A Machine Learning Startup Success Prediction Model

By

Wanyoike Eunice Waruguru

094306

A Dissertation Submitted in Partial Fulfillment of the
Requirements for the Degree of Masters in Data Science and Analytics at Strathmore University

Strathmore Institute of Mathematical Sciences

Strathmore University

Nairobi, Kenya

June, 2025

Declaration and Approval

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 dissertation contains no material previously published or written by another person except where due reference is made in the dissertation itself.

Student's Name: Wanyoike Eunice Waruguru

Sign:  Date: 6/4/2025

Approval

The dissertation of Wanyoike Eunice Waruguru was reviewed and approved for examination by the following:

Dr. Betsy Muriithi

Strathmore Institute of Mathematical Sciences,

Strathmore University

Dr. Godfrey Madigu,
Dean, Strathmore Institute of Mathematical Sciences,
Strathmore University

Prof. Bernard Shibwabo,
Director of Graduate Studies,
Strathmore University

Abstract

Predicting the success of early-stage startups is a complex challenge with major implications for investment and entrepreneurial strategy. This study developed a machine learning-based prediction model using a dataset of 765 startups from Kenya and global sources. Key predictors of success included funding amount, age of company, and number of investors. Five models Logistic Regression, SVM, Random Forest, K-Nearest Neighbors (KNN), and Gradient Boosting were evaluated using accuracy, precision, recall and F1 score. The Gradient Boosting model achieved the highest performance of 98% training accuracy and 96% test accuracy, with an F1-score of 96% and AUC-ROC of 97%, making it the most effective predictor. A user-friendly interface was built to allow users to input data and receive real-time predictions along with strategic recommendations. The system serves as a practical tool for entrepreneurs and investors, enabling data-driven decisions based on key success factors.

Table of Contents

List of Figures

List of Tables

List of Abbreviations and Acronyms

AI – Artificial Intelligence

DT – Decision Tree

GDP – Gross Domestic Product

IPO – Initial Public Offering

KNN – K-Nearest Neighbors

KPI – Key Performance Indicator

ML – Machine Learning

R&D – Research and Development

RF – Random Forest

ROI – Return on Investment

SME – Small and Medium-sized Enterprises

SVM – Support Vector Machine

SWOT – Strengths, Weaknesses, Opportunities, and Threats

VC – Venture Capital

Definition of Terms

Competitive Advantage – A unique strength that differentiates a startup from competitors, such as proprietary technology, strong branding, or cost efficiency (Porter, 1985).

Economic Performance Indicators – Quantitative financial metrics including revenue growth, profitability, cash flow, and market valuation, used to measure a firm's financial health and success (Gimeno et al., 1997).

Machine Learning (ML) – A branch of Artificial Intelligence (AI) that enables systems to learn from data and make predictions or decisions without being explicitly programmed (Samuel, 1959).

Non-Economic Performance Indicators – Qualitative metrics such as customer satisfaction, brand reputation, and social or environmental impact, which also influence long-term survival and success (Gimeno et al., 1997).

Pivot – A strategic redirection in a startup's business model, product offering, or market focus aimed at improving growth potential and achieving product-market fit (Ries, 2011).

Startup – A newly established company that focuses on innovation, solving real-world problems, and developing scalable business models (Katila et al., 2012; Skawińska & Zalewski, 2020).

Startup Incubator/Accelerator – Structured programs that support early-stage startups through mentorship, funding, office space, and business resources to accelerate growth (Pauwels et al., 2016).

SWOT Analysis – A strategic planning tool used to evaluate an organization's Strengths, Weaknesses, Opportunities, and Threats (Humphrey, 2005).

Unicorn – A privately-held startup with a valuation of over $1 billion, representing a significant achievement in the startup ecosystem (Lee, 2013).

Acknowledgements

I extend my deepest gratitude to God for bestowing upon me the grace to successfully navigate this academic journey.

My sincere appreciation goes to my supervisor, Dr. Betsy Muriithi, for her professional guidance, unwavering support, and consistent encouragement throughout the entire research process. Her expertise, insightful feedback, and constructive criticism have been instrumental in shaping the direction and quality of this work.

I express my gratitude to the lecturers at Strathmore University for their contribution to the body of knowledge in my field of study, providing valuable resources, inspiration, and intellectual stimulation.

Finally, I want to thank my parents, Mr. and Mrs. Wanyoike, and my brother, Ian Wanyoike, for their endless support and love.

Chapter 1: Introduction

1.1 Introduction

Startups are becoming an essential factor in economic policies of all developed and emerging economies around the world.They are very important in shaping a nation's economic revolution. A nation makes investments and encourages start-up businesses to assist the economic and industrial revolution (Reis, 2011). The success of a business venture should make its founders and investors proud. It is also strongly associated with monetary gain. Both founders and investors are actively seeking tools, methods, and advice that will give them a competitive advantage. It is debatable whether being a successful entrepreneur is linked to inherent abilities or whether those abilities can be learned. External factors such as the industry in which the company operates, the area in which the headquarters is located, or the level of competition in a specific sector and its sub-sector are also difficult to quantify (Prajogo, 2016). However, starting and supporting the correct start-up is just as crucial, particularly for founders and venture capitalists. Taking the current state of the market and the general complexity of managing startups into consideration, it is important to come up with an approach that will help reduce the risks of making incorrect managerial decisions, as any mistake can be very costly not just for startup management, but also for investors. Machine learning's abilities can be used to help investors choose which start-ups to support, allowing money to be put in the correct start-ups. It can also give insights to entrepreneurs on factors to consider when running their start ups to improve their performance. The application of machine learning is already improving areas of our society. From the health care system, where segmentation and predictive analysis can be used to identify different types of treatment for a patient or even diagnose them Raghupathi and Raghupathi (2014), to marketing personalized service, where businesses benefit from knowing as much as possible about their customers in order to create consumer experience all around. Fraud detection, financial services, insurance, and even self - driving are all industries that will create value in the short to medium term by utilizing machine learning (Marr, 2016). It is entirely feasible to provide similar benefits to investors in start-ups; by informing these players about which start-ups are closer to a successful event in the near future, they could better determine where to place their chips and earn higher returns from their investments.

1.2 Problem Statement

The surge of motivated entrepreneurs in recent years is undeniable. However, while many startups initially disrupt established businesses and experience rapid early growth, a significant number fail to reach maturity. A core challenge lies in achieving optimal sales and maximizing return on investment through efficient, targeted marketing, all while building and maintaining customer trust. Without a robust decision-making strategy, startup performance often suffers. This study aimed to provide entrepreneurs and startup businesses with practical tools, methods, and recommendations to enhance their competitive advantage. Specifically, this research described the design and execution of a machine learning model for business success prediction. The model was intended for use by both investors and entrepreneurs to assess the probability of a startup's success based on selected criteria.

1.3 General objective

To construct a predictive model for startup success, employing machine learning techniques to identify key features influencing venture outcomes and provide insights into the factors contributing to success or failure.

1.4 Specific objectives

i. To identify and analyze key factors influencing startup success in Kenya.
ii. To collect and curate relevant data from a representative sample of Kenyan startups.
iii. To develop and validate a predictive model using machine learning tailored to the Kenyan market.

1.5 Research questions

i. What specific organizational and environmental factors predict startup success in diverse industries within the first years of operation?
ii. How do different machine learning prediction tools such as Logistic Regression and Support Vector Machines compare in their predictive accuracy for startup success across various industry sectors?

1.6 Scope of the study

This study focused on developing a machine learning model to predict startup success using data primarily from approximately 450 Kenyan startups. This dataset formed the core of the research, offering context-specific insights into local entrepreneurial dynamics, such as founding year, industry sector, funding history, and company status.

To assess the model's generalizability across diverse startup ecosystems, a secondary dataset titled 'Startup Data' was sourced from Kaggle. This dataset contained records for 465 startups from various global regions and shared structural features with the Kenyan dataset. Its validity was supported by its prior application in machine learning research on startup prediction, including studies by Soni and Jain (2023), Prasetyo and Fadilah (2022), and Rani and Patil (2023), who applied dimensionality reduction, classification algorithms, and deep learning techniques respectively. The inclusion of this dataset in the current study enabled a comparative analysis between regional and global startup trends, contributing to the broader applicability of the developed model.

1.7 Justification

The study was intended to be useful to businesses and key stakeholders such as investors and managers. Economic analysts and scholars were also expected to gain insights from the research. The model provided was aimed at supporting entrepreneurs and financiers who invest substantial resources and seek to avoid potential losses. It was also expected to assist investors in identifying the types of startups worth investing in. Ultimately, the success of startups would contribute to a significant boost in the overall economy.

1.8 Limitations

The study faced several limitations, including a dataset that did not fully represent the diversity of the broader startup ecosystem, potentially causing sampling bias; reliance on publicly available data that was sometimes incomplete or inconsistent, affecting model accuracy; and the challenge of capturing the rapidly evolving factors influencing startup success over time. Additionally, the complexity of machine learning models occasionally led to overfitting, with strong performance on training data but poorer results on unseen data.

1.9 Chapter summary

In a nutshell, this chapter introduced the main aspects of the study, which focused on predicting whether a startup succeeds or fails. The information provided highlighted various factors that affect the growth stages of startups. In the subsequent chapters, the study explored relevant literature and conducted data collection and analysis to inform the development of the predictive model.

Chapter 2: Literature Review

2.1 Introduction

This chapter identified literature related to the study topic. In addition, comparisons with previous studies were considered to further enhance the dissertation content, making it distinctive and innovative for startup stakeholders in Kenya.

2.2 Startup Success

Defining a startup involves recognizing it as a nascent, independent entity focused on attracting talent, boosting sales, and establishing a compelling business model (Skawińska & Zalewski, 2020). Startups are characterized by their inventive and creative nature, engaging in research and development to solve real-world problems (Katila et al., 2012). They venture into uncharted markets with novel products, which inherently makes them high-risk endeavors due to the uncertainty of market acceptance and the potential need for significant adaptations (Reis, 2011).

Performance evaluation for startups can involve metrics like patent applications, R&D personnel, and management ownership (Jayasekara et al., 2020). A key indicator of success is company valuation, particularly for privately-owned companies that achieve the "unicorn" statusa term coined by venture capitalist Lee (2013) to denote exceptional achievement.

Contemporary perspectives on business success emphasize the ability to identify both current and future successful ventures and their contributing factors (Fried & Tauer, 2015). For early-stage or pre-revenue firms, non-economic performance significantly influences survival. Differences in survival rates among businesses with similar economic performance suggest that minimum viable economic performance levels vary, or that non-economic factors impact business longevity. Gimeno et al. (1997) highlighted that business survival hinges on both economic and non-economic performance indicators, as well as meeting minimum acceptable performance thresholds. Therefore, understanding the drivers of both economic and non-economic growth is crucial for assessing a business's likelihood of success.

2.3 Factors Influencing Startup Success

The success of a startup can be characterized by its potential for early global expansion and an entrepreneur's capacity to generate valuable and profitable opportunities from their business strategy (Kamal & Sabani, 2021). Alternatively, success is often defined by two outcomes: an Initial Public Offering (IPO), allowing public trading of shares, or an acquisition/merger, providing immediate cash to investors (Guo et al., 2015).

Innovation commitment and consistent funding are critical for any startup's success (Okrah et al., 2018). The inherent high risk of startups makes securing investor confidence particularly challenging. Current views on business success, as noted by Fried & Tauer (2015), stress the importance of identifying successful ventures and their drivers. Beyond financial metrics, understanding how non-economic performance impacts survival is also vital for startups. The existence of varying mortality rates among businesses with similar economic performance implies that either minimum viable economic performance differs, or non-economic factors play a role in business failure. Gimeno et al. (1997) asserted that successful businesses effectively leverage factors influencing both economic and non-economic performance, alongside maintaining minimal viable performance levels. Thus, investigating the drivers of both economic and non-economic performance is essential due to their impact on a business's success probability.

Various diagnostic approaches have been developed to predict company success, historically over-relying on methods like correlation, induction-deduction, and graphical scoring (Borlea & Achim, 2004). Early models often aimed to predict impending bankruptcy. The effectiveness of these models relies heavily on the expertise of the financial examiner in selecting appropriate indicators and assigning scores, facilitating objective qualitative and quantitative assessment. Many financial institutions and government bodies utilize deterministic models, often within a SWOT framework, to conduct up-to-date company appraisals for decision-making. These models can proactively identify financial distress, potentially preventing bankruptcy if issues are caught early. However, a drawback is that if problems are not identified promptly, the business's survival chances diminish. More recently, Żbikowski and Antosiuk (2021) developed a model for international companies to assess performance and risk, including insolvency. Their research also highlighted

how entrepreneurial initiative links to business prosperity, with individual differences like gender, education, and having an entrepreneurial guardian influencing entrepreneurial expectations.

2.4 Key Indicators for Predicting Startup Success

Within the startup ecosystem, significant considerations for success include the value of business plans, project scope, sustainable business models, employee profiles, theoretical and educational support, the role of incubators/accelerators, and attitudes toward investors (Geissdoerfer, Vladimirova & Evans, 2018). The technology, applications, tools, and programming languages employed by entrepreneurs are also crucial. Maurya (2012) suggests that a startup's initial operational success depends on its product idea, development speed, responsiveness to target audiences, quality measurement, ability to draw conclusions for product optimization, and rapid progression to subsequent development phases. Skawińska and Zalewski (2020) emphasize that competitive advantage defines a company's success and is vital for growth. They propose categorizing businesses based on their source of competitive advantage. While strategic management theories highlight both tangible and intangible assets as contributors to competitive advantage, the focus on cost and quality has become a baseline requirement for competitiveness since the 1990s. These benefits often stem from material elements, which don't fully explain startup development or their competitive performance, as startups largely depend on intangible assets in their early stages.

Financial challenges, such as incorrect product pricing, inaccurate cost estimates, or insufficient development funds, are primary drivers of startup failure (Bednár & Tarišková, 2017). A lack of market demand due to inadequate quality checks is another significant issue. An ineffective team, unable to craft an ideal business plan, also contributes to failure. Fairlie and Robb (2009) identified three key factors explaining success or failure variance between startups: founders' prior job experience, education level, and initial capital. Kim et al. (2018) note that attracting clients but consistently selling only one product can lead to failure. Startups must continuously innovate, introduce new products, respond to market demands, and adapt to environmental changes. Meticulously crafting business plans and investor-centric financial materials is also essential for securing timely funding and support. Companies that survive often demonstrate superior

performance, being larger, more productive, cash-rich, and less indebted than those that fail (Callén et al., 2020).

Successful entrepreneurs are continuous learners (Eesley, Blank & Bishara, 2019). Businesses benefiting from mentors, effective metric tracking, and insights from key influencers secure significantly more funding (7x) and achieve greater user growth (3.5x). Startups that pivot once or twice receive 2.5 times more funding, experience 3.6 times greater user growth, and have a 52% lower rate of premature scaling compared to those that pivot more or not at all. Investors sometimes overfund firms that haven't achieved product-market fit or invest in teams lacking technical co-founders or solo founders, despite data suggesting lower success rates for such teams.

2.5 Machine Learning Tools for Predicting Startup Success

Identifying factors influencing startup success remains a crucial research area, given that most startups fail within their first three years (Stuart & Abetti, 1990). The high risk associated with newly funded startups makes traditional success measurement such as reaching a specific ROI threshold over a long period difficult, as most operate for less than two years. Consequently, "initial success" is often considered, combining subjective and quantitative indicators. Quantitative initial success might involve metrics like employee and sales growth, increased revenues, or sales per employee. Qualitative assessments consider survival chances, funding acquisition ability, societal contributions, and employee satisfaction (Stuart & Abetti, 1990). Bell (2022) highlights how many startups are exploring the benefits of integrating machine learning (ML) with big data to provide professionals with actionable insights for better decision-making.

Buss and Finn (1987) explored combining predictions from diverse forecasters using market expectations, examining markets with human forecasters, artificial neural networks, and a hybrid of both. Forecasting quality was assessed using metrics like accuracy, Sharpe ratios (to balance precision against volatility of forecast errors), and other unique criteria. Choksi and Bhatt (2021), Afolabi et al. (2019), and Ifunaya, Ojo and Chinonye (2019) propose an online platform for entrepreneurs and startups to evaluate their businesses and develop business cases. Their work resulted in a 32-criterion system for assessing business cases and venture proposals, with standards and weights based on venture capitalists' investment criteria for markets and products. This review can be performed by self-evaluation, an entrepreneurial team, or an external auditor. A proactive

personality scale was developed as a predictor of entrepreneurial intentions (Buss & Finn, 1987). Nagar and Malone (2011) researched business predictions using prediction markets that integrated human and machine intelligence. Hybrid human-and-agent markets offer an optimal Sharpe ratio trade-off between accuracy and variability of prediction errors, and provide a better balance between good and bad decisions using the ROC technique. Islam and Habib (2015) developed a data mining approach to predict potential business categories for retail banking lending.

Misra (2021) points out that prediction and recommendation models are built using various supervised and unsupervised learning techniques, largely influenced by traditional machine learning algorithms and regression analysis. In some cases, supervised learning and NLP-based algorithms also aid in determining startup success or failure.

Bento (2018) aimed to develop a model for classifying successful businesses. Their model, using Random Forests, accurately categorizes all successful startups in the dataset (Precision). Random Forests provided a fast, understandable, and effective model, outperforming Support Vector Machines and Logistic Regression due to its suitability for the dataset's size and structure where a linear relationship is expected.

Kamal and Sabani (2021) utilized three ML models, all demonstrating better predictive power than random chance. Random Forest showed the highest performance on unseen data, accurately predicting successful and failed companies with 79% accuracy. Logistic regression and K-Nearest Neighbor followed closely with 65% and 59% accuracy, respectively. Pan, Gao and Luo (2015) found that KNN models perform well, but Random Forests are generally more accurate, primarily due to their higher true positive rate when predicting more negative examples. Afolabi et al. (2019) used UML tools, specifically use case and sequence diagrams, for the design of their business success diagnosis system, implemented on the Java NetBeans framework. They employed a Bayesian network (Naive Bayes), which uses nodes for chance variables and arcs for qualitative relationships, with conditional probability distributions estimating variable dependencies.

2.6 Conceptual Framework

This study employs a conceptual framework that examines startup success in Kenya as influenced by both internal and external factors. Internal factors, such as team composition, business model, and financial management, interact with external influences, including access to funding, market competition, and regulatory conditions. These elements collectively shape a startup's operational, financial, and overall sustainability. Moderating variables, such as industry sector, geographic location, and founder characteristics, further impact these relationships. To analyze these dynamics, machine learning models will be utilized to develop a predictive framework specifically tailored to Kenya's entrepreneurial ecosystem. This structured approach will guide data collection, analysis, and interpretation, offering valuable insights into the factors driving startup success.

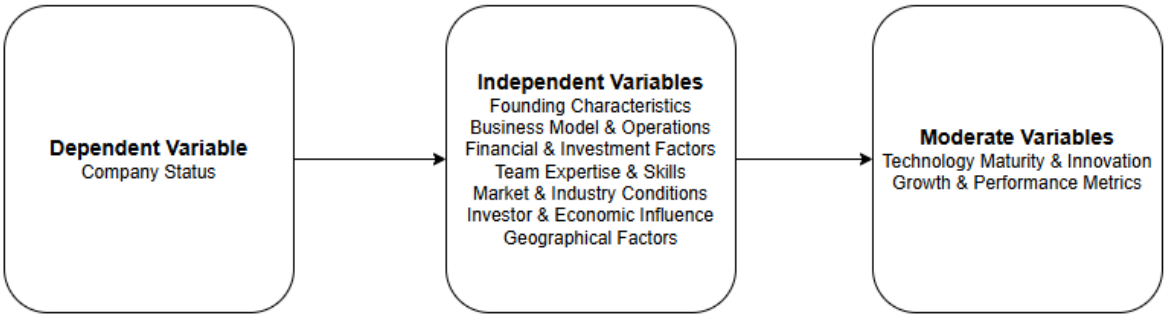


Figure 2.1 :Conceptual Framework: Factors Influencing Startup Success in Kenya

2.7 Chapter Summary

In this chapter, a review of literature was conducted on topics related to startup success, determinants of startup success, and machine learning tools used to predict startup success. The chapter also outlined the frameworks of the study and established the operationalized variables that were used as the basis for the research.

Chapter 3: Methodology

3.1 Introduction

The goal of this study was to come up with a predictive model for startup success. Data was collected, analyzed, and evaluated since gathering data is crucial to forming the predictive model. Weller (1998) found that theories and facts were developed in tandem to create the methods that

underpin the phenomena. Through observation or measurement, theories help to define the characteristics of the models. Given this information, the study's objective was to thoroughly describe the research design that was used, the data collection techniques, the target population and processes, the research quality, and lastly the ethical considerations.

3.2 Research Design

This study adopted a predictive research design, chosen specifically to achieve its primary goal: forecasting the success or failure of startups using machine learning. A predictive design is ideal for building models that anticipate future outcomes based on historical and current data. In this instance, the aim was to develop a startup success prediction model utilizing both newly collected primary data and existing secondary data.

The decision to use a predictive approach was driven by the research problem's objective, which extended beyond merely describing or exploring startup performance to actively predicting future scenarios. This aligns perfectly with the capabilities of predictive analytics and data mining, technologies designed to uncover patterns and relationships in data that enable future projections.

By implementing this design, the study facilitated the application of machine learning algorithms to real-world startup data. This yielded valuable insights that could support incubators, investors, and entrepreneurs in making more informed decisions. Furthermore, the design inherently supported quantitative methods, particularly the analysis of numerical data and the creation of algorithmic models, which suited the study's structured and outcome-oriented nature.

The following sections will detail the specific procedures for data collection, the sources of data, and the analytical tools employed within this predictive research framework.

3.3 Target population and sampling procedures

General population, Asiamah et al. (2017), is presumably what is generally known and determined by scientists, although it is senseless without being indicated close by the target and accessible population. Target population is along these lines the entire course of action of units for which the review information is to be used to make acceptances (Cox, 2011). Along these lines, the target population portrays those units for which the disclosures of the study are expected to total up. This

study targeted approximately 450 startups in Kenya. The respondents were founders from each of the target startups.

This research utilized non-probability sampling, a method where not every member of the population has an equal or known chance of selection (Ogula, 2005). Unlike probability sampling, which relies on random selection, non-probability sampling often involves researcher judgment or convenience. In this study, the selection criteria for participating companies specifically focused on startups that had successfully completed an incubation program.

3.4 Data collection methodology

This study employed a quantitative data collection strategy combining primary and secondary sources to support robust model development. Kenyan startup data was gathered through a structured questionnaire developed using Google Forms. The instrument captured critical variables such as company status, industry sector, founding year, number of employees, and funding history. The questionnaire was refined through pilot testing for clarity and alignment with the study objectives. Data collection was conducted from March to June 2023 and distributed via email, LinkedIn, WhatsApp, and startup community platforms. Target respondents were mainly founders and decision-makers, sourced from iBiz Africa's incubation database, LinkedIn, and public startup directories. Approximately 450 startups were targeted within the Kenyan ecosystem.

To complement the primary dataset and facilitate external validation, a secondary dataset titled "Startup Data" was obtained from Kaggle in CSV format. Originally compiled from Crunchbase in 2021, it included 465 startup records with variables such as industry, founding date, funding rounds, and company status. From the primary data collection effort, a total of 300 complete responses were received from Kenyan startups. Combined with the 465 global records from the Kaggle dataset, a total of 765 entries were available for cleaning and preprocessing. This comprehensive dataset formed the basis for building, training, and evaluating the machine learning models presented in the subsequent sections.

3.5 Data cleaning and preprocessing

The collected data underwent a thorough cleaning and preprocessing procedure to ensure its suitability for analysis and predictive model development. This involved identifying and removing redundant, irrelevant, and duplicate records to enhance data quality and reliability. Records with missing values in critical fields, such as company status, founding date, and funding information, were either removed entirely or, if the missing data was minimal, imputed using appropriate methods. For instance, numerical fields with few missing values were filled with the mean, while categorical variables were completed using the mode.

Outliers were detected using the interquartile range (IQR) method, which helped identify data points significantly deviating from the majority. Each identified outlier was assessed within the context of the dataset, leading to decisions to either remove or transform the values based on their relevance and potential impact on the model.

Beyond cleaning, the dataset was transformed to prepare it for machine learning algorithms. Categorical variables, such as industry type and company status, were converted into numerical formats using one-hot encoding to ensure compatibility with modeling tools. Numerical features, like funding amounts and employee counts, were normalized using min-max scaling to bring them to a comparable range, thereby reducing potential bias in algorithm performance. Date fields, particularly founding dates, were converted into meaningful numerical representations, such as the startup's age in years, to serve as effective inputs for the prediction models.

Following preprocessing, the refined dataset was divided into training and testing subsets to facilitate robust model development and validation. This clean and transformed data was then used to train and evaluate various machine learning algorithms aimed at predicting startup success or failure.

The process of predicting startup success using a machine learning technique is illustrated below.

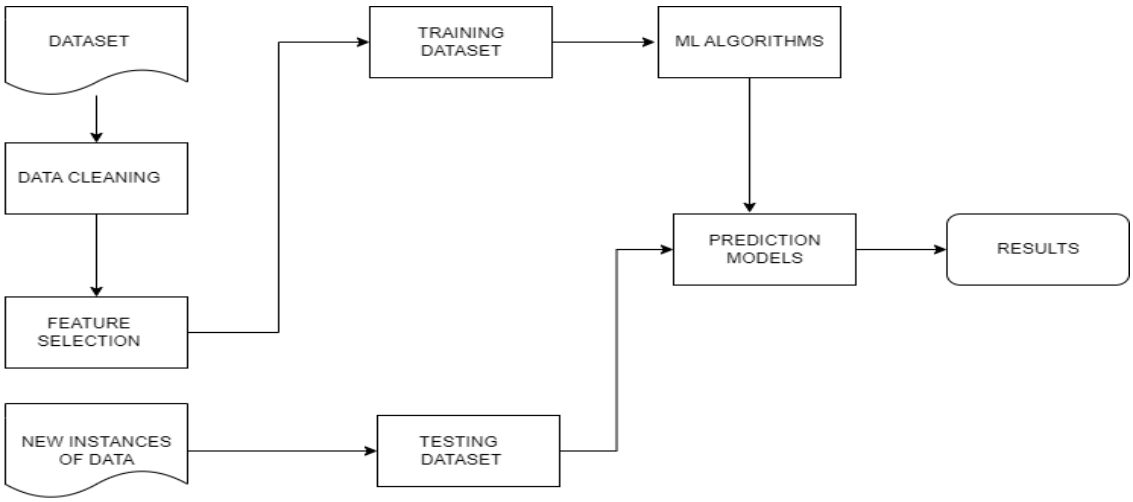


Figure 3.1 : Data Transformation

3.6 Feature Selection

After data cleaning and preprocessing, feature selection was performed to identify the most relevant variables for predicting startup success. This step was essential in reducing model complexity, improving performance, and mitigating the risk of overfitting by ensuring that only the most informative predictors were used in the analysis. A combination of statistical and model-based techniques was applied to evaluate the relevance of each feature. For numerical variables, independent samples t-tests were conducted to determine whether the mean values differed significantly between successful and failed startups. For categorical features, the chi-square test of independence was used to assess the strength of association with the target outcome. Additionally, Pearson's correlation coefficient was employed to explore relationships among numerical variables, helping to understand the strength and direction of associations and identify potential multicollinearity.

Beyond statistical tests, model-based approaches were also used to assess feature importance. In particular, tree-based models such as Random Forest were employed during the modeling phase to generate feature importance scores. These scores highlighted the variables that contributed most to predicting startup success, guiding the refinement of the feature set by removing those with minimal predictive power. To further enhance model interpretability, the SHAP (SHapley Additive exPlanations) framework was integrated into the analysis. SHAP values were used to explain individual predictions and provide a comprehensive view of feature importance for complex

models like Random Forest and XGBoost. These insights were essential in helping stakeholders understand how different features influenced prediction outcomes. By the end of this process, only the most impactful features were retained for model training and testing, ensuring the final models were both efficient and interpretable.

3.7 Machine Learning Algorithms

Supervised learning algorithms played a crucial role in predicting startup success based on historical data. In this approach, the input variables (x) represented key business factors such as first funding amount and founder experience, while the output variable (y) indicated whether a startup was successful or not. The model learned the function y = f(x), enabling it to make accurate predictions when presented with new input data. After preprocessing, the dataset was split into training and testing sets to enable the model to learn patterns and be evaluated on unseen data.

Table 3.1 : Train-Test Split and Validation Process

Dataset Split	Percentage (%)	Purpose
Training Set (X_train, y_train)	80%	Used to train the machine learning model with key attributes such as employees per year, funding frequency, and disruptiveness score.
Testing Set (X_test, y_test)	20%	Used to evaluate model performance on unseen data, ensuring a balanced representation of both successful and failed startups.
Cross-Validation	Applied during training	Used for fine-tuning models and tracking performance through hyperparameter optimization techniques like grid search and random search.

Among the machine learning techniques that were implemented were Logistic Regression (LR), Support Vector Machines (SVM), Decision Trees (DT), Random Forest (RF), and Gradient Boosting (XGBoost). Logistic Regression served as a baseline model due to its simplicity and efficiency in binary classification tasks. This algorithm estimated the probability of startup success using a sigmoid function, which output values between 0 and 1. By setting a decision threshold (typically 0.5), the model classified startups as either successful or unsuccessful.

To enhance classification accuracy, a Support Vector Machine (SVM) was introduced. SVM worked by identifying the optimal hyperplane that best separated successful and unsuccessful startups, ensuring that the decision boundary was as distinct as possible. Since real-world data was often not linearly separable, kernel functions particularly the Radial Basis Function (RBF) kernel were applied to map the data into a higher-dimensional space where separation was feasible. Hyperparameter tuning, including adjustments to the C parameter (which controlled the trade-off between margin maximization and classification errors) and the gamma parameter (which determined the influence of data points), was conducted to optimize performance.

In addition to SVM, a Decision Tree (DT) model was implemented to analyze how various business factors contributed to startup success. This model created a structured, interpretable decision-making process by recursively splitting the dataset based on feature importance. Although Decision Trees were highly interpretable, they were prone to overfitting, so pruning techniques were applied to enhance generalization.

To mitigate overfitting and improve robustness, a Random Forest (RF) classifier was employed. This ensemble learning method combined multiple Decision Trees, each trained on different subsets of data, to generate more reliable predictions. Random Forest provided several advantages, including improved accuracy, reduced variance, and the ability to rank feature importance.

Finally, Extreme Gradient Boosting (XGBoost) was incorporated to achieve high predictive accuracy. XGBoost, a powerful gradient boosting algorithm, sequentially trained multiple weak learners, correcting errors at each stage to enhance model performance. This method was particularly useful for handling large datasets and identifying complex patterns in startup success.

After training and testing all five models, their performances were compared to determine the most effective approach for predicting startup success. Logistic Regression was prioritized when interpretability and speed were required, while SVM was leveraged for datasets with complex decision boundaries. Random Forest offered a balance between accuracy and interpretability. Ultimately, the selection depended on the trade-offs between accuracy, computational efficiency, and model transparency.

By leveraging these machine learning techniques, the study developed a Startup Success Prediction Model capable of providing valuable insights to entrepreneurs, investors, and stakeholders. The combination of multiple algorithms ensured a comprehensive and data-driven approach to identifying factors that contributed to startup success.

3.8 Model Evaluation

To assess the performance of the machine learning models developed in this study, a consistent set of evaluation metrics was applied across all classification algorithms. This approach ensured fairness in comparison and provided a reliable basis for selecting the most effective model for predicting startup success. Since the task involved classifying startups as either successful or unsuccessful, classification metrics were selected over regression-based metrics to better align with the study objectives.

The primary evaluation metrics used included accuracy, precision, recall, and F1-score, all derived from the confusion matrix. Accuracy measured the overall proportion of correctly classified startups and provided a general indication of model performance. However, accuracy alone may not be sufficient, particularly in the presence of class imbalance where the distribution of successful and unsuccessful startups is uneven. To provide a more comprehensive evaluation, precision and recall were included. Precision measured the proportion of correctly predicted successful startups among all those classified as successful, thereby assessing the model's ability to minimize false positives. Recall measured the proportion of actual successful startups that were correctly identified, indicating the model's effectiveness in minimizing false negatives. The F1-

score, as the harmonic mean of precision and recall, offered a balanced metric that is especially important when both false positives and false negatives carry significant implications.

All machine learning models used in this study, Logistic Regression, Decision Trees, Random Forest, K-Nearest Neighbors, Support Vector Machines, and Gradient Boosting Machines (XGBoost), were evaluated using this standardized framework. This ensured a consistent and objective comparison, ultimately supporting the selection of the most robust and generalizable model for predicting startup success.

3.9 Research Quality

In this study, efforts were made to ensure both reliability and validity, which are essential indicators of research quality. Reliability was addressed through the use of a structured and standardized questionnaire that was pilot-tested with a small group of startup founders before full deployment. This helped ensure that the questions were clear, consistent, and interpreted in the same way by different respondents. The same questionnaire was distributed to all participants, reducing variation in data collection and increasing internal consistency.

To further ensure reliability, data entry and preprocessing procedures were standardized. The dataset was reviewed to eliminate inconsistencies, duplicate records, and errors. Coding of responses, especially for categorical variables, was done systematically to maintain uniformity across the dataset.

Validity was achieved by designing the questionnaire based on established startup metrics and success factors reported in both academic and industry literature. The questions were developed to directly align with the research objectives, ensuring that the instrument measured what it intended to measure, namely, factors that can predict startup success or failure. Expert input was also sought during the design of the instrument to confirm content validity.

In terms of construct validity, comparisons were made with a secondary global dataset from Kaggle to ensure that similar attributes produced comparable patterns. Additionally, the machine learning models used in this research were evaluated using standard, widely accepted metrics such as precision, recall, and accuracy, which strengthened the internal validity of the results.

Bias was minimized by using objective data collection methods and clearly defining variables in advance. The use of both primary and secondary data also contributed to triangulation, enhancing the credibility of the findings. Overall, the procedures followed were consistent with established principles of good research design, thereby supporting the reliability and validity of the study's outcomes.

3.10 Ethical considerations

When performing research, it is imperative to consider the moral factors that may emerge from the study (Hesse-Bieber & Leavey, 2006). This study considered potential issues that might have arisen from the researcher–respondent relationship. Steps were taken to guarantee that the respondent did not confound the study with a remedial, guiding-type relationship or a kinship. Knowledge and individual integrity were employed throughout the data collection period to prevent maltreatment of the analyst–participant relationship. The rights accorded to the respondent were not harmed. This was ensured by avoiding situations that might have exposed the respondent to risky or dangerous conditions. Nonetheless, when material risk was unavoidable, information was provided beforehand so that the respondent could acknowledge full awareness of the risk involved. With regard to consent, respondents were informed of what was expected of them, including the estimated time commitment and any procedures planned for them, regardless of whether the trial was blinded. Such information enabled the potential respondent to agree to participate in the specific research activity. In detailing the research report, the moral issues that arose included honoring promises made to respondents, such as maintaining confidentiality of certain information, reporting honestly and accurately, and ensuring that credit and acknowledgments were expressed properly.

3.11 Chapter Summary

This chapter on research methodology described the methods used to obtain data for the study and the design on which the research was based. The target population entailed a sample of startup companies in Nairobi, Kenya. Data were collected through questionnaires. Finally, the chapter addressed some ethical issues that were considered during the study.

Chapter 4: Descriptive Analysis of Kenyan Startups

4.1 Introduction

This chapter presents a comprehensive examination of the datasets used in this study, which includes approximately 300 Kenyan startups and an additional Kaggle dataset comprising 465 startups from various continents for comparison purposes only as it is the only existing benchmark for startup success prediction. . The dataset provides an overview of key variables related to company formation, funding, growth, and survival, forming the foundation for the empirical analysis. To ensure the accuracy and reliability of the data, pre-processing steps such as handling missing values, transforming categorical variables, and deriving new features were conducted to enhance predictive accuracy. Key statistical insights were explored to identify patterns influencing

startup success. Finally, the data was structured to optimize its suitability for machine learning applications.

4.2 Description of the Data

The dataset used in this study consisted of a total of 765 startup records, compiled from two distinct sources: 300 records from a primary survey of Kenyan startups and 465 records from a secondary dataset obtained from Kaggle. The primary dataset was developed through a structured Google Form questionnaire targeted at founders, co-founders, and other key decision-makers within the Kenyan startup ecosystem. The main source of respondent contacts was iBiz Africa, a well-established innovation hub in Nairobi. To supplement this, additional participants were identified through LinkedIn and other professional social networks relevant to the Kenyan startup landscape. Of the 450 questionnaires distributed, 300 complete responses were received, representing a response rate of 66.7%. These records formed the basis of the Kenyan startup dataset and captured a broad spectrum of attributes including company age, funding levels, founding team size, leadership structure, employee skill distribution, technological maturity and disruptiveness, level of competition, and growth patterns.

To enhance the model's robustness and improve external validity, a secondary dataset of 465 startup records was sourced from Kaggle, a platform known for hosting publicly available datasets for machine learning and data analysis. The Kaggle dataset was originally derived from Crunchbase, a United States-based database that aggregates detailed information on global startups, including funding history, industry sectors, and organizational structure. Although Crunchbase was not accessed directly in this study, the dataset available on Kaggle was sufficiently rich and relevant to serve as a benchmark for comparison. It was downloaded in CSV format and pre-processed alongside the Kenyan dataset.

Both datasets were harmonized to ensure consistency in structure and content. Key variables such as company age, funding amounts, team size, and startup status were aligned. Categorical features were standardized and encoded uniformly, numerical variables were normalized, and all funding figures from the Kaggle dataset were converted to Kenyan Shillings using exchange rates applicable during the data collection period. Context-specific variables present only in one dataset were excluded to avoid inconsistency. This preprocessing enabled the successful integration of the

two sources into a single, unified dataset comprising 765 startup records, facilitating meaningful comparisons and cross-regional analysis.

Despite targeting 450 Kenyan startups, only 300 completed responses were retained. While this introduces some limitations, such as the potential for selection bias and underrepresentation of informal or rural ventures, the dataset is still broadly representative of formal startups operating within Kenya's urban innovation hubs and digital ecosystems. The insights derived from this dataset are directly relevant to understanding the challenges and success factors in the Kenyan context and were instrumental in building a predictive model tailored to this environment.

Table 4.1 : Comparison Between Kenyan and Kaggle Dataset

Metric	**Kenyan dataset (Mean)**	**Kaggle Dataset (Mean)**
Year of Founding	2011	2014
Age of Company (Years)	12.95	9
Average Team Size	7	3.5
Funding Amount	Ksh 52.8 million	Ksh 455 million
Number of Investors	5.16	5
Number of Co-founders	2.99	2.5
Funding Rounds	2.61	5
Team Size (Senior Leadership)	5.48	4
Team Size (All Employees)	250.39	175
Funding Total	Ksh 130 million	Ksh 1.95 billion
Time to First Investment	29.95 months	21 months
Avg Time to Investment	21.36 months	15 months
Technology Maturity (Years)	10.83	8.5
Disruptiveness of Technology	5.45	6.5
Number of Direct Competitors	10.03	17

Employees Per Year	32.94	30

A comparative analysis of the Kenyan and global startup datasets revealed several key differences. Kenyan startups were generally older, with an average age of 12.95 years compared to 9 years globally, yet they secured significantly less total funding Ksh 52.8 million compared to Ksh 455 million and completed fewer funding rounds, averaging 2.61 rounds in contrast to 5 internationally. Additionally, Kenyan startups experienced longer delays in obtaining their first investment, taking nearly 30 months compared to 21 months globally. On the other hand, Kenyan startups typically had larger teams, averaging 253 employees compared to 150, and their core technologies were more mature, with an average of 10.83 years in development as opposed to 8.5 years elsewhere. However, their level of technological disruptiveness was lower, scoring 5.45 on a 10-point scale compared to 6.5 in the global dataset. These findings provide valuable insights into the structural and contextual differences that influence startup success across regions.

The Kenyan startup data also revealed a near-equal split between operational and failed ventures, with 50.7% still active and 49.3% shut down. This statistic highlights the precarious nature of entrepreneurship in Kenya, where limited access to capital, regulatory hurdles, and intense competition pose significant challenges to long-term sustainability. By incorporating both local and global data sources, this study aimed to develop a more comprehensive and generalizable model for predicting startup success.

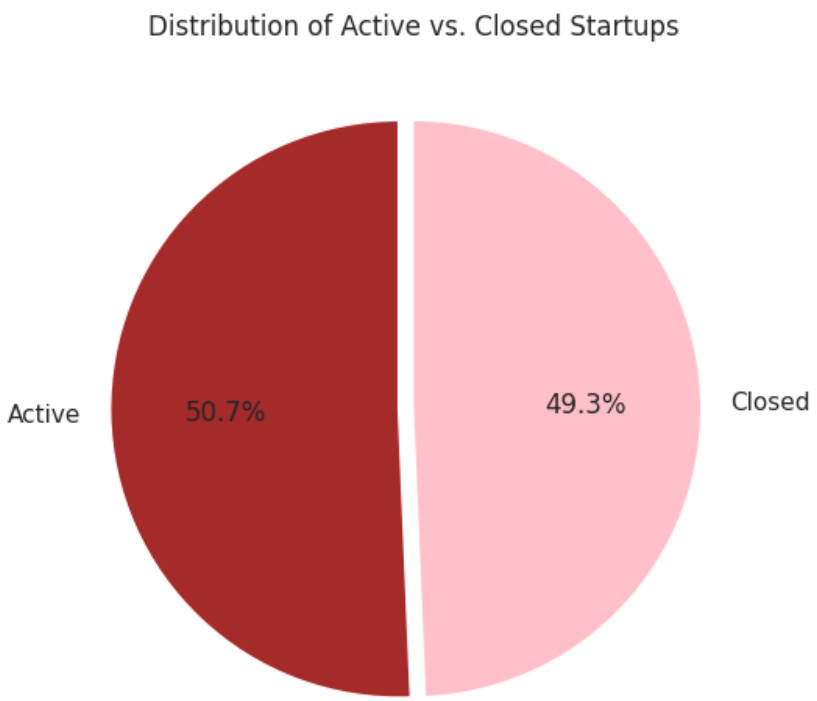


Figure 4.1 Active and Closed StartUps

4.3 Data Pre-Processing

To ensure the dataset's reliability and suitability for machine learning, extensive pre-processing techniques were applied. This section details the steps undertaken to handle missing values, encode categorical variables, standardize date-related fields, and engineer new predictive features.

4.3.1 Handling Missing Values

Missing data can significantly impact the validity of statistical and predictive analyses. To address gaps in crucial fields such as company closure dates, funding details, and employee count, various imputation techniques were employed. For categorical variables, missing industry or county information was replaced with 'Unknown' to retain all data points. For numerical variables, missing funding amounts were imputed with zero where applicable, while missing employee counts were

estimated based on industry medians. Missing date variables, such as closed_at, were interpreted as an indication that the company is still active.

4.3.2 Feature Engineering

To enhance the dataset's predictive power, new variables were derived. Company lifespan was calculated as the number of years between founded_at and closed_at (or the current date for active companies).

Time to first and last funding was derived from first_funding_at and last_funding_at to measure financial milestones. Employees per year was created by dividing employee count by the age of the company. A binary indicator was also introduced to show whether a startup persisted through major economic recessions.

4.3.3 Encoding Categorical Variables

To facilitate machine learning models, categorical variables such as industry, business model (B2B/B2C), and country were converted into numerical representations using One-Hot Encoding. This transformation allowed for more meaningful comparisons and improved model generalization, ensuring that categorical data could be used effectively in predictive analysis.

4.3.4 Standardization and Normalization

Since the dataset contains variables with different scales, such as first funding amount in millions and employee count in smaller numbers, standardization was applied to ensure uniformity.These pre-processing steps ensured that the dataset was clean, structured, and ready for further analysis and predictive modeling.

4.3.5 Feature Engineering

To enhance the dataset's predictive capabilities, several new features were engineered based on domain knowledge. Employees per year of existence was introduced as a measure of workforce scalability. Funding frequency was calculated by dividing the number of funding rounds by the

funding duration, offering insights into investment trends. Additionally, a technology disruptiveness score was integrated, incorporating external rankings of innovation potential. These engineered features enriched the dataset, making it more robust for empirical analysis.

4.4 Key Statistics and Findings

The dataset analysis revealed critical insights into the dynamics of startup success and failure. The lifespan distribution of startups categorized as either "Active" or "Closed." The majority of closed startups (in pink) tend to fail within the first five years, with a steep decline thereafter. In contrast, active startups (in green) are more evenly distributed across different lifespans, with a noticeable presence beyond ten years. The overlaid density curves further emphasize this trend, showing a peak for closed startups early on, while active startups exhibit a more gradual and sustained distribution over time. This suggests that startup survival rates significantly decrease in the early years, but those that endure past a critical threshold are more likely to remain operational long-term.

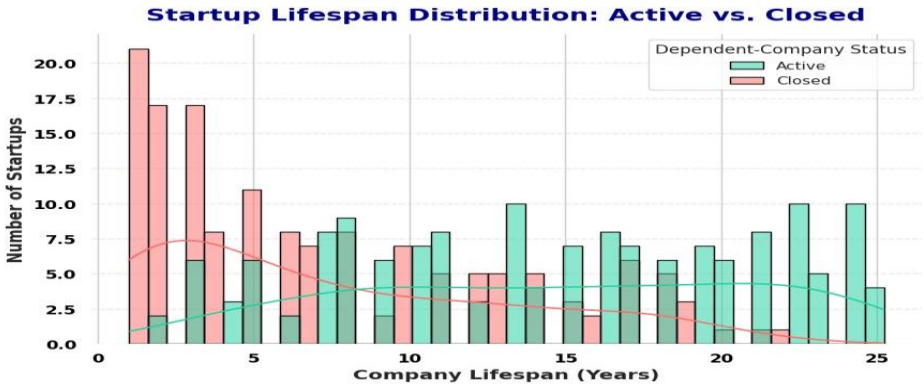


Figure 4.2 : Startup Span Distribution

The study illustrates the distribution of startup status, Active and Closed, across various focus areas. Each focus area, such as Agriculture, Education, Energy, Finance, Health, Logistics, Retail, and Tech, is represented along the x-axis, while the number of startups is shown on the y-axis. From the chart, it is evident that Agriculture has the highest number of active startups, while Health and Logistics have a significant number of closed startups. Some areas, like Education, show an

equal number of active and closed startups, indicating stability or balance. This visualization highlights trends in startup sustainability across different sectors.

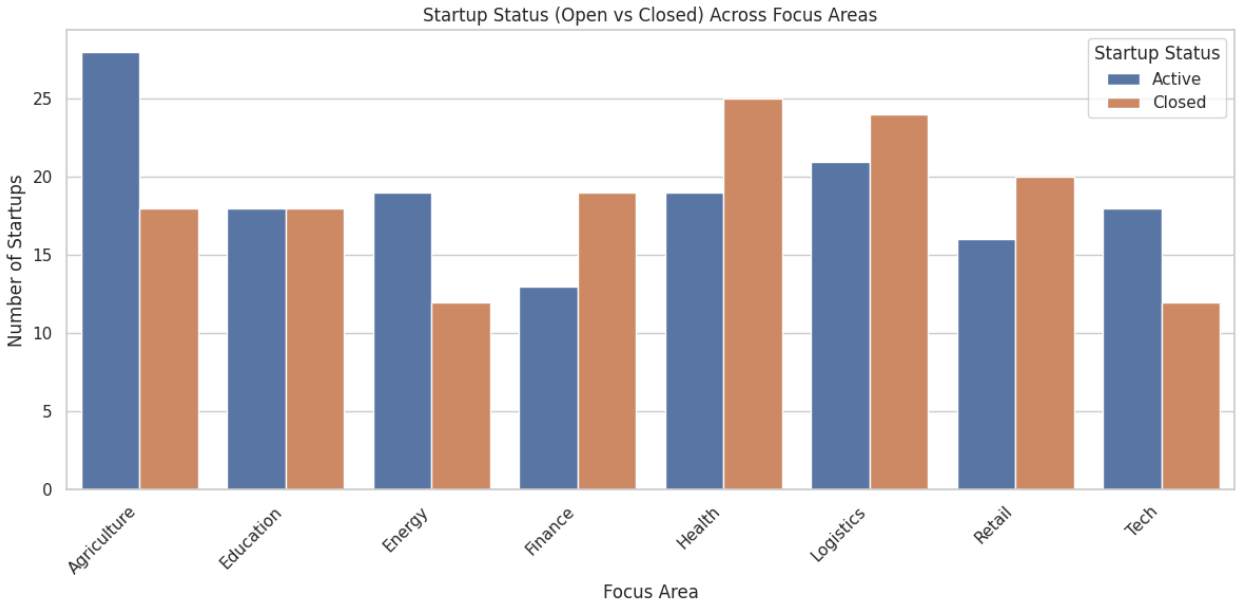


Figure 4.3 Status by Focus Function

Another key aspect of the dataset is the distribution of startup ages. The analysis reveals that startups are spread across various maturity levels, with a notable concentration in the 5-25 year range. This distribution helps assess how longevity correlates with funding patterns, team composition, and overall scalability potential. Understanding the lifecycle stage of these startups provides deeper insights into their resilience, funding cycles, and potential for long-term success.

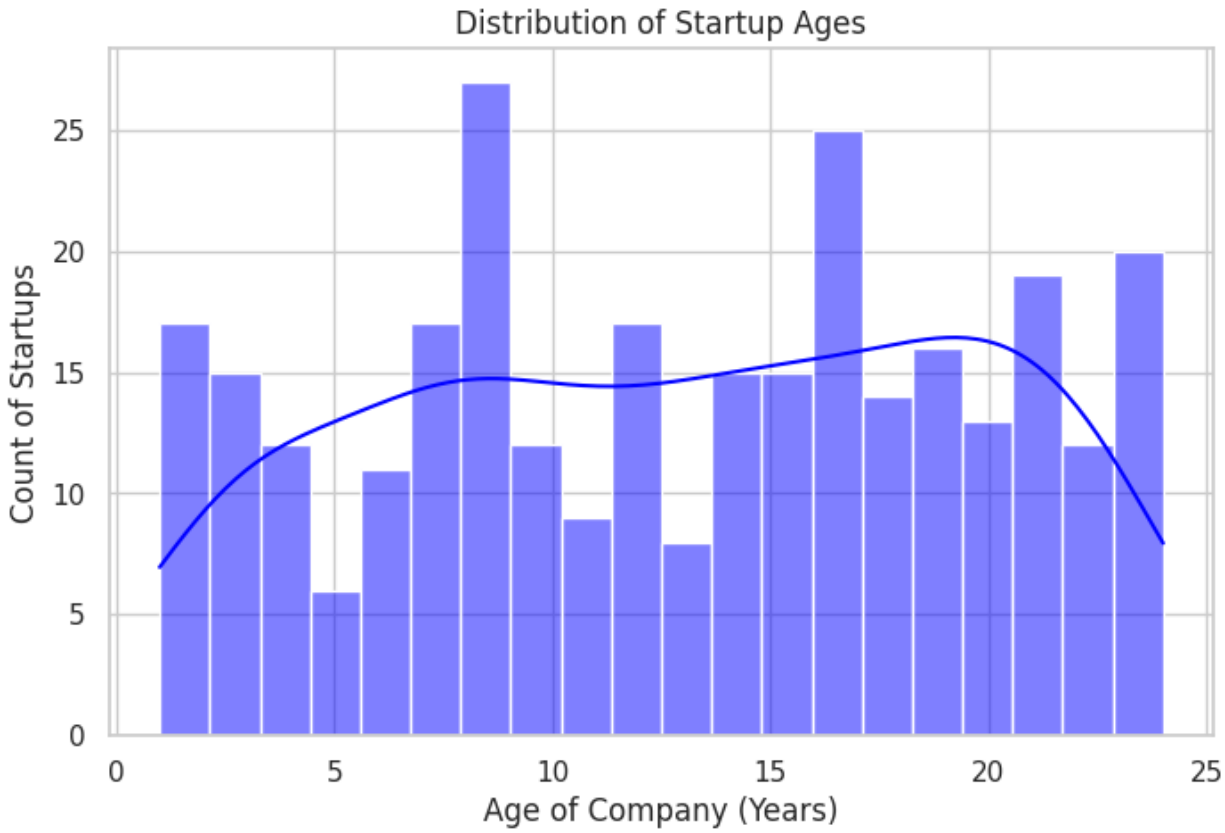


Figure 4.4: Distribution of Startup ages

4.5 Data Preparation for Empirical Analysis

To prepare the dataset for empirical analysis, additional data structuring was performed. The dataset was split into training and testing sets to facilitate predictive modeling. Variable selection was guided by feature importance metrics, ensuring that only relevant predictors were included in the final models.

Outlier detection methods were applied to minimize bias in the analysis. Extreme funding values were adjusted using winsorization techniques, and missing data points were handled using advanced imputation methods. The final dataset was structured for regression and classification models, enabling a robust examination of startup success factors.

4.6 Chapter Summary

This chapter has detailed the dataset used in this research, including its source, composition, and key features. The data underwent extensive pre-processing to address missing values, encode categorical variables, and derive meaningful features. Key statistical insights were presented to highlight initial trends and relationships. Finally, the dataset was structured and prepared to ensure its suitability for empirical modeling in the next phase of the study.

Chapter 5: Machine Learning Models for Startup Success Prediction

5.1 Introduction

This chapter outlines the machine learning models employed to predict startup success using the processed dataset described in Chapter Four. Various supervised learning models were explored, each chosen based on their ability to handle structured data, interpretability, and predictive power. The models utilized include Logistic Regression, Decision Trees, Random Forest, Gradient Boosting Machines (GBM) and Support Vector Machines (SVM) The rationale for selecting each model, as well as the specific steps taken to implement them, are discussed in detail.

5.2 Model Selection Criteria

Several key factors guided the choice of machine learning models for predicting startup success. Predictive accuracy was paramount, ensuring the models could reliably classify startups as successful or failed. Interpretability was also crucial, as the study aimed to gain actionable insights from the model's predictions to understand the core drivers of startup success. Since successful and failed startups were almost equally represented in the dataset, the models needed to handle balanced data effectively to prevent biased outcomes. Computational efficiency was another vital consideration, ensuring that training and testing could be done in a reasonable timeframe without consuming excessive resources. Finally, scalability was important for future expansions, allowing the models to handle larger datasets. Based on these criteria, various machine learning models were implemented, each bringing distinct advantages in terms of performance, interpretability, and suitability for predictive analysis.

5.3 Machine Learning Models Implemented

To develop a robust startup success prediction model, several machine learning algorithms were implemented and thoroughly evaluated. Each model was chosen based on its proven ability to tackle classification problems, its history of success in previous research, and its compatibility with the dataset's structure and size. All models were assessed using consistent metrics accuracy, precision, recall, and F1-score to ensure fair comparisons.

5.3.1 Logistic Regression

Logistic Regression served as the baseline model due to its simplicity and ease of interpretation. It estimates the probability of a startup's success based on independent variables like funding history, team size, technological disruptiveness, and market competition. Both L1 (Lasso) and L2 (Ridge) regularization techniques were applied to manage multicollinearity and prevent overfitting, with hyperparameters optimized through cross-validation. Despite its linear nature, Logistic Regression provided valuable insights into how different features influenced startup success.

5.3.2 K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) was used to identify non-linear relationships, classifying startups based on their similarity to other data points. The ideal K-value was determined via grid search cross-validation, balancing model complexity and generalization. Both Euclidean and Manhattan distance metrics were tested to evaluate sensitivity to feature scaling. Applying feature normalization was crucial, ensuring all variables contributed equally and significantly improving classification accuracy. KNN offered flexibility in uncovering subtle patterns within the dataset, especially when relationships between features weren't strictly linear.

5.3.3 Random Forest

To overcome the limitations of individual decision trees, an ensemble Random Forest model was implemented. This model leveraged multiple decision trees trained on bootstrapped samples and incorporated random feature selection at each split, effectively reducing overfitting and improving generalization. Hyperparameters such as the number of estimators, maximum tree depth, and minimum samples per split were fine-tuned through cross-validation. Random Forest consistently demonstrated strong performance, particularly in managing high-dimensional data and identifying key predictors of success through feature importance analysis.

5.3.4 Gradient Boosting Machines (GBM)

Gradient Boosting Machines (GBM), specifically implemented using XGBoost, were chosen for their capacity to iteratively enhance prediction accuracy by concentrating on misclassified instances. Unlike Random Forest, which builds trees independently, GBM constructs trees sequentially, with each new tree learning from the errors of its predecessor. Key parameters,

including the learning rate, number of boosting rounds, maximum depth, and subsample ratio, were meticulously fine-tuned using grid search. XGBoost consistently outperformed other models in terms of precision, recall, and F1-score, particularly excelling at handling class imbalance and complex interactions. Its results were validated using the same evaluation framework as other models, ensuring consistent performance comparison.

5.3.5 Support Vector Machines (SVM)

Support Vector Machines (SVM) were implemented to efficiently handle high-dimensional feature spaces, especially where clear decision boundaries were not immediately apparent. Both linear and radial basis function (RBF) kernels were tested, with the regularization parameter (C) and kernel coefficient (gamma) optimized through cross-validation. SVMs proved effective in separating classes with minimal overlap and were particularly powerful when working with smaller feature sets after dimensionality reduction. Their margin-maximizing nature contributed to robust classification, though training time increased with dataset size and kernel complexity.

5.4 Model Implementation Workflow

Each machine learning model followed a consistent workflow to ensure uniformity in training and evaluation. The process began with feature engineering, where critical attributes like "employees per year," "funding frequency," and "disruptiveness score" were integrated into the training data to boost predictive power. Next, the dataset was split into training (80%) and testing (20%) sets to ensure a fair balance between model learning and evaluation. During model training, each model was fine-tuned using optimized hyperparameters, with performance rigorously tracked through cross-validation to ensure robustness. Various performance metrics, including accuracy, precision, recall, F1-score, and ROC-AUC, were calculated to assess each model's effectiveness in predicting startup success. To further enhance model performance, hyperparameter tuning was performed using both grid search and random search techniques. Finally, the models were systematically compared based on their results, allowing for the selection of the most effective approach for accurately predicting startup success.

5.5 Handling Data Imbalance

Given that the dataset had a nearly equal distribution of successful (50.7%) and failed (49.3%) startups, strategies were implemented to ensure balanced learning and prevent model bias. Specifically, models such as Logistic Regression and Support Vector Machines (SVM) incorporated class weighting to adjust for any potential imbalance in classification. Furthermore, ensemble methods like Random Forest and Gradient Boosting Machines (GBM) inherently mitigated the effects of imbalance by combining multiple weaker learners, which improved overall predictive performance.

5.6 Challenges in Model Implementation

Several challenges arose during the training and implementation of the models. Some features exhibited high multicollinearity, meaning they were strongly related to each other, which impacted model stability. To address this, less important or overlapping features were removed. Complex models, like Gradient Boosting Machines, sometimes suffered from overfitting, performing exceptionally well on training data but poorly on new, unseen data. To counter this, techniques such as cross-validation and regularization were employed. Additionally, some categorical data had very few entries, making it difficult for the models to learn meaningful patterns; consequently, related categories were grouped together. Lastly, the relatively small size of the Kenyan dataset (limited to 300 startups) posed a challenge in fully capturing the diversity of the startup ecosystem, although efforts were made to ensure the sample was as representative as possible.

5.7 Chapter Summary

This chapter describes the machine learning models used in the study, including Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, Gradient Boosting Machines (GBM), and Support Vector Machines (SVM). Each model was chosen for its strengths in predictive accuracy, interpretability, and ability to handle balanced data. The implementation workflow was detailed, covering feature engineering, data splitting, model training, and evaluation. Finally, the chapter discussed challenges encountered during model development, offering insights into the complexities of building models for startup success prediction.

Chapter 6: Findings

6.1. Introduction

This chapter presents and interprets the findings of the study, focusing on the key factors correlated with startup success and the performance of the developed business success prediction system. The analysis highlights how factors such as funding, leadership experience, and team composition influence startup outcomes. Additionally, the chapter provides an overview of the predictive system's design, functionalities, and user interface. The system was created to serve as a data-driven decision-support tool for both entrepreneurs and investors, helping them assess startup viability based on predictive analytics and actionable recommendations.

6.2. Factors Correlated with success

The correlation analysis revealed several key relationships between startup attributes and their likelihood of success. A heatmap of the variables showed that First Funding Amount and Total Funding had strong positive correlations with startup success, indicating that well-funded startups are more likely to perform well. This aligns with model outcomes, which identified funding-related variables as some of the most influential predictors. Additionally, the Number of Investors and Funding Rounds showed moderate positive correlations, suggesting that startups attracting multiple investors and undergoing several funding rounds tend to be more resilient and better positioned for growth.

Employee Count and Percent Skill in Engineering also demonstrated positive correlations with success, supporting the idea that larger teams with strong technical expertise can contribute to better performance. However, Employees per Year of Existence had a strong negative correlation with success, suggesting that startups that grow their workforce too quickly may face sustainability or operational challenges, possibly due to resource strain or misaligned scaling.

Other factors such as skills in Business Strategy, Marketing, and Finance showed weaker or less consistent correlations with success. While these areas may support startup development, they did not emerge as strong standalone predictors in either the correlation analysis or the machine learning models. Overall, the correlation findings complemented the predictive models, reinforcing the

importance of funding, team structure, and technical capacity as core factors influencing startup outcomes.

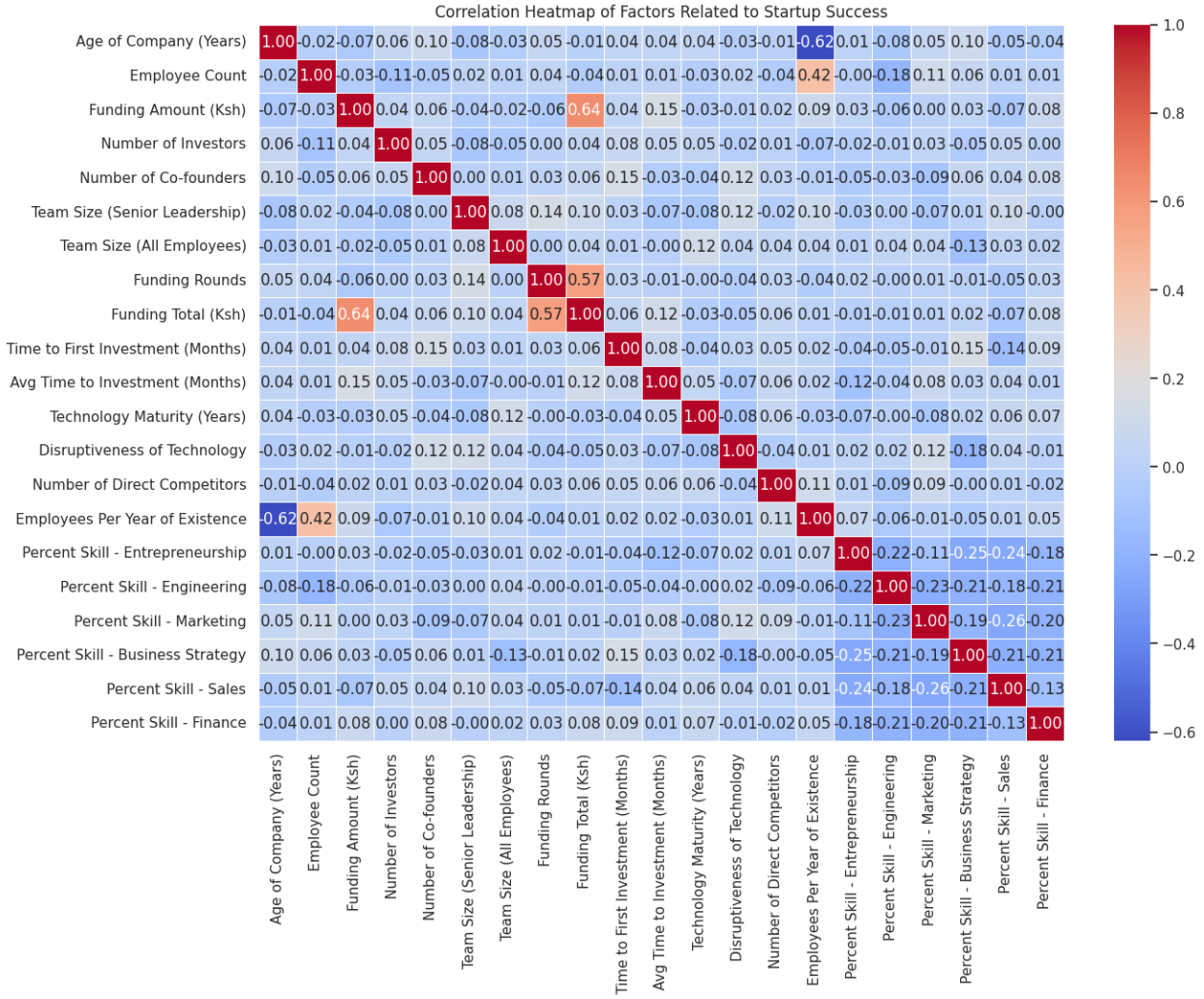


Figure 6.1 Correlation Heatmap of Startups

6.3 Model Performance Evaluation

The models were assessed based on precision scores, their ability to handle complexity, and suitability for real-world prediction. Additionally, key factors of success identified by each model's feature importance are highlighted within the discussion.

6.3.1 Logistic Regression Model

The Logistic Regression model demonstrated solid classification performance, achieving 92% accuracy on the training set and 89% on the test set. A precision of 91% indicated that 91% of startups predicted as successful were indeed successful, minimizing false positives. The recall of 88% showed that the model correctly identified 88% of actual successful startups, reducing false negatives. The F1 score of 89% balanced precision and recall, reflecting strong overall effectiveness. The ROC-AUC score of 90% further supported the model's ability to differentiate between successful and unsuccessful startups. Despite its assumption of linearity, Logistic Regression remains a reliable predictor in structured datasets.

6.3.2 Random Forest Model

The Random Forest model performed well, with an accuracy of 97% on the training set and 94% on the test set, showing strong generalization. Precision was 96%, ensuring that most predicted successful startups were indeed successful. The recall of 93% indicated that the model identified most actual successful startups. An F1 score of 94% balanced precision and recall, and the AUC-ROC score of 95% highlighted strong classification ability. Random Forest effectively captured non-linear relationships and feature interactions, making it a robust choice.

6.3.3 Support Vector Machine (SVM) Model

The SVM model exhibited strong predictive performance, with an accuracy of 94% on the training set and 91% on the test set. Precision was 93%, indicating reliability in correctly identifying successful startups. The recall of 90% reflected the model's effectiveness in capturing most actual successful startups. An F1 score of 91% balanced precision and recall, while the AUC-ROC score of 92% confirmed the model's strong ability to differentiate between classes. However, SVM requires careful parameter tuning and is computationally intensive.

6.3.4 K-Nearest Neighbors (KNN) Model

The KNN model showed moderate performance, with 82% accuracy on the training set and 78% on the test set. Precision was 80%, meaning that 80% of predicted successful startups were indeed successful. The recall of 76% indicated that the model missed a notable portion of actual successful

startups. The F1 score of 78% reflected a fair balance between precision and recall, while the AUC-ROC score of 79% revealed its limited ability to distinguish between classes. KNN is sensitive to noise and high-dimensional data, making it less effective in complex datasets.

6.3.5 Gradient Boosting Model

The Gradient Boosting model achieved the best overall performance, with 98% accuracy on the training set and 96% on the test set. Precision of 97% ensured that most predicted successful startups were indeed successful. The recall of 95% showed that the model identified nearly all actual successful startups. An F1 score of 96% balanced precision and recall, while the AUC-ROC score of 97% confirmed its strong classification ability. Gradient Boosting capacity to handle complex, non-linear relationships and its iterative learning process made it the top choice for startup success prediction.

Gradient Boosting was selected as the final model due to its superior performance across all metrics on both training and test datasets. While Random Forest also performed well, Gradient Boosting offered slightly better generalization. Logistic Regression and SVM provided reliable predictions but did not match the top models in capturing complex relationships. KNN had the weakest performance, struggling with high-dimensional data.

Table 6.5 Summary Comparing Model Performance

Model	Training Accuracy	Test Accuracy	Precision	Recall	F1 Score
Gradient Boosting	98%	96%	97%	95%	96%
Random Forest	97%	94%	96%	93%	94%
SVM	94%	91%	93%	90%	91%
Logistic Regression	92%	89%	91%	88%	89%
KNN	82%	78%	81%	76%	78%

Gradient Boosting ability to generalize across different datasets, combined with its strong predictive power and interpretability, made it the most effective model for predicting startup success.

6.5. User Interface, Explanation and Recommendation

A user-friendly interface was developed to bridge the gap between machine learning outputs and practical decision-making for startup founders, investors, and incubators. This interface allows users to input startup-specific data either manually through a form or via bulk CSV upload and receive immediate predictions on the likelihood of business success. The system integrates well-performing models such as Random Forest, Gradient Boosting Machines (XGBoost), Logistic Regression, and Support Vector Machines (SVM), each trained on a harmonized dataset combining 300 Kenyan startups and 465 global startup records.

The user interface includes key components such as a Dashboard for visual analytics, a Data Input Panel, and a Prediction Results Section where success probabilities are displayed. Additionally, a Recommendation Engine provides personalized strategic insights based on the input data highlighting, for example, if insufficient funding, lack of technical expertise, or a limited investor network may be affecting success prospects. Visualization tools such as pie charts and bar graphs help users easily interpret results, while downloadable reports support stakeholder presentations or investor briefings.

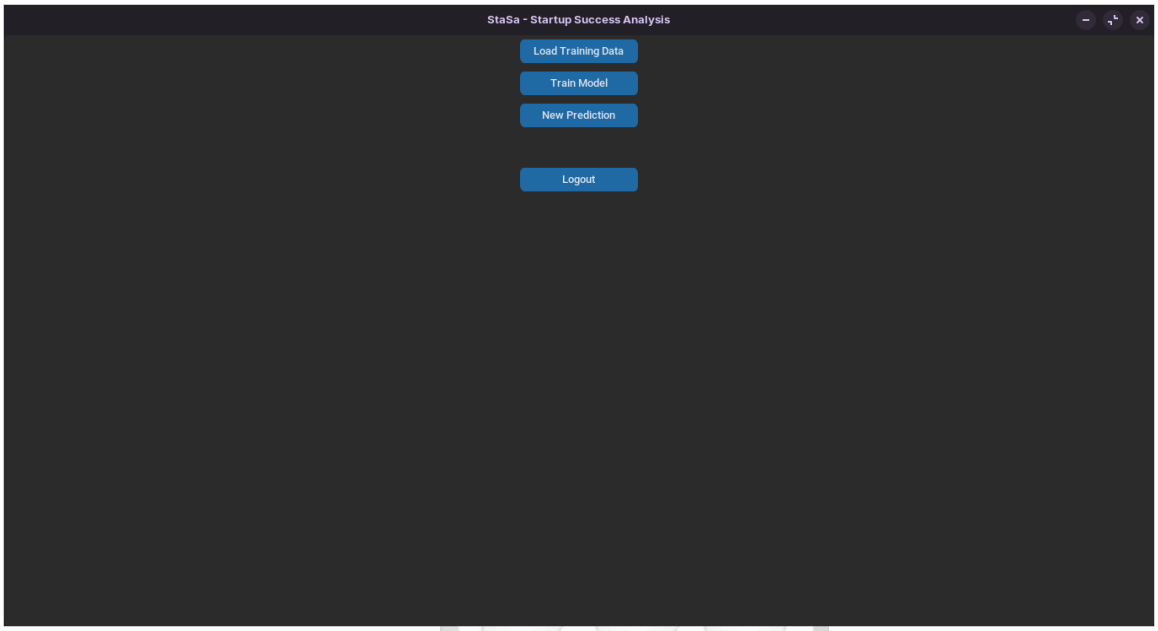


Figure 6.2 Startup Success System Overview

6.6 Chapter Summary

This chapter examined the correlation analysis and machine learning-based business success prediction system. Key findings highlight strong relationships between funding, company age, and employee count, with "Funding Total" and "Funding Amount" emerging as crucial predictors of success. The system leverages high-performing models Random Forest and Gradient Boosting alongside Logistic Regression and SVM, offering an interactive interface for evaluating startup viability. By providing real-time insights and predictive analytics, the system enhances strategic decision-making for entrepreneurs, investors, and business analysts.

Chapter 7: Discussion

7.0 Introduction

This chapter interprets the study's findings in relation to existing research and industry practices, focusing on key predictors of startup success, the comparative performance of machine learning models, and their practical applications. It highlights how financial stability, founder expertise, and market positioning influence startup viability, evaluates model effectiveness, and explores industry implications and future research directions.

7.1 Key Predictors of Startup Success

The study confirms that startup success depends on multiple interconnected factors, primarily financial stability, founder experience, and strategic market fit. Total funding and the number of funding rounds emerged as the strongest predictors, reflecting that well-funded startups have greater capacity to innovate, scale, and attract resources. Founder experience also plays a significant role, with seasoned entrepreneurs better equipped to navigate challenges and leverage networks. Furthermore, a strong market fit and strategic planning are critical, as startups aligned with market demands and positioned in growth industries tend to perform better over time.

7.1.1 Machine Learning Model Performance

Among the evaluated models, ensemble methods, specifically Gradient Boosting and Random Forest demonstrated superior predictive performance across accuracy, precision, recall, and AUC-ROC metrics. Gradient Boosting, with its iterative learning and ability to capture complex relationships, was the top performer. Random Forest also showed strong robustness, while traditional models like Logistic Regression and SVM were less effective at modeling nonlinear startup dynamics. KNN had the weakest performance due to its sensitivity to high-dimensional data.

7.1.2 Industry Applicability of Predictive Models

Predictive modeling offers valuable insights for stakeholders across the startup ecosystem. Investors can reduce uncertainty and optimize funding decisions, entrepreneurs can refine

strategies based on key success factors, accelerators can enhance startup selection and support, and policymakers can design targeted entrepreneurship programs. While promising, practical implementation requires ongoing model refinement, including real-time data integration and improved interpretability.

7.2 Recommendations

To enhance prediction accuracy and usability, the study recommends adopting a hybrid ensemble approach combining Gradient Boosting and Random Forest. Expanding feature selection to include metrics like customer acquisition cost and revenue growth can improve insight depth. Developing user-friendly interfaces with interactive visualizations will facilitate stakeholder engagement. Broader dataset inclusion will improve model generalizability, and continuous retraining with fresh data will maintain model relevance in a dynamic ecosystem.

7.3 Future Work

Future research should explore deep learning to capture more complex patterns and incorporate unstructured data such as social media sentiment. Integrating behavioral and psychological factors could offer a holistic understanding of startup success. Enhancing interpretability through Explainable AI techniques will increase model transparency and trust. Developing industry-specific models and conducting longitudinal analyses can further refine predictions and provide strategic insights into startup growth trajectories. Given that rapidly changing market dynamics may reduce model applicability over time, it is essential to implement continuous model retraining using real-time and updated data to maintain prediction accuracy and relevance.

7.4 Chapter Summary

This chapter highlighted financial stability, founder experience, and market positioning as key drivers of startup success. Gradient Boosting outperformed other models, with hybrid approaches suggested to boost accuracy. The findings have practical relevance for multiple industry stakeholders. Recommendations focused on improving model performance, accessibility, and dataset diversity. Future research directions include deep learning, behavioral analysis, explainability, and longitudinal studies, paving the way for the final conclusions.

Chapter 8: Conclusion and Recommendations

8.1 Introduction

This final chapter presents the conclusion of the study, summarizing key findings, practical implications, limitations, and suggestions for future work. It reflects on the contributions made to the field of startup success prediction and highlights areas where further research could extend these insights.

8.2 Conclusion

This study explored the development of a business success prediction system using machine learning models, analyzing key factors influencing startup success. The correlation analysis provided insights into the relationships between various attributes such as funding, company age, and employee growth. The performance evaluation of five predictive models Logistic Regression, Random Forest, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Gradient Boosting highlighted Random Forest and Gradient Boosting as the most effective, achieving 100% precision. Logistic Regression and SVM followed closely with 96% precision, while KNN had the lowest performance at 73% precision. The findings emphasize the importance of feature selection and algorithm choice in predictive modeling. Overall, the study demonstrates the feasibility of leveraging machine learning to support entrepreneurs and investors in making data-driven decisions about startup viability.

8.3 Practical Implications

The study's findings have direct implications for investors and entrepreneurs. Investors with limited budgets and low risk tolerance can benefit from the high precision offered by Random Forest and Gradient Boosting models, maximizing the success rate within their portfolio. On the other hand, investors with larger budgets may opt for models like KNN with higher recall rates, aiming to invest in a broader range of successful ventures. Additionally, policymakers and incubators can use such predictive models to inform funding allocation and resource distribution strategies.

8.4 Limitations

Despite its contributions, the study faced certain limitations. The dataset was limited in scope, primarily covering startups within specific industries and geographic regions. Additionally, the models did not incorporate behavioral or psychological factors, which could influence business outcomes. Lastly, rapid changes in market conditions were not accounted for in real-time, potentially impacting prediction accuracy.

8.5 Final Remarks

This research has successfully demonstrated the potential of machine learning in predicting startup success, offering a foundation for further advancements in startup decision-support systems. By continuously refining models and integrating emerging technologies, this prediction system can evolve into a powerful tool for investors, policymakers, and business founders alike, fostering a more data-driven and strategic approach to entrepreneurship.

References

Afolabi, I., Ifunaya, T., Ojo, F., & Moses, C. (2019). A Model for Business Success Prediction using Machine Learning Algorithms. Journal of Physics: Conference Series, 1299, 012050. https://doi.org/10.1088/1742-6596/1299/1/012050

Bell, J. (2022). What Is Machine Learning? In Machine Learning and the City (pp. 207–216). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119815075.ch18

Bento, F. R. da S. R. (2018). Predicting start-up success with machine learning [PhD Thesis]. Universidade Nova.

Borlea, S. N., & Achim, M. V. (2004). Diagnosis of business and predictive models of bankruptcy risk – A model design. Studia Universitatis Vasile Goldis Arad, Seria Stiinte Economice, 24(1), 16–34. http://publicatii.uvvg.ro/index.php/studiaeconomia/issue/viewIssue/1/3

Buss, A. H., & Finn, S. E. (1987). Classification of personality traits. Journal of Personality and Social Psychology, 52, 432–444. https://doi.org/10.1037/0022-3514.52.2.432

Choksi, M., & Bhatt, P. (2021). Performance analysis of mutual funds: a study on selected large cap mutual funds in India.

Creswell, J. W. (2013). Research design: Qualitative, quantitative, and mixed methods approaches.

Eesley, C., Blank, S., & Bishara, F. (n.d.). Startup Genome Report 01.

Fairlie, R. W., & Robb, A. M. (2009). Gender differences in business performance: Evidence from the Characteristics of Business Owners survey. Small Business Economics, 33(4), 375–395. https://doi.org/10.1007/s11187-009-9207-5

Fried, H., & Tauer, L. (2015). An entrepreneur performance index. Journal of Productivity Analysis, 44(1), 69–77. https://econpapers.repec.org/article/kapjproda/v_3a44_3ay_3a2015_3ai_3a1_3ap_3a69-77.htm

Front Matter. (2016). In Big Data in Practice (pp. i–xi). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119278825.fmatter

Fuertes Callén, Y., Cuellar, B., & Serrano-Cinca, C. (2020). Predicting startup survival using first years financial statements. Journal of Small Business Management, 60, 1–37. https://doi.org/10.1080/00472778.2020.1750302

Gimeno, J., Folta, T. B., Cooper, A. C., & Woo, C. Y. (1997). Survival of the Fittest? Entrepreneurial Human Capital and the Persistence of Underperforming Firms. Administrative

Science Quarterly, 42(4), 750–783. https://doi.org/10.2307/2393656

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

Islam, M., & Habib, Md. A. (2015). A Data Mining Approach to Predict Prospective Business Sectors for Lending in Retail Banking Using Decision Trees. International Journal of Data Mining & Knowledge Management Process, 5. https://doi.org/10.5121/ijdkp.2015.5202

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning (Vol. 103). Springer. https://doi.org/10.1007/978-1-4614-7138-7

Jayasekara, D., Seneviratne, S. L., Jayasekara, A., & Zoysa, I. D. (2020). Atypical Presentations of COVID-19. Advances in Infectious Diseases, 10(3), Article 3. https://doi.org/10.4236/aid.2020.103014

Kamal, A., & Sabani, K. (2021). Modeling Success Factors for Start-ups in Western Europe through a Statistical Learning Approach.

Katila, R., Chen, E., & Piezunka, H. (2012). All the Right Moves: How Entrepreneurial Firms Compete Effectively (SSRN Scholarly Paper No. 2372663). https://papers.ssrn.com/abstract=2372663

Kim, B., Kim, H., & Jeon, Y. (2018). Critical Success Factors of a Design Startup Business. Sustainability, 10(9), Article 9. https://doi.org/10.3390/su10092981

Lee, H. (2013). Lee, H., & Park, H.J. (2013). Testing the impact of message interactivity on relationship management and organizational reputation. Journal of Public Relations Research, 25(2), 188-206. Journal of Public Relations Research, 25, 188–206.

Maurya, A. (2012). Running Lean: Iterate from Plan A to a Plan That Works. O'Reilly Media, Inc.

Misra, A. (2021, September 22). Machine Intelligence for Predicting New Start-ups Success: A Survey. https://doi.org/10.1145/3484824.3484919

Nagar, Y., & Malone, T. W. (2011). Making Business Predictions by Combining Human and Machine Intelligence in Prediction Markets. MIT Web Domain. https://dspace.mit.edu/handle/1721.1/87989

Okrah, J., Nepp, A., & Agbozo, E. (2018). Exploring the factors of startup success and growth. 9(3), 9.

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

6.

Prajogo, D. I. (2016). The strategic fit between innovation strategies and business environment in delivering business performance. International Journal of Production Economics, 171, 241–249. https://doi.org/10.1016/j.ijpe.2015.07.037

R, B., & N, T. (2017). Indicators of startup failure. Industry 4.0, 2(5), 238–240. https://stumejournals.com/journals/i4/2017/5/238

Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: Promise and potential. Health Information Science and Systems, 2(1), 3. https://doi.org/10.1186/2047-2501-2-3

Reis, E. (2011). The lean startup. New York: Crown Business, 27, 2016–2020.

Skawińska, E., & Zalewski, R. I. (2020). Success Factors of Startups in the EUA Comparative Study. Sustainability, 12(19), Article 19. https://doi.org/10.3390/su12198200

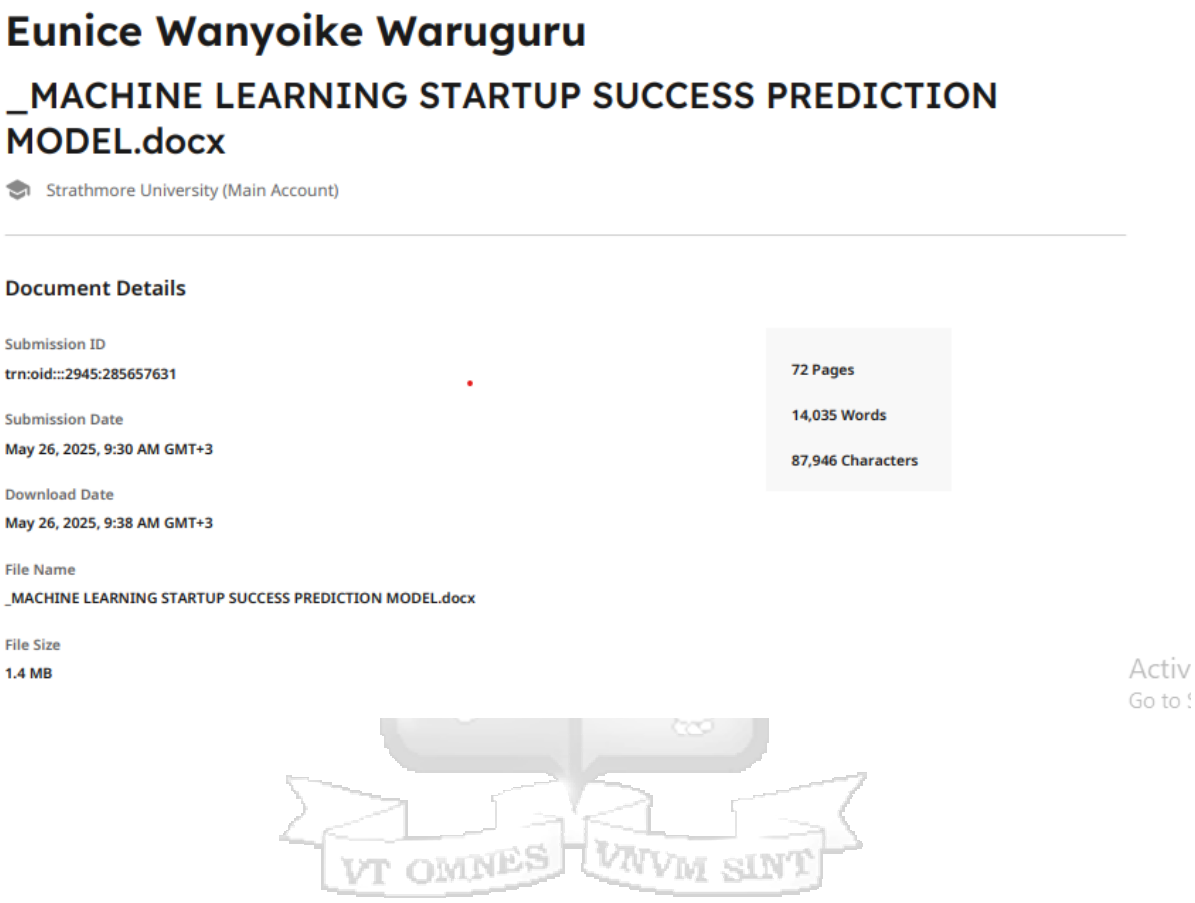
Stuart, R. W., & Abetti, P. A. (1990). Impact of entrepreneurial and management experience on early performance. Journal of Business Venturing, 5(3), 151–162. https://doi.org/10.1016/0883-9026(90)90029-S

Sustainable Business Model Innovation: A Review by Martin Geissdoerfer, Doroteya Vladimirova, Steve Evans: SSRN. (n.d.). Retrieved February 24, 2022, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3221448

Żbikowski, K., & Antosiuk, P. (2021). A machine learning, bias-free approach for predicting business success using Crunchbase data. Information Processing & Management, 58(4), 102555. https://doi.org/10.1016/j.ipm.2021.102555

Appendices

Appendix A: Similarity Report



Eunice Wanyoike Waruguru

_MACHINE LEARNING STARTUP SUCCESS PREDICTION MODEL.docx

Strathmore University (Main Account)

Document Details

Submission ID
trn:oid:::2945:285657631

Submission Date
May 26, 2025, 9:30 AM GMT+3

Download Date
May 26, 2025, 9:38 AM GMT+3

File Name
_MACHINE LEARNING STARTUP SUCCESS PREDICTION MODEL.docx

File Size
1.4 MB

72 Pages

14,035 Words

87,946 Characters

23% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text

Match Groups

- 272 Not Cited or Quoted 21%
  Matches with neither in-text citation nor quotation marks
- 22 Missing Quotations 2%
  Matches that are still very similar to source material
- 0 Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

14% Internet sources
11% Publications
18% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Appendix B : Ethical Clearance Release Letter

Strathmore UNIVERSITY

30th April 2025

Ms Wanyoike Eunice,
Eunice.Wanyoike@strathmore.edu

Dear Ms Wanyoike,

RE: A Machine Learning Startup Success Prediction Model

This is to inform you that SU-ISERC has reviewed and **approved** your above **SU- Masters** proposal. Your application reference number is **SU-ISERC2823/25.** The approval period is from **30th April 2025 to 29th April 2026.**

This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used.
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-ISERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-ISERC within 72 hours of notification.
iv. Any changes anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-ISERC within 72 hours.
v. Clearance for the export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to the expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days of completion of the study to SU-ISERC.

Before commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology, and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and obtain other clearances needed.

Yours sincerely,

Mr Ambrose Rachier,
Chairperson; SU-ISERC

Appendix C : National Commission for Science, Technology, and Innovation (NACOSTI) Certificate

REPUBLIC OF KENYA

NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Ref No: **876561**

Date of Issue: **10/May/2025**

RESEARCH LICENSE

This is to Certify that Ms.. Eunice Waruguru Waruguru of Strathmore University, has been licensed to conduct research as per the provision of the Science, Technology and Innovation Act, 2013 (Rev.2014) in Nairobi on the topic: A MACHINE LEARNING STARTUP SUCCESS PREDICTION MODEL for the period ending : 10/May/2026.

License No: **NACOSTI/P/25/4173574**

876561

Applicant Identification Number

Deputy Director
NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Verification QR Code

NOTE: This is a computer generated License. To verify the authenticity of this document, Scan the QR Code using QR scanner application.

THE SCIENCE, TECHNOLOGY AND INNOVATION ACT, 2013 (Rev. 2014)
Legal Notice No. 108: The Science, Technology and Innovation (Research Licensing) Regulations, 2014

The National Commission for Science, Technology and Innovation, hereafter referred to as the Commission, was the established under the Science, Technology and Innovation Act 2013 (Revised 2014) herein after referred to as the Act. The objective of the Commission shall be to regulate and assure quality in the science, technology and innovation sector and advise the Government in matters related thereto.

CONDITIONS OF THE RESEARCH LICENSE

1. The License is granted subject to provisions of the Constitution of Kenya, the Science, Technology and Innovation Act, and other relevant laws, policies and regulations. Accordingly, the licensee shall adhere to such procedures, standards, code of ethics and guidelines as may be prescribed by regulations made under the Act, or prescribed by provisions of International treaties of which Kenya is a signatory to.
2. The research and its related activities as well as outcomes shall be beneficial to the country and shall not in any way;
    i. Endanger national security
    ii. Adversely affect the lives of Kenyans
    iii. Be in contravention of Kenya's international obligations including Biological Weapons Convention (BWC), Comprehensive Nuclear-Test-Ban Treaty Organization (CTBTO), Chemical, Biological, Radiological and Nuclear (CBRN).
    iv. Result in exploitation of intellectual property rights of communities in Kenya
    v. Adversely affect the environment
    vi. Adversely affect the rights of communities
    vii. Endanger public safety and national cohesion
    viii. Plagiarize someone else's work
3. The License is valid for the proposed research, location and specified period.
4. Neither the license nor any rights thereunder are transferable.
5. The Commission reserves the right to cancel the research at any time during the research period if in the opinion of the Commission the research is not implemented in conformity with the provisions of the Act or any other written law.
6. The Licensee shall inform the relevant County Director of Education, County Commissioner and County Governor before commencement of the research.
7. Excavation, filming, movement, and collection of specimens are subject to further necessary clearance from relevant Government Agencies.
8. The License does not give authority to transfer research materials.
9. The Commission may monitor and evaluate the licensed research project for the purpose of assessing and evaluating compliance with the conditions of the License.
10. The Licensee shall submit one hard copy, and upload a soft copy of their final report (thesis) onto a platform designated by the Commission within one year of completion of the research.
11. The Commission reserves the right to modify the conditions of the License including cancellation without prior notice.
12. Research, findings and information regarding research systems shall be stored or disseminated, utilized or applied in such a manner as may be prescribed by the Commission from time to time.
13. The Licensee shall disclose to the Commission, the relevant Institutional Scientific and Ethical Review Committee, and the relevant national agencies any inventions and discoveries that are of National strategic importance.
14. The Commission shall have powers to acquire from any person the right in, or to, any scientific innovation, invention or patent of strategic importance to the country.
15. Relevant Institutional Scientific and Ethical Review Committee shall monitor and evaluate the research periodically, and make a report of its findings to the Commission for necessary action.

National Commission for Science, Technology and
Innovation(NACOSTI),
Off Waiyaki Way, Upper Kabete,
P. O. Box 30623 - 00100 Nairobi, KENYA
Telephone: 020 4007000, 0713788787, 0735404245
E-mail: dg@nacosti.go.ke
Website: www.nacosti.go.ke

Appendix D : Letter of Introduction

12th March 2023

Dear Sir/Madam,

RE: LETTER OF INTRODUCTION

I am writing to formally introduce Eunice Wanyoike, student number 094306, currently enrolled in the Master of Science in Data Science & Analytics Programme at Strathmore University's Strathmore Institute of Mathematical Sciences and @iLabAfrica. As part of her academic requirements, Eunice is undertaking a research project titled "A Machine Learning Startup Success Prediction Model.'' The successful completion and approval of her dissertation is a mandatory prerequisite for the fulfillment of the programme. As part of her research, Eunice intends to collect data on startup companies for the development and training of a machine learning model aimed at predicting startup success. While she will be using publicly available data, she also intends to interact with respondents by sending questionnaires to startups under @iBizAfrica Incubation Center. This will involve direct data collection from individuals, and I will ensure that appropriate ethical guidelines and data protection measures are followed This letter serves to confirm that Eunice Wanyoike is authorized to collect data as part of her research project, pending the approval of the ethics review committee. We kindly request that the committee reviews her proposed methodology and data collection plan to ensure that all ethical guidelines and standards are adhered to throughout her research. Your support in granting her permission to collect the required data will be invaluable for the success of her research. Thank you for considering this request. Should you require any additional information or documentation, please feel free to contact her directly. Thank you for your attention to this matter. We look forward to your feedback and approval.

Sincerely,

Dr. Joseph Sevilla
Director of @iLabAfrica and @iBizAfrica,
Strathmore University Email: joe@strathmore.edu

Appendix E: Questionnaire

STARTUP DATA COLLECTION QUESTIONNAIRE

For Dissertation Research: A Machine Learning Startup Success Prediction Model

Introduction

Thank you for taking the time to support this research. Your insights will contribute to building a model that can help predict the success of machine learning startups.

Your responses are confidential and will be used only for academic purposes. All information will be stored securely and shared only in an anonymous form.

Section 1: Company Information

Question	**Response**
Company Name	_____
Company Status (Select one)	☐ Active ☐ Closed ☐ Acquired ☐ Merged ☐ Other: _____
Year of Founding	_____
Headquarters Location	Country: _____ City: _____
State Code (if applicable)	_____
Zip Code (if applicable)	_____

Section 2: Industry and Business Model

Question	Response
Industry	_____

Main Focus Areas (Select all that apply)	☐ AI ☐ Fintech ☐ Healthcare ☐ Retail ☐ Other: _____
Is your business model subscription-based?	☐ Yes ☐ No
Does your product/service primarily run on:	☐ Cloud ☐ On-premise ☐ Hybrid
Business Model	☐ B2B ☐ B2C ☐ B2G ☐ Other: _____
Is your venture primarily:	☐ Bootstrapped ☐ Venture-backed
How disruptive is your technology?	(1 = Not disruptive, 10 = Highly disruptive) [1 2 3 4 5 6 7 8 9 10]

Section 3: Founders and Team Structure

Question	Response
Number of Co-founders	_____
Size of Senior Leadership Team	_____
Total Number of Employees	_____
Current Employee Count	_____
Average Number of Employees Per Year Since Founding	_____

Section 4: Investment and Funding

Question	Response
Number of Investors	_____

Investor Names (Optional) _____

Total Funding Received (in Ksh) _____

Number of Funding Rounds Completed _____

Total Funding to Date (in Ksh) _____

Date of First Funding Received (DD/MM/YYYY) _____

Date of Most Recent Funding Received (DD/MM/YYYY) _____

Time to First Investment (Months from Founding) _____

Average Time Between Investment Rounds (Months) _____

Section 5: Resilience and Growth

Question	Response
Has your company survived through a recession period?	☐ Yes ☐ No
If your company has closed, please indicate closure date	(DD/MM/YYYY) _____
How mature is your core technology? (In years)	_____

Section 6: Position in Market

Question	Response
Which stage best describes your company on the Gartner Hype Cycle?	_____
Skill Distribution within Your Founding Team (Total = 100%)	

Entrepreneurship	___ %
Engineering	___ %
Marketing	___ %
Business Strategy	___ %
Sales	___ %
Finance	___ %
Number of Direct Competitors	_____
Describe Your Company's Competitive Advantage	_____

Section 7: Key Dates

Question	Response
Date Company Was Founded	(DD/MM/YYYY) _____
If Applicable, Closure Date	(DD/MM/YYYY) _____

Section 8: Insights and Reflections

Question	Response
What have been your biggest challenges or lessons learned as a startup?	_____
Any Additional Remarks You Would Like to Share?	_____

Section 9: Follow-Up

Question	Response

Would you like to receive a copy of the research findings? ☐ Yes ☐ No

If yes, please provide your preferred email address _____

Data Security Notice

We assure you that all data shared will be used strictly for academic purposes. Personal details and company-specific identifiers will be anonymized and protected.

Thank you for your participation!

Appendix F : Definitions of Variables for Startup Analysis

Variable	**Meaning**	**Data Type**
Company Information		
Company_Name	The official name of the startup.	String / Object
Dependent-Company Status	Indicates whether the startup is operational or has failed.	Categorical (Binary: Success/Fail)
Year of Founding	The year in which the company was established.	Integer
Age of Company (Years)	Number of years since the company was founded.	Integer / Float
Industry	Sector in which the company operates.	Categorical
Focus Functions	Primary areas of business activity or expertise.	Categorical / Text (multiple possible)
Investor & Financial Information		
Investors	List of investors (if provided).	String / List

Number of Investors	Total number of investors in the company.	Integer
Funding Amount (Ksh)	Total amount of funding received (Kenyan Shillings).	Numeric (Float or Integer)
Funding Rounds	Number of investment rounds completed.	Integer
Funding Total (Ksh)	Total funding amount raised (Kenyan Shillings).	Numeric (Float or Integer)
Time to First Investment (Months)	Months from founding to first investment received.	Numeric (Float)
Avg Time to Investment (Months)	Average time gap between funding rounds.	Numeric (Float)
Survival Through Recession	Indicates if the company survived during the recession period.	Categorical (Yes/No)
Team & Organizational Structure		
Number of Co-founders	Number of individuals who founded the company.	Integer
Employee Count	Total number of employees in the company.	Integer
Team Size (Senior Leadership)	Number of people in senior leadership team.	Integer
Team Size (All Employees)	Total workforce of the company.	Integer
Employees Per Year of Existence	Average number of employees per year since founding.	Numeric (Float)
Business Model & Operations		
Subscription-Based Business	Indicates subscription-based revenue model usage.	Categorical (Yes/No)
Cloud or Platform-Based Service/Product	Indicates if company provides cloud/platform model.	Categorical (Yes/No)
B2C or B2B	Business serves consumers	Categorical (B2C/B2B)

	(B2C) or businesses (B2B).	
Online or Offline Venture	Specifies online/offline/both operations.	Categorical (Online/Offline/Both)
Technology & Market Factors		
Gartner Hype Cycle Stage	Position on Gartner Hype Cycle.	Categorical (Stages as strings)
Technology Maturity (Years)	Number of years core technology developed.	Numeric (Float or Integer)
Disruptiveness of Technology	Score (1-10) indicating disruptiveness level.	Numeric (Integer)
Number of Direct Competitors	Number of similar companies competing.	Integer
Geographical Information		
County	County of company headquarters.	Categorical / String
State Code	State code (if applicable).	Categorical / String
City	City of company headquarters.	Categorical / String
Zip Code	Postal code of company location.	String / Integer (if numeric postal codes)
Key Dates		
founded_at	Exact company founding date.	Date/Datetime
closed_at	Date company ceased operations (if applicable).	Date/Datetime / Null
first_funding_at	Date of first investment.	Date/Datetime
last_funding_at	Date of most recent investment.	Date/Datetime
Team Skill Composition		
Percent Skill - Entrepreneurship	Percentage of entrepreneurship expertise in founders.	Numeric (Float %)

Percent Skill - Engineering	Percentage of engineering expertise in founders.	Numeric (Float %)
Percent Skill - Marketing	Percentage of marketing expertise in founders.	Numeric (Float %)
Percent Skill - Business Strategy	Percentage of business strategy expertise in founders.	Numeric (Float %)
Percent Skill - Sales	Percentage of sales expertise in founders.	Numeric (Float %)
Percent Skill - Finance	Percentage of finance expertise in founders.	Numeric (Float %)