Not Just Luck: Predicting Startup Success with Machine Learning

Shobhanam Krishna*

Organisational Behaviour and Human Resources, Indian Institute of Management Shillong, India

Anita Choudhary

Organisational Behaviour and Human Resources, Indian Institute of Management Shillong, India

Rohit Dwivedi

Organisational Behaviour and Human Resources, Indian Institute of Management Shillong, India

Ashutosh Bishnu Murti

Organisational Behaviour and Human Resources, Indian Institute of Management Shillong, India

Abstract

Startups are central to innovation but face a staggering failure rate of nearly 90%, largely due to misaligned product-market fit, limited capital, flawed business models, or ineffective leadership. Traditional methods for predicting startup outcomes often fail to account for the complex, nonlinear dynamics of entrepreneurial ecosystems. This study aims to identify critical success factors and develop a predictive model to assess startup viability using supervised machine learning, thereby improving investment decision-making and supporting ecosystem resilience. Leveraging a global Crunchbase dataset with over 100 multidimensional features, the study employs Principal Component Analysis (PCA) for dimensionality reduction and Random Forest for classification. A five-step data mining framework was used to process, transform, and analyze the data. The model classifies startups as successful (acquired/operating) or failed (closed), with predictive accuracy evaluated through confusion matrices and F1-scores. The model achieved 93.9% test accuracy and 0.82 confusion matrix accuracy. Key predictors include founder background, funding patterns, industry, geography, and early-stage agility. Practical experience and consulting exposure were stronger predictors than formal education. PCA helped enhance interpretability and reduce noise. The study bridges Human Capital Theory with machine learning, offering a rare integration of theoretical grounding and algorithmic modeling. Unlike prior research, it combines PCA with Random Forest, enabling robust, interpretable predictions on a large-scale, global dataset. The findings inform investors, incubators, and policymakers on optimizing startup selection, funding strategies, and ecosystem development. The model serves as a scalable tool for enhancing startup evaluation, resilience, and long-term innovation impact.

Keywords: Crowdfunding Platforms, Data Mining, Machine Learning, Marketing Strategies, Startup Success.

JEL Classifications: M12, M13, M14, M15, M51, M53, M54, O15, E24

*Corresponding author
E-mail: shobhak.phd22@iimshillong.ac.in
Address: Iim Shillong, East Khasi Hills, Umsawli, Shillong.

1. Introduction

"The best way to predict the future is to create it." - Abraham Lincoln.

Startups are the engines of innovation, daring to create new futures amidst uncertainty and risk. Each successful venture from Amazon's humble beginnings in a garage to SpaceX's ambitious quest to colonize Mars began as an idea fuelled by vision and resilience. Yet, beneath these celebrated triumphs lies a stark reality: nearly 90% of startups ultimately fail, commonly due to challenges such as poor product-market fit, inadequate capital, flawed business models, or ineffective leadership (Forbes, 2024). At their core, startups are organizational entities founded to develop innovative products or services amid considerable uncertainty (Ries, 2011). They represent the earliest operational stages of a company, typically driven by one or more entrepreneurs devoted to transforming novel ideas into viable offerings. However, these ventures often encounter substantial initial costs and limited revenue streams, necessitating significant external financing from sources such as venture capitalists, angel investors, or crowdfunding platforms to sustain and scale their growth (Bonini & Capizzi, 2019). Startups undergo multiple funding rounds to secure capital, allowing external investors to invest funds in exchange for equity or partial ownership of the company. Additional investment avenues include debt, convertible notes, stocks, or dividends. Startups often begin with "seed" funding or backing from angel investors. The overarching goal for most startups is to either be acquired by another company or become publicly traded. For instance, Facebook's acquisition of WhatsApp for nineteen billion dollars resulted in a 50x return on investment for Sequoia, a venture capital fund (Ahluwalia & Kassicieh, 2024; Chang, 2003). However, it is essential to note that startups carry an estimated 90% probability of failure, translating to substantial investments without assured returns (Patel and Pearce II, 2018). A study by CB Insights examined the post-mortems of 101 startups to compile a list of the Top 20 Reasons Startups Fail, focusing on company-level causes for failure. This investigation's nine most significant findings include a lack of market need, running out of cash, assembling the wrong team, facing intense competition, pricing/cost issues, delivering a subpar product, lacking a viable business model, ineffective marketing, and neglecting customer input. The present study centers on how startups or investors can leverage this knowledge for informed decision-making in investment strategies and financial gains. Employing data mining and machine learning techniques, the study aims to create a predictive model with a dependent variable classifying whether or not a startup is successful. At the core of the problem is the persistently high failure rate of startups, commonly driven by misalignment with market needs, marketing inefficiencies, and team related shortcomings. While venture capitalists have traditionally borne the bulk of startup risk, the advent of crowdfunding now enables individual investors to participate in this high-stakes domain.

The rationale for employing machine learning techniques, supervised machine learning in predicting startup success is grounded in its superior capacity to learn from labelled datasets and model complex, non-linear interdependencies among a multitude of predictor variables. Startups operate within volatile, dynamic, and multifactorial environments wherein traditional statistical methodologies, often encumbered by assumptions of linearity, normality, and independence, are rendered inadequate to capture the nuanced and evolving intricacies of entrepreneurial ecosystems (Ermers et al., 2020; Rollmann et al., 2023). Supervised machine learning algorithms encompassing logistic regression, decision trees, support vector machines, and neural networks demonstrate pronounced suitability in this context, given their ability to systematically discern intricate patterns, latent structures, and multifaceted relational dynamics embedded within historical datasets of successful and failed ventures (Potanin et al., 2023; Zhang et al., 2021). Through rigorous training on expansive, labelled datasets, these models facilitate the accurate classification and prediction of startup outcomes, thereby offering a robust and scalable alternative to conventional predictive frameworks. Furthermore, the interpretability afforded by certain supervised models enhances their strategic applicability, providing decision-makers with critical insights into the underlying determinants of venture success or failure (Gu et al., 2024). The iterative refinement capacities inherent to supervised learning further augment its predictive fidelity, enabling continuous adaptation to novel data patterns and emergent market dynamics (Sen et al., 2021; Eckart et al., 2021). Advanced supervised architectures, notably multilayer perceptron networks, have demonstrated substantial efficacy in capturing non-linearities, high-order interactions, and complex variable interdependencies that traditional models frequently overlook (Zhang et al., 2021). The integration of supervised machine learning into entrepreneurship studies marks a profound and transformative shift, offering a fresh epistemological lens through which the intricacies of startup trajectories can be deciphered and forecasted (Jordan & Mitchell, 2015). Machine learning technologies including regression-based models, classification algorithms, neural networks, and ensemble techniques exhibit significant promise in identifying latent patterns, interrelationships, and critical success factors that influence firm trajectories (Basole, Park, & Seuss, 2023). Through the exploitation of extensive and heterogeneous datasets, these algorithms unearth the subtle and often hidden drivers of entrepreneurial outcomes, facilitating the creation of highly precise predictive models capable of projecting a venture's success or failure with remarkable accuracy. The necessity of adopting a machine learning driven approach is further accentuated by the limitations of conventional analytical methods in managing the voluminous and intricate data characteristic of contemporary business environments (Allioui & Mourdi, 2023). Unlike traditional statistical approaches, which often falter when confronted with non-linear relationships and complex interaction effects, machine learning algorithms are inherently designed to navigate and model such complexities (El Hajj & Hammoud, 2023). Consequently, supervised machine learning not only enhances predictive precision but also enriches the theoretical and practical frameworks through which entrepreneurial success is understood, offering profound implications for scholars, practitioners, and policymakers engaged in navigating the complexities of modern startup ecosystems (Graham & Bonner, 2022; Nacchia et al., 2021; Leo et al., 2019).

To aid startup investors in their decision-making process, this paper aims to identify critical features contributing to startup success and predict a company's success using supervised machine learning methods. The specific objective of this paper is to:

i. To explore the interactions and relationships among various variables influencing startup behavior and success.

ii. To develop a methodological framework for assessing the importance of variables in determining startup success.

iii. To outline a transparent and detailed process for building a predictive model to assess startup success using machine learning algorithms.

This study makes several significant contributions to the entrepreneurship and predictive analytics literature. First, it extends existing research on startup success prediction by combining Principal Component Analysis (PCA) with the Random Forest algorithm to improve model robustness, interpretability, and predictive power, an integration rarely explored in prior studies. Second, the study enriches the literature by using a global Crunchbase dataset with multidimensional variables, providing a much broader and deeper empirical base than earlier works, which were typically limited to smaller, regional samples. Third, the research addresses a critical theoretical gap by operationalizing Human Capital Theory within a machine learning framework, thereby offering a nuanced understanding of how founders' education, consulting experience, entrepreneurial skills, and strategic competencies influence startup outcomes. This advances prior models that predominantly emphasized formal education while neglecting the strategic importance of applied human capital and practical exposure. The research also shifts the traditional focus of algorithmic comparison studies toward practical implications for investors, entrepreneurs, and policymakers, providing a comprehensive strategic framework to enhance startup success rates. By linking empirical insights with actionable strategies, the study contributes to both theoretical advancement and real-world application, offering a foundation for future interdisciplinary research in entrepreneurship, data science, and innovation management. The paper is structured as follows: it begins with an introduction, followed by a literature review, methodology, data analysis, discussion of implications, and concludes with a final section synthesizing key insights and future research directions.

2. Literature Review

Startups represent companies that introduce innovative products or services in ways not previously explored. This inherent novelty renders startups precarious and unpredictable, as their offerings may only resonate with intended users, necessitating continual adjustments once achieving product/market fit. Fundamentally, a startup embodies a high-risk enterprise in its initial operational stage, often associated with technological products or services (Ries, 2011). Engaging with startups involves a gamble of higher risk for the potential of higher returns. Acknowledging the substantial risk and the considerable percentage of startups that face failure, coupled with the exponential surge in startup formations in the United States and Europe, along with their growing significance in national economies, underscores the importance of quantitatively studying a phenomenon that presents challenges on a global scale. The success of a startup typically follows a dual strategy, where a company can opt for an IPO by entering the public stock market, allowing shareholders to sell shares to the public. Alternatively, the company may pursue acquisition or merger with another entity, enabling previous investors to receive cash in exchange for their shares promptly. This strategic decision is called an exit strategy (Guo et al.,2015). In the startup ecosystem, going public or being acquired is commonly regarded as a success for the company, as these events bring substantial immediate monetary gains to founders, investors, and early employees (Guo et al.,2015). A prevalent motivation for larger companies to acquire smaller startups is to access their talent pool. In such acquisitions, known as "acquiring," the parent company not only acquires technology but also recruits the skilled employees of the startup, providing a swift strategy for growth in competitive markets (Marita Makinen et al.,2014).

2.1 Machine Learning Approaches to Predict Startup Success

The application of machine learning techniques to predict business success has a long-standing history. Early studies, such as Lussier (1995), employed logistic regression models based on survey data collected from U.S. companies to forecast the success of young firms. Similarly, other researchers explored the impact of macroeconomic factors on the success of Australian ICT companies, utilizing machine learning algorithms such as K-Nearest Neighbors (K-NN), Naive Bayes, and Support Vector Machines (SVM) (Tomy & Pardede, 2018). However, these early studies were limited by the relatively small size of their datasets, containing only 216 and 250 instances, respectively. The availability of larger datasets, particularly from platforms like Crunchbase, significantly expanded research possibilities. Using Crunchbase data allowed researchers to work with thousands of instances and incorporate detailed information about funding events into predictive models (Krishna et al., 2016). Bento (2018) utilized Crunchbase data on U.S. startups to predict outcomes such as acquisition or initial public offering (IPO) by applying Logistic Regression, SVM, and Random Forest algorithms. Building on these efforts, Sharchilev et al. (2018) developed a gradient boosted decision tree model to predict whether companies that had secured seed or angel funding would obtain Series A funding within the following year. Their study enhanced the Crunchbase dataset by incorporating additional data from LinkedIn profiles of employees working at the respective startups, thereby enriching the feature set used for prediction. These studies illustrate the evolving use of machine learning techniques in the domain of startup success prediction, highlighting the shift towards leveraging larger and richer datasets to improve model accuracy and generalizability.

2.2 Random Forest for Startup Classification

Ensemble learning techniques, particularly those based on decision trees, have proven highly effective in improving model accuracy and stability. Among these, the Bagging Classifier plays a pivotal role. As a meta-classifier, bagging enhances the performance of tree-based algorithms by reducing variance and mitigating the risks of overfitting (Subasi et al., 2020). The bagging method involves bootstrap sampling, wherein multiple tree models are trained on different subsets of the training data. By aggregating the predictions of these models, bagging produces a more robust and accurate overall prediction. This technique is particularly useful for improving the predictive capabilities when assessing the likelihood of startup success. Building upon bagging principles, the Random Forest Classifier extends the idea of ensemble learning by introducing further randomness. In Random Forest, each tree is trained on a bootstrap sample of the dataset, but crucially, each split in a tree considers only a random subset of features, not all available features. This additional layer of randomness increases the diversity among individual trees, effectively reducing overfitting and leading to higher generalization capabilities (Pandimurugan et al., 2022). As a result, Random Forest classifiers deliver stronger predictive performance across a wide range of classification tasks, including startup success prediction. Another important tree-based ensemble method is the Extra Trees Classifier. While conceptually similar to Random Forest, Extra Trees introduces even more randomness by selecting cut points for splitting nodes at random, rather than searching for the most discriminative thresholds (Sharma et al., 2022). Additionally, Extra Trees trains each base estimator on a random subset of features during the training phase. This heightened randomization increases the diversity among trees even further, leading to improved generalization and boosting the overall predictive power of the model. In this study, the Random Forest algorithm is employed for accurate prediction through several systematic steps:

Step 1: Bootstrap Sampling

The first step involves generating n bootstrap samples from the training dataset. Considering a dataset (A, V) where observations are represented by elements a1, a2, a3,...,ad and responses by v1,v2,..., vd, new datasets (Ac, Vc) are created by randomly selecting observations with replacement to form T datasets of size d.

Step 2: Model Training

Each bootstrap sample (Ac, Vc) is used to train a decision tree independently, for C=1,2, 3..., c

Step 3: Prediction Aggregation

For predicting new data points, each trained decision tree casts a vote, and the final prediction is determined by majority voting across all trees.

Step 4: Model Evaluation

After generating predictions from the out-of-bag (OOB) samples, the mean prediction error for each tree is calculated. This helps estimate the OOB error rate, offering an internal validation measure. Alternatively, the dataset can be divided into training and testing sets to calculate the model’s error rate more traditionally.

2.3 Principal Component Analysis (PCA) for Feature Reduction

As datasets become increasingly complex with a surge in features and dimensions, data scientists grapple with the curse of dimensionality, a phenomenon that results in overfitting, escalated computational demands, and decreased model accuracy (Wáng et al., 2023). High dimensional data challenges the extraction of statistically meaningful insights as algorithms must navigate expansive feature spaces, thereby exponentially amplifying the complexity and time needed for classification and clustering tasks (Wiriyathammabhum & Kijsirikul, 2012). To overcome these hurdles, dimensionality reduction techniques such as PCA are pivotal, simplifying datasets while preserving essential information (Jeon & Oh, 2020; Tokuda et al., 2018). PCA effectively transforms high dimensional data into a smaller set of uncorrelated variables, principal components by identifying directions that capture the greatest variance within the data (Nurhasanah & Safitri, 2012; Tan et al., 2021; Ringnér, 2008). This process involves structuring the dataset into a two-dimensional matrix, standardizing features to prevent dominance by those with larger ranges, calculating the covariance matrix, and extracting eigenvalues and eigenvectors (Smith, 2002; Jolliffe, 2005; D'Acquisto & Naldi, 2019; Ma & Dai, 2011). By maximizing variance retention and minimizing information loss, PCA generates informative low-dimensional projections, particularly valuable when data contain small independent and identically distributed (i.i.d.) noise or sparse errors (Zhou et al., 2010). However, PCA's sensitivity to outliers can distort eigendirections, necessitating robust alternatives that use techniques like robust scaling and Huber M-estimators to reduce outlier influence (Hubert et al., 2005; Podosinnikova et al., 2014; McCoy & Tropp, 2011; He et al., 2023). Developed originally to condense large variable sets while preserving maximum variance (Jolliffe, 1986), PCA extends beyond dimensionality reduction to applications in data visualization, feature identification, and fault detection (Velandia et al., 2021), and addresses correlated variable issues by transforming them into uncorrelated components (Sainani, 2014). Nevertheless, while PCA components with the largest eigenvalues capture significant variance, they may not always optimally expose the data's cluster structure, particularly in datasets affected by outliers (Banerjee et al., 2022; Gharibnezhad et al., 2014). Consequently, PCA remains a powerful but nuanced tool, widely applied in fields such as statistics, neuroscience, and image compression (Pankavich & Swanson, 2015).

2.4 Theoretical Underpinnings

This study is anchored in Human Capital Theory, emphasizing the critical role of founders' accumulated knowledge, skills, and experiences in shaping the trajectory of startup ventures. Human Capital Theory posits that education, skills, and practical experiences significantly enhance an individual's productivity and innovation capacity, thereby influencing organizational performance and economic prosperity (Unger et al., 2009; Chen & Chang, 2013).The research aligns with this theoretical perspective by identifying education, industry experience, consulting exposure, and entrepreneurial competencies as pivotal elements of human capital that determine startup success (Thomas & Bhasi, 2018; Wang et al., 2019). The importance of investing in human capital through education and skill development as a means of improving organizational outcomes is reinforced throughout the study (Torrech, 2018).

Central to the study's argument is the assertion that competency based assessments, rooted in real-time, behavioral, and contextual intelligence offer greater predictive power than static or qualification-based metrics (Østergaard & Marinova, 2018; Fini & Toschi, 2016). Just as applied human capital enhances startup resilience, the research suggests that organizational success hinges on the ongoing identification and activation of context-relevant skills across roles and functions. In entrepreneurial contexts, applied human capital manifested through real-world problem-solving and strategic decision-making skills often proves more critical than theoretical knowledge alone (Fini & Toschi, 2016). Moreover, the study operationalizes Human Capital Theory within a machine learning-based predictive model, integrating human capital variables to derive actionable insights into founder success factors. This methodological innovation bridges theoretical foundations with data-driven analytics, demonstrating how diverse human capital dimensions can be systematically linked to startup performance (Millán et al., 2013). Such integration advances the practical application of Human Capital Theory, offering valuable guidance to investors, policymakers, and startup incubators in identifying and nurturing high-potential ventures.

The research highlights that entrepreneurial competencies facilitate the translation of prior knowledge into practice through an effectual approach, with cognitive characteristics and strategic thinking playing a decisive role in determining startup success (Schmidt & Heidenreich, 2018). In today's competitive markets, the increasing emphasis on entrepreneurial human capital underscores the necessity for founders to possess not just educational credentials, but also adaptive, strategic, and service-oriented capabilities. Furthermore, differences in human capital influence the creation of social capital the networks and relationships that help overcome obstacles in venture development particularly among academic entrepreneurs (Mosey & Wright, 2007). This dynamic interplay between human and social capital demonstrates how diverse forms of capital mutually reinforce entrepreneurial performance and venture growth (Ashourizadeh et al., 2014). By empirically testing the impact of multiple human capital dimensions, the study offers a granular understanding of the specific skills, experiences, and competencies that matter most in entrepreneurial success. This address calls in the literature for more detailed investigations into how different units of human capital affect entrepreneurial development and economic growth (Torrech, 2018). The findings have significant practical implications. For investors, identifying founders with the right blend of educational and experiential backgrounds can enhance the probability of higher returns. Policymakers can design targeted programs to cultivate specific competencies such as consulting experience and strategic thinking, critical for startup success. Similarly, startup incubators can tailor their support to address skill gaps, enhancing the readiness and resilience of emerging ventures (Perez & Guevarra, 2020).

3. Methodology

This study aims to predict the likelihood of startup success or failure, where success is operationalized as a company being either Acquired or Operating, and failure as a company being Closed. To construct the predictive model, a comprehensive set of over 100 input variables was utilized (refer to Annexure for details). These variables captured a wide range of information, including company basics (such as industry, country, and business model type), founders' backgrounds (education, experience, and global exposure), team composition, funding history and growth metrics, technology orientation (such as use of machine learning or big data), founders' and teams' skillsets across different domains, and factors related to reputation and legal risks. This study employed a structured five-step process for data analysis, outlined by Fayyad et al. (1996). The methodology consists of: (1) Selection, identifying relevant target data; (2) Pre-processing, transforming target data into a clean, structured format; (3) Transformation, preparing data for model building; (4) Data Mining, applying machine learning algorithms to extract patterns; and (5) Interpretation/Evaluation, deriving knowledge from the discovered patterns.

3.1 Data Collection

The primary dataset was sourced from the Crunchbase platform (Crunchbase Inc., 2020), which offers comprehensive business information covering private and public companies, founders, leadership, investors, and funding rounds. The dataset consisted of approximately 54,000 records, encompassing 472 rows and 116 columns. Key attributes included company name, URL, market, country, state, region, city, founding date, first funding date, and last funding date. Further, the dataset contained detailed information on multiple types of funding activities such as seed funding, angel investment, venture capital, equity crowdfunding, convertible notes, debt financing, grants, private equity, post-IPO equity, and secondary market investments. Companies operational status was categorized into "acquired," "operating," or "closed." Definitions and characteristics of various funding types were adopted directly from Crunchbase descriptions, encompassing early-stage investments like angel and pre-seed rounds to later-stage private equity and post-IPO activities.

3.2 Data Preprocessing

Following data collection, extensive preprocessing procedures were conducted. First, extraneous spaces and unwanted characters (such as commas and hyphens) were removed from textual fields. Numeric columns were verified and corrected to ensure appropriate data types, and all date fields were converted to standard date formats. Rows containing null values were eliminated to maintain data consistency. Columns with a high proportion of missing data specifically 'state,' 'city,' 'region,' and 'founding date' were excluded from the analytical framework.

3.3 Feature Engineering

Feature engineering was undertaken to enhance the dataset's predictive capacity. Key transformations included:

i. Creation of a new feature capturing the time difference between the first and last funding dates.

ii. Construction of a Total Investment feature by summing across all investment types, including seed, venture, equity crowdfunding, convertible note, debt financing, angel, grant, private equity, post-IPO equity, post-IPO debt, secondary market, and product crowdfunding.

iii. Consolidation of the 753 distinct market categories into 43 broader Industry Groups based on Crunchbase's industry segmentation.

iv. Joining the dataset with an external source to create a Continent variable based on the country information.

v. Discretization of continuous numerical variables (e.g., total investment, funding duration) into categorical bins such as "low" and "high" based on distribution thresholds.

vi. Conversion of sparse binary investment indicators (e.g., equity crowdfunding, convertible notes, debt financing) into 0/1 format to simplify model input.

These transformations aimed to reduce dimensionality, normalize data distributions, and create meaningful features aligned with business context.

3.4 Model Building and Machine Learning Approach

The predictive modeling utilized the Random Forest algorithm, an ensemble method that aggregates the outputs of multiple decision trees to improve classification performance. Each decision tree independently predicts a class label, and the Random Forest aggregates these predictions through majority voting (Krishna and Sidharth, 2022). Random Forest was chosen due to its high accuracy, robustness to overfitting, and ability to handle complex interactions between features. It is particularly effective in classification tasks where variables may have nonlinear relationships with the target variable. The research focused on a binomial classification problem, where the goal was to predict whether a company would be classified as "acquired" or "closed." Companies still operating were excluded from the classification task to sharpen the focus on successful exits versus failures.

3.5 Model Evaluation

The model performance was evaluated using a confusion matrix (Krishna and Sidharth, 2023), which captures the true positives, true negatives, false positives, and false negatives. This evaluation metric provides a comprehensive view of the model's predictive ability, beyond simple accuracy, by highlighting how well the model distinguishes between the two target classes.

4. Data Analysis

The data analysis involves key exploratory data analysis (EDA) on various aspects of the dataset. Notably, most companies belong to the software and mobile industries, with the mobile industry securing the highest total funding. Conversely, the hospitality industry exhibits the second-lowest total funding (Figure 1). The landscape of startup funding is a multifaceted arena, influenced by a plethora of factors ranging from industry affiliation to geographical location and the intricacies of investor networks (Berns & Schnatterly, 2015). EDA serves as a pivotal tool in unravelling the complexities inherent in startup datasets, offering invaluable insights into the distribution of funding across various sectors and the underlying dynamics that govern investment decisions (Sudaryana et al., 2024). The concentration of funding within specific industries, such as software and mobile, underscores the prevailing trends in technological innovation and the enduring appeal of these sectors to investors seeking high-growth opportunities (Sudaryana et al., 2024). The prominence of the mobile industry in securing the highest total funding likely reflects the ubiquitous nature of mobile technology and its pervasive impact on various aspects of modern life, from communication and entertainment to commerce and productivity (Gui et al., 2017). This is consistent with the venture capital community's historical preference for scalable, high-return opportunities in tech sectors that promise rapid adoption and global applicability (Ghosh & Nanda, 2010). Conversely, the hospitality industry’s position as the recipient of the second-lowest total funding may be attributed to its relatively lower scalability compared to technology-driven sectors, its capital-intensive operational model, and its susceptibility to economic fluctuations and seasonal variations. This divergence in funding allocation highlights the heterogeneous nature of the startup ecosystem, where industries with greater perceived growth potential and scalability tend to attract a larger share of investment capital, shaping the trajectory of innovation and economic development. Venture capital plays a crucial role in propelling breakthrough innovations within the economy, particularly in the early stages of startup development (Howell et al., 2020). By analyzing investment patterns, stakeholders can gain a deeper understanding of the factors that drive startup success, informing strategies for fostering innovation and economic growth. The rise of alternative funding models such as crowdfunding, online angel syndicates, and online lending reflects the democratization of capital access and the increasing efficiency of capital formation (Foundation, 2016). These alternative models cater to a broader spectrum of startups, including those that may not align with the traditional venture capital paradigm, thereby fostering greater diversity and inclusivity within the entrepreneurial ecosystem. The critical role of venture capital in nurturing startups stems from its ability to provide not only financial resources but also invaluable expertise, mentorship, and networking opportunities, which are essential for navigating the challenges of early-stage growth (Ghosh & Nanda, 2010). Furthermore, venture capital activity significantly elevates the rate of patenting within an industry, suggesting that venture-backed companies are more likely to engage in innovative activities (Kortum & Lerner, 1998). Despite representing a small portion of overall R&D spending, venture capital accounts for a substantial proportion of industrial innovations, underscoring its catalytic role in technological advancement and economic prosperity. The significance of venture capital in the dynamic landscape of small entrepreneurial ventures cannot be overstated (Babouei, 2018). Venture capitalists provide not only the crucial financial backing necessary for these ventures to flourish, but also invaluable counsel that is indispensable for navigating the intricate challenges inherent in the business world (Cumming & Cumming, 2010; Gui, 2024). The contribution of social networks to the triumph of social enterprises has been well-documented, with evidence showing that venture capital firms with stronger networks tend to excel in identifying and supporting promising ventures (Bocken, 2015). These findings underscore the importance of robust networks in accessing information, resources, and opportunities, all of which are vital for the success of startups. Venture capitalists often leverage their networks to connect startups with potential customers, partners, and future investors, thereby enhancing their market traction and long-term sustainability. At the same time, they exhibit a dual nature by offering developmental support while exercising control, as seen in practices such as replacing founding CEOs to ensure business viability (Hellmann & Puri, 2002).

Beyond financial capital, human capital plays an equally pivotal role in the success of startups. Firms that invest in general human capital are more likely to generate novel, innovative ideas, thereby increasing the demand for venture capital services (Sevilir, 2009). The ability to secure capital from venture capital institutions hinges not only on tangible assets but also on the competencies and strategic vision of the management team (Xue & Yun, 2022). Human capital, particularly in the form of experienced entrepreneurs, bridges the gap between academic research and industry application, enabling the commercialization of innovative ideas (Mosey & Wright, 2007). It serves as the bedrock of knowledge-based economies, influencing how entrepreneurs organize and scale their ventures (Neergaard & Krueger, 2007). Moreover, entrepreneurial competence, motivation, and creativity-core components of human capital positively impact the performance of new ventures (Chen & Chang, 2013; Sevilir, 2009; Torrech, 2018). However, it is not merely the possession of human or social capital that guarantees success, but the effective utilization of these assets through strategic action (Gordon, 2012). Drawing from human capital theory, it becomes clear that nascent business ventures can achieve sustained growth and reduce failure risk through continuous investment in the skills, knowledge, and capabilities of their leadership (Rauch & Rijsdijk, 2011). This underscores the dynamic interplay between human capital and entrepreneurial action in the venture creation process, emphasizing that successful outcomes depend on the ability to harness internal competencies and external networks to overcome challenges and seize opportunities (Unger et al., 2009; Fertala, 2006).

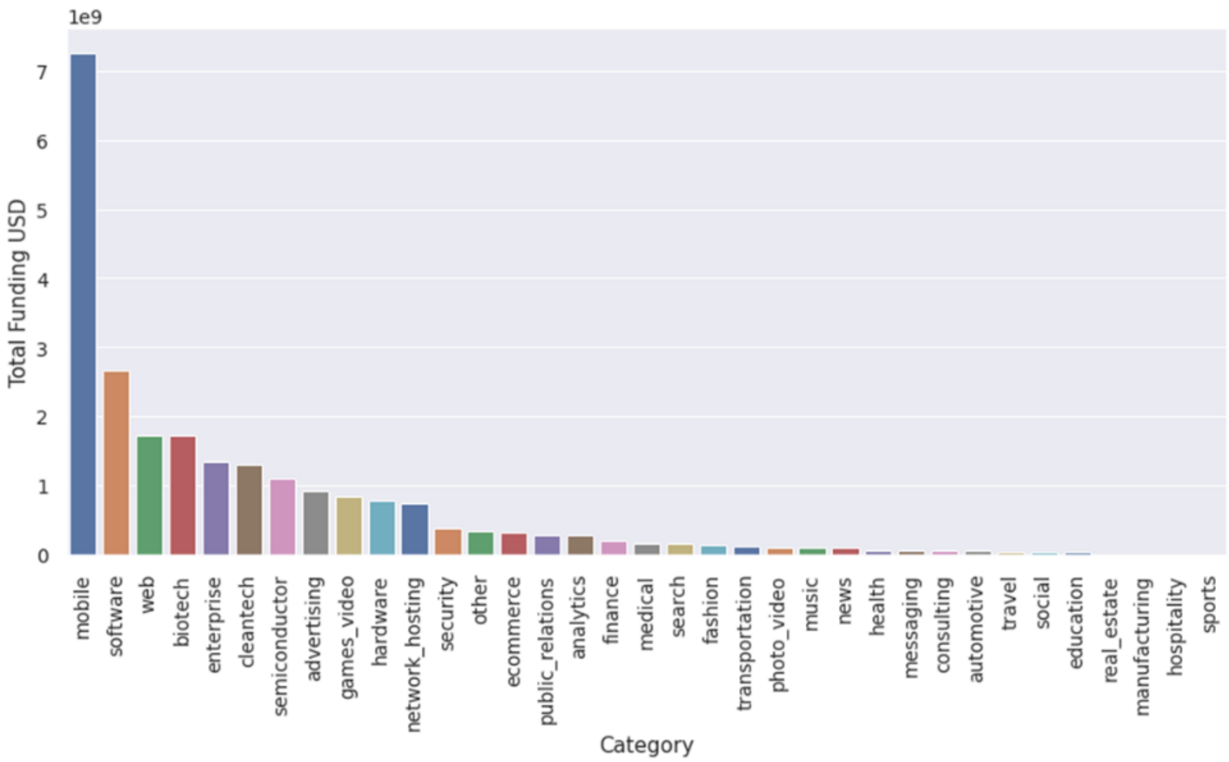

Figure 1. Depiction of Funding Distribution Across Sectors
Source: Author's own analysis.

Venture and seed funding are prevalent among startups at early stages; however, as firms advance toward later-stage funding rounds, particularly Series C and D, the number of qualifying companies diminishes significantly compared to those securing Series A and B investments. This sharp decline reveals a pyramidal structure within the venture capital landscape, where only a small subset of startups with demonstrable growth potential and strategic viability continue to attract investment. This trend aligns with the "selection hypothesis," which posits that only high-potential ventures with favorable trajectories are able to progress, while others are eliminated due to operational inefficiencies or external constraints (Sanchez, 2010; Bonini & Capizzi, 2019). Contributing to this attrition are increasingly stringent investment criteria and the inherent challenges of scaling in competitive environments. Data analysis further shows that acquired and currently operating companies are predominantly based in the United States, reflecting the maturity of the U.S. startup ecosystem in terms of venture creation, growth infrastructure, and acquisition pathways. Among these, acquired firms demonstrate higher mean and median funding levels and typically undergo more funding rounds than their closed or active counterparts, indicating the capital-intensive journey toward acquisition (Ghosh & Nanda, 2010). Founding year distributions span from 1902 to 2014, with a notable concentration around the early 2000s, a period marked by technological expansion and the dot-com boom, which fueled increased startup activity and venture capital deployment (Babouei, 2018).

The distribution of total funding is positively skewed, consistent with power law dynamics in startup ecosystems where a few high-performing ventures attract the majority of capital (Eisenmann, 2006). While aggressive capital deployment strategies are common, they do not necessarily ensure long-term returns and may lead to diminishing gains from marketing and expansion investments. These patterns support the "liability of newness" theory, which highlights how emerging ventures struggle with legitimacy, resource acquisition, and institutional acceptance in their formative years (Kerr & Nanda, 2008). Moreover, heightened venture capital (VC) activity correlates positively with innovation, as evidenced by increased patenting rates within supported industries , highlighting VC's critical role in catalyzing technological advancement and intellectual property creation (Kortum & Lerner, 1998). In addition to financial capital, venture capitalists contribute mentorship, governance expertise, and access to strategic networks, significantly enhancing the likelihood of startup survival and success. The evolving landscape also includes alternative funding models such as crowdfunding and angel syndicates, which have broadened capital access for startups that may fall outside traditional investment norms, promoting greater inclusivity within the entrepreneurial ecosystem (Foundation, 2016). Importantly, venture capitalists often shape startup trajectories by directly influencing human capital development. This includes appointing external leadership and steering strategic direction, occasionally resulting in the replacement of founding entrepreneurs, a reflection of their dual role as collaborators and controllers (Hellmann & Puri, 2002). Investments in general human capital have also been shown to expand the demand for venture capital by increasing the number of skilled employees seeking funding support (Sevilir, 2009). Thus, VCs act not just as financiers but as co-architects of startup vision and operations (Bocken, 2015). The U.S.'s dominance in startup acquisitions also reflects its robust M&A environment and the active role of corporate investors seeking strategic innovation opportunities (Gui, 2024). Historical trends suggest that the clustering of company foundings around 2000 was partly driven by speculative optimism, later corrected through market rationalization. Venture capital has long served as a cornerstone for both nascent and established enterprises in the United States (Levitsky, 1994) and is increasingly recognized globally as a key driver of innovation and economic development (Cumming & Cumming, 2010). As such, it remains a foundational force in shaping startup ecosystems and fostering technological and societal progress.

Drawing inferences from visualizations, such as bar charts, offers compelling insights into the dynamics of startup success. Notably, the data reveals that startups under five years of age account for approximately 64.62% of all successful ventures (Figure 3), reinforcing the critical role of a company's early years in shaping it's trajectory. This finding aligns with existing literature, which identifies the initial stages of a startup's life cycle as pivotal for developing a viable business model, acquiring customers, and establishing a competitive foothold (Santisteban & Mauricio, 2017). These formative years often represent a window of opportunity for adaptability, innovation, and strategic experimentation, all of which are essential for long-term sustainability. Furthermore, geographic analysis of success rates shows significant regional variation, with South America exhibiting a 100% success rate, followed by Europe (84.21%), Asia (73.33%), and North America (66.23%) (Figure 2). These regional disparities may reflect a complex interplay of market saturation, institutional support, cultural dynamics, and economic conditions. For example, the high success rate in South America could be attributed to emerging market potential, lower competitive pressure, or region-specific innovation niches, whereas North America's relatively lower success rate may be symptomatic of a saturated ecosystem and heightened competitive intensity. These patterns underscore the importance of considering both internal firm-level attributes and external environmental factors when evaluating startup potential. The local context and path dependencies such as regulatory environments, infrastructure, and ecosystem maturity play a crucial role in influencing startup outcomes (Fenwick & Vermeulen, 2016). A startup's ability to thrive depends not only on its internal strategy and capabilities but also on the broader regional and institutional frameworks that either support or constrain entrepreneurial growth (Sudaryana et al., 2024). Key ecosystem components such as access to incubators, proximity to customers, strategic location, and a collaborative business community are instrumental in supporting early-stage ventures (Bărbulescu & Constantin, 2019). Equally influential are the traits and experiences of the founding team, with research showing that entrepreneurial personality, motivation, innovation approach, and managerial acumen significantly impact the potential of incubated ventures (McCarthy et al., 2023; Wong et al., 2005).

Understanding why startups fail is as critical as identifying success factors. Research indicates that many failures occur not due to technological inadequacy but due to an inability to secure a customer base, particularly in the first two years of operation (Giardino et al., 2014). Therefore, strategic thinking, early customer identification, and value-driven innovation are vital (Muramalla &

Al-Hazza, 2019). In today's dynamic and increasingly digital business environment, startups must also capitalize on digital models and knowledge-intensive strategies to gain and sustain competitive advantage (Fatema et al., 2020). The early stages, marked by iterative product development and a sharp focus on user experience, often determine whether a venture can secure its place in the market (Szathmári et al., 2024; Zaheer et al., 2019).

The visual data confirms that startup success is intricately tied to age, location, and founder characteristics, which align with broader empirical and theoretical findings (Muramalla & Al-Hazza, 2019; Slávik et al., 2021). The fact that younger startups are more likely to succeed highlights the importance of agility and early-stage adaptability, while regional variations in success rates call attention to local market conditions and ecosystem maturity. The interplay between internal strengths-such as leadership, vision, and customer-centric strategies and external factors like geography and institutional support, necessitates a nuanced and context-specific approach to entrepreneurship. As emphasized by Chillakuri et al. (2020), entrepreneurial drive alone is insufficient; success also hinges on aligning internal competencies with external opportunities and constraints (Karani & Mshenga, 2021; Triono et al., 2022). These findings reinforce the need for a holistic understanding of startup ecosystems to better inform strategic decisions by founders, investors, and policymakers.

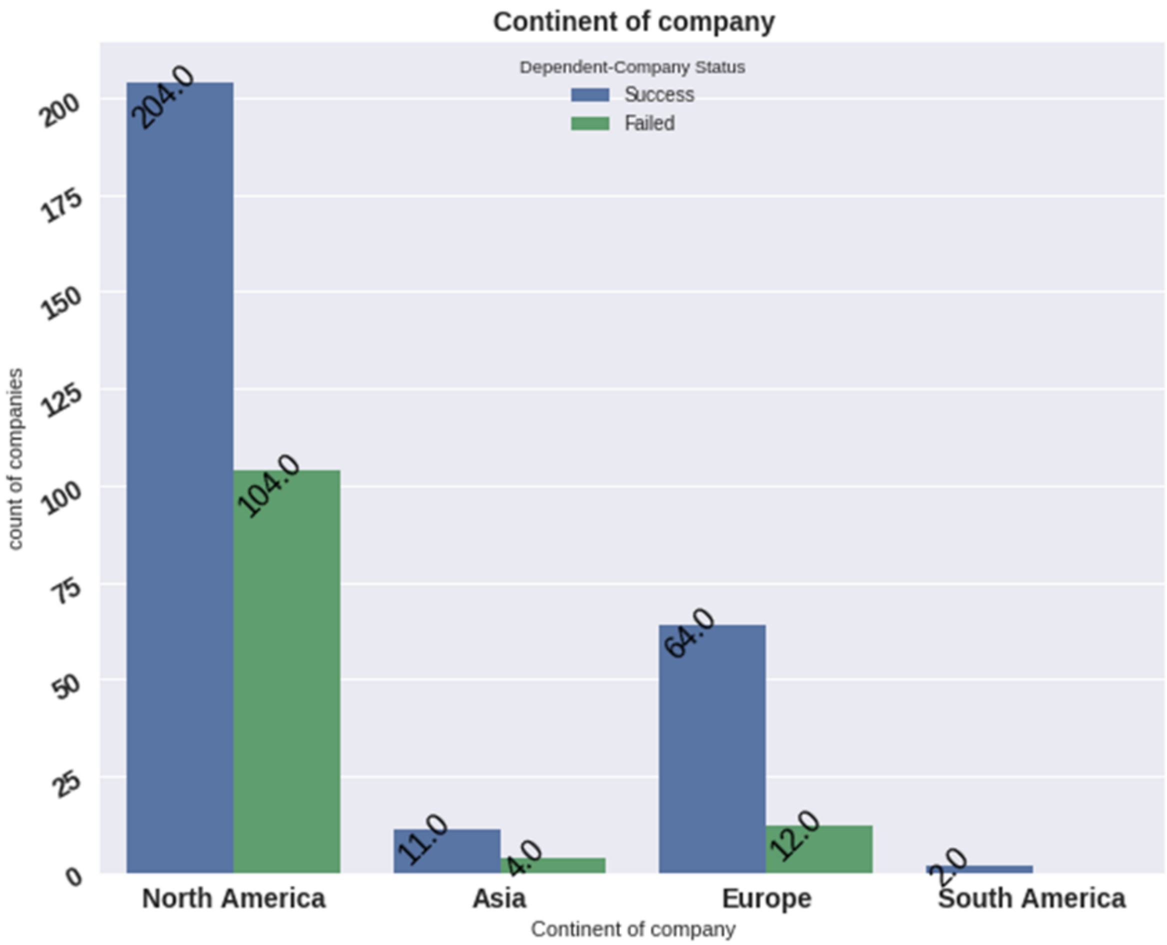

Figure 2. Bar graph illustrating the Number of Companies by Continent
Source: Author's own analysis.

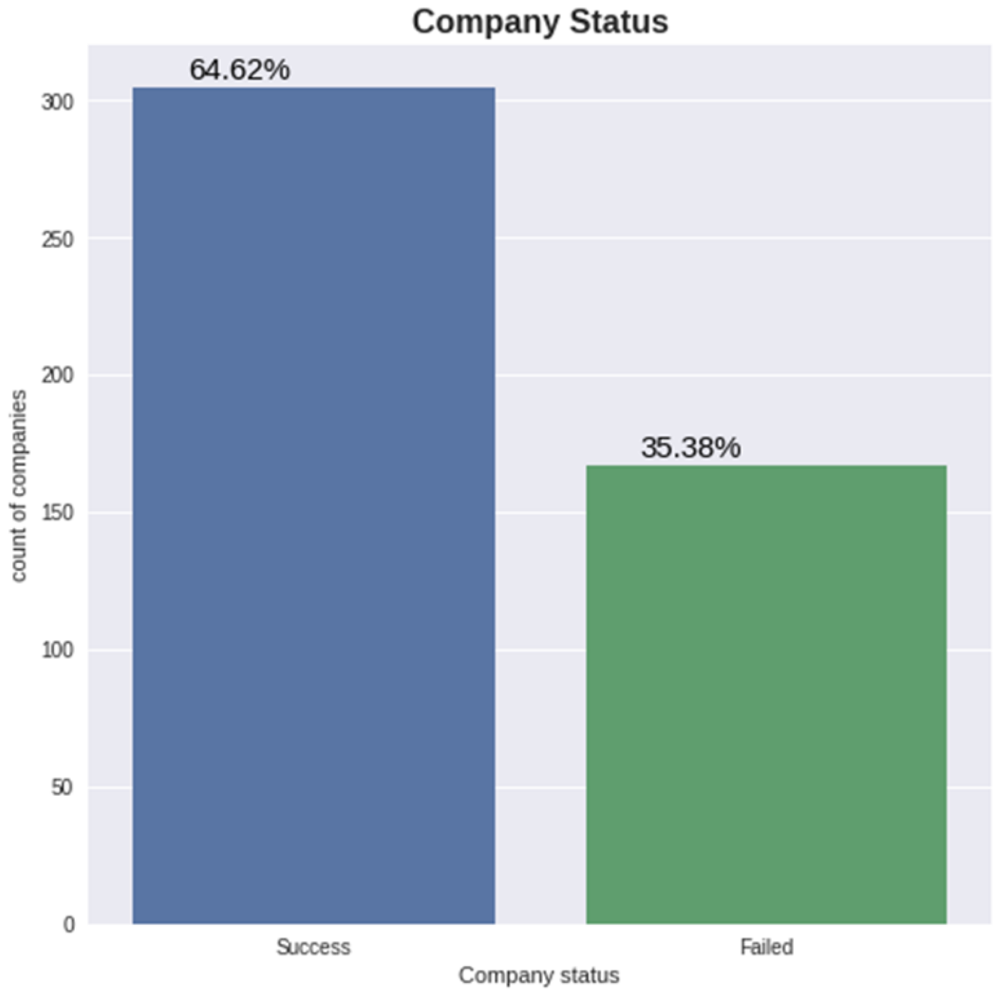

Figure 3. Bar graph illustrating the Status of Companies
Source: Author's own analysis.

The geographical and industrial distribution of startups offers key insights into the broader dynamics shaping global entrepreneurship. Analysis reveals that North America is home to 76.8% of startups, making it the dominant global hub for startup activity (Figure 4). This overwhelming concentration can be attributed to the region's mature venture capital infrastructure, robust intellectual property protection, innovation-friendly culture, and access to a highly skilled talent pool (Oliver et al., 2020; Cumming & Cumming, 2010). Europe follows distantly with 19%, reflecting growing support for entrepreneurship through government initiatives, R&D investments, and innovation networks. In contrast, India accounts for only 2.5% of startups, a gap that likely stems from structural challenges such as limited early-stage funding, infrastructural deficits, and regulatory barriers (Fenwick & Vermeulen, 2016). Despite India's rapidly expanding economy and tech-savvy population, these systemic issues restrict its startup potential relative to more established ecosystems. The success of startups is also closely tied to the industries in which they operate. Analysis shows that sectors such as analytics, commerce, marketing, advertising, and mobile have the highest proportion of successful startups (Figure 5). These domains are increasingly integral to the digital economy, with success driven by rising global demand for data-driven solutions, mobile-first platforms, and personalized marketing strategies (Sudaryana et al., 2024). Startups operating in these industries are well-positioned to leverage technological innovation and meet evolving customer expectations. This trend aligns with broader global shifts, where digital platforms dominate market capitalization and innovation is increasingly driven by the ability to harness big data and offer seamless digital experiences (Schneider, 2020). The regional disparity in startup activity not only highlights differences in entrepreneurial ecosystems but also underscores how ecosystem quality shapes entrepreneurial outcomes. North America's dominance is indicative of a well-established startup support system, including access to venture capital, policy support, and business incubators, all of which are essential for enabling innovation and reducing risk (Chillakuri et al., 2020; Bărbulescu & Constantin, 2019). In contrast, the limited startup activity in India, despite its population advantage and growing digital infrastructure, reflects the need for more conducive policy, infrastructure, and funding environments. As emphasized by Pukała (2019), competition within startup ecosystems drives innovation and necessitates adaptability, strategic thinking, and continuous development to secure market leadership. Startups operating in sectors like analytics and e-commerce benefit not only from rising demand but also from business models that are inherently scalable and technology-intensive, giving them a competitive edge in global markets.

Moreover, the success of startups is not only economic but also societal, as many ventures are focused on solving real-world problems and improving the quality of life (Santisteban & Mauricio, 2017). This highlights the importance of a customer-centric approach where the founder's ability to empathize with user needs, build strong relationships, and design tailored solutions is central to long-term success (Zaheer et al., 2019). At the heart of startup resilience is the entrepreneur's capacity for strategic learning, adaptability, and risk tolerance (Lee et al., 2023). Failures, when studied analytically, offer crucial lessons that shape future decisions and foster strategic thinking (Muramalla & Al-Hazza, 2019). Entrepreneurial orientation marked by innovation, proactiveness, and a willingness to take calculated risks has been found to be positively correlated with startup success, particularly in volatile and rapidly evolving markets.

Thus, the geographical concentration of startups in North America and the industry-level success seen in analytics, mobile, and digital commerce sectors reflect the interdependent relationship between ecosystems, market readiness, and technological capabilities. The findings underscore the necessity of aligning entrepreneurial resources, strategic vision, and supportive environments to foster startup success and innovation. For countries like India to improve their startup output and competitiveness, greater emphasis must be placed on removing infrastructural and regulatory barriers, fostering access to capital, and building stronger entrepreneurial networks (Karani & Mshenga, 2021; Triono et al., 2022). Ultimately, startup viability depends on continuous innovation, adaptability to external conditions, and the ability to create sustainable value within a dynamic global economy (Argaw & Liu, 2024; McCarthy et al., 2023; Szathmári et al., 2024).

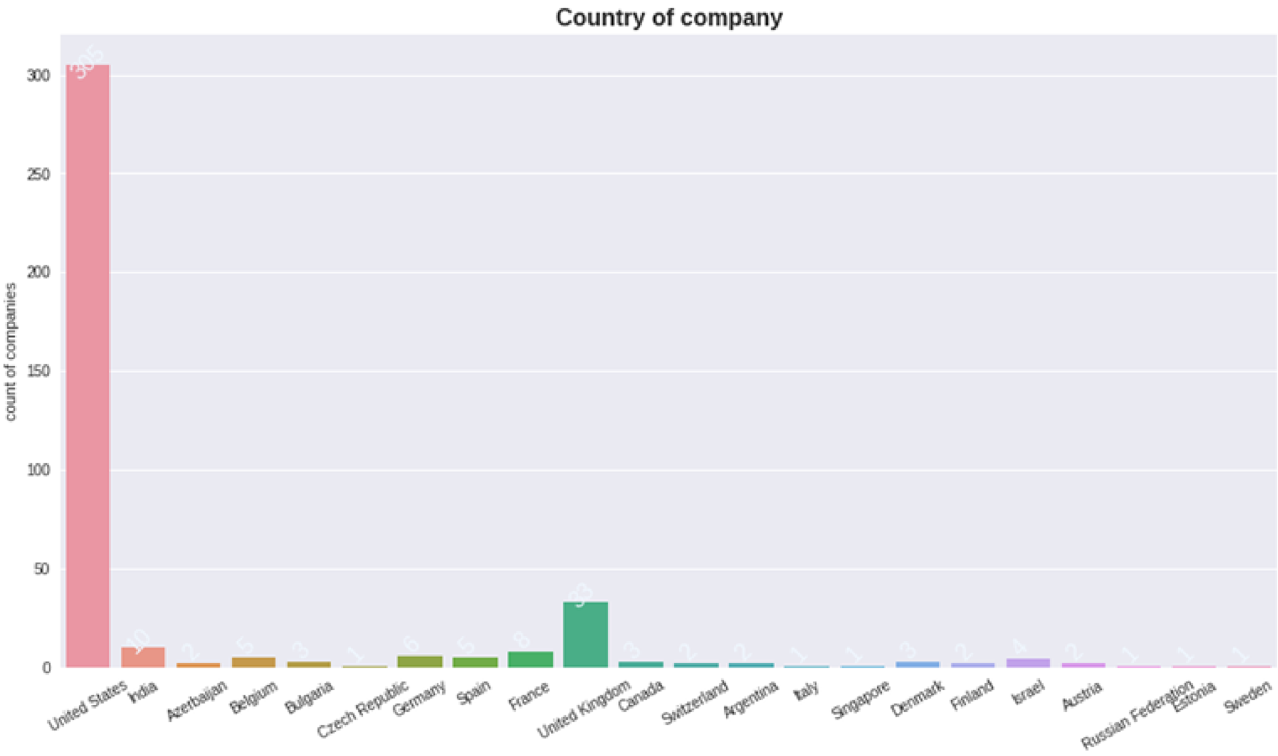

Figure 4. Examination of the number of companies by country
Source: Author's own analysis.

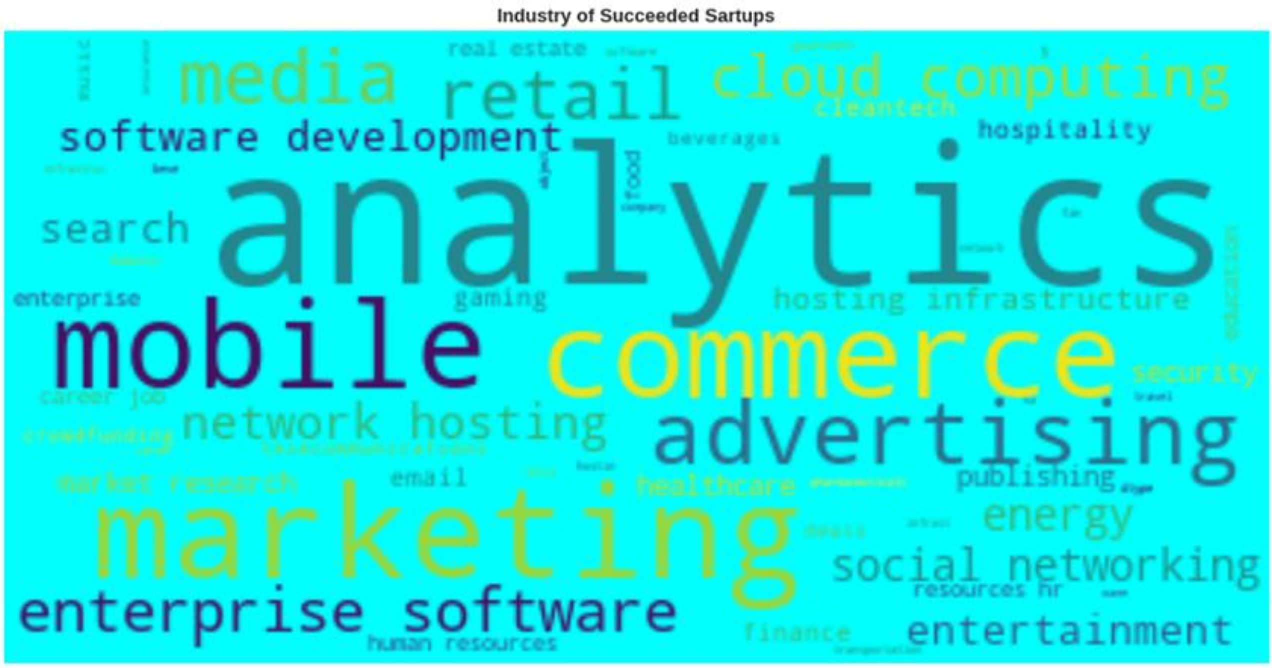

Figure 5. Word cloud depicting the industries of successful startups
Source: Author's own analysis.

To further examine the predictors of success, PCA was performed on 17 features after applying standard scaling, as illustrated in Figure 6. This was followed by the development of a Random Forest classification model, which achieved an accuracy of 0.93 on the test set and 0.82 on the training set. This high performance suggests the model's robustness in identifying key determinants of startup success (Figure 7; Table 1). Notably, The horizontal bar plot illustrates the relative importance of different features in predicting startup success, with each bar representing a feature and longer bars indicating greater significance. It is evident from the plot 7 that a few features contribute much more substantially to the model's predictions than others. Features 1 and 0 emerge as the most influential, each displaying an importance value around 0.20, suggesting a strong impact on the model's decision-making. Features 5 and 7 show moderate influence, while features such as 8, 6, 2, and 9 also contribute meaningfully, albeit to a lesser extent. In contrast, features like 3, 10, 13, 14, and 15 demonstrate minimal impact, as indicated by their very small importance values. When aligned with the accompanying analysis, it appears that critical determinants such as state code, latitude, city, name, and age at last funding year correspond to the top features identified, highlighting the pivotal role of geographical and temporal factors in shaping startup trajectories. These findings are consistent with earlier empirical studies (Argaw & Liu, 2024; Savin et al., 2022), which emphasize that location, access to regional capital ecosystems, and timing within the funding lifecycle are powerful contributors to startup viability.

The impact of founding team dynamics is also well-documented in entrepreneurship literature. Factors such as the CEO's integration within the team, shared entrepreneurial competencies, and prior founding experience significantly influence the likelihood of a successful startup launch (Leary & DeVaughn, 2009; Reese et al., 2020). In particular, teams composed of diverse skill sets frequently referred to as the "Hipster, Hacker, and Hustler" model are found to double the chances of startup success, emphasizing the importance of cross-functional collaboration (McCarthy et al., 2023). Entrepreneurial personality traits, such as risk-taking, innovativeness, and a proactive mindset, have consistently been associated with higher venture performance (Wong et al., 2005; Brandstätter, 2010). This aligns with broader psychological research on entrepreneurs, showing that they tend to exhibit greater risk tolerance, internal locus of control, and achievement motivation (Kerr et al., 2019). Moreover, founders with a high degree of self-efficacy are more likely to persevere through uncertainty and capitalize on emerging opportunities (Kerr et al., 2018). These findings also point to the importance of personality diversity and pre-launch planning within startup teams. Rather than conforming to a single ideal personality type, successful startups often feature complementary personalities across co-founders, suggesting that team composition and interpersonal dynamics are as crucial as individual brilliance (McCarthy et al., 2023). A statistically significant relationship has also been found between startup success and the level of strategic planning during the pre-startup phase, underscoring the role of foresight and preparation in venture performance (Khan, 2018). The startup phase is typically marked by resource constraints and market ambiguity, making traits like agility, vision, and customer-centricity particularly valuable (Szathmári et al., 2024).

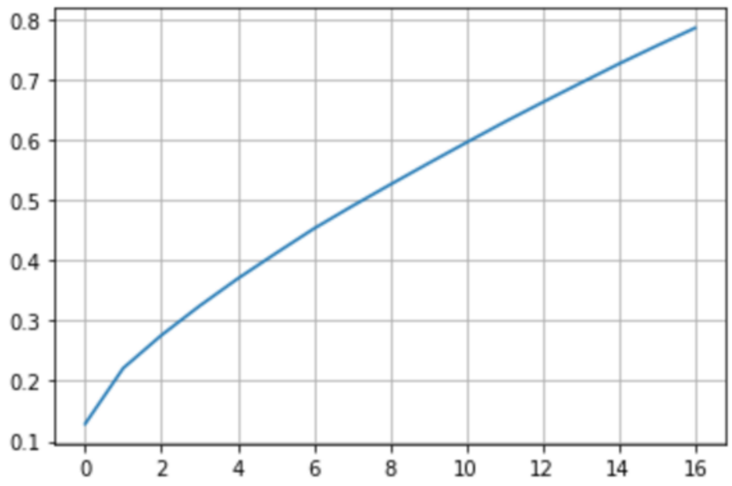

Figure 6. PCA analysis for n components=17
Source: Author's own analysis.

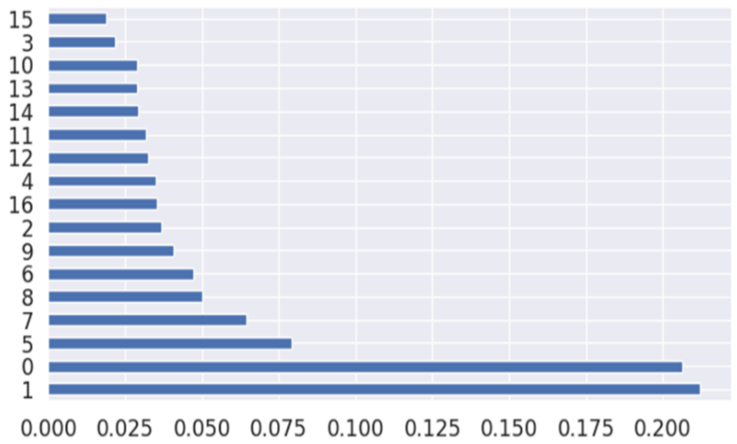

Figure 7. Graph illustrating the key features contributing to the accuracy of the model
Source: Author's own analysis.

Table 1. Train Test accuracy for the random forest model

Accuracy of Train Set	Accuracy of Test Set
0.815	0.939

Source: Author's own analysis.

Table 2. Confusion Matrix Results for the random forest model

	Precision	Recall	F1 Score	Support
0	0.81	0.93	0.87	177
1	0.84	0.61	0.71	100
Accuracy			0.82	277
Macro Average	0.82	0.77	0.79	277
Weighted Average	0.82	0.82	0.81	277

Source: Author's own analysis.

The evaluation of the model’s performance is based on key metrics: accuracy, precision, recall, and F1-score, each providing important insights into how well the model distinguishes between successful and unsuccessful startups.

Accuracy serves as one of the most intuitive evaluation metrics, addressing the fundamental question: “What proportion of the model’s predictions are correct?” It is calculated as the sum of true positives and true negatives divided by the total number of predictions. The model shows a training set accuracy of 81.5% and a significantly higher test set accuracy of 93.9%, suggesting that the model generalizes well to unseen data. The confusion matrix-based accuracy of 82% further aligns closely with the training set, reinforcing that the model’s predictive performance remains consistent across different datasets.

Precision focuses on the correctness of positive predictions. It measures how many of the startups predicted to succeed actually did succeed, calculated as true positives divided by the sum of true positives and false positives. The model exhibits a higher precision for successful startups (0.84) compared to unsuccessful ones (0.81), indicating that when it predicts a startup will succeed, the prediction is fairly reliable. However, a slightly lower precision for unsuccessful startups implies a presence of false positives, where some unsuccessful startups are incorrectly classified as successful.

Recall, also referred to as sensitivity or true positive rate, measures the model's ability to identify all actual positive cases, calculated as true positives divided by the sum of true positives and false negatives. The model achieves a high recall for unsuccessful startups (0.93), meaning it accurately identifies most startups that are likely to fail. Conversely, the recall for successful startups is significantly lower at 0.61, indicating that approximately four out of ten successful startups are being misclassified as failures. This poses a critical limitation, especially in contexts where identifying potential winners is important.

F1-Score is a harmonic mean of precision and recall and is particularly useful when there is an imbalance between classes. It helps balance the trade-off between precision and recall. The model’s F1-Score for unsuccessful startups is 0.87, higher than the F1-Score for successful startups (0.71). This imbalance reinforces the earlier observation that the model is more proficient at predicting failures than successes.

Further, when examining the macro and weighted averages:

The macro average treats both classes equally and reports a precision of 0.82, recall of 0.77, and F1-Score of 0.79. This slightly lower recall indicates that the model's overall ability to identify successful startups needs improvement. The weighted average, which accounts for class imbalance, reports similar precision (0.82), recall (0.82), and F1-Score (0.81), suggesting that the model’s overall performance is slightly inflated due to the majority class (unsuccessful startups) being easier to predict. The analysis of these performance metrics highlights both the strengths and weaknesses of the model. On the positive side, the model is highly effective at identifying unsuccessful startups, with a high recall and precision for the failure class. This makes it particularly valuable for investors or incubators that are risk-averse and aim to minimize investments in startups likely to fail. The high test set accuracy further strengthens confidence in the model's ability to generalize to new data. However, the model exhibits notable weaknesses when it comes to identifying successful startups. The lower recall for successful startups means that many promising ventures could be overlooked, potentially leading to missed investment opportunities. Even though the precision for successful startups is relatively high, the inability to identify all successful startups (as reflected in the recall and F1-score) limits the model's usefulness when the objective is to discover high-potential startups.

5. Managerial Implications

This study provides valuable insights for investors, marketers, entrepreneurs, and policymakers aiming to improve decision-making in the high-risk startup ecosystem. By identifying key success factors such as founder backgrounds, funding patterns, industry dynamics, and geographical considerations stakeholders can refine their strategies to optimize outcomes and mitigate risks (Zaheer et al., 2019; Muramalla & Al-Hazza, 2019). Investors, particularly those engaged in venture capital and crowdfunding platforms, can strategically allocate resources by using predictive models based on data mining and machine learning techniques. Understanding critical failure factors, such as product-market misalignment, marketing inefficiencies, and internal team dynamics (Bielialov, 2022), allows for more informed investment decisions, minimizing the likelihood of funding ventures that are prone to early-stage failure (Elsiefy & Foudah, 2015; Safari & Das, 2022). For marketers and industry analysts, the study highlights that sectors like software and biotechnology attract higher levels of funding, offering strategic advantages for targeted promotions and product positioning (McCarthy, Gong, Braesemann, et al., 2023). Moreover, analyzing startup success rates across continents such as South America's higher success rate compared to North America enables the integration of regional factors into investment and marketing strategies (Fenwick & Vermeulen, 2016). The findings also stress the importance of founders' educational and professional backgrounds. Practical experience and consulting exposure, rather than solely academic qualifications, were found to be stronger indicators of startup success (Cantamessa et al., 2018; Potanin et al., 2023). Investors and incubators can leverage this knowledge to better assess the viability of startups based on the human capital attributes of their founding teams. Additionally, understanding strategic motives like acquihiring where larger firms acquire startups primarily for their talent pool provides guidance to technology firms aiming to strengthen their human resources (Marita Makinen et al., 2014). For startups, this implies that cultivating a strong and versatile team can significantly enhance acquisition potential. Startups, given their dynamic nature, require continuous monitoring and strategic adjustments. Investors and marketers can use the developed prediction model and EDA insights to adapt promptly to industry trends and emerging opportunities (Pisoni et al., 2020). From a policy perspective, governments and entrepreneurship support institutions should focus on nurturing entrepreneurial ecosystems that provide access to funding, mentorship, and strategic resources. Considering the high failure rate of startups where approximately 80% fail within their first two years (Giardino et al., 2014) policymakers must prioritize interventions that address financial planning, team development, and market fit challenges (Fatema et al., 2020; Kartanaitė & Krušinskas, 2022). Moreover, recognizing that startup failures often result from internal team-related issues rather than purely external challenges (Brattström, 2019) reinforces the need for holistic startup evaluations. A multifaceted approach, addressing psychological factors, strategic planning, and realistic market assessment, is crucial to enhance startup resilience (Safari & Das, 2022; Gladwin et al., 1989).

Ultimately, the study's integration of human capital theory and machine learning into predictive modeling offers a powerful tool for improving startup success rates. By carefully evaluating factors such as founder experience, industry trends, and regional dynamics, stakeholders can foster more resilient startups, stimulate innovation, and contribute significantly to economic growth (Gupta & Rubalcaba, 2021; Skawińska & Zalewski, 2020). Building on these insights, Figure 8 illustrates a strategic framework designed to promote resilience and drive innovation in the startup ecosystem. It begins by emphasizing the importance of understanding startup dynamics through data mining and machine learning techniques, followed by identifying critical success factors using statistical models and predictive algorithms. Empowering stakeholders, including both venture capitalists and average investors, with accessible tools is highlighted as crucial for democratizing decision-making. The framework advocates for moving beyond traditional boundaries to support a diverse range of participants and identifies future research avenues to close knowledge gaps. Collaboration through industry-academic partnerships and knowledge-sharing initiatives is encouraged to drive richer insights. The roadmap stresses striving for sustainability by implementing research findings, promoting continuous advancement through learning and innovation, and fostering inclusive decision-making. Recognizing global considerations, the model encourages strategies tailored to different industries and regions. Ethical dimensions are integrated to ensure responsible innovation, culminating in the adoption of advanced predictive models to maintain a competitive edge in the dynamic startup environment. Overall, the model offers a holistic, forward-looking strategy for cultivating a vibrant and sustainable startup ecosystem.

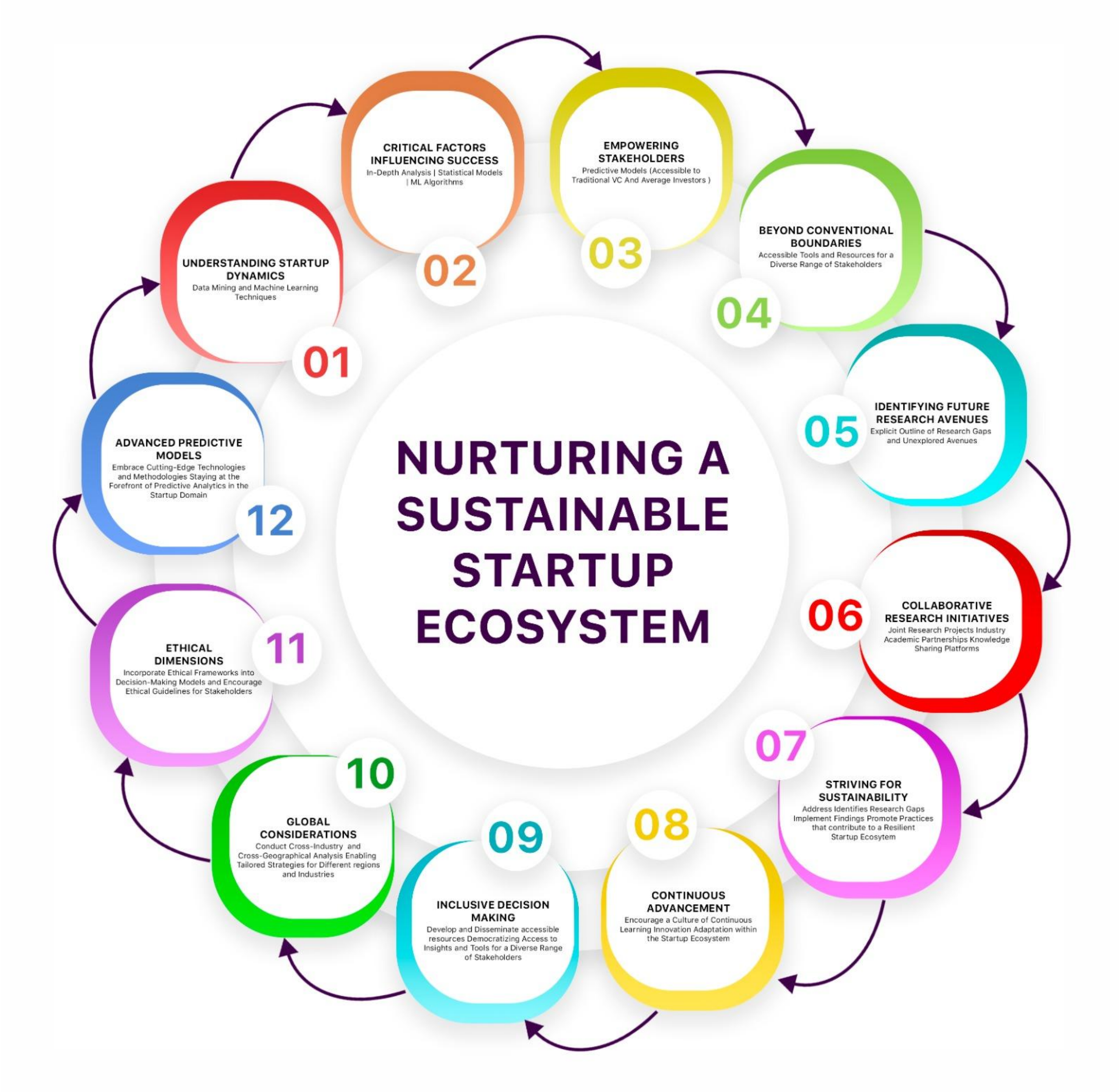

Figure 8. Blueprint for Cultivating a Resilient Startup Environment
Source: Author's own analysis.

6. Conclusion

This study provides a comprehensive exploration of startup success prediction by integrating human capital theory with advanced machine learning techniques. Recognizing the inherent uncertainty and high failure rates within startup ecosystems, the research underscores the necessity for innovative, data-driven approaches to enhance decision-making for investors, entrepreneurs, and policymakers. Drawing on a literature review, exploratory data analysis, and predictive modeling with Random Forest and PCA, this study identifies key success factors, including founder background, funding patterns, industry sectors, and regional dynamics. The findings reveal that practical experience and consulting exposure often outweigh formal academic qualifications in influencing startup outcomes, highlighting the critical role of applied human capital. Additionally, geographical factors, industry affiliations, and early-stage agility emerged as pivotal determinants of startup success. By systematically analyzing a real world Crunchbase dataset, the research demonstrates the efficacy of supervised machine learning models in capturing complex, non-linear relationships often overlooked by traditional statistical methods. Importantly, the study offers actionable implications for stakeholders: investors can refine resource allocation strategies; marketers can better target high-growth sectors; policymakers can design interventions that nurture entrepreneurial ecosystems; and incubators can tailor programs to bridge human capital gaps. The proposed strategic framework (Figure 8) further provides a roadmap for fostering a resilient and innovative startup environment by promoting continuous learning, ethical innovation, and inclusivity. However, the study is not without limitations. As noted, Crunchbase's focus on funded startups introduces a potential selection bias, as the sample may be skewed toward ventures that are more visible or already relatively successful. Consequently, the generalizability of the findings to early-stage or unfunded startups may be constrained. While PCA was employed to reduce noise and potential distortions arising from over-represented variables, fully eliminating selection bias remains a challenge. Future research should consider integrating data from alternative sources, such as AngelList, PitchBook, or regional startup incubator databases, to validate and generalize the predictive framework across diverse ecosystems.

Building on the findings of this study, several avenues for future research can be explored to enhance predictive accuracy and deepen the understanding of startup success dynamics. First, future studies could investigate the relationship between a startup's patent portfolio including the number, quality, and citation count of patents and its likelihood of achieving long-term success or acquisition. Patents often signal innovation capacity and competitive advantage, making them critical indicators worth integrating into predictive models. Second, applying time-series models such as ARIMA or LSTM could provide dynamic insights into how startups evolve across different funding stages, product iterations, or customer acquisition patterns over time. Modeling startup trajectories longitudinally would allow researchers to better capture growth patterns and early warning signs of success or failure. Third, the development of a real-time "Startup Health Score" by combining structured data (e.g., funding amounts, team size, location) and unstructured data (e.g., news mentions, employee reviews from platforms like Glassdoor, LinkedIn activity) presents a promising direction. Such a score could offer investors and stakeholders an up-to-date assessment of startup viability. Additionally, tracking the correlation between web analytics such as Google Trends activity, SimilarWeb traffic, or app store performance and startup funding rounds or exit events could reveal new leading indicators of market traction and investor interest.From a machine learning perspective, future research should address the class imbalance issue observed in startup datasets, where unsuccessful ventures typically outnumber successful ones. Techniques like SMOTE (Synthetic Minority Oversampling Technique), undersampling, or adjusting class weights in algorithms could help improve model sensitivity towards predicting successful outcomes. Moreover, exploring alternative algorithms such as Gradient Boosting methods (XGBoost, LightGBM, CatBoost) and SVM could offer better handling of complex and imbalanced data structures, potentially boosting predictive performance. Fine-tuning model hyperparameters using Grid Search or Random Search is recommended to optimize for metrics like Recall and F1-Score, especially for the successful startup class. This will help ensure the model's predictions are both sensitive and accurate. Finally, advanced feature engineering and selection techniques can be employed by incorporating novel variables such as founders' consulting experience, pivot timelines, geopolitical risk indices, team heterogeneity (including skill diversity and gender composition), and competency-based founder profiles derived from public resumes or LinkedIn data. These additions would enhance the dataset by capturing nuanced dimensions of founder experience, funding progression, and market characteristics, while simultaneously filtering out irrelevant or noisy features. To ensure the reliability and generalizability of results, Stratified K-fold Cross Validation should be employed, allowing consistent performance assessment across diverse data subsets. Pursuing these directions will enable the development of more robust, adaptive, and practically applicable predictive models that can better inform and support startup success across varied ecosystems.

References

Ahluwalia, S. and S. Kassicieh, (2024), "Pathways to Success: The Interplay of Industry and Venture Capital Clusters in Entrepreneurial Company Exits," *Journal of Risk and Financial Management*, 17(4), 159.

Allioui, H. and Y. Mourdi, (2023), "Unleashing the potential of AI: Investigating cutting-edge technologies that are transforming businesses," *International Journal of Computer Engineering and Data Science (IJCEDS)*, 3(2), 1-12.

Ashourizadeh, S., S. Rezaei, T. Schott, and J. Vang, (2014), "Entrepreneurs' human and social capital: direct and reinforcing benefits for export," *International Journal of Entrepreneurship and Small Business*, 21(2), 246-267.

Banerjee, P., T. Chattopadhyay and A.K. Chattopadhyay, (2024), "Investigation of the effect of bars on the properties of spiral galaxies: a multivariate statistical study," *Communications in Statistics-Simulation and Computation*, 53(3), 1216-1246.

Basole, R. C., H. Park, and C. D. Seuss, (2024), "Complex business ecosystem intelligence using AI-powered visual analytics," *Decision Support Systems*, 178, 114133.

Bielialov, T., (2022), "Risk management of startups of innovative products," *Journal of Risk and Financial Management*, 15(5), 202.

Bonini, S. and V. Capizzi, (2019), "The role of venture capital in the emerging entrepreneurial finance ecosystem: future threats and opportunities," *Venture Capital*, 21(2-3), 137-175.

Brattström, A., (2019), "Working with startups? these are the three things you ought to know about startup teams," *Technology Innovation Management Review*, 9(11).

Cantamessa, M., V. Gatteschi, G. Perboli, and M. Rosano, (2018), "Startups' roads to failure," *Sustainability*, 10(7), 2346.

Chang, S. J., (2004), "Venture capital financing, strategic alliances, and the initial public offerings of Internet startups," *Journal of Business Venturing*, 19(5), 721-741.

Chen, M. H. and Y. Y. Chang, (2013), "The impacts of human capital in enhancing new venture's performance: Competence, motivation and creativity," *Journal of Knowledge-based Innovation in China*, 5(2), 146-168.

D'Acquisto, G. and M. Naldi, (2019), "A conceptual framework for assessing anonymization-utility trade-offs based on principal component analysis," *arXiv preprint arXiv:1903*, 11700.

Eckart, L., S. Eckart, and M. Enke, (2021), "A brief comparative study of the potentialities and limitations of machine-learning algorithms and statistical techniques," In *E3S Web of Conferences* (Vol. 266, p. 02001), EDP Sciences.

El Hajj, M. and J. Hammoud, (2023), "Unveiling the influence of artificial intelligence and machine learning on financial markets: A comprehensive analysis of AI applications in trading, risk management, and financial operations," *Journal of Risk and Financial Management*, 16(10), 434.

Elsiefy, E. and M. L. Foudah, (2015), "Modelling the Behavior of the Venture Capital Industry-Evidence from Egypt," *Archives of Business Research*, 3(3).

Ermers, N. J., K. Hagoort, and F. E. Scheepers, (2020), "The predictive validity of machine learning models in the classification and treatment of major depressive disorder: State of the art and future directions," *Frontiers in Psychiatry*, 11, 472.

Fatema, K., M. M. Syeed, and M. S. U. Miah, (2020), "Demography of startup software companies: an empirical investigation on the success and failure," *International Journal of Computer Applications*, 975, 8887.

Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth, (1996), "From data mining to knowledge discovery in databases," *AI magazine*, 17(3), 37-37.

Fenwick, M. and E. P. Vermeulen, (2016), "Global Start-Up Communities," Lex Research Topics in Corporate Law & Economics Working Paper, 2016-9.

Fertala, N., (2007), "A study of immigrant entrepreneurship in Upper Bavaria," *International Journal of Entrepreneurship and Small Business*, 4(2), 179-206.

Fini, R. and L. Toschi, (2016), "Academic logic and corporate entrepreneurial intentions: A study of the interaction between cognitive and institutional factors in new firms," *International Small Business Journal*, 34(5), 637-659.

Forbes., (2024), "90% of startups fail-how to secure your place in the 10%."

Giardino, C., X. Wang, and P. Abrahamsson, (2014), "Why early-stage software startups fail: a behavioral framework," In *International conference of software business* (pp. 27-41), Cham: Springer International Publishing.

Gladwin, C. H., B. F. Long, E. M. Babb, L. J. Beaulieu, A. Moseley, D. Mulkey, and D. J. Zimet, (1989), "Rural entrepreneurship: One key to rural revitalization," *American Journal of Agricultural Economics*, 71(5), 1305-1314.

Graham, B. and K. Bonner, (2022), "One size fits all? Using machine learning to study heterogeneity and dominance in the determinants of early-stage entrepreneurship," *Journal of Business Research*, 152, 42-59.

Guo, B., Y. Lou, and D. Pérez-Castrillo, (2015), "Investment, duration, and exit strategies for corporate and independent venture capital-backed start-ups," *Journal of Economics & Management Strategy*, 24(2), 415-455.

Gupta, V. and L. Rubalcaba, (2021), "Competency-industry relatedness (C-IR) framework for sustained business growth in startups during and beyond pandemic: Myths and lessons from publicly funded innovative startups," *Sustainability*, 13(9), 4632.

He, Y., L. Li, D. Liu, and W. X. Zhou, (2025), "Huber principal component analysis for large-dimensional factor models," *Journal of Econometrics*, 249, 105993.

Hubert, M., P. J. Rousseeuw, and K. Vanden Branden, (2005), "ROBPCA: a new approach to robust principal component analysis," *Technometrics*, 47(1), 64-79.

Jeon, H. and S. Oh, (2020), "Hybrid-recursive feature elimination for efficient feature selection," *Applied Sciences*, 10(9), 3211.

Jolliffe, I. T., (2002), "Principal components as a small number of interpretable variables: some examples," In *Principal component analysis* (pp. 50-63), New York, NY: Springer New York.

Jolliffe, I. T., (1986), "Mathematical and statistical properties of sample principal components," In *Principal component analysis* (pp. 23-49), New York, NY: Springer New York.

Jordan, M. I. and T. M. Mitchell, (2015), "Machine learning: Trends, perspectives, and prospects," *Science*, 349(6245), 255-260.

Kartanaitė, I. and R. Krušinskas, (2022), "Financial efficiency of unicorns: regional and sector related aspects," *Engineering Economics*, 33(2), 200-214.

Krishna, S. and S. Sidharth, (2022), "HR analytics: Employee attrition analysis using random forest," *International Journal of Performability Engineering*, 18(4), 275.

Krishna, S. and S. Sidharth, (2023), "AI-powered workforce analytics: Maximizing Business and employee success through predictive attrition modelling," *International Journal of Performability Engineering*, 19(3), 203.

Leo, M., S. Sharma, and K. Maddulety, (2019), "Machine learning in banking risk management: A literature review," *Risks*, 7(1), 29.

Ma, S. and Y. Dai, (2011), "Principal component analysis based methods in bioinformatics studies," *Briefings in bioinformatics*, 12(6), 714-722.

Makinen, M., D. Haber, and A. Raymundo, (2012), "Acqui-hires for growth: planning for success," *Venture Capital Review*, 28, 31-42.

McCarthy, P. X., X. Gong, F. Braesemann, S. F. tephany, M. A. Rizoiu, and M. L. Kern, (2023), "The impact of founder personalities on startup success," *Scientific Reports*, 13(1), 17200.

McCoy, M. and J. A. Tropp, (2011), "Two proposals for robust PCA using semidefinite programming," 1123-1160.

Mosey, S. and M. Wright, (2007), "From human capital to social capital: A longitudinal study of technology–based academic entrepreneurs," *Entrepreneurship theory and practice*, 31(6), 909-935.

Muramalla, V. S. S. R. and A. M. Al-Hazza, (2019), "Entrepreneurial strategies and factors stimulate the business of tech startups," *International Journal of Financial Research*, 10(3).

Nacchia, M., F. Fruggiero, A. Lambiase, and K. Bruton, (2021), "A systematic mapping of the advancing use of machine learning techniques for predictive maintenance in the manufacturing sector," *Applied Sciences*, 11(6), 2546.

Nurhasanah, N. and D. Safitri, (2010), "Komponen Utama untuk Pengendalian Kualitas secara Statistik," *Media Statistika*, 3(1), 9-20.

Østergaard, A. and S. T. Marinova, (2018), "Human capital in the entrepreneurship ecosystem," *International Journal of Entrepreneurship and Small Business*, 35(3), 371-390.

Pankavich, S. and R. Swanson, (2015), "Principal component analysis: Resources for an essential application of linear algebra," *PRIMUS*, 25(5), 400-420.

Patel, P. C. and J. A. Pearce, (2018), "The survival consequences of intellectual property for retail ventures," *Journal of Retailing and Consumer Services*, 43, 77-84.

Perez, D. J. D. and J. G. Guevarra, (2020), "Entrepreneurial competencies and readiness of extension service beneficiaries of a state university," *Philippine Social Science Journal*, 3(3), 107-118.

Pisoni, A., E. Aversa, and A. Onetti, (2021), "The role of failure in the entrepreneurial process: a systematic," *International Journal of Business and Management*, 16(1).

Podosinnikova, A., S. Setzer, and M. Hein, (2014), "Robust PCA: Optimization of the robust reconstruction error over the Stiefel manifold," In *German Conference on Pattern Recognition* (pp. 121-131), Cham: Springer International Publishing.

Potanin, M., A. Chertok, K. Zorin, and C. Shtabtsovsky, (2023), "Startup success prediction and VC portfolio simulation using CrunchBase data," *arXiv preprint arXiv*, 2309, 15552.

Ringnér, M., (2008), "What is principal component analysis?" *Nature biotechnology*, 26(3), 303-304.

Rollmann, I., N. Gebhardt, S. Stahl-Toyota, J. Simon, M. Sutcliffe, H. C. Friederich, A. Safari, and A. Das, (2023), "Entrepreneurial failure analysis using quality management approaches," *Total Quality Management & Business Excellence*, 34(1-2), 235-260.

Sainani, K. L., (2014), "Introduction to principal components analysis," *PM&R*, 6(3), 275-278.

Schmidt, J. and S. Heidenreich, (2018), "The role of human capital for entrepreneurial decision-making-investigating experience, skills and knowledge as antecedents to effectuation and causation," *International Journal of Entrepreneurial Venturing*, 10(3), 287-311.

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

Smith, L. I., (2002), "A tutorial on principal components analysis."

Szathmári, E., Z. Varga, A. Molnár, G. Németh, Z. P. Szabó, and O. E. Kiss, (2024), "Why do startups fail? A core competency deficit model," *Frontiers in psychology*, 15, 1299135.

Thomas Asha E. and M. Bhasi (2018), "Investment in Higher Education Sector of India: A Review of Related Literature and Preliminary Investigation," *International Journal of Management Studies*, 5(2), 47-62.

Tokuda, T., J. Yoshimoto, Y. Shimizu, G. Okada, M. Takamura, Y. Okamoto, and K. Doya, (2018), "Identification of depression subtypes and relevant brain regions using a data-driven approach," *Scientific reports*, 8(1), 14082.

Unger, J. M., A. Rauch, M. Frese, and N. Rosenbusch, (2011), "Human capital and entrepreneurial success: A meta-analytical review," *Journal of business venturing*, 26(3), 341-358.

Wang, Y., Y. Cheng, H. Liu, Q. Guo, C. Dai, M. Zhao, and D. Liu, (2023), "A review on applications of artificial intelligence in wastewater treatment," *Sustainability*, 15(18), 13557.

Wang, Y., C. H. Tsai, D. D. Lin, O. Enkhbuyant, and J. Cai, (2019), "Effects of human, relational, and psychological capitals on new venture performance," *Frontiers in Psychology*, 10, 1071.

Wiriyathammabhum, P. and B. Kijsirikul, (2012), "Robust principal component analysis using statistical estimators," *arXiv preprint arXiv*, 1207, 0403.

Zaheer, H., Y. Breyer, J. Dumay, and M. Enjeti, (2019). "Straight from the horse's mouth: Founders' perspectives on achieving 'traction'in digital start-ups," *Computers in Human Behavior*, 95, 262-274.

Zhang, Y., S. Xu, L. Zhang, and M. Yang, (2021), "Big data and human resource management research: An integrative review and new directions for future research," *Journal of Business Research*, 133, 34-50.

Zhou, Z., X. Li, J. Wright, E. Candes, and Y. Ma, (2010), "Stable principal component pursuit," In *2010 IEEE international symposium on information theory* (pp. 1518-1522), IEEE.

Appendix

Description of Variables Used in the Study

Table 3. Outlining: Basic Company Information

Variable	Description
Company_Name	Name of the startup company.
Short Description of company profile	Brief overview of the company's business model and focus.
Industry of company	Industry or sector the startup operates in (e.g., Mobile, Software, Biotech).
Focus functions of company	Key functional areas the company focuses on (e.g., analytics, marketing, product development).
Country of company	Country where the startup is headquartered.
Continent of company	Derived continent based on the country.
Online or offline venture	Indicates if the startup operates online, offline, or both.
B2C or B2B venture	Business-to-Consumer (B2C) or Business-to-Business (B2B) classification.

Table 4. Outlining: Founders and Team Information

Variable	Description
Number of Co-founders	Total number of co-founders.
Number of of advisors	Total number of advisors associated with the startup.
Average Years of experience for founder and co-founder	Average professional experience (in years) of the founding team.
Exposure across the globe	Whether founders have global experience (Yes/No).
Breadth of experience across verticals	Experience across multiple industry verticals.
Highest education	Highest education level of founders (e.g., Bachelor's, Master's, Ph.D.).
Years of education	Total years of formal education completed.
Specialization of highest education	Field of study specialization (e.g., Engineering, Business, Marketing).
Relevance of education to venture	Whether education is related to the startup's domain (Yes/No).
Relevance of experience to venture	Whether prior experience is relevant to the current business.
Degree from a Tier 1 or Tier 2 university?	Educational background from prestigious institutions (Yes/No).
Worked in top companies	Experience in reputed firms (Yes/No).
Have been part of startups in the past?	Previous startup experience (Yes/No).
Have been part of successful startups in the past?	Experience in successful startups (Yes/No).
Consulting experience?	Experience in consulting roles (Yes/No).
Was he or she partner in Big 5 consulting?	Past partnership experience in major consulting firms (Yes/No).
Experience in Fortune 100 / 500 / 1000 organizations	Prior work experience in top global companies.
Top management similarity	Alignment in top leadership experience (Yes/No).

Table 5. Outlining: Company Performance and Growth Metrics

Variable	Description
Employee Count	Number of employees at the company.
Employees count MoM change	Month-on-Month change in employee count (growth/shrinkage).
Has the team size grown	Indicates whether the team has grown recently (Yes/No).
Team size Senior leadership	Size of the senior management team.
Team size all employees	Total number of employees.
Employees per year of company existence	Average number of employees added per year.
Last Funding Amount	Amount raised in the latest funding round (in USD).
Last Funding Date	Date when the most recent funding was secured.
Funding Duration	Time duration between first and last funding rounds.
Internet Activity Score	Level of online presence and visibility.
google page rank of company website	Search engine visibility of the company website.

Table 6. Outlining: Financial and Funding Variables

Variable	Description
Dependent-Company Status	Final outcome of the company: Acquired, Operating, or Closed.
Number of Investors in Seed	Number of investors participating at the seed stage.
Number of Investors in Angel and or VC	Number of investors involved at angel and venture stages.
Presence of a top angel or venture fund in previous round of investment	Whether a top VC or angel investor backed the company (Yes/No).
Number of of repeat investors	Number of investors who participated in multiple funding rounds.
Crowdfunding based business	Whether the startup raised funds via crowdfunding platforms.
Invested through global incubation competitions?	Indicates funding from global startup competitions.
Last round of funding received (in million USD)	Last funding round amount.
Time to 1st investment (in months)	Time taken (in months) from founding to first investment.
Avg time to investment	Average time gap between funding rounds.

Table 7. Outlining: Business Model and Strategy Variables

Variable	Description
Product or service company?	Indicates if the company offers a product, service, or both.
Cloud or platform based service/product?	Delivery model: cloud-based or platform service (Yes/No).
Local or global player	Focus on local markets or global expansion.
Linear or Non-linear business model	Revenue model: Linear (traditional) or Non-linear (scalable, tech-driven).
Capital intensive business	High capital requirements like e-commerce or engineering (Yes/No).
Crowdsourcing based business	Business model based on crowdsourced solutions (Yes/No).
Owns data or not? (monetization of data)	Whether the startup owns monetizable data assets (Yes/No).
Is the company an aggregator/marketplace?	Marketplace-based business model (Yes/No).
Hyper localisation	Focus on hyper-local markets (Yes/No).
Subscription based business	Whether revenue is subscription-based (Yes/No).
Pricing Strategy	Nature of startup pricing (e.g., premium, freemium, cost-leader).

Table 8. Outlining: Technology Focus Variables

Variable	Description
Machine Learning based business	Startup leverages machine learning technologies (Yes/No).
Predictive Analytics business	Startup offers predictive analytics solutions (Yes/No).
Prescriptive analytics business	Business involving prescriptive analytics (Yes/No).
Speech analytics business	Products related to speech data analysis (Yes/No).
Big Data Business	Focus on big data technologies (Yes/No).
Technical proficiencies to analyse and interpret unstructured data	Capability in handling unstructured data (Yes/No).
Solutions offered	Nature of technological solutions provided.
Disruptiveness of technology	Level of technology innovation and market disruption.
Gartner hype cycle stage	Position on Gartner's technology maturity cycle.
Time to maturity of technology (in years)	Estimated time for the technology to mature.

Table 9. Outlining: Founder and Team Skills Variables

Variable	Description
Skills score	Aggregate score of founder and team skills.
Team Composition score	Score reflecting team balance across functional areas.
Percent_skill_Entrepreneurship	Percentage weight of entrepreneurship skills.
Percent_skill_Operations	Percentage weight of operations management skills.
Percent_skill_Engineering	Percentage weight of technical/engineering skills.
Percent_skill_Marketing	Percentage weight of marketing skills.
Percent_skill_Leadership	Percentage weight of leadership qualities.
Percent_skill_Data Science	Percentage weight of data science skills.
Percent_skill_Business Strategy	Percentage weight of strategic thinking skills.
Percent_skill_Product Management	Percentage weight of product management skills.
Percent_skill_Sales	Percentage weight of sales expertise.
Percent_skill_Domain	Percentage weight of domain-specific knowledge.
Percent_skill_Law	Percentage weight of legal expertise.
Percent_skill_Consulting	Percentage weight of consulting skills.
Percent_skill_Finance	Percentage weight of financial management skills.
Percent_skill_Investment	Percentage weight of investment expertise.

Table 10. Outlining: Risk and Reputation Variables

Variable	Description
Controversial history of founder or co-founder	Any negative legal or public incidents related to founders (Yes/No).
Legal risk and intellectual property	Presence of legal challenges or IP-related issues.
Client Reputation	Market reputation among clients.
Company awards	Awards and recognitions received by the company.
Renown score	Reputation score of the founder/co-founder.
Number of Recognitions for Founders and Co-founders	Total awards or honors received by founders.
Number of of Research publications	Academic and professional research contributions.