Journal of Innovation and Entrepreneurship

RESEARCH

Open Access

Predicting the success of startups using a machine learning approach

Mona Razaghzadeh Bidgoli[1], Iman Raeesi Vanani[2*] and Mehdi Goodarzi[1]

*Correspondence:
imanraeesi@atu.ac.ir

[1] Department of Technology Management and Entrepreneurship, Faculty of Management and Accounting, Allameh Tabataba'i University, Dehkadeh-Ye-Olympic, Tehran, Tehran Province, Iran
[2] Department of Operations and Information Technology Management, Faculty of Management and Accounting, Allameh Tabataba'i University, Dehkadeh-Ye-Olympic, Tehran, Tehran Province, Iran

Abstract

Successful investment in early-stage companies has high uncertainty. More specifically, the tools available to investors need to be more robust to reduce the risk and manage the uncertainty of startups. This research aims to use machine learning methods to design a prediction solution to identify successful startups for investors. In order to design the predicting solution and provide policies, classification, and clustering algorithms have been utilized to predict the success of startups and perform feature importance analysis based on the SHAP and permutation methods. Subsequently, the performance of four classification algorithms, such as Random Forest, Gradient Boost, Multilayer Perceptron, Logistic Regression and Support Vector Machine, are compared to predict business success. Meanwhile, Random Forest and Gradient Boosting algorithms showed the best accuracy, which was equal to 82% and 80%, respectively. Based on the feature importance of Random Forest and Gradient Boosting, which is obtained from the SHAP method, indicated that the higher values of "Number of followers on LinkedIn", "Number of employees on LinkedIn", "Number of followers on Twitter", and "Last raised amount" have higher SHAP values and a more significant impact on the model output. Three clustering algorithms including hierarchy, K-means, and DBSCAN were also compared. Among them, the K-means algorithm performs best with 72% silhouette, and K-means was employed to explain each cluster's characteristics. Finally, an effective artificial intelligence-based prediction solution has been proposed to show the way for investors to apply machine learning concepts to predict the success of startups.

Keywords: Startups, Machine learning, Prediction solution, Crunchbase

Introduction

As one of the main drivers of economic growth and innovation, startups play a vital role in today's economy (Hunt, 2013). As defined by authoritative authors such as Steve Block, a startup is a temporary organization that was formed in order to find a scalable and repeatable business model (Blank & Dorf, 2010).

Despite startups' power, they have a high failure rate within the entrepreneurial ecosystem. According to the reports of the European Angel Investor Association, approximately 50 million new projects have been launched annually (137,000 per day), but 90% of these projects do not succeed (Bednár & Tarišková, 2017). According to Picken, over 75% of companies manage a marginal existence or fail, even companies with a significant venture-backed. One reason may be that the environment in which startups operate, and grow is complex and presents many challenges to the founders. Furthermore, a common problem of startups is financing to start and run a business, which is one of the important limitations for founders (Korosteleva & Mickiewicz, 2011). Financing is one of the most essential resources for new businesses to become successful (Woods et al., 2020). The lack of credit history and reputation differentiates startups from established enterprises and raises the issue of financing for startup companies (Huyghebaert et al., 2007).

Financial problems of startups create a gap and; therefore, an opportunity for savvy investors that are willing to bet on early-stage companies with less validation than their more mature peers. Venture capitalists and accelerators are among the investors who invest in startups. Venture capitalists and accelerators specialize in financing startups in development and accept a high level of risk (Drover et al., 2017).

However, one of the most challenging aspects of startups is predicting their success in order to continue their existence (Żbikowski & Antosiuk, 2021). Also, investing in early-stage companies is extremely difficult, especially when no data is available to support the decision-making process (Corea et al., 2021). Data can be used to assist investors and founders in the decision-making process. There is exponential growth in the volume of data produced and available, and many businesses are constantly utilizing the latest information gleaned from the data. In an environment where uncertainty is high for startups, having a forecasting tool can be very effective (Arroyo et al., 2019). Hence, using machine learning might be an advanced strategy for investors and founders to predict future events based on enormous amounts of data.

Despite many studies that focus on predicting whether a startup will ultimately succeed or not, those studies did not have a comprehensive view. Therefore, we strive to take a more holistic view and design a prediction solution for investors. Also, previous research just determined the feature importance, but they did not analyze the impact of each feature importance on the success of startups. For this reason, we use the SHAP method and permutation feature importance to have an advanced analysis of features importance. According to previous research, questions of this study consist of the main research question and sub-questions that are as follows:

RQ: How is a prediction solution for investors to predict the success of startups using a machine learning approach?

> A: What clusters of startups are there and what are the characteristics of each cluster?
>
> B: Which machine learning-based classification algorithm can perform better in predicting the success of startups?
>
> C: What characteristics and indicators should startups have for achieving more success?

The purpose of this research is to use machine learning methods to design a prediction solution to identify successful startups to investors. Classification and clustering algorithms were used to design the prediction solution. Using machine learning classification algorithms, it investigates what factors and characteristics increase the success of startups in order to reduce the percentage of investment failure. In fact, in this research, the factors affecting the success of startups are examined from the perspective of investors so that they can make more successful investments. In addition, by using clustering, questions of what clusters of startups exist and how to predict the correct cluster of startups are answered. The outcome of this investigation will provide answers to the design prediction solution.

As a result of the machine learning classification and clustering algorithm, policies have been presented to assist startups in their efforts to succeed. These policies may facilitate the decision-making process for founders and investors and put them on the road to success. Further, the prediction solution is intended to introduce startups that are most likely to succeed to investors and provide operational advice for achieving success in each cluster. The main objective of this study is an improved data-driven and machine-learning framework in order to provide investors with a more accurate method of selecting successful startups through designing the prediction solution.

Startup

The origin of the startup dates back to 1974. This history is related to the era when small and innovative companies emerged in developed countries and created a great transformation in traditional markets. At that time, the term "startup" referred to a number of small and advanced companies in the market, mostly operating in the fields of technology (Skawińska & Zalewski, 2020).

What differentiates startups is their product or service, innovative business model and rapid growth. Also, startups are important engines for creating jobs which play a key and vital role in the innovation and economic growth of the world. Previous research has shown that promoting startups is a key action point to encourage economic growth (Cavallo et al., 2019). Previous research has shown that promoting startups is a key action point to encourage economic growth (Cavallo et al., 2019). They contribute to the development of traditional sectors by combining innovation and knowledge (Blank, 2018).

Machine learning

As defined by Arthur Samuel, machine learning is the study of how computers can learn without being explicitly programmed. The goal of machine learning (ML) is to teach machines how to handle data more efficiently (Mahesh, 2020). Machine learning is the programming of machines to optimize a performance measure using past data or experience. There is a defined model with some parameters, and learning to run a machine to optimize the model parameters must use training data or past experience. This model may be for future prediction or a description for gaining knowledge from data (Alpaydin, 2020).

Literature review on startup success prediction using machine learning

Funding is a frequent problem for startups, and at the same time, investors face many challenges when investing in them. Unfortunately, the tools currently available to investors are not robust enough to reduce investment risk and help them make investment decisions. Meanwhile, methods based on machine learning can fill this gap (Arroyo et al., 2019). By using machine learning tools, Investors can make more informed investments.

On the other hand, investors can better identify unknown startups and startups with high potential by using machine learning tools. Existing literature primarily focuses on using classifier algorithms to analyze startup data and identify key factors influencing success.

Classifier algorithms for startup success prediction

Focus on Startup Survival and Identifying Key Investment Factors: Ghassemi et al. (2020) utilized logistic regression to predict startup survival, finding that teams with diverse backgrounds and targeting niche markets are more likely to succeed. The results of their research showed that teams with diverse professional and academic backgrounds are more likely to survive and that ideas targeting established markets are less likely to survive in a competitive market. Similarly, Bai et al. (2021) researched the issue of which criteria related to startups are more important in investors' decisions to invest. Their findings show that "planning strategy" and "team management" determine factors in the company's investment decisions.

Finding Feature Importance for Startup Success through Using Crunchbase Database: Arroyo et al. (2019) used classifier algorithms in their research. They used more than 120,000 companies in the initial stage through site Crunchbase. The primary objective of this project was to create and assess a data-driven approach for investors to identify and endorse the best companies to support. For this purpose, they divided their dataset into used warmup and the simulation window. The warmup window was considered 4 years. The simulation window considered 3 years and represented whether the company determines its success or not. They utilized multi-class predictions such as acquired, funding round, IPO, closed, and no event. This approach caused them to focus on early-stage companies, and apart from IPO, they considered multi-class to predict the success of startups. Using this approach for selecting data is a realistic setting to predict the success of startups compared to another research. They found that variables such as LinkedIn, the number of founders, different nationalities of the founders, and the lifespan of the startup increase the success of startups.

Corea et al. (2021) investigated venture capital investment decisions through the Crunchbase dataset, identifying 21 relevant features for evaluating startup success. Their research emphasized the importance of social media presence, previous investors and their reputations, and funding amount. A similar study used six classifier algorithms as machine learning models to forecast the success of startups and determine the key features that play a significant role in making such predictions. Their dataset consists of information companies in Crunchbase (Kim et al., 2023). They used feature importance to predict startup success, and their results indicated that media exposure, monetary funding, industry convergence level, and industry association level have vital roles in startup success. To determine success, they considered only IPO. They considered that IPO determined that a company is successful, and if companies achieve to IPO they are successful.

The results of Kim et al. (2023) research is consistent with the results of Arroyo et al. (2019) and Corea et al. (2021). In their results, all three of these researchers concluded that factors exposure and monetary funding are among the factors that affect the success of startups. These two factors are common in research and are considered to indicate the validity of previous research.

As is shown in Table 1, most of the approaches reviewed are based on classifier algorithms, and they utilized the Crunchbase dataset. However, fewer works have been explored based on cluster algorithms and design prediction solution.

As a result, that the literature reveals the potential of using machine learning to exploit Crunchbase data to predict successful startups. However, we consider that previous approaches focus on classifier algorithms, which are not enough for investors to have high performance to predict the success of startups. The following section presents our approach and differences in using machine learning algorithms to help investors evaluate the potential of startups.

Beyond classifier algorithms

Classifier algorithms have been widely used as promising ML methods for forecasting the success of startups and guiding investment choices. However, few studies have explored the use of clustering algorithms and prediction solutions to group startups based on similar characteristics. For instance, Skawinska et al. (2020) focused on key success factors of startups in the European Union, utilizing clustering algorithms for the

Table 1 Comparison of previous research

References	Dataset	Model	Model performance
Holmes et al. (2010)	Newly established manufacturing firms in north-east England	Log-logistic	–
Hoenen et al. (2012)	Biotechnology US companies	Linear Regression	–
Krishna et al. (2016)	Crunchbase	Random Forest	AUC = 0.972
Tomy and Pardede (2018)	2013 Victorian ICT Industry Statistics survey (Australia)	Naive Bayes	Accuracy = 77.5
Johnson et al. (2019)	SBIR	Random Forest	AUC = 96%

Zhang et al. (2019)	Beijing Innofund	Support Vector Machine	Accuracy = 86%
			F-score = 82%

Arroyo et al. (2019)	Crunchbase	Gradient Tree Boosting, Random Forests, Extremely Randomized Trees	Accuracy GTB = 82.2% Accuracy FR = 81.8% Accuracy ERT = 81.9%
Ghassemi et al. (2020)	MIT $100K Entrepreneurship Competition	Logistic Regression	AUC = 72%
Corea et al. (2021)	Crunchbase	Gradient Tree Boosting	–

Żbikowski and Antosiuk (2021)	Crunchbase	XGBoost	Accuracy = 85%
			F-score = 0.43

Thirupathi et al. (2021)	SBIR/STTR and Crunchbase	XGBoost	Accuracy = 84%
Ross et al. (2021)	Crunchbase, USPTO	Random Forest MLP XGBoost Ensemble	Accuracy RF = 0.88 Accuracy MLP = 0.88 Accuracy XGBoost = 0.89 Accuracy Ensemble = 0.89
Kim et al. (2023)	Crunchbase	Random Forest Gradient Boost Decision Tree Support Vector Machine	Accuracy FR = 87% Accuracy GB = 85% Accuracy SVM = 88% Accuracy DT = 86%

cluster analysis of countries. Their results showed that the structure of German startups is most similar to the European Startups Mean, while Romanian startups differ dramatically from all others. Additionally, Ahluwalia et al. (2021) examined the effect of venture capital clusters on the exit of a startup through mergers and acquisitions (M&A). Their findings indicated that when a startup is in a venture capital (VC) cluster, the probability of a successful exit increases. Therefore, more research is required to investigate ML algorithms for prediction and clustering applications, and to address the constraints of classifier algorithms. In this study, we have employed a combination of classification, clustering, and prediction solution algorithms.

Research method

Researchers utilize a diverse of methodological approaches in the field of AI. Also, in both academic and industrial areas there is no standard methodology for AI, and related to the type of problems and the solution is required (Aleisa et al., 2023). Beyond this, there have been few efforts to advance methodological guidelines describing the general process (François, 2008). In general, four main data mining guidelines consist of Fayyad's methodology or Knowledge Discovery in Databases (KDD), Sample Explore Modify Model Assess (SEMMA) methodology, Cios methodology, and cross-industry standard process for data mining methodology (CRISP-DM).

Depending on the problem, the research methodology should be chosen. This study focuses on using machine learning in order to design a prediction solution to predict the success of startups. In this research, we choose CRISP-DM as the research methodology. CRISP-DM consists of a cycle that comprises six stages that consist of business understanding, data understanding, data preparation, modeling, evaluation, and deployment (Azevedo & Santos, 2008). For the reason that CRISP-DM methodology considers all aspects of the project from initial business understanding to final deployment, and it consists of the complete approach. CRISP-DM is one of the latest research methods in the field of artificial intelligence. Because of the reversibility of the steps and it was important to business understanding concepts in this CRISP-DM method, we used this method as our research method.

CRISP-DM

Several data mining methodologies have been proposed to systematize the discovery of knowledge from data, including the Cross-Industry Standard Process for Data Mining (CRISP-DM), which is very complete and well documented (Clifton & Thuraisingham, 2001). It consists of a cycle and six stages. All his steps are properly organized, structured, and defined, which allows a project to be easily understood or revised (Santos & Azevedo, 2005). The CRISP-DM method has been widely accepted as the best methodology for data mining (Piatetsky, 2014; Vanani & Jalali, 2018). Furthermore, the CRISP-DM methodology is implemented as an existing stepwise model in this study to develop a comprehensive approach for modeling themes. Figure 1 shows the research model based on phases of the CRISP-DM methodology for designing a prediction solution for investors to predict the success of the startup research model.

Every data mining project begins with the definition of the project's goals, which is determined during the first phase, "Business Understanding". The purpose of this

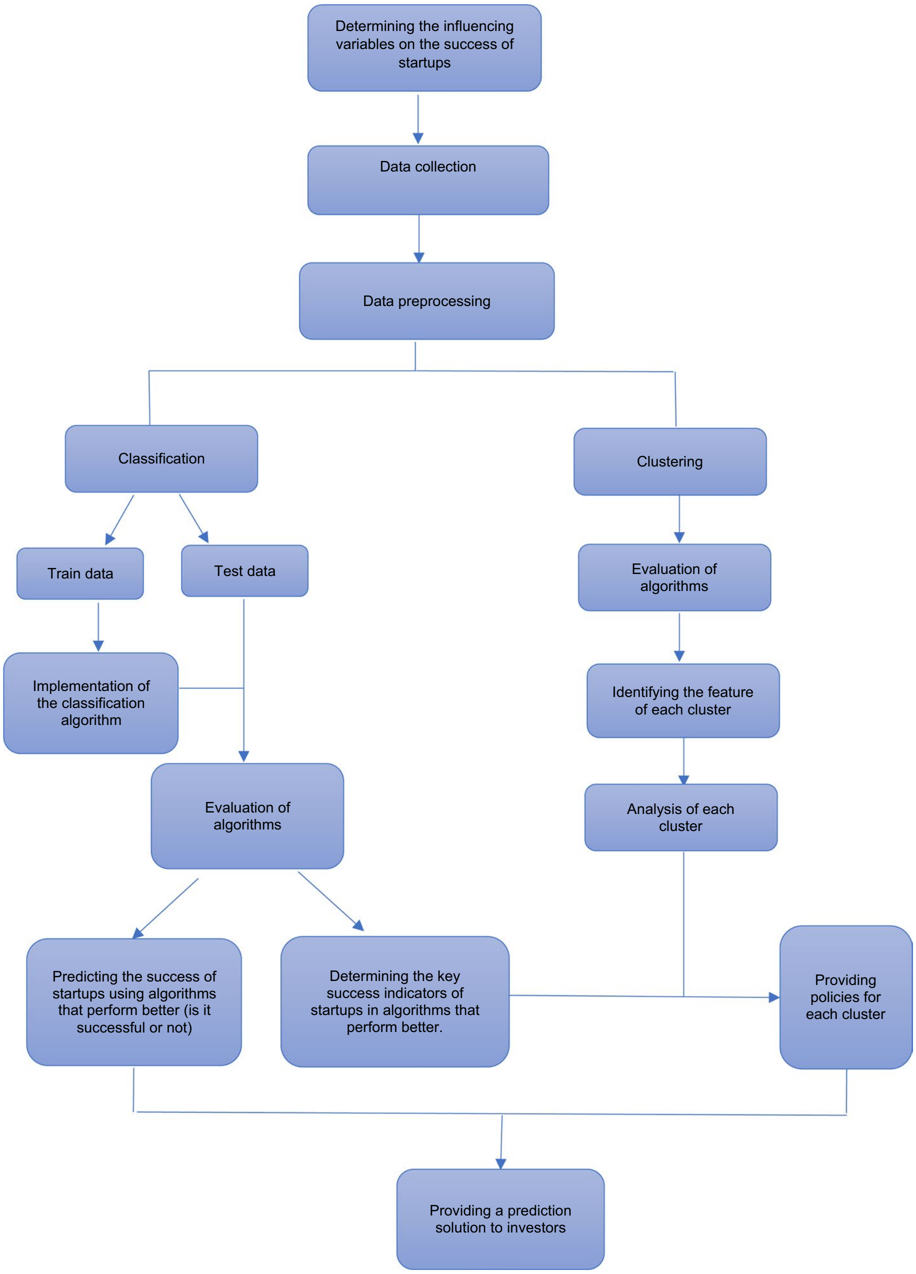
Determining the influencing variables on the success of startups
Data collection
Data preprocessing
Classification
Clustering
Train data
Test data
Evaluation of algorithms
Implementation of the classification algorithm
Identifying the feature of each cluster
Evaluation of algorithms
Analysis of each cluster
Predicting the success of startups using algorithms that perform better (is it successful or not)
Determining the key success indicators of startups in algorithms that perform better.
Providing policies for each cluster
Providing a prediction solution to investors

Fig. 1 Research steps based on CRISP-DM

research is to design a prediction solution, policy and determine the key success indicators of startups. The focus of this research is to create a tool for investors so that they can better evaluate companies. The "Data Understanding" phase involves collecting, describing, and understanding data related to the data mining project objectives. In this research, a period of 7 years has been considered for the data. This 7-year period includes two 4-year periods (March 2014 to March 2018) and 3-year periods (March 2018 to March 2021). Startups that were established between March 2014 and March 2018 are being considered. The data were collected through crunchbase.com and social networks.

"Data preprocessing" is one of the main principles of this research. The datasets produced during the preparation phase are used for modeling or the major analysis work of the project. There is a close relationship between modeling and preprocessing. As part of the "Modeling" phase, the preprocessed data set is analyzed using various data mining techniques. In modeling, clustering and classification algorithms have been used in this research. Clustering algorithm is used to determine the characteristics of each cluster. For clustering, the number of suitable clusters must be determined first. Subsequently, the condition of the clusters has been investigated. Classification algorithms have been implemented to determine the key success indicators of startups. According to the results, classification and clustering algorithms have been used to provide policies and prediction solution design.

As part of the "Evaluation" phase, the constructed model and the steps involved in its construction are evaluated for suitability and achieving the project's goals. The classification algorithms were checked using accuracy, precision, recall, and other related criteria in the evaluation part. For clustering algorithms, it has been evaluated how well the clustering algorithm works based on the number of clusters, which used the silhouette method. Once the model has been evaluated and proven successful, it is deployed to implement and develop in the "Deployment" phase.

Data collection

Our data are extracted from Crunchbase.com, one of the largest and most complete startup ecosystems databases, which includes data on investors, companies, key people, and events. The data of this research were gathered manually.

We created a unique dataset of startups by combining data from multiple databases. The major source of our data is created from Crunchbase. The startups were considered invested by the Y Combinator. Since the launch of Y Combinator, which is the first cohort of eight ventures in 2005 and is one of the best ventures in the world of entrepreneurship, we consider this category of startups. In this research, companies invested by Y Combinator between March 2014 and March 2018 are extracted, and then the main details of each startup are exploited from the Crunchbase database. In the end, in order to achieve more detailed variables, we utilize various social media like Facebook, Twitter, and LinkedIn. To prevent data from being dispersed, we consider limited data on the early-stage companies to focus on the evaluation of variables separately. In addition, to have a comprehensive dataset, the data are selected from various categories and types of companies.

Our data selection method is based on research by Arroyo et al. (2019). As Fig. 2 shows, we consider 7-year period for startup, which is between March 2014 and March 2021. This 7-year period is divided into 4- and 3-year periods. In the 4-year period, the startup performance is investigated. The variables are checked during this period. In addition, startups must have been established in this period. As a result, the list of startups that were established in the period from March 2014 to March 2018 is extracted. After that, in the second 3-year period, in fact, in the period from March 2018 to March 2021, it was checked whether this startup was among the successful or failed startups. At this point, we used the summary and financial sections of each startup in Crunchbase. In general, our data include 400 startups. Through the

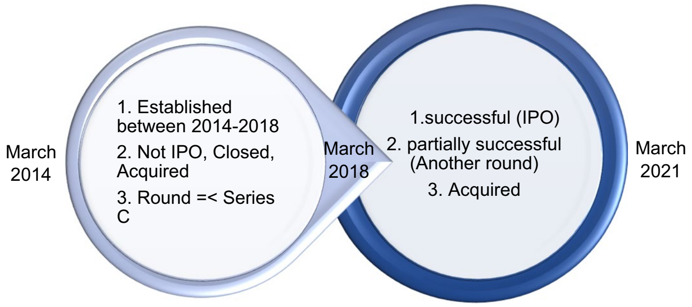
March 2014
1. Established between 2014-2018
2. Not IPO, Closed, Acquired
3. Round =< Series C
March 2018
1.successful (IPO)
2. partially successful (Another round)
3. Acquired
March 2021

Fig. 2 Selection of startups

website Crunchbase.com, information related to the brand names of the founders, investments, and patents was extracted.

First, startups that were established between March 2014 and March 2018 were considered. The startups that have reached the higher stage of the C series and have been closed, IPO, and acquired during this period have been excluded. Because they are not interested in early-stage investment.

The next step was to determine the startups that could be successful and startups are successful. IPO is the terminal goal of investors. However, achieving companies to the IPO level is rare, and 30% of seed-funded companies exited through an IPO (Insights, 2018). Investors seek out companies that will progress towards fresh infusions of funds and potentially greater results. Also, they looking for companies to spend money quickly to grow faster automatically (Arroyo et al., 2019). Therefore, in addition to successful startups, we considered partially successful startups and acquired startups for better evaluation for investment. Startups that reached the stage of IPO between March 2018 and March 2021 are considered successful companies. Moreover, companies that have reached another round of funding in this period are considered partially successful startups. The last category is startups that have been acquired. Among the 400 collected data, 232 data are considered the startups that could be successful and startups are successful; also, 168 of them are unsuccessful startups.

As we said, the dataset for this research was manually gathered from Crunchbase. In order to extract the founders' information, we extracted data from the summary section of each startup in Crunchbase. In the summary section, the name of the founders is visible. To extract the details of funders, we utilized the LinkedIn of each founder and Crunchbase. For the derivation of funding information, we used the financials section of each startup in Crunchbase. This section specifies the number of rounds, investor, lead investor, and founding for each round.

To achieve the company information, we used the social media of each startup to extract the followers on Facebook, Twitter, and LinkedIn. Another section of LinkedIn is people; in this section, we achieved number of employees. Also, in the technology section of each startup, the number of inventions and trade name are shown. To specify the type of company and the age of the startup, we used the summary section of each startup in Crunchbase. The types of startups are divided into nine categories including Real Estate and Construction, Industrials, Education, Financial, B2B Software and Services, Government, health care, consumer, and Unspecified.

Feature selection

According to previous research on venture capital criteria and the restrictions in the access of the crunchbase.com site, 37 variables are chosen as the input feature for the training of machine learning algorithms. Generally speaking, the predictor variables in this work contain three main parts, including information related to the founders of the companies, data related to the financing of the companies, and characteristics associated with the organization in Table 2.

The Founders Information category includes information about the people who founded a company. In their choice of investment, venture capitals often take into account entrepreneurs' education and industry experience (Kim & Lee, 2022). Ughetto (2016) indicates that the experience of the entrepreneur in having a previous firm positively affects focal firms' growth. For this reason, information about the founders has

Table 2 Input feature selection based on the information from founders, founding, and companies

Category	Variable	Definition
Founders information	Founders_count	Number of people who founded that startup
	Founders_dif_nationality	Different nationality of the founders
	Founders_entrepreneurial_background	Number of founders who have an entrepreneurial background
	Founders_male_count	Number of founders who are male
	Founders_female_count	Number of founders who are female
	Founders_degree	Education level of the founders
	University	
Funding information	Round_count	The number of funding rounds obtained between 2014 until March 2018
	Investor_count	The total number of investors in all funding rounds between 2014 until March 2018
	Lead_investor_count	The total number of lead investors in all funding rounds between 2014 until March 2018
	Total_founding_amount	The accumulated budget all funding rounds between 2014 until March 2018
	Last_round_Investors_count	The number of investors in the last funding round
	Last_raised_amount	The amount of capital raised in the last funding round
	Timelapse_until_fifth_year	The distance between the last funding round to the beginning of the fifth year
Company information	Facebook	Number of followers on the Facebook page of startups
	Twitter	Number of followers on the Twitter page of startups
	LinkedIn	Number of followers on the LinkedIn page of startups
	Employs	Number of employees on LinkedIn
	Trade name	The number of trade names that the company has registered
	Invention	The number of invention that the company has registered
	Type of company	Determining the type of company according to 9 specified groups
	Startup-age	The number of years since its establishment between 2014 until 2018

been considered in this research because their information can be very influential on the output results.

The variables of the Funding Information category relate to financial events that occurred between March 2014 and March 2018. According to several previous research, we use Crunchbase.com data as one of our sources with the main focus on early-stage companies. For this reason, the age of startups is important. Due to the selection of a specific time frame to select startups and determine their success, the time in this section has been carefully examined.

Company information consists of characteristics related to the organization. Early-stage companies are increasingly, using social media to communicate with their target stakeholders such as customers and investors.

Data preprocessing

For data preprocessing, we used three steps which include "Data cleaning", "Data integration", and "Data transformation". "Data cleaning" is the procedure to remove incomplete, incorrect, and inaccurate data from the datasets. In the data cleaning section, one of the concerns of the data cleaning process is due to the removal of startups that do not have all the necessary information considered in this research. Since much of the data in this research was extracted from the crunchbase.com database and the data entered into this site are composed by users, there are a large number of missing values. As such, we do not have the necessary information, which has led to the removal of some startups.

"Data integration" is the process of combining multiple sources into a univalent dataset. In the part of data integration, by analyzing and examining different startups, it is trying to identify new variables and features. The information on the considered variables was extracted from Crunchbase.com. Additional information has been collected through the startups' own social networks. In fact, new data sets will be created from multiple sources to consider more variables for startup success.

"Data transformation" is one of the basic steps in data processing. In this part, first, non-numerical data are converted into numerical data. Then the data are normalized between 0 and 1. For each attribute, the minimum value of that attribute is set to 0 and the maximum value is set to 1. The other point in this section is the significance of the interquartile range (IQR), which is utilized to eliminate the dispersal of data. In order to calculate the IQR, data should be grouped according to quartiles. These quartiles are indicated by Q_1 which corresponds with the 25th percentile and Q_3 corresponds with the 75th percentile. The amount of interquartile range is determined by using this formula: $IQR = Q_3 - Q_1$. After finding the IQR, the value "$Q_{(1,3)} - 3 \times IQR$" was obtained. The data are lowest or highest this value ($Q_{(1,3)} - 3 \times IQR$) should substituted with the mean value.

Data modeling

In the modeling of our research, clustering and classification algorithms have been used to provide policy to investors and design a prediction solution. Figure 3 shows the experiment schema of model process. In the following, it is explained how to use data classification and clustering algorithms for the purposes of this research.

- Classification algorithms

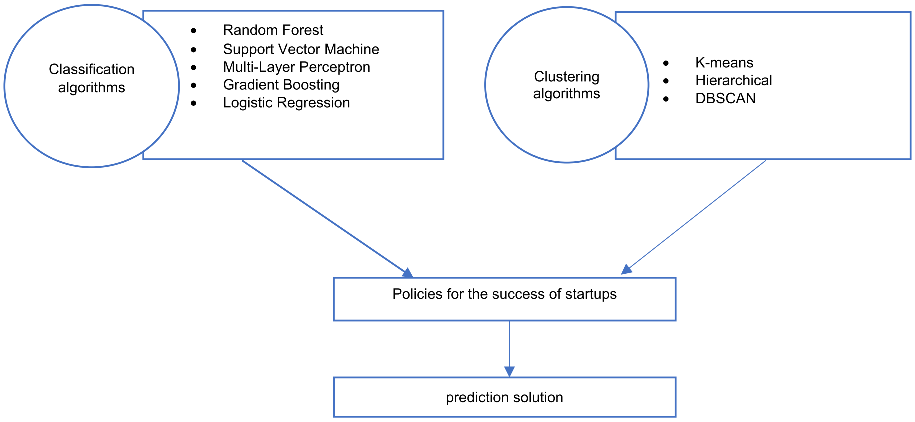

Fig. 3 Schema of model process

Classification algorithms have been chosen to predict the success of startups and to determine key success indicators. Algorithms were chosen based on their suitability to the problem, and these algorithms show better results. We have considered four different machine learning classifier algorithms, including Random Forest (RF), Gradient Boosting (GB), Logistic Regression, and Support Vector Machine (SVM). In addition to machine learning algorithms, deep learning algorithm has been used that includes Multilayer Perceptron (MLP).

Gradient Boost and Random Forest consist of ensemble algorithms. We consider ensemble algorithms because it has been successfully used in the previous research (Arroyo et al., 2019; Kim et al., 2023; Rossi et al., 2020). As a complement to the ensemble classifiers, we use a classification method based on a different paradigm which is Support Vector Machines; in addition, this method has already shown good results in plenty of fields, including our research area (Kim et al., 2023; Zhang et al., 2019).

Moreover, we utilize Multilayer Perceptron (MLP) as a neural network algorithm which consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Neural networks have been used effectively in several applications and solve problems in many areas, such as classification, pattern recognition, and image processing (Turkoglu & Kaya, 2020). Considering that MLP is less commonly used in previous research and its performance may not be as well-known as other classification algorithms, it is important to thoroughly evaluate its effectiveness in investment research.

In order to evaluate the performance of classification algorithms, the input dataset is randomly divided into the training dataset (80%, 320 records of data) and the test dataset (20%, 80 records of data). By dividing the data into training and test datasets, we can assess the performance of the trained algorithms by evaluating their predictions on the test dataset. This allows us to determine if the algorithms are capable of making accurate predictions on unseen data.

Evaluating classification algorithms is judged based on their confusion matrix. The confusion matrix, consisting of the actual classifications and the predicted classifications, typically requires four values to assess performance. True positive (TP): the predicted classification is as positive as the actual classification. True negative

(TN): both predicated and positive classifications are negative. False positive (FP): the predicted classification is positive while the actual classification is negative. False negative (FN): the predicted classification is negative and the actual classification is positive (Fahmy Amin, 2022).

As shown below, recall presents the percentage of real positive cases that are correctly predicted positive. This metric is known as the main source due to the fact that the aim of recall is to identify all real positive cases, and it is the basis of ROC. In contrast, the precision of the proportion of predicted positive cases that are correctly real positives (Powers, 2020). The F1_score is described as the harmonic mean of precision and recall. Accuracy presents the percentage among the correctly predicted and all the cases in the dataset (Chicco & Jurman, 2020):

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}},$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}},$$

$$\mathrm{F1\_score} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}},$$

$$\mathrm{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}.$$

The other metrics used to investigate the performance of classification algorithms are ROC and AUC. The AUC and ROC can be applied in decision-making processes to determine the best model. The ROC curve illustrates the performance over an arrangement of thresholds and can be briefed by the area under the curve (AUC) by using a single number (Muschelli III, 2020).

- Clustering algorithms

  Clustering algorithms create clusters that have the most similarity within each group and the least similarity between groups. The clustering algorithm produces high-quality clusters. Also, this algorithm is an unsupervised method. In this research, we use clustering algorithms to categorize and determine the characteristics of each cluster. The status of each cluster has also been examined. We have considered three different machine learning cluster algorithms, such as K-means, hierarchical, and DBSCAN.

  The silhouette method is used to evaluate the final quality of the clustering method. This method is used to find the mean silhouette co-efficient of all the samples for different numbers of clusters. The highest silhouette score demonstrates the ideal number of clusters (Shahapure & Nicholas, 2020).
- Policies for the success of startups

According to the results of the combination of clustering and classification algorithms, macro-policies have been written for the success of startups. Also, based on macro-policies, operational recommendations are written for each cluster.

- Prediction solution
  We have designed a prediction solution for investing in early-stage companies. In fact, choosing a for-profit company for investment can be difficult, especially in relation to companies that do not have complete information available. Therefore, the existence of a prediction solution can help investors make better decisions for choosing startups. The designed prediction solution suggests startups that have a higher success rate to investors and provides operational recommendations for each cluster.

Evaluation

A comparison of algorithms and a final summary of each model's performance criteria have been discussed, in this section. As shown in Table 3, four classification algorithms have been examined and compared based on six criteria: accuracy, precision, recall, F1_score and AUC.

Recall measures the proportion of successful startups that are correctly classified among all truly successful startups. Recall is also an important criterion for venture capitalists since a higher recall value will lead to fewer missed opportunities for investors. As shown in Table 3, random forest and gradient boost have a higher value of recall. Concerning this criterion, tree-based algorithms have demonstrated the best performance. The precision measure shows the true success rate of a startup. The other significant measure for the evaluation of algorithms is accuracy. Random forest and gradient boost algorithms illustrate the highest value in recall, precision, F1_score and accuracy metrics among the other algorithms, they have higher precise and authentic performance.

Additionally, since random forest and gradient boosting algorithms have higher performance, the ROC curve of these algorithms is investigated. Figure 4 provides a comparison ROC curve for them. ROC curve plots the proportion rate of the true positive rate into false positive rate. The area under this curve is the AUC, which is actually the ability of algorithms to distinguish between classes. AUC is used for more analysis. Under AUC criteria, random forest dominates the rest of the models. Similar to the evaluation under previous metrics, the gradient boost algorithm is the second-best-performing model under AUC.

Random Forest and Gradient Boosting are among the most robust classifiers for predicting startup success since it outperforms other classifiers in all selected metrics. Overall, Random Forest and Gradient Boosting performance metrics show high results.

Table 3 Comparison of results from classification algorithms

	Recall	Precision	F1_score	Accuracy	AUC
RF	0.92	0.82	0.87	0.82	0.79
Gradient Boost	0.96	0.78	0.86	0.80	0.73
MLP (deep learning)	0.80	0.75	0.77	0.73	0.72
Logistic Regression	0.79	0.76	0.77	0.73	0.71
SVM	0.77	0.76	0.76	0.71	0.69

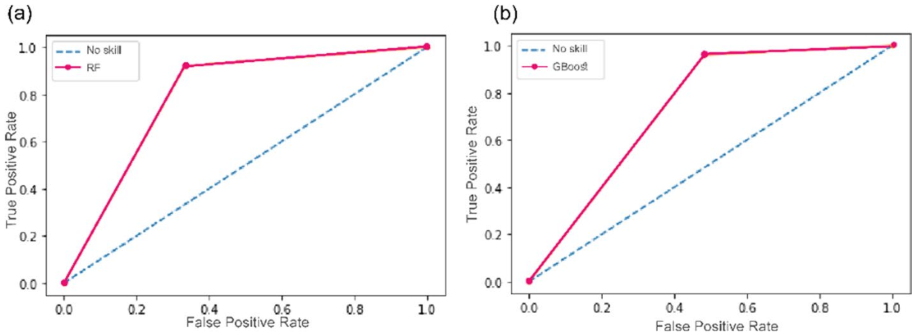

Fig. 4 Comparison of ROC curve. **a** Roc curve for *Random Forest algorithm*. **b** Roc curve for *Gradient Boost algorithm*

Table 4 Comparison of clustering algorithms

	Cluster	Silhouette
K-means	4	0.72
Hierarchical	4	0.71
DBSCAN	5	0.35

Furthermore, Random Forest and Gradient Boosting are appropriate for datasets with sparse data as well as categorical variables, and they reveal the significance of each feature variable. Therefore, classifier algorithms can provide a reliable prediction of startup success and be utilized to aid investment decisions in venture capital.

For evaluating the performance of clustering, it is necessary to determine the number of suitable clusters and the best clustering method. We used the K-means, hierarchical, and DBSCAN algorithms as clustering algorithms. By using the silhouette score cluster algorithms are examined and compared. Table 4 presents the comparison results of these three algorithms.

According to Table 4, the K-means algorithm has shown the best silhouette among the implemented clustering algorithms.

Results and discussion

Classifier

Among the implemented classification algorithms, Random Forest and Gradient Boosting algorithms showed the best accuracy. In this research we used SHAP and permutation methods to investigate the feature importance, and SHAP method had not been explored before in the literature. SHAP by Lundberg as a unified framework for interpreting predictions, is currently the most widely used approach for estimating the importance of input predictors, is (Lundberg & Lee, 2017). The advantage of SHAP is that based on the Shapley value which is theoretical results, showing that the Shapley value is the unique distribution solution with a set of desirable axioms such as symmetry and linearity (Roth, 1988). The permutation feature importance can apply to trained ML models and is commonly used to define how input features interact with the model output (Molnar et al., 2023).

Figure 5 presents the feature importance of Random Forest which provides the possibility to evaluate the feature importance using the SHAP method. Feature importance refers to the concept of which features contributed to the decision, not the extent of their influence on the target variable. The importance of the attribute can be measured.

Figure 5a displays the influence of each input feature on the model output for the whole 400 data using the SHAP value; in the order of feature importance, colored based on the features' values. This indicates that the higher values of LinkedIn, Employs, Twitter, Last raised amount, and Lead investor count possess higher SHAP values with higher impact on the model output. In contrast, the lower values of Timelapse until the fifth year, Startup-age, Last round Investors count, and Investor count possess higher SHAP values. In addition to SHAP values which indicate the impact of each input features on the model output, there are other opportunities to evaluate the feature as well.

Figure 5b shows the mean SHAP values of feature importance. LinkedIn's mean value as the most important feature is equal to 0.11. In Fig. 5c, the heatmap of SHAP shows the value of input feature for the whole 400 data instances, in the order of feature importance. The black bar plot displays the average SHAP values for each input feature across the data instances.

Besides the SHAP method, permutation feature importance is also utilized. Interestingly, certain company sectors appear more appealing and attractive. Although the type of company does not comprise the nine important features identified by the SHAP method, two types of companies exhibit higher importance values compared to the others. According to the permutation feature importance analysis conducted using the Random Forest algorithm, it can be concluded that among the nine categories of company types, "B2B Software and Services" and "Consumer" hold higher value and significance.

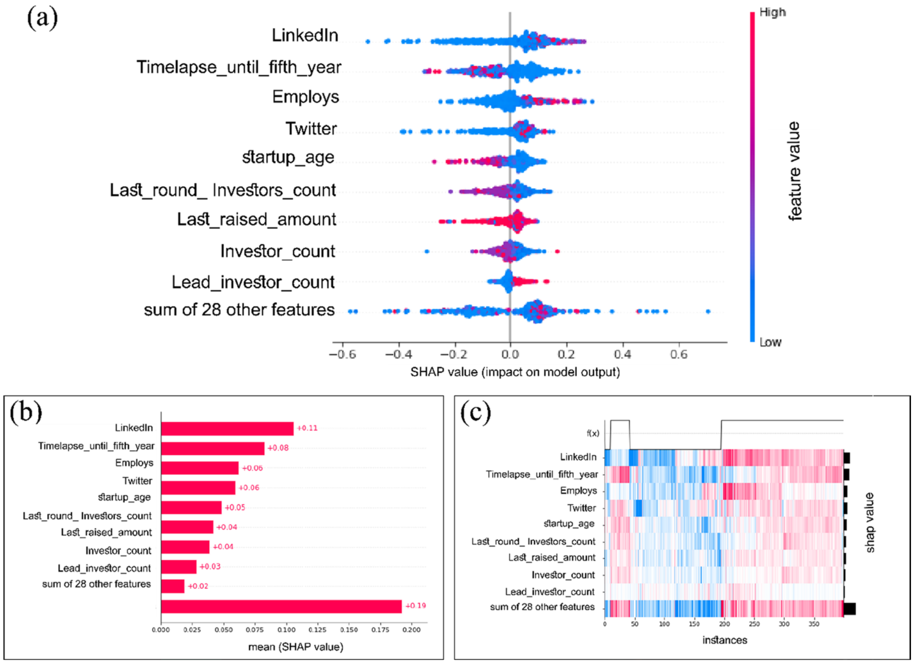

Fig. 5 The important features of Random Forest obtained from SHAP method. **a** The influence of each input feature on the model output. **b** The average of SHAP values. **c** The heat map of SHAP values over input data

Their feature importance score of “B2B Software and Services” and “Consumer”, respectively, are 0.01181 and 0.0139, while the other seven categories are lower than 0.00700. These findings suggest that these two company sectors have a greater impact on the prediction model and are more influential in determining investment attractiveness and can have better results.

In Fig. 6, the feature importance of the Gradient Boosting algorithm is discussed. By using the SHAP method, features importance and effective features have been identified in this algorithm. The importance of features addresses the concept of which features have played a role in decision-making and shows the feature’s importance can be measured.

Figure 6a shows the influence of each input feature on the model output. The influence of each input feature on the model output for the whole 400 data using the SHAP value; in the order of feature importance, colored based on the features’ values. This indicates that the higher values of LinkedIn, Employs, Founders count, Total funding amount, Twitter and Last raised amount possess higher SHAP values with higher impact on the model output. In contrast, the lower values of Timelapse until the fifth year, Last round Investors count, and Investor count possesses higher SHAP values. In addition to SHAP values which indicate the impact of each input feature on the model output, there are other opportunities to evaluate the feature.

As shown in Fig. 6b, the mean value of “Timelapse until fifth year”, the most important feature, is 0.13. In Fig. 6c, the heat map of SHAP indicates the value of the input feature for the whole 400 data instances, in the order of feature importance.

Furthermore, according to permutation feature importance values, “B2B Software and Services” and “health care” have the highest value among the other nine categories of

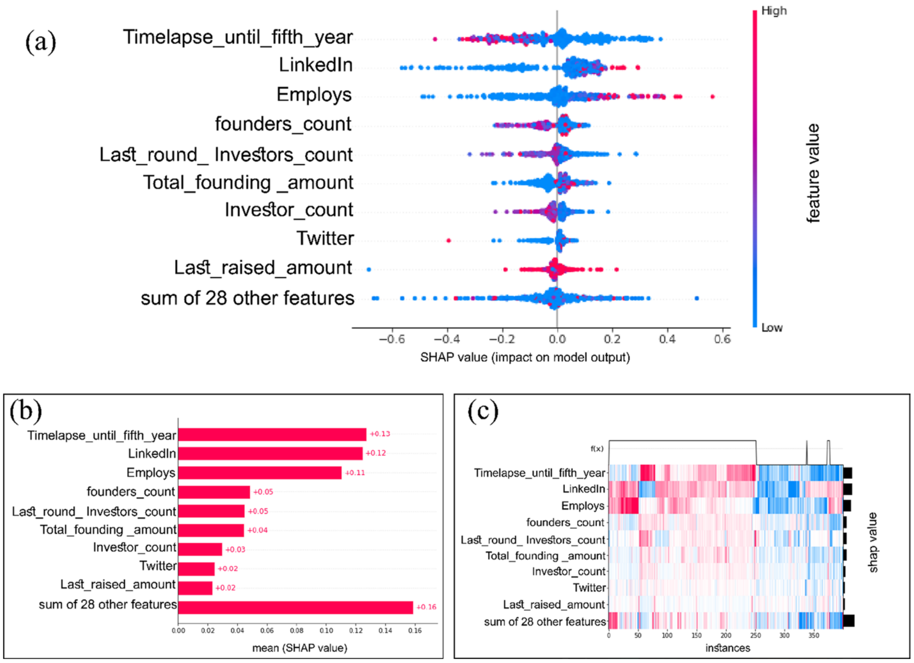

Fig. 6 Features important of Gradient Boost obtained from SHAP method. **a** The influence of each input feature on the model output. **b** The average of SHAP values. **c** The heat map of SHAP values over input data

companies. Their feature importance score of “B2B Software and Services” and “health care”, respectively, are 0.00468 and 0.00414, while the other seven categories do not have any score. This suggests that the “B2B Software and Services” and “Consumer” are more effective in investing and can have better results. Thus, these types of companies are more popular.

Clustering

In this research we used cluster algorithms, and cluster algorithms had not been explored before in the literature. Among the clustering algorithms, the K-means has shown the best accuracy, and in the following, the features of each cluster of the K-means clustering algorithm have been investigated. In Fig. 7, the K-means clustering algorithm is further explained and illustrates the status of each cluster.

Highly ranked investors in the last round (HIL): The average number of founders of cluster 1 is equal to 2.29. Also, the average number of male founders of this cluster is equal to 2.006 and the number of female founders is equal to 0.278. Regarding education degrees, the majority of the founders in this cluster have other education degrees besides master’s degrees and doctorates. Regarding the financial part of this cluster, it has an average number of 1.7 funding rounds with 6.2 investors and 0.8 lead investors.

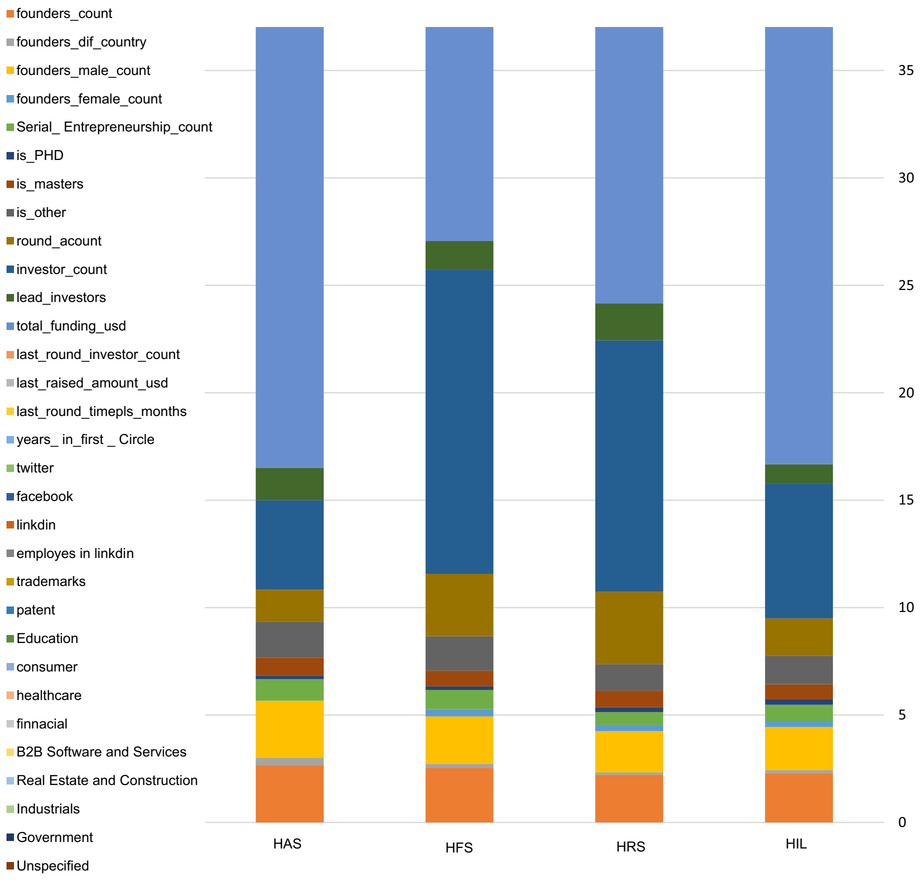

Fig. 7 Examining the status of clusters in the K-means

Its average amount of funding is 2,660,583. In this cluster, the number of followers in its social networks is low. In addition to the mentioned cases, this cluster is active in all areas of the industry.

In general, the first cluster includes startups with fewer founders. The funding round, the number of lead investors, and its amount of funding are average compared to other clusters 3 and 4. The distance between the last funding round to the beginning of the fifth year is high.

High round-count startup (HRS): In cluster two, the average number of cluster founders is 2.21. Also, the average number of male founders in this cluster is 1.92 and the number of female founders is 0.28. Regarding the financial part of this cluster, it has an average number of 3.36 founding rounds with 11.7 investors and 1.8 lead investors. Its average total investment budget is 5,976,995. In this cluster, the average of the variables related to the financial part is higher than the rest of the clusters. The average number of followers in the networks is almost the same as the cluster.

In general, cluster two includes startups that have fewer founders, but have more funding rounds and investors than other clusters.

Highly funded startups (HFS): The average number of founders of cluster 3 is equal to 2.5. Also, the average number of male founders in this cluster is 2.2. The majority of the founders of this cluster are male. Regarding the financial part of this cluster, it has an average number of 2.9 funding rounds with 14 investors and 1.3 lead investors. This cluster does not have an active presence on the LinkedIn social network.

In general, cluster 3 includes startups that have good funding rounds; however, it does not have a large amount of funding and a number of lead investors. Further, the appearance on LinkedIn social networks is not as good as other clusters.

Highly active on social media (HAS): In cluster 4, the average number of male founders is equal to 2.6, and this cluster does not include female founders. The average number of lead investors is equle1.5. On average, the total funding budget is 128,333 and is less than other clusters. And the number of Twitter and Facebook followers is higher than other clusters.

In general, this cluster includes startups with an active presence on Facebook and Twitter social networks, albeit they have fewer funding round and investors.

Policies for the success of startups

In this section, policies for the success of startups are presented, which make the path of identification and success of startups smoother for investors. At first, the macro-policies are presented and after that, according to the macro-policies, recommendations are provided for each cluster. In general, the policies presented in this section provide better investment signals to investors.

- Startups that have received positive feedback from other lead investors and have attracted more capital than competitors have made the startup's ability to obtain relevant funds greater during its founding date. This means that the more funding round and lead investors a startup has, the more successful it will be.
- Startups that started working in the right time frame and used the opportunities before them are among the successful startups. The shorter the age of the startup

in the 4-year period and the distance between the last investment period and the fifth year of the startup, the more successful the startup is. Why this group of startups has started working in the right time frame and the short distance between their last investment period and the specified time frame has made them create a better investment field for investors.

- The presence of startups in social networks allows the founders to obtain resources cheaper than what can be obtained in the market. For example, reputation, direct contact with customers, and so ones. Therefore, the more startups use social networks to develop and implement their strategies, the more advantages they can gain. In general, the presence of startups on social networks such as Facebook, Twitter, and LinkedIn have made them more successful.
- Startups that have an active and motivated team have a more sustainable advantage compared to their competitors and can put the company on the path to success by using their motivation and unique activities. The more employees a startup has on LinkedIn, the more successful it will be.

HIL: This cluster receives a good amount of funding; therefore, if the number of lead investors and funding rounds increases and performs better. If the distance between the last funding round to the beginning of the fifth year is less, startups can achieve better results. Also, increasing the activity of startups in social networks makes them more successful from an investment point of view, along with other features of startups.

HRS: Considering that most of the founders of this cluster are men and have more funding rounds and amount funding than other clusters; if they are more active on social networks and the number of their employees' presence on LinkedIn is more active, it will increase their success from an investment point of view. In addition to the mentioned items, if the number of investors increases, it will improve the financial situation of this startup.

HFS: Since this cluster receives good funding rounds, it will perform better if the number of lead investors increases. Moreover, enhancing the activity of startups in social networks makes them more successful in terms of investment, among other characteristics of startups.

HAS: The presence of this cluster in active networks; however, they have fewer funding rounds and lead investors. If they are formed in terms of characteristics related to investment and finance, it may help them to enhance their success process.

Prediction solution

In this section, the prediction solution is explained. The features importance of this prediction solution is based on Gradient Boosting algorithm and Random Forest because these algorithms perform better than other classification algorithms implemented in this research. This prediction solution is designed to assist investors in identifying and predicting the startup's success.

Although the type of company does not consist of main feature importance, the permutation feature importance analysis conducted using the Random Forest and Gradient Boost algorithms reveals that "B2B Software and Services", "consumer" and "health care" have higher values than the rest of categories of company. Thus, this prediction solution may have performed better in these three types of companies.

In Figs. 5 and 6, features importance of Random Forest and Gradient Boosting algorithms using the SHAP method are illustrated. As demonstrated in these figures, among the first ten features importance of Random Forest and Gradient Boosting algorithms, seven of them are the identical. These seven variables are as follows:

- LinkedIn: The number of followers on social networks is the key to the growth of companies. The presence of startups on social networks such as LinkedIn strengthens the brand of startups, and social networks can be used to achieve business goals. As a result, the higher the number of followers on LinkedIn, the higher the probability of success.
- Timelapse until fifth year: The lower the value of this variable, the greater the success rate. The closer the investment period is to March 2018; the more likely startups will be successful.
- Employees: Employees variable represents the number of employees at LinkedIn. The activeness of employees on LinkedIn is considered a success factor because the presence of employees on LinkedIn makes the business goals of the company more widely through employees. In fact, in addition to a large number of employees, it is possible to use diverse talents; their presence in the social network also helps to expand the business goals of the company.
- Twitter: Startups with more followers on Twitter are more likely to be successful since Twitter may provide an opportunity for startups to communicate directly with their audience. This platform makes it easy for startups to reach their audience, and due to the high number of users on Twitter, information can be transmitted to many people.
- Last round investors count: Fewer number of investors in the last round may lead to more positive effect on the success of startups. This body means that in the last round of investment, it is the larger amount of the investment budget that has a positive effect on success, not the larger number of investors.
- Last raised amount: The presence of last raised funding amount means that funding is an advantage and receiving more funding increases the probability of success from an investment perspective. As a result, one of the factors that influence the success of startups is the increase in the amount of funding in the last round.
- Investor count: Smaller numbers of investors are associated with the greater success of startups. As shown in Fig. 6, however, it is evident that larger number of main investors lead to the success of startups. This means that the number of investors does not lead to success, and these are the main investors, whose presence creates an advantage in the success of startups.

The seven selected variables have been obtained according to the sharing of important features of Random Forest classification algorithms and Gradient Boosting. Based on Random Forest classification algorithm and K-means clustering algorithm, which have demonstrated the best performance among other algorithms, the design of the prediction solution has been done. As illustrated in Fig. 8, engines of

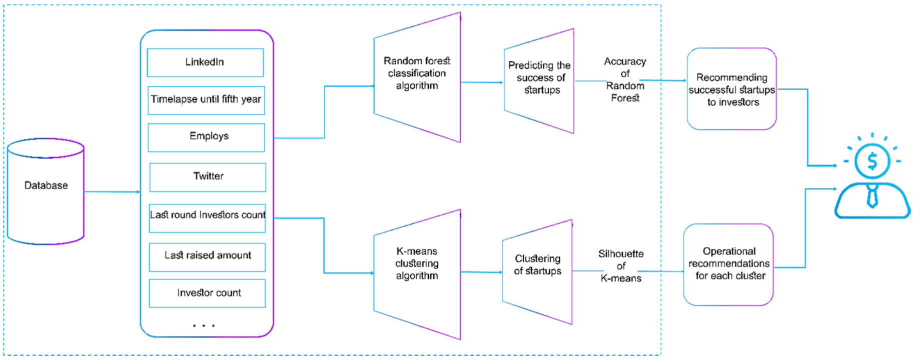
Database
LinkedIn
Timelapse until fifth year
Employs
Twitter
Last round Investors count
Last raised amount
Investor count
. . .
Random forest classification algorithm
Predicting the success of startups
Accuracy of Random Forest
Recommending successful startups to investors
K-means clustering algorithm
Clustering of startups
Silhouette of K-means
Operational recommendations for each cluster

Fig. 8 Prediction solution based on classifier and clustering algorithms

this prediction solution are the classification and clustering algorithms. Based on the results of this research, the presentation engine has been assessed in the prediction solution. Experts are expected to be able to use this prediction solution to build software for investors, and the output of this prediction solution is to predict successful startups to investors and operational recommendations for the cluster.

Based on Random Forest classification algorithm and K-means clustering algorithm, the architecture design of the prediction solution has been done. Random Forest predicts the success of startups well. Due to the high accuracy of this algorithm, successful startups are presented to investors. Using the Random Forest classification algorithm gives investors the opportunity to identify successful startups with high probability. K-means clustering algorithm clusters data with high silhouette percentage. Following clustering the data, operational advice should be provided according to the characteristics of each cluster, for the success of the startup.

This model can help investors to invest in startups by evaluating the success chances of startups. As mentioned earlier, predicting the success of startups is challenging and critical for investors. The final model of this research is the design of a prediction solution to predict the success of startups using clustering and classification algorithms. The general approach of this research can be applied to other similar data sets. In particular, experts can provide this prediction solution model in a practical way for investors to identify successful startups.

Additionally, our theoretical work can be complemented by empirical studies to validate the influence of the seven important variables identified in predicting startup success. By focusing on data points specifically related to early-stage companies, the trained prediction solution is anticipated to exhibit enhanced performance in accurately predicting the success of early-stage companies. Furthermore, it is expected that this prediction solution may demonstrate improved quality and effectiveness, particularly in the “B2B Software and Services”, “Consumer”, and “Healthcare” sectors. Thus, future empirical studies should focus on the early-stage companies and these three sectors. We believe that combining the power of artificial intelligence and empirical study approaches will significantly enhance investors’ ability to identify and support successful startups.

Comparison with previous work

Previous works related to predicting business success used different definitions of success. To determine the success and failure of startups, this research has followed Arroyo et al., (2019). Our work uses suitable predictor variables compared to previous studies, but due to the limitation of the number of data in this research, we did not expect it to be nearly the same as the performance of its models. However, despite the data limitations, we were able to compare our work to the rest of the research. In addition, we have used clustering algorithms and the prediction solution in our research, otherwise, these items were not used in another research. The results of the studies performed by Arroyo et al. (2019), Corea et al.(2021), and Kim et al.(2023) are comparable to ours.

According to Arroyo et al. (2019) and Corea et al. (2021), LinkedIn was found to be a significant factor to successful of startups, as discussed in previous studies, but with this difference in this study, we considered the number of followers on LinkedIn, while they considered whether one startup has a LinkedIn URL or not. The presence of startups on social media, especially LinkedIn reinforces the networks of startups, and they can achieve higher reputation among their followers. As a result, using LinkedIn can be utilized to accomplish trade goals.

The second variable discussed in previous studies, and it is common with our results is the investor account. Previous research has shown that the number of previous investors is a matter (Corea et al., 2021; Kim et al., 2023), while they did not specify how investors influence the success of startups. Our evidence with using the SHAP method shows us that lead investors cause an advantage in the success of startups, and the enormous number of investors does not lead to success.

Arroyo et al. (2019) showed that "Timelapse until fifth year" is one of the features importance in both Gradient Tree Boosting and Random Forests algorithms. In this research, we specifically using the SHAP method determined that a lower value of this variable leads to greater success.

Finally, Kim et al. (2023) exhibited considering employee variables is necessary to predict startup success. Our study indicates that higher employees cause the higher success of startups. The results of our research are in line with other research, with the difference that we have used SHAP to examine more precisely what the results of each variable are on the success of startups.

In this research, the various variables are considered by using multiple databases. However, it is important to acknowledge that there may be additional variables that were not discussed or included in the analysis. Another limitation of this research is determining the dependent or target variable. The target variable has been determined based on previous research, and 7 years is considered to predict the success of startups. In a practical application, every investor should know the optimal valuation period for their needs and make decisions accordingly. Finding more variables as inputs to the algorithms and expanding the categories of the output predictable variable to find more insights into the startup's success can strengthen the depth and breadth of findings in future research.

Conclusions

We proposed a machine learning approach for designing a prediction solution for investors to predict the success of startups. The design of this prediction solution is based on classifier and clustering algorithms. The data of this research are from the Crunchbase database. The selection of data includes startups that were established in the period from 2014 to 2018. Also, in the period from 2018 to 2021, the state of startups has been investigated. After data selection, data preprocessing is done. Classification algorithms have been used to predict the success of startups and determine features importance with using SHAP and permutation feature importance. To predict business success, we compared the performance of five algorithms: Random Forest, Gradient Boost, MLP, Logistic Regression, and SVM. Among them, Random Forest showed the best performance with accuracy equal to 82% and 80%, respectively.

For clustering, three cluster algorithms have been compared, which included hierarchy, K-means and DBSCAN. Among them, K-means showed the best performance, and K-means were used to explain the characteristics of each cluster. Policies are based on classification and clustering algorithms, and macro-policies are based on classification algorithms. According to the macro-policies provided for each cluster, operational recommendations are provided. The design of this prediction solution is based on Random Forest classifier and K-means clustering algorithms. According to the Random Forest classification algorithm, the output of this prediction solution is expected to predict startups that are likely to be successful to be introduced to investors. In addition, based on the K-means clustering algorithm, operational recommendations should be provided for each cluster. The main advantage of the research is the reproducibility of its outcomes for real-world scenarios. In the following section consist of three dimensions: first, the theoretical implications, second the managerial implications, and last, future research directions.

Theoretical implications

Our study has several theoretical implications. First, our research indicates that the higher values of "LinkedIn", "Employs", "Twitter" and "Last raised amount" possess higher SHAP values with higher impact on the model output. In contrast, the lower values of "time-lapse until fifth year", "Last round Investors count" and "Investor count" possess higher SHAP values. Second, our research provides macro-policies according to the cluster algorithms. Startups that have received positive feedback, started working in the right time frame and used the opportunities, have the presence of startups in social networks, and Startups that have an active and motivated team, have more potential for success in the market.

Managerial implications

This research provides some insight into designing of prediction solution in order to predict the success of startups. This prediction solution is based on clustering and classifier algorithms. By determining the key success factors of startups that are obtained through classification algorithms and the policies provided through clustering algorithms, investors can use the factors obtained in this research to make decisions on investing in startup. Therefore, in order to invest successfully in startups, it is better for investors to pay attention to these factors when investing so that they can reduce the risk of their investment.

Future research directions

Future work will need to be more focused on the seven important features that extract from this research. It should be investigated why these seven characteristics have been obtained and these seven characteristics should be evaluated on other startups so that other models can be obtained. Also, future work requires using the four clusters obtained from the results of the clustering algorithm, run the classification algorithm on those four clusters. Considering variables like economic conditions, and market trends could improve the result of future research. Future research could consider a determined period for startups and record all of the changes of features during the time manually, especially when they do not have access to the past feature. For example, features are time-varying, such as the number of followers on social media.

Abbreviations

CRISP-DM	Cross-Industry Standard Process for Data Mining
GB	Gradient Boosting
HAS	Highly active on social media
HFS	Highly funded startups
HIL	Highly ranked investors in the last round
HRS	High round-count startup
IPO	Initial public offering
ML	Machine Learning
MLP	Multilayer Perceptron
RF	Random Forest
SVM	Support Vector Machine
VC	Venture Capital

Acknowledgements

Not applicable.

Author contributions

MB was responsible for data preparation, cleaning and curation data, impalement of methodology and writing—original first draft manuscript. IV (corresponding author): methodology, supervision, analysis of survey data, research development, and research validation. MG was a major contributor in writing and validation of the manuscript. All authors read and approved the final manuscript.

Funding

This research is not supported by any financial source.

Data availability

The data sets generated and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could influence the work discussed in this article.

Received: 29 June 2023 Accepted: 30 September 2024
Published online: 28 October 2024

References

Ahluwalia, S., & Kassicieh, S. (2021). Effect of financial clusters on startup mergers and acquisitions. *International Journal of Financial Studies, 10*(1), 1.

Aleisa, M. A., Beloff, N., & White, M. (2023). Implementing AIRM: A new AI recruiting model for the Saudi Arabia labour market. *Journal of Innovation and Entrepreneurship, 12*(1), 59.

Alpaydin, E. (2020). *Introduction to machine learning*. MIT press.

Arroyo, J., Corea, F., Jimenez-Diaz, G., & Recio-Garcia, J. A. (2019). Assessment of machine learning performance for decision support in venture capital investments. *IEEE Access, 7*, 124233–124243.

Azevedo, A., & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. *IADS-DM*.

Bai, S., & Zhao, Y. (2021). Startup investment decision support: Application of venture capital scorecards using machine learning approaches. *Systems, 9*(3), 55.

Bednár, R., & Tarišková, N. (2017). Indicators of startup failure. *Industry, 2*(5), 238–240.

Blank, S. (2018). Why the lean start-up changes everything.

Blank, S., & Dorf, B. (2010). Startup. Handbook of the founder.

Cavallo, A., Ghezzi, A., & Balocco, R. (2019). Entrepreneurial ecosystem research: Present debates and future directions. *International Entrepreneurship and Management Journal, 15*(4), 1291–1321.

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics, 21*(1), 1–13.

Clifton, C., & Thuraisingham, B. (2001). Emerging standards for data mining. *Computer Standards & Interfaces, 23*(3), 187–193.

Corea, F., Bertinetti, G., & Cervellati, E. M. (2021). Hacking the venture industry: An Early-stage Startups Investment framework for data-driven investors. *Machine Learning with Applications, 5*, 100062.

Drover, W., Busenitz, L., Matusik, S., Townsend, D., Anglin, A., & Dushnitsky, G. (2017). A review and road map of entrepreneurial equity financing research: Venture capital, corporate venture capital, angel investment, crowdfunding, and accelerators. *Journal of Management, 43*(6), 1820–1853.

Fahmy Amin, M. (2022). Confusion matrix in binary classification problems: A step-by-step tutorial. *Journal of Engineering Research, 6*(5), Article 1.

François, D. (2008, April). Methodology and standards for data analysis with machine learning tools. In: *ESANN* (pp. 239–246).

Ghassemi, M. M., Song, C., & Alhanai, T. (2020). The automated venture capitalist: Data and methods to predict the fate of startup ventures. Association for the Advancement of Artificial Intelligence.

Hoenen, S., Kolympiris, C., Schoenmakers, W., & Kalaitzandonakes, N. (2012). Do patents increase venture capital investments between rounds of financing. Pobrane z: http://edepot.wur.nl/216191.

Holmes, P., Hunt, A., & Stone, I. (2010). An analysis of new firm survival using a hazard function. *Applied Economics, 42*(2), 185–195.

Hunt, R. A. (2013). Entrepreneurial tweaking: An empirical study of technology diffusion through secondary inventions and design modifications by start-ups. *European Journal of Innovation Management, 16*, 148–170.

Huyghebaert, N., Van de Gucht, L., & Van Hulle, C. (2007). The choice between bank debt and trace credit in business start-ups. *Small Business Economics, 29*(4), 435–452.

Insights, C. B. (2018). Venture capital funnel shows odds of becoming a unicorn are about 1%. *CB Research Briefs*.

Johnson, K., Pasquale, F., & Chapman, J. (2019). Artificial intelligence, machine learning, and bias in finance: Toward responsible innovation. *Fordham l. Rev., 88*, 499.

Kim, D., & Lee, S. Y. (2022). When venture capitalists are attracted by the experienced. *Journal of Innovation and Entrepreneurship, 11*(1), 31.

Kim, J., Kim, H., & Geum, Y. (2023). How to succeed in the market? Predicting startup success using a machine learning approach. *Technological Forecasting and Social Change, 193*, 122614.

Korosteleva, J., & Mickiewicz, T. (2011). Start-up financing in the age of globalization. *Emerging Markets Finance and Trade, 47*(3), 23–49.

Krishna, A., Agrawal, A., & Choudhary, A. (2016, December). Predicting the outcome of startups: less failure, more success. In: *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)* (pp. 798–805). IEEE.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR)., 9*, 381–386.

Molnar, C., Freiesleben, T., König, G., Herbinger, J., Reisinger, T., Casalicchio, G., ... & Bischl, B. (2023, July). Relating the partial dependence plot and permutation feature importance to the data generating process. In: *World Conference on Explainable Artificial Intelligence* (pp. 456–479). Cham: Springer Nature Switzerland.

Muschelli, J., III. (2020). ROC and AUC with a binary predictor: A potentially misleading metric. *Journal of Classification, 37*(3), 696–708.

Piatetsky, G. (2014). CRISP-DM, still the top methodology for analytics, data mining, or data science projects. KDD News.

Powers, D. M. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint* arXiv:2010.16061.

Ross, G., Das, S., Sciro, D., & Raza, H. (2021). CapitalVX: A machine learning model for startup selection and exit prediction. *The Journal of Finance and Data Science, 7*, 94–114.

Roth, A. E. (1988). Introduction to the Shapley value. The Shapley value, 1–27.

Santos, M. F., & Azevedo, C. S. (2005). Preâmbulo [a]" Data mining: descoberta de conhecimento em bases de dados". FCA-Editora de informática, Lda.

Shahapure, K. R., & Nicholas, C. (2020, October). Cluster quality analysis using silhouette score. In: *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)* (pp. 747–748). IEEE.

Skawińska, E., & Zalewski, R. I. (2020). Success factors of startups in the EU—A comparative study. *Sustainability, 12*(19), 8200.

Thirupathi, A. N., Alhanai, T., & Ghassemi, M. M. (2021, November). A machine learning approach to detect early signs of startup success. In: *Proceedings of the second ACM international conference on AI in finance* (pp. 1–8).

Tomy, S., & Pardede, E. (2018). From uncertainties to successful start ups: A data analytic approach to predict success in technological entrepreneurship. *Sustainability, 10*(3), 602.

Turkoglu, B., & Kaya, E. (2020). Training multi-layer perceptron with artificial algae algorithm. *Engineering Science and Technology, an International Journal, 23*(6), 1342–1350.

Ughetto, E. (2016). Growth of born globals: The role of the entrepreneur's personal factors and venture capital. *International Entrepreneurship and Management Journal, 12*, 839–857.

Vanani, I. R., & Jalali, S. M. J. (2018). A comparative analysis of emerging scientific themes in business analytics. *International Journal of Business Information Systems, 29*(2), 183–206.

Woods, C., Yu, H., & Huang, H. (2020). Predicting the success of entrepreneurial campaigns in crowdfunding: A spatio-temporal approach. *Journal of Innovation and Entrepreneurship, 9*, 1–23.

Żbikowski, K., & Antosiuk, P. (2021). A machine learning, bias-free approach for predicting business success using Crunchbase data. *Information Processing & Management, 58*(4), 102555.

Zhang, C., Zhang, H., & Hu, X. (2019). A contrastive study of machine learning on funding evaluation prediction. *Ieee Access, 7*, 106307–106315.

Publisher's Note