Predicting the Success of Early-Stage African Startups Using Machine Learning

Maureen Wambui Ndung'u

133482

Submitted in partial fulfillment of the requirements for the Degree of
Bachelor of Business Science in Financial Engineering at Strathmore University

Strathmore Institute of Mathematical Sciences
Strathmore University
Nairobi, Kenya

January 2025

DECLARATION

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the Research Project contains no material previously published or written by another person except where due reference is made in the Research Project itself.

Maureen Wambui Ndung'u [*Name of Candidate*]

_____ [*Signature*]

31st January 2025 [*Date*]

This Research Project has been submitted for examination with my approval as the Supervisor.

Edwin Adoyo Obonyo [*Name of Supervisor*]

_____[*Signature*]

31st January 2025 [*Date*]

Strathmore Institute of Mathematical Sciences

Strathmore University

ABSTRACT

Africa's share of global venture funding is estimated to be around 1%; meaning that only a very small portion of worldwide venture capital investment goes towards African startups. This presents a challenge for entrepreneurs, investors, and policymakers seeking to foster innovation and economic growth. This study aims to bridge this gap by leveraging machine learning models to predict the success of African startups based on key factors: business operating status, number of funding rounds, and business age. Unlike prior research, which has predominantly focused on Western markets and defined success through acquisitions or IPOs, this study specifically examines African startups, addressing the continent’s unique entrepreneurial landscape.

The research utilizes CrunchBase data spanning from 2000 to 2024, encompassing 28,851 startups, applying three machine learning models—Logistic Regression, Support Vector Machines, and Random Forest—to evaluate startup success. The dataset was split into training and validation sets, ensuring robust model performance assessment. Results indicate an exceptionally high accuracy of 99-100%, with strong sensitivity but lower specificity, highlighting potential dataset imbalance. Despite this, the machine learning models outperform traditional probability-based approaches by capturing non-linear relationships and complex interactions between startup success factors. This provides a more nuanced and data-driven approach to early-stage business evaluation compared to simplistic probabilistic models.

The findings offer practical implications for investors by enabling more informed decision-making, for entrepreneurs by identifying key success drivers, and for policymakers by informing strategies that enhance startup ecosystems in Africa. Future work should focus on balancing the dataset, incorporating additional predictive features, and expanding testing to ensure greater generalizability. This study contributes to the growing body of research on startup success prediction, offering a tailored approach for the African market and providing valuable tools for practitioners in the entrepreneurial and investment space.

ACKNOWLEDGMENT

This research would not have been possible without the support and guidance of many individuals. I would like to express my deepest gratitude to my supervisor, Dr. Edwin Obonyo, for his invaluable insights, encouragement, and constructive feedback throughout this project. I am also grateful to Dr. Marian Chatoro for her support with Chapters 4 and 5. A special thank you goes to Matilda Bosire for her expertise and assistance in reviewing the machine learning methodology.

Special thanks to my family and friends for their unwavering support and belief in my abilities. To my professors at Strathmore University and my peers, who inspired me with their commitment to academic excellence, I am truly grateful. Finally, I acknowledge the use of Crunchbase as a comprehensive data source, which significantly contributed to the depth and rigor of this study. This project is a testament to the collective efforts of all who supported me, and I am deeply appreciative of their contributions.

Table of Contents

Preliminary

List of Figures

List of Tables

Chapter 1: Introduction

1.1 Background

The success of early-stage startups has been a subject of extensive study, highlighting their importance to economic growth, innovation, and employment generation (Mehmeti & Musabelli, 2024). These studies span various geographical contexts, including Europe, where research has shown that early-stage companies are crucial drivers of the economy (Mehmeti & Musabelli, 2024). In Asia, the importance of startups has been recognized, particularly in the technology sector, where they contribute significantly to the digital economies (Misra, Jat, & Mishra, 2021). Similarly, in North America, research underscores the role of startups in fostering innovation and competition within markets (Mehmeti & Musabelli, 2024). In the Middle East, the government's support for startups has been linked to the rapid growth of tech-based ventures, contributing to the region's global economic standing (Żbikowski & Antosiuk, 2021). Common themes in these studies include the critical role of innovation, the importance of access to capital, the challenges of market entry, and the significant impact of government policies on the success of early-stage companies (Vasquez, Santisteban, & Mauricio, 2023).

Machine learning has increasingly been employed to predict the success of startups across various countries, leveraging a range of algorithms and data sources to enhance predictive accuracy. Ensemble models, which combine multiple machine learning algorithms to improve prediction accuracy, have demonstrated significant effectiveness in forecasting startup success. For instance, studies by Ross, Das, Sciro, & Raza (2021) utilized ensemble methods to merge data from Crunchbase and patent databases. These models often outperform individual algorithms by aggregating their strengths and mitigating their weaknesses. The use of ensemble techniques such as Random Forest, Gradient Boosting, and eXtreme Gradient Boosting (XGBoost) has shown high accuracy rates and robust performance in various studies (Arroyo, Corea, Jimenez-Diaz, & Recio-Garcia, 2019; Krishna, Agrawal, & Choudhary, 2016; Ünal & Ceasu, 2019; Corea, Bertinetti, & Cervellati, 2021; Bangdiwala, Mehta, Agrawal, & Ghane, 2022). Key outcomes include improved prediction accuracy and enhanced ability to handle complex, high-dimensional data.

Hybrid intelligence methods, which integrate human expertise with machine learning models, address the complexities and uncertainties inherent in startup predictions. Dellermann et al. (2017) explored these methods by combining collective human judgments with machine learning algorithms. This approach allows for the incorporation of intuitive insights from

experts alongside analytical rigor provided by algorithms. The results highlighted that hybrid models could effectively capture nuanced patterns that purely algorithmic or human-based approaches might miss (Dellermann, Lipusch, Ebel, Popp, & Leimeister, 2017). This integration has led to better identification of success factors and more informed decision-making for investors. Deep learning models have been employed to analyze large and complex datasets, such as those from Crunchbase and Kaggle (Ferrati, Chen, & Muffatto, 2021; Potanin, Chertok, Zorin, & Shtabtsovsk, 2023). These models, including neural networks and their variants, have shown promise in predicting startup success. For example, Ferrati et al. (2021) developed a deep learning model with high recall rates, indicating its effectiveness in identifying successful startups. Deep learning's ability to handle unstructured data and learn from intricate patterns has contributed to more precise predictions and deeper insights into startup performance.

The application of machine learning models has yielded several key insights. Models such as Random Forest and XGBoost have demonstrated impressive accuracy and precision, with some studies reporting precision levels exceeding 90% (e.g., Bangdiwala et al., 2022), indicating their effectiveness in reliably predicting startup success. Additionally, hybrid and ensemble models have shown a superior ability to handle uncertainty and complexity by integrating various sources of information and analytical approaches, offering a more detailed assessment of potential success. Furthermore, machine learning has proven crucial in identifying critical success factors, with features such as funding stages, company age, and market trends emerging as significant predictors. For instance, Ünal and Ceasu (2019) highlighted that XGBoost and Random Forest prioritized funding details and company age as important features in their analyses.

The growing application of machine learning in startup success prediction underscores its transformative potential. By utilizing high-quality data sources and advanced analytical techniques, machine learning offers a powerful tool for enhancing predictive accuracy and informing investment decisions. The integration of various models and methodologies continues to advance the field, promising even more refined and actionable insights for stakeholders in the startup ecosystem.

Startup success is a multidimensional concept, often defined subjectively depending on the perspective of the founder, venture capitalists, or investors (Baskoro, Prabowo, Meyliana, & Gaol, 2022). Researchers have identified various criteria to define this success. Some measure

success by whether a startup is acquired (Cholil, et al., 2024) or has issued an IPO (initial public offering) or attained unicorn status (Potanin, Chertok, Zorin, & Shtabtsovsk, 2023). Others assess it based on the startup's operational status, acquisition, or IPO issuance (Ünal & Ceasu, 2019). Additionally, success can be defined by achieving an IPO or undergoing a merger and acquisition (M&A) (Gangwani & Zhu, 2024; Thirupathi, Alhanai, & Ghassemi, 2022; Bangdiwala, Mehta, Agrawal, & Ghane, 2022). Securing Series A funding is also considered a milestone of success (Te, et al., 2022; Dellermann, Lipusch, Ebel, Popp, & Leimeister, 2017; Sharchilev, et al., 2018), as is profitability (Tomy & Pardede, 2018; Vasquez, Santisteban, & Mauricio, 2023). Alternatively, repeated financing rounds can indicate a startup's success (Piskunova, Ligonenko, Klochko, Frolova, & Bilyk, 2021; Arroyo, Corea, Jimenez-Diaz, & Recio-Garcia, 2019).

There is a growing need to apply machine learning to predict the success of early-stage African businesses, given the unique challenges and opportunities present in the continent (McKenzie & Sansone, 2019). African startups often face distinct obstacles, such as limited access to capital, infrastructural challenges, and diverse market conditions, which necessitate a tailored approach to predicting their success (African Scalecraft, n.d.). The use of machine learning in this context could help identify patterns and factors that are specific to African startups, thereby enabling more accurate predictions and better decision-making for investors and entrepreneurs. Moreover, the integration of localized data sources and contextual knowledge with machine learning algorithms could provide insights that are more relevant to the African market, ultimately contributing to the growth and sustainability of startups in the region (Gichohi, 2023).

African startups operate in a context marked by a youthful demographic, rapid urbanization, a growing middle class, and increasing mobile phone penetration, which collectively offer a fertile ground for entrepreneurial activities (United Nations, 2021; United Nations, 2023). There are notable trends in the African entrepreneurship space, with several emerging tech-enabled companies significantly shaping the landscape. The significance of digital technology in the global economy renders it a strategic sector, both economically and politically (Smart Africa, 2020). For example, in 2023, the total capitalization of Google, Apple, Microsoft, Meta, Amazon, and NVIDIA is more than 3 times the total GDP of the entire African continent (Smart Africa, 2020; STATISTA, 2024). This highlights the significant impact of ICT on the global

economy and underscores its role as a powerful driver of economic and social development in developing countries (Smart Africa, 2020). In recent years, technology has become a central focus in Africa's private capital investment sector, fundamentally altering traditional investment paradigms and industry trends (AVCA, 2024).

Another notable trend is the increasing interest of international investors in African startups. Global venture capital firms, development finance institutions, and corporate investors are recognizing the potential of Africa's innovation ecosystem and are making significant investments (AVCA, 2024). Local investors are also playing a crucial role in supporting African startups. African venture capital firms, angel investors, and corporate venture arms are providing not only funding but also mentorship and strategic support. Initiatives such as the African Business Angels Network (ABAN) and various startup incubators and accelerators are fostering a vibrant entrepreneurial ecosystem by connecting startups with investors and resources.

The exit environment for African startups has been difficult, with a 48% decrease in exits in 2023 compared to the previous year (AVCA, 2024). Mergers and acquisitions (M&A) are becoming an increasingly prevalent exit strategy for these startups (EAVCA, 2024; AVCA, 2024). The proportion of sales to private equity buyers rose to 33% in 2023 from 23% in 2022 (AVCA, 2024). This strategy is growing in popularity among fund managers aiming to expand their platform countries. Although initial public offerings (IPOs) remain relatively rare in Africa, they represent another potential exit route. Strengthening capital markets and regulatory frameworks could make IPOs a more attractive option for startups. Jumia's successful IPO on the New York Stock Exchange in 2019 has set a precedent, showing that African startups can garner significant global investor interest.

Investment and entrepreneurial activity vary significantly across different regions of Africa (AVCA, 2024). Southern Africa, with South Africa at its core, continues to be a major hub for startups and investment (AVCA, 2023). West Africa, home to vibrant startup ecosystems in countries like Nigeria and Ghana, has traditionally attracted significant venture capital investment (AVCA, 2023). North Africa, which includes countries like Egypt, Morocco, and Algeria, has also seen growing entrepreneurial activity (AVCA, 2024). East Africa, with Kenya as a leading hub, continues to attract investment and foster innovation (EAVCA, 2023).

The unique environment of African startups is shaped by a complex interplay of socio-economic and political factors (Ajayi-Nifise, Tula, Asuzu, Mhlongo, & Ibeh, 2024), along with

emerging trends in technology and investment (AVCA, 2024). Understanding these dynamics is crucial for predicting the success of early-stage businesses in Africa. Despite the challenges, the resilience and adaptability of African entrepreneurs, coupled with increasing technological advancements and a youthful demographic, present promising opportunities for growth and innovation in the continent's startup ecosystem (United Nations, 2021; Raj, 2023). By leveraging machine learning approaches, stakeholders can better predict and enhance the success of African startups, driving economic development and fostering sustainable growth.

1.2 Problem Statement

Understanding the location of a company is crucial for predicting startup success, as entrepreneurial ecosystems vary significantly across different regions (Ferrati & Muffatto, 2020). Most existing research on startup success has predominantly focused on developed economies such as North America, Europe, and Asia, often neglecting emerging markets like Africa (Azeem & Khanna, 2023). For instance, the study by Ünal and Ceasu (2019) on using machine learning to predict startup success, excluded data from Africa and Oceania due to zero or near-zero variance, effectively overlooking these regions' unique dynamics and potential. This geographical bias in research not only diminishes the relevance of predictive models for African startups but also misses out on valuable insights specific to the African entrepreneurial ecosystem. By incorporating African data into predictive models, we can develop a more nuanced understanding of local success factors, which is essential for accurate predictions and effective support for startups in the region. Failing to address this gap risks perpetuating a cycle of underinvestment and missed opportunities in Africa's burgeoning startup sector.

Moreover, the global venture capital (VC) market has been seen to continue underinvesting in the African VC ecosystem compared to other regions, as noted by Truman (2023). This underinvestment is further exacerbated by the tendency of American venture capital and private equity to predominantly fund white foreign founders, leaving African entrepreneurs at a disadvantage (The Guardian, 2020; Data Driven VC, 2024). The reliance on traditional investment methods, which often lack inclusivity and fail to address biases, limits the growth potential of African startups. Machine learning models present an opportunity to counteract these biases by providing data-driven insights that can encourage more equitable investment practices. By leveraging machine learning, we can create more inclusive and accurate predictive models that better reflect the realities of the African startup landscape. Without these advancements, Africa risks continued underrepresentation and inequality in global venture capital, stifling innovation and hindering economic development.

For African startups, this lack of tailored predictive models exacerbates challenges such as limited access to capital and support, which are critical for scaling businesses in emerging markets (BIC Africa, 2021; Azeem & Khanna, 2023). Furthermore, studies reveal that without localized data, predictive models may continue to perpetuate biases, leading to unequal investment opportunities and hindering the growth of the African venture capital ecosystem (Turman, 2023; Ganesan, Mahalingam, Nathan, Ware, & Weinberg, 2023). Machine learning

models, when adapted to include African-specific data, can mitigate these issues by offering more accurate predictions and promoting fairer investment practices. This is not merely a technological update but a necessary step towards fostering a more equitable entrepreneurial environment that can drive sustainable development and innovation across the continent (Data Driven VC, 2024).

1.3 Research Objectives

1.3.1 General Objective

The main objective of this study is to develop a model that predicts the success of early-stage business in Africa using supervised machine learning.

1.3.2 Specific Objectives

1. To identify and analyze key success factors for early-stage African startups.
2. To develop a supervised machine learning model that predicts the success of these businesses based on identified factors.

1.4 Research Questions

1. What are the critical success factors for early-stage African startups?
2. How can these success factors be quantified, weighted, and incorporated into a supervised machine learning predictive model?

1.5 Justification of Study

Investing in startups presents inherent risks, particularly in emerging markets like Africa, where the entrepreneurial landscape is both dynamic and under-researched. Despite the continent's rich reservoir of innovation and entrepreneurial talent, many African startups struggle to secure the necessary funding to scale their ventures. This difficulty often stems from the uncertainty surrounding their potential for success and the lack of predictive tools tailored to the unique characteristics of the African market. The absence of reliable predictive models exacerbates the risk for investors, who may face substantial losses from investing in ventures that fail to thrive. This uncertainty hampers the growth of promising startups and stifles the broader economic development of the region.

To address this challenge, the development of a robust predictive tool specifically designed for African startups is essential. By analyzing data from startups established between 2000 and 2023 across all 54 African countries and diverse industries, this study aims to create a predictive model that captures the distinctive factors influencing startup success in Africa. Such a model

would provide valuable insights into which startups are likely to succeed, thereby reducing the risk of investment and enhancing the confidence of investors.

Accurate prediction of startup success is crucial not only for mitigating investment risks but also for fostering a more vibrant and inclusive entrepreneurial ecosystem. With more precise and tailored predictions, investors, including angel investors and venture capitalists, will be better positioned to support high-potential startups. This increased confidence can lead to greater investment in early-stage ventures, providing African entrepreneurs with the financial backing they need to scale their businesses and contribute to economic growth. As a result, this research not only benefits investors and entrepreneurs but also contributes to the broader economic and social advancement of Africa.

1.6 Significance Of Study

The research on predicting the success of early-stage African businesses using a machine learning model holds substantial significance across multiple domains. This study aims to bridge the gap in understanding the unique success factors that drive the growth and sustainability of startups in the African context. By identifying these factors and developing a robust predictive model, the research will provide insights and tools that can benefit various stakeholders, including African entrepreneurs and startups, investors and venture capitalists, policymakers and economic planners, academia and researchers, and business support organizations and incubators.

This paper will be immensely beneficial for African entrepreneurs and startups. By understanding the critical success factors identified through this study, entrepreneurs can make informed decisions, adopt best practices, and strategically plan their business operations to enhance their chances of success. The predictive model developed will serve as a tool for self-assessment, enabling startups to evaluate their potential for success based on historical data and identified success factors.

Investors and venture capitalists will gain a deeper understanding of the elements that contribute to the success of early-stage African businesses. The predictive model will provide a data-driven approach to assessing the viability and potential of investment opportunities. This can lead to more informed investment decisions, reduced risk, and optimized allocation of resources. By identifying promising startups, investors can better support innovation and economic growth in Africa.

Policymakers and economic planners will benefit from insights into the success factors of African startups, which can inform the development of policies and programs that foster entrepreneurship growth. The research findings can guide the creation of supportive regulatory frameworks, financial incentives, and infrastructure development initiatives. Ultimately, this can lead to a more conducive environment for business development, economic diversification, and job creation.

The academic community and researchers will find this study valuable as it contributes to the body of knowledge on entrepreneurship and SME success in the African context. The research methodology, findings, and predictive model can serve as a foundation for further studies, facilitating scholarly discourse and the advancement of research in this field. Additionally, the study can be incorporated into academic curricula, enhancing the education of future entrepreneurs and business leaders.

Business support organizations and incubators play a crucial role in nurturing startups. The insights from this research will enable these organizations to tailor their support services, mentorship programs, and resources to address the specific needs of African entrepreneurs. The predictive model can be used as a diagnostic tool to identify areas where startups require assistance, thereby enhancing the effectiveness of incubation programs and increasing the overall success rate of supported businesses.

This research has the potential to make a significant impact on the African entrepreneurial ecosystem by providing actionable insights, reducing investment risks, informing policy, advancing academic research, and strengthening support structures for startups. The development and validation of a machine learning model to predict business success will not only empower individual entrepreneurs but also contribute to the broader economic development of the African continent.

1.7 Scope Of The Study

This study will examine African startups established between 2000 and 2024. By examining startups founded during this 24-year period, the study will capture the evolution of the African entrepreneurial landscape, from its early days of mobile technology adoption (GSMA, 2023) to the present era of digital transformation (World Bank Group, 2024) and increased global investment (AVCA, 2024). This broad timeframe allows for the identification of key factors that have influenced startup success over the years, providing valuable insights into both historical and contemporary dynamics in the African context.

The research will encompass all industries to ensure a broad understanding of startup success factors across various sectors. This inclusive approach aims to identify common success factors as well as industry-specific dynamics within the African entrepreneurial ecosystem.

Provided that a startup is based in Africa, it will be considered in the analysis. This approach ensures a comprehensive examination of the startup landscape across the entire continent, capturing the diverse entrepreneurial environments and regional variations within Africa.

Chapter 2: Literature Review

2.1 Startup Success Factors

The success of startups is influenced by multiple factors that determine their ability to survive, scale, and sustain a competitive edge. This section examines key success factors highlighted in recent literature, including product-market fit, financing, headquarters location, and team composition.

2.1.1 Product-Market Fit

Product-market fit is a fundamental determinant of a startup's success, representing the alignment between a product's offerings and market demands. Initially popularized by Andreessen in the early 2000s, this concept has evolved, with recent studies emphasizing the iterative process required to achieve and maintain this fit. Meijer (2019) argues that startups must engage in a continuous cycle of building, measuring, and learning to adapt their products to market needs. The lean startup methodology promotes launching a "minimum viable product" (MVP), a simplified version aimed at testing market demand with minimal effort and cost (Ries, 2011; Meijer, 2019; Maurya, 2016). This approach enables startups to gather critical customer feedback and refine their product to better achieve product-market fit (PMF) (Dennehy, Kasraian, O'Raghallaigh, & Conboy, 2016).

The Lean Startup Process underscores the importance of connecting deeply with the target audience, converting customer insights into actionable strategies. Maurya (2016) emphasizes that continuous customer feedback is essential for refining a product to meet evolving market needs. This iterative collaboration between the company and its customers accelerates the refinement of the MVP, ultimately enhancing its alignment with PMF. The timing of product-market fit is also crucial (Gross, 2015). Introducing a product too early can result in a mismatch between the product's capabilities and market needs, while late entry can lead to missed opportunities and market saturation (Gurbuz, 2018). Ahmad et al. (2024) highlight the role of agile methodologies in enabling startups to iterate and pivot rapidly, thereby increasing the chances of achieving timely product-market fit. This agility allows startups to capitalize on emerging opportunities while avoiding strategies that may no longer be viable.

Kartika (2024) expands on the concept by examining the role of product innovation and scalability in achieving product-market fit. While innovation is essential for differentiating a product in a crowded market (Faster Capital, 2024), scalability ensures that the product can

meet growing demand without compromising quality (Spacenco & Mandari, 2020). However, achieving product-market fit is not without challenges. Studies suggest that an overemphasis on innovation can lead to products that are too advanced for current market conditions (Pampillo, 2023; True Digital, 2017), while excessive focus on market demands can stifle innovation. Startups must carefully navigate these trade-offs to achieve and sustain product-market fit.

2.1.2 Financing

Access to adequate financing is another critical factor that influences startup success. Literature consistently emphasizes the importance of securing sufficient capital to support growth, attract talent, and scale operations. Kaplan and Lerner (2016) find that startups with greater access to funding are more likely to survive and achieve significant milestones, such as subsequent funding rounds or exits through acquisitions or IPOs. Startups secure funding from a variety of sources, such as angel investors, venture capital (both traditional and corporate), crowdfunding, friends and family, bootstrapping, grants, and debt financing (Janaji, Ibrahim, & Ismail, 2021).

Marullo, Casprini, Di Minin, and Piccaluga (2018) assert that access to venture capital significantly enhances a startup's chances of success. Venture capital-backed startups are more likely to achieve critical milestones, such as expanding into new markets or launching new products, due to the combination of financial resources and strategic guidance provided by venture capital firms (Zeng, 2023). However, excessive dependence on venture capital can cause founders to lose control, as investors may advocate for aggressive growth strategies that conflict with the startup's long-term goals (Sulillari, 2023; Stripe, 2024). Therefore, maintaining a balance between securing necessary funding and preserving strategic autonomy is crucial (LinkedIn Community, 2023).

Crowdfunding has also emerged as a viable financing option, offering both capital and market validation (Cornelius & Gokpinar, 2021). Mollick and Robb (2016) found that successful crowdfunding campaigns provide necessary funds and generate early customer engagement, which can be critical for refining products and achieving product-market fit. However, they caution that the pressure to deliver on promises made during the campaign can strain a startup's resources.

Sauvage, Zeisberger, and Varadan (2022) suggest that startups carefully evaluate whether a Corporate Venture Capital (CVC) fund aligns with their strategic goals, as CVCs offer unique benefits and risks compared to traditional venture capital and angel investors. Financial

management also plays a crucial role in sustaining a startup's growth. Ampong (2024) emphasizes the importance of closely monitoring cash flow, managing burn rate, and making strategic investments that align with long-term goals. Effective resource management, including strategic allocation of human and technological resources, is essential for maximizing output (Mahmudur, 2023; Symeonidou, Leiponen, Autio, & Bruneel, 2022).

2.1.3 Headquarters Location

The geographical location of a startup's headquarters significantly influences its access to critical resources, including capital, talent, and markets (Guzman, 2018). Startups located in established entrepreneurial ecosystems, such as Silicon Valley or London, often enjoy distinct advantages (Guzman & Stern, 2015; Ahluwalia & Kassicie, 2024). Research shows that location choice is relevant for entrepreneurship, as proximity to venture capital firms and skilled labor pools can facilitate easier access to funding, networking opportunities, and mentorship, all of which are crucial for early-stage growth (Yu & Artz, 2019; Stam, 2015; Díaz-Santamaría & Bulchand-Gidumal, 2021).

Geographical factors, such as infrastructure and resources, also play a significant role in shaping venture capital activities. A well-established infrastructure-comprising advanced transportation systems, modern communication technologies-and a favourable business climate, promotes entrepreneurial growth and attracts venture capital investments (Zeng, 2023). Areas with prestigious universities and research institutions often attract top talent, boosting the chances that startups in these regions will secure venture capital funding due to the abundance of highly skilled human resources (Zeng, 2023; Kézaia & Skalac, 2024). Conversely, startups in regions with less developed ecosystems may face challenges in accessing these resources (Nims, 2023).

The impact of location on startup success also extends to regulatory environments. Zeng (2023) emphasizes that sound laws and regulations provide a secure foundation for business operations, ensuring smooth growth by minimizing legal uncertainties and reducing regulatory obstacles. Additionally, cultural fit can influence a startup's operations and success. Bojadjiev, Mileva, Misoska, and Vaneva (2023) demonstrate that startups aligning with local cultural norms are more likely to gain traction in those markets, affecting everything from marketing strategies to product development. Hemmert et al. (2019) further note that variations in entrepreneurship across countries are shaped by market conditions and cultural values, which differ based on the entrepreneurial ecosystem.

2.1.4 Team Composition

The composition of the founding team is another critical factor in startup success. The literature suggests that the skills, experience, and diversity of the founding team are pivotal in determining a startup's ability to navigate challenges and seize opportunities. Mol (2019) argues that a successful startup team requires more than prior experience and industry-specific skills; shared entrepreneurial passion and a collective strategic vision are equally important. While experience enhances decision-making, alignment in vision and passion drives team performance. Mol (2019) finds that teams with high levels of experience but lacking in passion and vision tend to underperform in areas such as innovation and customer satisfaction, whereas teams with strong alignment in soft skills perform significantly better.

D'Acunto, Tate, and Yang (2019) argue that diverse founding teams—those with members from various backgrounds, industries, and skill sets—are more likely to succeed. Diversity enhances creativity and problem-solving abilities, particularly in the early stages of a startup when innovation and adaptability are crucial. The experience of the founding team also significantly impacts a startup's chances of success. Mol (2019) highlights that experienced entrepreneurs are better equipped to make strategic decisions, avoid common pitfalls, and build resilient businesses, as they are more likely to recognize patterns and trends that inform critical business decisions.

However, team dynamics can present challenges. Conflicts within the founding team can hinder a startup's progress. Faster Capital (2024) notes that misalignments in vision, strategy, or decision-making can lead to disputes that distract from the startup's objectives and erode team cohesion. Effective communication, clear role definitions, and shared goals are essential for maintaining a productive team dynamic. Access to influential networks and mentors is another critical aspect of team composition. Daradkeha and Mansoora (2023) argue that startups with strong networks can access valuable resources, insights, and opportunities unavailable to less connected teams. Mentorship provides guidance and support that helps startups overcome challenges and accelerate growth (Zeng, 2023). Kabatunzi (2022) adds that founders with a strong personal brand, characterized by leadership, vision, resourcefulness, and resilience, are more likely to lead successful businesses (Mol, 2019; Elsafty, Abadir, & Shaarawy, 2020; Indrianti, Sasmoko, Abdinagoro, & Rahim, 2024).

2.1.5 Business Strategy

The effectiveness of a startup's business strategy is often shaped by various external and internal influences. According to Teece, (2018), strategic decisions in startups are significantly

influenced by market conditions, technological advancements, and regulatory environments. Startups that can effectively adapt their strategies in response to these influences are better positioned to capitalize on emerging opportunities and mitigate risks. For instance, the ability to pivot—a concept popularized by Ries (2011) in *The Lean Startup*—allows startups to change their business models or product offerings in response to market feedback, thereby enhancing their chances of success. Moreover, Bradley, Hirt, & Smit (2018) highlight that startups with a clear understanding of their competitive landscape are more likely to develop strategies that differentiate them from competitors, thus securing a competitive edge.

Marketing and distribution are pivotal components of a startup's business strategy, directly impacting customer acquisition, retention, and overall market positioning. Research by Chaffey & Ellis-Chadwick (2022) suggests that startups must prioritize digital marketing strategies to reach broader audiences and create stronger brand recognition. The use of data analytics and customer insights enables startups to tailor their marketing efforts, thereby increasing the effectiveness of their campaigns and improving return on investment (ROI). Additionally, startups that invest in omnichannel distribution strategies, integrating both online and offline channels, are more likely to succeed in today's highly competitive markets (Kotler, Kartajaya, & Setiawan., 2017). Furthermore, Nagle & Müller (2018) emphasize the importance of value-based pricing strategies in aligning product offerings with customer perceptions of value, which can enhance profitability and customer satisfaction. In line with this, Ries (2011) argues that a startup's marketing strategy should be closely aligned with its product development process to ensure that the product-market fit is achieved early on, which is essential for sustained success.

The ultimate goal of a startup's business strategy is to create value for stakeholders while securing a competitive advantage in the market. According to Porter & Heppelmann (2018), startups can achieve this by leveraging innovative technologies and business models that disrupt traditional industries. The capacity to continuously innovate and adjust to evolving market conditions is vital for sustaining a competitive advantage (Kaniawati, Sukma, & Oktaviani, 2024). Moreover, Teece (2018) argues that startups should focus on building sustainable business models that not only generate immediate profits but also ensure long-term viability. This involves optimizing operational efficiency, managing supply chains effectively, and investing in technology infrastructure to scale the business. For instance, startups that adopt

lean operations and agile methodologies are better equipped to respond to market changes and customer needs, thereby enhancing their operational efficiency (Womack & Jones, 2015). In terms of branding and positioning, startups that successfully differentiate themselves from competitors through unique value propositions and strong brand identities are more likely to achieve market dominance (Keller & Swaminathan, 2020). This is particularly important in highly competitive industries where brand loyalty can be a significant driver of growth.

2.2 Predicting The Success of Businesses Using Machine Learning

Over the past decade, numerous machine learning models have been developed and applied to the prediction of startup success. These models typically leverage large datasets such as Crunchbase, CB Insights, and other similar repositories to predict whether a startup will succeed in terms of acquisition, repeated funding rounds, IPOs, or profitability. This section reviews eight common machine learning models used for this purpose, with a specific focus on the top three that will be used in this study: logistic regression, random forest and support vector machines (SVM). Each model is discussed in terms of its advantages, disadvantages, and overall performance in the prediction of startup success.

2.2.1 Logistic Regression

Logistic Regression is a fundamental classification algorithm used to predict binary outcomes, such as success or failure (James, Hastie, Witten, & Tibshirani, 2021). It operates by modeling the probability of a class through a logistic function, also referred to as the sigmoid function, which produces values ranging from 0 to 1 (Pan, Gao, & Luo, 2018). This model is both simple and interpretable, offering a clear understanding of how features contribute to the predicted probability (James, Hastie, Witten, & Tibshirani, 2021). Its computational efficiency is another advantage, making it a suitable choice for straightforward classification problems. However, logistic regression assumes a linear relationship between the features and the log-odds of the outcome, which may restrict its ability to capture more complex patterns (James, Hastie, Witten, & Tibshirani, 2021). Additionally, it can be sensitive to outliers and may perform poorly if the data does not meet its assumptions.

Table 1: Past Research on Predicting Startup Success Using Logistic Regression

Authors	Data Source	Definition of success	Machine Learning Model	Accuracy	Sensitivity	F1 Score
Krishna et al. (2016)	Crunchbase	Acquired	Logistic Regression			
Dellermann, Lipusch, Ebel, Popp, & Leimeister (2017)	Crunchbase, Mattermark, and Dealroom	Series A funding	Logistic Regression	not aim of study		
Pan, Gao, & Luo (2018)	Crunchbase	M&A or IPO	Logistic Regression	0.7254		0.442
Shah & Mcgaugh (2019)	Crunchbase	Acquired	Logistic Regression	0.859	0.76	
Piskunova, Ligonenko, Klochko, Frolova, & Bilyk, 2021	Ukrainian Dealroom	Repeated Funding rounds	Logistic Regression	0.6	0.45	0.486
Żbikowski & Antosiuk (2021)	Crunchbase and Web-based information	Operating with Series B financing, acquired or IPO	Logistic Regression	0.86	0.21	0.33
Bangdiwala, Mehta, Agrawal, & Ghane (2022)	Crunchbase	IPO or M&A	Logistic Regression	0.925		

2.2.2 Random Forest

Random Forest are an ensemble learning technique that builds multiple decision trees and merges their outputs to produce a final prediction (Piskunova, Ligonenko, Klochko, Frolova, & Bilyk, 2021). By using bootstrapping and feature randomness, Random Forest create diverse trees that collectively improve predictive accuracy and robustness (James, Hastie, Witten, & Tibshirani, 2021). This approach is less likely to overfit compared to individual decision trees and can manage both classification and regression tasks (James, Hastie, Witten, & Tibshirani, 2021; Krishna, Agrawal, & Choudhary, 2016). However, the complexity of Random Forest reduces their interpretability compared to single trees and demands more computational resources and memory (Cholil, et al., 2024). Training can also be slower, especially with a large number of trees.

Table 2: Past Research on Predicting Startup Success Using Random Forest

Authors	Data Source	Definition of success	Machine Learning Model	Accuracy	Sensitivity	F1 Score
Krishna et al. (2016)	Crunchbase	Acquired	Random Forest Classifier			
Dellermann, Lipusch, Ebel, Popp, & Leimeister (2017)	Crunchbase, Mattermark, and Dealroom	Series A funding	Random Forest Classifier	Not aim of the study		
Pan, Gao, & Luo (2018)	Crunchbase	M&A or IPO	Random Forest Classifier	0.843		0.391
Arroyo, Corea, Jimenez-Diaz, & Recio-Garcia (2019)	Crunchbase	Acquired, IPO or repeat funding round	Random Forest Classifier	0.818		
Ünal & Ceasu (2019)	Crunchbase	Operating, acquired or IPO	Random Forest Classifier	0.941		
Piskunova, Ligonenko, Klochko, Frolova, & Bilyk, 2022	Ukrainian Dealroom	Repeated Funding rounds	Random Forest Classifier	0.57	0.358	0.399
Bangdiwala, Mehta, Agrawal, & Ghane (2022)	Crunchbase	IPO or M&A	Random Forest Classifier	0.9243		
Cholil, et al. (2024)	Kaggle	Acquired	Random Forest Classifier	0.8393		

2.2.3 Support Vector Machine

Support Vector Machines (SVMs) are robust classifiers that identify the hyperplane which optimally divides data into distinct classes (James, Hastie, Witten, & Tibshirani, 2021). The goal is to maximize the margin between these classes. They perform well in high-dimensional spaces and can manage non-linear classification using kernel functions (Żbikowski & Antosiuk, 2021). Additionally, SVMs are resilient to overfitting, especially when there are many dimensions compared to the number of samples (Tomy & Pardede, 2018). Nevertheless, SVMs can be memory-intensive and require careful parameter tuning, such as the choice of kernel and regularization parameters (James, Hastie, Witten, & Tibshirani, 2021). The model's interpretability is also limited, particularly with complex kernels (Felgueiras, Batista, & Carvalho, 2020).

Table 3: Past Research on Predicting Startup Success Using Support Vector Machines

Authors	Data Source	Definition of success	Machine Learning Model	Accuracy	Sensitivity	F1 Score
Dellermann, Lipusch, Ebel, Popp, & Leimeister (2017)	Crunchbase, Mattermark, and Dealroom	Series A funding	Support Vector Machine (SVM)	Not aim of the study		
Tomy & Pardede (2018)	Australian dataset	Profitable	Support Vector Machine (SVM)	0.7347	0.75	
Arroyo, Corea, Jimenez-Diaz, & Recio-Garcia (2019)	Crunchbase	Acquired, IPO or repeat funding round	Support Vector Machine (SVM)	0.817		
Felgueiras et al. (2020)	Crunchbase		Support Vector Machine (SVM)		0.421	
Żbikowski & Antosiuk (2021)	Crunchbase and Web-based information	Operating with Series B financing, acquired or IPO	Support Vector Machine (SVM)	0.87	0.2	0.32
Vasquez, Santisteban, & Mauricio (2023)	Australian dataset	Profitability	Support Vector Machine (SVM)	0.97		

2.2.4 Gradient Boosting

Gradient Boosting is an ensemble technique that constructs models in a sequential manner, with each new model aiming to address the errors of the previous ones (Cholil et al., 2024). By aggregating multiple weak learners, usually decision trees, Gradient Boosting develops a robust predictive model that frequently achieves high accuracy (Arroyo, Corea, Jimenez-Diaz, & Recio-Garcia, 2019). This method is effective at capturing intricate patterns and feature

interactions. However, it is computationally intensive and may be susceptible to noisy data and outliers (James, Hastie, Witten, & Tibshirani, 2021). Additionally, Gradient Boosting has a potential risk of overfitting if not carefully tuned, particularly with an excessive number of boosting stages (Żbikowski & Antosiuk, 2021).

Table 4: Past Research on Predicting Startup Success Using Gradient Boosting

Authors	Data Source	Definition of success	Machine Learning Model	Accuracy	Sensitivity	F1 Score
Arroyo, Corea, Jimenez-Diaz, & Recio-Garcia (2019)	Crunchbase	Acquired, IPO or repeat funding round	Gradient Tree Boosting (GTB)	0.822		
Ünal & Ceasu (2019)	Crunchbase	Operating, acquired or IPO	Extreme Gradient Boosting	0.945		
Corea et al. (2021)	Crunchbase + LinkedIn	Acquired, IPO or repeat funding round	Gradient Boosting Machine	Precision: ~0.7		
Żbikowski & Antosiuk (2021)	Crunchbase and Web-based information	Operating with Series B financing, acquired or IPO	Extreme Gradient Boosting	0.86	0.17	0.28
Bangdiwala, Mehta, Agrawal, & Ghane (2022)	Crunchbase	IPO or M&A	Gradient Tree Boosting (GTB)	0.9196		
Thirupathi, Alhanai, & Ghassemi (2022)	Crunchbase	IPO or M&A	Extreme Gradient Boosting	0.84		
Vasquez, Santisteban, & Mauricio (2023)	Australian dataset	Profitability	Gradient Boosting	0.91		
Cholil, et al. (2024)	Kaggle	Acquired	Extreme Gradient Boosting	0.881		
			Light Gradient Boosting Machine	0.881		
			Gradient Boosting	0.875		

2.2.5 Neural Networks

Neural Networks are composed of layers of interconnected nodes (neurons) that emulate the structure of the human brain (James, Hastie, Witten, & Tibshirani, 2021). They offer significant flexibility and are adept at learning intricate patterns and representations from data through backpropagation (Bangdiwala, Mehta, Agrawal, & Ghane, 2022). Neural Networks are particularly effective with large datasets and varied data types, such as images, text, and time-series (James, Hastie, Witten, & Tibshirani, 2021). They can automatically extract relevant features from raw data, which is a notable advantage. However, they require considerable data and computational power for successful training (Gichohi, 2023). The complexity of Neural Networks often renders them a "black box," making it difficult to interpret how decisions are derived.

Table 5: Past Research on Predicting Startup Success Using Neural Networks

Authors	Data Source	Definition of success	Machine Learning Model	Accuracy	Sensitivity	F1 Score
Dellermann, Lipusch, Ebel, Popp, & Leimeister (2017)	Crunchbase, Mattermark, and Dealroom	Series A funding	Artificial Neural Network (ANN)	Not aim of the study		
Ferrati et al. (2021)	Crunchbase, United States and Patents Office and CB insights top investors	Acquired or IPO	Neural Network		0.93	
Bangdiwala, Mehta, Agrawal, & Ghane (2022)	Crunchbase	IPO or M&A	Neural Network	0.9186		
Gichohi (2023)	Crunchbase	Funding rounds, attributes of entrepreneur, M&A	Artificial Neural Network (ANN)	0.86		

2.2.6 Naive Bayes

Naive Bayes is a probabilistic classifier grounded in Bayes' theorem, assuming feature independence for simplicity (James, Hastie, Witten, & Tibshirani, 2021). It estimates the probability of a class based on the features and assigns the class with the highest probability (James, Hastie, Witten, & Tibshirani, 2021). This model is straightforward, fast, and works well with small datasets and text classification tasks. It can also manage missing data by disregarding absent features during training (Felgueiras, Batista, & Carvalho, 2020). However, the assumption of feature independence may be unrealistic, potentially leading to suboptimal performance if the features are dependent (James, Hastie, Witten, & Tibshirani, 2021). Additionally, Naive Bayes has limited capacity to model complex relationships between features.

Table 6: Past Research on Predicting Startup Success Using Naive Bayes

Authors	Data Source	Definition of success	Machine Learning Model	Accuracy	Sensitivity	F1 Score
Krishna et al. (2016)	Crunchbase	Acquired	Naive Bayes			
Dellermann, Lipusch, Ebel, Popp, & Leimeister (2017)	Crunchbase, Mattermark, and Dealroom	Series A funding	Naive Bayes	Not aim of the study		
Tomy & Pardede (2018)	Australian dataset	Profitable	Naive Bayes	0.7755	0.805	
Felgueiras et al. (2020)	Crunchbase		Naive Bayes			

2.2.7 Decision Trees

Decision Trees partition data into subsets based on feature values, forming a tree-like structure of decisions (James, Hastie, Witten, & Tibshirani, 2021). Each internal node represents a test of a feature, each branch signifies an outcome, and each leaf node denotes a class label or regression result (Piskunova, Ligonenko, Klochko, Frolova, & Bilyk, 2021). They are straightforward to understand and visualize, and they effectively handle both numerical and categorical data. Additionally, Decision Trees offer insights into feature importance (Bangdiwala, Mehta, Agrawal, & Ghane, 2022). However, they are susceptible to overfitting, especially with deep trees and complex datasets (James, Hastie, Witten, & Tibshirani, 2021). Decision Trees can also be unstable, as minor changes in the data may result in different tree structures.

Table 7: Past Research on Predicting Startup Success Using Decision Trees

Authors	Data Source	Definition of success	Machine Learning Model	Accuracy	Sensitivity	F1 Score
Krishna et al. (2016)	Crunchbase	Acquired	ADTrees	Precision: 0.88-0.97		
Arroyo, Corea, Jimenez-Diaz, & Recio-Garcia (2019)	Crunchbase	Acquired, IPO or repeat funding round	Decision Trees	0.746		
Piskunova, Ligonenko, Klochko, Frolova, & Bilyk, 2023	Ukrainian Dealroom	Repeated Funding rounds	Decision Trees (DT)	0.612	0.347	0.523
Bangdiwala, Mehta, Agrawal, & Ghane (2022)	Crunchbase	IPO or M&A	Decision Trees (DT)	0.9243		

2.2.8 K-Nearest Neighbours

K-Nearest Neighbors (KNN) is an instance-based learning algorithm that classifies a sample based on the majority class among its k nearest neighbors in the feature space (James, Hastie, Witten, & Tibshirani, 2021). It is easy to understand and implement, requiring no training phase as all computations are performed during prediction. KNN can be applied to both classification and regression tasks (James, Hastie, Witten, & Tibshirani, 2021). However, it can be computationally intensive and slow, particularly with large datasets due to the need for distance calculations (James, Hastie, Witten, & Tibshirani, 2021). Additionally, KNN is sensitive to noisy data and irrelevant features, and its performance is greatly influenced by the choice of the parameter k and the distance metric (James, Hastie, Witten, & Tibshirani, 2021).

Table 8: Past Research on Predicting Startup Success Using K-Nearest Neighbours

Authors	Data Source	Definition of success	Machine Learning Model	Accuracy	Sensitivity	F1 Score
Pan, Gao, & Luo (2018)	Crunchbase	M&A or IPO	K-Nearest Neighbors	0.7333		0.464
Tomy & Pardede (2018)	Australian dataset	Profitable	K-Nearest Neighbors	0.7143	0.777	

In this research on predicting early-stage startup success, the selection of Logistic Regression, Random Forest and Support Vector Machines (SVM) as the top three models was driven by their complementary strengths and the ability to address each other's limitations. Logistic Regression offers simplicity and interpretability, essential for understanding the impact of different features on predictions. Random Forest deliver robustness and high accuracy through ensemble learning, reducing the risk of overfitting seen in individual decision trees. SVMs are adept at managing high-dimensional data and complex relationships via kernel functions, making them effective for capturing intricate patterns

The downsides of one model are effectively complemented by the strengths of the others; for example, the linear assumptions of Logistic Regression are offset by the non-linear capabilities of SVMs. Similarly, the computational intensity of Support Vector Machines is balanced by the efficiency of Logistic Regression, while the robustness of Random Forest can counterbalance the interpretability challenges posed by SVMs. This combination of models

allows for a comprehensive approach to predicting startup success, leveraging the unique strengths of each while addressing their individual limitations.

2.3 The Relationship Between Startup Success Factors And Machine Learning

The application of machine learning models in the private capital space is rapidly gaining traction (Ferrati & Muffatto, 2020; Gerdin, 2022). These models are increasingly being used to predict startup success, forecast the timing of future funding rounds, and assess the likelihood of an investor committing to a specific company (Ferrati & Muffatto, 2020). Furthermore, machine learning models assist equity investors in their decision-making by categorizing companies into industries. This classification helps identify similar firms and aligns investors with companies within their sectors of interest (Ferrati & Muffatto, 2020).

This emerging trend, often referred to as "data-driven venture capital," is praised for its efficiency, effectiveness, and inclusivity compared to traditional venture capital practices (Data Driven VC, 2024). Traditional venture capital process, as described by Gompers et al. (2020), follows a funnel-like process where the number of startups progressing to the next stage decreases exponentially, with only one out of every 101 deals considered ultimately closing (Gompers, Gornall, Kaplan, & Strebulaev, 2020). This approach is resource-intensive, requiring significant time, personnel, and financial investment, and may prove ineffective if a venture capital (VC) firm overlooks outlier opportunities (Data Driven VC, 2024). Additionally, traditional venture capital is prone to bias, resulting in a disproportionate allocation of capital (Data Driven VC, 2024).

Globally, disparities in funding are evident, with venture capital in North America being 52 times greater than in the Latin American (LATAM) region (Data Driven VC, 2024). In Africa, most venture capital firms are led by individuals of American and European descent, resulting in a preference for funding white founders, often sidelining African entrepreneurs with equally fund-worthy startups (Obonyo & Zeisberger, 2024; Turman, 2023; The Guardian, 2020; Ganesan, Mahalingam, Nathan, Ware, & Weinberg, 2023). As noted, “Companies led by white males or Africans with strong ‘Western’ backgrounds attract more VC investment than comparable companies in the African startup ecosystem, as investors undervalue the need and advantages of local knowledge” (Turman, 2023).

In response to the challenges and limitations inherent in the traditional venture capital process, data-driven venture capital introduces three significant changes (Gerdin, 2022). First, VC firms can source deals on a larger scale (Weibl & Hess, 2019). Second, the quality of scrutinized

deals is enhanced, as each company is scored and evaluated before being presented to the investment team (Corea, Bertinetti, & Cervellati, 2021; Weibl & Hess, 2019). Third, by automating the sourcing process, the investment team can focus more on value-added activities (Weibl & Hess, 2019). These changes would significantly address the downsides of the traditional venture capital process.

Many researchers utilize Crunchbase data to train machine learning models for predicting startup success (Krishna, Agrawal, & Choudhary, 2016; Pan, Gao, & Luo, 2018; Arroyo, Corea, Jimenez-Diaz, & Recio-Garcia, 2019; Ünal & Ceasu, 2019; Żbikowski & Antosiuk, 2021; Bangdiwala, Mehta, Agrawal, & Ghane, 2022; Felgueiras, Batista, & Carvalho, 2020; Thirupathi, Alhanai, & Ghassemi, 2022; Gichohi, 2023). However, the definition of startup success varies among stakeholders. Piskunova et al. (2021) outline different success metrics, including investment success (securing additional financing), customer success (achieving target user growth), market success (reaching sales targets or market share), adaptive success (surviving beyond five years), and financial success (achieving an IPO or acquisition, allowing founders and investors to exit and monetize their investments Financial success aligns with the "classic understanding" of startup success (Piskunova, Ligonenko, Klochko, Frolova, & Bilyk, 2021). It is widely recognized that the key milestone marking a venture-backed company as financially successful is the exit event (Ferrati & Muffatto, 2020). A venture-backed company can achieve an exit through two primary strategies: by conducting an Initial Public Offering (IPO) or by being acquired by a larger company through mergers and acquisitions (M&A).

Crunchbase's datasets are extensive, including seventeen .csv files that cover five major areas (Ferrati & Muffatto, 2020). These datasets provide essential information, including company status (operating, closed, acquired, or IPO), total funding amounts, most recent funding dates, employee counts, and geographical details, making them highly valuable for machine learning classification models.

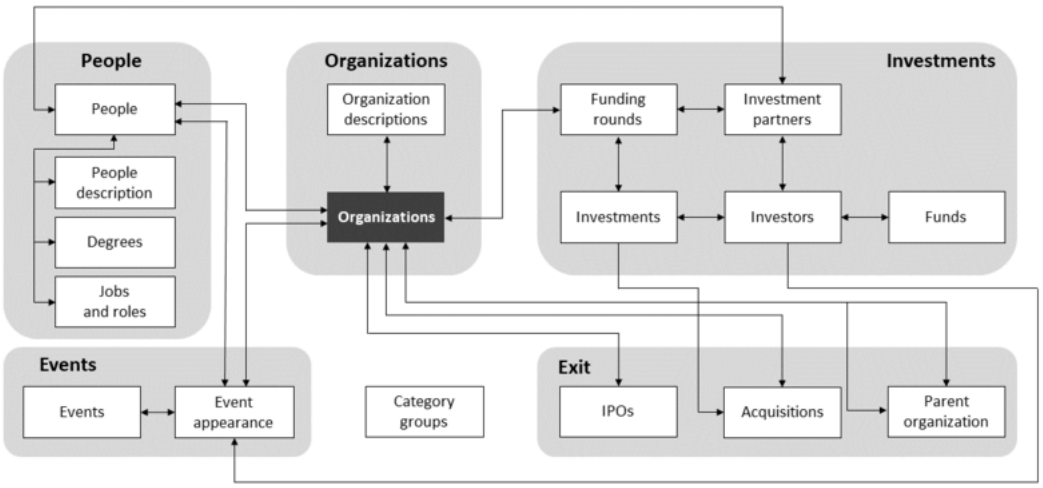

Figure 1: Relationships of the Crunchbase's datasets

In reviewing the literature, 14 out of 20 studies used the company status variable as the dependent variable for predicting financial success, with the remaining studies focusing on repeat funding and profitability as alternative success metrics. With a well-constructed machine learning model, access to reliable and sufficient data, and a clearly defined target variable, it is possible to predict the success of any startup with greater accuracy.

2.4 Gaps Found In The Literature

In the realm of predicting startup success, existing literature presents several critical gaps that undermine the comprehensiveness and accuracy of current models. This section highlights these gaps, underscoring the need for more nuanced approaches that incorporate varied datasets, sectoral diversity, geographical representation, and larger, more robust datasets, especially in the African context.

Most studies in this area rely on cross-sectional data, which captures information about startups at a single point in time. This approach fails to account for the dynamic nature of startups, whose trajectories and outcomes evolve over time. The lack of longitudinal data, which tracks startups over extended periods, limits our understanding of how and why startups fail. Incorporating panel data that captures growth metrics, such as changes in employee count and funding rates, could significantly enhance the accuracy of predictive models by offering insights into the triggers of startup success or failure over time. The scarcity of longitudinal

studies tracking startup performance is a significant gap, as it overlooks the temporal dimensions critical to understanding the evolution of startups.

Much of the research on startup success is centered around regions with well-established venture capital ecosystems, such as North America and Europe. There is a notable scarcity of studies focusing on Africa, where the venture capital scene is still emerging. The literature available often excludes African data points since they have zero or near-zero variance, as noted by Ünal & Ceasu (2019). This exclusion could suggest that African data points may act as outliers on the downward side when compared to those from other continents, further skewing the analysis. The underrepresentation of African startups in the literature is a significant gap, as it ignores the unique challenges and opportunities within this rapidly growing market. Furthermore, the nascent use of machine learning for predicting startup success in Africa exacerbates this gap, as existing models may not be well-suited to the African context.

Another critical gap in the literature is the potential misrepresentation of data. For instance, AVCA (2024) suggests that the global mean sizes of funding rounds portray Africa as being on par with other continents. However, this may be misleading, as Africa has fewer companies, and the mean could create the illusion of parity with more developed markets. This misrepresentation could skew the understanding of Africa's startup ecosystem, leading to inaccurate conclusions and predictions.

As of 2020, Crunchbase data indicated that only 4.4% of companies represented were closed businesses, despite a high known failure rate for startups (Ferrati & Muffatto, 2020). This discrepancy may be due to profiles being deleted upon failure, leading to an incomplete picture of startup success and failure. This underreporting of failures is a significant gap, as it prevents a full understanding of the factors contributing to startup demise, which is crucial for developing accurate predictive models.

Beyond these points, the literature also fails to adequately address the differences in different entrepreneurial ecosystems. The one-size-fits-all approach often seen in existing models does not account for these variables, further limiting the applicability of the findings to diverse contexts, particularly in Africa. By addressing these gaps, future research can develop more accurate and context-specific models for predicting startup success, particularly in underrepresented regions and sectors.

Although this study relies on cross-sectional data, it will deepen its analysis by incorporating a comprehensive set of variables that capture a wide range of factors influencing startup success. To address potential data misrepresentation, the study will conduct a rigorous analysis that accounts for the unique economic and venture capital dynamics in Africa. Instead of relying solely on average funding sizes, the study will examine the distribution of funding amounts, the variance between startups, and the implications of these factors on the perceived success of African startups.

Even within the cross-sectional framework, the study will include data on both successful and failed startups to address the gap of underreported failures and survivorship bias. By ensuring the dataset includes startups that did not succeed, the study will analyze the factors contributing to failure, providing a more balanced and realistic model of startup success. This approach will help develop predictive models that are more robust and reflective of the true dynamics within the startup ecosystem.

2.5 Conceptual Framework

This conceptual framework is designed to explore the relationship between key success factors and the eventual success of startups, utilizing supervised machine learning models. It identifies critical variables, such as company status, financial metrics, industry, headquarters location, and founder details, as the primary inputs for predicting startup success. The framework demonstrates how these factors are analyzed through machine learning techniques, including Logistic Regression, Random Forest, and Support Vector Machines, to predict outcomes such as repeated funding, acquisition, IPO, and sustained operations beyond five years. This approach aims to enhance the understanding of startup success and improve investment decision-making processes by leveraging data-driven insights.

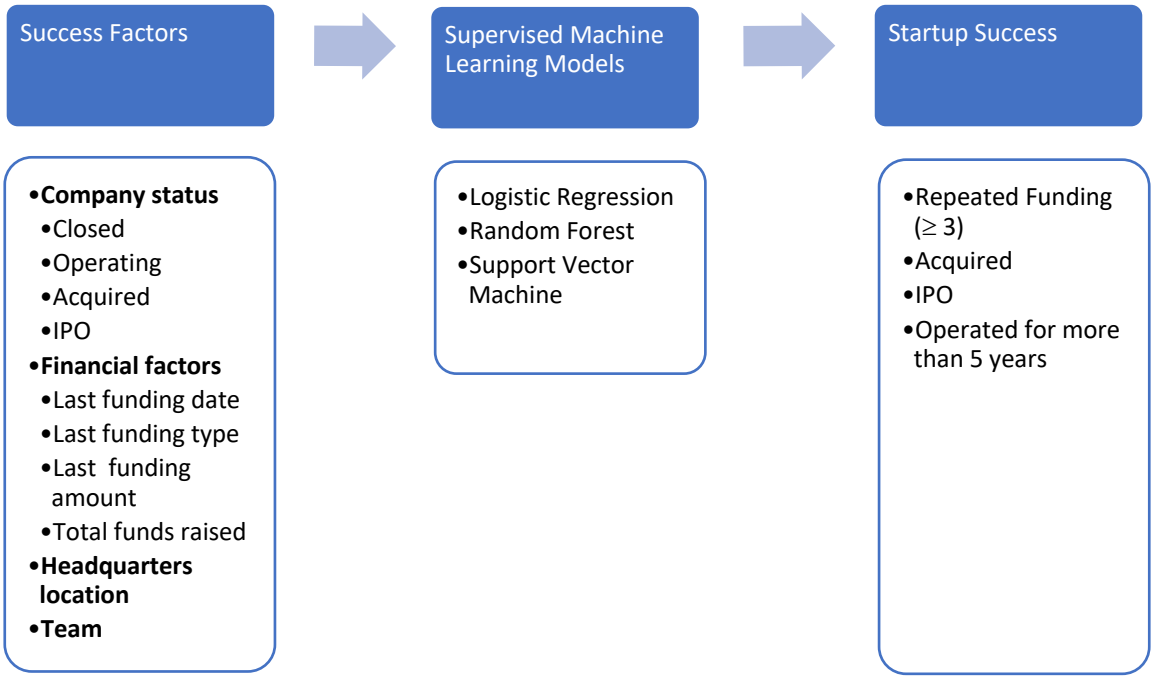

Figure 2: Conceptual Framework.

Chapter 3: Methodology

3.1 Introduction

This section introduces the methodology employed in the study, detailing the research design, population and sampling methods, data collection, and analysis procedures. The study uses a purely quantitative approach, utilizing machine learning to analyze the success of African startups.

3.2 Research Design

The research employed a quantitative research design to analyze the success of African startups using machine learning approaches. This design was chosen for its ability to manage large datasets while integrating numerical analysis with deeper contextual insights. By utilizing secondary data from Crunchbase, which includes startups from various continents with different funding rounds, the study focused on a comprehensive sample of startups actively operating within the African market. Crunchbase is a credible database for African data points due to its comprehensive coverage of global startups, including those in Africa, and its focus on emerging markets. The platform's data is regularly updated and verified daily, ensuring accuracy, and it provides detailed information on funding rounds, which is crucial for understanding the financial landscape of African startups. Widely recognized and used by investors and researchers, Crunchbase offers reliable and structured data that is well-suited for integration into machine learning models, making it a valuable resource for analyzing the unique dynamics of the African startup ecosystem. This approach provided detailed insights into the broader startup ecosystem in Africa. The cross-sectional nature of the data allowed for a snapshot analysis of the startups at a specific point in time.

3.3 Population And Sampling

The population for this study comprised startups listed on Crunchbase globally, representing various continents and stages of funding. The sample selected includes 44,831 startups operating in the African market as of August 31, 2024, across all funding stages. This broad inclusion allows for a comprehensive analysis, incorporating quantitative metrics. Of these, 28,851 startups, founded between 2000 and 2024, will be used for exploratory data analysis. The representativeness of the sample was ensured by including a wide range of startups from various sectors and regions across Africa. This approach provided a comprehensive view of the startup landscape within the African market, minimizing sampling bias and enhancing the generalizability of the research findings.

3.4 Data Collection

The data collected for this study were both quantitative and qualitative in nature, drawn from multiple datasets provided by Crunchbase, a comprehensive platform that aggregates information on organizations, investors, and related entities. The quantitative data encompassed numerical information such as financial metrics, funding rounds, valuations, investment amounts, and the number of investors involved. Qualitative data provided deeper contextual insights, including details on the startups' operating status, industry classifications, investor profiles, team structures, geographical locations, and significant events such as acquisitions and IPOs.

The datasets utilized in this research included:

1. Organization: This dataset provided detailed information on each startup, including its name, country of operation, industry, status, and financial metrics such as total funding received.
2. Funding Rounds: This dataset detailed the various funding rounds that startup underwent, including the amount raised, the type of funding, and the investors involved. It also included temporal data such as the date of funding announcements.
3. Investors: This dataset provided profiles of investors, including their investment history, types, and geographical focus.
4. Acquisitions: Data on acquisitions were used to track exit events, detailing both the acquiring and acquired entities, as well as the transaction value.
5. IPOs: Information on startups that went public, including their IPO dates, share prices, and market valuations.
6. Events: This dataset captured significant events in the lifecycle of the startups, such as product launches, partnerships, and other notable occurrences.

Each dataset is joined using unique identifiers such as UUIDs, allowing for a comprehensive analysis across multiple dimensions.

3.4.1 Data Collection For Objective 1

For the first objective, identifying and analyzing key success factors, the data was meticulously extracted from the aforementioned datasets. The focus was on gathering information that could elucidate the critical elements influencing the success of early-stage African startups. This involved selecting key variables such as the total funding amount, number of funding rounds, operating status and business age.

3.4.2 Data Collection For Objective 2

For the second objective—developing a supervised machine learning model to predict startup success—the data collected was tailored to feed into the models' training processes. This involved preparing a dataset that included both the predictor variables (such as financial metrics, operating status and business age) and the target variable, the success or failure of the startup, as indicated by key outcomes like successful exits (acquisitions or IPOs), repeated funding or ongoing operations for more than 5 years.

3.5 Data Analysis

The analysis involved a systematic approach to identifying success factors and developing a predictive model based on the data collected.

3.5.1 Identifying Success Factors

To achieve the first objective of identifying critical success factors for early-stage African startups, a thorough and methodical approach was employed, combining both exploratory data analysis (EDA) and advanced statistical techniques.

3.5.1.1 Exploratory Data Analysis

The initial phase involved conducting an EDA to understand the underlying patterns, distributions, and relationships within the dataset. This step was crucial for uncovering any potential correlations or trends that could influence startup success. The EDA process included:

1. Descriptive Statistics: Calculating measures such as mean, median, standard deviation, and interquartile ranges for independent variables. This provided a summary view of the data, helping to identify any outliers or anomalies.
2. Correlation Analysis: Evaluating the pairwise correlations between different variables, such as the relationship between the amount of funding received and the likelihood of success. This helped in pinpointing variables that have a strong linear relationship with startup success.
3. Visualization: Utilizing visual tools like histograms, box plots, and scatter plots to visually inspect the data. Heatmaps were also used to display the correlation matrix, making it easier to identify significant relationships at a glance.
4. Segment Analysis: Breaking down the dataset into different segments based on categorical variables such as geographical region. This segmentation allowed for a more granular analysis, revealing how specific factors might contribute to success in different contexts.

3.5.2 Model Selection

To achieve the study's second objective, several supervised learning algorithms were considered, including Logistic Regression, Random Forest and Support Vector Machines (James, Hastie, Witten, & Tibshirani, 2021). These algorithms were selected for their proven effectiveness in handling classification tasks and their ability to manage complex relationships between predictor variables (Żbikowski & Antosiuk, 2021; Krishna, Agrawal, & Choudhary, 2016; Arroyo, Corea, Jimenez-Diaz, & Recio-Garcia, 2019; Piskunova, Ligonenko, Klochko, Frolova, & Bilyk, 2021; Pan, Gao, & Luo, 2018). A comprehensive comparative analysis of all these algorithms will be conducted, and the final model will be chosen based on performance metrics such as accuracy, precision, recall, and the F1 score (Pan, Gao, & Luo, 2018; Piskunova, Ligonenko, Klochko, Frolova, & Bilyk, 2021). Additionally, the selection process will consider each model's interpretability and its ability to generalize to new data. Following the guidelines from the textbook *Introduction to Statistical Learning with Applications in R*, each model will be implemented and evaluated in detail, and the best-performing model will be selected for the study (James, Hastie, Witten, & Tibshirani, 2021).

3.5.2.1 Logistic Regression

Logistic regression is a fundamental classification technique used in machine learning, particularly when the dependent variable is categorical (Bai & Zhao, 2021). It is a linear model that predicts the probability that a given input belongs to a particular class. The model uses the logistic function, also known as the sigmoid function, to map the output of a linear combination of input features to a probability.

Mathematically, the logistic function is defined as $h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$,where θ represents the model parameters (weights) and x denotes the input feature vector. The output of this function, $h_\theta(x)$ is a value between 0 and 1, which can be interpreted as the probability of the input belonging to the positive class (James, Hastie, Witten, & Tibshirani, 2021).

To make a prediction, logistic regression applies a decision boundary at 0.5. If the predicted probability $h_\theta(x)$is greater than or equal to 0.5, the model predicts the input belongs to the positive class (label 1); otherwise, it predicts the negative class (label 0) (James, Hastie, Witten, & Tibshirani, 2021). The model's parameters are optimized by minimizing the cost function, which is defined as the binary cross-entropy or log loss:

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\log\left(h_\theta\left(x^{(i)}\right)\right) + \left(1-y^{(i)}\right)\log\left(1-h_\theta\left(x^{(i)}\right)\right)\right]$$

Here, m is the number of training examples, $y^{(i)}$ is the actual label for the i -th example, and $h_\theta\left(x^{(i)}\right)$ is the predicted probability for that example (James, Hastie, Witten, & Tibshirani, 2021).

The parameters θ are iteratively updated using gradient descent to minimize the cost function. The update rule for gradient descent is given by $\theta_j := \theta_j - \alpha\frac{\partial}{\partial\theta_j}J(\theta), where\ \alpha$ is the learning rate and j indexes the parameters. This process continues until the cost function converges to a minimum, at which point the model is considered trained and ready to make predictions on new data (James, Hastie, Witten, & Tibshirani, 2021).

3.5.2.2 Random Forest

Random Forest are a powerful ensemble learning method that builds multiple decision trees during training and outputs the mode of the classes (in classification) or the mean prediction (in regression) of the individual trees (Bai & Zhao, 2021). The basic idea behind random Forest is to create a "forest" of decision trees, where each tree is trained on a random subset of the data, and the final prediction is made by aggregating the predictions of all the trees. This process helps to reduce the variance and improve the overall accuracy of the model.

Each decision tree in a random forest is built using a process called bagging, or bootstrap aggregating, where a random subsample of the training data is drawn with replacement (James, Hastie, Witten, & Tibshirani, 2021). This means that some training examples may appear multiple times in the same tree, while others may not appear at all. Additionally, at each split in the tree, a random subset of features is considered for determining the best split, which introduces further randomness and diversity among the trees.

The splitting criterion used in classification trees is often the Gini impurity, defined as $\text{Gini}(D) = 1 - \sum_{k=1}^{K} p_k^2, where\ p_k$ is the proportion of instances of class k in the dataset D, and K is the number of classes (James, Hastie, Witten, & Tibshirani, 2021). The goal is to select the split that results in the highest reduction in impurity, thereby creating the most homogeneous child nodes (James, Hastie, Witten, & Tibshirani, 2021).

Once all the trees are built, the random forest makes a prediction by combining the outputs of the individual trees. For classification tasks, the final prediction is made by majority voting,

where the class with the most votes is selected. For regression tasks, the final prediction is the average of the predictions from all the trees. The ensemble nature of random Forest allows them to achieve high accuracy and robustness, particularly in situations where individual decision trees might be overfit to the training data (James, Hastie, Witten, & Tibshirani, 2021).

3.5.2.3 Support Vector Machine

Support vector machines (SVMs) are a class of supervised learning algorithms that are particularly well-suited for classification tasks in high-dimensional spaces (Bai & Zhao, 2021). The core idea behind SVMs is to find the optimal hyperplane that separates the data points of different classes with the maximum margin (James, Hastie, Witten, & Tibshirani, 2021). The margin is defined as the distance between the hyperplane and the nearest data points from either class, which are known as support vectors. SVM seeks to maximize this margin while correctly classifying the training data.

The decision function for an SVM is given by $f(x) = w^T x + b$, where w is the weight vector and b is the bias term. The optimization problem that SVMs solve is to minimize the norm of the weight vector $|w|^2$, subject to the constraint that all data points are correctly classified with a margin of at least 1 (James, Hastie, Witten, & Tibshirani, 2021). Mathematically, this can be expressed as $\min_{w,b} \frac{1}{2}|w|^2$, subject to $y^{(i)}\left(w^T x^{(i)} + b\right) \geq 1$ for all i, where $y^{(i)}$ is the label of the i-th data point and $x^{(i)}$ is its feature vector (James, Hastie, Witten, & Tibshirani, 2021).

In practice, data is often not perfectly separable, so SVMs introduce slack variables ξ_i to allow for some misclassification (James, Hastie, Witten, & Tibshirani, 2021). This leads to the soft margin SVM, where the optimization problem becomes $\min_{w,b,\xi} \frac{1}{2}|w|^2 + C\sum_{i=1}^{m} \xi_i$, subject to $y^{(i)}\left(w^T x^{(i)} + b\right) \geq 1 - \xi_i$ and $\xi_i \geq 0$ for all i. The parameter C controls the trade-off between maximizing the margin and minimizing the classification error (James, Hastie, Witten, & Tibshirani, 2021).

For non-linearly separable data, SVMs can be extended using the kernel trick, which maps the input features into a higher-dimensional space where a linear separator can be found (James, Hastie, Witten, & Tibshirani, 2021). Common kernel functions include the linear kernel $K(x, x') = x^T x'$, the polynomial kernel $K(x, x') = (x^T x' + c)^d$, and the radial basis function (RBF) kernel $K(x, x') = \exp(-\gamma|x - x'|^2)$. The optimization problem can also be

reformulated in its dual form, where the solution depends only on the support vectors. In the dual formulation, the objective is to

maximize $\sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} K\left(x^{(i)}, x^{(j)}\right)$, subject to $0 \leq \alpha_i \leq C$ and $\sum_{i=1}^{m} \alpha_i y^{(i)} = 0$, where α_i are the Lagrange multipliers. This formulation allows SVMs to efficiently handle large datasets and complex decision boundaries (James, Hastie, Witten, & Tibshirani, 2021).

3.5.3 Model Training And Validation

The dataset was divided into training and test sets, typically using an 70-30 split. The training set was used to build the model, while the test set provided an unbiased evaluation of the model's performance. To enhance the model's predictive power, hyperparameter tuning was performed using grid search and cross-validation techniques (Pan, Gao, & Luo, 2018). These methods helped in finding the optimal parameters that improve model accuracy while avoiding overfitting. Regularization techniques were also employed to further mitigate overfitting, ensuring that the model performed well on unseen data (James, Hastie, Witten, & Tibshirani, 2021).

3.5.4. Model Evaluation

Once trained, the model was evaluated using a range of metrics. Accuracy, precision, recall, and the F1 score provided a comprehensive view of the model's performance (Pan, Gao, & Luo, 2018). These metrics were interpreted in the context of the African startup ecosystem, with a focus on identifying the key factors that contribute to business success. The insights gained from this analysis not only validated the model but also offered valuable guidance for entrepreneurs and investors (James, Hastie, Witten, & Tibshirani, 2021).

3.5.4.1 Evaluation Metrics

1. Accuracy

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Where:

- TP: True Positives
- TN: True Negatives
- FP: False Positives
- FN: False Negatives

2. Precision (Positive Predictive Value)

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

3. Recall (Sensitivity or True Positive Rate)

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

4. F1-Score (Harmonic Mean of Precision and Recall)

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

5. Confusion Matrix

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

6. Cross-Validation Error

In k -fold cross-validation, the dataset is split into k subsets, and the model is trained kk times, each time using $k - 1$ subsets as the training set and the remaining subset as the validation set:

$$\text{CV-Error} = \frac{1}{k}\sum_{i=1}^{k} \text{Error}_i$$

Chapter 4: Results And Analysis

4.1 Dataset Overview: Focus on African Startups

The CrunchBase dataset provides a clear representation of the distribution of startups across African countries, as illustrated in the bar chart and the map visualization. The map highlights Africa as the focal point of this analysis, with a noticeable concentration of startup activity in specific regions.

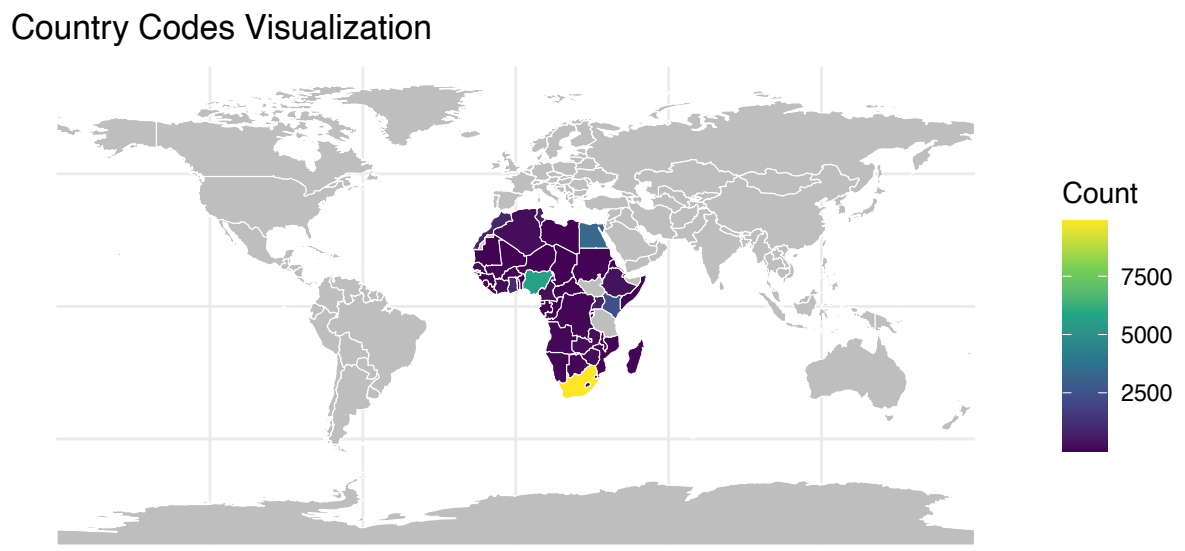

Figure 3: Continent Visualization

The bar chart further emphasizes the predominance of startups in four countries: Kenya, Nigeria, South Africa, and Egypt. These nations collectively account for the majority of the startup ecosystem in the dataset, underscoring their leadership roles in fostering innovation and entrepreneurship across the continent. The geographic and numerical disparities in startup representation highlight the unequal distribution of resources, funding, and entrepreneurial activity across Africa, making it evident why these four countries are often regarded as the hubs of technological and business advancements within the region.

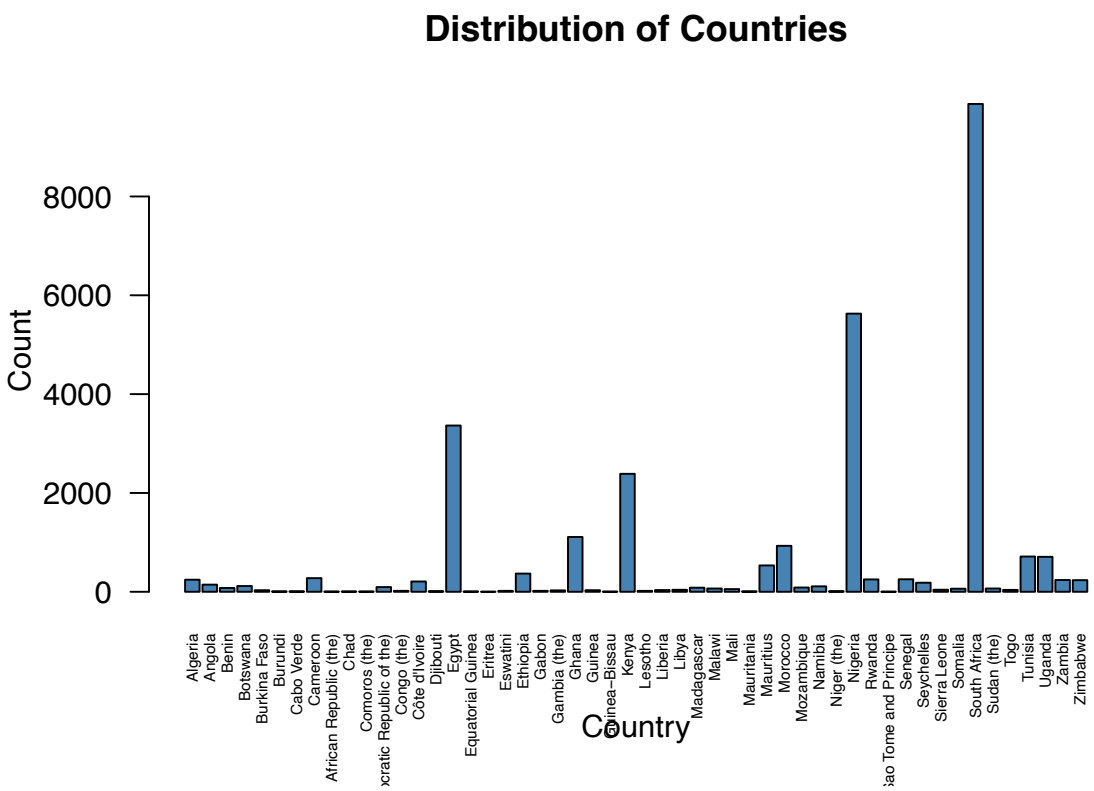

Figure 4: Distribution of Startups per Country

4.1.1 Data Preparation and Feature Selection

An extensive review of the dataset was conducted to identify and exclude extraneous variables that did not contribute directly to the analysis of factors influencing the success of startups. The primary goal of this step was to refine the dataset, ensuring that it included only the most critical features relevant to predicting success. After careful consideration and based on findings from prior research, the dataset was narrowed down to include features such as total funding, operating status, business age, and the number of funding rounds. These variables were selected due to their high relevance in determining the success trajectory of early-stage businesses.

4.1.2 Data Transformation and Encoding

In its original form, the dataset contained both numeric and non-numeric variables, which posed challenges for the seamless application of machine learning algorithms. To address this, a transformation process was employed to encode non-numeric variables into numeric formats. For example, the categorical variable representing business status was recoded for consistency and computational purposes. Specifically, the statuses "operating," "acquired," and "IPO" were

encoded as 1,2 and 3 respectively, while the status "closed" was encoded as 4. This binary encoding system allowed for better integration of categorical data into the modeling pipeline, eliminating potential incompatibilities and ensuring that the dataset was well-suited for predictive analytics. The histogram below shows the data distribution among the top variables.

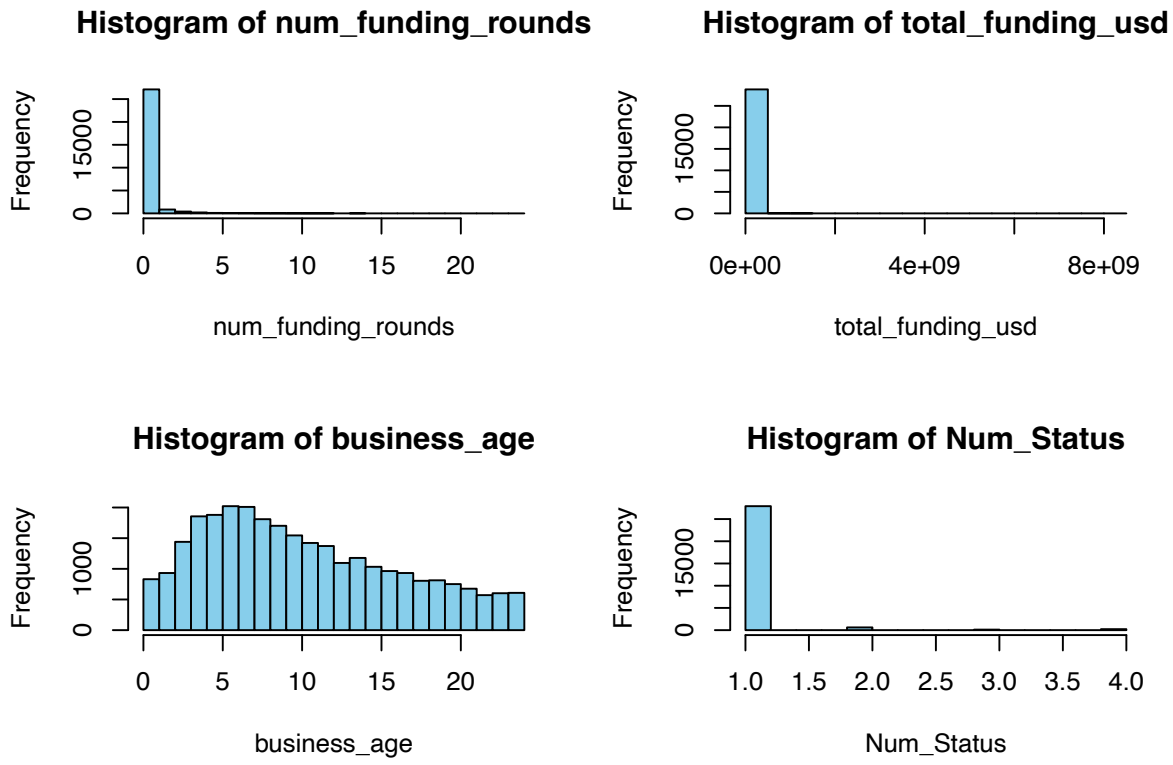

Figure 5: Data Distribution of Original Variables

4.1.3 Success Criteria

The definition of "success" in the context of this analysis was operationalized using three binary specific metrics. The first metric was longevity, defined as whether a business was still operational after five years. Businesses that survived beyond this threshold were classified as successful, given that five years is widely considered a significant milestone indicating resilience, adaptability, and market acceptance. The second metric involved whether the business had achieved notable milestones such as acquisition or an initial public offering (IPO). These milestones represent critical junctures in the lifecycle of a business and are often regarded as markers of success. Finally, the third metric focused on the number of funding

rounds a startup secured. Startups that had completed three or more funding rounds were deemed successful, as this indicated sustained investor confidence and effective resource mobilization over time. Together, these three metrics provided a robust framework for assessing success in a comprehensive manner by creating three scenarios where they are dependent variables.

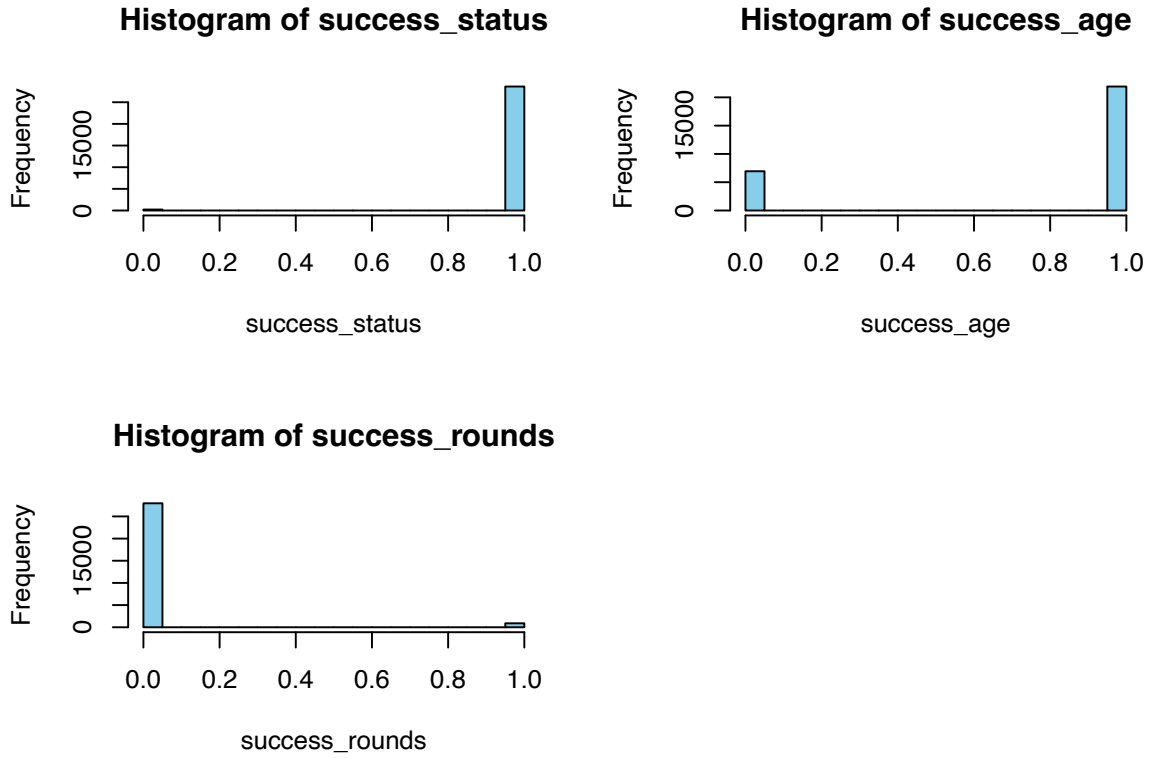

Figure 6: Data Distribution of Derived Variables

4.1.4 Dataset Overview After Implementing the Success Criteria

One of the most significant challenges encountered during the analysis was the issue of data imbalance. Applying the success criteria revealed a stark disparity in the frequency of observations across the different classes. For instance, the distribution of success status, success age, and success rounds was heavily skewed. Specifically, the dataset revealed that only 222 businesses fell into the "not successful" category based on their status, compared to 28,629 businesses classified as "successful." Similarly, for success age, there were 6,939 observations in the "not successful" category and 21,912 in the "successful" category. The success rounds

metric exhibited the most severe imbalance, with only 894 successful businesses compared to 27,957 in the "not successful" group.

Table 9: Dataset overview after applying the success criteria

Data Frequency		
Dependent Variable	0 (not success)	1 (success)
success_status	222	28629
success_age	6939	21912
success_rounds	27957	894

4.1.4.1 Mitigation Strategies for Imbalanced Data

This significant imbalance posed challenges for model training and evaluation, as machine learning algorithms are generally biased toward the majority class. To address this issue, multiple strategies were implemented. Class weighting was employed, wherein higher weights were assigned to the minority class to ensure that it received adequate attention during the training process. Additionally, stratified sampling was used to create balanced training and validation subsets, maintaining proportional representation of the minority class in each fold of the data. While these methods helped mitigate the effects of imbalance, the severe underrepresentation of the minority class in certain metrics remained a persistent challenge, limiting the model's ability to generalize effectively across all scenarios.

4.1.5 Advanced Data Transformation and Model Optimization Techniques

This captures the advanced processes you're discussing, emphasizing both feature engineering and dimensionality reduction while maintaining alignment with the goal of improving model performance.

4.1.5.1 Feature Engineering

To further enhance the predictive accuracy of the analysis, feature engineering was undertaken. This process involved deriving new variables from the existing dataset to capture additional

dimensions of information that were not explicitly present in the original data. One such feature was the funding-to-age ratio, which provided insights into how efficiently startups utilized their funding relative to their operational age. By dividing the total funding amount by the number of years the business had been operational, this ratio offered a measure of funding efficiency and resource utilization.

Another derived feature was funding per year, which normalized the total funding amount by the number of operational years. This normalization accounted for temporal effects, offering a clearer picture of the velocity at which funding was mobilized. A third engineered feature was the funding-rounds-age interaction, which captured the interplay between funding activity and business age. This interaction variable was designed to identify startups that exhibited consistent growth trajectories by considering both their funding patterns and operational maturity. Together, these engineered features enriched the dataset, providing deeper insights into the factors that contribute to startup success and enhancing the interpretability and predictive accuracy of the analysis.

4.1.5.2 Dimensionality Reduction with PCA

To simplify the dataset while retaining critical information, dimensionality reduction techniques were applied. Principal Component Analysis (PCA) was employed to reduce the number of features to a smaller set of uncorrelated components, thereby streamlining the modeling process and reducing computational complexity. The PCA analysis revealed that the first two principal components captured 100% of the variance in the dataset, indicating a high degree of informational efficiency. Specifically, the first principal component accounted for 57.71% of the variance, while the second component explained the remaining 42.29%. This result demonstrated that the dimensionality reduction process was successful in preserving the essential characteristics of the dataset while eliminating redundancy.

Table 10: PCA Components Analysis

Summary of PCA		
Importance of components:		
	PC1	PC2
Standard deviation	1.0743	0.9197
Proportion of Variance	0.5771	0.4229
Cumulative Proportion	0.5771	1

4.1.5.3 Data Splitting and Cross-Validation

The refined dataset was then divided into three subsets to ensure robust model evaluation and prevent overfitting. These subsets included a training set, comprising 70% of the data; a validation set, accounting for 20%; and a test set, consisting of the remaining 10%. Stratified sampling was applied during the splitting process to maintain consistent class distributions across all subsets, ensuring that the imbalance observed in the original dataset was proportionally represented in each subset.

A rigorous 10-fold cross-validation strategy was adopted to evaluate the performance of the models. In this approach, the training data was partitioned into ten equal folds, with nine folds used for training and the remaining fold used for validation. This process was repeated ten times, with each fold serving as the validation set once. The average performance across all iterations was used as the final evaluation metric. This cross-validation strategy minimized the risk of overfitting and underfitting, ensuring that the models were robust and reliable.

4.2 Objective 1: To identify the critical factors that influence startup success.

4.2.1 Correlation Analysis and Key Insights

Correlation analysis was conducted to examine the relationships between key variables, providing valuable insights into their relative importance.

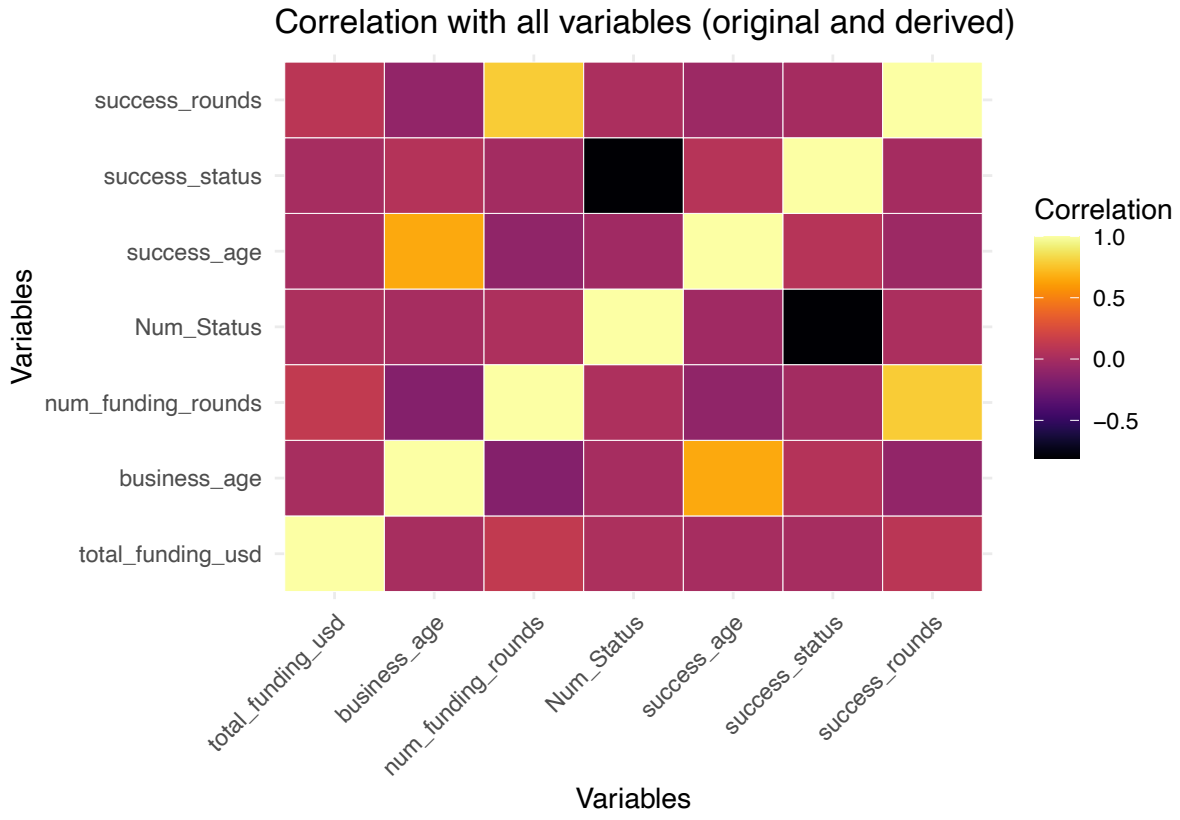

Figure 7: Heatmap of Data Variables

The analysis revealed a strong positive correlation between the number of funding rounds and success. Startups with a higher number of funding rounds consistently exhibited higher success rates, underscoring the importance of sustained financial backing in achieving growth and stability.

Another significant finding was the relationship between business age and success. Longevity emerged as a crucial predictor, with older businesses demonstrating greater resilience and adaptability to market conditions. This finding aligned with the operational definition of success, which emphasized the importance of surviving beyond five years as a key indicator of business viability.

Total funding was also positively correlated with success, although its predictive power was amplified when normalized by operational age. This observation highlighted the importance of considering temporal factors in the analysis, as businesses that secured substantial funding over a shorter period were more likely to exhibit rapid growth and market penetration.

4.3 Objective 2: Performance Analysis of the Machine Learning Predictive Models

Three supervised machine learning models—Logistic Regression, Support Vector Machine (SVM), and Random Forest—were evaluated using various performance metrics. The findings are summarized below:

4.3.1 Performance Analysis for the Dependent Variable: Success_Status

4.3.1.1 Validation Data Results

The accuracy of all three models—Logistic Regression, SVM, and Random Forest—on the validation dataset for success_status was 99%. This high accuracy suggests that the models were able to correctly predict the success status for nearly all instances in the dataset. The 95% confidence interval (CI) for the accuracy ranged from 99% to 99%, reflecting extremely low variance in the predictions. The No Information Rate (NIR), which indicates the accuracy that could be achieved by always predicting the majority class, was also 99%. This highlights a critical limitation: the models' performance was largely driven by the heavily imbalanced dataset, where the majority of the instances (approximately 99%) belonged to the “success” category.

Sensitivity, which measures the models' ability to correctly identify successful startups, was 100% across all three algorithms. This means that the models perfectly identified all startups labeled as “success” in the validation data. However, specificity, which measures the ability to correctly identify startups that were not successful, was 0% across all models. This indicates that none of the models could correctly predict the minority class (“not success”).

Metrics such as the Kappa statistic, which evaluates agreement between predicted and actual classifications beyond chance, were 0%, indicating that the models’ performance was no better than chance for the minority class. The McNemar’s test p-value was also 0%, reinforcing the notion that the models exhibited no discriminative ability for the minority class. Positive Predictive Value (PPV) was 99%, which aligns with the dataset’s imbalance. However,

Negative Predictive Value (NPV) could not be calculated due to the absence of correctly predicted instances for the “not success” class. The Balanced Accuracy, which averages sensitivity and specificity, was 50%, a clear reflection of the models' inability to generalize beyond the majority class.

Table 11:Results for success_status Using Validation Data

Validation Data - Success_Status			
Metrics	Logistic Regression	Support Vector Machine	Random Forest
Accuracy	99%	99%	99%
95% CI Lower	99%	99%	99%
95% CI Upper	99%	99%	99%
No Information Rate	99%	99%	99%
P-Value [Acc > NIR]	53%	53%	53%
Kappa	0%	0%	0%
Mcnemar's Test P-Value	0%	0%	0%
Sensitivity	100%	100%	100%
Specificity	0%	0%	0%
Positive Predictive Value	99%	99%	99%
Negative Predictive Value	NA	NA	NA
Prevalence	99%	99%	99%
Detection Rate	99%	99%	99%
Detection Prevalence	100%	100%	100%
Balanced Accuracy	50%	50%	50%

4.3.1.2 Test Data Results

The results on the test dataset for success_status mirrored those of the validation dataset. Accuracy remained at 99%, with a 95% CI ranging from 99% to 100%. Sensitivity remained at 100%, specificity at 0%, and Balanced Accuracy at 50%. Metrics like Kappa and McNemar's test p-value also remained unchanged, confirming that the models' over-reliance on the majority class persisted. The consistency between validation and test data results indicates that the models were not overfitting but were instead fundamentally constrained by the class imbalance.

Table 12:Results for success_status Using Test Data

Test Data - Success_Status			
Metrics	Logistic Regression	Support Vector Machine	Random Forest
Accuracy	99%	99%	99%
95% CI Lower	99%	99%	99%
95% CI Upper	100%	100%	100%
No Information Rate	99%	99%	99%
P-Value [Acc > NIR]	57%	57%	57%
Kappa	0%	0%	0%
Mcnemar's Test P-Value	0%	0%	0%
Sensitivity	100%	100%	100%
Specificity	0%	0%	0%
Positive Predictive Value	99%	99%	99%
Negative Predictive Value	NA	NA	NA
Prevalence	99%	99%	99%
Detection Rate	99%	99%	99%
Detection Prevalence	100%	100%	100%
Balanced Accuracy	50%	50%	50%

4.3.2 Performance Analysis for the Dependent Variable: Success_Age

4.3.2.1 Validation Data Results

For the dependent variable success_age, the models—Logistic Regression, SVM, and Random Forest—achieved perfect accuracy (100%) on the validation dataset. The 95% CI for accuracy ranged from 100% to 100%, reflecting absolute consistency in predictions. The No Information Rate (NIR) was 76%, meaning that a naive model predicting the majority class could achieve 76% accuracy. The high accuracy of the models thus represents a significant improvement over the NIR.

Kappa values for Logistic Regression and Random Forest were 100%, indicating perfect agreement between predictions and actual values. For SVM, the Kappa value was 99%, slightly lower due to minor inconsistencies in specificity. Sensitivity and specificity were both 100% for Logistic Regression and Random Forest, while SVM achieved a specificity of 99%. This suggests that SVM occasionally misclassified instances belonging to the minority class.

Metrics such as PPV and NPV were 100% across all models, except for SVM, where PPV was slightly lower at 99%. The Balanced Accuracy for all models was 100%, indicating robust and unbiased performance across both classes. McNemar's test p-value was not applicable for Logistic Regression and Random Forest, as no discrepancies were observed in their classifications. For SVM, the McNemar's test p-value was 0%, reflecting some misclassifications.

Table 13:Results for Success_Age Using Validation Data

Validation Data - Success_Age			
Metrics	Logistic Regression	Support Vector Machine	Random Forest
Accuracy	100%	100%	100%
95% CI Lower	100%	100%	100%
95% CI Upper	100%	100%	100%
No Information Rate	76%	76%	76%
P-Value [Acc > NIR]	0%	0%	0%
Kappa	100%	99%	100%
Mcnemar's Test P-Value	NA	0%	100%
Sensitivity	100%	100%	100%
Specificity	100%	99%	100%
Positive Predictive Value	100%	100%	100%
Negative Predictive Value	100%	100%	100%
Prevalence	76%	76%	76%
Detection Rate	76%	76%	76%
Detection Prevalence	76%	76%	76%
Balanced Accuracy	100%	100%	100%

4.3.2.2 Test Data Results

The test dataset results for success_age were consistent with those of the validation dataset. Accuracy, sensitivity, specificity, PPV, NPV, and Balanced Accuracy remained at 100% for Logistic Regression and Random Forest. SVM again showed minor inconsistencies, with a specificity of 99% and a slightly lower PPV of 97%. Kappa values remained at 100% for Logistic Regression and Random Forest and 98% for SVM. The consistency between validation and test data results reinforces the reliability of the models' performance for this dependent variable.

Table 14:Results for Success_Age Using Test Data

Test Data - Success_Age			
Metrics	Logistic Regression	Support Vector Machine	Random Forest
Accuracy	100%	100%	100%
95% CI Lower	100%	100%	100%
95% CI Upper	100%	100%	100%
No Information Rate	76%	76%	76%
P-Value [Acc > NIR]	0%	0%	0%
Kappa	100%	100%	100%
Mcnemar's Test P-Value	NA	7%	100%
Sensitivity	100%	100%	100%
Specificity	100%	99%	100%
Positive Predictive Value	100%	100%	100%
Negative Predictive Value	100%	100%	100%
Prevalence	76%	76%	76%
Detection Rate	76%	76%	76%
Detection Prevalence	76%	77%	76%
Balanced Accuracy	100%	100%	100%

4.3.3 Performance Analysis for the Dependent Variable: Success_Rounds

4.3.3.1 Validation Data Results

For success_rounds, the models achieved perfect accuracy (100%) on the validation dataset. The 95% CI for accuracy ranged from 100% to 100%, and the NIR was 97%, reflecting the heavily imbalanced nature of the dataset. Despite the high NIR, the models' performance was significantly better, indicating their ability to correctly classify both classes.

Kappa values were 100% for Logistic Regression and Random Forest and 99% for SVM, indicating near-perfect agreement between predictions and actual values. Sensitivity and specificity were 100% across all models, suggesting that they correctly identified instances from both the majority and minority classes. PPV and NPV were also 100% for Logistic Regression and Random Forest, while SVM achieved a slightly lower PPV of 98%.

The Balanced Accuracy for all models was 100%, highlighting their ability to perform equally well across both classes. McNemar's test p-value was not applicable for Logistic Regression and Random Forest, as no discrepancies were observed. For SVM, the McNemar's test p-value was 7%, reflecting minor misclassifications.

Table 15:Results for Success_Rounds Using Validation Data

Validation Data - Success_Rounds			
Metrics	Logistic Regression	Support Vector Machines	Random Forest
Accuracy	100%	100%	100%
95% CI Lower	100%	100%	100%
95% CI Upper	100%	100%	100%
No Information Rate	97%	97%	97%
P-Value [Acc > NIR]	0%	0%	0%
Kappa	100%	99%	100%
Mcnemar's Test P-Value	NA	7%	NA
Sensitivity	100%	100%	100%
Specificity	100%	100%	100%
Positive Predictive Value	100%	98%	100%
Negative Predictive Value	100%	100%	100%
Prevalence	3%	3%	3%
Detection Rate	3%	3%	3%
Detection Prevalence	3%	3%	3%
Balanced Accuracy	100%	100%	100%

4.3.3.2 Test Data Results

The test dataset results for success_rounds were consistent with those of the validation dataset. Accuracy, sensitivity, specificity, PPV, NPV, and Balanced Accuracy remained at 100% for Logistic Regression and Random Forest. SVM showed similar minor inconsistencies, with a specificity of 99% and a PPV of 97%. Kappa values were 100% for Logistic Regression and Random Forest and 98% for SVM. The consistency between validation and test data results suggests that the models generalized well to unseen data for this dependent variable.

Table 16:Results for Success_Rounds Using Test Data

Test Data - Success_Rounds			
Metrics	Logistic Regression	Support Vector Machine	Random Forest
Accuracy	100%	100%	100%
95% CI Lower	100%	100%	100%
95% CI Upper	100%	100%	100%
No Information Rate	97%	97%	97%
P-Value [Acc > NIR]	0%	0%	0%
Kappa	100%	98%	100%
Mcnemar's Test P-Value	NA	25%	NA
Sensitivity	100%	100%	100%
Specificity	100%	100%	100%
Positive Predictive Value	100%	97%	100%
Negative Predictive Value	100%	100%	100%
Prevalence	3%	3%	3%
Detection Rate	3%	3%	3%
Detection Prevalence	3%	3%	3%
Balanced Accuracy	100%	100%	100%

The results highlight both the strengths and limitations of the evaluated models. For success_status, the high accuracy across all models was largely driven by the imbalanced nature of the dataset, where the majority class dominated. The inability of the models to correctly predict the minority class ("not success") was evident from the 0% specificity and 50% Balanced Accuracy. This underscores the need for addressing class imbalance through techniques such as oversampling the minority class, undersampling the majority class, or using more sophisticated algorithms designed to handle imbalanced data.

For success_age and success_rounds, the models performed exceptionally well, achieving perfect accuracy, sensitivity, and specificity in most cases. The slightly lower performance of SVM for these variables suggests that it may be more sensitive to minor variations in the data compared to Logistic Regression and Random Forest. Nonetheless, the high consistency between validation and test data results indicates that the models were not overfitting and were able to generalize well to unseen data.

4.3.4 Further Performance Analysis: F1 Score and Matthew's Correlation Coefficient (MCC)

4.3.4.1 Dependent Variable: Success_Status

For the validation data on "success_status," all three models yield MCC and F1 scores of 0. This result points to an inability of the models to accurately classify success in this dimension. A deeper examination of the data distribution reveals a significant imbalance, with only 222 instances labeled as "not success" compared to 28,629 labeled as "success." Such a skewed dataset can lead to overfitting towards the majority class, which may explain the poor performance metrics. The models default to predicting the majority class, effectively failing to identify true negatives and false positives.

Table 17:Further Results for Success_Status Using Validation Data

Validation Data - Success_Status		
ML Model	MCC	F1 Score
Logistic Regression	0	0
Support Vector Machine	0	0
Random Forest	0	0

4.3.4.2 Dependent Variable: Success_Age

In predicting "success_age," the models exhibit significantly better performance. Logistic Regression and Random Forest both achieve perfect scores (MCC = 1, F1 = 1), while the Support Vector Machine model closely follows with MCC = 0.99 and F1 = 0.996. The improved performance in this dimension may be attributed to a relatively balanced distribution of the labels, with 6,939 instances labeled "not success" and 21,912 labeled "success." This balance provides the models with sufficient data to learn both classes effectively, resulting in superior predictions.

Table 18:Further Results for Success_Age Using Validation Data

Validation Data - Success_Age		
ML Model	MCC	F1 Score
Logistic Regression	1	1
Support Vector Machine	0.99	0.99621206
Random Forest	1	1

4.3.4.3 Dependent Variable: Success_Rounds

When evaluating "success_rounds," Logistic Regression and Random Forest again achieve perfect scores (MCC = 1, F1 = 1), with the Support Vector Machine model slightly lagging behind at MCC = 0.99 and F1 = 1. However, the data frequency reveals a stark imbalance, with 27,957 instances labeled "not success" compared to just 894 labeled "success." Despite this imbalance, the high performance of the models suggests that they may have overfitted to the majority class, raising questions about their generalizability to unseen data.

Table 19: Further Results for success_rounds Using Validation Data

Validation Data - Success_Rounds		
ML Model	MCC	F1 Score
Logistic Regression	1	1
Support Vector Machine	0.99	1
Random Forest	1	1

Chapter 5: Conclusion

5.1 Introduction

This study aimed to predict the success of early-stage African startups using machine learning, focusing on key indicators such as funding rounds, business age, and business operating status The research contributes to the field by identifying key success factors for early stage African startups providing a structured ML framework for evaluating startup sustainability, surpassing the capabilities of simpler probabilistic models

5.1.1 Discussion on the Analysis of Key Success Factors

5.1.1.1 Success_rounds as a Key Success Factor

The **Success_rounds** variable demonstrated the highest predictive power in identifying successful startups. It encapsulates a startup's growth trajectory and reflects investor confidence, as startups that secure multiple rounds of funding tend to have well-defined market opportunities and scalable business models. These startups typically exhibit long-term growth potential. Each additional funding round validates the startup's strategic direction and operational capabilities, attracting more investor support and expanding market share.

The balanced distribution of this variable across the dataset further enhanced the model's performance. A balanced distribution reduced the risk of bias and variance, allowing the model to generalize effectively and distinguish successful startups from those that failed. This was crucial for building a reliable predictive model.

Challenges with Other Variables

1. **Success_status**: This variable, which categorizes startups as successful or unsuccessful based on their operational status, presented significant challenges due to class imbalance. The dataset had a higher proportion of successful startups, leading to inflated accuracy scores. This imbalance hid underlying issues, such as low specificity, and caused the model to misclassify unsuccessful startups as successful. Additionally, the dataset was biased toward startups that survived or thrived, limiting the representation of failed businesses, thus impairing the model's ability to predict failure accurately.
2. **Success_age**: While intuitively relevant, the binary nature of this variable oversimplified the concept of startup longevity. This led to overfitting, where the model performed well on the training data but struggled to generalize to new datasets.

Furthermore, Success_age had a strong correlation with the number of funding rounds, creating redundancy and reinforcing the model's reliance on the Success_rounds variable. This made it harder for the model to identify broader success patterns across a more diverse range of startups.

5.1.1.2 Importance of Selecting Robust Variables

The challenges encountered with Success_status and Success_age highlight the importance of choosing robust and multidimensional variables for building predictive models. The Success_rounds variable, due to its high predictive power and balanced distribution, emerged as the most effective factor for predicting startup success. Unlike the other variables, which were impacted by bias or oversimplification, Success_rounds provided a comprehensive view of a startup's growth and scalability.

The success of the model demonstrated the significance of carefully curating predictive inputs to ensure they holistically capture the dynamics of startup success. Avoiding the pitfalls associated with variables like Success_status and Success_age led to a more reliable and accurate model, capable of making informed predictions.

5.1.2 Discussion of Machine Learning Model Results

The performance of the machine learning models—Logistic Regression, Support Vector Machine (SVM), and Random Forest—on predicting the success of early-stage startups in Africa presented several insights into both the strengths and limitations of the models. This section reflects on the results obtained, emphasizing the key challenges encountered and potential avenues for improving the models in future iterations.

5.1.2.1 Performance on Success_Status: The Challenge of Class Imbalance

One of the key findings from the analysis of success_status was the pronounced issue of class imbalance, which significantly impacted the models' ability to distinguish between successful and unsuccessful startups. Despite achieving high accuracy, the models predominantly predicted the majority class, "successful," at the expense of the "unsuccessful" class. This is a common challenge in predictive modeling, particularly when dealing with imbalanced datasets, where the majority class heavily outnumbers the minority class.

The accuracy scores were misleadingly high due to the models' tendency to classify most instances as "success." While accuracy in the 99% range might initially appear promising, it masked the models' inability to detect the minority class, "not success." The specificity of 0% for all models further underscored this issue, as no successful predictions were made for the

unsuccessful class. This result was consistent across all models, including Random Forest, which is generally known for its ability to handle imbalanced data better than simpler models like Logistic Regression and SVM.

The Balanced Accuracy score was notably low (50%), as it takes both sensitivity and specificity into account. In this case, the absence of predictive power for the minority class led to an inflated sense of performance, making it clear that improvements were necessary in how the models treated class imbalance.

5.1.2.1.1 Improving Performance on Success_Status

Several strategies could enhance model performance on the success_status prediction. First, resampling techniques**,** such as oversampling the minority class using SMOTE (Synthetic Minority Over-sampling Technique) or undersampling the majority class, could help balance the data. These approaches would ensure the model has enough information to recognize patterns within both classes. Another promising avenue is the use of cost-sensitive learning**.** By assigning higher misclassification penalties to the minority class, the model may become more sensitive to instances of startup failure, improving its overall predictive power.

Furthermore, ensemble methods such as boosting (e.g., AdaBoost) or balanced Random Forest could be explored to improve model performance by focusing more on difficult-to-predict minority instances. These techniques could help the model better account for the imbalance and improve predictive accuracy for the minority class.

5.1.2.2 Performance on Success_Age: Robust Predictive Power

In contrast to success_status**,** the model's performance on success_age—which focuses on the longevity of startups—was markedly stronger. The models demonstrated near-perfect accuracy in predicting whether a startup was operational after a specific period, indicating a robust ability to distinguish between startups that survived and those that did not.

This superior performance can be attributed to the more balanced nature of the success_age variable. Unlike success_status, the time period in which a startup operated did not present a highly imbalanced distribution, allowing the models to effectively learn from the data. The near-perfect sensitivity and specificity scores of 100% for both the validation and test datasets indicated that the models could predict both startup success and failure with high confidence.

One of the standout features of the performance on success_age was the perfect Kappa values achieved by both Logistic Regression and Random Forest. The Kappa statistic, which measures

agreement between predicted and observed values, further confirmed that the models were making accurate predictions that matched the true outcomes.

5.1.2.3 Performance on Success_Rounds: Predicting Funding Milestones

The success_rounds variable, which measured the number of funding rounds a startup underwent, also showed promising results, with high accuracy and excellent sensitivity and specificity scores. Although the dataset for success_rounds exhibited a mild class imbalance, it was less severe than that of success_status, and the models managed to identify both classes with significant accuracy.

The Kappa values for Logistic Regression and Random Forest were nearly perfect, suggesting the models were effectively capturing the relationship between funding rounds and startup success. The slight drop in performance for SVM, with a lower specificity and a marginally lower positive predictive value, suggests that SVM may be more sensitive to certain variations in the data, potentially overfitting to minor patterns that are not as relevant for generalization.

5.1.2.4 Evaluating Model Effectiveness: F1 Score and MCC

When examining model effectiveness through F1 Score and Matthew's Correlation Coefficient (MCC), it became evident that the models faced difficulty predicting success_status accurately, especially for the minority class. Both F1 Scores and MCC values were low for success_status, underscoring the challenge of achieving a meaningful classification of the "unsuccessful" class. These metrics, which provide a balanced view of a model's precision and recall, further illustrated the limitations posed by class imbalance.

In contrast, the F1 Scores and MCC for success_age and success_rounds were much higher, reflecting the models' ability to accurately predict startup longevity and funding milestones. The higher scores indicate that the models performed better on these variables, both of which had more clearly defined boundaries for success and failure.

5.2 Limitations and Challenges

The process of predicting startup success using machine learning presents several critical challenges, particularly due to the inherent limitations of the dataset.

5.2.1 Class Imbalance

A major issue lies in the disproportionate representation of classes within the dataset, which impacts both model training and evaluation. This imbalance becomes especially pronounced when considering the criteria used to define success. Based on age and operating status, the dataset predominantly features successful startups, with very few examples of early-stage failures. Conversely, when funding criteria are applied, the dataset highlights a significant number of startups as failed due to insufficient or absent funding. These contrasting dynamics create inconsistencies that complicate the development of reliable predictive models.

The imbalance based on age and operating status poses significant challenges. The overrepresentation of startups that have survived for several years skews the model toward recognizing long-term survival as the primary indicator of success. This bias undermines the model's ability to detect patterns associated with early-stage failures, which are critical for identifying potential risks in startups less than a year old. On the other hand, the funding-related imbalance emphasizes a different kind of skew. Startups that fail to secure funding are disproportionately categorized as unsuccessful, creating the false impression that funding alone determines success. These contrasting imbalances reduce the model's generalizability, as it becomes attuned to specific criteria while overlooking the multifaceted nature of startup success.

The lack of generalizability is exacerbated by conflicting predictions that arise when multiple criteria are considered. For instance, a startup that has operated for several years but received minimal funding might be classified as successful based on its operating status but unsuccessful under funding criteria. These inconsistencies complicate the interpretability of the predictions and limit the model's practical applicability. Addressing this issue requires a more nuanced approach to defining and evaluating success, where all relevant criteria are considered collectively rather than in isolation.

5.2.2 Survivorship Bias

Survivorship bias further compounds the challenges in predicting startup success. This bias stems from the quality of the data collected and its influence on the models. The dataset predominantly represents startups that have already achieved a degree of success, particularly those that have survived for several years or secured substantial funding. Startups that failed early in their lifecycle are significantly underrepresented, leading to an incomplete and distorted view of the startup ecosystem. This survivorship bias skews the models toward emphasizing factors that correlate with long-term survival or funding success while overlooking the warning signs of early-stage failure.

The impact of survivorship bias is particularly evident in how it shapes feature importance. Variables associated with established startups, such as large funding amounts or extended operating history, are often ranked as highly influential. In contrast, critical factors that may predict early-stage failure, such as market fit, execution challenges, or resource constraints, are undervalued or ignored entirely. This bias limits the utility of the models in real-world scenarios, especially for startups that have yet to reach significant milestones. Moreover, the overemphasis on long-term survival reduces the ability to generalize findings to early-stage startups, which are often the most vulnerable to failure.

5.2.3 Insufficient Feature Granularity

The lack of granularity in the dataset further restricts the scope and depth of the analysis. Key variables, particularly those related to funding, lack sufficient detail to capture the complexities of startup success. Funding data, for instance, is often aggregated without distinguishing between different types of investors, funding stages, or the timing of funding rounds. This lack of granularity obscures critical dynamics, such as the influence of strategic investors or the impact of early-stage funding on long-term outcomes. Similarly, features related to the founding team, product characteristics, or market conditions are often missing or poorly defined, limiting the ability to identify nuanced patterns that differentiate successful startups from failed ones.

The oversimplification of success factors due to insufficient feature granularity restricts the depth of the insights that can be derived from the models. For instance, total funding raised is often treated as a singular metric of success, overlooking the importance of how and when the funding was secured. This lack of detail leads to models that fail to capture the diverse factors influencing startup outcomes, such as the quality of the founding team, the competitive

landscape, or the adaptability of the business model. These oversights diminish the accuracy and reliability of the predictions.

Addressing these limitations requires thoughtful approaches to data collection and modeling. Expanding the dataset to include a more balanced representation of both successful and failed startups, particularly those that failed early in their lifecycle, is crucial for reducing class imbalance and mitigating survivorship bias. Enriching the dataset with additional details on funding stages, investor profiles, and market conditions can improve feature granularity and provide a more comprehensive understanding of the factors driving startup success. Incorporating temporal variables, such as the time between funding rounds or the age of the startup at the time of funding, can further enhance the models' ability to capture dynamic relationships.

By addressing these issues, the predictive models can be refined to provide actionable insights that better inform decision-making in the dynamic and uncertain world of startups.

5.3 Future Actions for Reconsideration

To enhance the reliability and predictive power of the models, several targeted actions should be taken to address the dataset's limitations and improve modeling approaches.

5.3.1 Data Enhancements

The dataset must be expanded to include a more diverse and representative range of startups. A significant gap exists in capturing startups that are in earlier stages of development or have failed. Collecting data on these underrepresented groups is essential to reduce survivorship bias and to provide a fuller picture of the startup ecosystem. By integrating data on early failures, the model can better identify the characteristics and warning signs that correlate with unsuccessful ventures, especially in their initial stages.

In addition, the dataset requires greater granularity in its features, particularly with respect to funding. This includes capturing variables such as financial traction achieved, market traction, product subscriptions, detailed investor profiles, funding types (e.g., debt, equity, grants), and stage-specific funding amounts. Incorporating such granular data will enrich the feature set, allow for deeper exploration of funding dynamics, and enhance the model's interpretability. Granularity can also provide insight into how specific funding decisions and patterns influence long-term startup outcomes.

Efforts should also prioritize the dataset's size and diversity. This means increasing data collection efforts, particularly for startups operating in niche industries, less well-represented geographic regions, or unique business models. In particular, focusing on collecting data from startups across African markets will ensure the model captures the nuances of these regions. Gathering this Africa-specific data will help the model address local challenges, such as access to funding, regulatory environments, and market conditions, which may significantly differ from global or developed market standards. Expanding data collection in underrepresented regions will help mitigate bias and improve the model's ability to generalize across diverse scenarios.

5.3.2 Alternative Success Definitions

The definition of startup success should be broadened to capture the multi-dimensional nature of performance. Current definitions, which may rely heavily on binary criteria like survival or funding acquisition, fail to account for the diverse ways in which startups achieve success. Alternative metrics such as revenue growth, customer acquisition rates, profitability milestones, geographic expansion, and brand influence should be considered. Adopting a composite definition of success that incorporates multiple dimensions would provide a more holistic view of startup performance, allowing the model to capture nuances and varying trajectories.

Furthermore, success should be evaluated dynamically over time. For instance, tracking startups' performance across specific time intervals (e.g., year-over-year revenue growth or funding milestones) can provide richer data to inform the model. A more dynamic, time-sensitive evaluation framework would allow for more granular and precise predictions.

5.3.3 Exploring Alternative Models

In light of the current models' varying performance, experimenting with alternative machine learning algorithms is essential. Gradient boosting models, such as XGBoost or LightGBM, have demonstrated success in imbalanced datasets and are known for their ability to model complex interactions among features. These models also offer strong regularization techniques that prevent overfitting, which is especially valuable in datasets with limited granularity.

Deep learning approaches, particularly neural networks, may also provide opportunities to uncover hidden patterns in the data. While these methods require larger datasets and greater

computational resources, they are capable of capturing intricate relationships that traditional algorithms might miss. Recurrent neural networks (RNNs) or temporal models could be particularly useful for analyzing longitudinal data, such as funding patterns over time or changes in market conditions.

Ensemble methods should also be refined to maximize their potential. Blending models with complementary strengths, such as Random Forest for feature importance and Logistic Regression for interpretability, could lead to a more robust solution. Stacking ensemble techniques could also improve predictive performance by combining the outputs of multiple models into a meta-model optimized for overall accuracy.

5.3.4 Threshold Tuning

Optimizing the decision threshold is another critical area for improving classification performance. The current models may rely on default thresholds (e.g., 0.5 for binary classification), which may not adequately capture the trade-off between sensitivity and specificity in imbalanced datasets. Adjusting the threshold to prioritize the detection of minority class instances, such as failed startups, can improve the model's ability to identify early-stage risks.

Threshold tuning should be carried out in conjunction with performance metrics like precision, recall, and F1-score to ensure that the adjustments do not compromise the overall accuracy of the model. For instance, lowering the threshold may increase the recall of failed startups but could lead to more false positives. A balanced approach to threshold selection, guided by specific use-case requirements, will help fine-tune the model's predictions.

5.3.5 Incorporating Temporal Features

Adding temporal features to the dataset is another critical step toward capturing the dynamics of startup success. Time-based attributes, such as the duration between funding rounds or the frequency of funding events, offer unique insights into a startup's performance trajectory. For instance, startups that raise successive funding rounds quickly may demonstrate strong market demand, operational efficiency, or heightened investor confidence. Similarly, the time taken to secure a startup's first funding round after its founding can signal its ability to attract investor attention in its early stages. Temporal features provide a timeline-based perspective that enables the differentiation of startups poised for rapid growth from those with slower

development trajectories. These insights can also be contextualized within industry norms, as certain sectors have distinct funding cycles that align with their developmental timelines.

5.3.6 Segmenting Aggregated Variables

Breaking down aggregated variables into meaningful subcategories can reveal hidden patterns within the dataset. For instance, instead of treating total funding as a single variable, it can be divided into categories such as domestic versus international investors or strategic versus financial backers. This segmentation allows for a more nuanced analysis of the types of investors contributing to startup success. Startups backed by international investors might benefit from global networks, access to new markets, and expertise, while domestic investors may provide localized knowledge and alignment with regional market trends. Similarly, distinguishing between strategic investors, such as corporations with industry expertise, and financial investors, like private equity firms, can uncover varying impacts on startups. Strategic investors might offer operational synergies and long-term guidance, whereas financial investors could focus on maximizing short-term returns. These segmented variables provide insights into the differing roles and priorities of investors and how they shape startups' growth strategies.

5.3.7 Adding Qualitative Features

Qualitative features represent a vital dimension for understanding the factors influencing startup success. Beyond quantitative metrics, qualitative aspects such as the founding team's experience, market conditions, and the competitive landscape play a significant role in shaping outcomes. The founding team's background, including previous entrepreneurial experience, industry expertise, and educational qualifications, can serve as indicators of their ability to navigate challenges and secure funding. For example, founders with prior success in similar ventures may be better equipped to scale their startups and attract investors.

Market conditions, such as regulatory environments, demand trends, and macroeconomic factors, also significantly influence outcomes. Favorable conditions may enable rapid scaling, while adverse factors could impede growth, regardless of the startup's internal capabilities. Additionally, the competitive landscape within the industry provides critical context. Startups in highly competitive sectors may struggle to differentiate themselves and gain market share, while those in emerging or less saturated industries may benefit from first-mover advantages and lower barriers to entry. Incorporating these qualitative variables offers a richer, more

realistic understanding of the complex interplay between internal and external factors that drive startup success, especially in emerging markets such as Africa.

5.4 Contribution of Predictive Machine Learning to Startup Success

In recent years, machine learning has been extensively applied to predict the success of early-stage startups. Various studies have leveraged different datasets, methodologies, and definitions of success to model startup viability. However, many existing studies focus on Western markets, often overlooking the unique challenges faced by African startups. This project aims to fill this gap by employing machine learning models to predict the success of early-stage African startups based on their operating status, number of funding rounds, and business age.

The results show exceptionally high accuracy across logistic regression, support vector machines, and random Forest, with accuracy rates close to 99-100% for predicting startup success. While these results indicate strong predictive capability, they also highlight potential concerns regarding the balance of the dataset and the generalizability of the models. Specifically, the models demonstrate high sensitivity but low specificity, suggesting potential overfitting to the dominant class within the dataset.

A comparison with previous studies shows that the model performance exceeds that of earlier research. For example, Krishna et al. (2016) applied logistic regression to Crunchbase data to predict startup acquisitions and achieved an accuracy significantly lower than these models. Similarly, Pan, Gao, & Luo (2018) employed logistic regression and reported an accuracy of 0.7254, with an F1 score of 0.442. Meanwhile, research by Ünal & Ceasu (2019) achieved an accuracy of 0.941 using random Forest, a figure comparable to these results but still slightly lower.

One of the key contributions of this work lies in its focus on African startups, a relatively underexplored area in predictive modeling. Most prior research has examined success using criteria such as acquisitions, IPOs, or repeated funding rounds, predominantly in Western contexts. This research expands this scope by including African startups and introducing business operating status, business age and success rounds as primary indicators of success.

A critical concern in this study is the imbalance of the dataset, which raises the question of whether simpler probabilistic models could perform equally well. While probabilistic models

may offer reasonable predictive power, machine learning models provide several key advantages. First, machine learning models can capture complex, non-linear relationships between variables that traditional probability models might overlook. Second, they offer scalability and adaptability to additional features, allowing for more nuanced and dynamic predictions. Finally, techniques such as resampling, ensemble learning, and cost-sensitive modeling can be applied to improve performance on imbalanced datasets, ensuring more robust predictions than simple probabilistic approaches.

Additionally, these findings have direct implications for investors, entrepreneurs, and policymakers. Investors can utilize the model to assess the viability of African startups before committing financial resources, reducing the uncertainty associated with early-stage investments. Entrepreneurs can gain insights into key success factors, helping them strategize for long-term growth. Policymakers can leverage these insights to formulate supportive policies that enhance startup ecosystems.

By bridging the gap between existing research and the unique dynamics of the African startup ecosystem, this study offers a novel contribution to the field. Future work should focus on refining model balance, incorporating additional features such as market size and team composition, and testing models on larger, more diverse datasets to improve generalizability. This research lays the foundation for more accurate, region-specific startup success predictions, providing valuable tools for both academic researchers and industry practitioners alike.

Bibliography

African Scalecraft. (n.d.). *The Enabling Ecosystem*. From African Scalecraft: https://www.africanscalecraft.com/stalledacceleration

OECD. (n.d.). *SMEs and Entrepreneurship*. From OECD: https://www.oecd.org/en/topics/policy-issues/smes-and-entrepreneurship.html

Smart Africa. (2020). *Africa's Blueprint for the Development of An ICT Start-Ups and Innovation Ecosystem.*

GSMA. (2023). *The Mobile Economy Sub-Saharan Africa 2023.* From GSMA: https://www.gsma.com/solutions-and-impact/connectivity-for-good/mobile-economy/sub-saharan-africa/

World Bank Group. (2024, January 18). *Digital Transformation Drives Development in Africa .* From World Bank Group: https://www.worldbank.org/en/results/2024/01/18/digital-transformation-drives-development-in-afe-afw-africa#:~:text=Over%20160%20million%20Africans%20gained,payment%20between%202014%20and%202021.

EAVCA. (2023). *The Evolution Of Private Capital In East Africa.* EAVCA.

Exeed College. (2023, October 9). *Key Success Factors for Startup Founders*. From Exeed College: https://exeedcollege.com/blog/key-success-factors-for-startup-founders/#:~:text=Startup%20success%20requires%20more%20than,finances%20and%20manage%20risks%20effectively.

Faster Capital. (2024, June 17). *Co founder and team building: Navigating Co founder Dynamics: Building a Cohesive Team*. From Faster Capital: https://fastercapital.com/content/Co-founder-and-team-building--Navigating-Co-founder-Dynamics--Building-a-Cohesive-Team.html

LinkedIn Community. (2023, May 31). *Tips to Balance Equity and Control of Your Startup.* From LinkedIn: https://www.linkedin.com/advice/0/how-do-you-balance-trade-off-between-giving-up

Stripe. (2024, April 25). *Angel investors vs. venture capitalists: What founders need to know*. From https://stripe.com/zh-my/resources/more/angel-investors-vs-venture-capitalists-what-founders-need-to-know?__hstc=33229921.bbee3a8b127f235496b213a4ef0d3449.1680825600275.1680

825600276.1680825600277.1&__hssc=33229921.1.1680825600278&__hsfp=541859945#:~:text

True Digital. (2017, May 19). *The time lag of innovation.* From True Digital Limited: https://truedigital.co.uk/our-insight/the-time-lag-of-innovation/

Faster Capital. (2024, June 9). *Product differentiation: Standing Out from the Crowd: Strategies for Product Differentiation in Startups.* From Faster Capital: https://fastercapital.com/content/Product-differentiation--Standing-Out-from-the-Crowd--Strategies-for-Product-Differentiation-in-Startups.html

The Guardian. (2020, July 17). *Silicon Valley has deep pockets for African startups – if you're not African*. From The Guardian: https://www.theguardian.com/business/2020/jul/17/african-businesses-black-entrepreneurs-us-investors

Data Driven VC. (2024). *Data-Driven VC Landscape.*

Crunchbase Inc. (2024). From Crunchbase: https://www.crunchbase.com/home

United Nations. (2023, May 10). *Policy Brief.* From United Nations: https://www.un.org/osaa/sites/www.un.org.osaa/files/ads2023_policy_brief_2.pdf

United Nations. (2021). *Young People's Potential, the Key to Africa's Sustainable Development.* From United Nations: https://www.un.org/ohrlls/news/young-people%E2%80%99s-potential-key-africa%E2%80%99s-sustainable-development

BIC Africa. (2021). Early Investment in Africa.

Oxford Business Group. (2022). *African Stock Exchanges Focus Report.* African Securities Exchanges Association.

World Economic Forum. (2022, August 24). *These four countries are leading Africa's start-up scene — here's why*. From World Economic Forum: https://www.weforum.org/agenda/2022/08/africa-start-up-nigeria-egypt-kenya-south-africa/

AVCA. (2024, March). *African Private Capital Activity Report 2023.* From AVCA: https://www.avca.africa/media/hvtdpiei/avca23-20-apca-annual-report-public.pdf

STATISTA. (2024, June). *The 100 largest companies in the world by market capitalization in 2023.* From STATISTA: https://www.statista.com/statistics/263264/top-companies-in-the-world-by-market-capitalization/

AVCA. (2023, April). *Venture Capital in Africa Report 2022.* From AVCA: https://www.avca.africa/media/m3db4yt0/02175-avca-vc-report-2023_4-final-1.pdf

AVCA. (2021, December). *Pension Funds and Private Equity in DECEMBER 2021 Nigeria.* From https://www.avca.africa/media/523mzpsh/avca-penop-pension-funds-study-nigeria-2021.pdf

World Intellectual Property Organization. (2023). *Global Innovation Index*. From World Intellectual Property Organization: https://www.wipo.int/edocs/pubdocs/en/wipo-pub-2000-2023-en-main-report-global-innovation-index-2023-16th-edition.pdf

Kaniawati, K., Sukma, A., & Oktaviani, D. (2024). Leveraging Strategic Orientations In Achieving A Competitive Advantage Among MSMEs: A Cross-Country Marketing Analysis. *Jurnal Ekonomi Bisnis dan Kewirausahaan (JEBIK), 13*(1), 40-66.

Azeem, M., & Khanna, A. (2023, June). A systematic literature review of startup survival and future research agenda. *Journal of Research in Marketing and Entrepreneurship*.

Ferrati, F., & Muffatto, M. (2020, June). Using Crunchbase for Research in Entrepreneurship: Data Content and Structure.

Raj, A. (2023, May 24). *Startup boom in Africa.* From World Business Outlook: https://worldbusinessoutlook.com/startup-boom-in-africa/

Ajayi-Nifise, A., Tula, Asuzu, O. F., Mhlongo, & Ibeh, P. (2024, February 1). The Role Of Government Policy In Fostering Entrepreneurship: A USA And Africa Review. *International Journal of Management & Entrepreneurship Research, 6*.

Womack, J., & Jones, D. (2015). *Lean Solutions: How Companies and Customers Can Create Value and Wealth Together.* Simon & Schuster.

Keller, K. L., & Swaminathan, V. (2020). *Strategic Brand Management Building, Measuring and Managing Brand Equity.* Pearson.

Porter, M. E., & Heppelmann, J. E. (2018). How Smart, Connected Products Are Transforming Companies. *Harvard Business Review*, 96–112.

Nagle, T. T., & Müller, G. (2018). *The Strategy And Tactics Of Pricing A Guide To Growing More Profitably.* Taylor & Francis.

Bradley, C., Hirt, M., & Smit, S. (2018). *Strategy Beyond the Hockey Stick: People, Probabilities, and Big Moves to Beat the Odds.* Wiley.

Teece, D. J. (2018). Business models and dynamic capabilities. *Long Range Planning, 51*(1), 40-49.

Kotler, P., Kartajaya, H., & Setiawan., I. (2017). *MARKETING 4.0 Moving from Traditional to Digital.* John Wiley & Sons.

Chaffey, D., & Ellis-Chadwick, F. (2022). *Digital Marketing: Strategy and Implementation. 8th ed.* Pearson Education.

Guzman, J., & Stern, S. (2015). Innovation economics. Where is Silicon Valley? *Science (New York, N.Y., 347*.

Guzman, J. (2018). Go West Young Firm: The Value of Entrepreneurial Migration for Startups and Their Founders. *SSRN Electronic Journal*.

Kézaia, P. K., & Skalac, A. (2024, May 1). Remarks on the location theories of startups: A case study on the Visegrad countries. *Regional Science Policy & Practice, 16*.

Albourini, F., Ahmad, A., Abuhashesh, M., & Nusairat, N. (2020). The effect of networking behaviors on the success of entrepreneurial startups. *Management Science Letters*.

Daradkeha, M., & Mansoora, W. (2023). The impact of network orientation and entrepreneurial orientation on startup innovation and performance in emerging economies: The moderating role of strategic flexibility. *Journal of Open Innovation*.

Kabatunzi, R. (2022). Entrepreneurial Strategies for the Survival of Small Business Entrepreneurial Strategies for the Survival of Small Business Enterprises in Uganda .

D'Acunto, F., Tate, G., & Yang, L. (2019). Entrepreneurial Teams: Diversity of Experience and Firm Growth.

Mol, E. d. (2019, March 21). What Makes a Successful Startup Teaml.

Hemmert, M., Cross, A. R., Cheng, Y., Kim, J.-J., Kotosak, F. K., Waldenberger, F., & Zheng, L. J. (2019, July). The distinctiveness and diversity of entrepreneurial ecosystems in China, Japan, and South Korea: an exploratory analysis. *Asian Business & Management, 18*(3), 211-247.

Bojadjiev, M. I., Mileva, I., Misoska, A. T., & Vaneva, M. G. (2023, April). Entrepreneurship addendums on Hofstede's dimensions of national culture. *The European Journal of Applied Economics , 20*(1), 122-134.

Crnogaj, K., & Rus, M. (2023). From Start to Scale: Navigating Innovation, Entrepreneurial Ecosystem, and Strategic Evolution. *Administrative Sciences*.

Nims, R. (2023, October 5). *Startup Deserts — How access to startup ecosystems is lacking, and the disparity in different parts of the world.* From Medium: https://medium.com/@robert.nims/startup-deserts-how-access-to-startup-ecosystems-is-lacking-and-the-disparity-in-different-f57d5daaec3c

Kézaia, P. K., & Skalac, A. (2024). Remarks on the location theories of startups: A case study on the Visegrad countries. *The Regional Science Association International*.

Ahluwalia, S., & Kassicie, S. (2024, April 15). Pathways to Success: The Interplay of Industry and Venture Capital Clusters in Entrepreneurial Company Exits. *Journal of Risk and Financial Management, 17*(4).

Yu, L., & Artz, G. M. (2019, October). Does rural entrepreneurship pay? *Small Business Economics, 53*(3), 647-668.

Stam, E. (2015, September 2). Entrepreneurial Ecosystems and Regional Policy: A Sympathetic Critique. *European Planning Studies, 23*.

Sauvage, N., Zeisberger, C., & Varadan, M. (2022, July 28). *Is Corporate Venture Capital Right for Your Startup?* From Harvard Business Review: https://hbr.org/2022/07/is-corporate-venture-capital-right-for-your-startup

Ampong, M. (2024, June 21). *Essential Guide to Burn Rate Management for Startups*. From The Business & Financial Times: https://thebftonline.com/2024/06/21/dr-maxwell-ampong-essential-guide-to-burn-rate-management-for-startups/#google_vignette

Symeonidou, N., Leiponen, A., Autio, E., & Bruneel, J. (2022, July). The origins of capabilities: Resource allocation strategies, capability development, and the performance of new firms. *Journal of Business Venturing, 37*(4).

Mahmudur, R. M. (2023). Operational Efficiency in Start-up Tech Companies.

Mollick, E., & Robb, A. (2016, February 1). Democratizing Innovation and Capital Access: The Role of Crowdfunding. *California Management Review, 58*(2), 72-86.

Cornelius, P., & Gokpinar, B. (2021, March 23). *Crowdfunding Can Deliver More Than Just Money.* From Harvard Business Review: https://hbr.org/2021/03/crowdfunding-can-deliver-more-than-just-money

Sulillari, J. (2023). An analysis of the funding challenges that a start- up has to deal with and the impact that it can have on the future of the company.

Zeng, B. (2023, October 16). Venture Capital Research—Investor Preferences and Success Factors for Startups. *Open Journal of Business and Management, 11*, 2743-2762.

Ganesan, V., Mahalingam, R., Nathan, A., Ware, A., & Weinberg, A. (2023). *Underestimated start-up founders: The untapped opportunity.* McKinsey.

Janaji, S., Ibrahim, F., & Ismail, K. (2021). Startups and Sources of Funding.

Kaplan, S. N., & Lerner, J. (2016). Venture Capital Data: Opportunities And Challenges.

Marullo, C., Casprini, E., Di Minin, A., & Piccaluga, A. (2018). 'Ready for Take-off': How Open Innovation influences startup success.

Pampillo, S. (2023, January 24). *TechDisrupt: Innovation vs. Risk Aversion: Striking the Right Balance in Startups.* From GitHub: https://santiagopampillo.github.io/TechDisrupt/Articles/58-Startups-148-innovation-vs.-risk-aversion.html

Spacenco, A., & Mandari, J. (2020). A Comprehensive Study Into The Emergence Of Scale-Ups: Success Factors And Barriers To Scaling.

Kartika, F. (2024). The Role of Innovation in Startup Success: A Comprehensive Review.

Dennehy, D., Kasraian, L., O'Raghallaigh, P., & Conboy, K. (2016). Product Market Fit Frameworks for Lean Product Development.

Ahmad, I. A., Akagha, O. V., Dawodu, S. O., Obi, O. C., Anyanwu, A. C., & Onwusinkwue, S. (2024). Innovation management in tech start-ups: A review of strategies for growth and sustainability.

Gurbuz, E. (2018). Theory of New Product Development and Its Applications.

Meijer, M. (2019). Strategizing the ideation phase of the startup studio model.

Maurya, A. (2016). *Scaling Lean.* New York: Penguin.

Ries, E. (2011). *The Lean Startup.* New York: Crown Business.

Turman, R. E. (2023, December 20). From TechCabal: https://techcabal.com/2023/12/20/venture-capital-in-africa-investment-trends-and-forecasts/

Felgueiras, M., Batista, F., & Carvalho, J. P. (2020). Creating classification models from textual descriptions of companies using crunchbase. *Communications in Computer and Information Science*, 695-707.

Gompers, P. A., Gornall, W., Kaplan, S. N., & Strebulaev, I. A. (2020). How do venture capitalists make decisions? *Journal of Financial Economics, 135*, 169-190.

Weibl, J., & Hess, T. (2019). Finding the next unicorn: When big data meets venture capital. *Proceedings of the Annual Hawaii International Conference on System Sciences*, (pp. 1075–1084).

Felgueiras, M., Batista, F., & Carvalho, J. (2020). Creating Classification Models from Textual Descriptions of Companies Using Crunchbase.

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

Ferrati, F., Chen, H., & Muffatto, M. (2021). A Deep Learning Model for Startups Evaluation Using Time Series Analysis.

Bonaventura, M., Ciotti, V., Panzarasa, P., Liverani, S., Lacasa, L., & Latora, V. (2020). Predicting success in the worldwide start-up network. *Scientific Reports, 10*.

Gerdin, L. W. (2022). Predicting Success in Early-Stage Start-ups using Founding and Executive Team Characteristics.

Indrianti, Y., Sasmoko, S., Abdinagoro, S. B., & Rahim, R. (2024). A Resilient Startup Leader's Personal Journey: The Role of Entrepreneurial Mindfulness and Ambidextrous Leadership Through Scaling-Up Performance Capacity. *Heliyon, 10*.

Sevilla-Bernardo, J., Sanchez-Robles, B., & Herrador-Alcaide, T. C. (2022). Success Factors of Startups in Research Literature within the Entrepreneurial Ecosystem. *Administrative Sciences*.

Díaz-Santamaría, C., & Bulchand-Gidumal, J. (2021). Econometric Estimation of the Factors that Influence Startup Success. *Sustainability*.

Elsafty, A., Abadir, D., & Shaarawy, A. (2020). How Does the Entrepreneurs' Financial, Human, Social and Psychological Capitals Impact Entrepreneur'S Success? *Business and Management Studies, 6*.

Kim, B., Kim, H., & Jeon, Y. (2018). Critical Success Factors of a Design Startup Business.

Krishna, A., Agrawal, A., & Choudhary, A. (2016). Predicting the Outcome of Startups: Less Failure, More Success. *IEEE 16th International Conference on Data Mining Workshops*.

Gross, B. (2015, June 2). *The single biggest reason why start-ups succeed | Bill Gross | TED*. From TED talks: https://www.youtube.com/watch?v=bNpx7gpSqbY

Patel, D. (2018, March 14). *Leadership*. From Entrepreneur: https://www.entrepreneur.com/leadership/10-personality-traits-of-legendary-entrepreneurs/310026

James, G., Hastie, T., Witten, D., & Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in R.* Springer.

Ross, G., Das, S., Sciro, D., & Raza, H. (2021). CapitalVX: A machine learning model for startup selection and exit prediction. *7*, 94-114.

Gautam, L., & Wattanapongsakorn, N. (2024). Machine Learning Models to Investigate Startup Success in Venture Capital Using Crunchbase Dataset.

Sharchilev, B., Roizner, M., Rumyantsev, A., Ozornin, D., Serdyukov, P., & de Rijke, M. (2018). Web-based Startup Success Prediction.

McKenzie, D., & Sansone, D. (2019). Predicting entrepreneurial success is hard: Evidence from a business plan competition in Nigeria. *Journal of Development Economics, 141*.

Gichohi, B. W. (2023). A Machine learning tool to predict early- stage start-up success in Africa.

Tomy, S., & Pardede, E. (2018). From Uncertainties to Successful Start Ups: A Data Analytic Approach to Predict Success in Technological Entrepreneurship.

Pan, C., Gao, Y., & Luo, Y. (2018). Machine Learning Prediction of Companies' Business Success.

Żbikowski, K., & Antosiuk, P. (2021). A machine learning, bias-free approach for predicting business success using Crunchbase data .

Baskoro, H., Prabowo, H., Meyliana, & Gaol, F. L. (2022). Predicting Startup Success, a Literature Review.

Kusumaningtyas, A., Bolo, E., Istianah, Chua, S., Wiratama, M., & Tirdasari, N. L. (2021). Why Start-ups Fail: Cases, Challenges, and Solutions. *Advances in Economics, Business and Management Research, 198*.

Mehmeti, V., & Musabelli, E. (2024). Start-Ups: Importance and Role in the Economy. *Interdisciplinary Journal of Research and Development, 11*(2).

Ünal, C., & Ceasu, I. (2019). A Machine Learning Approach Towards Startup Success Prediction.

Gangwani, D., & Zhu, X. (2024). Modeling and prediction of business success: a survey.

Potanin, M., Chertok, A., Zorin, K., & Shtabtsovsk, C. (2023). Startup Success Prediction And Vc Portfolio Simulation Using Crunchbase Data.

Shah, V., & Mcgaugh, M. (2019). Predicting the success of a startup company.

Cholil, S. R., Gernowo, R., Widodo, C. E., Wibowo, A., Warsito, B., & Hirzan, A. M. (2024). Predicting Startup Success Using Tree-Based Machine Learning Algorithms.

Huang, L., & Pearce, J. L. (2015). Managing the unknowable: The effectiveness of early-stage investor gut feel in entrepreneurial investment decisions. 634–670.

Attenberg, J., Ipeirotis, P., & Provost, F. (2015). Beat the Machine: Challenging Humans to Find a Predictive Model's "Unknown Unknowns". *Journal of Data and Information Quality* , 1-15.

Dellermann, D., Lipusch, N., Ebel, P., Popp, K. M., & Leimeister, J. M. (2017). Finding the Unicorn: Predicting Early Stage Startup Success through a Hybrid Intelligence Method.

Te, Y.-F., Wieland, M., Frey, M., Pyatigorskaya, A., Schiffer, P., & Grabner, H. (2022, August 31). Predicting the Success of Startups using Crunchbase and LinkedIn Data.

Arroyo, J., Corea, F., Jimenez-Diaz, G., & Recio-Garcia, J. (2019). Assessment of machine learning performance for decision support in venture capital investments.

Shi, Y., Eremina, E., & Long, W. (2023, December 13). Machine learning models for early-stage investment decision making in startups.

Bai, S., & Zhao, Y. (2021). Startup Investment Decision Support: Application of Venture Capital Scorecards Using Machine Learning Approaches.

Misra, A., Jat, D., & Mishra, D. (2021). Machine Intelligence for Predicting New Start-ups Success: A Survey.

Vasquez, E., Santisteban, J., & Mauricio, D. (2023). Predicting the Success of a Startup in Information Technology Through Machine Learning.

Varma, S. (2021). Machine Learning based Outcome Prediction of New Ventures: A review.

Piskunova, O., Ligonenko, L., Klochko, R., Frolova, T., & Bilyk, T. (2021). Applying Machine Learning Approach to Start-up Success Prediction.

Thirupathi, A., Alhanai, T., & Ghassemi, M. M. (2022). A machine learning approach to detect early signs of startup success.

Bangdiwala, M., Mehta, Y., Agrawal, S., & Ghane, S. (2022). *Predicting Success Rate of Startups using Machine Learning Algorithms.*

Christodoulou, I., Rizomyliotis, I., Konstantoulaki, K., Alfiero, S., & Has, S. (2024). Investigating the key success factors within business models that facilitate long-term value creation for sustainability-focused start-ups. *Business Ethics, the Environment & Responsibility*, 1–15.

Obonyo, E., & Zeisberger , C. (2024, July 4). *Entrepreneurship.* From INSEAD Knowledge: https://knowledge.insead.edu/entrepreneurship/how-africa-can-embrace-venture-capital

Ekefre, M. (2023, May). Start-up Success in Rwanda Whitepaper.

Gutterman, A. (2022). *Seed Capital.*

Amokeoja, O. (2024). *Current Affairs*. From Forbes Africa: https://www.forbesafrica.com/current-affairs/2024/04/21/africas-startup-funding-sees-decline-amidst-global-trends-experts-speak/#

Kabonga, I. (2016). Dependency Theory And Donor Aid: A Critical Analysis.

Nzibonera, E., & Waggumbulizi, I. (2020, May 31). Loans and growth of small-scale enterprises in Uganda: A case study of Kampala Central business area. *African Journal of Business Management, 14*.

Kuwonu, F. (2017). *Africa Renewal*. From United Nations: https://www.un.org/africarenewal/magazine/august-november-2017/alternative-financing-strategies-boost-small-businesses-africa

Oturu, D. (2023, July 17). *M&A set to become the norm in the African startup ecosystem*. From African Business: https://african.business/2023/07/long-reads/ma-set-become-the-norm-in-the-african-startup-ecosystem

Nzekwe, H. (2022, April 25). *For Kenyan Fintech Startups, Safaricom Is Both A Blessing And A Curse*. From Wee Tracker: https://weetracker.com/2022/04/25/safaricom-and-kenyan-fintech-startups/

Kato, A. I., & Chiloane-Tsoka, G. E. (2022). Venture Capital Financing as A Driver for Entrepreneurship Development in Africa. A Literature Review and Future Research Agenda. *International Journal of Entrepreneurship and Business Development, 5*.

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science, 2*(160).