ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 11, Issue 10, October 2024

Predicting Startup Success using Machine Learning: A Comprehensive Review of Current Research

[1] Roa Aleid*, [2] Majed Al-Mashari

[1] Researcher, Information System, King Saud University, Riyadh, Saudi Arabia
[2] Professor, Information System, King Saud University, Riyadh, Saudi Arabia
Corresponding Author Email: [1] Roa.aleid@gmail.com, [2] malmashari@yahoo.com

Abstract*— In recent years, the application of machine learning (ML) models has gained significant traction in different fields, such as the prediction of startup success. Startups face many challenges, and the ability to predict their success or failure is of paramount importance for investors, entrepreneurs, and stakeholders. This paper provides a comprehensive literature review of existing research on the use of ML models in predicting the success of startups. By examining a wide array of studies, this review highlights the key factors influencing startup success, such as financial performance, team composition, market conditions, and business models. Various ML algorithms—such as logistic regression, decision trees, support vector machines (SVM), and deep learning techniques—have been employed across these studies. The review also explores the datasets, features, and evaluation metrics commonly used in predicting outcomes. This paper aims to synthesize the state of the art in this field and identify current trends, challenges, and future research opportunities.

Index Terms*— Startup, Predictive Analytics, Machine Learning, Decision Tree Model, Gaussian Naive Bayes Model, Logistic Regression Model, K-NN Model.

I. INTRODUCTION

Startups, by nature, are characterized by high risk and uncertainty, with most failing in the first few years [1]. Predicting successful startups has long been a challenge for investors, entrepreneurs and market researchers. Traditional methods rely heavily on economic considerations, market research, and experienced stakeholder intuition. However, these methods often fail to account for the complexity and uncertainty inherent in startup firms [2]. In recent years, the availability of large datasets and advances in computational techniques have provided new approaches to this problem. Machine learning (ML), with its ability to process huge amounts of data and uncover hidden patterns, has emerged as a promising tool for predicting startup outcomes [3]. By analyzing various factors, such as financial performance, team structure, market dynamics, and even social media presence, ML models can identify correlations that may not be obvious through traditional methods. Several studies have explored the potential of ML to enhance the understanding of what drives startup success. Different algorithms, such as decision trees, support vector machines, neural networks, and ensemble methods, have been applied to this problem. These models utilize a diverse database from various sources, including web databases. Each model attempts to balance prediction accuracy with generalizability, as startups operate in rapidly changing environments where small variables can have significant impacts. As a result, the reliability and scalability of ML in this domain are still open questions, with varying results depending on the methodology and dataset used.

The purpose of this paper is to provide a comprehensive review of the current research on the application of machine learning to predict startup success. We examine the different approaches that researchers have taken, the models and algorithms used, and the types of data that have proven effective. Ultimately, the goal is to better understand how machine learning can contribute to more accurate and actionable predictions for startups.

II. LITERATURE REVIEW

In this section, the main concepts will be explained includes clarifying the definitions of startups and their impact on the real world, along with the general reasons for their failure, specifically in the Saudi market. Additionally, the current mechanisms for assessing the success of startups will be discussed. Finally, a comprehensive survey of relevant research studies.

A. Definition of startups:

In general terms, a startup is commonly described as a new business initiated by entrepreneurs who combine ideas and available resources [4]. According to Steve Blank [5], a startup is essentially a temporary organization with the specific purpose of exploring and establishing a repeatable and scalable business model. Eric Ries further clarifies that startups are organizations launched with the intent of creating new products or services in conditions of extreme uncertainty [6]. Ries contends that this definition encompasses various entities, including new business units of governments, large corporations, non-profit organizations, and business ventures, as long as they are engaged in creating innovative products or services under conditions of uncertainty [6] . According to Cohen et al., this process describes the events

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 11, Issue 10, October 2024

before the business becomes an organization and after it becomes a fully functional organization [7]. Some researchers label the period before the creation of the business and after its full inception as a startup [8]. However, it is still difficult to directly pinpoint the exact definition of a start-up. In fact, the focus in this project will be on any existing businesses referred as startups.

B. Impact of startups in real world:

In fact, startups have a significant impact on the economic and social aspects. Start-ups have contributed greatly to the field of creating job opportunities, as they have become a major player, which has contributed to reducing unemployment [9]. One study indicated that start-ups that are less than one year old have created an average of 1.5 million job opportunities annually over the past three decades in the world [10]. Not only that, but these start-ups also contribute to economic dynamism by injecting competition into markets and stimulating innovation. Furthermore, startups often foster a culture of creativity and entrepreneurship, encouraging individuals to pursue their passions and contribute to the job market. Beyond employment, start-ups play an essential role in economic growth. Introducing new products, services and business models stimulates competition and stimulates overall productivity. The inherent dynamism of startup environments can lead to groundbreaking developments, fostering a culture of continuous improvement and technological innovation. This in turn has a ripple effect on the broader economy, impacting existing industries and inspiring them to adapt and innovate to remain competitive.

In addition, start-ups play an active role in addressing societal challenges, as many start-ups are established with the aim of addressing specific issues, whether in the field of health care, education, environmental sustainability, or other social fields. In fact, the innovative solutions developed by start-up owners have far-reaching impacts, as they contribute to raising the quality of life for individuals and the general well-being of communities.

C. Why startups fail:

The extreme level of failure among startups is an element often of keen interest to startup investors. Assessments reveal that most startups fail to get beyond their 3rd year of success after inception. This failure is due to the extreme levels of competition in the real markets, as the huge companies are often more likely to take over and dominate, having the financial muscle and resources to outcompete the smaller startups. Numerous assessments have been conducted on the nature and characteristics associated with the failure and success of different startups. Thus, several theories have been introduced to describe why most startups fail. The first theory is the population ecology theory, introduced by Hannan and Freeman, which examines the dynamic changes in an organization [11]. The theory examines how organizations are born, their development, growth, and even mortality by considering the economic, political, and structural elements of a business. The theory proposed that organizations' probability of death decreases with the increase in their age since their initial learning expands and increases exponentially with age. The population ecology as such supposes that it is not entirely tied to the entrepreneur of a specific business that determines the success but rather the total composition of all the elements within the environment likely to affect the business [12]. The population ecology theory suggests that the dependency of startups on organizational populations for their success leads them to tremendous risks, especially in the external market environment. Examining the process through which firms emerge and cease to exist in an ecosystem reveals that there are both time-dependent factors and other external stressors that tend to be extreme for most organizations [13]. This theoretical perspective is further supported by research that demonstrates that organizations generally become more resilient with time, and it is not a matter of size but rather experience within the marketplace [14]. An assessment by Vest and Menachemi revealed that smaller organizations that have existed for extended periods are more likely to thrive amidst tough economic times compared to more prominent startups that are still new in the marketplace, which is the presupposition instituted by the population ecology theoretical perspective [15]. In contrast, Schumpetersian proposed that the inception of startups into a market begins with the inception of a new service or a new product [16]. According to Schumpeterian, the startup may also be accepted after or due to the discovery of an opening a new market [16]. Schumpeterian model or theory seems to be the most complete in helping describe what startups are and their nature in the marketplace [16]. This perspective or theoretical framework also offers ample ground from which the nature of the success and failure of startups can be concretely examined and assessed [16]. Schumpeterian believed that innovation and creativity were thus the center sage or the core of startups [16]. However, the very nature of innovation and its uncertainty instituted the enormous threat or increased probability through which startups fail. On the other hand, when it comes to the Saudi market, there is a lack of sufficient literature and theoretical framework that can be used to classify or assess factors behind the failure of startups in Saudi Arabia. Even so, one of the main challenges facing startups in Saudi Arabia includes a lack of proper market orientation [16]. Market orientation encompasses the proper research of a specific market to understand its needs and thus incept businesses based on these needs and perspectives. According to Alsolaim, the other challenge is the difficulty of start-ups to obtain sufficient funding in Saudi Arabia after their inception [18]. Overall, the high failure rate among startups attributing to several challenges such as the intense competition in real markets. In the context of the Saudi market, challenges include a lack of market orientation and difficulties in securing funding post-inception.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 11, Issue 10, October 2024

D. Current mechanisms of assessing the success of startups:

The failure of most startups less than three years after their inception has questioned venture capitalists and investors. As such, numerous frameworks have been developed to help potential investors examine the chances of a firm success by observing how other successful firms have failed in the past and the similarities in their characteristics as a mechanism of determining their overall success. Numerous theories have been incepted to help investors determine the risks of their money when investing in startups worldwide, including financial projections, business model canvas, Timmon's entrepreneurial process, and machine learning [19].

Financial projections are one of the easiest models and frameworks used to determine the success of businesses, especially startups and investors. Financial projections, according to Avagyan et al., encompass a wide array of financial positions in businesses, including their balance sheet, cash flow, and loan prediction, which are simply financial analysis projections [20]. It is one of the most accurate ways of determining the need for investments or businesses as it helps unveil the book value of the company, and its balance sheets describe the stability and financial security of the business [21]. While these methods are perfect and accurate, they often do not favor startup companies [20]. This is because startups generally lack sufficient financial background from which the balance sheets and cash flows can be projected. With less than three years in the market, it becomes difficult to judge the potential success of businesses and companies simply by looking at their financial projections. There is also not sufficient latitude from which the financial successes of the businesses can be examined over a specific period of time to determine potent fluctuations. As such, in the very inception of startups, it becomes difficult for them to be examined using their financial projections.

One of the most perfect methods developed to assess potential success is the “Business Model Canvas”. The Business Model Canvas (BMC) predicts projects the potential success of a business by looking at the model being utilized in a business and examining whether the method is viable and whether it is stackable to allow for the growth of a company [20]. The business model canvas provides a framework that helps institute value into the products of a business by examining finances, customers, and infrastructure developed by a business. A business model canvas offers a wholesome picture of a business or company, describing the customer base and how one intends to instill discipline into this category of persons. While it is an excellent strategic tool used for assessing the success of a business, it tends to be too simplistic [21]. It is done on a single sheet of paper yet highlights almost all aspects of the business. It is also based on presumptions regarding the potential value of the product and purported market segmentation [22] The business's strategy is often not included in the BMC [23]. According to Fakieh et al., BMCs do not consider the strategy being employed by a business and the ambitions and goals of a business within a specific marketplace [24]. It also fails to consider the rate at which either profits or losses are being made, which brings about inaccuracies. Furthermore, BMCs do not highlight interconnections and do not consider the position of a company in an ecosystem element, which further disadvantages a business.

Other methods that are used include Timmon's process of entrepreneurial process. This model considers the availability of opportunity, resources, and the team that operates a specific business [19]. The teams that constitute the inception of a startup become the primary elements from which the overall success of the investments and the businesses are examined and investigated. It is one of the most successful methods employed and used by Venture capitalists to determine the potential success of the business [24]. It entails assessing the track records of the teams behind a project and, judging by their previous success, coming up with excellent prediction regarding their potential for future success given a startup they currently hold [25]. The model also encompasses examining the market the startup is targeting and assessing whether there is sufficient skill within the startup to grow, develop, and scale. He examines whether the market intended by the startup is dwindling, stagnant, or growing and whether that market has unmet needs that the company of interest examines. This is one of the most successful methods and is often used to assess startups in the current startup ecosystems worldwide. However, it is relatively complicated and immediately dissociates startups that have been incepted by completely new individuals [26]. It becomes disadvantageous in that these startups fail to consider the dynamics of innovations and just how revolutionary certain ideas in a market can become [25]. Nonetheless, there is a relatively new model for assessing the potential success of startups and new businesses worldwide using Artificial intelligence. It encompasses the use of machine learning. Machine learning encompasses software programs that can continually improve and update their databases and analysis methods and use this increasing knowledge to provide better prediction for the future [27]. This software continually evolves and becomes accurate over time [27]. There has been an exponential surge in the use of AI for several business ventures which has exponentially grown worldwide. These mechanisms of machine learning can also be used to assess the potential for success or failure of startup businesses by examining numerous amounts of data for different startups that have both succeeded and failed and coming up with specific elements that constitute the startups that either fail or succeed in the long run [27]. This can then provide a baseline regarding the specific characteristics of a business concerning their market that investors and venture capitalists will use to predict the potential for success or failure of a

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 11, Issue 10, October 2024

specific startup and use it in their investment assessments.

E. Survey of related studies:

Numerous studies have focused on predicting startup success, with recent research increasingly leveraging machine learning techniques. In this review, we focus specifically on these recent advancements in applying machine learning to this problem. In [28], the researchers conducted a systematic review of existing studies concentrating on predicting startup success. Their aim was to consolidate insights and identify key factors influencing startup success. The primary contribution of this work was providing a comprehensive overview of the significant factors found and used in the previous literature concerning the startup success prediction. The paper concluded that the important determinants for startup success include funding rounds, market specifics, and geographical indicators. Overall, this literature review provides invaluable information that could be an initial point for researchers and practitioners aiming to understand the nature of predicting startup success tasks. On the other hand, the main contribution of the conducted study in [29] was to examine the key features of startup success using data from the Kaggle website with nearly 22,000 startups. Leveraging Random Forest and Support Vector Machine techniques, the study explored the most significant features that determine the success of startup companies. Thus, the study concluded that the most ten effective features in the success of any start-up include venture, market, city, region, state code, seed, funding rounds, round A, round B, and round C. Accordingly, the researchers in [30] developed a supervised learning predictive model in order to address the bias issues in the process of predicting the start-up's success where these issues usually occur due to poor data objectivity. The study was conducted using data from one of the largest business information sources (Crunchbase) with a final training set containing 213,171 companies. Moreover, they compared three algorithms—logistic regression, SVM, and XGBoos where the XGBoost model yielded the highest results with precision, recall, and F1 scores of 57%, 34%, and 43%, respectively highlighting its potential for predicting business success. Notably, the model primarily focused on using geographic, demographic, and fundamental company data to serve as a decision support mechanism tailored for venture capital funds. In the other hand, the authors in [31] developed a predictive model using supervised learning to classify successful and unsuccessful startups. for this task, they used K-Nearest Neighbours (KNN) model and then compared it with Logistic Regression (LR) and Random Forests (RF) models from a previous work. They found that K-Nearest Neighbours (KNN) has a better performance in terms of F1 score matric achieving 44.45%. Meanwhile, Random Forests (RF) outperformed in accuracy, reaching 84.03%. The authors conclude that this result made RF model more suitable for cases that have limited investment budget and aim to maximize the success across the portfolio. In contrast, KNN is more suitable for cases that have unlimited investment budget and aim to maximize successful investments. Notably, the used data was from (Crunchbase) with 60,000 total of Companies. Furthermore, the paper in [32] proposed a model aimed at predicting the success of startup companies by considering financial and managerial variables. Indeed, this study covered several factors such as investments, valuation, market value, total funds, acquisitions, and the financial background of key individuals. The study was conducted using data from over 15,000 companies collected from Crunchbase. Five models including random forest, text parsing, logistic regression, decision tree, and survival analysis, were employed to determine the most effective model for such goal. Survival analysis model indicates a significant relationship between degree and startup success. The Logistic Model, selected based on the Receiver Operating Characteristics (ROC) index, achieved a commendable 0.81, indicating strong performance. However, the researchers in [33] focused on developing a predictive model for startup success based on descriptive characteristics including the time from foundation to the first financing, business model, and applied technologies to predict startup investment success. Notably, the data used was from the Dealroom platform with a sample comprised of 123 startups operating in Ukraine. Additionally, three different models that are Logistic Regression, Decision Tree, and Random Forest models were compared. Among these models, the Decision Tree was identified as the most effective in predicting startup success, achieving average accuracies of 61.2%, 54.7%, and 52.3% for Accuracy, Sensitivity, and F-score, respectively. The study in [34] presented six models for startup success prediction based on a number of key variables such as last funding to date, first funding lag, and company age. Indeed, logistic regression, Decision Trees, random forest, and extreme gradient boosting models were used along with data from Crunchbase. The study showed that random forest, and extreme gradient boosting, emerge as the best with a test set prediction accuracy of 94.18%and 94.45%, and AUC of 92.22% and 92.91%, respectively. While the work in [35] proposed six predictive models for predicting the success/failure of early-stage startups. The models were proposed based on several key factors involved at various stages in the life of a startup such as funding stages, timeframes, and success/failure contributors at each company milestone. These factors were extracted by analyzing the data of over 11,000 companies that sourced from different sources such as Crunchbase and Tech Crunch. Specifically, the models used were Lazy lb1, Random Forest, NaiveBayes, ADTree, Bayesian Network, and Simple Logistic. These models achieved notable performance in terms of recall, and precision where the precision accuracies ranged from 73.3% to 96.3% while recall accuracies ranged from 78.3% to 96.6% across the models. Also, in [36] the researchers used

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 11, Issue 10, October 2024

several machine learning algorithms such as Decision Trees, Random Forest, Gradient Boost, Logistic Regression, and MLP Neural networks to create models for success/failure prediction of early-stage startups. Their definition of success encompassed instances where startups either initiated an Initial Public Offering (IPO) or underwent mergers/acquisitions. The used data in this study was historical data of startups from CrunchBase platform that included key factors such as valuations, funding rounds, and investments. Impressively, their models achieved a precision accuracy of around 92%, showcasing robust predictive capabilities across all the utilized methodologies. Furthermore, the study in [37] aims to construct a predictive model via supervised learning that predicts successful/ unsuccessful startups. The used data was sourced from CrunchBase where used the model was the Random Forests (RF) model. In this work, the experiments were applied in three stages i. a general model that includes all categorical features, the result achieved a True Positive Rate (TPR) of 94% at this stage. ii. specific models per each company category such as the results achieved ranged between 61% and 96% at this stage. iii. specific models for each of geographical regions such as CA (California), NY (New York), TX (Texas), MA (Massachusetts), the results achieved TPRs ranging between 90% and 96%. Additionally, the thesis in [38] proposed a reliable model to predict the success of Startups specifically located in the United States U.S. This research focused on the comparison between the performances of different predictive models, such as Logistic Regression, Linear Discriminant Analysis, Extreme Gradient Boosting, and Support Vector Machine with linear Kernel. The author concludes that with Area Under the Curves (AUCs) of roughly 0.93, the performance of Logistic Regression, Support Vector Machine, and Extreme Gradient Boosting in the 30% testing dataset is quite similar. However, Logistic Regression offered more reliable coefficients and more conservative outcomes. In [39] An idea was formulated by the researchers to predict the long-term success of startups. Machine learning classification algorithms were employed to ascertain effective strategies for startup success. The dataset sourced from Kaggle was used, with the pivotal determinant being the "status," encompassing two values: "acquired" signifying successful acquisition by a company and "closed" indicating startup failure. Eight distinct algorithms were applied to this dataset to determine the most suitable algorithm. These algorithms included Decision Trees, Random Forest, K-Nearest Neighbors, MLP, Naive Bayes, SVM, Logistic Regression, and SGD. Diverse efficiency scores were obtained from these algorithms, revealing Random Forest as the most suitable fit for the dataset. Also, the researchers in [40] designed two models for predicting startup success one predicting investor-profitable exits and another assessing funding potential exceeding 1 million Euros. They used a dataset of 406 Dutch startups from Techleap.nl. However, the first model underperforms with no true positives, while the second model achieved a 73.1% accuracy rate. In contrast of other studies that focused on the region rather than the field, the study in [41] introduced a systematic ML-based approach to predict success of startup in information technology SIT with high precision. The used data was 265 Australian SITs. The study encompassed a hybrid model and inventoried 79 critical success factors then mainly used 20 factors of them. Seven ML algorithms were used: support vector machine (SVM), multilayer perception (MLP), Decision Tree (DT), Naïve Bayes (NB), k-nearest neighbors (KNN), Random Forest (RF), Gradient Boosting (GB). MLP, GB, and SVM yielded the best results. Moreover, Employing the GreedyStepwise algorithm reduced the twenty factors to five key factors were startup size, company revenue, R&D, financial capital, and global economic environment, achieving 88% accuracy, 82% precision, and 94% specificity, proving effective with limited data.

In below table encapsulates a collection of essential information, offering insights into the landscape of the previous research. The table is carefully constructed to offer a comprehensive perspective. For each research study, a Reference is provided, linking to the respective research study. The Model column describes the machine learning techniques employed in the research, used to analyze the dataset and make predictions. The Training Factor column covers the factors used in the learning process and impacts the model's predictive capabilities. The Target Variable of Success column defines the variable that the model aims to predict as successful. The Performance column signifies the performance metrics used to assess the effectiveness and accuracy of the machine learning models. These metrics may include accuracy, recall, F1-score, and precision. Additionally, the Dataset Source column identifies the provider of the dataset used in the research, while the Dataset Region column specifies the geographical scope of the dataset. Lastly, the Dataset Size column denotes the number of instances or records in the dataset, representing the volume of data available for training and testing the machine learning model.

Table I: Summary of studies

Ref	Model	Considered Factors	Success/ Target Variables	Dataset			Performance
				Source	Region	Size	
[29]	Random Forest	Market, City, Venture, Region, State code, Funding round, Round A, Round B of financing, round C of financing, and seed.	N/A	Kaggle	Worldwide	22,000 Startups	Accuracy=89%
	Support Vector Machine (SVM)						Accuracy=88%

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 11, Issue 10, October 2024

Ref	Model	Considered Factors	Success/ Target Variables	Dataset			Performance
				Source	Region	Size	
[30]	Logistic regression,	Company subcategories, Company categories, Founder's gender, Founder education degree, Region size, City size, Years between founder's graduation and companies, Years of studying.	IPO, Operating and received B funding, Acquisition	Crunchbase	Worldwide	213,171 Startups	Accuracy=86%
	Support Vector Machine (SVM)						Accuracy=87%
	XGBoos						Accuracy=86%
[31]	K-Nearest Neighbors (KNN)	Category, Country, Funding Rounds, funding Total (USD), First funding, Last Funding at, and the difference between when First funding at and Last Funding at.	Merger, Acquisition	Crunchbase	Worldwide	60,000 Startups	Accuracy=73.7%
[32]	Logistic Regression	Burn Rate, Total Valuation, Total number of Milestones, Average days between each Milestone, Total Funding rounds, Average Days between each Funding Rounds, Time to Get Seed Funding, Domain, and Location.	Closure, Acquisition	Crunchbase	Worldwide	15,000 Startups	Accuracy=85.9%
	Neural Network						N/A
[33]	Logistic Regression	HQ Country, Time to first funding, Client Type (B2B_B2C), Industries, Yearly Growth digital activity in the SW rating, The number of visits to the site on average per year, First Round Type, First Round Amount, APP Downloads, Top Rank SW, Income Streams, Revenue Model, and Technologies.	Repeated Investment	Dealroom	Ukraine	123 Startups	Accuracy=60%
	Decision Tree						Accuracy=61.2%
	Random Forest						Accuracy=57%
[34]	Full Logistic Regression	Country Code, Status of company, Category Group List, Funding rounds, Total Funding (USD), Founded date, First funding date, Last funding date, Last funding to date , Twitter URL, and Facebook URL.	Operation, Acquisition, IPO	Crunchbase	Worldwide	215,729 Startups	Accuracy=77.45%
	Reduced Logistic Regression						Accuracy=77.41%
	Rpart Tree						Accuracy=93.63%
	Conditional Inference Tree						Accuracy=85.61%
	Random Forest						Accuracy=94.18%
	Extreme Gradient Boosting						Accuracy=94.45
[35]	Naïve Bayes	Start Date, Initial Funds, Total Rounds of Funding, Time for Seed(in months), Severity Scores Factors for company's growth/fallout, Average Severity Score, Average Severity Score , Venture Round funding , Valuation of the company after each round of funding , Defunct Date when the company dead-pooled(failed companies), Months Active , Market Value , Total Funds , and Burn Rate .	N/A	Crunchbase	Worldwide	11,000 Startups	Recall= 88%
	Alternative Decision Tree						Recall= 95%
	Bayes Net						Recall=92%
	LazyIb1						Recall=69%
	Random Forest						Recall= 97%
	Simple Logistics						Recall= 95%
[36]	Decision Tree	Permalink, Name, Homepage URL, Category list, Market, Funding total USD, Status, Country code, State code, Region, City, Funding rounds, Founded Date, founded month, Founded quarter, Founder year, First founding date, and Last founding date.	Merger, Acquisition	Crunchbase	Worldwide	41,835 Startups	Accuracy=92.43%
	Random Forest						Accuracy=92.43%
	Logistic Regression						Accuracy=92.59%
	Gradient Boosing Classifier						Accuracy=91.96%
	Neural Network						Accuracy=91.86%
[37]	Random Forest	Competitor acquired IOP, Age first funding year, Competitor count, Customer count, Funding rounds, Funding total USD, Investment per round, Investors per round, Round A age, Round A raised amount, Round B age, Round B raised amount, Round C age, Round C raised amount, Round D age, Round D raised amount, Top 500 investor, Total acquisitions, Total exp founders years, Total experience jobs years, Total investments, Total Founders, Total jobs, Category general, USA state code	Acquisition	Crunchbase	United States	143,348 Startups	F1-score=93.2%
[38]	Logistic Regression	Number of Angel Funding Rounds, Number of Milestones, Months from the foundation to the first milestone, Percent of firms in the industry before the year of foundation, Type of education, Years from graduation, The	IPO, Acquisition, Operation (+4 years new milestones)	Crunchbase	United States	100,000 Startups	Accuracy=90.68%
	Linear Discriminant Analysis						Accuracy=82.74%
	Support Vector Machine (SVM)						Accuracy=90.68%
	Extreme Gradient Boosting						Accuracy=90.14%

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 11, Issue 10, October 2024

Ref	Model	Considered Factors	Success/ Target Variables	Dataset			Performance
				Source	Region	Size	
		size of the team, Participants in the Angel investment rounds, Amount of Venture capital investment, Amount of Angel investment, Months to raise angel investment, GDP growth, Type of industry, The situation of the industry, and Location.					
[39]	Decision Tree	N/A	Acquisition	Kaggle	Worldwide	N/A	Accuracy=98.26%
	Random Forest						Accuracy=97.83%
	K-Nearest Neighbour (KNN)						Accuracy=69.70%
	Multilayer Perceptron (MLP)						Accuracy=35.93%
	Naïve Bayes						Accuracy=39.39%
	Support Vector Machine (SVM)						Accuracy=64.06%
	Logistic Regression						Accuracy=64.06%
	Stochastic Gradient Descent						Accuracy=35.93%
[40]	Logistic Regression	Unique identifier, the total amount of funding, Is success, Is Sustainable, Is Tech company, Client Type (B2B_B2C), Number of Founders, Total Degrees, Ratio of founders that have previous business experience, University Rank, Average amount of prior company experience among founders, and Amount of missing values per company.	Successful Exit	Techleap	Dutch	406 Startups	Accuracy=46.8%
	Linear model		Total funding				Accuracy=73.1%
[41]	Multilayer Perceptron (MLP)	Location, Financial capital, Age, R&D, Startup size, Availability of infrastructure, Amount employee skills, Innovation environment, Company revenue, Government regulation, Export products, Access to target market, Innovation of product/ service, Global economic environment, Size of investment, Exchange rates, Environment, Competition, Availability of skilled employees, and Access to export market.	Profitability	Various Source such as company incubators, investment funds, government programs tax agencies and surveys	Australia	265 Companies	Accuracy=71.68%
	Gradient Boosing						Accuracy=76%
	Support Vector Machine (SVM)						Accuracy=84%
	Random Forest						Accuracy=72%
	K-Nearest Neighbours (KNN)						Accuracy=64%
	Neural Network						Accuracy=68%
	Decision Tree						Accuracy=60%

III. DISCUSSION

Generally, we highlighted the efforts of several studies related to predicting the success of startups, showing the development of research in this field and the different methods used to achieve such a goal. While some studies focused on identifying the main factors that affect the success of startups, others highlighted the use of machine learning techniques to build models that predict the success of the startups. However, as can be noticed Table 1 offers a comprehensive overview of diverse machine learning models employed in predicting the success of start-up across varying datasets and regions. Predominantly, the acquisition attribute prominently featured as a significant success determinant across the majority of studies. Notably, Random Forest, Logistic Regression, Support Vector Machine (SVM), and Decision Tree models emerged as frequent choices among these studies, underlining their power in such tasks. Analyzing the considered factors, most of the studies directed their focus toward the geographical, financial, and industrial factors. Specifically, factors such as country, city, location, funding rounds, total funding, and funding dates constituted the primary hubs across most of the research endeavors. Indeed, the papers [40] and [33] explored the prediction of startup success using various models, including Logistic Regression, Linear models, Decision Trees, and Random Forests. Specifically, [40] aimed to predict startup success by focusing on both cases of successful exits or total funding, while [33] predicting the success based on the gained investment. Both studies considered a diverse range of factors; however, [40] included more specific founder-related, university rank, and data completeness factors, while [33] focused on broader aspects such as digital activity and industry specifics. [40] showcased higher accuracy rates for its models around 46.8% to 73.1% compared to the models in [33] with accuracies between 57% to 61.2%. One can notice that the Dutch study [40] provided deeper insights into specific success factors with higher

accuracies, whereas the Ukrainian study [33] explored a wider range of startup success aspects but with comparatively lower accuracies. Moreover, [36] explored Decision Trees, Random Forest, Logistic Regression, Gradient Boosting, and Neural Network models. The models were trained on several factors such as company details, funding information, geographical data, and temporal markers. the study aimed to predict the success of startups in terms of acquisition, achieving notable accuracies ranging from 91.86% to 92.59%. In contrast, [34] employed a diverse set of models—Full and Reduced Logistic Regression, Rpart Tree, Conditional Inference Tree, Random Forest, and Extreme Gradient Boosting. This study covered a range of factors including company status, funding details, social media presence, and historical dates. The used models targeted various success outcomes, including operational status, acquisition, and IPO issuance, achieving accuracies between 77.41% and 94.45%. Additionally, the study in [30] explored Logistic Regression, Support Vector Machine (SVM), and Gradient Boosting. where it delved into various organizational, categorical, founder-related, and temporal factors, aiming to predict success in terms of IOP, Funding, and Acquisition. The used models achieved accuracies ranging between 86% and 87%. While [31] focused on one model, K-Nearest Neighbors (KNN). The study considered factors such as category, country, funding rounds, and temporal differences related to funding where the KNN model aimed to predict startup success based on acquisition and IOP, achieving an accuracy of 73.7%. Notably, the work referenced in [29] focused on factors related to market, city, venture, and financing rounds. The models employed, Random Forest and Support Vector Machine (SVM), achieved commendable accuracies of 89% and 88%, respectively. Moving to [32], the study adopted factors such as burn rate, total valuation, and funding-related metrics. Logistic Regression and Neural Network models were employed for predicting closed or acquired outcomes. as shown Logistic Regression achieves an accuracy of 85.9%. Also, in [27] considered factors related to the start date, funding rounds, and severity scores. Models such as Naïve Bayes, Alternative Decision Tree, and Random Forest were employed, achieving recall metrics ranging from 69% to an impressive 97%. It should be noted that [21] focused on a wide range of factors using Random Forest and SVM, while [32] delved into financial and milestone-related metrics using Logistic Regression and Neural Networks. Additionally, [35] explored the temporal and financial aspects of company growth, employing a diverse set of models. Also, we can clearly see that both studies [37] and [38] operated within the U.S. startup ecosystem and share a common reliance on Crunchbase datasets, they differ in their dataset sizes, factors considered, and success variables predicted. Particularly, the primary focus of [37] was on predicting acquisitions. The Random Forest model employed achieved an impressive F1-score of 93.2%, showing its power in predicting successful acquisitions within the U.S. startup landscape. In contrast, The considered factors in [38] span venture capital and angel investments, industry-related metrics, and geographical aspects. The models used, including Logistic Regression, Support Vector Machine (SVM), Linear Discriminant Analysis, and Extreme Gradient Boosting, aimed to predict various outcomes such as IOP, acquisition, continued operation, and achieving new milestones after four years of operation. The reported accuracy metrics ranged from 82.74% for Linear Discriminant Analysis to 90.68% for both Logistic Regression and SVM, demonstrating robust predictive performance. On the other hand, the [39] study employed a comprehensive set of models, including Decision Tree, Random Forest, K-Nearest Neighbour (KNN), Multilayer Perceptron (MLP), Naïve Bayes, Support Vector Machine (SVM), Logistic Regression, and Stochastic Gradient Descent. The reported accuracy metrics ranged widely, from a remarkable 98.26% for Decision Tree to 35.93% for both Multilayer Perceptron and Stochastic Gradient Descent. This variability in accuracy emphasizes the inherent challenges in predicting acquisitions on a global scale. Lastly, researchers in [41] focused on predicting profitability in the Australian startup landscape. The models used, such as Multilayer Perceptron (MLP), Gradient Boosting, Support Vector Machine (SVM), Random Forest, K-Nearest Neighbours (KNN), Neural Network, and Decision Tree, exhibit diverse accuracy metrics ranging from 60% to 84%. This study underscores the complexity of predicting profitability and the need for a diverse set of models to capture the nuances of factors influencing financial success. Overall, the reviewed studies have contributed valuable insights into the complex nature of predicting startup success. Also, the studies showed the potential of machine learning techniques and highlighted the importance of considering various factors in such prediction tasks. Generally, these studies enriched our understanding of the complex dynamics involved in predicting startup success. it should be noticed that the implications of these findings can extend far beyond academic purposes.

IV. CONCLUSION

In conclusion, predicting startup success remains a complex challenge, characterized by uncertainty and a high failure rate, particularly in the early years of operation. Traditional methods, while valuable, often fall short in addressing the intricacies and dynamic nature of startups. The rise of machine learning offers a promising alternative, leveraging large datasets and advanced algorithms to uncover hidden patterns and provide more accurate predictions. As reviewed in this paper, various approaches ranging from decision trees to neural networks have been employed, each bringing unique insights into the factors that contribute to startup success.

However, despite the progress made, there is still

significant room for improvement. The scalability and reliability of machine learning models in this context are not yet fully realized, and challenges such as data availability, model generalizability, and the rapidly changing environments in which startups operate must be addressed. Future research should focus on refining these models, expanding the diversity of data sources, and integrating contextual factors unique to specific markets, such as the Saudi market. By doing so, machine learning can become an even more powerful tool for investors, entrepreneurs, and market researchers, providing actionable insights and contributing to more informed decision-making in the startup ecosystem.

REFERENCES

[1] Alsolaim M. Barriers to Survival for Small Start-up Businesses in Saudi Arabia. PhD diss., University of Brighton, 2019.

[2] Tomy S, Pardede E. From Uncertainties to Successful Start Ups: A Data Analytic Approach to Predict Success in Technological Entrepreneurship. Sustainability. 2018; 10(3): 602. doi:10.3390/su10030602.

[3] Tufail S, Riggs H, Tariq M, Sarwat AI. Advancements and Challenges in Machine Learning: A Comprehensive Review of Models, Libraries, Applications, and Algorithms. Electronics. 2023; 12(8): 1789. doi:10.3390/electronics12081789.

[4] Kim B, Kim H, Jeon Y. Critical Success Factors of a Design Startup Business. Sustainability. 2018; 10(9): 2981. doi:10.3390/su10092981.

[5] Blank S, Dorf B. The Startup Owner's Manual. John Wiley & Sons, 2020.

[6] Ries E. The Lean Startup. Crown Business, 2017.

[7] Cohen B, Amorós JE, Lundy L. The Generative Potential of Emerging Technology to Support Startups and New Ecosystems. ScienceDirect. 2017; 60(6): 714–745. doi:10.1016/j.bushor.2017.06.004.

[8] Gimpel H, Rau D, Röglinger M. Understanding Fintech Startups – A Taxonomy of Consumer-Oriented Service Offerings. Electronic Markets. 2017; 28(3): 245–264. doi:10.1007/s12525-017-0275-0.

[9] Díaz-Santamaría C, Bulchand-Gidumal J. Econometric Estimation of the Factors that Influence Startup Success. Sustainability. 2021; 13(4): 2242. doi:10.3390/su13042242.

[10] Kuzmianok D. Socioeconomic Impact of Startup Companies: The Republic of Belarus - Prospects and Challenges of Startup Ecosystem. 2016. Available online: https://monami.hs-mittweida.de/frontdoor/index/index/docId/8383.

[11] Hannan MT, Freeman J. Organizations and Social Structure in Organizational Ecology. Scott articles. Available online: https://faculty.babson.edu/krollag/org_site/org_theory/scott_articles/han_free_orgec.html#:~:text=Population%20ecology%20is%20the%20study,the%20population%20over%20long%20periods (accessed Oct. 30, 2023).

[12] Ford MR. Population Ecology Theory of Organizations. Global Encyclopedia of Public Administration, Public Policy, and Governance. 2018; pp. 4830–4834. doi:10.1007/978-3-319-20928-9_74.

[13] Henderson K, Loreau M. An Ecological Theory of Changing Human Population Dynamics. People and Nature. 2019; 1(1): 31–43. doi:10.1002/pan3.8.

[14] Xu J, Peng B, Cornelissen J. Modelling the Network Economy: A Population Ecology Perspective on Network Dynamics. Technovation. 2021; 102: 102212. doi:10.1016/j.technovation.2020.102212.

[15] Vest JR, Menachemi N. A Population Ecology Perspective on the Functioning and Future of Health Information Organizations. Health Care Management Review. 2017; 44(4): 344–355. doi:10.1097/hmr.0000000000000185.

[16] Casanova L, Dutta S, Cornelius PK. Financing Entrepreneurship and Innovation in Emerging Markets. 1st ed. Academic Press, 2018.

[17] Alsolaim MA. Barriers to Survival for Small Startup Businesses in Saudi Arabia. Thesis, 2019.

[18] Cantamessa M, Gatteschi V, Perboli G, Rosano M. Startups' Roads to Failure. Sustainability. 2018; 10(7): 2346. doi:10.3390/su10072346.

[19] Rezaei J, Ortt R. Entrepreneurial Orientation and Firm Performance: The Mediating Role of Functional Performances. Management Research Review. 2018; 41(7): 878–900. doi:10.1108/mrr-03-2017-0092.

[20] Avagyan V, Camacho N, Van der Stede WA, Stremersch S. Financial Projections in Innovation Selection: The Role of Scenario Presentation, Expertise, and Risk. International Journal of Research in Marketing. 2022; 39(3): 907–926. doi:10.1016/j.ijresmar.2021.10.009.

[21] Shang Z. The Research of Financial Forecasting and Valuation Models. Proceedings of the 2021 International Conference on Enterprise Management and Economic Development (ICEMED 2021). 2021. doi:10.2991/amber.k.210601.012.

[22] Fakieh B, AL-Malaise AL-Ghamdi AS, Ragab M. The Effect of Utilizing Business Model Canvas on the Satisfaction of Operating Electronic Business. Complexity. 2022; 2022: 1–10. doi:10.1155/2022/1649160.

[23] Shakeel J, Mardani A, Chofreh AG, Goni FA, Klemeš JJ. Anatomy of Sustainable Business Model Innovation. Journal of Cleaner Production. 2020; 261: 121201. doi:10.1016/j.jclepro.2020.121201.

[24] Weber P, Carl KV, Hinz O. Applications of Explainable Artificial Intelligence in Finance—A Systematic Review of Finance, Information Systems, and Computer Science Literature. Management Review Quarterly. 2023. doi:10.1007/s11301-023-00320-0.

[25] S. Social Entrepreneurship from the Perspective of Opportunity: Integration Analysis Based on Timmons Process Model. Journal of Human Resource and Sustainability Studies. 2019; 07(03): 438–461. doi:10.4236/jhrss.2019.73029.

[26] Ghee WY. An Application of Timmons Model in the Mini Entrepreneurial Logistics Project. Advances in Social Sciences Research Journal. 2018; 5(10). doi:10.14738/assrj.510.5541.

[27] Sarker IH. Machine Learning: Algorithms, Real-World Applications, and Research Directions. SN Computer Science. 2021; 2(3). doi:10.1007/s42979-021-00592-x.

[28] m (SIPLah) Using the UMEGA Model: Based on the Perspective of Users in Education Units. International Conference on Information Science and Technology Innovation (ICoSTEC). 2022; 1(1): 117–122. doi:10.35842/

icostec.v1i1.10.

[29] Li J. Prediction of the Success of Startup Companies Based on Support Vector Machine and Random Forest. 2020 2nd International Workshop on Artificial Intelligence and Education. 2020. doi:10.1145/3447490.3447492.

[30] Żbikowski K, Antosiuk P. A Machine Learning, Bias-Free Approach for Predicting Business Success Using Crunchbase Data. Information Processing & Management. 2021; 58(4): 102555. doi:10.1016/j.ipm.2021.102555.

[31] Pan C, Gao Y, Luo Y. Machine Learning Prediction of Companies' Business Success. CS229: Machine Learning. Fall 2018. Stanford University, CA.

[32] Shah V. Predicting the Success of a Startup Company. support.sas.com. 2019. Available online: https://support.sas.com/resources/papers/proceedings19/3878-2019.pdf (accessed Nov. 27, 2023).

[33] [33] Piskunova O, Ligonenko L, Klochko R, Frolova T, Bilyk T. Applying Machine Learning Approach to Start-up Success Prediction. Scientific Horizons. 2022; 24(11): 72–84. doi:10.48077/scihor.24(11).2021.72-84.

[34] [34] Ünal C, Ceasu I. A Machine Learning Approach Towards Startup Success Prediction. Humboldt-Universität zu Berlin. International Research Training Group 1792 'High Dimensional Nonstationary Time Series'. Berlin, 2019.

[35] [35] Krishna A, Agrawal A, Choudhary A. Predicting the Outcome of Startups: Less Failure, More Success. 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW). Dec. 2016. doi:10.1109/icdmw.2016.0118.

[36] 36] Bangdiwala M, Mehta Y, Agrawal S, Ghane S. Predicting Success Rate of Startups Using Machine Learning Algorithms. 2022 2nd Asian Conference on Innovation in Technology (ASIANCON). Aug. 2022. doi:10.1109/asiancon55314.2022.9908921.

[37] da Silva Ribeiro Bento FR. Predicting Start-up Success with Machine Learning. 2018.

[38] Veloso F. Predicting Startup Success in the US. The University of North Carolina at Charlotte, 2020.

[39] Zoayed MT, Arshe S, Rahman F. Startup Success Prediction Using Classification Algorithms. Jun. 2022.

[40] Fidder D. Finding the Most Significant Predictors of Startup Success with Machine Learning. Eindhoven University of Technology. Jan. 2022.

[41] Vasquez E, Santisteban J, Mauricio D. Predicting the Success of a Startup in Information Technology Through Machine Learning. International Journal of Information Technology and Web Engineering. 2023; 18(1): 1–17. doi:10.4018/ijitwe.323657.