business Intelligence

Introduction

Crowdfunding, a new generation investment system, is a good alternative for entrepreneurs and projects. In particular, the process of financing ventures over the internet has gained great importance with the maturity of technologies that come with web 2.0 and the success of resource utilization (Zvilichovsky et al., 2015). Accordingly, the interest in crowdfunding platforms has been increasing over the years, as it provides easier access to large audiences. The crowdfunding market size, which was 597 million dollars in 2014, was stated as 10.2 billion dollars in 2018. In 2025, the crowdfunding market size is expected to be 28.8 billion

[1] **Corresponding Author:** Murat Kılınç (PhD Candidate), Dokuz Eylul University, Faculty of Economics and Administrative Sciences, Department of Management Information Systems, Izmir, Turkey. E-mail: kilinc.murat@cbu.edu.tr ORCID: 0000-0003-4092-5967

[2] Can Aydın (Assoc. Prof.), Dokuz Eylul University, Faculty of Economics and Administrative Sciences, Department of Management Information Systems, Izmir, Turkey. E-mail: can.aydin@deu.edu.tr ORCID: 0000-0002-0133-9634

[3] Çiğdem Tarhan (Assoc. Prof.), Dokuz Eylul University, Faculty of Economics and Administrative Sciences, Department of Management Information Systems, Izmir, Turkey. E-mail: can.aydin@deu.edu.tr ORCID: 0000-0002-5891-0635

To cite this article: Kilinc, M., Aydin, C., & Tarhan, C. (2021). Do Machine Learning and Business Analytics Approaches Answer the Question of 'Will Your Kickstarter Project be Successful?. *Istanbul Business Research, 50*(2), 255-274. http://doi.org/10.26650/ibr.2021.50.0117

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License

dollars (Szmigiera, 2019). Many crowdfunding platforms have been developed nowadays for the increasing market. Thanks to platforms like Kickstarter, Indiegogo, GoFundMe, Fundable, and Patreon, project developers can collect hundreds of millions of dollars of support each year (Etter et al., 2013). Only within the Kickstarter platform, approximately 180,000 projects have received a total of $ 5 billion since the platform was established (Kickstarter Project Stats, 2020). After financial and tactical support provided through crowdfunding platforms, the ideas of entrepreneurs; come to life thanks to supporters, angel investors, and capital funds (Kuppuswamy and Bayus, 2018). Ideas that are disliked or who cannot find adequate support can fail because they cannot find adequate support financially. Within the Kickstarter ecosystem, 300.000 unsuccessful projects have not found adequate support since its establishment. Accordingly, statistics published in July 2019 show that the success rate on the Kickstarter platform is 37.3% (Statista, 2020a). For this reason, getting sufficient support in the venture ecosystem is very important in terms of realizing that idea. Therefore, providing a prediction on whether the idea of the initiative or the project will be successful before presenting the project to the supporters has a positive impact on the entrepreneurs. This prediction can be used to attract backers on crowdfunding platforms. Factors forming the dynamics of the project such as information, content, texts, images used in the project profile, in short, all the features of the project are directly related to the interest of the backers (Cheng and Others, 2019). Because the backers give more priority to the projects that were well expressed in the stage before the detailed review of the project. However, entrepreneurs on crowdfunding platforms cannot test their projects for success before presenting them to the supporters. For this reason, it is not possible to compare projects in the crowdfunding ecosystem with other projects before they are added to the platform and to have a prediction about their success. This problem can be solved by providing forecasting, data visualization, reporting, listing, querying operations to the entrepreneurs in a hybrid way on the web. Similarly, in a study conducted in 2015, 12 years of historical data of a consultancy company were analyzed and then an interactive decision support system was created using business analytics and machine learning (Cook et al., 2015). In this direction, the research question we have raised has been “do business analytics and machine learning methods affect the decision support processes for entrepreneurs in a hybrid way”. In our study, which provides a prediction in terms of success for the Kickstarter initiatives, a data set containing the project features were analyzed and evaluated primarily within the scope of business analytics. Afterward, the data on the system was trained and made ready for classification by machine learning methods. In the last stage, the features of the project of the user were taken as input, and a comparison was made with the other projects in the business analytics processes, and the project success estimation was made by using machine learning methods. To provide decision support, user interaction should be planned within the system development cycle. For this reason, user interaction was created by ensuring that this entire process is shown on the dashboard.

Related Works

In the research, crowdfunding platforms were examined within the scope of business analytics and machine learning, and many different studies were encountered. The support vector machine method was used in the system development work by Chen in 2013 to predict whether a Kickstarter project would be successful in advance. With the study, using the initial features of the projects, a success estimation of 67% accuracy was provided (Chen et al., 2013). In the research carried out by Kindler in 2019, propagation mechanisms on crowdfunding platforms such as Kickstarter, Indiegogo, and Sellaband were investigated. Because the spread of the project is directly related to virality and success (Kindler et al., 2019). In the research put forward by Chung, Kickstarter datasets, supporter-campaign graphics and Naive Bayes, Random Forest, and Adaboost classification methods were used. Adaboost classification method, which gives the highest value, made a success estimate with an accuracy rate of 76% according to the data set examined (Chung & Lee, 2015). In the study put forward by Rao and his team in 2014, it was emphasized that the success rate of the projects in the crowdfunding structure is less than 50%. Also, the relationship between money pledged and campaign success, which was made using decision trees, was examined. According to the review, it was determined that the duration of the campaigns had a significant impact on the success of the project. Besides, it was emphasized that with a predictor to be created, the success of the campaigns can be estimated correctly by 84%. (Rao et al., 2014). In the study by Jensen and Özkil in 2018, factors that may cause failure in crowdfunding platforms were examined. According to the review, the ability of campaign starters to make promises about product features and the project features created in this context plays an important role in the success of the project. The study also shows how crowdfunding platforms can be used in research with both data libraries and product development cases (Jensen & Özkil, 2018). The research conducted by Qianzhou and his team focuses mainly on the main points of the projects such as category and target. In this context, a large data set obtained from Kickstarter was used. According to the research result, there is a direct proportion between the information provided in the project descriptions and the financial support obtained. The model introduced can predict the financial success of the project with an accuracy rate of 73%. Also, the research proposes the Support Vector Machine (SVM) classification method to further increase the predictive accuracy rate (Du et al., 2015). In another study by Sheng Bi and his team in 2016, based on a detailed probability model, it was investigated how online information in the ventures affected investors' decisions. In the research carried out with the data of crowdfunding websites operating in China, it was revealed that a higher number of likes, online feedback, more detailed project description and video introduction of the project had a positive effect on fund investment decisions. The data analysis in the study also emphasized that different project categories should be evaluated with different perspectives (Bi et al., 2016). In the research carried out by Mortensen and his team, the method of learning and defining the success factor was done with machine learning methods. In the research, it is examined what drives the success of a packaging company on the Fortune 500 list and a model is presented accordingly. It is observed in the study that using statistical modeling techniques improves both revenue and profit revenues by affecting shortening sales cycles and decreasing sales costs. The best model with the use of machine learning methods, decision tree, gradient boost, and random forest algorithms; accuracy 80%, precision 86%, recall 77% results (Mortensen et al., 2019). Finally, in another study, it was observed that the use of ensemble artificial neural networks in the prediction of success of crowdfunding projects yielded better results than other algorithms. With the accuracy values varying according to the parts, Logistic Regression can make predictions with 88.38%, Artifical Neural Networks (ANN) 88.62% and Ensemble Neural Networks (ENN) 89.18% (Yeh & Chen, 2020).

Summarizing the studies examined, it is observed that the presentation of the projects in crowdfunding platforms to the backers with a detailed explanation increases the success. This explanation can be effective by determining the project features correctly. Also, another remarkable point in the literature is that the success rate in crowdfunding is in a downward trend. One of the main reasons for this can be described as the increasing interest in crowdfunding and the sloppy preparation of projects uploaded to crowdfunding platforms. On the other hand, the success of the classification results for crowdfunding projects has been increasing over the years. Because, especially in 2018 and after, machine learning algorithms, which can give better results when operated together (ensemble), have been used more. For this reason, a better classification can be made with both the selection of algorithms suitable for the data set and the ensemble approach. When the literature is evaluated in terms of business analytics, it is observed that decision support increases with visual reporting, analysis and user interaction.

Methodology

In the process of creating a decision support with machine learning and business analytics, the Cross-Industry Standard Process Model for Data Mining (CRISP-DM) was used and 6 steps in the cycle were applied in order. Since the chosen method is accepted as a flexible and circular model, it is frequently preferred in data science projects. Also, it is possible to go to the previous step in the model and make changes. Because there may be changes in needs or data structure. Therefore, re-evaluation should be made when similar situations are encountered. In other words, the CRISP-DM method; It can be considered as a supportive tool for improvement, error analysis and quality management, data analysis and mining projects (Schäfer et al., 2018, Weimer et al., 2019).

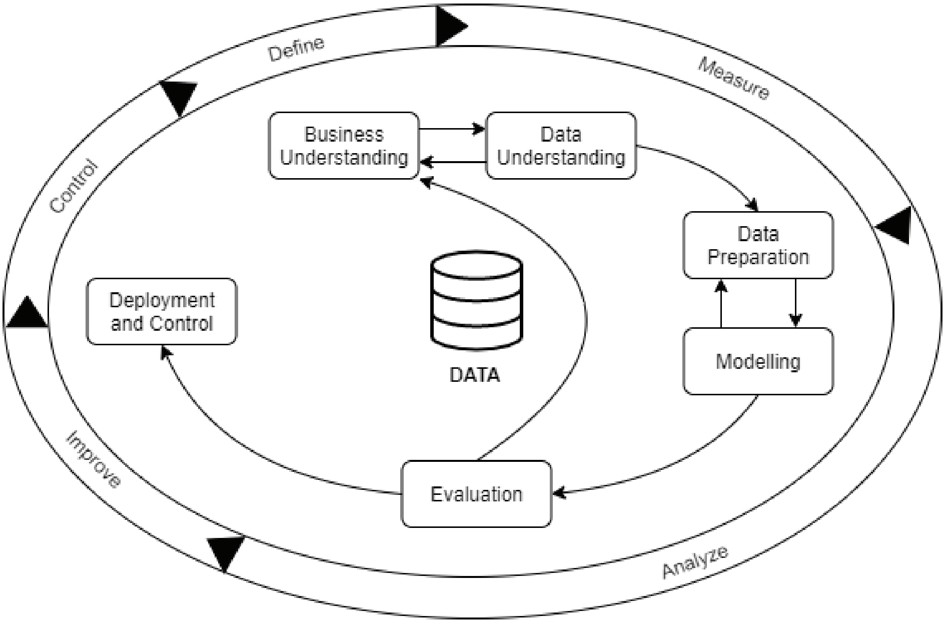


Figure 1. Cross Industry Standard Process Model for Data Mining (Huber et al., 2019)(Fahmy et al.,2017)

In the CRISP-DM process used; After evaluating possible problems at the first stage, a literature review was conducted and which software libraries to use were determined (Table 1). In the second stage, the quality, accessibility, and sustainability of the data were discussed. In the third stage, the data set that has undergone pre-processing has been applied to a model in terms of the path and analysis structures that the data will follow. In the evaluation section, the results of the classification algorithms are compared and the algorithms that can make the best classification are used for the application. Also, the functions such as querying the data, user interaction, listing, and reporting were inspected and the business analytics process was evaluated. Finally, during the deployment process, the model is provided to be displayed on the dashboard. Accordingly, a web application was developed and transferred to entrepreneurs through the control of all processes, business analytics, and machine learning methods. The aim at this point is to create an inference mechanism by combining the reports of business analytics with the results of classification algorithms.

Table 1
Software Libraries and Tools

	Web Application Section	Machine Learning Section	Data Preprocessing
Python Libraries and Structures	Flask Framework	Sklearn Library	Pandas, Numpy
Object-Oriented Programming Languages	PHP, Javascript, Python (Compiled by VS Code.)	Python (Compiled by Spyder VS Code.)	Python (Compiled by Jupyter)
Other Tools and Libraries	HTML, CSS, Bootstrap, Chart.js	Google Colab	Weka

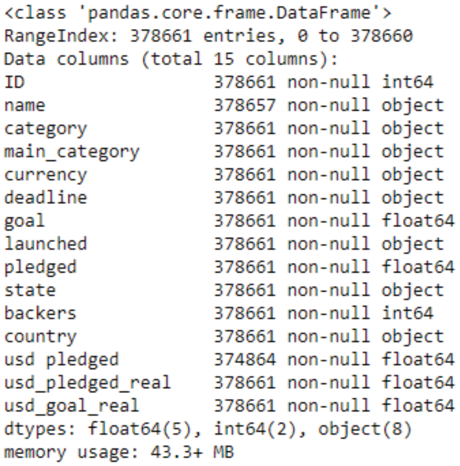
Data

Kickstarter is the most popular among crowdfunding platforms, thousands of projects are added to the platform daily. The features of the added projects constitute the focus of our study for the analysis processes. The results of business intelligence, business analytics, and machine learning methods have been put forward based on the characteristics of these projects. In this context, the data file provided by Kaggle, which covers more than 300.000 Kickstarter initiatives with various features, was used in the research. The projects created in 2017 and after are filtered out and the number of analyzed project data is reduced to around 50,000 (Mouillé, 2018). This filtering process was carried out to research with the most recent project data in the data set.

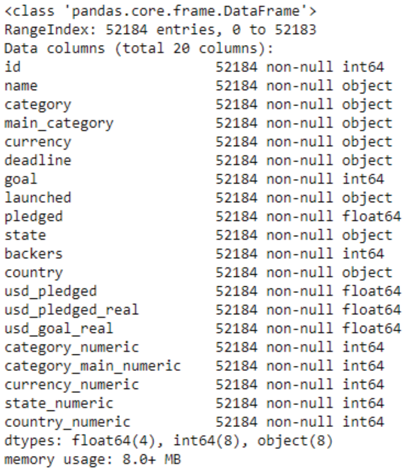
Data Preprocessing: During the preparation of the data in the CRISP-DM method, the data set used was prepared for analysis (Table 2).

Table 2
Data Set States Before and After Data Preprocessing

Before Preprocessing Data			After Preprocessing Data		
378.661 line	15 column	Size: 43.3 MB	52.184 line	20 column	Size: 8.0 MB



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 378661 entries, 0 to 378660
Data columns (total 15 columns):
ID                  378661 non-null int64
name                378657 non-null object
category            378661 non-null object
main_category       378661 non-null object
currency            378661 non-null object
deadline            378661 non-null object
goal                378661 non-null float64
launched            378661 non-null object
pledged             378661 non-null float64
state               378661 non-null object
backers             378661 non-null int64
country             378661 non-null object
usd pledged         374864 non-null float64
usd_pledged_real    378661 non-null float64
usd_goal_real       378661 non-null float64
dtypes: float64(5), int64(2), object(8)
memory usage: 43.3+ MB
```



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 52184 entries, 0 to 52183
Data columns (total 20 columns):
id                      52184 non-null int64
name                    52184 non-null object
category                52184 non-null object
main_category           52184 non-null object
currency                52184 non-null object
deadline                52184 non-null object
goal                    52184 non-null int64
launched                52184 non-null object
pledged                 52184 non-null float64
state                   52184 non-null object
backers                 52184 non-null int64
country                 52184 non-null object
usd_pledged             52184 non-null float64
usd_pledged_real        52184 non-null float64
usd_goal_real           52184 non-null float64
category_numeric        52184 non-null int64
category_main_numeric   52184 non-null int64
currency_numeric        52184 non-null int64
state_numeric           52184 non-null int64
country_numeric         52184 non-null int64
dtypes: float64(4), int64(8), object(8)
memory usage: 8.0+ MB
```

In the data set consisting of 15 columns, there are columns like name, category, main category, currency, deadline, goal, launched, pledged, state, backers, country, USD pledged, USD pledged real, USD goal real. For efficient analysis, rows of data with missing content were cleaned from within the data set. Also, columns with string values have been converted into numerical data within the scope of data pre-processing and made ready for analysis. Within the scope of data pre-processing, the following 2 different determinants were emphasized:

1) State: The State column is a feature that shows how the projects are in terms of success. Kickstarter projects data set includes 5 different project statuses: successful, failed, live, suspended, canceled. As part of our study, all project statuses were used for business analytics. But for classification methods, only successful and failed states were used to make the binary classification. For this reason, the number of data used for business analytics is 52,184, while the number of data used for machine learning is 43,304.

2) Data Profile: The data set used is in a scattered structure. Therefore, the application should be transferred to the database in a relational way. For this, the relationships between the main-sub categories have been numerically established and each project feature has been transferred correctly into the application.

Along with the 2 determinants, the pre-processing stage was completed and a data set ready for the analysis stage was obtained.

Data Splitting: Within the scope of the study, the data set is divided into 2 separate sections in order to train our classification models and to understand how they perform. In this context, the data set for training and testing processes is divided into 70% training and 30% testing. The ratios used are observational and determined in the most appropriate way for the data set. This use, random subsampling, is probably one of the most used methods to divide the data set in a study.

Data Scaling: Since the values in the columns in the data set cannot be converted into each other, they need to be rescaled. Scaling enables classification methods to work more accurately and give meaningful results (1).

$$X\_std = \frac{(X - X.min\ (axis = 0))}{(X.max(axis = 0) - X.min\ (axis = 0))} \quad (1)$$

For this reason, the data set was scaled in the range of 0-1 using the min-max scaler.

Modeling

The modeling process refers to a decision support process created by including the most appropriate elements in the business structure desired to be developed. In this context, how to use business analytics for a web application developed, which algorithms to use from machine learning methods, and how to follow it are determined at this stage (Figure 2). According to the modeling, the user who wants to test his interference on the system is included in the system with the project features he enters. After the Kickstarter dataset, which consists of approximately 50 thousand data, has been pre-processed with the Pandas and Numpy libraries of Python, the database of the web application has been created. The data in the created database are included in the extraction process in two different ways. Firstly, data analysis and reporting are provided with PHP. Later, with the chart.js library compiled with Javascript, the data was visualized in graphics and transferred to the web interface.

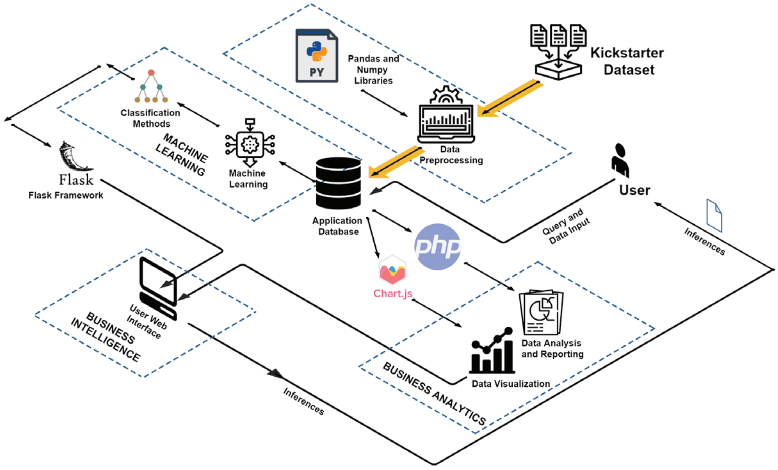


Figure 2. System Flow Chart Modeling Process

Secondly, the data kept numerically in the database were tested by machine learning methods after normalization. The methods are then displayed on the web interface with their accuracy rates. The decision support platform, which combines the reporting, data analysis, and data visualization capabilities of business analytics with the predictive ability of machine learning methods, provides inferences to the user. In this way, the query and data input made by the user on the system is returned to him as inference and decision support (Figure 2).

Business Analytics and Intelligence

Business analytics and intelligence emerged as an important field of study for both implementers and researchers, reflecting the magnitude and impact of problems related to data to be resolved in contemporary business organizations (Chen et al., 2012). In other words, to make more accurate decisions about the future, it is the business processes that enable the data to be transformed into information by examining the past or current data. Especially in the last 20 years, the usage rates of business intelligence and analytics have increased significantly both academically and industrially (Forbes, 2017; Statista ,2020b; Laudon, 2014). These methods, which contribute significantly to the decision support processes of large-scale formations, are also among the important branches by the companies that direct technology. Business analytics; It increases the interaction of users with data such as predictive modeling, making the data meaningful and optimization. All the methods used by transferring the whole process onto the dashboard assist the decision support processes, making the data interpretable and analyzable. This structure, which has been used frequently by organizations until recently, has also become a supporter of individual uses over time. Decision support processes that are revealed by taking certain data from users are the best examples of this. For example, in our research, machine learning and business analytics were brought together and entrepreneurs received decision support about their projects.

Selection of the Algorithm

Scope of work; Decision Tree (J48), Random Forest (RF), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Naive Bayes (NB), Logistic Regression (LR) algorithms were used to classify Kickstarter projects.

Decision trees are one of the basic classification algorithms and multiple decision trees come together to form the Random Forest algorithm (Leiva et al., 2019). The basis of the algorithm is based on a hierarchical tree structure (Berhane et al., 2018). However, although this structure is simple and understandable, in cases where one of the biggest problems of the algorithm is overfitting, the number of decision trees in the Random Forest method should be determined and the results obtained should be reviewed.

Random Forest Classification is one of the most successful classification methods used. The algorithm consists of decision trees, independent of the input vector of each classifier. Each tree in the hierarchy gives a unit vote to classify the input vector (Pal, 2005). This classification method gives better results in datasets and categorical variables with an unbalanced distribution. The data set used during the development of the application has many data categorically and has been tested within the scope of this algorithm since it has an unbalanced distribution (Ahmad et al., 2017).

The KNN algorithm is one of the simplest methods used to solve classification problems. It has important advantages over some data mining methods since it generally gives competitive results (Adeniyi et al., 2016). In particular, it can provide fast results with Decision Tree and Random Forest among the methods run on the webserver. This provides increased usability on the web interface. Also, KNN performs well on large data sets because it does not have scalability problems.

SVM is a classification algorithm that finds the best line at the point of separating two classes. Although it is an easy-to-train algorithm, it has two types of linear and nonlinear (Jain et al. 2020). But it usually tries to classify the data on the class linearly. In nonlinear cases, a third dimension (Kernel Trick) can be classified using the SVM algorithm.

The NB algorithm is a classifier that calculates the probability set by counting the frequency and value sets in a given data set (Saritas and Yasar, 2019). Good results with fewer education data and the ability to work on unbalanced datasets are among the advantages of the NB algorithm.

The LR algorithm is frequently used to reveal the binary state of the result after training the used dataset. Since the Kickstarter data set is classified as binary classification, the LR algorithm is also preferred among the methods analyzed. With the LR algorithm, the effects on the independent variable result variables can be calculated probabilistic.

Evaluation

In the evaluation phase, which constitutes the 5th part of the CRISP-DM cycle, the methods used in the research and their performances are discussed. The web platform where entrepreneurs will test their projects in terms of success has been developed with both business analytics and machine learning methods. For this reason, the evaluation section was examined under two separate topics.

Evaluation of Classification Algorithms: It is necessary to measure how accurately the classification methods used can classify. Confusion Matrix was used to test the classification methods. The properties of the Confusion Matrix are compatible with integration into machine learning classification methods and provide more semantic explanations (Xu et al. 2020). Besides, the Confusion Matrix is used to describe the performance of the classification method on test data with true values. Accordingly, the data set containing two classes, successful and failed, was analyzed with Decision Tree, Random Forest, KNN, SVM, Naive Bayes, and Logistic Regression algorithms. Afterward, the heat map analyzes were visualized and Confusion Matrix values appeared (Figure 3).

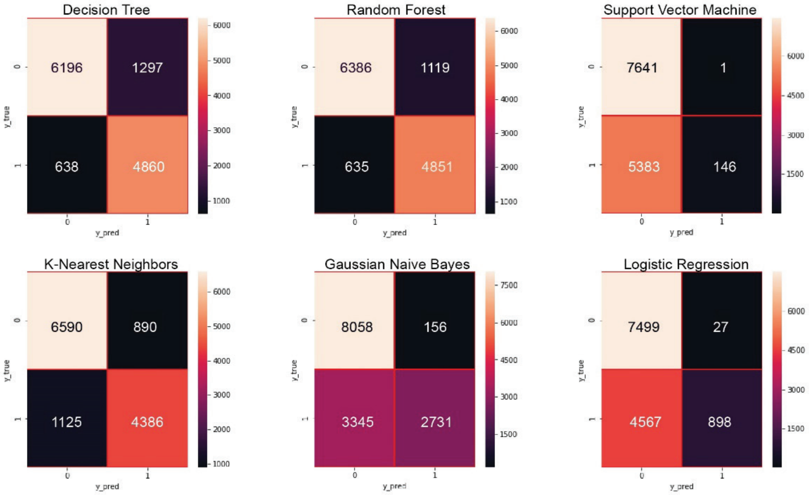


Figure 3. Confusion Matrix Results of the Classification Algorithms Used

(Failed = 0, Successful =1)

After exposing the confusion matrix, classification metrics need to be evaluated. In this direction, it should be stated how much of the results obtained from the model that was first put forward is estimated correctly. For this, accuracy rate measurement is used for classification methods (Yang et al., 2019), (2).

$$AccuracyRate = \frac{TP + TN}{TP + FP + FN + TN} \quad (2)$$

The other two of the calculations using the confusion matrix after the accuracy rate are the Recall and Precision measurements (Fawcett, 2004), (3), (4).

$$Recall = \frac{TP}{TP + FN} \quad (3), \quad Precision = \frac{TP}{TP + FP} \quad (4)$$

F1 Score is a classification method evaluation metric in which extreme situations are not ignored (5). The main reason for looking at the F1 score value instead of the accuracy rate in the classification method choices is to not make a wrong model selection in the data sets that are not evenly distributed. Since the Kickstarter data set consists of unevenly distributed samples, the F1 score of the 6 classification methods used was determined.

$$F1Score = \frac{2 * Recall * Precision}{Recall + Precision} \quad (5)$$

$$MeanSquaredError(MSE) = \frac{1}{n}\sum_{i=0}^{n}(y_i - y_i')^2 \quad (6)$$

MSE measures the average size of errors in the classification made (6). Separating from the complexity matrix, Cohen's Kappa coefficient (κ) is a statistical metric that measures the mismatch between two different values (Cohen, 1960). The resulting measurement value is between -1 and +1. The closer the κ value closes to the +1 value, the better the compatibility between the two different values (Table 4).

$$KappaScore\ (\kappa)\ =\frac{(P_0-P_c)}{(1-P_0)}\quad (7)$$

Similarly, the closer the coefficient κ is to the value -1, the incompatibility between the two values is high and does not make sense in terms of reliability. If κ = 0, it is stated that the agreement between the two evaluators may depend on chance.

Table 3

Evaluation of the Kappa Score

κ **(Kappa Score)**	**Evaluation**
$\boldsymbol{\kappa}$ **< 0**	Poor
$\boldsymbol{\kappa}$ > 0.0 ve ≤ 0.20	Slight
$\boldsymbol{\kappa}$ ≥ 0.21 ve ≤ 0.40	Fair
$\boldsymbol{\kappa}$ ≥ 0.41 ve ≤ 0.60	Moderate
$\boldsymbol{\kappa}$ ≥ 0.61 ve ≤ 0.80	Substantial
$\boldsymbol{\kappa}$ ≥ 0.81 ve ≤ 1	Almost Perfect

Source: Landis and Koch, 1977.

Kappa value is found with the above equation, $P_0 = AcceptedRate,\ P_c = ExpectedRate$(7). κ ≥ 0.4 appears to be an appropriate value (Table 3). The model evaluation metrics mentioned above were used to test how many classifications of the methods used for Kickstarter attempts can be done. The results in this context are compared with each other (Table 4).

Table 4

Comparison of the Results of the Classification Algorithm

Machine Learning Algorithms	Accuracy Score	Precision	Recall	F1 Score	AUC	Mean Squared Error (MSE)	Kappa Score
Decision Tree	0.85	0.86	0.86	0.86	0.860	0.143	0.710
Random Forest	0.91	0.91	0.91	0.91	0.912	0.088	0.819
K-Nearest Neighbors (KNN)	0.84	0.85	0.85	0.84	0.837	0.154	0.681
Support Vector Machine (SVM)	0.58	0.76	0.59	0.44	0.513	0.414	0.030
Gaussian Naive Bayes	0.76	0.81	0.76	0.74	0.728	0.239	0.484
Logistic Regression	0.63	0.77	0.64	0.55	0.576	0.364	0.169

As a result of the comparison, Decision Tree, Random Forest, and KNN algorithms perform the best classification processes. SVM, Naive Bayes, and Logistic Regression algorithms give lower values in the classification of Kickstarter projects than other methods. Finally, Decision Tree, Random Forest, and KNN algorithms that give the best results were also evaluated in terms of ROC Curves (Figure 4).

$$TruePositiveRate(TPR) = \frac{TP}{TP + FN}(7), \qquad FalsePositiveRate(FPR) = \frac{FP}{FN + TP}(8)$$

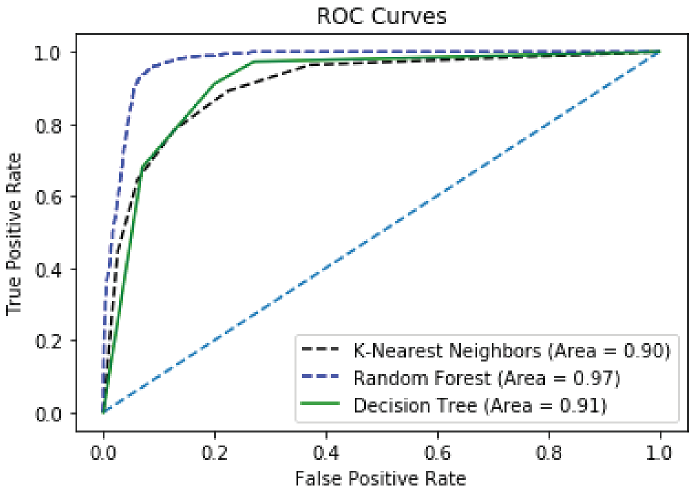


Figure 4. ROC Curves for Classification Algorithms

As a result of the evaluation, it is seen that Decision Tree, Random Forest, and KNN algorithms can successfully classify Kickstarter projects. For this reason, these 3 classification methods were used in the web application developed.

Evaluation of Business Analytics Methods: During the evaluation of the business analytics we use while developing the web application, the visualization, querying, listing, and meaningful reporting of the data recorded in the database were analyzed. Thanks to the relational database structure created, each recorded data takes an active role in the process leading to decision support.

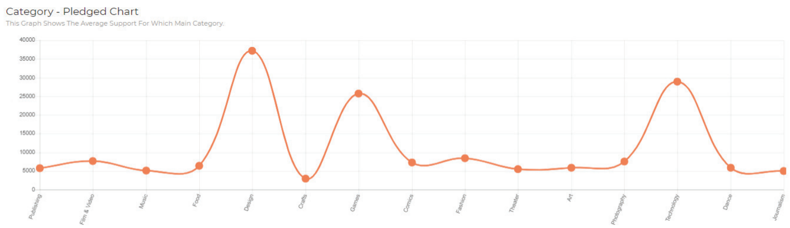


Figure 5. Visualization of Kickstarter Data

After the recorded data in the database is displayed on the dashboard with SQL queries, the relationships between the data are visualized and made easier to interpret (Figure 5). So much so, with this interpretation, entrepreneurs can make inferences about their projects and see what they need to do to make their projects successful thanks to the reporting screen. Users who can also see their deficiencies in creating a successful project can make detailed inquiries about other projects within the web application. Queries also cover the status of other projects. The table below shows the project status distributions in the categories (Table 5).

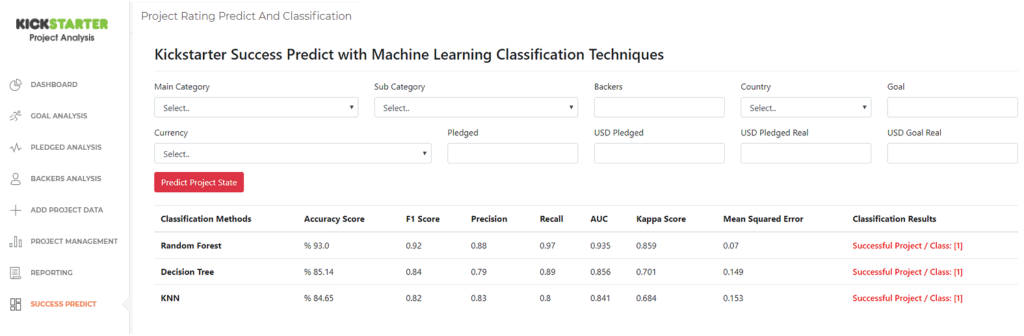
Table 5
Kickstarter Category and Project States Table

Categories	Project States					Project Count
	Successful	Failed	Canceled	Live	Suspended	
Publishing	%34.24	%51.28	%8.52	%5.68	%0.29	5.248
Film&Video	%34.5	%49.58	%9.53	%5.97	%0.42	5.508
Music	%43.34	%44.04	%6.41	%5.76	%0.45	4.864
Food	%22.29	%63.17	%8.13	%5.71	%0.69	3.185
Design	%38.54	%39.7	%15.78	%5.19	%0.8	5.862
Crafts	%25.88	%58.63	%9.73	%4.93	%0.84	1.542
Games	%41.49	%37.23	%16.53	%4.16	%0.58	6.895
Comics	%61.87	%27.84	%6.35	%3.8	%0.15	2.001
Fashion	%27.35	%54.96	%11.15	%5.92	%0.62	4.205
Theater	%53.6	%35.33	%6.42	%4.54	%0.11	903
Art	%42.46	%44.23	%7.86	%4.88	%0.58	3.957
Photography	%33.94	%52.07	%8.58	%4.63	%0.77	1.037
Technology	%19.4	%58.77	%14.02	%6.27	%1.54	5.984
Dance	%55.01	%33.06	%6.78	%4.88	%0.27	369
Journalism	%19.71	%62.18	%12.5	%4.97	%0.64	624

When the database is analyzed within the scope of business analytics, it is seen that the success rates of the projects put forward in the comics, theater, and dance categories are high (Table 5). There is a great failure in the categories of publishing, food, crafts, fashion, photography, technology, and journalism. In order to make such inferences more determinative, data visualization and reporting have been applied for other features such as location, currency, targeted amount of support, and the number of backers in the data set. Thus, all the data in the process was presented to the user in an analyzed format within the dashboard.

Deployment and Control

In the deployment process of the CRISP-DM cycle, the developed web application is implemented ready to use. In this context, the classification methods used are shown in the web interface via the Flask Framework (Figure 6). The results that appear with the user interaction in the web application change as the project features differ. In other words, a classification based estimate is made for the success of the project, taking into account the previous data. Therefore, examples of the contribution of methods of machine learning to business intelligence have been seen (Wang et al., 2005).



Project Rating Predict And Classification

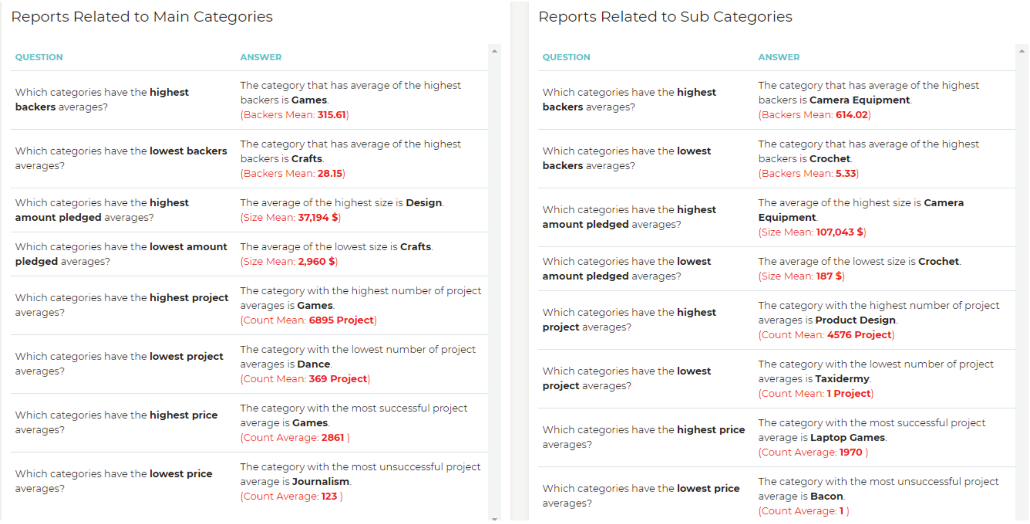
Kickstarter Success Predict with Machine Learning Classification Techniques

Main Category: Select..
Sub Category: Select..
Backers:
Country: Select..
Goal:
Currency: Select..
Pledged:
USD Pledged:
USD Pledged Real:
USD Goal Real:

Classification Methods	Accuracy Score	F1 Score	Precision	Recall	AUC	Kappa Score	Mean Squared Error	Classification Results
Random Forest	% 93.0	0.92	0.88	0.97	0.935	0.859	0.07	Successful Project / Class: [1]
Decision Tree	% 85.14	0.84	0.79	0.89	0.856	0.701	0.149	Successful Project / Class: [1]
KNN	% 84.65	0.82	0.83	0.8	0.841	0.684	0.153	Successful Project / Class: [1]

Figure 6. Demonstration of Classification Algorithms with Flask Framework in the User Interface

It is also observed that by enabling decision support, solutions to complex problems are provided through the dashboard (Cook et al., 2015).



Reports Related to Main Categories

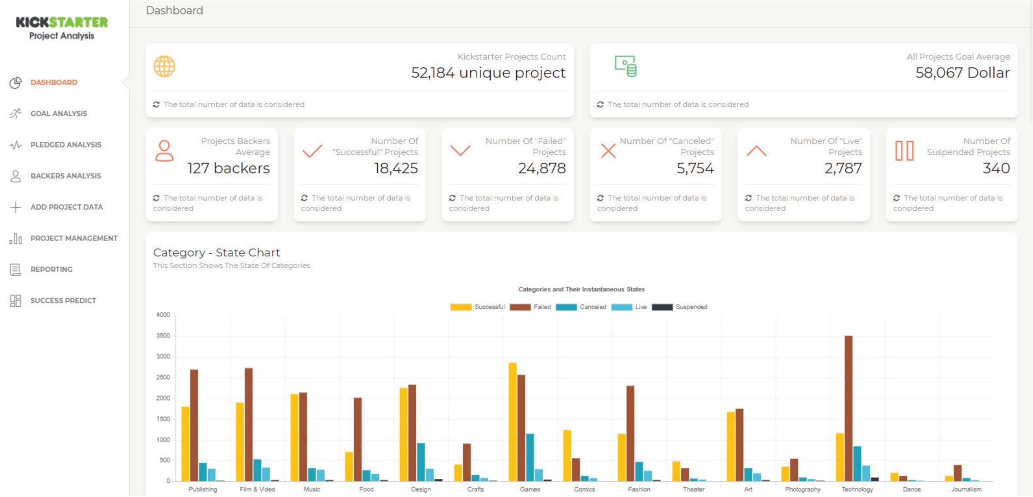
QUESTION	ANSWER
Which categories have the **highest backers** averages?	The category that has average of the highest backers is **Games**. (Backers Mean: **315.61**)
Which categories have the **lowest backers** averages?	The category that has average of the highest backers is **Crafts**. (Backers Mean: **28.15**)
Which categories have the **highest amount pledged** averages?	The average of the highest size is **Design**. (Size Mean: **37,194 $**)
Which categories have the **lowest amount pledged** averages?	The average of the lowest size is **Crafts**. (Size Mean: **2,960 $**)
Which categories have the **highest project** averages?	The category with the highest number of project averages is **Games**. (Count Mean: **6895 Project**)
Which categories have the **lowest project** averages?	The category with the lowest number of project averages is **Dance**. (Count Mean: **369 Project**)
Which categories have the **highest price** averages?	The category with the most successful project average is **Games**. (Count Average: **2861**)
Which categories have the **lowest price** averages?	The category with the most unsuccessful project average is **Journalism**. (Count Average: **123**)

Reports Related to Sub Categories

QUESTION	ANSWER
Which categories have the **highest backers** averages?	The category that has average of the highest backers is **Camera Equipment**. (Backers Mean: **614.02**)
Which categories have the **lowest backers** averages?	The category that has average of the highest backers is **Crochet**. (Backers Mean: **5.33**)
Which categories have the **highest amount pledged** averages?	The average of the highest size is **Camera Equipment**. (Size Mean: **107,043 $**)
Which categories have the **lowest amount pledged** averages?	The average of the lowest size is **Crochet**. (Size Mean: **187 $**)
Which categories have the **highest project** averages?	The category with the highest number of project averages is **Product Design**. (Count Mean: **4576 Project**)
Which categories have the **lowest project** averages?	The category with the lowest number of project averages is **Taxidermy**. (Count Mean: **1 Project**)
Which categories have the **highest price** averages?	The category with the most successful project average is **Laptop Games**. (Count Average: **1970**)
Which categories have the **lowest price** averages?	The category with the most unsuccessful project average is **Bacon**. (Count Average: **1**)

Figure 7. Business Analytics Reporting and Inference Screen

Within the scope of business analytics, after the data used has been made meaningful with PHP and SQL queries, it has been transferred to the web interface for decision support on the reporting screen (Figure 7), (Figure 8). In this way, together with user interaction, entrepreneurs will be able to compare their projects with other projects and see their shortcomings better.



Dashboard

Kickstarter Projects Count
52,184 unique project
The total number of data is considered

All Projects Goal Average
58,067 Dollar
The total number of data is considered

Projects Backers Average
127 backers
The total number of data is considered

Number Of "Successful" Projects
18,425
The total number of data is considered

Number Of "Failed" Projects
24,878
The total number of data is considered

Number Of "Canceled" Projects
5,754
The total number of data is considered

Number Of "Live" Projects
2,787
The total number of data is considered

Number Of Suspended Projects
340
The total number of data is considered

Category - State Chart
This Section Shows The State Of Categories

Figure 8. Dashboard Developed for Kickstarter Users

Also, as a dynamic structure is created, the reporting interface can update itself as data is added to the application. Therefore, the validity of the system has been maintained.

Discussion and Conclusion

The growth of interest in crowdfunding platforms tends to continue in the coming years, as online resources can be used practically for creative entrepreneurs (Szmigiera, 2019). However, the decrease in project success rates is striking, inversely proportional to this increase in interest (Rao et al., 2014). In this context, there are many factors that affect the success rate. While some of these factors are in the data set we use; factors such as project visuals, audio materials and past experiences of the project owner are not included in the data set we use. For this reason, more accurate estimates can be obtained by performing studies with other factors in similar studies. On the other hand, the classification results we have revealed are at a satisfactory level with an upper limit of 91% at the point of predicting success. Other results shown in the web interface are also close to the upper limit and are shown in Table 4. This upper limit can be further increased with ensemble models that can be used and data to be added to the web application. As a matter of fact, in a recent study; It is seen that a better classification is made compared to previous years by using decision tree, gradient boost and random forest algorithms together (Mortensen et al., 2019). Another important point we noticed in the study is that the selection of the appropriate algorithm for the data set can affect the success of the classification. Especially good results can be obtained with random forest and decision tree algorithms in analyzes to be made with a scattered and irregular data set. Therefore, choosing an algorithm suitable for the data set to be used increases the success of the results. In the KNN algorithm, a high ratio was obtained by determining the most appropriate K value and classifying it. For this reason, it is necessary to find the value that gives the best result among the K selections during the study. When we look at the data visualization, listing and reporting features of the application we have developed, decision support is provided with the active use of user interaction. This decision support becomes more active with the use of the web (Shim et al., 2002). In other words, the classification results; since user interaction can be evaluated in terms of visualizing summaries and seeing alternative scenarios, users in the web application can have a prediction about their projects.

As a result, although there are many studies in the literature about crowdfunding projects, it was not possible for end users to test success because these studies were not presented on the web. The application we developed to overcome this deficiency creates a business intelligence environment by combining business analytics and machine learning methods. Users who can test the features of the project on the platform we developed can get decision support that will enable them to make necessary changes in their projects in line with the resulting report and success predict. Accordingly, if we summarize the results;

1) Using business analytics and machine learning methods together is highly effective for creating decision support. For this reason, similar processes should be carefully considered when analyzing other crowdfunding platforms.

2) A web-based decision support system has been established that entrepreneurs can use while preparing their projects. As the data is added to the established system, there is a dynamic structure because the forecast values can change.

3) Supervised learning algorithms that give the best results in web application are used for success prediction. In this context, Random Forest can classify with 91%, Decision Tree 85% and KNN 84% accuracy. In addition, the consistency of these results was tested with Recall, Precision, F1 Score, MSE and ROC curves and the agreement between the variables of the data set was measured with Kappa Score. The fit and test results support the accuracy values (Table 3, Table 4).

4) Query speed is of significant importance for effective user interaction. For this reason, structures that can be queried quickly should be used in similar systems. Django and Flask Frameworks were tried during the display of our work on the web application and the fastest query results were obtained with Flask. This situation may vary depending on the size of the data set.

5) An analysis was made according to the project characteristics, and the data were visualized, listed and reported. Thanks to the data categorized on the Kickstarter platform and with a certain success; detailed graphics, lists, reports and decision support were provided.

Future Works

There are many platforms in the crowdfunding ecosystem, aside from Kickstarter. The data on these platforms can also be analyzed and integrated into the system. Also, since the Ensemble algorithm approaches can generally classify better, they should be tested and monitored within the system. Apart from these, images containing project promotion can be evaluated within the scope of the web-based decision support system. In particular, it should be investigated whether the colors used correspond to the project category, whether the objects in the image express the project, and image processing should be done within the scope of the project's success.

Peer-review: Externally peer-reviewed.

Conflict of Interest: The authors have no conflict of interest to declare.

Grant Support: The authors declared that this study has received no financial support.

Author Contributions: Conception/Design of study: M.K., C.A.; Data Acquisition: M.K.; Data Analysis/Interpretation: M.K., C.A., Ç.T.; Drafting Manuscript: M.K.; Critical Revision of Manuscript: M.K., C.A., Ç.T.; Final Approval and Accountability: M.K., C.A., Ç.T.

References

Chen, K., Jones, B., Kim, I., & Schlamp, B. (2013). Kickpredict: Predicting Kickstarter Success. *Technical report, California Institute of Technology*.

Kickstarter Projects Stats (2020). Access address: https://www.kickstarter.com/help/stats

Shim, J. P., Warkentin, M., Courtney, J. F., Power, D. J., Sharda, R., & Carlsson, C. (2002). Past, present, and future of decision support technology. Decision Support Systems, 33(2), 111–126. doi:10.1016/s0167-9236(01)00139-7

Cheng, C., Tan, F., Hou, X., & Wei, Z. (2019, August). Success Prediction on Crowdfunding with Multimodal Deep Learning. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence, Macao, China* (pp. 10-16).

Statista (2020a). Overview of projects and dollars on crowdfunding platform Kickstarter as of July 2019, Access address: https://www.statista.com/statistics/251727/projects-and-dollars-overview-on-crowdfunding-platform-kickstarter/

Statista (2020b). Business Intelligence Software, Access address: https://www.statista.com/outlook/14230/100/business-intelligence-software/worldwide#market-marketDriver

Forbes (2017). State of Cloud Business Intelligence, Access address: https://www.forbes.com/sites/louiscolumbus/2017/04/09/2017-state-of-cloud-business-intelligence

Leiva, R. G., Anta, A. F., Mancuso, V., & Casari, P. (2019). A Novel Hyperparameter-Free Approach to Decision Tree Construction That Avoids Overfitting by Design. *IEEE Access*, *7*, 99978-99987.

Berhane, T. M., Lane, C. R., Wu, Q., Autrey, B. C., Anenkhonov, O. A., Chepinoga, V. V., & Liu, H. (2018). Decision-tree, rule-based, and random forest classification of high-resolution multispectral imagery for wetland mapping and inventory. *Remote sensing*, *10*(4), 580.

Yang, T. L., Lin, C. H., Chen, W. L., Lin, H. Y., Su, C. S., & Liang, C. K. (2019). Hash Transformation and Machine Learning-based Decision-Making Classifier Improved the Accuracy Rate of Automated Parkinson's Disease Screening. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*.

Huber, S., Wiemer, H., Schneider, D., & Ihlenfeldt, S. (2019). DMME: Data mining methodology for engineering applications–a holistic extension to the CRISP-DM model. Procedia CIRP, 79, 403-408.

Schäfer, F., Zeiselmair, C., Becker, J., & Otten, H. (2018, November). Synthesizing CRISP-DM and Quality Management: A Data Mining Approach for Production Processes. In *2018 IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD)* (pp. 190-195). IEEE.

Mouillé, M. (2018). Kickstarter Projects Dataset, Kaggle. More than 300,000 kickstarter projects (Version 7). Access address: https://www.kaggle.com/kemical/kickstarter-projects

Wiemer, H., Drowatzky, L., & Ihlenfeldt, S. (2019). Data Mining Methodology for Engineering Applications (DMME)—A Holistic Extension to the CRISP-DM Model. Applied Sciences, 9(12), 2407. doi:10.3390/app9122407

Chung, J., & Lee, K. (2015, August). A Long-Term Study of a Crowdfunding Platform: Predicting Project Success and Fundraising Amount. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media* (pp. 211-220).

Rao, H., Xu, A., Yang, X., & Fu, W. T. (2014, April). Emerging dynamics in crowdfunding campaigns. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction* (pp. 333-340). Springer, Cham.

Etter, V., Grossglauser, M., & Thiran, P. (2013, October). Launch hard or go home! Predicting the success of Kickstarter campaigns. In *Proceedings of the first ACM conference on Online social networks* (pp. 177-182).

Jensen, L. S., & Özkil, A. G. (2018). Identifying challenges in crowdfunded product development: a review of Kickstarter projects. *Design Science*, *4*.

Du, Q., Fan, W., Qiao, Z., Wang, G., Zhang, X., & Zhou, M. (2015). Money talks: a predictive model on crowdfunding success using project description.

Kindler, A., Golosovsky, M., & Solomon, S. (2019). Early Prediction of the Outcome of Kickstarter Campaigns: Is the Success due to Virality? *Palgrave Communications*, *5*(1), 1-6.

Yeh, J.-Y., & Chen, C.-H. (2020). A machine learning approach to predict the success of crowdfunding fintech project. Journal of Enterprise Information Management, ahead-of-print(ahead-of-print). doi:10.1108/jeim-01-2019-0017

Zvilichovsky, D., Inbar, Y., & Barzilay, O. (2015). Playing both sides of the market: Success and reciprocity on crowdfunding platforms. *Available at SSRN 2304101*.

Bi, S., Liu, Z., & Usman, K. (2017). The influence of online information on investing decisions of reward-based crowdfunding. *Journal of Business Research*, *71*, 10-18.

Kuppuswamy, V., & Bayus, B. L. (2018). Crowdfunding creative ideas: The dynamics of project backers. In *The economics of crowdfunding* (pp. 151-182). Palgrave Macmillan, Cham.

Ahmad, F. S., Tyagi, D., & Kaur, S. (2017). Predicting crowdfunding success with optimally weighted random forests. 2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions) (ICTUS). doi:10.1109/ictus.2017.8286110

Laudon, K. C. (2014). *Management Information Systems: Managing the Digital Firm*. Pearson Education India.

Mortensen, S., Christison, M., Li, B., Z hu, A., & Venkatesan, R. (2019, April). Predicting and Defining B2B Sales Success with Machine Learning. In *2019 Systems and Information Engineering Design Symposium (SIEDS)* (pp. 1-5). IEEE.

Pal, M. (2005). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, *26*(1), 217-222.

Xu, J., Zhang, Y., & Miao, D. (2020). Three-way confusion matrix for classification: a measure driven view. *Information Sciences*, *507*, 772-794.

Sasikala, B. S., Biju, V. G., & Prashanth, C. M. (2017, May). Kappa and accuracy evaluations of machine learning classifiers. In *2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)* (pp. 20-23). IEEE.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37-46.

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 159-174.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS quarterly*, 1165-1188.

Wang, J., Wu, X., & Zhang, C. (2005). Support vector machines based on K-means clustering for real-time business intelligence systems. *International Journal of Business Intelligence and Data Mining*, *1*(1), 54-64.

Cook, A., Wu, P., & Mengersen, K. (2015). Machine learning and visual analytics for consulting business decision support. In *2015 Big Data Visual Analytics (BDVA)* (pp. 1-2). IEEE.

Fahmy, A. F., Mohamed, H. K., & Yousef, A. H. (2017). A data mining experimentation framework to improve six sigma projects. In *2017 13th International Computer Engineering Conference (ICENCO)* (pp. 243-249). IEEE.

Mitra, T., & Gilbert, E. (2014, February). The language that gets people to give: Phrases that predict success on kickstarter. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing* (pp. 49-61).

Greenberg, M. D., Pardo, B., Hariharan, K., & Gerber, E. (2013). Crowdfunding support tools: predicting success & failure. In *CHI›13 Extended Abstracts on Human Factors in Computing Systems* (pp. 1815-1820).

Fawcett, T. (2004). ROC graphs: Notes and practical considerations for researchers. *Machine learning*, *31*(1), 1-38.

Adeniyi, D. A., Wei, Z., & Yongquan, Y. (2016). Automated Web Usage Data Mining and Recommendation System Using K-Nearest Neighbor (KNN) Classification Method. *Applied Computing and Informatics*, *12*(1), 90-108.

Jain, M., Narayan, S., Balaji, P., Bhowmick, A., & Muthu, R. K. (2020). Speech Emotion Recognition Using Support Vector Machine. *arXiv preprint arXiv:2002.07590*.

Saritas, M. M., & Yasar, A. (2019). Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification. *International Journal of Intelligent Systems and Applications in Engineering*, *7*(2), 88-91.

Szmigiera, M. (2019). Crowdfunding - Statistics & Facts, Statistica. Access address: https://www.statista.com/topics/1283/crowdfunding/

Roshan Joseph, V., & Vakayil, A. (2020). SPlit: An Optimal Method for Data Splitting. *arXiv e-prints*, arXiv-2012.